UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD ECONOMÍA Y PLANIFICACIÓN

"VALIDACIÓN DEL MODELO ESTIMADOR DE INGRESOS PARA LA BANCA MINORISTA"

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERA ESTADÍSTICA INFORMÁTICA

KENIA NORA CALISAYA MALLCO

LIMA - PERÚ

2024

VALIDACIÓN DEL MODELO ESTIMADOR DE INGRESOS PARA LA BANCA MINORISTA

INFORME DE ORIGINALIDAD

12%	12%	3%	7%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	fedefliguer.github.io Fuente de Internet	2%
2	eio.usc.es Fuente de Internet	1%
3	repositorio.udec.cl Fuente de Internet	1%
4	repositorio.umsa.bo Fuente de Internet	1%
5	online-tesis.com Fuente de Internet	1%
6	www.auditcontsa.com Fuente de Internet	1%
7	repositorio.uchile.cl Fuente de Internet	<1%
8	pdfs.semanticscholar.org Fuente de Internet	<1%
9	revistas.udea.edu.co Fuente de Internet	

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD ECONOMÍA Y PLANIFICACIÓN

"VALIDACIÓN DEL MODELO ESTIMADOR DE INGRESOS PARA LA BANCA MINORISTA"

PRESENTADO POR
KENIA NORA CALISAYA MALLCO

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE INGENIERA ESTADÍSTICA INFORMÁTICA

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

……	……
Dr. Jaime Carlos Porras Cerrón	Mg. Sc. Jesus Eduardo Gamboa Unsihuay
PRESIDENTE	ASESOR
……	..……
Dr. Carlos López de Castilla Vásquez	Dra. Frida Rosa Coaquira Nina
MIEMBRO	MIEMBRO

LIMA – PERÚ
2022

DEDICATORIA

Para mis padres que siempre me inculcaron que con empeño y dedicación todo se puede, haciendo posible la culminación del TSP. A mi novio que siempre estuvo acompañándome en las amanecidas, aunque a veces se quedaba dormido. A mi Toffy, que siempre estuvo a mi lado, durmiendo, pero acompañándome.

AGRADECIMIENTO

A mi familia por siempre preocuparse por cómo iba avanzando.

A mi novio Heder que ha estado conmigo en todo momento por su apoyo constante y preocupación.

A mi amiga Yaquelin, que constantemente me motivaba a seguir avanzando y conseguir la ansiada titulación.

A mi asesor Jesús Eduardo que, sin su ayuda, sus conocimientos y su apoyo no hubiera sido posible el desarrollo del presente documento de TSP.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ANEXOS

RESUMEN

El presente trabajo de suficiencia profesional, describe la evaluación de los controles de calidad de datos, la metodología de cálculo aplicada y los procedimientos de implementación de la Calibración del Modelo Estimador de Ingresos para Dependientes + RCC de la Banca Minorista; con el fin mitigar un incorrecto uso de la metodología y mejorar los modelos; además de cumplir con las normas que exige la Política de Gestión de Riesgos basados en las mejores prácticas conforme lo exige la Norma Internacional para la Práctica Profesional de la Auditoría Interna. Para ello, se revisaron el adecuado funcionamiento de los controles de calidad, se realizó la réplica del modelo machine learning basado en árboles llamado XGBoost. Utilizando los softwares SQL, SAS y Rstudio. En ese sentido, se concluye que el modelo cumple con los estándares establecidos por las mejores prácticas, en auditoría de Validación a la implementación. Además, dicho modelo fue implementado de forma adecuada siguiendo los lineamientos definidos por las unidades del banco; sin embargo, en los procedimientos metodológicos de la implementación del modelo se encontraron deficiencias, los mismos que fueron comunicados y subsanados por las áreas o unidades correspondientes.

Palabras Clave: Calibración, Dependientes, Reporte Consolidado Crediticio (RCC), Machine Learning, XGBoosst.

ABSTRACT

The present work of professional sufficiency describes the evaluation of the data quality controls, the calculation methodology applied and the implementation procedures of the Calibration of the Income Estimator Model for Dependents + RCC of Retail Banking; in order to mitigate incorrect use of the methodology and improve the models; in addition to complying with the standards required by the Risk Management Policy based on best practices as required by the International Standard for the Professional Practice of Internal Auditing. To this end, the proper functioning of quality controls was reviewed, and the tree-based machine learning model called XGBoost was replicated. Using SQL, SAS and Rstudio software. In this sense, it is concluded that the model complies with the standards established by the best practices, in the implementation validation audit. In addition, this model was adequately implemented following the guidelines defined by the bank's units; However, deficiencies were found in the methodological procedures of the implementation of the model, which were reported and corrected by the corresponding areas or units.

Keywords: Calibration, Dependents, Consolidated Credit Report (RCC), Machine Learning, XGBoosst.

I. INTRODUCCIÓN

El sector bancario es uno de los más representativos del Sistema Financiero en el Perú, debido a que juega un papel fundamental en la economía, desarrollo del país y de las personas; además, funciona como intermediario financiero que canaliza los recursos excedentes de las personas (como los productos de ahorro y cuentas corrientes), en préstamos para atender las diferentes necesidades de los clientes, de consumo, compra de vivienda, carro, viajes, estudios entre otros.

En ese sentido, y debido a la coyuntura actual, el sistema financiero bancario ha sufrido diferentes cambios a lo largo de los últimos 3 años, tiempo en el que el mundo ha sufrido efectos por la pandemia del COVID -19. Para afrontar los efectos de la pandemia, las empresas que se encuentran en el sector bancario han optado por mejorar los aspectos financiero-contables de las operaciones creando procesos de implementación, validación y actualización, pero sin dejar de lado procesos que van más allá de una transacción operacional como: mejorar los modelos de clasificación de un cliente, validar los modelos de predicción de los ingresos de un cliente que trabaja en un determinado sector, calibrar los modelos de predicción del *Ratio Default* de un préstamo, entre otros. En relación con esto, muchos bancos cuentan con un área de auditoría interna independiente que evalúa las mejoras e implementaciones.

"La auditoría interna es una actividad independiente y objetiva de aseguramiento y consulta, concebida para agregar valor y mejorar las operaciones de la organización, al ayudarla a cumplir sus objetivos, aplicando un enfoque sistemático y disciplinado para evaluar y mejorar la eficacia de los procesos de gestión de riesgos, control y gobierno" (Instituto de Auditores Internos, 2017). En tal sentido, una auditoría basada en riesgos ayuda a identificar deficiencias y falencias que se puedan tener en los procesos de gestión de riesgos; además, de la revisión y cuestionamiento de la metodología usada en cada uno de los procesos. Asimismo, dará a la alta dirección la seguridad de que los procesos mencionados se realizan de forma efectiva y correcta.

Desde hace más de 50 años, opera en el Perú una entidad bancaria que ofrece una gran gama de productos y servicios para personas, pymes y empresas, como también para entidades gubernamentales, microfinancieras y organismos internacionales. Su división de banca mayorista abarca las operaciones de banca de negocios, corporativa e institucional. Sus servicios y productos incluyen leasing, financiamiento corporativo, comercio exterior y

servicios para empresas. El banco también entrega una serie de servicios de banca minorista tales como créditos hipotecarios, tarjetas de crédito, productos de ahorro, y productos para pymes incluyendo crédito rotativo, crédito fijo, leasing y financiamiento de activos fijos.

El banco tiene como una de sus aspiraciones ser referente regional en gestión empresarial, potenciando el liderazgo histórico y transformador de la industria financiera; por lo que, es de vital importancia el mantenerse en la vanguardia de las transformaciones y cambios, así como nuevos avances tecnológicos y aplicaciones de técnicas avanzadas de análisis de datos. Además de validar las aplicaciones metodológicas en el empleo de nuevas técnicas estadísticas avanzadas e incluso verificar el cumplimiento de supuestos mínimos necesarios que exigen las técnicas estadísticas clásicas.

En tal sentido, en líneas anteriores se presenta una de las aspiraciones del banco, que con la colaboración de profesionales que cuentan con las habilidades y competencias técnicas adquiridas, se logra cumplir. Es por ello que durante mi permanencia en el banco pertenecí al equipo de Auditoría de Modelos de Riesgos que validaba la implementación de modelos estadísticos.

1.1 Problemática

A lo largo de la historia del ser humano, han ocurrido diferentes pandemias, donde las más relevantes han sido la Peste Negra (1348-1400), la Gripe española (1918-1920), el VIH/SIDA, el SARS, la Gripe porcina, la Gripe aviar, Ébola, Zika,y la más reciente, la provocada por el virus SARS-COV-2 conocida como la COVID-19 (Huremović,2019; Tisdell,2020). Particularmente, en el siglo XXI se han presentado brotes epidémicos y pandémicos cuyo impacto sobre la economía ha sido importante, sobre todo por sus características de zoonosis, lo cual amerita que la academia, la ciencia y la tecnología se unan para producir respuestas efectivas que reduzcan las consecuencias en el ámbito socioeconómico y de salud pública (Dipaola,2020; Huremović, 2013).

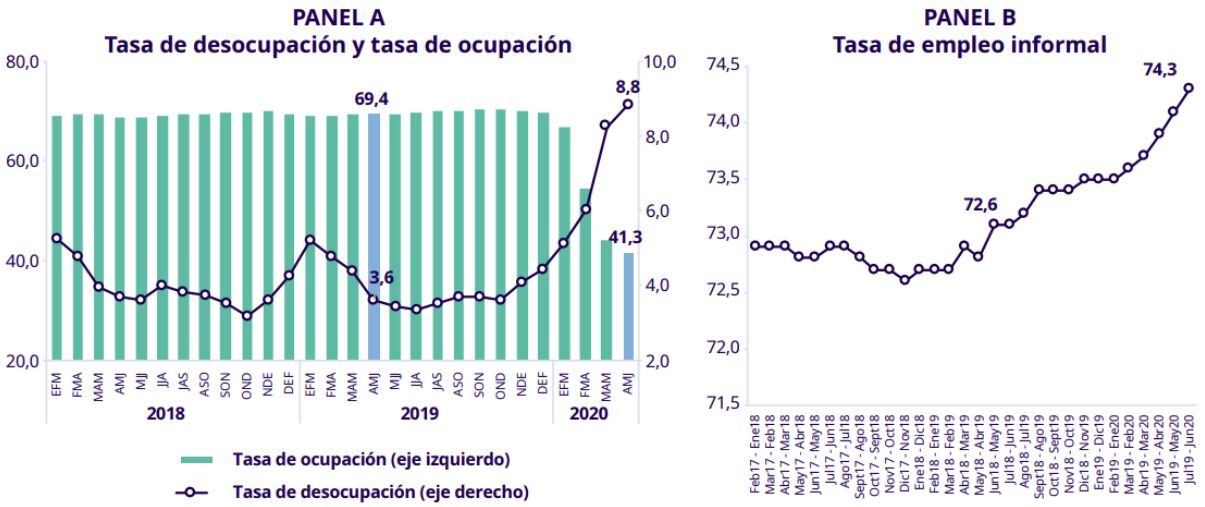
El primer caso de presencia del virus SARS-COV-2 en el Perú fue reportado oficialmente el 6 de marzo del 2020 y el 25 del mismo mes se promulgó el Decreto Supremo N.º 094-2020-PCM, el cual estableció "*las medidas de aislamiento social destinadas hacia una nueva convivencia social; y que prorrogó el Estado de Emergencia Nacional por las graves circunstancias que*

afectan la vida de la Nación a consecuencia del COVID-19". Estas medidas sacaron a la luz, no solo las graves deficiencias del sistema de salud sino también cuál debía ser el papel de la industria y la sociedad para hacer frente a la pandemia (Maguiña, 2020).

Debido al impacto de la COVID-19, con el paso del tiempo y bajo las medidas de aislamiento social se vieron afectados todos los sectores económicos (agricultura, ganadería, minería, educación, entre otros); y sobre todo el sector financiero. En ese sentido, dicho sector tuvo que tomar nuevas medidas ante el reciente panorama económico; ya que el comportamiento social y económico de una persona sufrió muchos cambios como consecuencia del aislamiento social; como se puede observar en la Figura1, en el segundo trimestre del 2020 (meses donde impactó mucho las medidas de aislamiento social y la crisis sanitaria) más de 6 millones de personas perdieron su empleo y aumentó la precariedad laboral; en ese mismo periodo, la tasa de ocupación cayó en 28 pp (puntos porcentuales) y la tasa de desocupación subió en más de 5 pp.

Figura 1

Desocupación, Ocupación y Empleo Informal, Trimestre Móvil EFM2018-AMJ2020



Fuente: INEI, Encuesta Nacional de Hogares
Nota: El gráfico B se encuentra en años móviles

Como consecuencia de estos cambios, el sector bancario estuvo obligado a tomar medidas de cambio en los modelos con los que cuenta (predicción, clasificación, segmentación, entre otros) para los servicios bancarios que brinda a personas, pymes y empresas, como también para entidades gubernamentales, microfinancieras y organismos internacionales. Frente a este contexto, y los nuevos cambios que ocurrían en la economía, empleo y cierres de empresas;

estos modelos necesitaban una actualización y/o cambio; ya que las condiciones para emplearlos eran diferentes a las que se vivía en un contexto COVID. Dicho esto, el banco tuvo la necesidad de priorizar la implementación de los cambios e incluso creación de nuevos modelos.

Debido a que el contexto COVID era muy cambiante, las actualizaciones o cambios de los modelos también lo eran; y fue ahí donde se empezó a generar con mayor frecuencia la evaluación del riesgo del modelo dentro de sus procesos de actualización y/o creación con el fin de no incurrir en errores en el desarrollo e implementación. Por eso fue necesario validar la metodología estadística usada en la implementación mediante una auditoría de modelos, con el fin mitigar un incorrecto uso de la metodología y mejorar los modelos; además de cumplir con las normas que exige la Política de Gestión de Riesgos basados en las mejores prácticas conforme lo exige la Norma Internacional para la Práctica Profesional de la Auditoría Interna No. 2210.

1.2 Objetivos

Objetivo principal:

Evaluar los controles de calidad de datos, la metodología de cálculo aplicada y los procedimientos de implementación de la Calibración al Modelo Estimador de Ingresos para Dependientes + RCC (Reporte Consolidado Crediticio) de la Banca Minorista.

Objetivos específicos:

- Verificar el adecuado funcionamiento de los controles de calidad de la información que garanticen la integridad de los datos utilizados en la implementación y ejecución del modelo.
- Comprobar que el análisis estadístico realizado en la selección de las variables del modelo permite identificar adecuadamente el nivel de contribución de cada variable sobre la variable respuesta, y cuenten con adecuados niveles de estabilidad y significancia.
- Validar que los procedimientos metodológicos utilizados en el desarrollo y calibración de los modelos cuenten con sustento analítico, y permiten mejorar el grado de precisión de la estimación.

II. REVISIÓN DE LA LITERATURA

Con el fin de poder resolver la problemática mencionada en el capítulo anterior, se realizó la validación metodológica estadística usada en la implementación de modelos, para no incurrir en errores que se materialicen en pérdidas para el banco. Para un mejor entendimiento de los conceptos y terminologías se presentarán diferentes conceptos que ayudarán a un mejor entendimiento del presente documento. Cabe recalcar que la mayoría de las definiciones se manejan de manera interna en el banco.

2.1 Conceptos, fases y terminologías de Auditoría.

Para iniciar con la descripción de los conceptos se comenzará con la definición de auditoría y sus fases. En ese sentido, se puede definir a la auditoría como un proceso o herramienta de control y supervisión que permite descubrir fallas o vulnerabilidades en los procesos o flujos que existen en la organización; para ello, se evalúa de forma objetiva algunas evidencias plasmadas en informes (Álvarez et al. (2006)). Además, contribuye a la creación de una cultura disciplinaria en la organización y existen varios tipos de auditoría, sin embargo, esta investigación se centra básicamente en la Auditoría interna, la cual se realiza en el banco, se encuentra definida como una actividad independiente y objetiva que presta servicios de aseguramiento y de consultoría, además, tiene como objetivo la generación de valor y mejora de las operaciones de la organización. También ayuda a la organización en el logro de sus objetivos, aplicando un enfoque sistemático y disciplinado para evaluar y mejorar la eficacia de sus procesos de administración de riesgo, de control y de gobierno corporativo. Esta auditoría tiene 3 fases, Planificación, ejecución y conclusión y se detallan en la Figura 2 y 3. Por otro lado, en el banco existen terminologías como:

Comité de Auditoría Interna (CAI) que es un órgano de soporte y supervisión, que apoya las funciones realizadas por la Junta Directiva en materia de control interno.

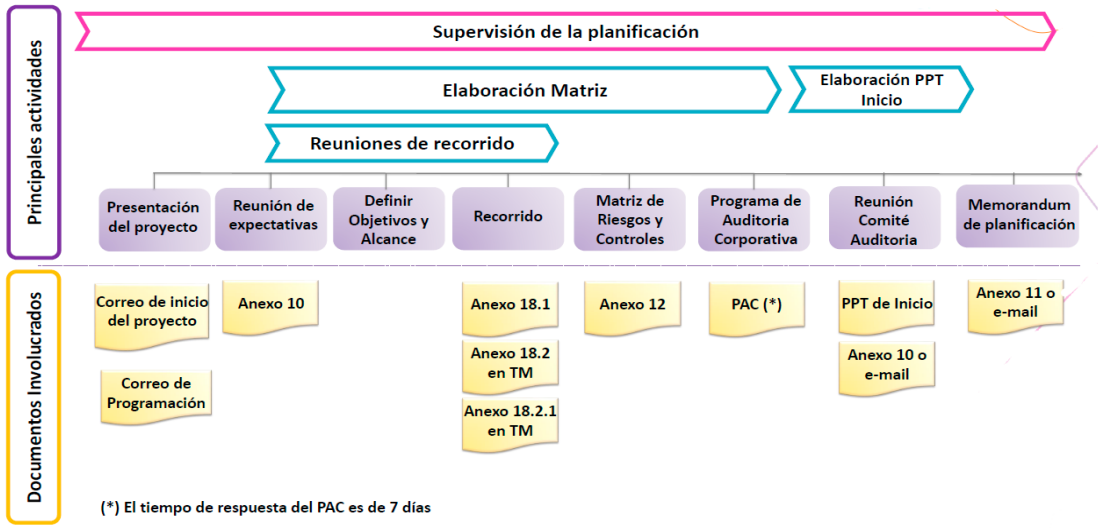
Riesgo de banca minorista (RBM) es aquel riesgo que se origina mayoritariamente a la actividad de financiación a personas físicas y PYMES.

Reporte consolidado crediticio (RCC) es un registro que contiene información sobre los deudores del banco en el sistema financiero, con el fin de contar con información consolidada y clasificada sobre los deudores del banco a efectos de promover la solidez de los sistemas, evitando el sobreendeudamiento y la morosidad de sus usuarios.

Por otro lado, existen actividades relacionadas con la atención del cliente interno (del mismo banco), con quienes se acuerda la naturaleza y alcance de estas actividades, las cuales pretenden agregar valor y mejorar los procesos de gobierno, administración de riesgos y de control de una organización, sin que el auditor interno asuma la responsabilidad de gestión. Por ejemplo, opinión, consejo y entrenamiento; a esto de le denomina consultoría interna.

Figura 2

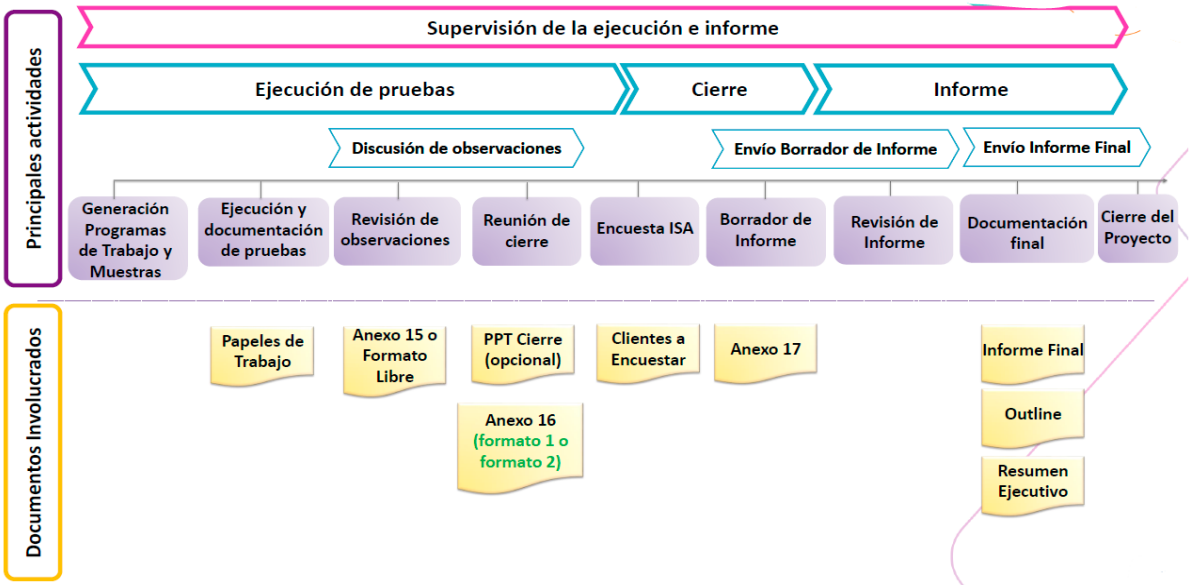
Fase de Planificación



Nota: Reproducida de Fase de Planificación. 2022. Elaboración del banco.

Figura 3

Fase de Ejecución y Conclusión



Nota: Reproducida de Fase de Planificación. 2022. Elaboración del banco.

Ahora que se conoce que la auditoría interna es una herramienta de control, se tiene que definir qué es un control y qué es un riesgo, ambos se encuentran muy relacionados, ya que el control nace debido a la existencia un **riesgo**. En ese sentido, este último se encuentra definido como la posibilidad de que ocurra un acontecimiento que tenga un impacto en el alcance de los objetivos. El riesgo se mide en términos de impacto y probabilidad. Mientras que el **control** se define como cualquier medida que toma la dirección, el directorio y otras partes, para gestionar riesgos y aumentar la probabilidad de alcanzar los objetivos y metas establecidos. La dirección planifica, organiza y dirige la realización de las acciones suficientes para proporcionar una seguridad razonable de que se alcanzaran los objetivos y metas.

2.2 Conceptos enlazados a los modelos

Luego de conocer qué es un riesgo y control se debe conocer la definición de modelo, se refiere a un proceso o procedimiento que busca replicar lo que ocurriría en la realidad mediante una simplificación de ésta. Dicho proceso o procedimiento aplica una metodología que usualmente proviene de teorías o técnicas (econométricas, estadísticas, económicas, financieras o matemáticas, etc.) y/o del conocimiento experto, pero también de la formulación de determinados supuestos. Un **modelo** busca por lo general explicar o predecir algo que ocurriría en la realidad sea de naturaleza cuantitativa o cualitativa. La modelización estadística es la aproximación a la realidad mediante un modelo matemático que cuenta con datos obtenidos mediante la experiencia o la experimentación. En resumen, un modelo es un proceso que utiliza determinados insumos o datos en el marco de una metodología con el objetivo de predecir o explicar algo de la realidad. Bajo la definición del modelo, puede que exista errores en su desarrollo e implementación lo que conlleva a posibles pérdidas que una empresa puede incurrir como consecuencia de decisiones que podrían basarse principalmente en la producción de esos modelos (internos); a esto se le denomina **riesgo de modelo**.

En el **seguimiento de modelos** se incluye cada uno de los procesos que aseguren un sistema eficiente de administración de información para monitorear un modelo en el tiempo; permitiendo mayor precisión en la toma de decisiones que implica el uso de este. A su vez es un proceso mensual, trimestral u otro de acuerdo con la política interna del banco correspondiente al comité de modelos. A medida que el tiempo transcurre un modelo puede tornarse inestable debido a diversos factores como, por ejemplo: cambios en la población sobre

el cual fue construido, pérdida del poder discriminativo de las variables, disminución de su poder predictivo, uso indebido del mismo, entre otros. El seguimiento permite detectar a tiempo alguna inconsistencia desde el momento de la implementación del modelo siendo la mejor forma de conocer el desempeño de un modelo al monitorear cualitativa y cuantitativamente su comportamiento y poder predictivo en el tiempo.

Por otro lado, cuando un modelo ya implementado con el paso del tiempo ya no estima correctamente, es porque necesita una actualización; es decir, necesita un ajuste o cambio con el fin de mejorar el modelo; a esto se le denomina **calibración de modelos.** Asimismo, el proceso donde se evalúa y revisa que los procedimientos usados en la implementación y/o calibración de un modelo se haya realizado correctamente, es el de **validación de modelos**.

Para llevar a cabo una validación de modelos, el banco realiza diferentes tipos de pruebas como:

Prueba de cumplimiento: tiene como principal objetivo proporcionar seguridad razonable de que los procedimientos relativos a controles internos están siendo aplicados tal como fueron establecidos. Estas pruebas son necesarias si se va a confiar en los procedimientos descritos. Sin embargo, se puede decidir no confiar en los mismos si se ha llegado a la conclusión de que:

a) Los procedimientos no son satisfactorios para este propósito.
b) El trabajo necesario para comprobar el cumplimiento de los procedimientos descritos es mayor que el trabajo que se realizaría en el caso de no confiar en dichos procedimientos.

Esta última conclusión puede resultar de consideraciones relativas a la naturaleza o número de transacciones o saldos involucrados, los métodos de procedimiento de datos que se esté usando y los procedimientos de auditoría que puedan ser aplicados al realizar las pruebas sustantivas.

Prueba sustantiva: Son aquellas pruebas que diseña el auditor con el objetivo de conseguir evidencia referida a la información financiera y operativa. Están relacionadas con la integridad, exactitud y la validez de la información financiera y operativa. Los procedimientos sustantivos intentan dar validez y fiabilidad a toda información que generan los estados contables y en concreto a la exactitud monetaria de las cantidades reflejadas en los estados financieros.

Prueba multipropósito: Cuando los procedimientos sustantivos, por si solos, no proporcionan evidencia de auditoría suficientes y adecuadas, el auditor tiene la obligación de realizar pruebas de control para obtener evidencias de auditoría de su eficacia operacional.

