UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"IDENTIFICACIÓN DE LA PROPENSIÓN A LA ADQUISICIÓN DE UN SUBPRODUCTO DE UNA TARJETA DE CRÉDITO EN UNA ENTIDAD BANCARIA"

PRESENTADO POR

CINTHIA ZARABIA YUPANQUI

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE

INGENIERA ESTADÍSTICA E INFORMÁTICA

Lima – Perú

2020

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"IDENTIFICACIÓN DE LA PROPENSIÓN A LA ADQUISICIÓN DE UN SUBPRODUCTO DE UNA TARJETA DE CRÉDITO EN UNA ENTIDAD BANCARIA"

PRESENTADO POR

CINTHIA ZARABIA YUPANQUI

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL

TÍTULO DE INGENIERA ESTADÍSTICA E INFORMÁTICA

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

.....	.....
M.A. Fernando René Rosas Villena Presidente	Mg. Jesús Walter Salinas Flores Asesor
.....	.....
Mg. Iván Dennys Soto Rodríguez Miembro	MS. Grimaldo Febres Huamán Miembro

Lima – Perú
2020

DEDICATORIA

A mis padres, Gladys Yupanqui y Eloy Zarabia, por motivarme siempre a poder lograr mis metas y enseñarme que a base esfuerzo y dedicación todo es posible. Y a mis 8 hermanos por ser mi fuente de inspiración para seguir creciendo profesionalmente.

AGRADECIMIENTOS

Agradezco a Dios por su amor infinito y guía constante a lo largo de mi vida, y por darme fortaleza en momentos de debilidad o desánimo.

A la empresa donde pude aplicar estos conocimientos adquiridos en mi carrera profesional, y por haberme proporcionado los datos necesarios para llevar a cabo mi investigación.

A la Universidad Nacional Agraria La Molina por ser mi alma mater y permitirme convertirme en una gran profesional. Al Ing. Mg. Jesús Walter Salinas Flores, asesor de la presente tesis, mi enorme agradecimiento, por su orientación y consejos durante el desarrollo del presente trabajo.

A mi hermana mayor Elizabeth Zarabia por ser mi ejemplo y referente desde muy niña y siempre motivarme a seguir luchando por mis sueños.

A toda mi familia por su confianza y motivación para ser una gran profesional, pero sobre todo por su amor y apoyo en cada decisión tomada.

A mis amigos por sus palabras de ánimos, consejos y por su apoyo incondicional durante esta etapa profesional.

Índice de Contenido

1. PRESENTACIÓN 1

2. INTRODUCCIÓN 2

3. OBJETIVOS 4

3.1. Objetivo General 4

3.2. Objetivos Específicos 4

4. CUERPO DEL TRABAJO 5

4.1. Funciones Desempeñadas 5

4.2. Puesta en práctica de lo aprendido en la carrera 6

4.2.1 Descripción de las técnicas estadísticas 6

4.2.2 Revisión de artículos científicos 11

4.2.3 Propuesta de alternativa de solución a la situación problemática 16

4.3. Contribución en la solución de situaciones problemáticas 27

4.4. Análisis de la contribución en términos de competencia y habilidades 29

4.5. Nivel de beneficio obtenido por el centro laboral 29

5. CONCLUSIONES Y RECOMENDACIONES 30

5.1. Conclusiones: 30

5.2. Recomendaciones: 31

6. REFERENCIAS BIBLIOGRÁFICAS 32

7. ANEXOS 34

Índice de Tablas

Tabla 1: Representación de una Matriz de Confusión 9
Tabla 2: Tasa de adquisición de Extra Línea en la data de análisis 19
Tabla 3: Diccionario de variables 20
Tabla 4: Resumen de variables cuantitativas 21
Tabla 5: Criterio para la selección de dos variables correlacionadas 24
Tabla 6: Lista de variables seleccionadas 24
Tabla 7: Datos Balanceados mediante SMOTE 25
Tabla 8: Comparación de Modelos 26
Tabla 9 : Evaluación del Modelo en un mes de campaña 28

Índice de Gráficos

Gráfico 1: Esquema del proceso interno del algoritmo Random Forests 8
Gráfico 2: Representación de la curva ROC 10
Gráfico 3: Vista del diseño de una cosecha 18
Gráfico 4: Gráfica de la importancia de variables 23
Gráfico 5: División de la base de datos 25
Gráfico 6: Ventas acumuladas por deciles de propensión 28

Índice de Anexos

Anexo 1: Códigos de Procesamiento en R 34

1. PRESENTACIÓN

El caso aplicativo se desarrolló en el área de Inteligencia Comercial de una entidad financiera, que tiene como principal objetivo ayudar a la entidad en el crecimiento del número de colocaciones y captaciones de los productos activos y pasivos respectivamente, elevando así el número de ventas y desembolsos en los distintos productos ofrecidos. Las funciones realizadas fueron: el desarrollo e implementación de modelos predictivos, las propuestas de estrategias comerciales y el seguimiento de los modelos.

Se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que incluye estándares de mejores prácticas en el desarrollo de proyectos de Data Mining; mediante la cual se desarrollan modelos analíticos para las distintas fases del ciclo de vida del cliente y para cada producto de la entidad, modelos como: adquisición de tarjeta de crédito, desembolso de préstamos personales, propensión a la apertura de una cuenta de ahorros, propensión a la adquisición de una Extra Línea, etc., con el fin de identificar y priorizar a los clientes más propensos a la toma de productos y encontrar los principales drivers que influyan en la toma del producto, de manera que permita a la entidad realizar estrategias diferenciadas de acuerdo a la propensión del cliente a la toma de un producto en específico. Todo los modelos desarrollados se implementaron en el negocio enfocados en estrategias comerciales, permitiendo incrementar los principales indicadores de las campañas en la entidad financiera (efectividad de ventas y montos de desembolso).

También se trabajó con el equipo de Analytics Internacional de la entidad financiera en el codesarrollo del modelo de Customer Lifetime Value (CLV) para los clientes Retail de la entidad con el fin de calcular el CLV a nivel cliente, como el valor actual neto de los ingresos obtenidos del cliente para los próximos 12 meses, permitiendo este modelo aplicar estrategias de retención, fidelización y priorización de acuerdo al valor del cliente.

Actualmente, se desarrollan modelos de pricing analytics para los distintos sub productos de préstamos personales de la entidad, con el fin de asignar las tasas óptimas a nivel cliente, de modo que se maximice el beneficio obtenido por la entidad y la satisfacción del cliente se mantenga, ya que se le asignará una tasa de acuerdo a su sensibilidad de precio respecto al producto.

2. INTRODUCCIÓN

Es importante ser competitivos en el mercado de créditos del sistema financiero para mejorar la experiencia del cliente, y esto se puede lograr através de la analítica en los datos.

En este sentido el CRM (Customer Relationship Management) analítico utiliza el Data Mining o la explotación de datos para conocer el comportamiento del cliente para poder ofrecerles un mejor servicio, por medio de la segmentación y la identificación del perfil efectivo de los clientes de una entidad financiera. Esto permite a la entidad diseñar acciones comerciales a medida para cada uno de los segmentos, con el fin de satisfacer mejor las necesidades del cliente. Estos clientes que se sienten valorados y reciben un servicio de calidad son menos propensos a buscar otra entidad, mejorando la retención de clientes y maximizando las oportunidades de ventas.

Las entidades del rubro financiero se esfuerzan cada vez más por conocer el comportamiento de sus clientes en los distintos productos de la entidad, trabajando en la identificación de clientes que tienen una alta propensión a responder positivamente a la adquisición de un producto, permitiéndoles enfocar sus campañas de adquisición en dichos clientes, controlando así los costos de adquisición y aumentando la efectividad de ventas en las campañas. En este rubro, los modelos de propensión se utilizan para identificar a los clientes con el perfil adecuado ante la aceptación de un producto, desarrollando así modelos para cada producto de la entidad como para las distintas fases del ciclo de vida del cliente, clasificándolos y gestionándolos en las campañas comerciales según su tendencia a comprar un producto.

En base a lo anterior, el objetivo general de este proyecto es identificar a los clientes con tarjeta de crédito de una entidad financiera más propensos a la adquisición de un subproducto de tarjeta de crédito llamado Extra Línea, que es una línea paralela aprobada en la tarjeta de crédito con la cual se puede disponer de efectivo a tasas y plazos preferenciales sin alterar su línea de crédito. Es decir, se podrá seguir realizando consumos con la tarjeta de crédito con total normalidad. Para este análisis se usó como universo, la base de los clientes con tarjetas de crédito de la entidad financiera, analizando la información histórica de estos clientes durante los 11 últimos meses. También se buscó identificar a los factores que influyen en la propensión de los prospectos a responder positivamente a esta oferta (adquisición de una Extra Línea).

Esto permitió a la entidad financiera limitar y enfocar sus esfuerzos en aquellos clientes que son más propensos a la adquisición del sub producto llamado Extra Línea, constribuyendo en la solución de la situación problemática de tener una mejor gestión de registros en las campañas comerciales de este subproducto, ya que tener el desarrollo de una herramienta analítica como es el modelo de propensión permitió tener una gestión más eficiente en la distribución de registros para las campañas comerciales de este sub producto, ya que se tenían recursos limitados de tiempo y dinero para la gestión de esta campaña, y sin embargo se tenía una población de posibles prospectos muy amplia, ya que son 400 mil clientes aproximadamente con tarjeta de crédito en la entidad financiera a los cuales se les puede ofrecer este subproducto.