En ese sentido, como parte del desarrollo de una auditoría de modelos es necesario contar con evidencias que sustenten los diferentes tipos de pruebas realizadas en la auditoría. Es decir, la **evidencia de auditoría** es la suma de la información usada por el auditor para llegar a las conclusiones en las cuales se basan los resultados de sus trabajos y comprende la información y los datos subyacentes a la información financiera, riesgos, controles y tecnología de la información. La evidencia de auditoría es de naturaleza acumulativa e incluyen las evidencias obtenidas durante el trabajo, como también de otras fuentes, como trabajos de auditorías anteriores. Existen diferentes tipos de evidencia de auditoría como:

- **Evidencia documental:** Consiste en verificar documentos (financieros, nóminas, etc.).
- **Evidencia física:** Permite identificar la existencia física de activos. cuantificar las unidades en poder de la empresa, y en ciertos casos especificar la calidad de los activos.
- **Evidencia por medio de cálculos:** Realización de cálculos y pruebas globales para verificar la precisión aritmética de saldos, registros y documentos.
- **Evidencia por medio de comparaciones y ratios:** Es un medio de localizar cambios significativos que deberán ser explicados al auditor.
- **Evidencia Verbal:** Por medio de preguntas a empleados y ejecutivos.

2.3 Conceptos y definiciones estadísticas

Por otro lado, para el desarrollo del presente trabajo, es necesario conocer la definición de un **Modelo estimador de ingresos,** que es una herramienta estadística que permite estimar el monto de ingresos de una persona a través de su información. En un modelo estadístico existen diversas etapas en la construcción; y para ello es necesario conocer algunas definiciones y procedimientos que ayudaran a un mejor entendimiento, como:

- **Convolución de variables:** Proceso de transformación de variables para la creación de un modelo; estos pueden ser convolución para variables de tipo categóricas o continuas. Por ejemplo:

  Se cuentan con dos variables, estado civil y edad; la convolución de ambas sería una variable denominada “cv_estcivil_edad”; donde esta se encuentra definida como:

  cv_estcivil_edad = f (Estadocivil, Edad, Ingreso)

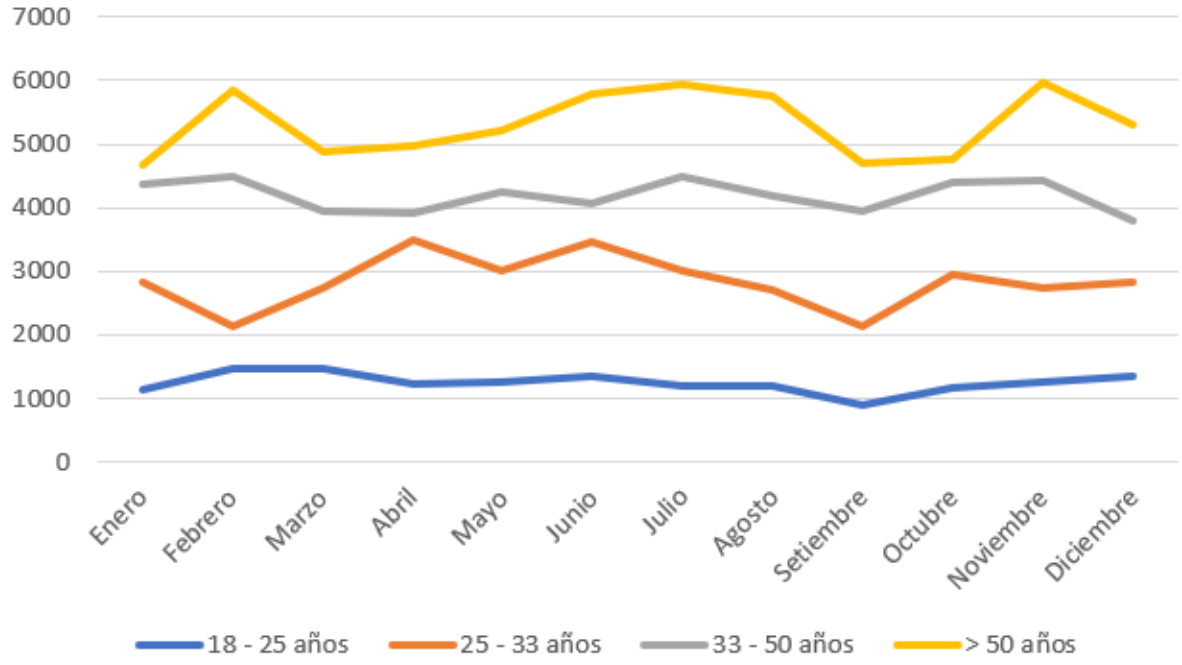
  Es un tipo de transformación, donde se crea la variable convolucionada. Para el ejemplo, la variable convolucionada “cv_estcivil_edad” se calcula promediando la variable de

interés “Ingreso” en función a todos los subconjuntos creados de la combinación de categorías de las variables Edad y Estadocivil.

- **Sentido económico**: es la evaluación del comportamiento de una variable, en función a un criterio y/o lógica razonable, por ejemplo: cómo se puede observar en la Figura4, el ingreso promedio de una persona de 33 años a más es mayor al ingreso promedio de un joven de entre 18 a 25 años. Puede existir casos atípicos, pero la tendencia común es la que predomina en el comportamiento de una variable.

Figura 4

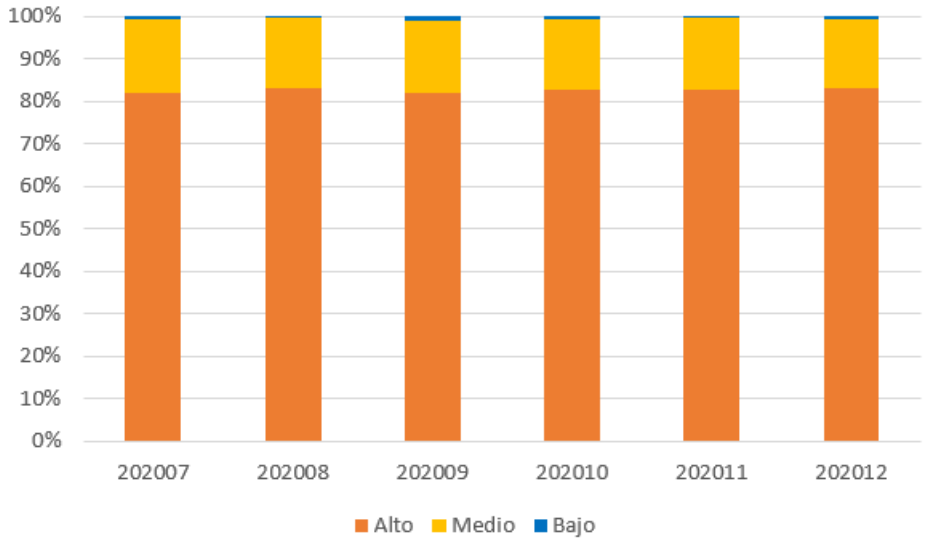
Ingreso Promedio Según Grupo de Edad



- **Significatividad o materialidad:** Se dice que una variable tiene significatividad o materialidad, cuando al evaluar la proporción de sus categorías, estas son representativas. Por ejemplo, como se puede observar en la Figura5 cuando se evalúa la significatividad de una variable categórica que tiene 3 valores (bajo, mediano, alto), y uno de los valores solo representa el 1% del total de registros de la variable, se dice que esta categoría de la variable tiene muy poca significatividad o materialidad. Cabe recalcar que no se debe de confundir significatividad con significancia estadística.

Figura 5

Gráfica de Materialidad de la Variable Categorizada Ingresos



- **Tiering del modelo:** esta definición se basa en repartir todos los modelos en cuatro niveles en orden de relevancia, siendo el Tier I el grupo de los modelos más importante y el Tier IV el de los menos relevantes. El tiering se construye a partir de cuatro indicadores que se integran entre ellos para definir el tier final de un modelo. Los factores que se han tenido en cuenta a la hora de clasificar los modelos han sido:
  i) Impacto sobre los resultados del banco.
  ii) Materialidad en función del tamaño de la cartera subyacente.
  iii) Criticidad de las decisiones en función del uso del modelo.
  iv) Importancia estratégica con visión de futuro en función de planes de inversión o desinversión sobre la cartera a la que el modelo se aplica.

- **Precisión ±25% y Sobreestimación/ Subestimación:**

  - El descriptivo de la métrica: Ésta fue implementada por el banco y se define como la proporción de casos que se encuentran dentro del ± 25% del ingreso observado, mientras que la segunda métrica de sobrestimación/ subestimación se incluye en el indicador como una métrica adicional, básicamente para conocer en qué porcentaje

el modelo estima más o menos al real. Por ejemplo, como se muestra en la Tabla1, si tenemos 3 ingresos observados de 1000, 2000 y 1600**.**

Tabla 1

Ejemplo de Precisión ±25%

Ingreso Observado	±25% Obs (Banda Sup)	±25% Obs (Banda Inf)	Ingreso Est Infinity	INDICADOR	IN +-25%	SOB +25%
1000	1250	750	1350	SOB		
2000	2500	1500	2250	IN	66.7%	33.3%
1600	2000	1200	1400	IN		

Para el ejemplo, 2 de los 3 ingresos estimados se encuentran dentro del intervalo de ±25% del ingreso observado. Por lo tanto, el porcentaje de valores que se estimaron dentro de los valores de ±25% del ingreso observado (IN) es de 66.7%, mientras que la sobreestimación es de 33.3%.

Siguiendo la línea, según García et al. (2017) la etapa de **selección de variables** ayuda a evitar problemas de multicolinealidad, es decir es de gran ayuda que las variables seleccionadas cuenten con pruebas de confirmación. Para ello se selecciona el conjunto de variables explicativas y significativas, evitando entonces que se incluyan variables poco significativas o con información redundante (colinealidad), lo que puede distorsionar la capacidad predictiva de la función discriminante estimada.

Finamente, luego de conocer las terminologías de auditoría básicas, se presenta las técnicas estadísticas más importantes que se deben conocer. En ese sentido, cabe mencionar que existen tres categorías de análisis de datos: el análisis univariante, el análisis bivariante y el análisis multivariante

Según Everitt et al. (2010). el **Análisis Univariado** es la forma más sencilla de análisis de datos, en la que los datos analizados sólo contienen una variable y solo se preocupa de las causas no de las relaciones. Este análisis de datos tiene como objetivo principal describir los datos y encontrar los patrones que existen en ellos. Un ejemplo de variable en el análisis univariante puede ser la edad y/o altura. El análisis univariante no examinaría estas dos variables al mismo tiempo, ni examinaría la relación entre ellas. La forma en que se pueden describir los patrones encontrados en los datos univariados pueden ser: la observación de la media, la moda, la mediana, el rango, la varianza, el máximo, el mínimo, los cuartiles y la desviación estándar.

Además, se cuenta con otras formas de mostrar los datos univariantes son las tablas de distribución de frecuencias, los gráficos de barras, los histogramas, los polígonos de frecuencias y los gráficos circulares, entre otros. Siguiendo la línea, también afirma que el **Análisis Bivariado** es utilizado para conocer la relación que existe entre dos variables diferentes. Algo tan sencillo como crear un gráfico de dispersión trazando una variable frente a otra en un plano cartesiano (considerar los ejes X e Y) a veces puede dar una idea de lo que los datos están tratando de decirle. Si los datos parecen ajustarse a una línea o curva, entonces existe una relación o correlación entre las dos variables. Por ejemplo, se puede elegir representar la ingesta calórica frente al peso.

Machine Learning: Es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo). Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados.

Durante los últimos años, ha ganado mucha relevancia debido al aumento de la capacidad de computación y al boom de los datos. Las técnicas de aprendizaje automático son, de hecho, una parte fundamental del Big Data.

Los algoritmos de Machine Learning se dividen en tres categorías, siendo las dos primeras las más comunes:

- Aprendizaje supervisado: estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. Un ejemplo es un detector de *spam* que etiqueta un *e-mail* como *spam* o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.). Entre los algoritmos de aprendizaje supervisado se tiene: Arboles de decisión, Clasificación de Naive Bayes, Regresión por mínimos cuadrados, Regresión Logistica, Support Vector Machines (SVM), XGBoost, entre otros.
- Aprendizaje no supervisado: estos algoritmos no cuentan con un conocimiento previo de los datos. Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna manera. La mayoría de las técnicas de aprendizaje no supervisado están dedicadas a las tareas de agrupamiento, también llamadas clustering

o segmentación**,** donde su objetivo es encontrar grupos similares en el conjunto de datos. Por ejemplo, en el campo del marketing se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas.

Existen dos grupos principales de métodos o algoritmos de agrupamiento:

1. Los métodos jerárquicos, que producen una organización jerárquica de las instancias que forman el conjunto de datos, posibilitando de esta forma distintos niveles de agrupación.

2. Los métodos particionales o no jerárquicos, que generan grupos de instancias que no responden a ningún tipo de organización jerárquica.

- Aprendizaje por refuerzo: su objetivo es que un algoritmo aprenda a partir de la propia experiencia. Esto es, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo con un proceso de prueba y error en el que se recompensan las decisiones correctas. En la actualidad se está utilizando para posibilitar el reconocimiento facial, hacer diagnósticos médicos o clasificar secuencias de ADN.

Estos conceptos se encuentran definidos en de Iberdrola N.N (2022). Iberdrola. Recuperado el día 08 de Agosto del 2022, de https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico#:~:text=El%20Machine%20Learning%20es%20una,elaborar%20predicciones%20(an%C3%A1lisis%20predictivo).

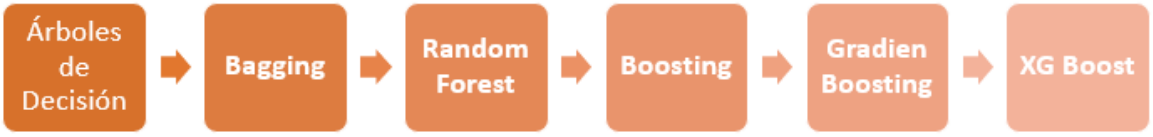
Algoritmo Extreme Gradient Boosting (XGBoost):

Según Leo (1996), para conocer el concepto del XGBoost, inicialmente se define el concepto de Bagging, donde se define como un método que combina las predicciones de varias secuencias de datos para crear un resultado más preciso. Esto puede ser aplicado a árboles de decisión, separando un conjunto de datos en N partes, ajustando un árbol de decisión a cada uno y luego promediando las salidas para obtener un resultado final. Posteriormente surgió el concepto de *Booosting* que tiene una lógica parecida al método *Bagging*, es decir, se separa un conjunto de datos en N partes, donde los predictores (en nuestro caso, árboles de decisión) aprenden de los errores de los árboles anteriores, obteniendo un resultado final a través del último árbol. Con esto, se define un *modelo de ensamble* como aquel que está formado por un conjunto de árboles de decisión individuales, los cuales son entrenados de forma secuencial, de

forma de que cada nuevo árbol utiliza la información del árbol anterior para aprender de sus errores. La predicción de una nueva observación se obtiene sumando las predicciones de todos los árboles que componen al modelo. Como se puede apreciar en la Figura 6, la evolución que siguen los modelos Gradient Boosting y Extreme Gradient Boosting.

Figura 6

Evolución de los Algoritmos Basados en Arboles de Decisión



Según Chen y Guestrin (2016), XGBoost es un algoritmo de aprendizaje automático supervisado basado en árboles de decisión y que es considerado el estado del arte en la evolución de estos algoritmos.

El algoritmo XG Boost tiene las siguientes características

a. Consiste en un ensamblado secuencial de árboles de decisión (este ensamblado se conoce como CART, acrónimo de “Classification and Regression Trees”). Los árboles se agregan secuencialmente a fin de aprender del resultado de los árboles previos y corregir el error producido por los mismos, hasta que ya no se pueda corregir más dicho error (esto se conoce como “gradiente descendente”.
b. La principal diferencia entre los algoritmos XGBoost y Random Forest es que en el primero el usuario define la extensión de los árboles mientras que en el segundo los árboles crecen hasta su máxima extensión.
c. Utiliza procesamiento en paralelo, poda de árboles, manejo de valores perdidos y regularización (optimización que penaliza la complejidad de los modelos) para evitar en lo posible sobreajuste o sesgo del modelo.

Asimismo, Brasa Pedro (2019) define al XGBoost como uno de los algoritmos de boosting más recientes es el XGBoost (Extreme Gradient Boosting). Es un método de código abierto muy utilizado en retos y trabajos de data mining como las competiciones de machine learning de Kaggle (es la plataforma de Data Science más grande del mundo con más de 1 millón de

usuarios, y es una plataforma excelente para que estudiantes crezcan en el campo de Data Science y Machine Learning.) donde en 2015, 17 de los 25 métodos presentados utilizaron XGBoost para su resolución. Además, en KDDCup 2015 (es un congreso virtual que conecta a los científicos de datos y expertos a través de una competición en la que se proponen soluciones inteligentes a desafíos y casos reales) los 10 equipos con mejor puntuación utilizaron este algoritmo. A través del boosting, el XGBoost ejecuta K iteraciones donde en cada una se añade un nuevo árbol de decisión que intentará corregir el error cometido en las anteriores. En cada iteración del algoritmo XGBoost se añade un nuevo árbol de decisión, que se desarrolla con el objetivo de minimizar el error acumulado por los árboles anteriores. El objetivo del XGBoost es minimizar el error de predicción optimizando la ecuación 1

$$obj(\theta) = \sum_i^N l(y_i, \hat{y_i}) + \sum_{k=1}^k \Omega(f_k) \quad \text{..... (Ecuación 1)}$$

Donde el primer término es una función que representa el error cometido entre las predicciones estimadas por XGBoost y los valores reales. Esta función por defecto suele ser el error cuadrático medio (MSE) en el caso de regresión o la precisión (ACC) para clasificación. Este primer término es personalizable por el usuario en función del objetivo del modelo concreto. El segundo término de la fórmula es una función que penaliza la complejidad del modelo para reducir el sobreajuste del modelo, ya que es un problema recurrente al trabajar con XGBoost. Este sobreajuste se debe a la relación de dependencia estadística entre los árboles construidos. Además, el tiempo de construcción del modelo aumenta al aumentar el número de iteraciones.

Boruta

Inicialmente se empezará a conocer los diferentes problemas que surgen en la selección de variables, debido a la gran cantidad de información existente en las bases de datos, lo que conlleva a una disminución de la precisión en los modelos cuando el número de variables usadas no es el óptimo (Alsahaf et al., 2022; Kohavi y John, 1997). La presencia de variables irrelevantes para el modelo conlleva a un aumento en los gastos de almacenamiento y costo computacional. El proceso de selección de variables puede mitigar estos problemas seleccionando el conjunto de variables más importantes y eliminando las variables irrelevantes.

Según Kursa y Rudnicki (2010), Boruta es un método utilizado para obtener un subconjunto de variables “importantes” para el modelo.

Además, utiliza Random Forest como algoritmo subyacente. La idea es generar en cada iteración una serie de variables sombra a partir de los predictores, copiando cada uno de ellos y permutando entre sí los elementos de cada nueva columna. Se ajusta un modelo por Random Forest y se calculan las importancias relativas de cada variable. Si una variable sistemáticamente queda por debajo de las sintéticas (ruido), será indicativo de que su aportación al modelo será dudosa y por tanto se elimina. El proceso continúa hasta que todas las variables son aceptadas, rechazadas o se alcanza un número de iteraciones límite. (Guerrero, 2016).

Hiperparameters tunning

Antes de definir el tuneo de hiperparámetros, primero se definirá lo que son los hiperparámetros; estos son las configuraciones de un modelo de machine learning donde el algoritmo no aprende a partir de los datos, sino que éstas deben ser definidas previamente por el analista antes de la construcción del modelo. Por ejemplo: en un Random Forest los hiperparámetros son: Número de árboles, profundidad de los árboles, numero de variables a usar en cada árbol. Por otro lado, el tuneo de hiperparámetros es el proceso mediante el cual se busca el conjunto de hiperparámetros que logra maximizar el desempeño de un modelo a la hora de realizar predicciones. Es decir, tuneo de hiperparámetros se refiere a la optimización mediante iteraciones para encontrar el set de hiperparámetros que consigan el menor valor de la función de pérdida en la data de validación.

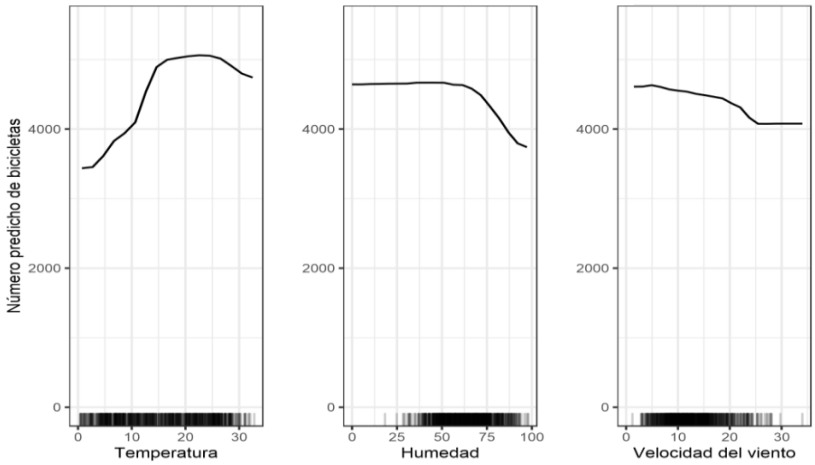
Gráficos de Dependencia Parcial (PDP)

Según Molnar (2020) son uno de los métodos más populares para explorar la relación entre una variable continua y el resultado del modelo. Puede mostrar si la relación entre el objetivo y una característica es lineal, monótona o más compleja. Por ejemplo, cuando se aplica a un modelo de regresión lineal, los gráficos de dependencia parcial siempre muestran una relación lineal. Por ejemplo, en un modelo de regresión, en el que se predice el número de bicicletas que se alquilarán en un día determinado, primero se ajustó a un modelo de aprendizaje automático, luego, se analizó las dependencias parciales. Para el ejemplo se ajustó a un Random Forest para predecir el número de bicicletas y utilizar el diagrama de dependencia parcial para visualizar las relaciones que el modelo ha aprendido. La influencia de las características climáticas en el conteo previsto de bicicletas se visualiza en la siguiente Figura7.

Para el modelo de predicción de conteo de bicicletas y temperatura, humedad y velocidad del viento. Las mayores diferencias se pueden ver en la temperatura. Cuanto más caliente, más bicicletas se alquilan. Esta tendencia sube a 20 grados centígrados, luego se aplana y cae ligeramente a 30. Las marcas en el eje x indican la distribución de datos. Para clima cálido, pero no demasiado caluroso, el modelo predice en promedio un gran número de bicicletas alquiladas. Los ciclistas potenciales se inhiben cada vez más en el alquiler de una bicicleta cuando la humedad supera el 60%. Además, cuanto más viento, a menos personas les gusta andar en bicicleta, lo que tiene sentido. Curiosamente, el número previsto de alquileres de bicicletas no disminuye cuando la velocidad del viento aumenta de 25 a 35 km/h.

Figura 7

Gráfico PDP del Número Predicho de Bicicletas.



Nota: PDP para el modelo de predicción de conteo de bicicletas y temperatura, humedad y velocidad del viento. Las mayores diferencias se pueden ver en la temperatura. Cuanto más caliente, más bicicletas se alquilan. Esta tendencia sube a 20 grados centígrados, luego se aplana y cae ligeramente a 30. Las marcas en el eje x indican la distribución de datos. Tomada de *Interpretable Machine Learning* por Christoph Molnar, 2022.

Por otro lado, este método tiene sus ventajas y desventajas. Si la variable para la que calculó el PDP no está correlacionada con las otras variables, entonces los PDP representan perfectamente cómo la característica influye en la predicción en promedio. En el caso no correlacionado, la interpretación es clara; mientras que entre sus desventajas se encuentra la suposición de independencia de variables, los efectos heterogéneos pueden estar ocultos, es decir, se tiene

solo una línea horizontal, esto se puede deber a que puede existir una asociación positiva con la variable respuesta y la otra mitad tiene una asociación negativa lo que hace que los efectos de ambas mitades de la variable se cancelen entre sí.

III. DESARROLLO DEL TRABAJO

3.1 Delimitación temporal y de ámbito geográfico y la naturaleza de trabajo

El banco de donde se obtuvo la información y donde se realizó el presente trabajo, se encuentra ubicado en Perú, departamento de Lima. Para el universo de modelamiento, se extrajo de las fuentes del banco la información para todos los clientes con ingresos por pago de haberes con información histórica desde octubre del 2020 hasta marzo del 2021, procedente de la base de ingresos observados Pago de Haberes (PDH) de Riesgo de Banca Minorista (RBM); además, dichos clientes son trabajadores dependientes y tienen experiencia en el sistema financiero por lo menos un año, finalmente se contó con 727,945 registros de clientes PDH.

El tipo de investigación que se aplicó es descriptivo, debido a que se realizó la réplica del proceso de implementación de un modelo y analizó los resultados obtenidos de la validación de los modelos; además, se describió y mostró los resultados obtenidos y se contribuyó al mejoramiento de éste.

Por otro lado, la naturaleza del estudio es no experimental, transversal, ya que se tomó las variables de las fuentes internas del banco y no se manipularon las variables. El diseño es transversal, debido a que analizó los datos en un solo periodo de tiempo (octubre 2020 – marzo 2021).

3.2 Fuentes de Información

La fuente de información que se utilizó fue la primaria, la cual se especifica a continuación:

- El universo de modelamiento se generó de la información del banco, procedente de la base de ingresos observados PDH de Riesgo de Banca Minorista (RBM); dicho universo, fueron todos los clientes con ingresos por PDH con información histórica desde octubre 2020 hasta marzo del 2021; además, tenían que ser trabajadores dependientes y contaban con experiencia en el sistema financiero por lo menos un año. Los filtros usados en este universo de modelamiento se detallarán en el siguiente subcapítulo.

- Entrevistas con el personal involucrado en el proceso: estas entrevistas se realizan con el fin de obtener información de proceso de implementación y/o calibración de un modelo; además de mejorar el entendimiento del negocio y/o fin del modelo.
- Documentación del área de auditoría interna: hace referencia a la documentación que se obtiene de los auditados; donde se encuentran los archivos de trabajo con el paso a paso de cada uno de los controles que realizan al momento de la implementación y/o calibración de los modelos.

3.3 Procedimientos:

Considerando la experiencia profesional de la bachillera en estadística, quien suscribe este documento, cuyo cargo viene siendo el de Auditor senior de Modelos de Riesgo en el Área de Auditoría de Riesgos del banco; se ha conseguido evaluar la metodología aplicada, los procedimientos de cálculo realizados y los resultados de los procesos de implementación y/o calibración de modelos. Además, el cargo y la función que se viene desempeñando hasta la actualidad en el banco han permitido conocer las diferentes etapas en la construcción y/o calibración de un modelo que va más allá de la metodología de cálculo; además, evaluar el proceso global de su implementación en la gestión del banco separados por ámbitos; por ejemplo:

- La documentación, donde se revisa la existencia y suficiencia de un documento metodológico de la construcción y/o calibración del modelo.
- El gobierno, donde se revisa políticas definidas en el banco, como: la aprobación y/o modificación de un modelo, monitoreo o seguimiento y la definición del tiering del modelo.
- Finalmente, los ámbitos de Calidad de datos y metodología, que son los ámbitos donde se tiene mayor conocimiento, debido a la formación profesional que se tiene como bachiller en Ciencias – Estadística Informática.

En ese sentido, el procedimiento para el logro de objetivos planteados se empieza con la Fase de planificación en Auditoría, en esta fase el proceso de auditoría de Validación al Modelo Estimador de Ingresos para Dependientes + RCC de la Banca Minorista, empieza con un correo de inicio, donde se detalla el inicio del proyecto de validación, los integrantes del equipo, y algunos objetivos por ámbito; como lo detalla la Figura8; con el propósito de dar a conocer a

las auditados (internos) los puntos antes mencionados. Además, se tienen las reuniones de recorrido (entendimiento) donde se recopilan las expectativas de las unidades auditadas, se definen objetivos y alcance. Después de tener las reuniones de recorrido los auditores elaboran la matriz de riesgos y controles y elaboran el memorándum de planificación. En esta fase (Planificación) se van documentando diferentes Anexos con el fin de cumplir con las políticas internas del banco, y la Norma Internacional para la Práctica Profesional de la Auditoría Interna No. 2210.

Figura 8

Correo de Inicio de la Auditoría de Modelos Estimador de Ingresos

RV: Correo de Inicio DAP-012-22 "Auditoría de Validación del Modelo Estimador de Ingresos - Banca Minorista"

Asunto: Correo de Inicio DAP-012-22 "Auditoría de Validación del Modelo Estimador de Ingresos - Banca Minorista"

Estimados Señores:

En cumplimiento de nuestro Plan Anual, se efectuará una Auditoría de Validación al Modelo Estimador de Ingresos de la Banca Minorista, a fin de evaluar los controles de calidad de datos, la metodología de cálculo aplicada, los procedimientos de implementación, el seguimiento a los resultados, la validación efectuada y los controles del entorno tecnológico de soporte, verificando que se cumplan los estándares establecidos por las mejores prácticas, en tal sentido presentamos a ustedes a nuestros auditores:

Liz Suarez (Jefe de Equipo)
Oscar Zúñiga
Alessandra Carrasco
Franz Villanueva
Kenia Calisaya
Francisco Aira

Quienes conformarán el equipo de trabajo encargado de realizar la Auditoría, por lo que agradeceremos se sirvan brindar las facilidades correspondientes, para el mejor desempeño de sus funciones. Asimismo, conforme a la Norma Internacional para la Práctica Profesional de la Auditoría Interna No. 2210, planteamos de manera preliminar los siguientes objetivos:

Base de Datos:
Verificar el adecuado funcionamiento de los controles de calidad de la información que garanticen la integridad de los datos utilizados en la ejecución del modelo.
Verificar la existencia de documentación completa referida al proceso de generación de datos y ejecución del modelo, que permita la trazabilidad del proceso.

Metodología de Cálculo:
Comprobar que existan procedimientos y métricas adecuadas para asegurar la calidad de los datos utilizados en el desarrollo y calibración del modelo.
Comprobar que el análisis estadístico realizado en la selección de las variables del modelo permite identificar adecuadamente el nivel de contribución de cada variable sobre la variable respuesta, y cuenten con adecuados niveles de estabilidad y significancia.
Validar que los procedimientos metodológicos utilizados en el desarrollo y calibración de los modelos cuenten con sustento analítico, y permiten mejorar el grado de precisión de la estimación.
Comprobar que la documentación referente a la metodología de cálculo y ajuste de los modelos cuenten con un nivel de detalle suficiente que permita replicar de forma correcta el proceso.
Verificar que el proceso de validación del modelo se realice de forma oportuna e independiente y que sus resultados se presenten ante un foro pertinente.

Integración a la Gestión
Verificar que se haya realizado un adecuado proceso de implementación del modelo y la correcta certificación de dicho proceso.
Verificar que se realice un adecuado seguimiento de los resultados del modelo y se comunique de forma correcta y oportuna a las Gerencias usuarias y/o los Comités pertinentes.

Entorno Tecnológico:
Verificar que se realicen procedimientos de respaldo y restauración del entorno SAS así como un monitoreo efectivo de su capacidad de procesamiento y almacenamiento.
Comprobar que existan controles que aseguren la confidencialidad de la información y la disponibilidad de los sistemas utilizados en todo el ciclo de vida de los modelos.

Cabe indicar, que el examen se efectuará sobre la información presentada a enero 2022.

A la espera de poder contar con vuestra colaboración y apoyo.

Saludos
JE

Luego, se empezó con la fase de la ejecución, en la cual se realiza la generación de los programas de trabajo, ejecución y documentación de las pruebas, revisión de observaciones y reunión de cierre. Esta fase es la principal del proceso de auditoría; aproximadamente el 40% del tiempo de duración del proyecto se usa en la ejecución de las pruebas. Es aquí donde se realizaron las pruebas en función a los objetivos. Cabe mencionar que en esta etapa se pondrá a prueba el conocimiento teórico adquirido en los diferentes cursos llevados en la carrera.

Como primer paso para la ejecución de las pruebas, se revisó toda la documentación solicitada posterior a las reuniones de entendimiento; como se muestra en la Tabla2; con estos documentos

será mucho más fácil hacer la reconstrucción de las bases de datos utilizando las fuentes oficiales y confiables del banco.

Tabla 2

Descripción de los Documentos Usados para la Ejecución de las Pruebas.

Documento	Descripción
Manual de implementación del Estimador Ingresos RCC - Dependiente - Rev1.1	El documento contiene información donde se detalla los pasos necesarios para la correcta implementación del nuevo modelo del estimador de ingresos RCC – dependientes. Los acápites del presente documento son: construcción del universo de clientes; en segundo lugar, especificamos las fuentes y variables, en tercer lugar, la segmentación; luego el tratamiento de las variables; por último, mostramos el proceso de puntuación y calibración.
Manual_metodológico_Estimador_ Ingresos_RCC_Dep 1.1 - k (1)	Documento que detalla el proceso metodológico utilizado en la construcción del Nuevo Estimador de Ingresos para personas dependientes con experiencia en el sistema financiero. Considerando tal proceso desde la extracción de las fuentes, su tratamiento estadístico, la determinación de las decisiones y procedimientos para ello y el análisis hasta la obtención del Ingreso Estimado Final.

Entonces se inicia el proceso de réplica basada en los objetivos del presente trabajo; para ello es necesario definir el universo de modelamiento y la población objetivo.

El universo de modelamiento son todos los clientes con ingresos por pago de haberes (PDH) con información histórica desde octubre del 2020 hasta marzo del 2021, procedente de la base de ingresos observados Pago de Haberes (PDH) de Riesgo Banca Minorista (RBM). Además, dichos clientes tienen que ser trabajador dependiente y tener experiencia en el sistema financiero por lo menos 1 año. Los filtros usados son:

- Que el mes de observación se encuentre entre octubre del 2020 y marzo del 2021 usando la variable CODMES.
- Que el registro corresponda a un trabajador dependiente de acuerdo con la base de datos comprada a Experian usando la variable TIP_TRAB_EXP=1.DEP.

- Que cuente con el segmento de estimador de ingresos entre los valores 2(Pasivos + RCC) y 3(solo RCC) usando la variable flg_seg_i_F in (2,3).
- Que el ingreso observado (última definición de modelamiento (MONTO_POND) y última definición de gestión (ING_14_EXT8)) se encuentre entre los 700 y 50000 soles.
- Que tenga saldo en el sistema financiero mayor a 100 soles en los últimos 12 meses usando la variable n_meses_100_12.

Se puede ver un ejemplo de este procedimiento ejecutado en SQL en la Figura9.

Figura 9

Códigos en SQL Para la Obtención del Universo de Modelamiento

```
where 202010<=CODMES<=202103 and
TIP_TRAB_EXP="1.DEP" and
flg_seg_i_F in (2,3) and
n_meses_100_12=12
and 700< MONTO_POND <50000 and 700< ING_14_EXT8 <50000
```

Para este objetivo se filtró a la población de clientes que se encuentren en la base PDH de RBM cuyo identificador fue el CODCLAVECIC. Además, se seleccionaron a los clientes con ingresos en los meses desde octubre 2020 y marzo 2021. Con estos filtros, la población a modelar cuenta con 727,945 registros de clientes PDH. Y se verificó que no exista duplicidad de clientes por mes. Para el modelo de estimador de ingresos para personas dependientes + RCC, además de contar con los filtros antes mencionados, se cuenta con las siguientes fuentes de variables.

- **Matriz de modelamiento:** Se compone de dos tablas mensuales (RCC y Resumen Saldo).
- **Agregados RCC:** Esta compuesta de 4 tablas (Tiers, Reprogramados SBS, Cuota SBS, Antigüedad en la SBS).
- **Empresa proveedora de información:** Fuente compuesta de la tabla de situación laboral y patrón vehicular.

- **Maestras demográficas:** Compuesta por las tablas de Relación cliente y manzana, información de persona natural, zona geográfica, relación cliente y empresa, AEMA, información del censo.
- **Datos del empleador:** Contiene información económica del cliente, CIIU del cliente e información de ventas del empleador.
- **Equifax:** Se extraen las variables asociadas al censo.
- **Sunedu:** Se extraen variables relacionadas al grado de educación obtenido.

Debido a que las fuentes se información, así como las tablas que lo componen y sus variables, son muy extensas, estas se encuentran detalladas en el Anexo A.

Luego de que se haya verificado las fuentes de información confiables y oficiales que cuenta el banco; y que corresponde al primer objetivo específico planteado al capítulo de introducción; se observó la etapa de segmentación.

La segmentación del modelo fue la primera etapa de selección, ya que segmentar una población se entiende también como la exigencia de generar un modelo diferenciado por cada grupo de interés. En esta etapa se buscaron variables de segmentación que permitan generar grupos de comportamiento diferenciado, donde el resultado fue, modelos diferentes para cada grupo de comportamiento. La metodología utilizada para la segmentación fue árboles de decisión y criterio experto para definir la segmentación en: ingresos bajos, ingresos medios e ingresos altos.

Se hicieron múltiples pruebas de segmentación para encontrar el mejor set de variables que nos ayude a definir los ingresos bajos, medios y altos. Finalmente se encontró variables balanceando información demográfica y RCC. Esta segmentación se basa en la disponibilidad de información de los clientes, de tal modo que se optimiza el uso de variables en aquellos clientes que realmente tienen poblados dichos campos. Se confirmó que, bajo este enfoque, se cumple con tener materialidad. Luego de que se observó y entendió la segmentación se procedió a replicar el análisis univariado, tratamiento de las variables categóricas y el análisis bivariado.

Análisis Univariado

Se espera que para que una variable sea considerada para el modelo final, debería tener pocos valores perdidos (missings). Cabe resaltar que, dado que la correlación es un indicador que se ve afectado por la presencia de outliers, todas las variables fueron acotadas al percentil 1% y 99% antes del cálculo de los indicadores de correlación.

Se realizó la evaluación de los missing mediante el indicador de Fill Rate, que el banco define como el total de valores missing de la variable como porcentaje total de registros de la base. Los umbrales para que una variable se mantenga o se descarte, se han definido con el fin de que no se descarte de manera drástica conceptos de variables o fuentes que podrían ser importantes al interactuar con otras variables. En la Tabla 3, se encuentran definidos los 4 criterios/umbrales aceptables máximos, para aceptar una variable.

Tabla 3

Umbrales Definidos Según Correlación & Fill Rate

Criterios
I CORR< 0.20 & FR < 0.05
II CORR < 0.15 & FR < 0.10
III CORR < 0.10 & FR < 0.20
IV CORR < 0.05 & FR < 0.40

Cabe recalcar que la idea es mantener un equilibrio entre la correlación y el porcentaje de missing; si se cuenta con un alto nivel de correlación, se debe tener bajo porcentaje de missing (criterio I); y viceversa, si se tiene baja correlación se acepta un porcentaje de missing alto (criterio IV).

Tratamiento de variables categóricas

Para ello, primero se transformó a las variables categóricas (mediante dos técnicas), luego se hizo la selección de éstas. En ese sentido, y debido que para diversas técnicas de machine learning, las variables no pueden contener caracteres como valores, se procede a transformar las variables en números mediante dos técnicas.

i) En la primera técnica se asignan valores numéricos a aquellas variables que son del tipo categórica, la transformación que se ha utilizado es asignar el promedio del ingreso de los registros que caen en dicha categoría en la muestra de entrenamiento, como por ejemplo se puede observar en la Tabla4, donde se categoriza la variable Estado Civil.

En caso de que alguna categoría tenga menos de 100 registros en su respectivo segmento o la categoría sea un missing, se procederá en asignar el valor del ingreso promedio del segmento que pertenece el cliente.

Las variables tratadas tienen el sufijo "_CAT1" para el segmento de ingresos bajos, "_CAT2" ´para ingresos medios y "_CAT3" para ingresos altos.

Tabla 4

Ejemplo de Tratamiento de Variable Categórica

Cliente	Variable Dependiente	Estado Civil	Estado Civil_CAT1
1	1200	SOLTERO	1400
2	1400	SOLTERO	1400
3	1600	SOLTERO	1400
4	2000	CASADO	2500
5	2500	CASADO	2500
6	3000	CASADO	2500

Cabe mencionar que las categorías deben discriminar en cada segmento. Por ejemplo, en la variable "Macrozona", el ingreso promedio de las categorías del segmento de ingresos altos es mayor al de ingresos medios y este también es mayor al de ingresos bajos.

ii) En el caso que la variable categórica cuente con muchas categorías, se procede a transformar la variable a través del promedio del target (conocido como "target statistics" o TS). En el caso que haya alguna categoría sin suficiente materialidad para asignar un promedio y además se cuente con otra variable de otro nivel agrupando dichas categorías, se pondera dependiendo de la materialidad de la categoría en ambas variables. Esta asignación de promedios se realiza mediante la siguiente fórmula:

$$S_i = \lambda(n_i)\frac{\sum_{k\in L_i} Y_k}{n_i} + (1-\lambda(n_i))\frac{\sum_{k=1}^{N_{TR}} Y_k}{n_{TR}}\,, \qquad \text{……..(Ecuación 2)}$$

Donde ∑Y/n es el promedio del target en una categoría y en el total de registros (TR). Lambda es una función que depende exclusivamente del número de registros con los que cuenta cada categoría en la muestra de entrenamiento. Esta función se define de la siguiente manera:

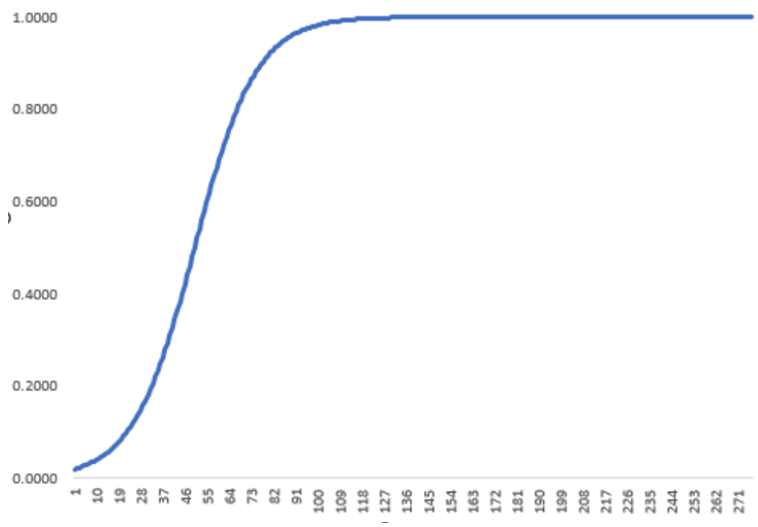
$$\lambda(n) = \frac{1}{1+e^{-\frac{(n-k)}{f}}}\,, \qquad \text{…….…….(Ecuación 3)}$$

Donde k es la mitad del número de observaciones que consideramos para tener un promedio confiable, en este caso está definido con un valor de 50 (cabe mencionar que este valor fue definido por el analista, es una definición empírica) con el fin de evitar el sobreajuste, y f es una constante que determina la pendiente de la curva, definido en este caso con el valor de 25.

Como es de esperarse, a valores muy pequeños de λ, la ponderación que se da al promedio de la categoría es muy baja, mientras que a valores cercanos a 100, la ponderación es cercana a 1; esto se puede observar en la Figura10. Para el cálculo del valor de n que es input de la función lambda se consideraron la cantidad de registros por categoría.

Figura 10

Gráfica de la Ecuación 3



Nota: El eje X es n, el eje Y es λ

En el caso de la ubicación geográfica del cliente, se consideró transformar la variable hasta nivel de distrito, en caso no se cuente con materialidad del distrito, se toma el promedio de la provincia, en caso la provincia no cuente con materialidad se toma el del departamento y finalmente si el departamento no cuenta con materialidad suficiente se toma el promedio de la población de su respectivo segmento. Este enfoque permite aprovechar la información hasta el nivel de distrito, al mismo tiempo que se evita el sobreajuste debido a que las categorías con

baja materialidad se colapsan hacia el promedio poblacional. Las variables que se trataron mediante dichos procedimientos son las siguientes: profesión, distrito y provincia.

Para tener un mejor entendimiento de este proceso, se presenta un ejemplo aplicado a la variable provincia en el segmento de ingresos altos, para ello es necesario contar con información preliminar de:

num_seg1(Nro. de registros que corresponden al segmento alto) = 133053

mnto_pond_seg1 (Promedio de la variable target/ dependiente en el segmento alto) = 4088.1

num_tot (Nro. de registros que corresponden a todos los segmentos) = 727945

mnto_pond_tot (Promedio de la variable target/ dependiente en todos los segmentos) = 2385.5

Como se ha mencionado anteriormente, 100, es el número de registros que mínimamente debe tener una categoría; y como se puede aprecia en la Tabla5 en la provincia de Ambo, solo se tiene 5 registros (n_p); entonces, se aplica la metodología de transformación a través del promedio target, considerando el siguiente nivel de provincia, que es departamento. Los valores de la ecuación 3 se ven reflejadas en peso prov, mientras que los valores de la ecuación 2 es el PROF_ST. Finalmente se obtienen los promedios target.

Tabla 5

Proceso de Transformación a través de la Variable Target

PROV	DEP	n_p	MONTO_POND_p	num_dp	MONTO_POND_d	peso prov	peso depa	PROF_ST
Alto Amazonas	Loreto	88	2467.2	469	2907.4	0.95	0.05	2487.3
Ambo	Huanuco	5	2239.2	283	3034.6	0.03	0.97	3013.4
Andahuaylas	Apurimac	37	5353.7	177	5886.9	0.26	0.74	5747.6

Nota:
n_p: Número de registros por provincia.
num_dp:Número de registros por departamento.
MONTO_POND_p: Promedio de la variable Target (dependiente) por provincia.
MONTO_POND_d: Promedio de la variable Target (dependiente) por departamento.
peso_prov: Es el λ calculado de provincia.
peso_depa: Es el λ calculado de departamento.
PROF_ST: Asignación del promedio con la transformación a través del promedio target.

Luego de tratar las variables categóricas se realizó una selección de estas variables para cada uno de los segmentos; para ello, primero se agruparon variables categóricas por afinidad y se

elige la variable con mayor asociación con el target y/o la variable con menor missing. Por ejemplo, como se puede observar en la Tabla6, las siguientes variables tienen la misma definición "Estado civil" y/o provienen de distintas fuentes, se seleccionó la variable con mayor asociación con el target y además también tiene menor missing que las demás variables; es decir, la variable TIPESTCIVIL_CAT3:

Tabla 6

Ejemplo de Selección de Variables por Afinidad y Correlación

Variables	Fuente de Información	Número de Missing	% de Missing	Correlación con el Target
TIPESTCIVIL_CAT3	ODS	30	0	0.185
dem_des_estadocivil1_cat3	AEMA	34	0	0.184
dem_des_estadocivil2_cat3	AEMA	133053	100%	0
dem_tip_estadocivilbcp_cat3	AEMA	34	0	0.184

Finalmente, dentro del tratamiento de variables categóricas está el proceso de Convolución de Variables, en esta etapa se crean variables que dependen de 2 o más variables categóricas con el fin de crear una nueva variable que recoja los efectos no lineales sobre la variable dependiente. Las variables finalistas posterior a la convolución fueron elegidas en base a sentido económico; éstas son las siguientes: edad (rango), situación de la casa, estado civil, género, novel educacional, actividad económica y N° de trabajadores.

Para transformarlas a un valor numérico, se le atribuye la media del ingreso a aquellas combinaciones de variables categóricas que cuenten con más de 100 registros. Como se puede observar en la Tabla7, a las variables convolucionadas con dos variables que tienen imputada una media se les asignó el prefijo "cv_".

Tabla 7

Listado de Variables Convolucionadas con su Prefijo

Variables originales		Convolución
Variable 1	Variable 2	
Estado Civil	Género	cv_estcivil_sexof
Estado Civil	Edad agrupada	cv_estcivil_edadn
Estado Civil	Nivel Educativo	cv_estcivil_nivedu
Estado Civil	Sectores	cv_estcivil_sector
Estado Civil	Región	cv_estcivil_region

Variables originales		Convolución
Variable 1	Variable 2	
Género	Edad agrupada	cv_sexo_edadn
Género	Nivel Educativo	cv_sexo_niveledu
Género	Sectores	cv_sexo_sector
Género	Región	cv_sexo_region
Edad agrupada	Nivel Educativo	cv_edadn_niveledu
Edad agrupada	Sectores	cv_edadn_sector
Edad agrupada	Región	cv_edadn_region
Nivel Educativo	Sectores	cv_nivedu_sector
Nivel Educativo	Región	cv_nivedu_region
Sectores	Región	cv_sector_region

Las variables convolucionadas deben cumplir un sentido económico, por ejemplo, se puede observar en la Tabla8, la siguiente variable "cv_estcivil_edadn" del segmento de ingresos medios, muestra que los casados con bienes separados (CBS) tienen en promedio más ingresos que los solteros y en cada subcategoría muestra que las personas mayores tienen mayor ingreso promedio que los más jóvenes:

Tabla 8

Ingresos Promedios Según Estado Civil y Edad

Estado civil	Edad		
	< 30 años	33-50 años	>50 años
Casado Bienes Separados	2847	3701	3767
Separado	2847	3051	3308
Casado	2536	3292	3174
Soltero	2387	2675	2688
Viudo	2847	2366	2310
conviviente	2069	2375	2375

Para poder elegir las variables convolucionadas en la modelación, se corrió una regresión lineal versus el target y se seleccionó a dichas variables donde se tenga una ganancia de R^2 mayor al 5% respecto al R^2 del target versus cada una de las variables originales:

También se ha construido otras dos variables convolucionadas con 4 variables:

- CONV_1_ZN: Combina las categorías de las siguientes variables:

i) sector,
ii) número de trabajadores de la empresa donde labora el cliente (micro, pequeñas, medianas y grandes empresas)
iii) edad
iv) macrozona.

- CONV_1_ED: Combina las categorías de las siguientes variables: i) sector, ii) número de trabajadores de la empresa donde labora el cliente (micro, pequeñas, medianas y grandes empresas), iii) edad y iv) nivel educativo.

Para estas variables convolucionadas también se evalúa el sentido económico.

Análisis Bivariado

En esta etapa se busca agrupar las variables que tienen una alta correlación y seleccionar la mejor variable de cada grupo o familia de variables de acuerdo con la asociación que tengan en relación con el target que se está tratando de predecir. En esta etapa se utilizan dos inputs. Primero se construye la matriz de correlaciones (Pearson) y se reordena de acuerdo con la correlación entre las distintas variables. Como segundo paso, se calcula la correlación lineal (Pearson) de cada variable con el target y de cada grupo de variable se selecciona la variable que tenga mayor correlación (en valor absoluto) para ser candidata en el modelo final. El umbral de correlación utilizado en la modelación fue de 0.6. Esta revisión se realizó para las variables continuas, a las cuales se le adicionaron variables discretas y variables convolucionadas.

Cabe resaltar que, dado que se utilizaron técnicas basadas en árboles, la correlación entre las variables (multicolinealidad) no es un problema por resolver en cuanto a la predicción del modelo, sin embargo, es deseable que, por motivos de parsimonia y facilidad de implementación, el modelo tenga sólo las variables necesarias para no perder poder predictivo.

Luego de replicar el análisis univariado, tratamiento de las variables categóricas y el análisis bivariado se realizó el muestreo estratificado para el modelo de entrenamiento (60%), muestra de prueba (20%) y la muestra para validación (20%). Estas tres muestras se estratifican considerando que no se repitan clientes en las distintas muestras. Posterior a ello se procedió a realizar la réplica de la etapa de selección de variables; a fin de verificar el correcto procedimiento del análisis, el mismo que corresponde al 2do objetivo específico planteado.

Selección de Variables

En la etapa de selección de variables, se realizan 3 pasos como parte de la metodología usada en el modelo, las cuales son: Sentido Económico I, Boruta y sentido Económico II. Para la revisión en la auditoría, se seleccionó el Top 5 variables, según Gain para verificar que, no presenten un sentido económico diferente al obtenido por la Unidad de Modelamiento. Ello permitirá validar el adecuado sentido económico de las variables. Dicha prueba se realizará utilizando los gráficos PDP (Partial Dependence Plot) que muestra la relación del output del modelo y la variable de análisis.

i) Sentido Económico I

Debido a que la metodología utilizada es un XGBoost, técnica de machine learning basada en árboles, las relaciones entre las variables y el target no es necesariamente lineal. Frente a esto se utilizó partial dependency plots (PDP), los cuales extrapolan las curvas de la variable vs la predicción del modelo de cada observación para luego promediar las curvas. Extrapolan las curvas al probar el modelo cambiando el valor de las variables (X) y revisando los cambios que esta produce en la estimación (Y).

Se revisaron los PDPs para cada una de las principales variables por segmento, identificando su sentido (positivo, negativo), además de identificando aquellas que tenían diferentes sentidos según intervalos; es decir, aquellos que no tenían un sentido definido respecto a la predicción del modelo. Luego de eso se eliminaron aquellas variables que no cumplían con el sentido de familia.