Para el desarrollo de este modelo de propensión se utilizó la metología CRISP-DM (Cross Industry Standard Process for Data Mining) que inicialmente permitió tener un conocimiento del negocio correcto, permitiendo que los objetivos del modelo analítico estuviesen muy alineados al objetivo del negocio. A continuación se realizó un entendimiento de los datos por medio de un análisis exploratorio sobre la información histórica del cliente, ya que la propensión a responder positivamente a una oferta específica depende de una serie de factores, como la información demográfica del cliente (edad, sexo, estado civil, nivel de ingresos, situación laboral, etc.), así como la información registrada en el banco (cantidad de tarjetas de crédito, límites de crédito, historial de crédito, clasificación de riesgo) y la información en el sistema financiero (saldos en préstamos hipotecarios, préstamos vehiculares, préstamos personales, etc.). Luego, se trabajó en la limpieza de la data, transformación y creación de variables adicionales para el análisis, continuando con la selección de variables más significativas por medio de técnicas de selección de variables, usando así este conjunto de datos para el desarrollo del modelo predictivo. Posteriormente se trabajó en la fase del modelado donde se seleccionaron y aplicaron las técnicas predictivas de clasificación, haciendo una calibración de sus hiper parámetros a valores óptimos del olgoritmos utilizado. En la siguiente fase del proyecto, se aplicaron distintas pruebas para selección del modelo final donde se usaron distintos criterios de evaluación para elegir el mejor modelo, y adicionalmente se realizó una revisión con las personas del negocio para validar si hay algún punto (regla de negocio o variables) que no haya sido considerada suficientemente. Al final de esta fase, se obtuvo el modelo propensión final.

Como última fase se tuvo la puesta en producción del modelo, donde a partir del modelo elegido se logró estimar la probabilidad de adquisición del producto bancario y agrupar

a los prospectos en deciles de acuerdo a sus probabilidad de adquisición, de mayor a menor propensión de adquisición. Permitiendo ello tener la priorización de los clientes en el universo de prospecto y así enfocar las acciones comerciales en los deciles con mayor propensión.

El éxito de la correcta clasificación del cliente para el negocio se midió según su tendencia a comprar el producto del banco, es decir la tasa de respuesta que expresa la proporción de clientes que realmente compraron el producto. Esta tasa de respuesta se debe dar en orden decreciente para los clientes con deciles altos (mayor propensión) respecto a los deciles más bajos (menor propensión).

3. OBJETIVOS

3.1. Objetivo General

Identificar a los clientes con tarjeta de crédito más propensos a la adquisición del subproducto Extra Línea de la tarjeta de crédito en una entidad financiera.

3.2. Objetivos Específicos

- Identificar las principales variables que influyen en la adquisición de una Extra Línea, para clientes con tarjeta de crédito de una entidad financiera.
- Determinar el perfil de clientes con alta propensión a la adquisición de la Extra Línea.
- Comparar el desempeño del algoritmo Random Forest en datos sin balancear y con balanceo usando el algoritmo SMOTE mediante los indicadores: Tasa de Correcta Clasificación (TCC), Sensibilidad y Especificidad .

4. CUERPO DEL TRABAJO

4.1. Funciones Desempeñadas

Empresa de BPO (Business process Outsourcing), desde agosto 2014 hasta mayo 2016

Desarrollo de proyectos de Business Analytics, utilizando técnicas de Data Mining y Machine Learning, desde la etapa inicial de levantamiento de información hasta la etapa de implementación del modelo, así también el seguimiento a la medición de los resultados de los proyectos a través de indicadores comerciales de las campañas (efectividad de ventas).

Entidad Financiera, desde junio 2016 hasta la actualidad

Desarrollo de modelos predictivos para las distintas fases del ciclo de vida del cliente y vista productos. Así como también la realización de segmentaciones que permitan a la entidad financiera enfocarse en cierto grupo de clientes de acuerdo a objetivos específicos de la entidad. Se utilizó la metodología CRISP, realizando modelos de propensión como el modelo de adquisición de tarjeta de crédito y una Extra Línea, que tuvieron como objetivos identificar a las personas con mayor probabilidad de adquirir una tarjeta de crédito y una Extra Línea respectivamente, el modelo de desembolso de préstamos personales, que tuvo como objetivo encontrar a las personas potenciales a realizar un desembolso de préstamos personales, así también el modelo de propensión a la apertura de una cuenta de ahorros cuyo objetivo fue encontrar a las personas con un perfil propensos a tener excedentes con el fin de que le permitiera aperturar una cuena de ahorros. Usando para el desarrollo de estos modelos los algoritmos de Regresión Logística, Árboles de Decisión y Random Forest.

Se desarrollaron distintas segmentaciones tal como la segmentación de comportamiento en canales, por medio de la cual se buscó identificar los canales de preferencias de los clientes, y en base al segmento que pertenecía el cliente en función a su uso canales, a estos segmentos se les realizó distintas acciones comerciales como el crossell de productos por medio del canal de preferencia. Esta segmentación se realizó mediante la técnica de K-Means.

Se trabajó con el equipo de Internacional de Analytics de la entidad financiera en el codesarrollo de la construcción del Modelo de Customer Lifetime Value (CLV) para los clientes Retail del banco, el cual tuvo como objetivo calcular el valor del cliente, como

el valor actual neto de los ingresos obtenidos del cliente para los próximos doce meses. Este cálculo se realizó en función a dos modelos de Machine Learning. El primer modelo permitió predecir si un cliente tendrá ingreso neto de provisiones positivo para los próximos doce meses y el segundo modelo predecir el valor presente del ingreso neto de provisiones para los próximos doce meses (para clientes con ingresos positivos). Para ambos modelos se usaron los algoritmos de Gradiente Boosting Machine y Random Forest respectivamente, permitiendo estos modelos conocer el CLV del cliente, lo que ayudó a direccionar las estrategias de retención, fidelización y priorización de acuerdo al valor del cliente.

Actualmente se trabaja en el desarrollo de modelos de pricing analytics para los distintos sub productos de préstamos personales. Estos modelos permiten sugerir tasas ópimas para los clientes de manera que se maximice el beneficio del banco y a la vez mantenga la satisfacción del cliente, ya que la asignación de la tasa óptima también esta alineada a la sensibilidad del cliente. Para el desarrollo de estos modelos usualmente se realizan segmentaciones por medio de la técnica de k-Means o PAM dependiendo de la naturaleza de las variables y modelos de optimización para maximizar el beneficio.

4.2. Puesta en práctica de lo aprendido en la carrera

4.2.1 Descripción de las técnicas estadísticas

En el presente trabajo se utilizó el algoritmo Random Forest debido a que se trató de un problema de clasificación y a los buenos resultados que esta técnica ha mostrado en diversas aplicaciones en los últimos años.

Random Forest es un método versátil de aprendizaje automático capaz de realizar tanto tareas de regresión como de clasificación. Es un tipo de ensamble en máquinas de aprendizaje automático que surge como combinación de las técnicas de árboles de Clasificación y/o Regression (CART) mediante el empleo del Bootstrap y Bagging para realizar la combinación de árboles predictores en la que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

Breiman (2001) menciona que Random Forest utiliza varios árboles de decisión (también denominado bosque aleatorio) donde cada uno de ellos se entrena con un subconjunto aleatorio de casos y variables predictoras (obtenidos mediante bootstrapping) denominado in-bag, el resto de los casos forman el out-of-bag. A partir de los casos en el out-of-bag se obtiene una estimación del error de clasificación (OOB).

James (2013), afirma que con un número de árboles suficientemente grande la estimación de OOB es prácticamente equivalente a la obtenida con validación cruzada. La aleatoriedad introducida en el Random Forest disminuye la correlación entre árboles dando más sentido al uso de un conjunto de clasificadores y al utilizar varios predictores disminuye el error de generalización y se obtienen mejores resultados que con otros algoritmos (Breiman, 2001).

Por otro lado, el principal problema de Random Forest, en comparación con el análisis de un único árbol de clasificación, es que es más difícil de interpretar, ya que no se dispone de un único árbol en el que pueda verse el efecto de cada variable. Sin embargo, Random Forest permite obtener medidas acerca de la importancia que las variables predictoras han tenido en el modelo, las muestras OOB también son usadas en Random Forest para calcular la fuerza de predicción de cada una de las variables usadas, conociéndose esto como la importancia de las variables, que está condicionada a su interacción con el resto de las variables.

Williams (2011), indica que para identificar las variables de mayor importancia, el modelo Random Forest tiene en consideración dos medidas: MDA (Mean decrease accuracy) y MDG (Mean Decrease Gini).