Luego del análisis bivariado quedaron un poco más de 100 variables por cada segmento. Se eliminaron entre 15 y 49 variables (depende del segmento) por no tener un sentido económico definido. Por ejemplo, la edad debería de tener un sentido positivo contra el target pero los PDP nos muestra todo lo contrario cuando se supone que a mayor edad debería de predecir mayor ingreso.

ii) Boruta

Para la selección de variables se utilizó el algoritmo Boruta con un modelo base de XGBoost. El algoritmo Boruta consiste duplicar las variables y aleatorizar las observaciones en cada una de las variables duplicadas. Se corre un modelo base, en este caso un XGBoost, y se compara la importancia entre la variable con las observaciones

originales y la variable con las observaciones aleatorizadas. Mientras más alta es la diferencia entre la variable original y la variable con las observaciones en Random mejor. Las variables que tienen diferencia significativa según el Z test entra en un vector de "hits". Boruta itera este proceso y va eliminado o confirmando las distintas variables ordenándolas, comparando el número de veces en las que las variables originales son mejores que las aleatorizadas.

iii) Sentido Económico II

En esta etapa se realiza la misma metodología enunciada en el ítem de "sentido económico" solo que esta vez se desarrolla la revisión para todas las variables finales ya que se cuenta con menor cantidad de variables respecto a la cantidad de la salida del bivariado. En la revisión se descartan aquellas variables que:

- No cuenten con un sentido económico claro, o que no se considere relevante para predecir el ingreso.
- O, cuya correlación lineal sea contraria al sentido esperado, dicha revisión se realiza para cada segmento del modelo y se realiza en conjunto por expertos de los equipos de MMGR, VI, y RBM.

Asimismo, se identificaron aquellas variables cuya importancia era muy pequeña, dado que no permitía generar PDPs ya que los percentiles estaban muy cercanos entre sí. Las variables sobrantes se agruparon en familias de variables para confirmar el sentido de la familia con el negocio.

Técnica de Hyperparameters Tunning

En esta etapa se busca optimizar los hiperparámetros mediante iteraciones para encontrar el set de hiperparámetros que consigan el menor valor de la función de pérdida en la data de validación. En el caso del estimador de ingresos, se utilizó un modelo machine learning basado en árboles llamado XGBoost. El método que se utilizó para optimizar estos hiperparámetros de configuración fue el "grid search". Esta metodología consiste en realizar una selección de valores de un conjunto de opciones especificadas manualmente en el espacio de los hiperparámetros.

Los hiperparámetros tuneados fueron 7 que se explicarán a continuación: Learning rate (Eta), Max Depth, Colsample by tree, Subsample, Min child, Lambda y Gamma.

- Learning rate: indica cuánto va a aprender el algoritmo en cada ronda. Es representado por la letra "eta", si esta es cercana a 0 será lento, si es cercano a 1 el modelo aprenderá rápido (usualmente se utiliza un eta entre 0.01 y 0.05). Este hiperparámetro debe estar balanceado con el número de iteraciones. Mientras más pequeño el learning rate debe haber un número más grande de iteraciones para evitar que caiga en mínimos locales. Se probaron los valores 0.05, 0.1 y 0.15.
- Max Depth: Especifica cuántos cortes de hojas se quieren hacer. Como se mencionaba, siempre se parte en 2, pero con la profundidad se especifica el número de veces de esa partición doble. Su finalidad es reducir el sobreajuste controlando la profundidad del árbol. Se probaron los valores 11, 13 y 15.
- Colsample by tree: indica que porcentaje de las variables serán seleccionadas en cada iteración. Este hiperparámetro permite mejorar la velocidad de entrenamiento y ayuda a prevenir el overfitting. Se probaron los valores 0.6, 0.8 y 1.
- Subsample: parecido al hiperparámetro colsample, subsample especifica el porcentaje de las observaciones seleccionadas sin remplazo aleatoriamente en cada iteración. Este hiperparámetro también permite mejorar la velocidad de entrenamiento y ayuda a prevenir el overfitting. Se probaron los valores 0.6, 0.8 y 1.
- Min_child: Tamaño mínimo de una hoja del árbol. Se probaron los valores 60, 120 y 360.
- Lambda: indica el valor de la L_1, regularización en términos de peso. Incrementar este valor vuelve al modelo más conservador. Se probaron los valores 0 y 1.
- Gamma: Se requiere una reducción mínima de pérdidas para realizar una nueva partición en un nodo hoja del árbol. Cuanto más grande sea gamma, más conservador será el algoritmo. Se probaron los valores 0 y 1.

Finalmente se ejecutaron 972 iteraciones y en cada una de ellas se guardó el valor del R^2 de la base de test y train y el porcentaje de aciertos ±25% (indicador de precisión). Se muestra a continuación las Lógicas de programación en RStudio del análisis de optimización de hiperparámetros; donde se evidencia el uso de los 7 hiperparámetros, en las Figuras 11, 12 y 13.

Figura 11

Códigos en RStudio de la Optimización de Hiperparámetros en Ingresos Bajos

```
# DMatrices generation
DTrain <- xgb.DMatrix(data = data.matrix(dt_train[, explicative_vars, with = F]),
                      label = data.matrix(dt_train[, target, with = F]))
DDev <- xgb.DMatrix(data = data.matrix(dt_test[, explicative_vars, with = F]),
                    label = data.matrix(dt_test[, target, with = F]))

#Grid Search

searchGridSubCol <- expand.grid(subsample = c(0.6,0.8,1),
                                colsample_bytree = c(0.6,0.8,1),
                                max_depth = c(11,13,15),
                                min_child = c(60,120,360),
                                eta = c(0.05,0.1,0.15),
                                gamma = c(0,1),
                                lambda =  c(0,1)
)
ntrees <- 500

system.time(
  rmseErrorsHyperparameters <- apply(searchGridSubCol, 1, function(parameterList){
    #Extract Parameters to test
    currentSubsampleRate <- parameterList[["subsample"]]
    currentColsampleRate <- parameterList[["colsample_bytree"]]
    currentDepth <- parameterList[["max_depth"]]
    currentEta <- parameterList[["eta"]]
    currentMinChild <- parameterList[["min_child"]]
    currentGamma <- parameterList[["gamma"]]
    currentLambda <- parameterList[["lambda"]]
    set.seed(1234)
    xgboostModel <- xgb.train(data =  DTrain, nrounds = ntrees, watchlist = list(train=DTrain, test=DDev), showsd = TRUE,
                              verbose = 1, "eval_metric" = "rmse", maximize = F,
                              "objective" = "reg:linear", "max.depth" = currentDepth, "eta" = currentEta,
                              "subsample" = currentSubsampleRate, "colsample_bytree" = currentColsampleRate
                              , print_every_n = 10, "min_child_weight" = currentMinChild, "gamma" = currentGamma,"lambda" = currentLambda,
                              booster = "gbtree",
                              early_stopping_rounds = 10)

    xvalidationScores <- as.data.frame(xgboostModel$evaluation_log)
    bestiter <- xgboostModel$best_iteration
```

Figura 12

Códigos en RStudio de la Optimización de Hiperparámetros en Ingresos Medios

```
# DMatrices generation


DTrain <- xgb.DMatrix(data = data.matrix(dt_train[, explicative_vars, with = F]),
                      label = data.matrix(dt_train[, target, with = F]))
DDev <- xgb.DMatrix(data = data.matrix(dt_test[, explicative_vars, with = F]),
                    label = data.matrix(dt_test[, target, with = F]))

#Grid Search

searchGridSubCol <- expand.grid(subsample = c(0.6,0.8,1),
                                colsample_bytree = c(0.6,0.8,1),
                                max_depth = c(11,13,15),
                                min_child = c(60,120,360),
                                eta = c(0.05,0.1,0.15),
                                gamma = c(0,1),
                                lambda =  c(0,1)
)
ntrees <- 500

system.time(
  rmseErrorsHyperparameters <- apply(searchGridSubCol, 1, function(parameterList){

    #Extract Parameters to test
    currentSubsampleRate <- parameterList[["subsample"]]
    currentColsampleRate <- parameterList[["colsample_bytree"]]
    currentDepth <- parameterList[["max_depth"]]
    currentEta <- parameterList[["eta"]]
    currentMinChild <- parameterList[["min_child"]]
    currentGamma <- parameterList[["gamma"]]
    currentLambda <- parameterList[["lambda"]]
    set.seed(1234)
    xgboostModel <- xgb.train(data =  DTrain, nrounds = ntrees, watchlist = list(train=DTrain, test=DDev), showsd = TRUE,
                              verbose = 1, "eval_metric" = "rmse", maximize = F,
                              "objective" = "reg:linear", "max.depth" = currentDepth, "eta" = currentEta,
                              "subsample" = currentSubsampleRate, "colsample_bytree" = currentColsampleRate
                              , print_every_n = 10, "min_child_weight" = currentMinChild, "gamma" = currentGamma,"lambda" = currentLambda,
                              booster = "gbtree",
                              early_stopping_rounds = 10)

    xvalidationScores <- as.data.frame(xgboostModel$evaluation_log)
    bestiter <- xgboostModel$best_iteration
```

Figura 13

Códigos en RStudio de la Optimización de Hiperparámetros en Ingresos Altos

```
searchGridSubCol <- expand.grid(subsample = c(0.6,0.8,1),
                                colsample_bytree = c(0.6,0.8,1),
                                max_depth = c(11,13,15),
                                min_child = c(60,120,360),
                                eta = c(0.05,0.1,0.15),
                                gamma = c(0,1),
                                lambda =  c(0,1)
)
ntrees <- 500

system.time(
  rmseErrorsHyperparameters <- apply(searchGridSubCol, 1, function(parameterList){

    #Extract Parameters to test
    currentSubsampleRate <- parameterList[["subsample"]]
    currentColsampleRate <- parameterList[["colsample_bytree"]]
    currentDepth <- parameterList[["max_depth"]]
    currentEta <- parameterList[["eta"]]
    currentMinChild <- parameterList[["min_child"]]
    currentGamma <- parameterList[["gamma"]]
    currentLambda <- parameterList[["lambda"]]
    set.seed(1234)
    xgboostModel <- xgb.train(data =  DTrain, nrounds = ntrees, watchlist = list(train=DTrain, test=DDev), showsd = TRUE,
                              verbose = 1, "eval_metric" = "rmse", maximize = F,
                              "objective" = "reg:linear", "max.depth" = currentDepth, "eta" = currentEta,
                              "subsample" = currentSubsampleRate, "colsample_bytree" = currentColsampleRate
                              , print_every_n = 10, "min_child_weight" = currentMinChild, "gamma" = currentGamma,"lambda" = currentLambda,
                              booster = "gbtree",
                              early_stopping_rounds = 10)

    xvalidationScores <- as.data.frame(xgboostModel$evaluation_log)
    bestiter <- xgboostModel$best_iteration
```

IV. RESULTADOS Y DISCUSIÓN

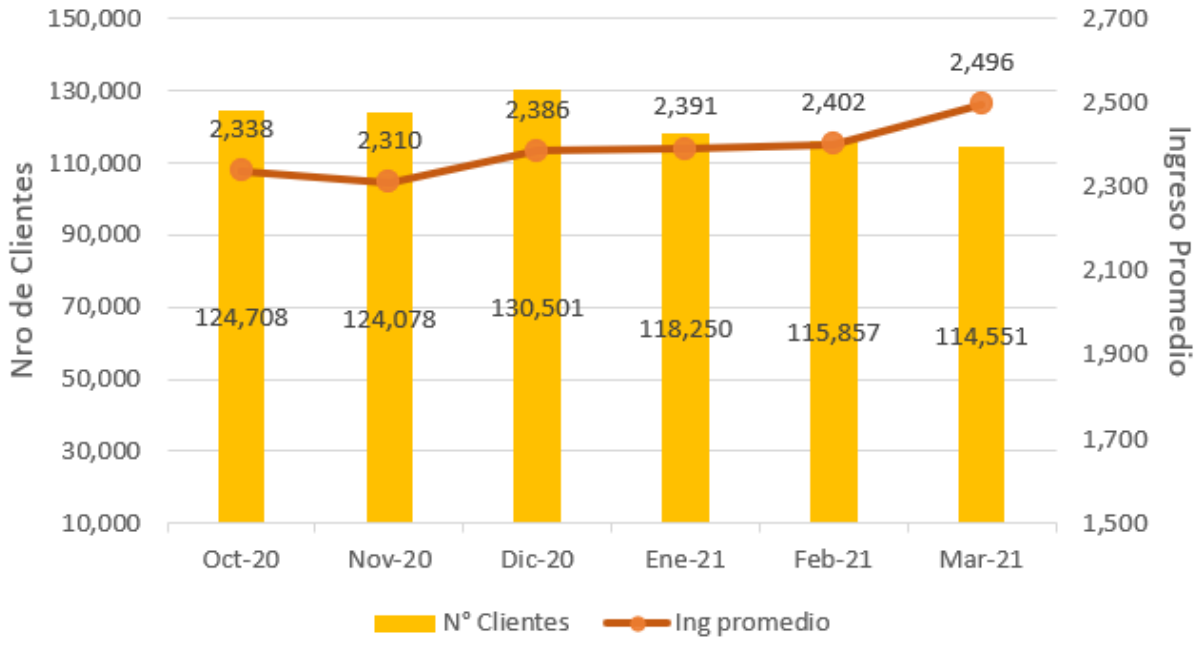
4.1 Definición del Universo

Se realizó el proceso de definición del universo; donde finalmente se obtiene la tabla única de modelamiento que cuenta con 727,945 registros y 1,463 variables; donde la variable dependiente es: MONTO_POND y en el banco se le denomina Target. Posterior a obtener la base de modelamiento se procedió a realizar algunas estadísticas descriptivas a la base con el fin de conocer el comportamiento del ingreso promedio y el Nro. de clientes en los meses de evaluación.

Se puede observar en la Figura14, que durante los meses de evaluación (octubre 2020 – marzo 2021) el número de clientes tiene una tendencia decreciente marcada en los meses de enero a marzo; mientras que el ingreso promedio tiende a crecer.

Figura 14

Número de Clientes e Ingreso Promedio de la Población por Mes



Nota: El ingreso promedio en soles.

Por otro lado, con respecto a la cantidad de registros por PDH, cabe señalar que, para la base de modelación, no necesariamente todos los clientes tienen ingresos PDH en los últimos 12 meses, sino también clientes que hayan tenido abonos de por lo menos 8 de los últimos 12 meses (PDH 8 de 12); asimismo, abonos en por lo menos 6 de los últimos 9 meses (PDH 6 de 9); en ese sentido, como se puede observar en la Tabla9, la base cuenta con un 67% de clientes que han recibido 12 pagos de haberes en los últimos 12 meses, mientras que se cuenta solo con un 5% de clientes que han recibido pago en al menos 6 de los 9 meses. Esta proporción cuenta con materialidad, debido a que, en los meses de evaluación, esta proporción se mantiene o no hay un cambio brusco en el comportamiento.

Tabla 9

Cantidad de Registros por Tipo de PDH

Mes	Oct-20	Nov-20	Dic-20	Ene-21	Feb-21	Mar-21	Total	%
NPH0 (12 de 12)	85,434	84,825	85,674	79,574	77,179	74,474	487,160	67%
NPH1 (8 de 12)	35,482	34,450	37,370	31,056	31,180	32,475	202,013	28%
NPH2 (6 de 9)	3,792	4,803	7,457	7,620	7,498	7,602	38,772	5%
Total	124,708	124,078	130,501	118,250	115,857	114,551	727,945	100%

Con respecto al ingreso promedio por tipo de PDH, se observa una diferenciación entre los diferentes meses de pago de haberes, es decir los clientes que han percibido su pago de haberes en los 12 meses, tienen un ingreso promedio mayor a los que han percibido sus ingresos en 8 de los 12 últimos meses; y los clientes que han percibido pago en 6 de los 9 meses tienen aún un ingreso menor a los otros dos grupos. Se puede observar esto, en la Tabla10, donde esta misma diferenciación se mantiene en los meses de evaluación.

Tabla 10

Promedio de Ingreso por Tipo PDH

Mes	Oct-20	Nov-20	Dic-20	Ene-21	Feb-21	Mar-21
NPH0 (12 de 12)	2,497	2,475	2,563	2,580	2,607	2,713
NPH1 (8 de 12)	2,010	1,983	2,092	2,055	2,028	2,113
NPH2 (6 de 9)	1,839	1,761	1,820	1,784	1,858	1,996
Total	2,338	2,310	2,386	2,391	2,402	2,496

Nota: El ingreso promedio en soles

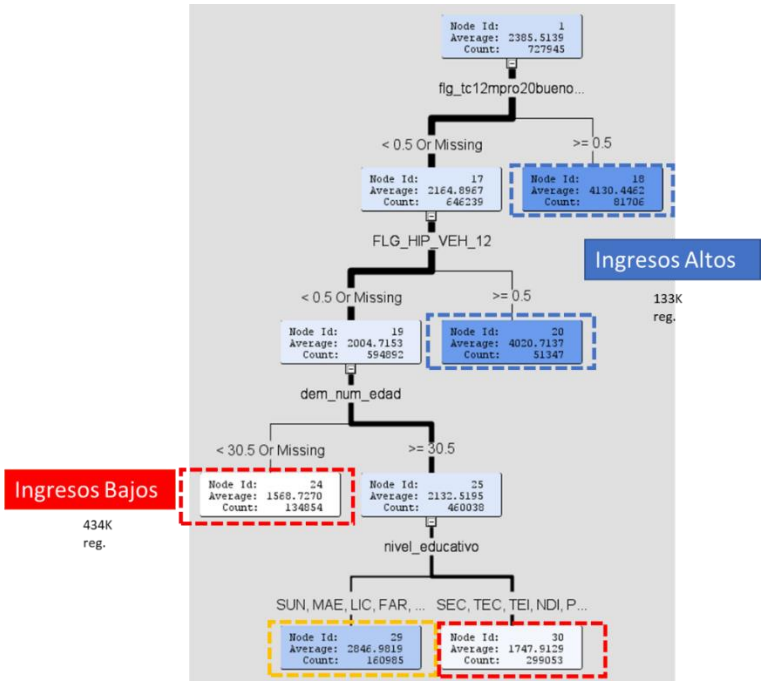
4.2 Segmentación

En la etapa de segmentación mediante árboles de decisión se obtuvo las siguientes variables para definir cada segmento; esto se puede visualizar de forma gráfica en la Figura15; sin embargo, a continuación, se detalla las variables de cada segmento:

1) **Ingresos Bajos:** Este segmento tiene en promedio ingresos de S/ 1,692 con un 60% de toda la población de modelación. Las variables que definen a este segmento son: Promedio de Línea Tarjeta de Crédito últimos 12 meses< S/ 20 mil o con mal comportamiento de pago. Y sin crédito hipotecario ni vehicular últimos 12 meses.
   y (edad <= 30 años o con educación primaria, secundaria, técnica, analfabeto).
2) **Ingresos Medios:** Este segmento tiene en promedio ingresos de S/ 2,847 con un 22% de toda la población de modelación. Las variables que definen a este segmento son: Promedio de línea tarjeta de crédito últimos 12 meses< S/ 20 mil o con mal comportamiento de pago. Y sin crédito hipotecario ni vehicular últimos 12m. Edad > 30 años y con educación superior, doctorado, bachiller, titulado, licenciado, maestría, Fuerzas armadas, o perteneciente a la base SUNEDU.
3) **Ingresos Altos:** Este segmento tiene en promedio ingresos de S/ 4,088 con un 18% de toda la población de modelación. Las variables que definen a este segmento son: Promedio Línea Tarjeta de Crédito últimos 12 meses >= S/ 20 mil con buen comportamiento de pago en los últimos 12 meses o con tenencia de crédito vehicular o hipotecaria en los últimos 12meses.

Figura 15

Gráfico del Árbol de Decisión diferenciados por Segmentos

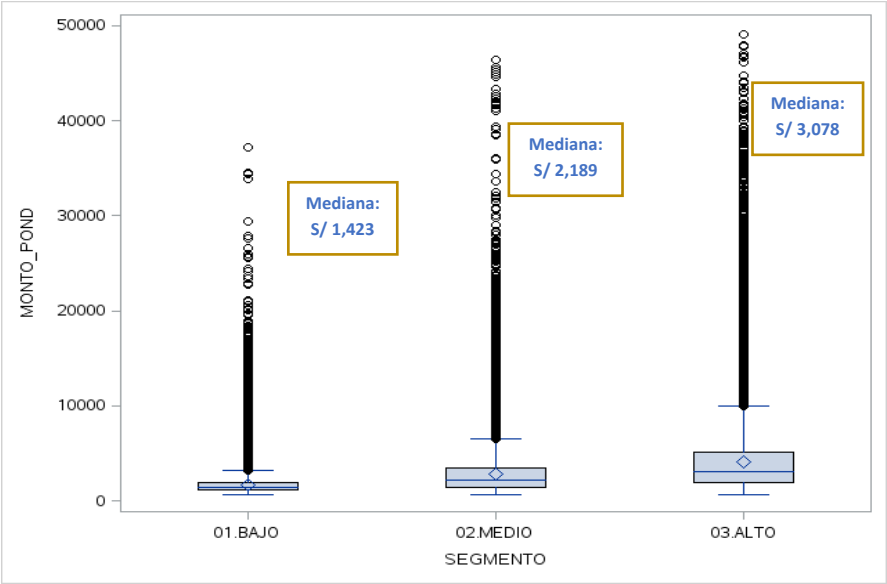


Nota: Gráfico de salida en SAS.

En la Figura16 se puede observar que el segmento de ingresos altos tiene una concentración de clientes con ingresos altos respecto a los otros dos segmentos, entonces cada segmento tiene un nivel de ingreso diferenciado.

Figura 16

Gráfico de Cajas Diferenciado por Segmentos



Nota: Gráfico de salida en SAS.

4.3 Análisis Univariado

Luego de aplicar los criterios y umbrales definidos en el análisis univariado, a continuación, se muestran las salidas en SAS del proceso de análisis univariado que se detalla en la Figura 17, 18 y 19 en los segmentos de ingresos bajos, medios y altos respectivamente.

Figura 17

Gráfico del proceso en SAS del Análisis Univariado en Segmento

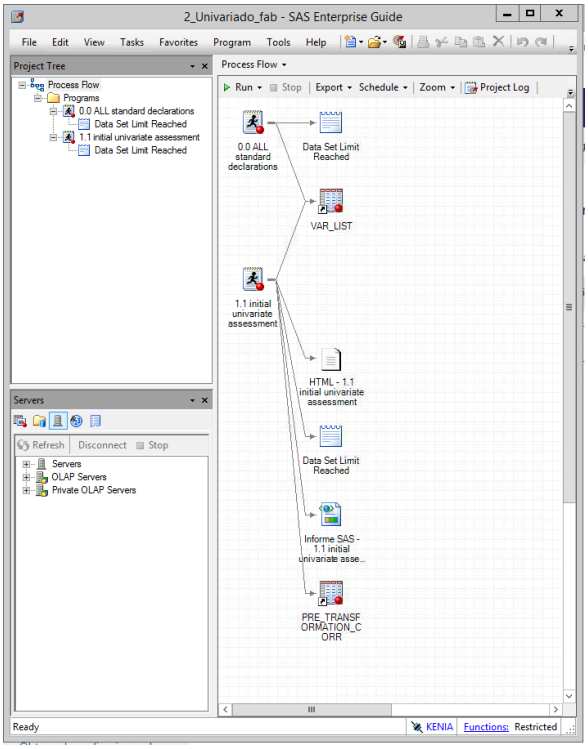


Figura 18

Gráfico del proceso en SAS del Análisis Univariado en Segmento Medio

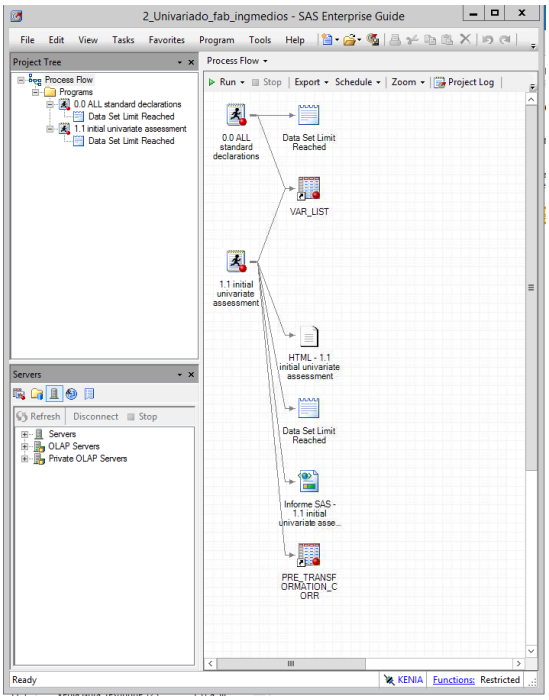
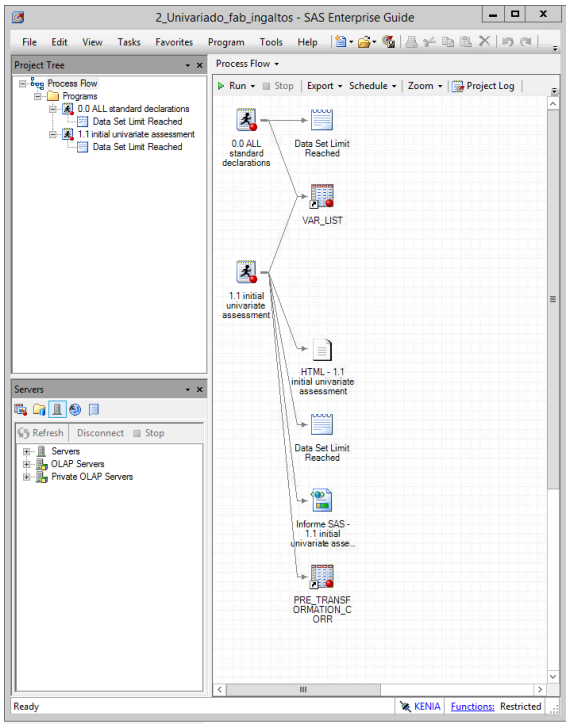


Figura 19

Gráfico del proceso en SAS del Análisis Univariado en Segmento



Finalmente se muestra en la Tabla11 el número de variables finales que se mantienen en cada uno de los segmentos, después del análisis univariado.

Tabla 11

Número de Variables Finales por Segmento Después del Análisis Univariado

Segmento	Variables Univariado
Ingresos Bajos	716
Ingresos Medios	1,013
Ingresos Altos	749

4.4 Tratamiento de variables categóricas

Los resultados del tratamiento de las variables categóricas, después de aplicar las técnicas de tratamiento, selección de variables y la convolución de estas se muestran a continuación. Para las técnicas de tratamiento y selección de variables se mostraron con ejemplos de algunas variables como se realizaron estos procedimientos; por otro lado, para la convolución de variables sí se cuenta con un detalle de resultados que es más específico y se detalla a continuación.