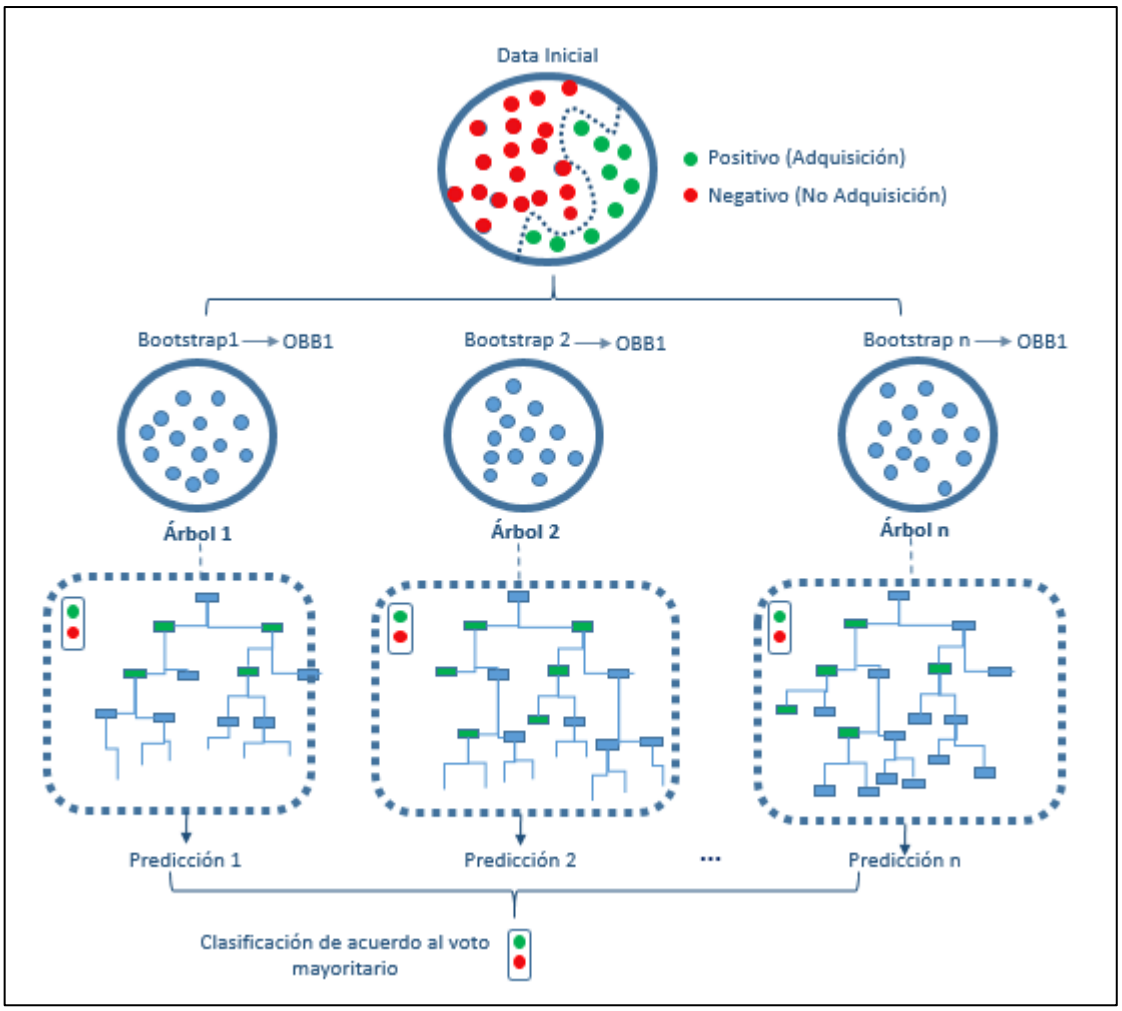
Una gran ventaja de Random Forest es la reducción de la dependencia entre árboles en la determinación de los nodos mediante la elección aleatoria de conjuntos de predictores en cada árbol. En torno a esto, Breiman (2001) sugiere determinar previamente el número de variables a elegir en cada nodo, por lo que sugiere hasta tres valores a probar: : $\sqrt{P}$,1/ 2$\sqrt{P}$ *y* 2$\sqrt{P}$, siendo P el número total de variables de predictor.

En forma resumida el algoritmo de de Ramdom Forest sigue este proceso:

1. Selecciona casos al azar con reemplazo para crear diferentes subconjuntos de datos. Cada subconjunto debe ser aproximadamente 66% del conjunto total.
2. En cada subconjunto crea un árbol donde:
    a. Las variables predictoras son seleccionados al azar entre todas las variables predictoras P.
    b. La variable de predicción que proporciona la mejor división, de acuerdo con una función objetiva, se utiliza para hacer una división binaria en ese nodo.
3. Para cada subconjunto elije otras variables predictoras al azar entre todas las variables y hace lo mismo.

4. Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva.

Gráfico 1: Esquema del proceso interno del algoritmo Random Forests



Fuente:Elaboración propia

Existen varios criterios de precisión los cuales permiten el análisis y evaluación del desempeño del modelo predictivo construido en un determinado conjunto de datos. A continuación, se presentan los principales criterios de precisión extraídos de POWERS (2011).

Tabla 1: Representación de una Matriz de Confusión

		Clase Predecida	
		No Adquisición	**Adquisición**
Clase Real	**No Adquisición**	Verdadero Negativo(VN)	Falso Negativo(FN)
	Adquisición	Falso Positivo(FP)	Verdadero Positivo(VP)

Fuente: Elaboración propia

Donde se analizan los siguientes criterios:

- **Tasa de Correcta Clasificación(TCC):** Es la proporción entre los casos correctamente clasificadas y el total de casos.

$$\text{Éxito} = \frac{VP + VN}{VP + FP + FN + VN}$$

- **Error:** Es la proporción entre los casos incorrectamente clasificadas y el total de casos.

$$\text{Error} = \frac{FP + FN}{VP + FP + FN + VN}$$

- **Sensibilidad:** Es la proporción del total de predicciones positivas sobre el total de casos positivos reales de la base.

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

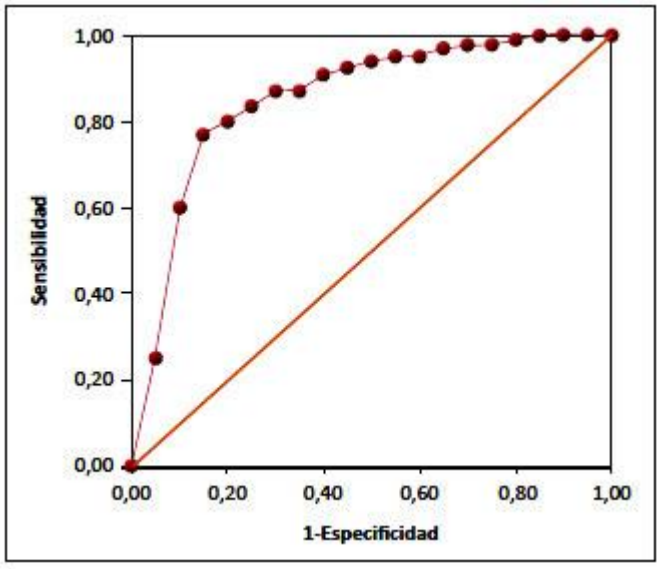
- **Especificidad :** Es la proporción de las predicciones negativas y el total de casos negativos reales de la base.

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

Otro criterio importante de evaluación es la curva ROC (curva de características operativas del receptor) que se basa en los conceptos de Sensibilidad y Especificidad.

La curva ROC poblacional representa 1-especificidad frente a la sensibilidad para cada posible valor umbral o punto de corte en la escala de resultados de la prueba en estudio. El área bajo la curva, llamado AUC proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles. El AUC es el estadístico por excelencia para medir la capacidad discriminante de la prueba. También para comparar pruebas entre sí y determinar cual és la más eficaz. Su rango de valores va desde 0.5, siendo este valor el correspondiente a una prueba sin capacidad discriminante, hasta 1, que es cuando los dos grupos están perfectamente diferenciados por la prueba. Por tanto, podemos decir que cuanto mayor sea el AUC mejor será la prueba. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio.

Gráfico 2: Representación de la curva ROC



Fuente: Revista Pontificia Universidad Católica de Chile, Santiago

También, cabe resaltar la importancia de revisar la proporción de la categoría a predecir de la variable target, ya que que las clases desbalanceadas terminan siendo un problema complejo al ser analizados debido a que la precisión de las mismas dependerá de la cantidad de datos en cada clase. Para abordar el problema de desbalanceo de clases, existen métodos de re-muestreo para balancear las clases. Los métodos aplican diferentes técnicas, entre las que se tienen:

- Sobre-muestreo: Agrega objetos de la clase minoritaria.
- Sub-muestreo: Elimina objetos de la clase mayoritaria.

- Híbrido: Combinación de técnicas de sobre-muestreo y sub-muestreo.

En el presente trabajo, se aplicó la técnica de balanceo SMOTE (Syntetic Minority Over-sampling Technique) que es un método de sobre-muestreo con el cual se genera observaciones sintéticas de la clase minoritaria con el fin de balancear las clases del conjunto de datos por clasificar, basado en la regla del vecino más cercano. (Chawla et al., 2002).

4.2.2 Revisión de artículos científicos

Los modelos predictivos se han vuelto muy usuales y de gran importancia en las entidades financieras, los modelos de propensión se utilizan para clasificar a los clientes según su tendencia a la adquisición de un producto bancario, riesgo crediticio, fuga de clientes, etc.

Yap Bee Wah & Irma Rohaiza Ibrahim (2010), indican el uso de técnicas de Minería de Datos para el desarrollo de un modelo de propensión al riesgo crediticio. Señalan que los modelos de propensión requieren técnicas de minería de datos, usan datos históricos en pagos y variables demográficas que puedan ayudar a identificar las características más importantes relacionadas con el riesgo de crédito y así proporcionar un score para cada cliente. Un modelo de calificación crediticia proporciona una estimación de riesgo de crédito del prestatario, es decir, la probabilidad de que el prestatario reembolsar el préstamo según lo prometido, en función de una serie de características cuantificables del prestatario. El objetivo de la investigación fue desarrollar un modelo de calificación crediticia para clasificar a los solicitantes de tarjetas de crédito. Se comparó la capacidad predictiva de tres modelos de calificación crediticia: Regresión logística (LR), árbol de clasificación y regresión (CART) y modelos de red neuronal (NN) en la clasificación de solicitantes de tarjetas de crédito. Las variables edad, ingresos, género, estado civil, número de hijos, número de otras tarjetas de crédito retenidas y si el solicitante tiene un préstamo hipotecario monto del préstamo, duración del préstamo, estado civil, salario mensual, ingresos adicionales, casa propia o alquiler, y nivel educativo fueron evaluadas para construir los modelos de calificación crediticia para evaluar el riesgo de crédito (pagado o impago) para la aplicación a la solicitud de tarjeta de crédito. La variable de interés (dependiente) objetivo fue una variable binaria con dos categorías: aceptada o rechazado ante la solicitud de la tarjeta de crédito. La muestra constó de 4,305 solicitantes de tarjetas de crédito mediante los cuales hay 1,330 (31%) rechazados y 2,975 (69%) aceptados solicitantes. Los datos de la muestra se dividieron primero en un

entrenamiento muestra (70%) y una muestra de validación (30%). El entrenamiento los datos de muestra se utilizan para construir los modelos, mientras que la validación para seleccionar al mejor modelo. Se hizo una comparación entre estos tres modelos para determinar el mejor modelo para predecir el estado de la aplicación. La precisión predictiva de los tres modelos es bastante comparable con el modelo de red neuronal que tiene un poco mayor porcentaje de clasificación correcta. Los resultados muestran que el modelo de red neuronal tiene una predicción de validación ligeramente más alta en la tasa de precisión (RL = 74.56%, NN = 76.46%, CART = 73.66%).

Ling Kock Sheng & Teh Ying Wah (2011) , investigan sobre el uso de modelos predictivos, tales como regresión logística, neural redes, C5, naive bayes, IBk (aprendiz basado en instancias, k vecino más cercano) y logit incremental para obtener el mejor clasificador que se usó para mejorar la precisión predictiva de riesgo de tarjeta de crédito de los consumidores de un banco en Malasia. La industria de tarjetas de crédito en Malasia ha experimentado algunos cambios importantes en la última década, como la competencia entre entidades y el aumento de los riesgos crediticios. Sin embargo los bancos todavía se sienten muy atraídos por la tarjeta de crédito, ya que es uno de los servicios más rentables para participar aunque sea muy competitivo. Por lo tanto, es importante que los bancos gestionen su proceso de mitigación de riesgos adecuadamente, para maximizar los márgenes a obtener. En los bancos, se basa con mayor frecuencia en información que pueden encontrar o extraer de su base de datos histórica sobre el prestatario y su tendencia a incumplir en su pago. El impacto de mejorar la precisión predictiva de pago puntual del cliente es esencial en el camino hacia asegurar que el banco siga siendo rentable. Dependiendo de los usos comerciales de la calificación crediticia, la metodología para construir modelos de calificación crediticia varía de banco en banco. Puede implicar, en primer lugar, una muestra de historia registros clasificados como "buenos" y "malos" (o como malos pérdida, mala ganancia y buen riesgo dependiendo del número de categorías requeridas) dependiendo de su reembolso. En la investigación se realizó un análisis cuantitativo de los datos para derivar un modelo de calificación crediticia. Con el modelo de calificación crediticia adecuado, el banco puede evaluar cualquier perfil nuevo o existente de clientes con precisión, permitiéndoles minimizar los riesgos potenciales que podrían ser inminente . Tales modelos de puntuación, junto con el información proporcionada por CCRIS (referencia de crédito central sistema de información) del Banco Central y otros proveedores de servicios de información forman la base para el crédito que se

establecieron sistemas de calificación. Los sistemas de calificación crediticia se utiliza para clasificar el riesgo de una persona como alto, medio o bajo. Esto permite el soporte de decisiones por aceptar, extender o rechazar cualquier solicitud de crédito.

En la clasificación crediticia , se utilizaron un conjunto de datos de entrenamiento como entrada para construir un modelo que describa el conjunto predeterminado de clases de datos. Una vez que la precisión predictiva del modelo es aceptable, el modelo puede usarse para predecir el futuro. Las técnicas empleadas fueron: red neuronal , regresión logística y árbol de decisión. Se utilizó una encuesta para observar el nivel de adopción y madurez en el uso de herramientas y técnicas de minería de datos en la industria bancaria en Malasia, metodología de prototipos y experimentación para la solución de minería de datos. La población objetivo de la encuesta abarca el crédito de tarjetas de todos los bancos en Malasia. Los bancos con servicios en tarjetas de crédito incluyen nueve bancos locales de anclaje y siete bancos extranjeros calificados. Las técnicas utilizadas son C5.1, red neuronal y regresión logística, naive bayes, IBk y logit incremental. Se utilizaron atributos como edad, ubicación, sexo, acumulados monto del crédito, límite de crédito / nivel de ingresos. Cada atributo es asignado como "entrada", el campo predictor mientras que el campo "información de clase" se establece como el campo "salida", una clasificación binaria de dos categorías: solicitud pago o incumplimiento en el pago, los campos predichos para un aprendizaje automático proceso de la herramienta de minería de datos. Para lograr la mejor precisión predictiva, cada modelo es entrenado y probado para su puntaje más alto. Luego, la prueba se extiende para tener una submuestra más grande de 120,000 registros que representan 24 meses de datos de 5000 registros de clientes. Cada mes de datos se agrega de forma incremental. El entrenamiento y los tamaños de las particiones de prueba se establecen en 90 y 10% respectivamente. La última parte de la prueba implica el uso del aprendizaje incremental. Los clasificadores utilizados para la prueba fueron naive bayes, IBk y logit incremental. Todos los modelos fueron entrenados y probados utilizando una muestra de 5,000 registros con el 90% de muestra utilizada para entrenamiento y 10% elegida al azar para obtener los resultados de la prueba. Las variables analizadas fueron: la edad, ubicación, sexo, monto adeudado y límite de crédito y nivel de ingresos.
La red neuronal logró el valor de precisión predictiva con 92.46% en la base de prueba. El C5 alcanzó un 94.68% usando el clasificador C5 de Clementine con una submuestra de tamaño de 120,000 registros. IBk basado en instancias tiene la mejor precisión predictiva de los tres modelos de aprendizaje incremental generados, la tasa de precisión

es del 93,63%. Naive Bayes y el Logit incremental solo podría obtener un nivel de precisión de 90.24 y 90.23%. Esto significa que el esquema de aprendizaje incremental no está funcionando mejor que el clasificador por lotes C5 que utiliza este conjunto de datos y ajuste.

Kruppa, Schwarz, Arminger y Ziegler (2013), indican que los sistemas de calificación crediticia son una parte integrante de la gestión de riesgos de las empresas, ya que su objetivo es prevenir la pérdida de deuda incobrable mediante la identificación, análisis y monitoreo del riesgo de crédito del cliente. Por lo tanto, para medir el riesgo de incumplimiento involucrado por ventas a crédito, los clientes son asignados a cierto riesgo basado en sus propensiones individuales al incumplimiento de pago. La principal fuente interna de la información sobre solvencia crediticia es la contabilidad propia de una empresa que puede proporcionar datos sobre el pago anterior de un cliente, comportamiento y características individuales, como edad, educación, profesión y residencia. Las empresas también pueden recurrir a las agencias de crédito comercial que recopilan datos de consumidores sobre criterios, como facturas impagas, solicitudes de pago emitidas por orden judicial, procedimientos de ejecución y cheques descubiertos. Estos criterios normalmente sirven como criterios de exclusión ya que entregan hechos directos en la propensión del consumidor al incumplimiento de pago. Estas variables se utilizan para construir una calificación crediticia mediante un modelo para predecir la probabilidad predeterminada de nuevos créditos. La calificación crediticia del consumidor a menudo se considera una tarea de clasificación en la que los clientes reciben un estado crediticio bueno o malo. Las probabilidades de incumplimiento proporcionan información más detallada sobre la solvencia de los consumidores, y generalmente se estiman por regresión logística.