Como se mencionó en el capítulo anterior, la convolución de variables usa una regresión lineal, donde se seleccionó a dichas variables donde se tenga una ganancia de R^2 mayor al 5% respecto al R^2 del target versus las variables originales. Los resultados se pueden visualizar por segmentos (Bajo, Medio y Alto). Entonces se puede observar en la Tabla 12, que para el segmento de ingresos bajos se seleccionó dos variables convolucionadas: i) género con la edad y ii) edad con el sector. Para el segmento de ingresos medios se seleccionó dos variables convolucionadas: i) estado civil con la edad y ii) edad con el sector. Y finalmente para el segmento de ingresos altos se seleccionó 4 variables convolucionadas: i) género con la edad, ii) estado civil con el sector, iii) edad con el sector y iv) nivel educativo con el sector.

Tabla 12

Comparativo de R^2 en Regresión Lineal de las Variables Individuales y Convolucionadas

Convolución	Ingreso Bajo (R^2)			Ingreso Medio (R^2)			Ingreso Alto (R^2)		
	Ind.	Conv.	Var.	Ind.	Conv.	Var.	Ind.	Conv.	Var.
cv_estcivil_sexof	0.043	0.044	3%	0.023	0.023	0%	0.037	0.039	4.8%
cv_estcivil_edadn	0.018	0.019	4.4%	0.020	0.021	5.1%	0.042	0.044	4.5%
cv_estcivil_nivedu	0.036	0.031	-13%	0.042	0.042	0%	0.107	0.107	0%
cv_estcivil_sector	0.124	0.126	2%	0.067	0.066	0%	0.081	0.086	7%
cv_estcivil_region	0.019	0.018	-5%	0.019	0.019	1%	0.035	0.035	-2%
cv_sexo_edadn	0.036	0.040	12%	0.011	0.011	2%	0.025	0.026	6%
cv_sexo_niveledu	0.049	0.046	-6%	0.032	0.033	4%	0.090	0.092	2%
cv_sexo_sector	0.132	0.138	4.6%	0.057	0.058	2%	0.054	0.057	4.4%
cv_sexo_region	0.031	0.032	1%	0.008	0.008	0%	0.006	0.006	4%
cv_edadn_niveledu	0.033	0.031	-6%	0.030	0.031	3%	0.099	0.099	0%
cv_edadn_sector	0.119	0.125	5.03%	0.056	0.060	7%	0.068	0.075	9%
cv_edadn_region	0.012	0.011	-3%	0.006	0.006	0%	0.021	0.021	-1%
cv_nivedu_sector	0.124	0.110	-11%	0.069	0.070	1%	0.120	0.127	6%
cv_nivedu_region	0.021	0.021	1%	0.026	0.027	1%	0.083	0.083	0%
cv_sector_region	0.113	0.111	-2%	0.053	0.054	3%	0.053	0.053	1%

Nota: Ind. = Individual, Conv.= Convolución, Var.=Variación.

En la convolución de variables se usa la regresión lineal, la misma que debe de cumplir los supuestos de linealidad y normalidad; en ese sentido se evaluó el cumplimiento de los supuestos, encontrándose que no se cumple, como se puede observar en el Anexo B.

Por otro lado, en esta etapa, también se evalúan el sentido económico de las variables convolucionadas, en la Tabla13, se tiene como resultado de la variable “Sectores” evaluado con el Rango Edad de los clientes que pertenecen al sector minería e hidrocarburos o construcción tienen en promedio más ingresos que el sector agropecuario; por otro lado, dentro de cada sector, el promedio de ingresos de las personas que pertenecen a la categoría de mayor a 50 años es mayor a la de los jóvenes (Menor <30 años). De esta manera se evalúan diferentes combinaciones de variables que expliquen un buen sentido económico; es decir, que tengan un sentido lógico.

Tabla 13

Ingreso Promedio Según sector y Rango de Edad

Sector	Rango edad			Total general
	Menor < 30	33 - 50	Mayor a 50	
Minería e Hidrocarburos	2,371	2,957	3,343	2,890
Construcción	2,237	2,637	2,597	2,490
Electricidad y agua	2,535	2,448	2,165	2,383
Transporte-Aéreo	1,851	2,979	1,692	2,174
Manufactura-No esencial 1	1,756	1,985	2,163	1,968
Manufactura-Esencial 2	1,685	1,930	2,253	1,956
Servicios-No esencial 1	1,819	1,962	2,039	1,940
Servicios-Esencial	1,859	1,948	1,928	1,912
Transporte-Logística	1,608	1,933	2,068	1,870
Pesca	1,570	1,747	1,997	1,771
Manufactura-Esencial 1	1,554	1,740	1,834	1,709
Comercio-No esencial	1,533	1,779	1,773	1,695
Otros	1,685	1,692	1,692	1,690
Comercio-Esencial 2	1,471	1,773	1,764	1,669
Transporte-Otros	1,455	1,734	1,452	1,547
Comercio-Esencial 1	1,323	1,548	1,581	1,484
Servicios-No esencial 2	1,442	1,477	1,479	1,466
Hoteles y Restaurantes	1,439	1,523	1,431	1,464
Manufactura-No esencial 2	1,356	1,488	1,470	1,438
Agropecuario	1,303	1,361	1,335	1,333
Total general	1,692	1,932	1,903	1,842

4.5 Análisis Bivariado

Como resultado del análisis bivariado se puede observar en la Tabla17 que el número de variables en cada segmento no excede de los 110; además, entre dichas variables, se verifica que no existe correlación mayor a 60%. En las Tablas 14, 15 y 16, se muestran algunas variables en la matriz de correlaciones en los diferentes segmentos:

Tabla 14

Correlación de las primeras 8 Variables de Ingresos Bajos

Variables	monto_pond	conv_1_ed	conv22_ing_bajo	conv32_ing_bajo	ubicacion_cat1	dem_cod_ciiu_cat1	cuota_rcc_max_12	flgsexo_cat1
monto_pond	1.000	0.341	0.263	0.255	0.242	0.256	0.244	0.171
conv_1_ed	0.341	1.000	0.163	0.166	0.234	0.261	0.136	0.077
conv22_ing_bajo	0.263	0.163	1.000	0.321	0.160	0.132	0.187	0.012
conv32_ing_bajo	0.255	0.166	0.321	1.000	0.139	0.133	0.328	0.072
ubicacion_cat1	0.242	0.234	0.160	0.139	1.000	0.162	0.109	0.018
dem_cod_ciiu_cat1	0.256	0.261	0.132	0.133	0.162	1.000	0.102	0.106
cuota_rcc_max_12	0.244	0.136	0.187	0.328	0.109	0.102	1.000	0.049
flgsexo_cat1	0.171	0.077	0.012	0.072	0.018	0.106	0.049	1.000

Tabla 15

Correlación de las primeras 8 Variables de Ingresos Medios

Variables	monto_pond	mtototdeu_d_i_max24_pj	conv_1_zn	mto_profesion	cuota_compras_f_max_12	dem_des_mejorvehiculo_cat2	cuota_rcc_sum_12	dem_cod_ciiu_cat2
monto_pond	1.000	0.314	0.329	0.215	0.212	0.265	0.199	0.228
mtototdeu_d_i_max24_pj	0.314	1.000	0.191	0.142	0.343	0.213	0.455	0.143
conv_1_zn	0.329	0.191	1.000	0.101	0.160	0.231	0.099	0.205
mto_profesion	0.215	0.142	0.101	1.000	0.080	0.093	0.090	0.136
cuota_compras_f_max_12	0.212	0.343	0.160	0.080	1.000	0.159	0.098	0.059
dem_des_mejorvehiculo_cat2	0.265	0.213	0.231	0.093	0.159	1.000	0.107	0.094
cuota_rcc_sum_12	0.199	0.455	0.099	0.090	0.098	0.107	1.000	0.070
dem_cod_ciiu_cat2	0.228	0.143	0.205	0.136	0.059	0.094	0.070	1.000

Tabla 16

Correlación de las primeras 8 Variables de Ingresos Altos

Variables	monto_pond	mtodeudamax24_ind	conv_1_zn	conv_1_ed	dem_cod_ciiu_cat3	dem_des_mejorvehiculo_cat3	CUOTA_COMPRAS_F_max_12	mto_profesion
monto_pond	1.000	0.390	0.379	0.386	0.262	0.287	0.206	0.297
mtodeudamax24_ind	0.390	1.000	0.246	0.255	0.122	0.239	0.261	0.238
conv_1_zn	0.379	0.246	1.000	0.559	0.259	0.269	0.144	0.197
conv_1_ed	0.386	0.255	0.559	1.000	0.298	0.215	0.079	0.420
dem_cod_ciiu_cat3	0.262	0.122	0.259	0.298	1.000	0.104	0.056	0.160

dem_des_mejorvehiculo_cat3	0.287	0.239	0.269	0.215	0.104	1.000	0.164	0.164
cuota_compras_f_max_12	0.206	0.261	0.144	0.079	0.056	0.164	1.000	0.098
mto_profesion	0.297	0.238	0.197	0.420	0.160	0.164	0.098	1.000

Tabla 17

Número de Variables Finales por Segmento Después del Análisis Bivariado

Segmento	Variables Bivariado
Ingresos Bajos	101
Ingresos Medios	107
Ingresos Altos	106

4.6 Selección de Variables

i) Sentido Económico

La revisión del sentido económico del Top 5 principales variables, se realizó según Gain, y se realizó para cada segmento del modelo.

Ingreso Bajo: Los resultados de la validación del sentido económico de las variables obtenidas del análisis fueron, una lista de variables con su importancia en el modelo con R^2 de 31.41%; estas variables se encuentran enlistadas en la Tabla18.

Tabla 18

Variables Obtenidos por la Unidad de Modelamiento en el Segmento Bajo

N°	Variables	Gain
1	CONV_1_ED	27.27%
2	conv22_ING_BAJO	7.63%
3	ubicacion_cat1	7.25%
4	conv32_ING_BAJO	6.33%
5	CUOTA_RCC_max_12	5.34%
6	dem_cod_ciiu_cat1	5.13%
7	FLGSEXO_cat1	3.77%
8	CUOTA_COMPRAS_F_max_12	3.05%
9	conv52_ING_BAJO	2.37%
10	dem_des_mejorvehiculo_cat1	2.32%
11	RT_REFRI_HOGARES	2.13%
12	FLG_MIN_CONST	2.10%
13	SOW_MTODEUDATOT_T1	2.09%
14	CUOTA_OTROS_F_max_12	1.64%
15	mto_profesion	1.52%

N°	Variables	Gain
16	MTODEU_CEF_T2	1.49%
17	RT_TIENE_DOCUMENTO	1.43%
18	MTODEU_CEF_T1	1.35%
19	CTDEMPREPORTADOCLIMED24	1.31%
20	FLG_DISEF24	1.26%
21	PORC_REPROG_TOT	1.21%
22	CTD_PER_MENOR_18	1.18%
23	PRC_PER_MENOR_18	1.12%
24	nivel_educativo_f_cat1	1.10%
25	TIPSITUACIONCASA_cat1	0.75%
26	ING_D_20	0.74%
27	MONTOADE_ACT3_MAX3_S_HIP	0.74%
28	conv61_ING_BAJO	0.69%
29	DEUDIR1_DEUTOT1	0.48%
30	MTODEUDAVCDO_MAX24	0.47%
31	CTD_EMPRESAS_24	0.44%
32	MONTOADE_ACT_MAX3_IND	0.44%
33	TIPESTCIVIL_cat1	0.43%
34	SF_DIS_CAL_CPP12	0.36%
35	CTDPROD_T1	0.36%
36	PRED_EQX2021_cat1	0.33%
37	dem_flg_basesunat	0.33%
38	flg_ind12	0.31%
39	CTD_PRODTOTAL_6	0.31%
40	SOW_MTODEU_TC_T3	0.29%
41	FLG_DISEF3	0.28%
42	flg_inf12	0.22%
43	MTODEU_TC_T3	0.20%
44	DEUDA_NEG_RCC	0.15%
45	FLG_BAJA_RUC	0.10%
46	CUOTA_HP_FINAL_min_6	0.07%
47	flg_lintcpro20_12m	0.07%
48	MTODEU_CEF_T3	0.03%
49	CTD_CONSUMO	0.02%
50	CUOTA_VEH_FINAL_sum_6	0.00%
51	CUOTA_HP_FINAL_sum_6	-
52	CUOTA_VEH_FINAL_min_12	-

A continuación, se presentan los resultados de los Top 5 principales variables de los Partial Dependence Plot (PDP) del segmento Bajo, algoritmo que fue usado para expresar si una variable está afectando positiva o negativamente al target (variable respuesta). La lógica de comparación de los resultados de auditoría (gráficos de la izquierda) vs los resultados que obtuvo la unidad de modelamiento (gráficos de la derecha), son para verificar que los resultados mostrados por modelamiento sean los verídicos, y se haya realizado un correcto análisis estadístico en la selección de variables, asegurando un adecuado nivel de contribución de las variables; además, de validar que los procedimientos metodológicos utilizados cuenten con sustento analítico, y así cumplir con los objetivos específicos 2 y 3.

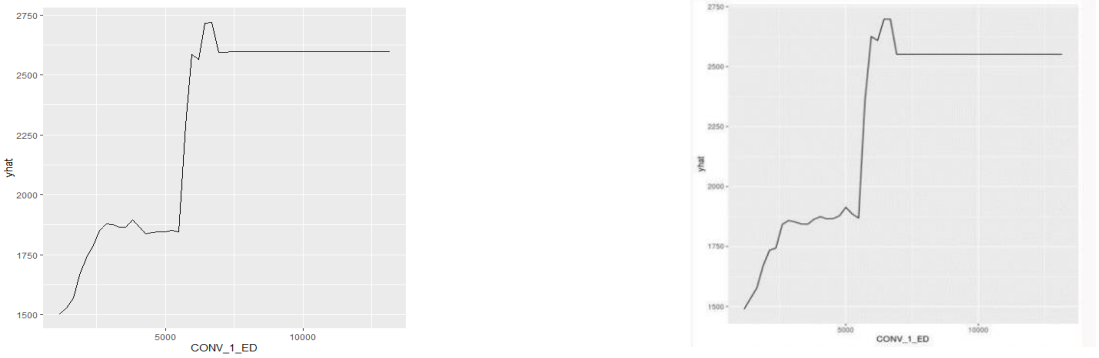
Cabe recalcar que además de saber qué variables son importantes, también es de interés saber en cómo las variables influyen en el resultado predicho. En ese sentido, el estudio de Castañeda et al. (2019), donde se estima el ingreso promedio, se elabora una encuesta con preguntas relacionadas a las condiciones educativas del hogar y del trabajo; con lo que se demuestra que estas variables si son generalmente usadas para la estimación del ingreso.

Un punto importante para considerar en los resultados de los gráficos de dependencia parcial (PDP) es que las variables categóricas fueron previamente tratadas por dos métodos, que se mencionaron previamente, por lo que se muestran ahora como variables continuas.

- Variable 1 CONV_1_ED: Es una variable convolucionada que recoge los efectos de las variables: rango de trabajadores, edad, sector y nivel educativo. Como se puede observar en la Figura20 el sentido observado es positivo hasta un rango específico, esto puede ser debido a para ciertos rangos de la variable CONV_1_ED el modelo de aprendizaje automático probablemente no pueda obtener una predicción significativa; sin embargo, se muestra que el modelo está respondiendo a la señal del predictor similarmente (en los resultados de Auditoría y Modelamiento); es decir, tiene un sentido creciente positivo hasta un rango.

Figura 20

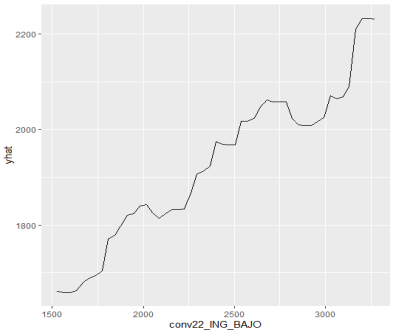
PDP de Auditoría vs Modelamiento de la Variable CONV_E_ED

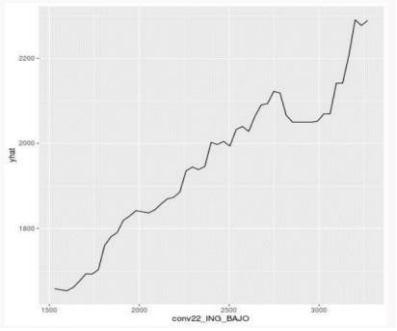


- Variable 2 conv22_ING_BAJO: Es una variable convolucionada continua que representa los meses con mal comportamiento en los últimos 12 meses y la deuda indirecta máxima de los últimos 24 meses. Se observa en la Figura21 que el sentido observado es positivo; es decir, mientras aumente el mal comportamiento del cliente, así como su deuda; mayor será su ingreso promedio.

Figura 21

PDP de Auditoría vs Modelamiento de la Variable conv22_ING_BAJO

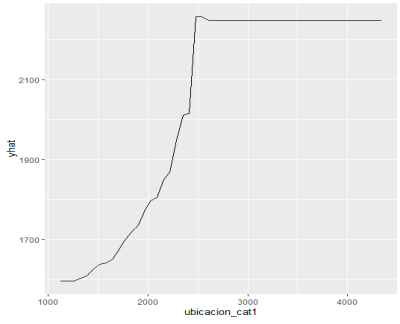


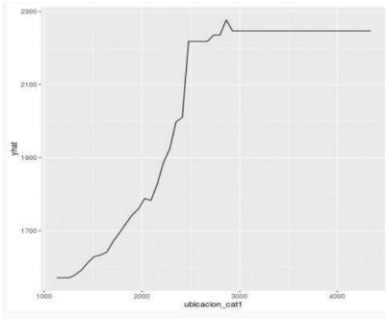


- Variable 3 ubicacion_cat1: Es una variable de distrito, y como se puede observar en la Figura22 el sentido observado es positivo hasta un rango específico, hasta 2500 aproximadamente; lo que indica que algunos distritos (que tienen un ingreso promedio de 2500 a más) perciben el mismo monto promedio de ingreso.

Figura 22

PDP de Auditoría vs Modelamiento de la Variable ubicacion_cat1





- Variable 4 conv32_ING_BAJO: Es una variable convolucionada continua entre meses con mal comportamiento de pago en los últimos 12 meses; y deuda la máxima de los últimos 24 meses/ deuda actual. Como se puede observar en la Figura23 el sentido

observado es positivo; y al igual que la variable 2 de ingresos bajos, tiene el mismo comportamiento; es decir, mientras aumente el mal comportamiento del cliente, así como su deuda (en ratio); mayor será su ingreso promedio, ya que ambos están en función del comportamiento de pago y la deuda, aunque con distintos ratios.

Figura 23

PDP de Auditoría vs Modelamiento de la Variable conv32_ING_BAJO

- Variable 5 CUOTA_RCC_max_12: Es la variable con la máxima cuota RCC (por RBM) de los últimos 12 meses. Como se puede observar en la Figura24 el sentido observado no es claro, por lo que el modelo de aprendizaje automático probablemente no pueda obtener una predicción significativa para este rango, debido a que en este segmento (ingresos bajos) los ingresos son bajos y no se cuenta con información de clientes que tengan una cuota alta.

Figura 24

PDP de Auditoría vs Modelamiento de la Variable CUOTA_RCC_max_12

Después de observar los PDP de los Top 5 principales variables, se puede concluir que las variables evaluadas tienen el mismo sentido; es decir, la predicción cambia en sentido positivo

creciente, cuando la variable cambia en el mismo sentido, sin embargo, 2 de las 5 variables evaluadas no se muestra un sentido claro en su totalidad, es decir, la variable ubicacion_cat1 tiene un sentido claro hasta un rango específico, mientras que CUOTA_RCC_max_12 no muestra un sentido claro en casi todo el rango de la variable.

Ingreso Medio: Los resultados de la validación del sentido económico de las variables obtenidas del análisis fueron, una lista de variables con su importancia en el modelo con R^2 de 33.65%; estas variables se encuentran listadas en la Tabla19.

Tabla 19

Variables Obtenidos por la Unidad de Modelamiento en el Segmento Medio

N°	Variables	Gain
1	MTOTOTDEU_D_I_MAX24_PJ	15.09%
2	CONV_1_ZN	14.14%
3	mto_profesion	4.28%
4	CUOTA_RCC_sum_12	4.28%
5	CUOTA_COMPRAS_F_max_12	4.25%
6	BACHILLER	4.13%
7	dem_cod_ciiu_cat2	3.70%
8	CONV_1_ED	3.36%
9	dem_des_mejorvehiculo_cat2	3.32%
10	ING_PROM_MZNA_20	2.58%
11	dem_num_antigrucmes	2.48%
12	dem_num_antigminvehiculomes	2.31%
13	MONTOADE_ACT6_MAX24	2.16%
14	BACHILLER_SUNEDU	2.03%
15	Sectores_MMGR_cat2	1.91%
16	CUOTA_OTROS_F_max_12	1.60%
17	CTDEMPREPORTADOCLIMED24	1.56%
18	NMESES_MTOTJCNSDSEF24	1.50%
19	PRC_PER_60_MAS	1.45%
20	PRC_PER_45_59	1.43%
21	MTOLINTCDMI0_PRO24	1.43%
22	conv42_ing_med	1.39%
23	REFTEN	1.34%
24	cv_estcivil_edadn2	1.28%
25	PRC_PER_MENOR_18	1.16%
26	FLG_LUJO	1.10%
27	ING_B_20	1.09%
28	FLGSEXO_cat2	1.05%
29	SOW_MTODEU_CEF_T1	0.84%
30	TIPSITUACIONCASA_cat2	0.82%
31	PORC_REPROG_TOT	0.77%
32	SOW_MTODEU_TC_T1	0.73%

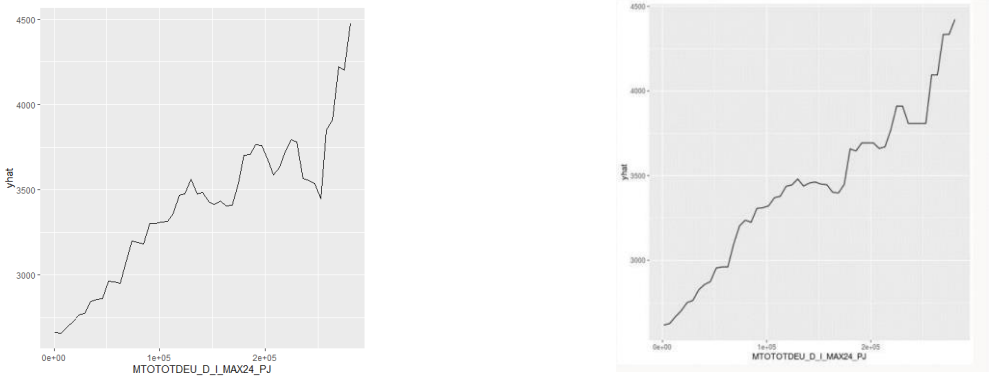
N°	Variables	Gain
33	conv32_ing_med	0.72%
34	ubicacionpro_cat1	0.71%
35	POB_D_20	0.68%
36	LINEA_TC_T3	0.61%
37	MONTOADE_ACT6_MAX6_IND	0.59%
38	CUOTA_MV_FINAL_sum_6	0.47%
39	SF6_SF12	0.46%
40	CUOTA_RCC_count_12	0.45%
41	MTODEU_TC_T2	0.45%
42	SF_DIS_CAL_CPP12	0.40%
43	RT_E_20	0.40%
44	CTD_PRODTOTAL_12	0.38%
45	MTODEUDAVCDO_MAX24	0.36%
46	flg_ind12	0.30%
47	CUOTA_OTROS_F_min_6	0.28%
48	MONTOADE_ACT3_MAX3	0.26%
49	CTDPROD_T1	0.23%
50	SF_DIS_VENCIDO24	0.19%
51	UTLCS1_UTLCS12_V1	0.17%
52	N_MESES_5000_12	0.16%
53	CTD_EMPRESAS_24	0.14%
54	FLAG_NUEVOS	0.13%
55	SOW_MTODEU_TC_T3	0.12%
56	FLG_TIT_EXT	0.10%
57	SOW_MTODEU_CEF_T3	0.09%
58	DEU_VCDA_TOT_MAX_3	0.09%
59	CUOTA_COMPRAS_F_count_12	0.08%
60	SFENT1_SFENT12	0.08%
61	ANTIGUEDAD_RCC48	0.06%
62	SF1_SF3	0.06%
63	dem_flg_modelovehiculartop	0.05%
64	SOW_MTODEU_HIP_T2	0.05%
65	flg_inf12	0.04%
66	dem_des_regiondem_cat2	0.04%
67	FLG_EMP_CON_FLUJO	0.04%
68	PORC_REPROG_MP	0.02%
69	dem_flg_basesunat	0.02%
70	ATRASOMAX_VEH_24	0.01%
71	FLG_BACHILLER	-
72	CUOTA_HP_FINAL_sum_6	-
73	CUOTA_VEH_FINAL_med_6	-

A continuación, se presentan los resultados de los Top 5 principales variables de los Partial Dependence Plot (PDP) del segmento Medio, algoritmo que fue usado para expresar si una variable está afectando positiva o negativamente al target (variable respuesta).

- Variable 1 MTOTOTDEU_D_I_MAX24_PJ: Es la variable de la deuda total máxima de los últimos 24 meses de las personas jurídicas. Como se puede observar en la Figura25 el sentido observado es positivo, es decir, mientras mayor sea la deuda, mayor será monto promedio de ingreso.

Figura 25

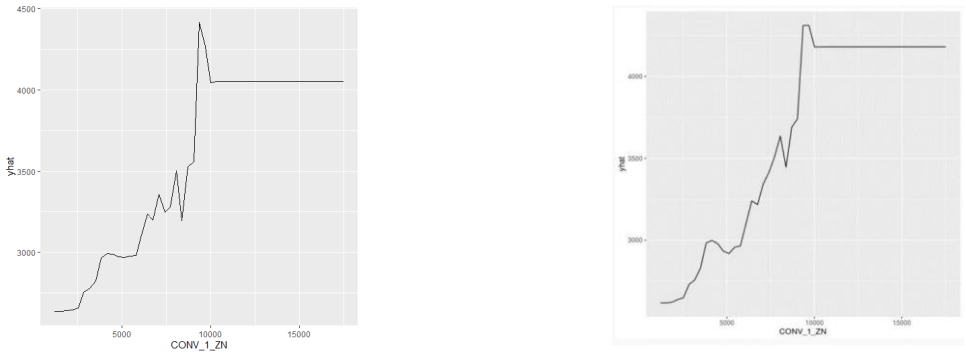
PDP de Auditoría vs Modelamiento de la Variable MTOTOTDEU_D_I_MAX24_PJ



- Variable 2 CONV_1_ZN: Es una variable convolucionada del rango de trabajadores, edad, departamento y macrozona. Y como se había mencionado anteriormente, las variables convolucionadas que tienen variables de sitios geográficos, pueden estar afectas a que en ciertos rangos el modelo de aprendizaje automático probablemente no pueda obtener una predicción significativa como se puede apreciar en la Figura26, que a partir del rango de 10 000 aproximadamente el ingreso promedio se mantiene estable.