Sin embargo, varios supuestos subyacen a estos métodos, y estos supuestos son bastante estrictos. Primero, se deben ingresar variables importantes y supuestas interacciones correctamente en el modelo. De lo contrario, pueden surgir problemas de especificación incorrecta del modelo. En segundo lugar, el modelo estándar de regresión logística no puede tratar multicolinealidad, es decir, alta correlación entre variables independientes. Siendo una alternativa al modelo paramétrico los métodos de aprendizaje automático. Por lo que en el paper se utiliza para estimar los riesgos de crédito al consumo individual los métodos de aprendizaje automático, mediante algortimo de Random Forest (RF), los k-nearest neighbors (kNN neighbors), bagged k-nearest neighbors (bNN) y adicionalmente una regresión logística optimizada, a un gran conjunto de datos de

historiales de pago completos de créditos a plazos a corto plazo. Demostrando que el algoritmo Random Forest supera al modelo de regresión logística optimizado, kNN y bNN en los datos de prueba de los créditos a plazos a corto plazo.

También señalan que los enfoques de aprendizaje automático se pueden usar fácilmente para lograr la estimación de los riesgos de crédito al consumo individual cuando se aplica a un gran conjunto de datos de calificación crediticia, ya que son computacionalmente rápidos, simples de implementar, y ya han demostrado su buen desempeño en otros problemas por lo que deberían ser considerados como grandes competidores de los modelos clásicos.

CHIRAG, Soni (2019), indica que la industria bancaria está atravesando por una transformación con el uso integral de los algoritmos de analítica avanzada en el negocio cotidiano de la banca. La adquisición de clientes a través de varios canales, la participación de los clientes existentes, la predicción de incumplimientos en las solicitudes de tarjetas de crédito o préstamos, etc., son algunas de las áreas en las que la analítica está haciendo un trabajo tremendo. De su experiencia pasada de trabajar en el equipo de Advanced Analytics de un banco multinacional líder donde usó conceptos de análisis y aprendizaje automático para resolver problemas comerciales complejos.

Resalta que en los bancos hay disponibilidad de mucha información sobre sus patrones de compra, demografía, transacciones, solicitudes de servicio, etc. y que esto se está utilizando de manera eficiente para predecir la propensión de un cliente a comprar un producto específico. Las campañas realizadas en la entidad bancaria generalmente contienen una oferta atractiva específica para el producto, como una tasa de interés más baja para una tarjeta de crédito, un interés más alto en una cuenta de ahorro, etc., que se cubre con la confianza de que un número significativo de clientes de la lista clasificada estaría dispuesto a comprar el producto, a diferencia de los contactados al azar. No sólo se predicen clientes potenciales para un producto, sino también qué clientes tienen la mentalidad de cerrar su cuenta (fuga del cliente).

Los algoritmos de ciencia de datos, como la regresión logística, funcionan bien para predecir la probabilidad de propensión del cliente a comprar o la probabilidad de deserción del cliente. Por ejemplo, un banco líder de MNC quiere formular una estrategia para frenar el desgaste del cliente, que está en constante aumento, para su producto de cuenta de ahorro. Para esto, se comunica con el departamento de Advanced Analytics para ayudarlos a retener a sus mejores clientes al predecir cuál de ellos tiene

una propensión a cerrar su cuenta para que puedan conectarse con ellos y ofrecer ofertas atractivas para continuar su relación invaluable.

El equipo de Advanced Analytics comienza primero reduciendo el problema y definiendo la fuga, por ejemplo, para clientes que sacan su dinero, cierran su cuenta de ahorros y no reinvierten su dinero con otro producto ofrecido por el mismo banco. El negocio quiere predecir el desgaste con 3 meses de anticipación para obtener el tiempo suficiente para diseñar una estrategia de retención.

Luego, el equipo continúa recopilando los datos requeridos, que generalmente ocupan la mayor parte del tiempo en el proyecto, ya que preguntas como qué datos recopilar, durante cuánto tiempo, qué datos adicionales se necesitan, etc. son requisitos muy críticos. Una vez que el escenario está listo, los científicos de datos realizan su acto final: el aprendizaje automático. Los algoritmos leerán los datos y descubrirán patrones que conducen a un comportamiento de deserción basado en cierres de cuentas anteriores. Luego aprovechará este aprendizaje sobre la nueva información disponible del cliente en el escenario actual y predice la probabilidad de fuga.

Los algoritmos de aprendizaje automático como Random Forest o Gradient Boosting también funcionan bien aquí. Estos algoritmos son eficientes en el manejo de grandes cantidades de datos y pueden identificar patrones con buena precisión. Los científicos de datos también usan una técnica llamada ensamble de modelos donde diferentes algoritmos se combinan y se introducen en otro modelo para calcular la probabilidad como una combinación de varios algoritmos de aprendizaje automático.

Las instituciones bancarias y financieras han estado produciendo datos para generar información precisa sobre lo que está buscando exactamente el cliente, y muestran de manera eficiente los productos que les interesaría.

4.2.3 Propuesta de alternativa de solución a la situación problemática

a) Fase de compresión del negocio

En la gerencia de Tarjetas de la entidad financiera, existía la necesidad de incrementar el número de colocaciones de un subproducto de tarjetas de crédito llamado Extra Línea, que es una línea de crédito paralela aprobada en la tarjeta de crédito con la cual se puede disponer de efectivo a tasas y plazos preferenciales, sin alterar la línea de crédito principal de la tarjeta. Es decir, se podrá seguir realizando consumos con la tarjeta de

crédito con total normalidad, inclusive si ha dispuesto de forma parcial o total de la extra línea paralela ofertada.

Este incremento es importante para elevar la participación de mercado de la entidad Financiera debido a que este es el principal subproducto dentro del producto de tarjetas, pues es el único que te permite dar un tipo de préstamo para tarjeta habientes. La manera de lograr este incremento de colocaciones es por medio de un aumento en la efectividad de las ventas en campañas comerciales de la entidad financiera. Es por ello la importancia para el negocio de tener una herramienta que permita identificar a los clientes con tarjeta de crédito más propensos a responder positivamente a esta oferta (adquisición de una Extra Línea). Logrando mejorar el control de los costos de adquisición, ya que se tiene recursos limitados de dinero y tiempo para esta campaña, pero a la vez se tiene una población de posibles prospectos amplia equivalente a 400 mil clientes, de los cuales se debe seleccionar 90 mil aproximadamente para que puedan ser gestionados en campañas por los distintos canales de venta (Call center, Telemarketing, etc.) y a los cuales se les puede ofrecer este subproducto. Presentándose en esta base mensualmente una efectividad del producto de 5%, que equivale a 4,500 clientes que adquieren un Extra Línea mensualmente.

En este sentido el identificar a los prospectos con mayor propensión distribuidos en las campañas comerciales permite direccionar la capacidad operacional de las fuerzas de ventas y en consecuencia mejorar la rentabilidad del negocio de tarjetas. Reduciendo costos de adquisición y aumentando las posibilidades de ventas al enfocarse en los clientes más efectivos, cumpliendo así con los objetivos de ventas, pero maximizando los beneficios e incrementando la participación de tarjeta de créditos en el mercado financiero.

Para el desarrollo de este modelo de propensión se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que permitió inicialmente tener un conocimiento del negocio correcto, por lo que los objetivos del modelo analítico estuvieron muy alineados al objetivo del negocio que es incrementar el número de colocaciones de un subproducto de tarjetas de crédito llamado Extra Línea.

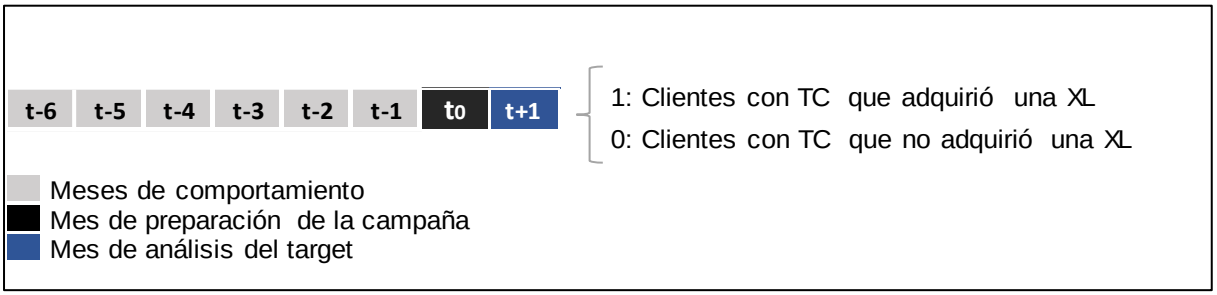
b) Fase de comprensión de los datos

Universo de Análisis:

Para el desarrollo de este caso aplicativo se usó como universo de análisis a los clientes con tarjeta de crédito presentes en las campañas comerciales, desde el mes de febrero del 2016 hasta diciembre del 2016.

En base a esa información se construyó 4 cosechas consecutivas de las campañas comerciales del subproducto de Extra Línea en una entidad financiera. Definiendo como una consecha a una ventana de análisis de periodos, donde para un periodo de tiempo donde se prepara la campaña (t_0) se analizan 6 meses de compartamiento del cliente (t_{-1},t_{-2},…t_{-6}) y un mes hacia adelante (t_{+1}) que es el mes donde se desplegó la campaña y se evalua el target.

Gráfico 3: Vista del diseño de una cosecha



Fuente: Elaboración propia

Se consideró como unidad de estudio a un cliente que cuenta con una Tarjeta de Crédito en la entidad financiera y que haya sido un contacto efectivo en la campaña.

La distribución y proporción del target en los datos es la siguiente.

Tabla 2: Tasa de adquisición de Extra Línea en la data de análisis

Target	Clientes(#)	Clientes (%)
Adquisición de una Extra Línea	19,529	5.42%
No adquisición de una Extra Línea	341,114	94.58%
Total	**360,643**	**100.0%**

XL: Extra Línea

Fuente: Elaboración propia

La variable target tiene como ratio de éxito un 5.42% de clientes con adquisición del producto frente a un 94.58% de no adquisición del producto, existen varias teorías sobre el desbalanceo, es decir, la frecuencia muy baja de una categoría de la variable dependiente y los posibles problemas que enfrentemos antes esta situación. Por lo que en este caso aplicativo se utilizó la técnica de balanceo de clases SMOTE, ya que se tenía una baja proporción de la variable respuesta.

Descripción de los datos: Se analizó la información sociodemográfica, del sistema financiero e información interna en la entidad financiera, ya que fueron las variables disponibles, y que fueron entregadas por la empresa para mostrar el caso aplicativo.

Tabla 3: Diccionario de variables

Tipo	Variables Predictoras	Tipo Variable
Ent.	SegmentoRFMTC	Cualitativas
Ent.	Antigüedad de meses de la TC	Cuantitativas
Ent.	Línea de crédito en los 6 Últ.meses	Cuantitativas
Ent.	Tipo de Tarjeta en el Últ.meses	Cualitativas
SSFF	Saldo Bancos grandes(BCP, SBP, IBK y BBVA)/ Saldo Total SSFF	Cuantitativas
SSFF	Saldo de Prestamo Personal en el SSFF en el penúltimo trimestre	Cuantitativas
SSFF	Saldo de Prestamo Personal en el SSFF en el Últ.trimestre	Cuantitativas
SSFF	Nro. de prod.financieros en los 6 Últ.meses	Cuantitativas
SSFF	Apalancamiento=Deuda/Ing. en el Últ. Trimestre	Cuantitativas
SSFF	Apalancamiento=Deuda/Ing. en los 6 Últ.meses	Cuantitativas
SSFF	Género	Cualitativas
SSFF	Línea de crédito en TC en el SSFF en el Últ. Trimestre	Cuantitativas
SSFF	Saldo de Tarjeta de Crédito en el SSFF en el penúltimo trimestre	Cuantitativas
SSFF	Saldo de Tarjeta de Crédito en el SSFF en el Últ.trimestre	Cuantitativas
SSFF	Saldo de Tarjeta de Crédito en el SSFF Últ. Mes	Cuantitativas
SSFF	Saldo de Prestamos Personales en Bancos Grandes en el SSFF Últ. Mes	Cuantitativas
SSFF	Saldo de Tarjeta de Crédito en Bancos Grandes en el SSFF Últ. Mes	Cuantitativas
SSFF	Saturación=Deuda/ LíneaTC en el Últ. Trimestre	Cuantitativas
SSFF	Ratio Ingreso/DeudaTC en el Últ. Trimestre	Cuantitativas
SSFF	Ratio Saldo de Prestamo Personal entre los 2 Últ.trimestres	Cuantitativas
SSFF	Nro. de bancos en el Últ. Trimestre	Cuantitativas
SSFF	Nro. de prod.financieros en el Últ.mese	Cuantitativas
SSFF	Apalancamiento=Deuda/Ing. en el Últ.meses	Cuantitativas
Sociod.	Ingreso de la Persona	Cualitativas
Sociod.	Edad de la persona	Cualitativas
Sociod.	SituacionLaboral	Cualitativas
Sociod.	Macrozona (agrupación departamentos y distritos)	Cualitativas

Tipo	Variable a Predecir	Tipo Variable
Banco	1: Adquisición de una Extra Línea	Cualitativas

Sociod : Sociodemográficos
SSFF: Sistema Financiero
Ent: Entidad Financiera

Fuente: Elaboración propia

De las 27 variables predictoras, 21 variables son de tipo numéricas y 6 variables son de tipo categóricas, por lo que se procedió a realizar la exploración diferenciada por el tipo de variable.

Exploración de los datos:

En el siguiente cuadro se observa las estadísticas descriptivas para las variables cuantitativas del universo de análisis. Las variables cuantitativas no presentan valores missing por lo que no es necesario realizar ninguna imputación.

Los valores máximos difieren bastante de los valores del percentil 75 de las variables, por lo que se realizó un análisis de Outliers para retirar a esos valores atípicos.

También se observó que la asimetría es grande para tres variables, por lo cual, se deduce que tiene un sesgo elevado, en estos casos se probó la transformación para las variables de ratio de saldo de Prestamo Personal (PersonalLoanTotalBalance_T2_T1), el ratio de ingreso respecto al saldo en su Tarjeta de Crédito (Rat_Ing_DeudaTC_T1) y el ratio de saldo de su Tarjeta de Crédito respecto a su línea (Rat_SaturacionTC_T1) .

Tabla 4: Resumen de variables cuantitativas

Variable	Q1	Median	Q3	Mean	Min	Max	NAs	SD	Asim
AntMeses	5	14	33	28	0	261	0	39	2.7
Rat_SOWBG_T1	0	1	1	1	0	1	0	0	-0.1
PersonalLoanTotalBalance_T2	0	0	116	970	0	310,238	0	3,425	12.8
PersonalLoanTotalBalance_T1	0	0	80	972	0	208,844	0	3,323	8.7
NRO_PROD_FIN_6UM	1	1	2	1	0	4	0	1	1.1
Mean_Rat_Apalancamiento_T1	0	1	3	2	0	126	0	3	7.1
Mean_Rat_Apalancamiento_6UM	0	1.1	2.9	2.0	0.0	127.6	0.0	2.7	6.9
CreditCardTotalCreditLimit_T1	2,530	6,203.8	14,180.7	10,521.1	0.0	133,052.1	0.0	11,892.2	2.2
CreditCardTotalBalance_T2	78	503	1,853	1,901	0	125,462	0	3,981	5.2
CreditCardTotalBalance_T1	89	518	1,831	1,877	0	104,495	0	3,875	5.0
CreditCardTotalBalance_1	57	483	1,792	1,856	0	109,276	0	3,895	5.0
BGPersonalLoanBalance_1	0	0	0	753	0	198,390	0	3,234	9.0
BGCreditCardBalance_1	0	105.8	1,039.8	1,370.8	0	82,825	0	3,541	5.6
Rat_SaturacionTC_T1	0	0.1	0.2	0.2	0.0	3,903.8	0.0	5.9	663.9
Rat_Ing_DeudaTC_T1	0	0.8	2.2	2.0	0.0	2,007.9	0.0	6.2	121.9
PersonalLoanTotalBalance_T2_T1	0	0.0	0.0	5.5	0.0	1,222,242	0.0	2,056.4	519.5
NroBanks_01_T1	2	4.0	5.0	3.7	0.0	12.0	0.0	1.8	0.4
NRO_PROD_FIN_1	1	1.0	2.0	1.4	0.0	4.0	0.0	0.5	1.1
LCreditoUSD_6UM	900	2,216.7	6,300.0	4,948.5	0.0	72,390.0	0.0	6,511.7	2.4
Rat_Apalancamiento_1	0	1.0	2.9	2.0	0.0	127.2	0.0	2.8	7.0

Fuente: Elaboración propia

Se realizó un análisis descriptivo univariado y bivariado para las variables cualitativas. A continuación se indican algunos primeros hallazgos importantes:

Del análisis bivariado de la variable Segmento RFM y el target, se observó que los clientes pertenecientes a la categoría 0 (clientes con menos de tres meses de antiguedad), tienen una efectividad que es casi el doble de las efectividades en las otras categorías.

Del análisis bivariado de la variable macrozona y el target, se observó que los clientes pertenecientes a la Lima Norte y Provincia tienen una efectividad mayor respecto a las efectividades en las otras categorías.

Del análisis bivariado de la variable rango ingreso y el target, se observó que los clientes con un ingreso menor a S/1,000 tienen una efectividad mayor respecto a las efectividades en las otras categorías.

Del análisis bivariado de la variable rango edad y el target, se observó que los clientes con un edad menor a 35 años tienen una efectividad mayor respecto a las efectividades en las otras categorías.

Adicionalmente se creó una variable FlagM1 que me indica si el registro estuvo en la capaña anterior, que también fue incluida para los análisis posteriores.

c) Fase de preparación de los datos

Limpieza de Datos:

Luego de revisar el cuadro resumen de univariados para las variables cuantitativas donde se observó que no teníamos variables con missing o sin variación en sus registros(sd=0), se procedió a identificar los outliers para las variables cuantitativas. Se utilizó el paquete outliers del software R para identificar los outliers para las 21 variables numéricas. Este paquete identifica a un valor atípico considerando el valor con la mayor diferencia respecto a la media y lo define como umbral para identificar valores extremos. Luego de la identificación de estos valores se procedió a retirarlos de la base de análisis, por lo que el universo de clientes disminuyó en 199 registros, lo que representó un 0.06 % del total de registros.

También se procedió a verificar la normalización de las variables, ya que dependiendo de ello se eligió la técnica más óptima para el análisis de correlación de variables.