Figura 26

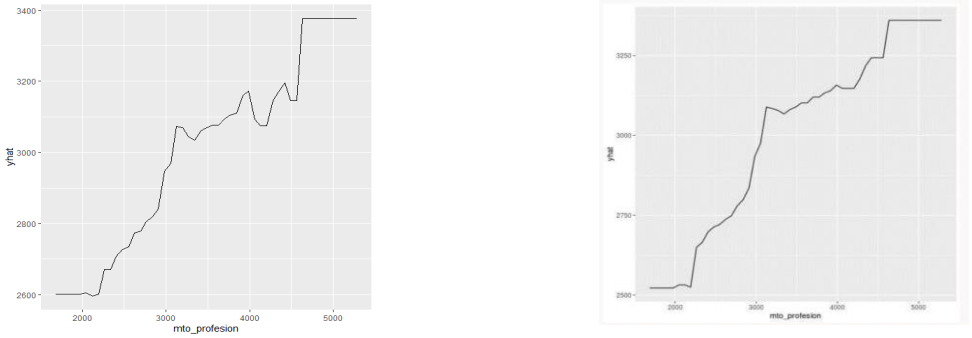
PDP de Auditoría vs Modelamiento de la Variable CONV_1_ZN



- Variable 3 mto_profesion: Es la variable que representa el ingreso promedio según la profesión del cliente; en ese sentido, se puede observar en la Figura27 el sentido observado es positivo creciente. Es decir, esta variable discrimina el ingreso promedio entre las profesiones de los clientes que se encuentran en el segmento medio.

Figura 27

PDP de Auditoría vs Modelamiento de la Variable mto_profesion



- Variable 4 CUOTA_RCC_sum_12: Es la variable que representa la suma de cuotas del Reporte Consolidado Crediticio (RCC) de los últimos 12 meses. Como se puede observar en la Figura28 el sentido observado es positivo creciente solo en un pequeño tramo inicial, sin embargo, luego solo tiene una línea horizontal, esto se puede deber a los efectos heterogéneos de los gráficos PDP. Es decir, la mitad de los datos de la variable tiene una asociación positiva, mientras que la otra mitad tiene una asociación negativa, lo que conlleva a que los efectos se anulen entre sí.

Figura 28

PDP de Auditoria vs Modelamiento de la Variable CUOTA_RCC_sum_12

- Variable 5 CUOTA_COMPRAS_F_max_12: Es la variable que representa la máxima cuota de compras de los últimos 12 meses. Como ya se mencionó en las variables anteriores la línea horizontal se puede deber a los efectos heterogéneos, sin embargo, existe un sentido creciente positivo al inicio de los rangos, tal como se aprecia en la Figura29.

Figura 29

PDP de Auditoría vs Modelamiento de la Variable CUOTA_COMPRAS_F_max_12

Después de observar los PDP de los Top 5 principales variables del segmento medio, se puede concluir que 3 de las variables evaluadas tienen un sentido creciente positivo, es decir, la predicción cambia en sentido positivo creciente, cuando la variable cambia en el mismo sentido, mientras que 2 de ellas no muestran un sentido claro.

Ingreso alto: Los resultados de la validación del sentido económico de las variables obtenidas del análisis fueron, una lista de variables con su importancia en el modelo con R^2 de 39.5%; estas variables se encuentran listadas en la Tabla20.

Tabla 20

Variables Obtenidos por la Unidad de Modelamiento en el Segmento Alto

N°	Variables	Gain
1	MTODEUDAMAX24_IND	22.22%
2	CONV_1_ZN	11.07%
3	CONV_1_ED	7.34%
4	dem_cod_ciiu_cat3	4.72%
5	dem_des_mejorvehiculo_cat3	3.71%
6	CUOTA_COMPRAS_F_max_12	3.66%
7	mto_profesion	3.07%
8	MTOLINTC_MAX24_V1	2.64%
9	MTODEUDAMAX24_MTODEUDAACTUAL	2.41%

N°	Variables	Gain
10	cv_estcivil_sector3	2.36%
11	CUOTA_RCC_max_6	2.32%
12	MAESTRIA	2.31%
13	CUOTA_HP_FINAL_max_12	1.99%
14	dem_num_antigrucmes	1.66%
15	BACHILLER_SUNEDU	1.59%
16	ING_PROM_MZNA_20	1.59%
17	CUOTA_CEF_FINAL_sum_12	1.34%
18	MONTOADE_ACT24_MAX24_IND	1.30%
19	MTODEUTC_MAX24	1.12%
20	MONTOADE_ACT6_MAX24	1.01%
21	RT_REFRI_HOGARES	1.00%
22	CTDEMPREPORTADOCLIMAX12	0.90%
23	PRC_PER_MENOR_18	0.85%
24	LINEA_TC_T2	0.84%
25	GAS_VES_C	0.84%
26	SOW_LINEA_TC_T2	0.79%
27	PRC_PER_45_59	0.78%
28	RT_C_20	0.77%
29	SOW_MTODEUDATOT_T1	0.76%
30	MTOLINTOT_DMI0_12_V1	0.70%
31	RT_NO_TIENE_DOCUMENTO	0.69%
32	CUOTA_VEH_FINAL_min_12	0.69%
33	MTODEUDATOT_T2	0.68%
34	PRC_PER_18_44	0.66%
35	CTD_PER_60_MAS	0.61%
36	ATRASOMAX_TDE_12	0.59%
37	SF12_SF24_IND	0.57%
38	N_AUTOS_NUEVOS	0.53%
39	CUOTA_OTROS_F_min_12	0.52%
40	cv_sexo_edadn3	0.48%
41	PORC_REPROG_CON	0.47%
42	TIPSITUACIONCASA_cat3	0.46%
43	SOW_LINEA_TC_T3	0.44%
44	ubicacionpro_cat1	0.40%
45	POB_D_20	0.37%
46	flg_ind12	0.31%
47	N_AUTOS_BAJA	0.30%
48	MONTOADE_ACT6_MAX12_IND	0.28%
49	CTD_CONSUMO_24	0.22%
50	SFENT12_SFENT24	0.22%
51	FLGSEXO_cat3	0.20%
52	MTODEU_TC_T2	0.19%
53	SF6_SF12_IND	0.18%
54	ATRASOMAX_VEH_12	0.18%
55	N_MESES_5000_12	0.17%
56	FLG_TITULO	0.17%
57	RT_E_20	0.16%
58	SF6_SF12	0.16%
59	SF3_SF6	0.15%
60	MONTOADE_ACT_MAX6_IND	0.15%
61	UTLCS_1	0.14%
62	FLAG_NUEVOS	0.13%
63	ATRASOMAX_TCSMC_24	0.13%
64	MTODEU_CEF_T2	0.10%
65	N_AUTOS_GLUJO	0.09%
66	ATRASOMAX_CRNOR_12	0.07%

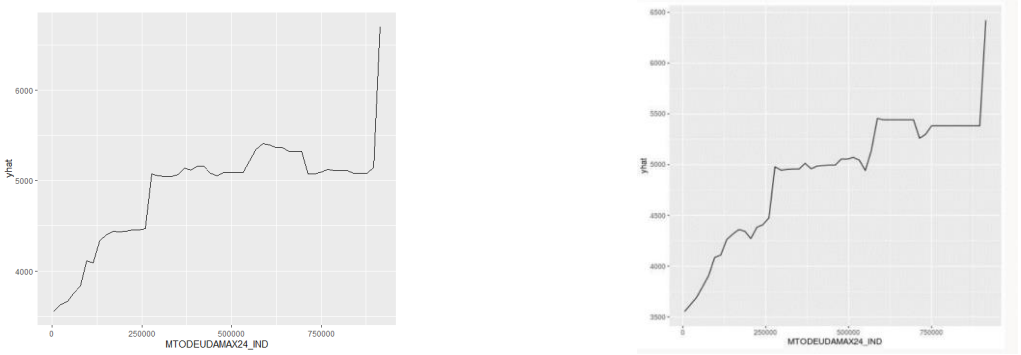
N°	Variables	Gain
67	CTDPROD_T1	0.07%
68	MONTOADE_ACT3_MAX3_IND	0.07%
69	dem_flg_basesunat	0.06%
70	ATRASOMAX_CRNENR_24	0.06%
71	SFENT1_SFENT12	0.05%
72	CTDPROD_T2	0.04%
73	CUOTA_MV_FINAL_max_12	0.03%
74	MTODEUDAVCDO_MAX24	0.03%
75	CTD_MESES_HIPOTECARIO_24	0.03%
76	ANTIGUEDAD_RCC48	0.02%
77	MTODEU_TC_T3	0.01%
78	SF1_SF3	0.01%
79	MESES_ACTIVO_SF_BUENO3	0.01%
80	ATRASOMAX_1	0.00%
81	FLG_TOPM_SECTOR	0.00%
82	DEU_VCDA_TOT_PRO_6	0.00%
83	SF_NUM_CAL_PER12	0.00%
84	N_MESES_1000_3	0.00%
85	SOW_CTDPROD_T3	0.00%
86	DEUDA_NEG_RCC	-
87	PORC_REPROG_MP	-
88	FLG_REV_24	-
89	SF_NUM_VENCIDO1	-
90	VECESCONTINUO	-
91	dem_flg_direccionbcp	-
92	dem_des_regiondem_cat3	-

A continuación, se presentan los resultados de los Top 5 principales variables de los Partial Dependence Plot (PDP) del segmento Alto, algoritmo que fue usado para expresar si una variable está afectando positiva o negativamente al target (variable respuesta).

- Variable 1 MTODEUDAMAX24_IND: Esta variable representa la deuda máxima en los últimos 24 meses incluyendo deuda indirecta. Como se puede observar en la Figura30 el sentido observado es positivo creciente en el gráfico de auditoría (gráfico izquierdo) y de igual forma en el gráfico de modelamiento (gráfico derecho); es decir mientras mayor sea la deuda, mayor será el ingreso promedio.

Figura 30

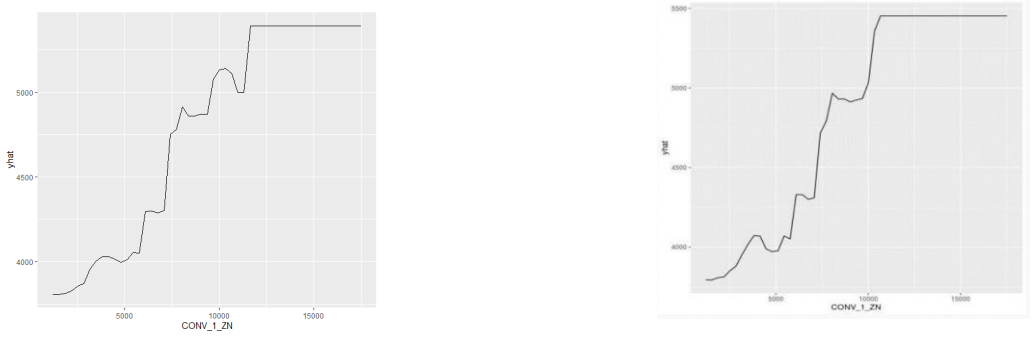
PDP de Auditoría vs Modelamiento de la Variable MTODEUDAMAX24_IND



- Variable 2 CONV_1_ZN: Se trata de una variable convolucionada que abarca las variables: rango de trabajadores, edad, departamento y macrozona. Al igual en los ingresos medios esta variable tiene un sentido creciente positivo como se puede ver en la Figura31. De igual forma se aprecia que se tiene el mismo sentido en ambas gráficas de PDP, de auditoría y de modelamiento.

Figura 31

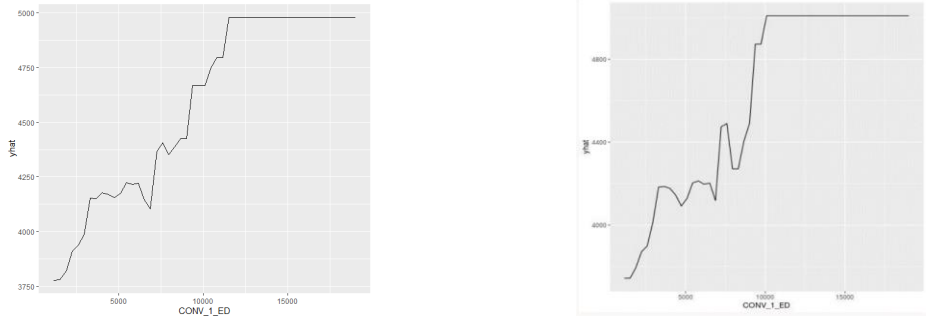
PDP de Auditoría vs Modelamiento de la Variable CONV_1_ZN



- Variable 3 CONV_1_ED: Se trata de una variable convolucionada que abarca las variables: rango de trabajadores, edad, sector y nivel educativo. Y se puede observar en la Figura32 el sentido observado es positivo creciente hasta un rango específico, al igual que en la misma variable en el segmento bajo. Además, se valida con los resultados de auditoría (gráfica izquierda) que se tiene el mismo sentido del gráfico de PDP de modelamiento.

Figura 32

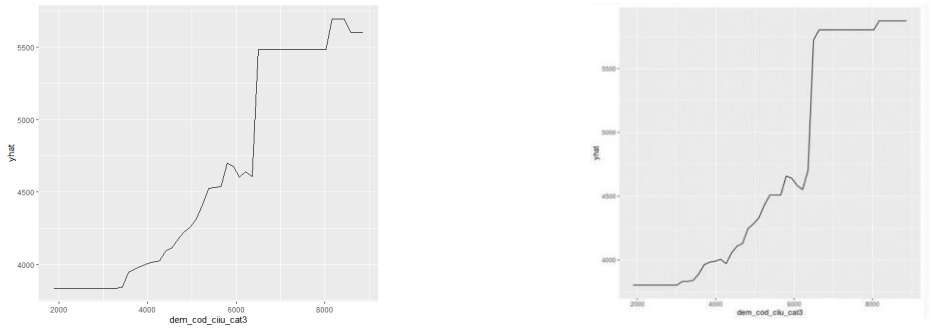
PDP de Auditoría vs Modelamiento de la Variable CONV 1 ED



- Variable 4 dem_cod_ciiu_cat3: Es una variable que indica la clasificación industrial internacional uniforme de todas las actividades económicas, es decir, indica el sector de actividad económica en el que se encuentra el empleador de la persona. Y como se muestra en la Figura33 el sentido observado es positivo.

Figura 33

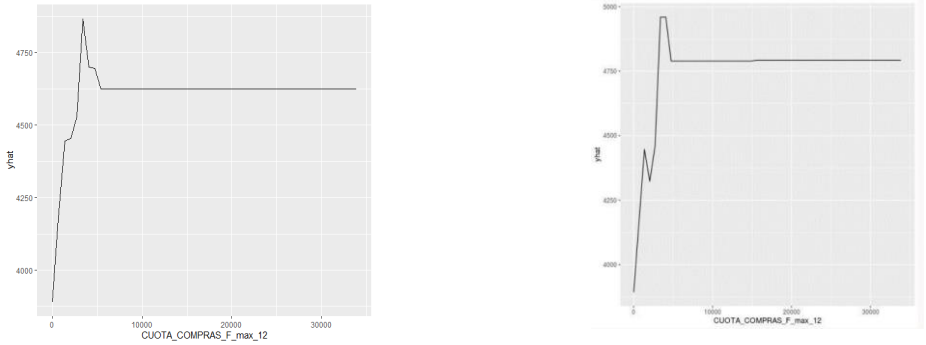
PDP de Auditoría vs Modelamiento de la Variable dem_cod_ciiu_cat3



- Variable 5 CUOTA_COMPRAS_F_max_12: Es la variable que representa la máxima cuota de compras del reporte consolidado crediticio (RCC) de los últimos 12 meses. Como se puede observar en la Figura34 el sentido observado no es claro, esto puede ser por una de las desventajas que presentan los gráficos PDP, los efectos heterogéneos que hacen que el sentido creciente de un rango se anule con el sentido negativo de otro rango, sin embargo, se aprecia que ambos gráficos tienen el mismo sentido, es decir el gráfico de auditoría y de modelamiento.

Figura 34

PDP de Auditoría vs Modelamiento de la Variable CUOTA_COMPRAS_F_max_12



Al igual que en los segmentos bajo y medio se cuentan con variables que tienen el mismo sentido; es decir, la predicción cambia en sentido positivo creciente cuando la variable cambia en el mismo sentido. Después de observar el comportamiento de las variables más importante en cada uno de los segmentos con respecto a la variable respuesta (promedio de ingreso), estas variables en su mayoría son demográficas y contienen información del reporte crediticio consolidado (RCC) del banco, por lo que es necesario mencionar que, como se muestra en la tesis de Ivanovna et al. (2013). "Propuesta metodológica para la estimación de ingresos en otorgamiento de créditos", para tener una mejor estimación del ingreso y que estos sean acordes a la realidad, se deben usar variables sociodemográficas y de endeudamiento en el sistema financiero; las mismas variables que se han demostrado que se encuentran en el modelo del presente trabajo.

Finalmente, luego de revisar el sentido económico de las variables finalmente se obtuvo un número de variables en cada segmento que se detallan en la Tabla21; donde quedan como finalistas 52 variables en el segmento bajo, y 73 variables en el segmento medios; y finalmente 92 variables en el segmento alto.

Tabla 21

Número de Variables Finales por Segmento Después del Sentido Económico

Modelo	Variables Bivariado	Revisión Sentido Econ. I
Ingresos Bajos	101	52
Ingresos Medios	107	73
Ingresos Altos	106	92

ii) Boruta

Los resultados de realizar la réplica con las bases de desarrollo del segmento bajo, el set de variables utilizadas en el análisis Boruta y verificar que las variables con la denominación "Confirmed" de la Tabla22 no hayan sido descartadas o eliminadas por la Unidad de Modelamiento en el post análisis, y se contrasta con la Tabla20 para validar que las variables "confirmadas" no hayan sido descartadas. Cabe mencionar que, réplica se aplicó sobre el segmento de mayor proporción "Ingresos Bajos" (60%), debido al elevado costo computacional que implicó el reproceso de este análisis, sobre los demás segmentos se verificó en los archivos de trabajo que las variables con el calificativo "Confirmed" no hayan sido eliminadas.

A continuación, en la Tabla22 se muestra información de la salida en R, donde la propia técnica nos arrojó que variables deben ser utilizadas "Confirmed", qué variables podrían ser incluidas "Tentative" y cuáles no deben ser consideradas "Rejected". Los resultados de la comparación con las variables elegidas por la unidad de Modelamiento son conformes en su totalidad (esto se puede contrastar con la Tabla23); se encontraron coincidencia de 40 variables "Confirmed" y 1 variable "Tentative"; sumando un total de 41 variables seleccionadas por Boruta.

Tabla 22

Tabla obtenida del Análisis Boruta - Auditoría

Variables Auditoría	meanImp	medianImp	minImp	maxImp	normHits	decision
CONV_1_ED_imp	0.35	0.35	0.34	0.35	1.00	Confirmed
conv22_ING_BAJO	0.09	0.09	0.09	0.09	1.00	Confirmed
conv32_ING_BAJO	0.07	0.07	0.07	0.08	1.00	Confirmed
conv52_ING_BAJO	0.01	0.01	0.01	0.01	1.00	Confirmed
conv61_ING_BAJO_imp	0.01	0.01	0.01	0.01	1.00	Confirmed
CTD_EMPRESAS_24	0.00	0.00	0.00	0.00	0.74	Confirmed
CTD_PER_MENOR_18_imp	0.00	0.00	0.00	0.01	0.97	Confirmed
ctdempreportadoclimed24	0.00	0.00	0.00	0.01	0.97	Confirmed
CTDPROD_T1_imp	0.00	0.00	0.00	0.01	0.97	Confirmed
cuota_compras_f_max_12_imp	0.02	0.02	0.02	0.02	1.00	Confirmed
cuota_hp_final_min_6_imp	0.00	0.00	0.00	0.00	0.84	Confirmed
cuota_otros_f_max_12_imp	0.00	0.00	0.00	0.00	0.80	Confirmed
CUOTA_RCC_max_12_imp	0.05	0.05	0.05	0.06	1.00	Confirmed
dem_cod_ciiu_cat1	0.08	0.08	0.07	0.08	1.00	Confirmed
dem_des_mejorvehiculo_cat1	0.03	0.03	0.03	0.04	1.00	Confirmed
dem_flg_basesunat_imp	0.00	0.00	0.00	0.00	0.75	Confirmed
DEUDA_NEG_RCC_imp	0.00	0.00	0.00	0.00	0.96	Confirmed
DEUDIR1_DEUTOT1	0.00	0.00	0.00	0.00	0.83	Confirmed

Variables Auditoría	meanImp	medianImp	minImp	maxImp	normHits	decision
FLG_DISEF24	0.01	0.01	0.01	0.02	1.00	Confirmed
FLG_DISEF3	0.00	0.00	0.00	0.00	0.60	Confirmed
flg_inf12	0.01	0.01	0.01	0.01	0.99	Confirmed
FLG_MIN_CONST	0.03	0.03	0.03	0.04	1.00	Confirmed
FLGSEXO_cat1	0.05	0.05	0.05	0.05	1.00	Confirmed
ING_D_20_imp	0.00	0.00	0.00	0.00	0.69	Confirmed
montoade_act3_max3_s_hip_imp	0.00	0.00	0.00	0.01	0.79	Confirmed
mto_profesion	0.01	0.01	0.01	0.01	1.00	Confirmed
MTODEU_CEF_T1_imp	0.03	0.03	0.02	0.03	1.00	Confirmed
MTODEU_CEF_T2_imp	0.00	0.00	0.00	0.01	0.84	Confirmed
MTODEU_CEF_T3_imp	0.00	0.00	0.00	0.00	0.67	Confirmed
MTODEUDAVCDO_MAX24	0.00	0.00	0.00	0.00	0.74	Confirmed
nivel_educativo_f_cat1	0.00	0.00	0.00	0.01	0.98	Confirmed
PORC_REPROG_TOT_imp	0.00	0.00	0.00	0.00	0.79	Confirmed
PRC_PER_MENOR_18_imp	0.00	0.00	0.00	0.01	0.98	Confirmed
PRED_EQX2021_cat1	0.00	0.00	0.00	0.00	0.79	Confirmed
RT_REFRI_HOGARES_imp	0.01	0.01	0.01	0.01	1.00	Confirmed
RT_TIENE_DOCUMENTO_imp	0.00	0.00	0.00	0.00	0.76	Confirmed
SOW_MTODEU_TC_T3_imp	0.01	0.01	0.00	0.01	0.99	Confirmed
SOW_MTODEUDATOT_T1_imp	0.02	0.02	0.01	0.02	1.00	Confirmed
TIPSITUACIONCASA_cat1	0.00	0.00	0.00	0.00	0.79	Confirmed
ubicacion_cat1	0.07	0.07	0.07	0.08	1.00	Confirmed
CTD_CONSUMO	0.00	0.00	0.00	0.00	0.00	Rejected
CTD_PRODTOTAL_6	0.00	0.00	0.00	0.00	0.00	Rejected
CUOTA_HP_FINAL_sum_6_imp	0.00	0.00	0.00	0.00	0.00	Rejected
cuota_veh_final_min_12_imp	0.00	0.00	0.00	0.00	0.00	Rejected
cuota_veh_final_sum_6_imp	0.00	0.00	0.00	0.00	0.00	Rejected
FLG_BAJA_RUC	0.00	0.00	0.00	0.00	0.00	Rejected
flg_lintcpro20_12m	0.00	0.00	0.00	0.00	0.00	Rejected
MTODEU_TC_T3_imp	0.00	0.00	0.00	0.00	0.00	Rejected
SF_DIS_CAL_CPP12_imp	0.00	0.00	0.00	0.00	0.06	Rejected
TIPESTCIVIL_cat1	0.00	0.00	0.00	0.00	0.00	Rejected
flg_ind12	0.00	0.00	0.00	0.00	0.54	Tentative
MONTOADE_ACT_MAX3_IND	0.00	0.00	0.00	0.00	0.44	Tentative

Tabla 23

Variables del Segmento Ingresos Bajos y su Importancia en el modelo con R^2=31.4%

N°	Variables de la Unidad de Modelamiento	Gain
1	CONV_1_ED	27.10%
2	conv22_ING_BAJO	7.80%
3	ubicacion_cat1	7.30%
4	conv32_ING_BAJO	6.40%

N°	Variables de la Unidad de Modelamiento	Gain
5	CUOTA_RCC_max_12	5.30%
6	dem_cod_ciiu_cat1	5.20%
7	FLGSEXO_cat1	3.60%
8	CUOTA_COMPRAS_F_max_12	3.10%
9	conv52_ING_BAJO	2.50%
10	dem_des_mejorvehiculo_cat1	2.30%
11	RT_REFRI_HOGARES	2.20%
12	SOW_MTODEUDATOT_T1	2.10%
13	FLG_MIN_CONST	2.10%
14	CUOTA_OTROS_F_max_12	1.80%
15	mto_profesion	1.70%
16	MTODEU_CEF_T2	1.60%
17	MTODEU_CEF_T1	1.50%
18	CTDEMPREPORTADOCLIMED24	1.40%
19	PRC_PER_MENOR_18	1.40%
20	RT_TIENE_DOCUMENTO	1.30%
21	FLG_DISEF24	1.20%
22	CTD_PER_MENOR_18	1.20%
23	PORC_REPROG_TOT	1.20%
24	nivel_educativo_f_cat1	1.20%
25	ING_D_20	0.90%
26	MONTOADE_ACT3_MAX3_S_HIP	0.80%
27	TIPSITUACIONCASA_cat1	0.80%
28	DEUDIR1_DEUTOT1	0.70%
29	conv61_ING_BAJO	0.70%
30	MTODEUDAVCDO_MAX24	0.60%
31	CTD_EMPRESAS_24	0.50%
32	SOW_MTODEU_TC_T3	0.40%
33	dem_flg_basesunat	0.40%
34	CTDPROD_T1	0.40%
35	PRED_EQX2021_cat1	0.30%
36	FLG_DISEF3	0.30%
37	flg_ind12	0.30%
38	flg_inf12	0.20%
39	DEUDA_NEG_RCC	0.20%
40	MTODEU_CEF_T3	0.00%
41	CUOTA_HP_FINAL_min_6	0.00%

Como resultado de la selección de variables Boruta, se puede observar en la Tabla24 que queda un óptimo con 41 variables para el segmento de ingresos bajos, 54 variables para el segmento de ingresos medios y 63 variables para el segmento de ingresos altos.

Tabla 24

Nro de Variables Posterior a la Selección de Variables por Boruta

Segmentos	Variables Bivariado	Revisión Sentido Econ. I	Post Boruta
Ingresos Bajos	101	52	41
Ingresos Medios	107	73	54
Ingresos Altos	106	92	63

iii) Sentido Económico II

En esta etapa se realiza la misma metodología enunciada en el ítem de "revisión de sentido económico I" solo que esta vez se desarrolla la revisión para todas las variables finales ya que se cuenta con menor cantidad de variables respecto a la cantidad de la salida del bivariado. En la revisión se descartan aquellas variables que: (1) no cuenten con un sentido económico claro, o que no se considere relevante para predecir el ingreso, o (2) cuya correlación lineal sea contraria al sentido esperado, dicha revisión se realiza para cada segmento del modelo y se realiza en conjunto por expertos de los equipos de MMGR, VI, y RBM.

Asimismo, se identificaron aquellas variables cuya importancia era muy pequeña, dado que no permitía generar PDPs ya que los percentiles estaban muy cercanos entre sí. Las variables sobrantes se agruparon en familias de variables para confirmar el sentido de la familia con el negocio. A continuación, en las Figuras 35, 36 y 37, mostramos los PDP de algunas variables finales por segmento.

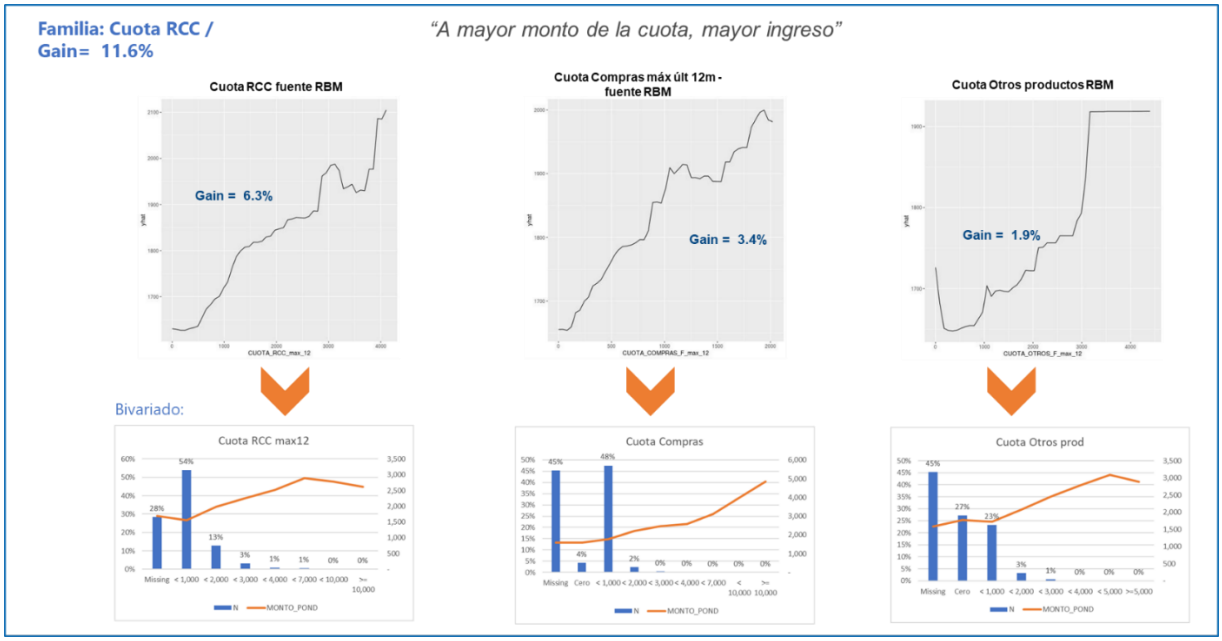
Ingreso Bajo:

Como se puede apreciar en la Figura35 se muestra los PDP de las variables CUOTA_RCC_max_12 (máxima cuota del RCC de los últimos 12 meses), CUOTA_COMPRAS_F_max_12 (máxima cuota de compras de los últimos 12 meses) y CUOTA_OTROS_F_max_12 (máxima cuota de productos asociado a la tarjeta de crédito a excepción de compras y disposición de efectivo de los últimos 12 meses) y el análisis bivariado de las mismas. En las 3 variables se muestra un sentido creciente positivo, es decir cuando el máximo monto de la cuota de los diferentes tipos (por compras, por productos de TC) aumenta,

el ingreso promedio ponderado también aumenta; en otras palabras “a mayor monto de la cuota, mayor ingreso”

Figura 35

Resultados de PDP y Análisis Bivariado de algunas Variables del Segmento Bajo

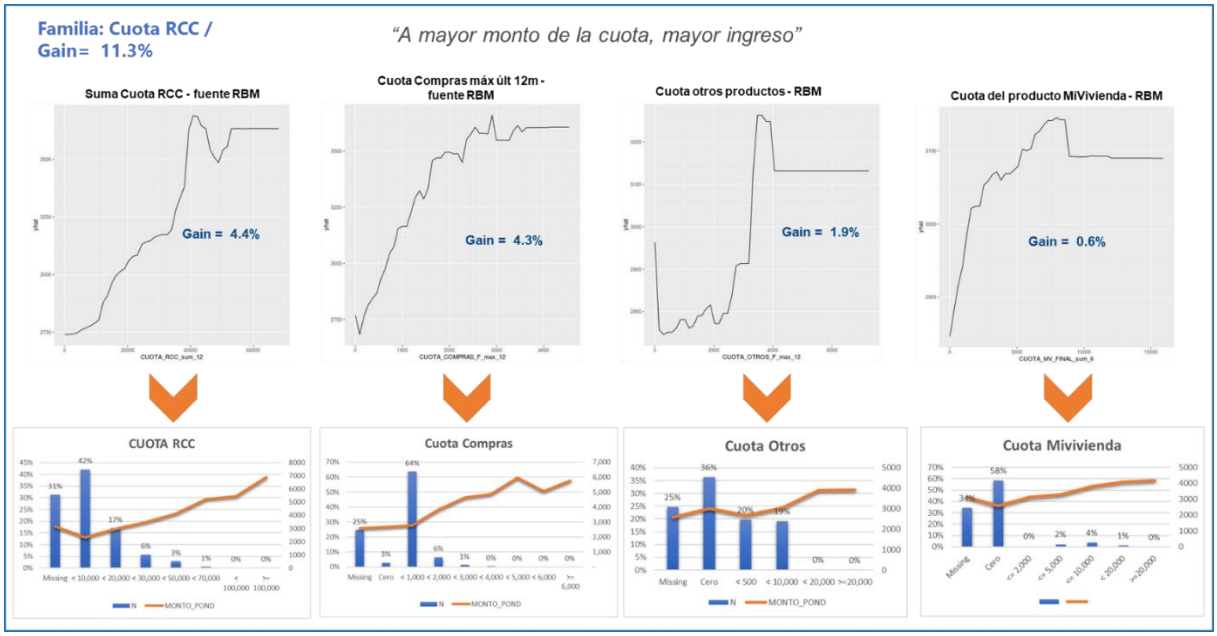


Ingreso Medio:

En la Figura36 se muestra los PDP de las variables CUOTA_RCC_sum_12 (máxima cuota del RCC de los últimos 12 meses), CUOTA_COMPRAS_F_max_12 (máxima cuota de compras de los últimos 12 meses) CUOTA_OTROS_F_max_12 (máxima cuota de productos asociado a la tarjeta de crédito a excepción de compras y disposición de efectivo de los últimos 12 meses) y CUOTA_MV_FINAL_sum_6 (suma de cuotas vehicular RCC de los últimos 6 meses) y el análisis bivariado de las mismas. Se muestra que las 4 variables muestran el mismo sentido creciente positivo, es decir “a mayor monto de la cuota, mayor ingreso”.

Figura 36

Resultados de PDP y Análisis Bivariado de algunas Variables del Segmento Medio

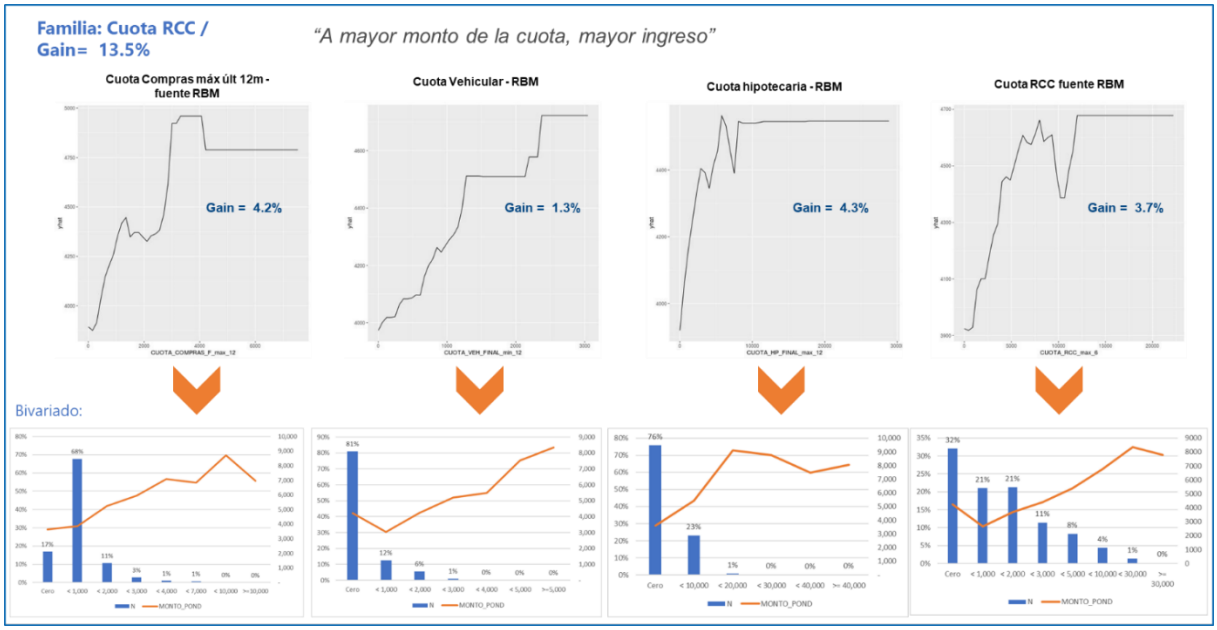


Ingreso Alto:

En la Figura37 se muestra los PDP de las variables CUOTA_HP_FINAL_max_12 (máxima cuota hipotecaria del RCC de los últimos 12 meses), CUOTA_COMPRAS_F_max_12 (máxima cuota de compras de los últimos 12 meses) CUOTA_RCC_max_6 (máxima cuota de RCC de los últimos 6 meses) y CUOTA_VEH_FINAL_min_12 (mínima cuota vehicular del RCC de los últimos 12 meses) y el análisis bivariado de las mismas. En los gráficos PDP de las variables se muestran que las 4 variables tienen el mismo sentido creciente positivo, es decir "a mayor monto de la cuota, mayor ingreso".

Figura 37

Resultados de PDP y Análisis Bivariado de algunas Variables del Segmento Alto



Finalmente, luego de revisar las variables, se puede mostrar en la Tabla25 que quedan 33 variables en el modelo del segmento bajo, 40 en el segmento medio y 28 en el segmento alto.

Tabla 25

Nro de Variables Posterior a la Revisión de Sentido Económico II

Modelo	Variables Bivariado	Revisión Sentido Econ. I	Boruta	Revisión Sentido Econ. II
Ingresos Bajos	101	52	41	33
Ingresos Medios	107	73	54	40
Ingresos Altos	106	92	63	28

Técnica de Hyperparameters Tunning

En esta técnica se ejecutaron 972 iteraciones y en cada una de ellas se guardó el valor del R^2 de la base de test y train y el porcentaje de aciertos ±25. A continuación se mostrará los resultados de los modelos en cada uno de los segmentos.

Segmento de Ingresos Bajos:

Como resultado se obtuvieron un set de modelos en base a distintas combinaciones de hiperpámetros. De estos se eligió aquel que tuviera mayor R^2 del test y aquel que tuviera la menor diferencia de R^2 entre el test y train. Como se puede observar en la Figura38 se eligieron dos posibles alternativas; y ante ello se buscó la opción donde el porcentaje de aciertos ±25 sea mayor y por ello se elige la opción 1, según la Tabla26. Este modelo tiene un performance de 29.8% de R^2 y 53.6% de aciertos ±25 en la base de validación.

Figura 38

Gráfico por Resultado del Tuneo de Hiperparámetros Segmento Bajo

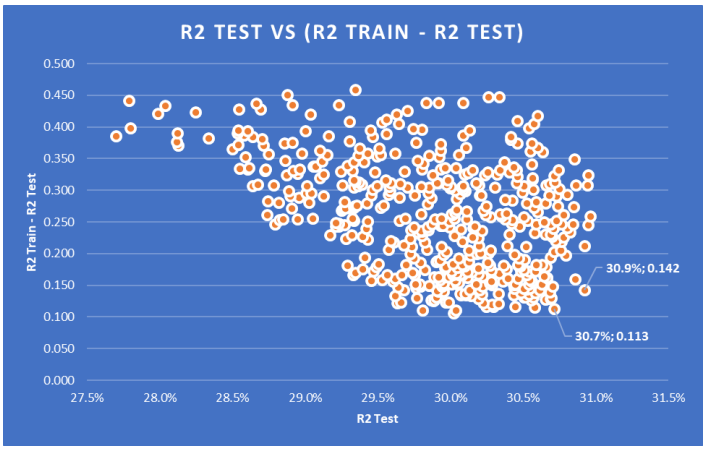


Tabla 26

Comparativo de Modelos en el Segmento de Ingresos Bajos con sus Hiperparámetros

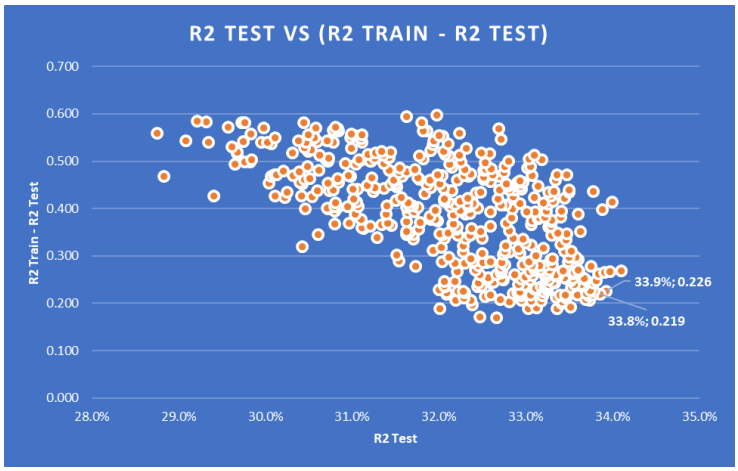
Opciones	Hiperparámetro	Valor final	R^2	R^2 Train - R^2 Test	Aciertos_25	Sobreestimación
1	eta	0.05	30.90%	0.142	53.50%	31.70%
	max_depth	11				
	colsample_bytree	0.6				
	subsample	1				
	min_child	360				
	lambda	0				
	Gamma	0				
2	eta	0.05	30.70%	0.113	53.30%	31.90%
	max_depth	11				
	colsample_bytree	0.6				
	subsample	0.6				
	min_child	360				
	lambda	1				
	Gamma	0				

Segmento de Ingresos Medios:

Como resultado se obtuvieron un set de modelos en base a distintas combinaciones de hiperparametros. De estos se eligió aquel que tuviera mayor R^2 del test y aquel que tuviera la menor diferencia de R^2 entre el test y train. Como se puede observar en la Figura39 se eligieron dos posibles alternativas; y se buscó la opción donde el porcentaje de aciertos ±25 sea mayor y por ello se elige la opción 2, según la Tabla27. Este modelo tiene un performance de 34.8% de R^2 y 40.1% de aciertos ±25 en la base de validación.

Figura 39

Gráfico por Resultado del Tuneo de Hiperparámetros Segmento



Según el gráfico superior se escogieron dos posibles alternativas y ante ello se buscó la opción donde el porcentaje de aciertos ±25 sea mayor y por ello se elige la opción 2 según el siguiente cuadro:

Tabla 27

Comparativo de Modelos en el Segmento de Ingresos Medios

Opciones	Hiperparámetro	Valor final	R^2	R^2 Train - R^2 Test	Aciertos_25	Sobreestimación
1	eta	0.05	33.90%	0.226	38.70%	43.20%
	max_depth	11				
	colsample_bytree	0.8				
	subsample	0.6				
	min_child	360				
	lambda	0				
	Gamma	0				

2	eta max_depth colsample_bytree subsample min_child lambda Gamma	0.05 11 0.6 0.6 360 1 0	33.80%	0.219	38.90%	43.00%

Segmento de Ingresos Altos:

De igual forma con en los segmentos bajos y medios, se obtuvieron un set de modelos en base a distintas combinaciones de hiperparámetros. De estos se eligió aquel que tuviera mayor R^2 del test y aquel que tuviera la menor diferencia de R^2 entre el test y train. Como se puede observar en la Figura40 se buscó la opción donde el porcentaje de aciertos ±25 sea mayor y por ello se elige la opción 2, según la Tabla28. Este modelo tiene un performance de 36.5% de R^2 y 39.8% de aciertos ±25 en la base de validación.

Figura 40

Gráfico por Resultado del Tuneo de Hiperparámetros Segmento

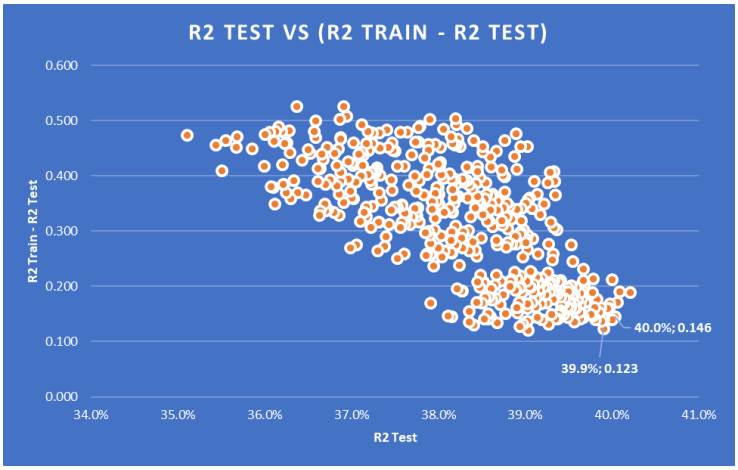


Tabla 28

Comparativo de Modelos en el Segmento de Ingresos Altos

Opciones	Hiperparámetro	Valor final	R2	R2 Train - R2 Test	Aciertos_25	Sobreestimación
1	eta max_depth colsample_bytree	0.1 11 0.6	40.00%	0.146	39.60%	40.30%

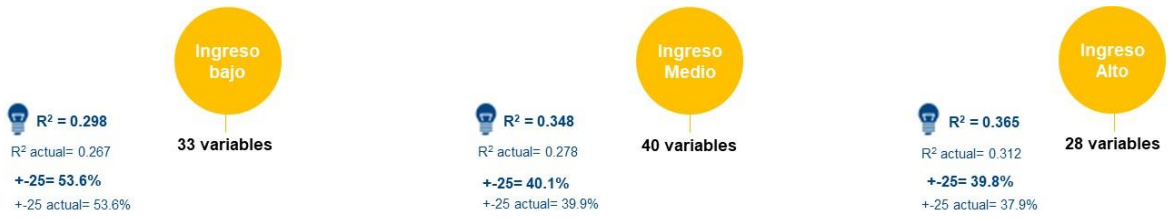
	subsample	0.6				
	min_child	360				
	lambda	0				
	Gamma	0				
	eta	0.05				
	max_depth	13				
	colsample_bytree	0.6				
2	subsample	0.6	39.90%	0.123	39.10%	40.40%
	min_child	360				
	lambda	1				
	Gamma	0				

4.7 Modelos Finales

De acuerdo con la optimización de los hiperparámetros, se ejecutó los tres modelos finales para cada segmento, obteniendo un R^2 de 29.8% y un porcentaje de aciertos de ±25 en 53.6% en el segmento bajo, un R^2 de 34.8% y un porcentaje de aciertos de ±25 en 40.1% en el segmento medio; y un R^2 de 36.5% y un porcentaje de aciertos de ±25 en 39.8% en el segmento alto. Se puede observar en la Figura41 que los indicadores de discriminación y precisión de los nuevos modelos en cada segmento son superiores a los modelos vigentes; lo que mejora la estimación de los ingresos.

Figura 41

Comparativo de R^2 en Todos los Segmentos y R^2 General



A continuación, en las Tablas 29, 30 y 31 se muestran la lista de variables finalistas para cada segmento.

Tabla 29

Variables del segmento Ingresos Bajos e importancia en el modelo de R^2=29.8%

Variables	Gain	Descripción
CONV_1_ED	21.50%	Convolución de educación (rango de trabajadores + edad + sector + nivel educativo)
conv22_ING_BAJO	9.60%	Convolución continua entre meses con mal comp. Los ult. 12 meses y deuda indirecta maxima los ult.24 meses.
ubicacion_cat1	8.00%	Distrito
conv32_ING_BAJO	7.20%	Convolución continua entre meses con mal comp. Los ult. 12 meses y deuda maxima los ult.24 meses/deuda actual.
dem_cod_ciiu_cat1	6.40%	CIIU
CUOTA_RCC_max_12	6.30%	Máxima Cuota RCC (por RBM) de los últimos 12 meses
FLG_MIN_CONST	5.00%	FLG si actividad económica es relacionada a Minería o Construcción
CUOTA_COMPRAS_F_max_12	3.40%	Máxima Cuota de Compras (por RBM) de los últimos 12 meses
conv52_ING_BAJO	2.80%	Convolución continua entre meses con mal comp. Los últimos 12 meses y deuda TC máxima los ult.24 meses.
dem_des_mejorvehiculo_cat1	2.60%	Mejor Vehículo
RT_REFRI_HOGARES	2.40%	ratio de hogares que tienen refrigerador en la manzana
mto_profesion	2.20%	Profesión
CUOTA_OTROS_F_max_12	1.90%	Máxima Cuota de productos asociado a TC a excepción de compras y disposición de efectivo (por RBM) de los últimos 12 meses
MTODEU_CEF_T1	1.90%	Monto deuda de crédito efectivo en tier 1 que corresponde
SOW_MTODEUDATOT_T1	1.80%	Porcentaje de deuda total en entidades TIER 1
MTODEU_CEF_T2	1.70%	Monto deuda de crédito efectivo en tier 2 que corresponde
FLG_DISEF24	1.70%	Flag que indica saldos correspondientes a disposición de efectivo en los últimos 24 meses
nivel_educativo_f_cat1	1.50%	Nivel educativo
CTDEMPREPORTADOCLIMED24	1.50%	Número promedio de empresas que reportan al cliente en los últimos 24 meses
CTD_PER_MENOR_18	1.50%	Cantidad de personas de menos de 18 años
RT_TIENE_DOCUMENTO	1.30%	porcentaje de personas con algún documento DNI/Carnet extranjería
TIPSITUACIONCASA_cat1	1.30%	Situación casa
MONTOADE_ACT3_ MAX3_S_HIP	1.20%	Ratio de la deuda en el sistema financiero de los 3 últimos reportes de RCC sobre la deuda máxima en los 3 meses sin incluir deuda del producto hipotecario
PORC_REPROG_TOT	1.20%	Porcentaje de deuda reprogramada total.