Para las variables cuantitativas después de relizar las pruebas de normalidad de Anderson Darling y Kolmogorov Smirnov a un nivel de significancia de 0.05 como criterio de decisión se observó que existe evidencia estadística para rechazar Ho. Es decir, podemos afirmar que la variable no tiene una distribución normal. Esto se cumple para todas las variables cuantitativas. (Ver Anexo 1)

Selección de variables:

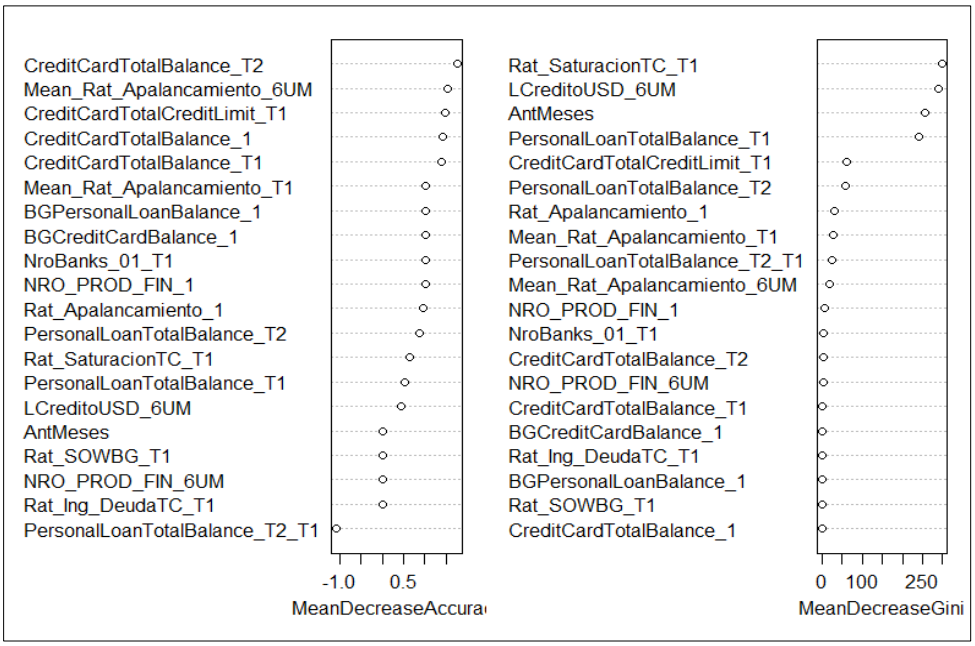
Para las variables cualitativas se realizó el test de bondad de ajuste de Chi-cuadrado con respecto a Y (adquisición de Extra Línea), para ver el tema de la dependencia. A un nivel de significancia de 0.05 como criterio de decisión se observó que existe evidencia estadística para rechazar Ho. Es decir, podemos afirmar que la variable categórica (predictora) y la variable Y (target) son dependientes. Esto se cumple para todas las variables cualitativas. (Ver Anexo 1)

En el caso de las variables cuantitativas debido a que no satisfacen la condición de normalidad requerida por el coeficiente de Pearson se usó el coeficiente de correlación de Spearman, que es una medida de correlación alternativa para estos casos.

En la matriz de correlación resultante se observó un gran número de variables altamente correlacionadas (>0.7), así que para decidir con que variable nos quedamos para cada par de variables correlacionadas, también se analizó la importancia de variables, de modo que ingresó al modelo las variable que tenía menos correlación con las demás variables y mayor peso en la importancia de variables respecto a la variable dependiente.

Se analizó la importancia de variables usando el algoritmo Random Forest, que determina la importancia de las variables según el indicador mean decrease accuracy y mean decrease gini como se observa en la siguiente figura.

Gráfico 4: Gráfica de la importancia de variables



Fuente: Elaboración propia

Para eliminar a las variables correlacionadas, se consideró el grado de correlación y la importancia de la variable :

Tabla 5: Criterio para la selección de dos variables correlacionadas

Variable 1	Variable 2	Coeficiente de Correlación
PersonalLoanTotalBalance_T2_T1	PersonalLoanTotalBalance_T2	0.92
PersonalLoanTotalBalance_T2_T1	PersonalLoanTotalBalance_T1	0.88

Fuente: Elaboración propia

En este caso se seleccionó a la variable PersonalLoanTotalBalance_T1, ya que presentó menor correlación respecto a las demás y tuvo mayor peso en la importancia de variables.

Al finalizar la etapa de preparación de datos, las variables seleccionadas fueron las siguientes:

Tabla 6: Lista de variables seleccionadas

Variables Predictoras	Tipo Variable
Antigüedad de meses de la TC	Cuantitativas
Tipo de Tarjeta en el Últ.meses	Cualitativas
Flag si estuvo en campaña en el mes anterior	Cualitativas
Género	Cualitativas
Línea de crédito en los 6 Últ.meses	Cuantitativas
Macrozona (agrupación departamentos y distritos)	Cualitativas
Apalancamiento=Deuda/Ing. en el Últ.meses	Cuantitativas
Nro. de prod.financieros en el Últ.mese	Cuantitativas
Nro. de bancos en el Últ. Trimestre	Cuantitativas
Logaritmo del Saldo de Prestamo Personal en el SSFF en el Últ.trimestre	Cuantitativas
Logaritmo de la Saturación(Deuda/ LíneaTC en el Últ. Trimestre)	Cuantitativas
SegmentoRFMTC	Cualitativas

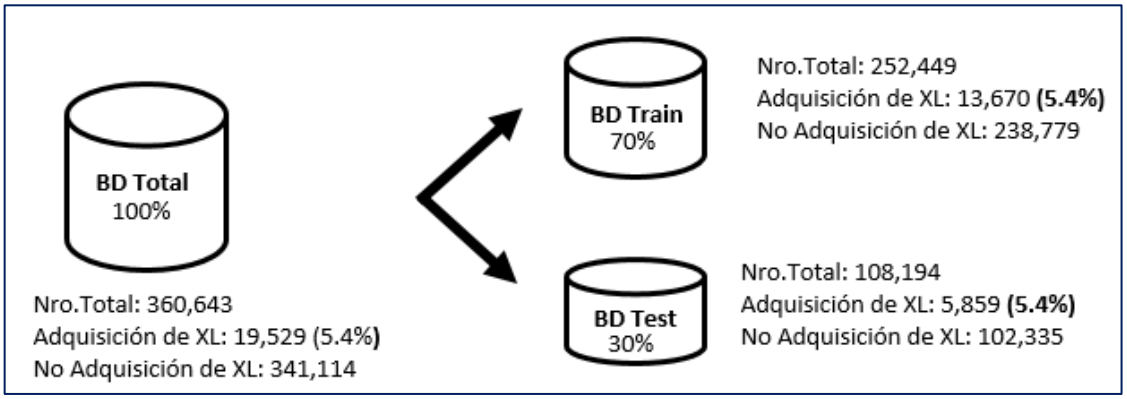
Fuente: Elaboración propia

d) Fase de modelado

División de la base de datos

Antes de la construcción de los modelos se seleccionó una muestra de train y test, de 70% y 30% respectivamente de la data obtenida con la preparación de datos. Donde la base de train se usó para el desarrollo del modelo y la base de test para probar el desempeño que tendría el modelo en una base nueva.

Gráfico 5: División de la base de datos



Fuente: Elaboración propia

Técnicas de Balanceo de Clases

Se observó un desbalanceo en las clases de la variable target de los datos de aproximadamente 95% de no adquisición frente a un 5% de adquisición del producto de Extra Línea, por lo que fue necesario aplicar una técnica de balanceo de clases para mejorar la clasificación de la clase minoritaria a predecir. Es por ello que en el desarrollo del presente trabajo se realizó un modelo de Random Forest con los datos originales y otro modelo de Random Forest con los datos balanceados mediante la aplicación del algoritmo de balanceo de clase SMOTE.

La técnica de sobre-muestreo se realizó con la muestra de train y se probó sobre la muestra de test. En este trabajó se probó distintos % para el balanceo de clases, siendo el de 50% y 50% el que nos permitió obtener los mejores indicadores.

Tabla 7: Datos Balanceados mediante SMOTE

Target	Clientes(#)	Clientes (%)
Adquisición de una Extra Línea	27,340	50%
No adquisición de una Extra Línea	27,340	50%
Total	**54,680**	**100.0%**

XL: Extra Línea

Fuente: Elaboración propia

Calibración de Hiperparámetros

Para poder elegir el mejor modelo se realizó una calibración de los hiperparámetros del modelo de Random Forest donde se probó el modelo con distintos valores de los hiperparámetros en números de árboles, número de variables y tamaño del nodo, de

donde se seleccionó para cada caso los hiperpáramteros que me permitan obtener el menor OOB. (Ver Anexo 1)

Los parámetros elgidos fueron los siguientes:

Número de árboles: 100

Número de variables: 6

Tamaño del Nodo:4

e) Fase de evaluación

Resultado de los criterios de Evaluación

Para analizar los resultados del modelo se analizaron los indicadores: Tasa de Correcta Clasificación (TCC), Sensibilidad, Especificidad y el área bajo la curva ROC (AUC).