Variables	Gain	Descripción
ING_D_20	0.80%	Ingresos totales de personas de nivel socioeconómico dado en la proyección 2020
CTDPROD_T1	0.70%	cantidad de productos en tier 1 que corresponde con deuda mayor a 500 soles
DEUDIR1_DEUTOT1	0.60%	Ratio entre la deuda directa promedio del último mes entre la deuda total (directa más indirecta) promedio del último mes
SOW_MTODEU_TC_T3	0.50%	Porcentaje de deuda TC en entidades TIER 3
dem_flg_basesunat	0.50%	Flag si existe en la base de sunat
flg_ind12	0.40%	Flag que indica si el trabajador ha sido independiente en el último año
FLG_DISEF3	0.40%	Flag que indica saldos correspondientes a disposición de efectivo en los últimos 3 meses
PRED_EQX2021_cat1	0.30%	NSE estimado por equifax
MTODEU_CEF_T3	0.10%	Monto deuda crédito efectivo en TIER 3 que corresponde

Tabla 30

Variables del segmento Ingresos Medios e importancia en el modelo de R^2=34.8%

Variables	Gain	descripción
MTOTOTDEU_D_I_MAX24_PJ	14.80%	Deuda total (directa más indirecta calculada para Personas jurídicas) máxima del cliente en los últimos 24 meses
CONV_1_ZN	14.30%	Convolución de (rango de trabajadores + edad +departamento + macrozona)
CUOTA_RCC_sum_12	4.40%	Suma de Cuotas RCC (por RBM) de los últimos 12 meses
mto_profesion	4.40%	Profesión
BACHILLER	4.40%	N° de bachilleres
CUOTA_COMPRAS_F_max_12	4.30%	Máxima Cuota de Compras (por RBM) de los últimos 12 meses
dem_des_mejorvehiculo_cat2	4.10%	Mejor vehículo
dem_cod_ciiu_cat2	3.90%	CIIU
CONV_1_ED	3.50%	Convolución de (rango de trabajadores + edad + sector + nivel educativo)
MTOLINTOT_DM0_PRO24_V1	3.40%	Corrección en línea total(=0) que presentan saldo positivo en deuda en los últimos 24 meses
ING_PROM_MZNA_20	2.90%	Ingreso promedio en la manzana suma (ingresos por nse) / poblacio total por nse
BACHILLER_SUNEDU	2.60%	N° de bachilleres licenciadas por SUNEDU
MONTOADE_ACT6_MAX24	2.30%	Ratio de la deuda en el sistema financiero de los 6 últimos reportes de RCC sobre la deuda máxima en los 24 meses
dem_num_antigminvehiculomes	2.30%	N° de meses de antigüedad del vehículo

Sectores_MMGR_cat2	2.10%	Sector
dem_num_antigrucmes	2.10%	Antigüedad del ruc en meses
CUOTA_OTROS_F_max_12	1.90%	Máxima Cuota de productos asociado a TC a excepción de compras y disposición de efectivo (por RBM) de los últimos 12 meses
CTDEMPREPORTADOCLIMED24	1.80%	Número promedio de empresas que reportan al cliente en los últimos 24 meses
NMESES_MTOTJCNSDSEF24	1.80%	Número de meses con disposición de efectivo en los últimos 24 meses
PRC_PER_45_59	1.50%	porcentaje de personas de 45-59
cv_estcivil_edadn2	1.40%	Convolución de estado civil y edad
conv42_ing_med	1.40%	Convolución continua entre meses con mal comportamiento Los ult. 24 meses y deuda TC promedio los ult.24 meses.
MTOLINTCDMI0_PRO24	1.40%	Línea promedio con tarjeta de crédito de personas que registran deuda negativa en los últimos 24 meses
PRC_PER_60_MAS	1.30%	porcentaje de personas de 60 a más adulto mayor
FLG_LUJO	1.20%	Flag que indica 1 si el auto del cliente es de gama media (no incluye marcas chinas).
ING_B_20	1.20%	Ingresos totales de personas de nivel socieconomico dado en la proyección 2020
TIPSITUACIONCASA_cat2	1.10%	Situación casa
REFTEN	1.10%	Cantidad de hogares que tienen refrigerador en la mzna
PORC_REPROG_TOT	1.00%	Porcentaje de deuda reprogramada total.
SOW_MTODEU_CEF_T1	0.90%	Porcentaje de deuda CEF en entidades TIER 1
ubicacionpro_cat1	0.80%	Provincia
LINEA_TC_T3	0.70%	Linea de Tarjeta en tier 3 que corresponde
SOW_MTODEU_TC_T1	0.70%	Porcentaje de deuda TC en entidades TIER 1
CTD_PRODTOTAL_12	0.70%	Número promedio de productos del tipo consumo y empresa en los últimos 12 meses
SF6_SF12	0.60%	Ratio de saldos medios de deuda de los últimos 6 meses respecto a los 12 últimos meses
CUOTA_MV_FINAL_sum_6	0.60%	Suma de Cuotas vehicular RCC (por RBM) de los últimos 6 meses
MTODEU_TC_T2	0.50%	Monto deuda de tarjeta en tier 2 que corresponde
flg_ind12	0.40%	Flag que indica si el trabajador ha sido independiente en el último año
dem_flg_modelovehiculartop	0.20%	Flag que indica si el modelo del vehículo es un top.
FLG_BACHILLER	0.10%	Flag que indica si tiene o no tiene bachiller

Tabla 31

Variables del segmento Ingresos Medios e importancia en el modelo de R^2=34.8%

Variables	Gain	Descripción
MTODEUDAMAX24_IND	19.90%	Deuda máxima en los últimos 24 meses incluyendo deuda indirecta
CONV_1_ZN	12.70%	Convolución de (rango de trabajadores + edad +departamento + macrozona)
CONV_1_ED	9.20%	Convolución de (rango de trabajadores + edad + sector + nivel educativo)
dem_cod_ciiu_cat3	5.60%	CIIU
BACHILLER_SUNEDU	4.90%	N° de bachiller por Sunedu
CUOTA_HP_FINAL_max_12	4.30%	Máxima Cuota hipotecaria RCC (por RBM) de los últimos 12 meses
CUOTA_COMPRAS_F_max_12	4.20%	Máxima Cuota de Compras RCC (por RBM) de los últimos 12 meses
cv_estcivil_sector3	4.20%	Convolución de estado civil y sector
dem_des_mejorvehiculo_cat3	4.00%	Mejor Vehículo
MTOLINTC_MAX24_V1	3.90%	Línea máxima con tarjetas de crédito en los últimos 24 meses
CUOTA_RCC_max_6	3.70%	Máxima Cuota RCC (por RBM) de los últimos 6 meses
mto_profesion	3.60%	Profesión
MTODEUDAMAX24_ MTODEUDAACTUAL	3.50%	Diferencia entre la deuda directa máxima de los últimos 24 meses y la deuda directa actual
ING_PROM_MZNA_20	2.20%	Ingreso promedio en la mzna suma (ingresos por nse) / pob total por nse
MONTOADE_ACT6_MAX24	1.90%	Ratio de la deuda en el sistema financiero de los 6 últimos reportes de RCC sobre la deuda máxima en los 24 meses
RT_REFRI_HOGARES	1.80%	ratio de hogares que tienen refrigerador en la manzana
CTDEMPREPORTADOCLIMAX12	1.40%	Número máximo de empresas que reportan al cliente en los últimos 12 meses
CUOTA_VEH_FINAL_min_12	1.30%	Mínima Cuota vehicular RCC (por RBM) de los últimos 12 meses
CTD_PER_60_MAS	1.10%	Cantidad de personas de 60 a más adulto mayor
MTOLINTOT_DMI0_12_V1	1.00%	Línea de crédito total de clientes que registran deuda negativa en los últimos 12 meses
SF12_SF24_IND	1.00%	Ratio de saldos medios de deuda incluyendo deuda indirecta de los últimos 12 meses respecto a los 24 últimos meses
N_AUTOS_BAJA	1.00%	N° de autos de baja categoría
PORC_REPROG_CON	0.90%	Porcentaje de deuda reprogramada consumo

ATRASOMAX_TDE_12	0.80%	Número máximo de días de atraso en tarjeta de disposición de efectivo en los últimos 12 meses
N_AUTOS_GLUJO	0.60%	N° de autos de lujo
ubicacionpro_cat1	0.60%	Provincia
TIPSITUACIONCASA_cat3	0.60%	Situación casa
flg_ind12	0.40%	Flag que indica si el trabajador ha sido independiente en el último año

Como se puede apreciar en las tablas anteriores, donde se muestra las diferentes variables seleccionadas en cada uno de los segmentos, estas son variables que hacen referencia al nivel de educación, comportamiento de pago, deuda directa o indirecta, número de cuotas máxima, monto de deuda, porcentaje de deuda, ratio de deuda en el sistema financiero, entre otros; estas son variables que se encuentran dentro de los indicadores financieros como lo mencionan los autores del Análisis de riesgo crediticio (2018), propuesta del modelo credit scoring de la revista Facultad de Ciencias Económicas: Investigación y Reflexión.

V. CONCLUSIONES

Como resultado de evaluar la calidad de los datos, la metodología aplicada, los procedimientos de cálculo realizados y los resultados de los procesos de implementación del Modelo Estimador de Ingresos (Personas) para Dependientes + RCC de la Banca Minorista, se puede concluir que la implementación del modelo cumple con los estándares establecidos por las mejores prácticas, en auditoría de Validación a la implementación. Además, dicho modelo fue implementado de forma adecuada siguiendo los lineamientos definidos por las unidades del banco (Unidad de Modelamiento y el Squad de Data).

Con respecto al objetivo de los controles de calidad de información; se concluye que se han incluido las variables seleccionadas de forma correcta y los resultados se encuentran alineados a las necesidades del negocio. Sin embargo, con respecto a la calidad de datos, no se evidenció que en el proceso de generación de datos se esté realizando el control de calidad a algunas variables relevantes para el modelo incumpliendo lo establecido en la norma N°4202.010.09 (norma interna del banco); en el Anexo C se encuentra el extracto específico de la norma.

Asimismo, con respecto al objetivo del análisis estadístico en la selección de variables, se cumplió con todos lineamientos mínimos necesarios para la selección idónea de variables utilizando el nivel de Gain en las variables de cada segmento, posterior a ello se aplicó la técnica de Boruta y finalmente una evaluación de sentido económico para las variables seleccionadas.

Finalmente, el ultimo objetivo relacionado a los procedimientos metodológicos durante la implementación del modelo, se ha encontrado el incumplimiento de algunos supuestos necesarios (linealidad y normalidad) para utilizar métodos de regresión lineal; además de aplicar un umbral diferenciado por tipo de variable (continua o categórica) del coeficiente de determinación (R^2) en el proceso de convolución de variables; e incluso en el proceso de selección del mejor “Modelo Estimador de Ingresos Dependiente + RCC”.

VI. RECOMENDACIONES

En base a las observaciones encontradas en el proceso de validación al Modelo Estimador de Ingresos para Dependientes + RCC de la Banca Minorista se han emitido algunas recomendaciones:

Referente a la calidad de datos, se deberá incluir en el “Framework de Calidad" la totalidad de variables pertenecientes a los modelos implementados, de acuerdo con lo indicado en la norma N°4202.010.09 o priorizar aquellas que se definan como criticas o de mayor relevancia en conjunto con las áreas modeladoras y usuarias. Asimismo, debido a que las variables provienen de distintas fuentes de información, se deberá definir los responsables de validar la calidad de datos de las tablas utilizadas por el modelo y establecer un flujo de comunicación entre los responsables de verificar la calidad de las variables input y el responsable de las variables output, a fin de que con ello se pueda identificar oportunamente variaciones inusuales o errores en las variables que forman parte del modelo.

Con respecto al uso diferenciado del coeficiente de determinación (R^2) en la convolución de variables, se deberá evaluar y posteriormente incluir en los estándares metodológicos, los umbrales necesarios que regulen el uso del R^2 en la selección de variables. Además, se deberá incluir en el manual metodológico respectivo, el sustento del uso de diferentes umbrales de R^2 en la selección de variables según su tipo o naturaleza de la variable, a fin de sustentar el criterio utilizado, mejorando el entendimiento del proceso y sus resultados.

Finalmente, se deberá evaluar, en futuros desarrollos, el cumplimiento de los aspectos necesarios para el uso del Coeficiente de determinación - R^2 y verificar para este modelo la existencia de algún impacto en los resultados. En caso de encontrar alguna debilidad material relacionada se deberá comunicar a las instancias pertinentes. Asimismo, se deberá evaluar la inclusión de algún estadístico de bondad de ajuste que no requiera el cumplimiento de algún supuesto de linealidad o normalidad como el RSME (Error cuadrático medio) y MAPE (Error porcentual absoluto medio) o algún otro, que complemente el indicador actualmente utilizado. Además, se deberá incluir dichos estadísticos en el estándar de construcción y calibración de modelos.

VII. REFERENCIAS BIBLIOGRÁFICAS

Muñoz, J. (1999). Calidad de cartera del sistema bancario y el ciclo económico: una aproximación econométrica para el caso peruano. *revista de estudios económicos*, *4*, 107-118.

Cortez, G. C. (2006). Competencia y Eficiencia en el Sector Bancario en el Perú 1990-2005. *Pensamiento Crítico*, *6*, 097-112.

Armebianchi, R. B. (2013). *Análisis de los factores motivacionales de los funcionarios del sector bancario peruano* (Doctoral dissertation, Pontificia Universidad Catolica del Peru-CENTRUM Catolica (Peru)).

García, J., Molina, J., Berlanga, A., Patricio, M., Bustamante, A., & Padilla, W. (2018). Ciencia de datos. *Técnicas Analíticas y Aprendizaje Estadístico. Bogotá, Colombia. Publicaciones Altaria, SL*.

Dipaola, E. (2020). Individualismo y pandemia: consecuencias y riesgos globales. *Reflexiones marginales*. Número especial 8: coronavirus. https://revista.reflexionesmarginales.com/individualismo-y-pandemia-consecuencias-y-riesgos-globales/

Huremović, D. (2019). Brief history of pandemics (pandemics throughout history). En D. Huremović (Ed.), *Psychiatry of Pandemics: A Mental Health Response to Infection Outbreak*. Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-15346-5_2

Tisdell, C. A. (2020). Economic, social and political issues raised by the COVID-19 pandemic. *Economic Analysis and Policy 68*, 17-28. https://doi.org/10.1016/j.eap.2020.08.002

Banco Central de Reserva del Perú (2020). *Estadísticas*. https://www.bcrp.gob.pe/estadisticas.html

MTPE (2020b). Boletín Mensual de Leyendo Números. Mayo 2020 Referido a https://cdn.www.gob.pe/uploads/document/file/911115/Boletin_Total_LN_Mayo2020.pdf

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package. Journal of Statistical Software, 36(11):1–13.

Everitt, B. S., & Skrondal, A. (2010). The Cambridge dictionary of statistics.

Maguiña, C. (2020). Reflexiones sobre el COVID-19, el Colegio Médico del Perú y la Salud Pública. *Acta Médica Peruana 37*(1), 8-10 https://doi.org/10.35663/amp.2020.371.929

Fiz Mayo, M. L. (2018). La gestión del riesgo de modelo: aspectos cualitativos y cuantitativos.

García, J. C. T., García, M. Á. M., & Martínez, F. V. (2017). Administración del riesgo crediticio al menudeo en México: una mejora econométrica en la selección de variables y cambios en sus características. *Contaduría y administración*, *62*(2), 377-398.

García, H. (1996). Qué es análisis estadístico bivariado. *Sigma*, (7), 33-40.

CEVALLOS TORRES, L. J., VALENCIA MARTINEZ, N. A., & BARROS MORALES, R. L. (2017). Análisis Estadístico Univariado.

Balzarini, M., Bruno, C., Córdoba, M., & Teich, I. (2015). Herramientas en el análisis estadístico bivariado. *Escuela Virtual Internacional CAVILA. Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba. Córdoba, Argentina*.

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, *21*(3). https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Guerrero, J. (2016). El problema de la dimensionalidad. *Revista Indice* (68) 22-24

Rouhiainen, L. (2018). Inteligencia artificial. *Madrid: Alienta Editorial*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/10.1007/bf00058655

Brasa Estévez, P. (2019) en galego: Análise estatístico de datos para o mantemento preditivo de. http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1564.pdf

Alsahaf, A., Petkov, N., Shenoy, V., and Azzopardi, G. (2022). A framework for feature selection through boosting. Expert Systems with Applications, 187

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1):273–324.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data 785-794. KDD '16.

Castañeda, J., Télleza, C., & Fúquene, J. (2019). Una alternativa para la estimación del ingreso promedio mediante métodos de estimación en áreas pequeñas An alternative for the average income estimation using small area methods. *arXiv preprint arXiv:1907.05387*.

Ivanovna, A., & Lucionovna, V. (2013). Propuesta *metodológica para la estimación de ingresos en otorgamiento de créditos.*

Álvarez, M. D. C. V., & Rivera, Z. (2006). La auditoría como proceso de control: concepto y tipología. *Ciencias de la Información*, *37*(2-3), 53-59.

VIII.- ANEXOS

Anexo A. Fuentes de información

Matriz de Modelamiento

1) RCC.

La fuente de RCC se compone por tablas mensuales que tienen la estructura: RCC_NEW_VARS_AAAAMM. Dicha fuente se encuentra almacenada en la ruta: /sasdata10/SI_ADR_RECU_02/RCC. De dicha fuente se extraen todas las variables de la fuente y se cruzan con el mes de referencia con un desfase de 1 mes. El cruce se realiza a nivel de CODCLAVECUC.

2) Resumen Saldo.

La fuente de RCC se compone por tablas mensuales que tienen la estructura: VARIABLES_RS_AAAAMM. Dicha fuente se encuentra almacenada en la ruta: /sasdata10/SI_ADR_RECU_02/ResumenSaldo. De dicha fuente se extrae las variables asociadas a saldos tanto en activos y pasivos en los últimos 12 meses y se cruzan con el mes de referencia. El cruce se realiza a nivel de CODCLAVECIC con el fin de replicar la segmentación. Las variables extraídas son:

- PMACTIVO_MED_12
- PMPASIVO_MED_12

Agregados RCC

1) Tiers.

De la tabla PROY_RBP.HM_SBS_TIER_PER se obtienen los siguientes campos:

- CTDPRODUCTOSSBS
- CTDPROD_T1
- CTDPROD_T2
- CTDPROD_T3
- LINEASBS
- LINEA_TC_T1
- LINEA_TC_T2

- LINEA_TC_T3
- MTODEUDATOTALSBS
- MTODEUDATOT_T1
- MTODEUDATOT_T2
- MTODEUDATOT_T3
- MTODEU_CEF_T1
- MTODEU_CEF_T2
- MTODEU_CEF_T3
- MTODEU_COM_T1
- MTODEU_COM_T2
- MTODEU_COM_T3
- MTODEU_HIP_T1
- MTODEU_HIP_T2
- MTODEU_HIP_T3
- MTODEU_MICRO_T1
- MTODEU_MICRO_T2
- MTODEU_MICRO_T3
- MTODEU_TC_T1
- MTODEU_TC_T2
- MTODEU_TC_T3
- MTODEU_VEH_T1
- MTODEU_VEH_T2
- MTODEU_VEH_T3
- MTOSALDOSBS
- MTOSALDOTOT_T1
- MTOSALDOTOT_T2
- MTOSALDOTOT_T3

El cruce se realiza a nivel de CODCLAVECUC.

2) Reprogramados SBS.

De la tabla PROY_RBP.HM_REPROG_COVID_SBS se obtienen los siguientes campos:

- MTOREPROGCV
- MTOREPROGCV_CONS
- MTOREPROGCV_MICRO_PEQ

El cruce se realiza a nivel de CODCLAVECUC.

3) Cuota SBS

RBM nos compartió una tabla UM_CUOTAS_RCC7_1909_2102 y UM_CUOTAS_RCC7_TC_1909_2102 con información de cuotas RCC (Compras con tarjeta de crédito, disposición de efectivo, cuota de crédito efectivo, hipotecario, mi vivienda, vehicular y total RCC) el cual fue almacenada y tratada agregando variables sumarizadas en una tabla SAS AGREGADOS_NEW_CUOTA_RCC_AAAMM. De dicha fuente se extraen todas las variables y se cruzan con el mes de referencia con un desfase de 1 mes. El cruce se realiza a nivel de CODCLAVECUC.

4) Antigüedad en la SBS

De la tabla PROY_RBP.HM_BE_BURO_SCORE se obtiene el siguiente campo:

- ANTIGÜEDAD_RCC48

El cruce se realiza a nivel de CODCLAVECUC.

Empresa proveedora de información

1) Situación laboral

De la base de datos comprada a una empresa (materializada en una tabla) se obtienen los siguientes campos:

- FLAG_DEPENDIENTE
- FLAG_INDEPENDIENTE
- FLAG_INFORMAL
- RUC_EMPLEADOR

El cruce se realiza a nivel de RUC.

2) Patrón vehicular.

Se extraen todas las variables asociadas a la información vehicular de los clientes desde una base facilitada por Experian "TP_VEHICULAR_EMPRESA", a ello se realizó un tratamiento para crear nuevas variables asociadas a la clasificación del tipo de auto que cuenta el cliente (alta, media, baja categoría, auto nuevo, etc.). El cruce se realiza a nivel de cliente.

Maestras Demográficas

1) Relación cliente y manzana.

De la tabla s43181.UM_GEO_CIC se obtienen los siguientes campos:

- CODUBIGEO
- IDZMZNA_CENSO2017

El cruce se realiza a nivel de CODCLAVECIC.

2) Información de persona natural.

De la tabla ODS.MD_PERSONANATURAL se obtienen los siguientes campos:

- Edad
- Sexo
- Estado civil
- Nivel de educación
- Tipo de situación de la casa
- Profesión

El cruce se realiza a nivel de CODCLAVECIC.

3) Zona geográfica.

De la tabla PROY_RBP.de_zonageografica_peru se obtienen los siguientes campos:

- Departamento

El cruce se realiza a nivel de CODUBIGEO.

4) Relación cliente y Empresa.

De la tabla PROY_RBP.HM_CLIENTE_EMPRESA_RELACION se obtienen los siguientes campos:

- CODACTECONOMICA
- CODSUBSEGMENTO
- FLG_EXPORTADOR
- FLG_IMPORTADOR

El cruce se realiza a nivel de CODCLAVECIC.

5) AEMA

De la tabla HM_DETCONCEPTODEMOGRAFICO (del datalake) fue almacenada en una tabla SAS AEMA_DEMO. De dicha fuente se extraen todas las variables y se cruzan con el mes de referencia. Dichas variables son:

- CODINTERNOCOMPUTACIONAL_F
- dem_cod_subsegmento
- dem_cod_segmento
- dem_des_segmento
- dem_cod_sector
- dem_des_grupocarterizado
- dem_flg_carterizado
- dem_cod_banca
- dem_des_banca
- dem_num_edad
- dem_des_sexo
- dem_des_estadocivil1
- dem_des_estadocivil2
- dem_des_niveleducacional2
- dem_des_niveleducacional4
- dem_des_tipsituacionlaboral
- dem_des_tipsituacioncasa
- dem_ctd_hijos
- dem_cod_profesion1
- dem_des_profesion1
- dem_des_grupo1profesion1
- dem_des_grupo2profesion1

- dem_cod_profesion2
- dem_des_profesion2
- dem_des_grupo1profesion2
- dem_des_grupo2profesion2
- dem_flg_direccion
- dem_cod_distrito
- dem_des_distrito
- dem_cod_provincia
- dem_des_provincia
- dem_cod_departamento
- dem_des_departamento
- dem_des_macrozonadem
- dem_des_regiondem
- dem_cod_ciiu
- dem_flg_direccionbcp
- dem_cod_distritobcp
- dem_des_distritobcp
- dem_cod_provinciabcp
- dem_des_provinciabcp
- dem_cod_departamentobcp
- dem_des_departamentobcp
- dem_des_macrozonadembcp
- dem_des_regiondembcp
- dem_fec_nacimientobcp
- dem_fec_nacimiento
- dem_tip_sexobcp
- dem_tip_estcivilbcp
- dem_cod_profesionbcp
- dem_des_profesionbcp
- dem_cod_ciiubcp
- dem_tip_niveleducacionalbcp
- dem_des_mejorvehiculoreciente

- dem_des_mejorvehiculo
- dem_des_peorvehiculo
- dem_flg_marcavehiculartop
- dem_flg_modelovehiculartop
- dem_cod_marcavehiculartop
- dem_cod_modelovehiculartop
- dem_ctd_vehiculototal
- dem_ctd_vehiculoclases
- dem_num_antigminvehiculomes
- dem_ctd_antigmaxvehiculomes
- dem_flg_ingreso
- dem_des_grupofuenteingreso
- dem_cod_fuenteingr
- dem_mto_ingresosol
- dem_des_grupofuenteingresoant
- dem_cod_fuenteingrant
- dem_mto_ingresoantsol
- dem_flg_fadcliente
- dem_flg_fadncliente
- dem_flg_fadtotalcliente
- dem_mto_fadestimado
- dem_flg_basesunat
- dem_flg_rucactivo
- dem_num_antigrucmes
- dem_tip_contribuyente
- dem_tip_estcontribuyente
- dem_tip_clasifcomercioexteriorsu

El cruce se realiza a nivel de CODINTERNOCOMPUTACIONAL.

6) Información de censo.

De la tabla ME_CENSO_REFIN se extraen todas las variables asociadas al censo (número de viviendas, hogares, hombres, mujeres, ingreso promedio por mzna, etc.) y

se cruzan con el mes de referencia. El cruce se realiza a nivel de ID_MZNA (código de manzana).

Datos del empleador

1) Información económica del cliente.

De la tabla ODS_V.MD_CLIENTE se obtiene el siguiente campo:

- CODACTECONOMICA

El cruce se realiza a nivel de CODCLAVECIC con la tabla S83672.HM_TABLON_EMPRESA.

2) Información CIIU del cliente.

De la tabla S83672.HM_TABLON_EMPRESA se obtiene el siguiente campo:

- CIIU_REV3

El cruce se realiza a nivel de RUC.

3) Información de ventas del empleador.

De la tabla S75701.hm_flujoventas se obtiene el siguiente campo:

- FLUJOVENTA_EMPLEADOR

El cruce se realiza a nivel de CODCLAVECIC.

EQUIFAX

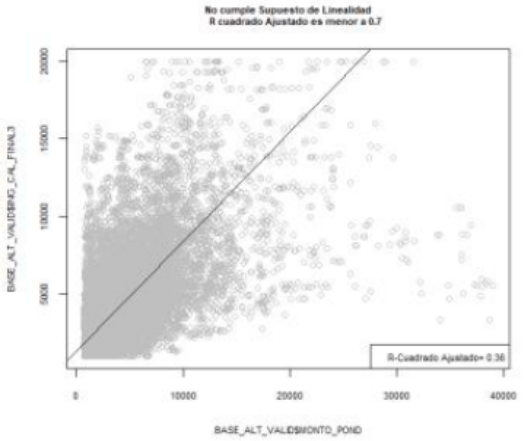
De la tabla PROY_RBP. MM_PROYECTADA_EQX se extraen todas las variables asociadas al censo elaboradas por Equifax y se cruzan con el mes de referencia. El cruce se realiza a nivel de CODCLAVECIC.

SUNEDU.

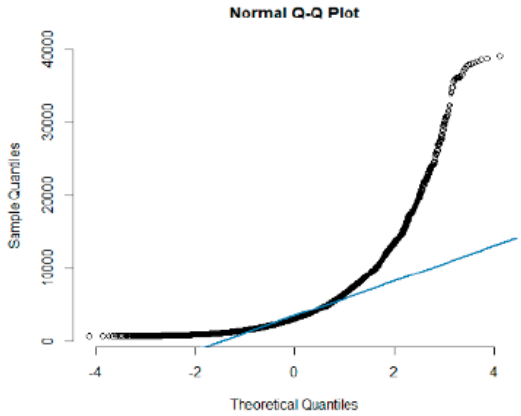
Se extraen todas las variables asociadas a la información de la SUNEDU desde la tabla TP_SUNEDU2 para los clientes y se realiza un tratamiento creando nuevas variables (número de títulos, número de bachiller, doctorado, flag de universidad pública, etc.). El cruce se realiza a nivel de CODCLAVECIC.

Anexo B. Evidencias del no cumplimiento de los supuestos en el uso de una regresión lineal

Revisión de linealidad



Revisión de normalidad



Nota: Revisión de supuestos del modelo de regresión convolucionada de las variables Edad con Sector del segmento de ingresos bajos.

Anexo C. Extracto de la Norma N°4202.010.09 - norma interna del banco

Unidad Proveedora de Datos

Las funciones de esta unidad son las siguientes:

a. En la fase de desarrollo del modelo, extraer los datos necesarios y brindar soporte a la unidad modeladora para su entendimiento.

b. Establecer los controles que fueran necesarios para asegurar la calidad de los datos. Estos controles deben extenderse de manera continua a las variables que utilizan los modelos implementados para asegurar el funcionamiento de los modelos durante toda su vida.