Tabla 8: Comparación de Modelos

Modelo	TCC	Sensibilidad	Especificidad	AUC
Data Sin Balanceo	94.5%	0.9%	99.9%	70.2%
Data con Balanceo	67.4%	64.1%	67.6%	71.4%

Fuente: Elaboración propia

Si bien es cierto que la Tasa de Correcta Clasificación (TCC) es más alta en la data sin balancear con un 94.5%, no se podría considerar este modelo, puesto que tiene una sensibilidad muy baja, es decir la proporción de predicciones positivas sobre el total de casos positivos reales de la base fue de tan sólo 0.9%. Sin embargo como se puede observar, el modelo de Random Forest con SMOTE, obtiene un mejor índice de 64.1% de sensibilidad, es decir predice mejor la categoría de adquisición de la Extra Línea, respecto al modelo de Random Forest sin balanceo.

Además, también obtiene mejor índice de área bajo la curva (AUC). Por ende se puede concluir que el modelo de Random Forest con SMOTE posee mejor desempeño como clasificador.

f) Fase de Implementación

Finalizado el desarrollo del modelo de propensión a la adquisición de una ExtraLínea, se coordinó una reunión con el área de campañas y productos de la entidad financiera, donde se presentó el modelo y se coordinaron puntos importantes para la fase de implementación:

Uso del Modelo

Para las campañas comerciales de Extra Línea, el modelo empezó a ser aplicado mensualmente, ya que permitía actualizar la probabilidad de propensión en base a los cambios de comportamiento del cliente en el sistema financiero .

Herramienta de Campaña

El modelo fue usado como herramienta vital en el momento de priorización de registros para la campañas (selección de registros de acuerdo a la capacidad de fuerzas de ventas), en la estrategia de gestión de registros (mayor intensidad de llamadas a los más propensos), en la asignación de registros por canales (los más propensos a canales proactivos) y estrategia de pricing (a mayor propensión se le podría dar una mayor tasa). El modelo fue presentado a los gerentes de Agencias y fuerzas de ventas, donde se les mostró a través de indicadores del negocio la potencialidad de tener un modelo de propensión a la toma del producto de Extra Línea, y motivo a que ellos puedan compartir esa información con los asesores de venta del producto para que generarles una mayor motivación al saber que hay un modelo estadístico que ayuda a priorizar los registros que se les envía y por tanto que les ayuda en el momento de realizar sus ventas.

4.3. Contribución en la solución de situaciones problemáticas

El tener un modelo de propensión para el producto de Extra Línea permitió brindar un apoyo estratégico, al área de campañas comerciales, basado en la analítica de los datos. De modo que se logró identificar entre el universo de clientes con tarjeta de crédito a los clientes con mayor potencial (probabilidad) de adquisición del producto. Así también ayudó a conocer el perfil de estos clientes propensos por medio de las variables más importantes del modelo, siendo esto muy importante cuando se quiera agregar o considerar algún cambio en los criterios comerciales para alguna estrategia diferenciada

a los clientes en base a su perfil. El tener esta identificación de clientes potenciales ayudó al equipo de campañas a realizar una distribución más eficiente de sus registros al momento de seleccionar a los registros mensuales que deben ser enviados en las campañas comerciales para que puedan ser gestionados por las fuerzas de ventas (call center), como también a tener una gestión de registros más direccionada en clientes potenciales.

Tabla 9 : Evaluación del Modelo en un mes de campaña

Decil	Efectividad	%Venta	Lift
10	15.8%	31.2%	3
9	9.2%	17.9%	1.7
8	7.0%	13.4%	1.3
7	5.1%	9.8%	1
6	4.1%	7.9%	0.8
5	3.6%	6.7%	0.7
4	2.6%	4.7%	0.5
3	1.9%	3.5%	0.4
2	1.7%	3.0%	0.3
1	1.0%	1.9%	0.2
Total	**5.3%**	**100.0%**	**1.0**

Fuente: Elaboración propia

Gráfico 6: Ventas acumuladas por deciles de propensión

Fuente: Elaboración propia

En la tabla 9 se puede observar la evaluación del modelo en un mes de campaña, donde se puede visualizar y validar que la efectividad de ventas, esta ordenada de acuerdo a los deciles de propensión del modelo, siendo 10 el de mayor propensión y 1 el de menor propensión. Apartir de esta tabla se construye el gráfico 6 donde se observa que con el 40% de los clientes con mayor probabilidad se obtiene el 72% de las ventas.

En este sentido, el tener identificado en la campaña un menor número registros pero con mayor propensión de la adquisición del subproducto ofrecido, ayudó a cubrir la capacidad operacional en las fuerzas de ventas y por lo tanto mejorar la rentabilidad del negocio de tarjetas (reduciendo costos de adquisición y aumentando posibilidades de ventas al enfocarse en los clientes más efectivos), cumpliendo así con los objetivos de ventas y maximizando los beneficios de la entidad finanicera.

4.4. Análisis de la contribución en términos de competencia y habilidades

La realización de este proyecto contribuyó tanto en el desarrollo de habilidades blandas y técnicas. Respecto a las habilidades técnicas el desarrollar todas las fases necesarias para la elaboración del modelo de propensión permitieron profundizar en el conocimiento en métodos de correlación para variables no paramétricas, criterios de selección de variables, métodos de balanceo, tunning de hiperpárametros para el algortimo de Random Forest, así como tener en consideración distintos criterios de evaluación para elegir al mejor modelo.

Respecto a las habilidades blandas, el desarrollo del presente trabajo permitió potenciarlas, ya que durante la fase del conocimiento del negocio se tuvieron distintas reuniones con las áreas de producto y campañas para entender el flujo comercial de este producto como algunas consideraciones en la elaboración del modelo (por ejemplo, debido al periodo de elaboración de las campañas las variables a ser usadas en el modelo deben construirse con un periodo de desface de dos meses hacia atrás). Estas reuniones permitieron también mejorar la habilidad de comunicación con colaboradores de distintos niveles jerárquicos, y con distintas habilidades (entre técnicas y de negocio).

4.5. Nivel de beneficio obtenido por el centro laboral

El universo de registros potenciales en base a criterios de riesgo crediticios que son enviados al área de campañas comerciales para que pueda realizarse la elaboración de de base de datos para las campañas del producto de Extra Línea, tiene registros limitados y con un porcentaje de repetidos entre meses continuos.

Por lo tanto la eficiencia es clave al momento de seleccionar esos registros que deben de ser enviados a las fuerzas de ventas, seleccionando así a los registros con mayor posibilidad de adquisción del producto.

El incluir el modelo de propensión como un input clave en la matriz de efectividad considerada en el área de campañas que tiene como variables principales la contactabilidad y frescura del registro en la campañas fue muy importante, ya que permitió afinar los criterios de selección de registros, es decir priorizar a los registros con mayor probabilidad de adquisición del producto de Extra Línea. Permitiendo que la efectividad de ventas pueda incrementar en promedio de %5.28 a %5.59 en los tres primeros meses de uso en las campañas comerciales.

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones:

- Usando el algoritmo Random Forest con datos balanceados con SMOTE se identifacron a los cliente más propensos a la adquisición del producto Extra Línea con una sensibilidad de 64.1%, especificidad de 67.9% y tasa de correcta clasificación del 67.4%.

- Se determinó que las principales variables que influyen en la adquisición de una Extra Línea, para clientes con tarjeta de crédito de una entidad financiera son principalmente: ratio de saturación(deuda/línea) de la tarjeta de crédito en la entidad financiera, flag de presencia en la campanas en el último mes, saldo el préstamo personal en el sistema financiero en el último trimestre , antigüedad del cliente en la entidad financiera, línea de crédito de la tarjeta de crédito en la entidad financiera, ratio de apalancamiento(dedua/ingreso) y número de productos en el sistema financiero en ese orden respectivamente.

- Se identificó que el perfil de clientes propensos a la Adquisición una Extra Línea:
  - Con deudas de consumo tarjetas y préstamos es 2 veces respecto a su Ingreso
  - Con una línea de crédito menor a $1,200 en los 6 últimos meses
  - Con deudas que representan la mitad de la línea
  - Con préstamo personal mayor a $425 en el últimos trimestre

- Según los indicadores de clasificacion mostrados; clasificación global, curva ROC sensibilidad y especificidad. El modelo de Random Forest aplicando la técnica de

balanceo: algoritmo de SMOTE, obtiene mejores resultados en la clasificacion de adquisición del producto de clientes con tarjeta de crédito en una entidad financiera que el modelo obtenido sin realizar el balanceo de clases .

5.2. Recomendaciones:

- Para el desarrollo del modelo se recomienda que adicionalmente se pruebe con otros algoritmos de balanceo y otros algoritmos de clasificación, como: Gradient Boosting Machine, Support Vector Machine, Redes Neuronales, entre otras. De modo que se pueda comparar y elegir a la de mayor eficiencia.

- En el presente trabajo se consideró la curva ROC como uno de los indicadores para seleccionar al modelo, sin embargo para futuras investigaciones podría considerarse también la curva de precisión de recuperación, ya que según la literatura las curvas de recuperación de precisión son apropiadas para conjuntos de datos con desbalanceo de clases.

- Para la implementación del modelo, se recomendaría usarlo para algunas otras estrategias comerciales. Por ejemplo: uso de los canales más costos como el call center sólo para los de alta probabilidad y uso de canales más baratos como los canales digitales para los de menor probabilidad.

6. REFERENCIAS BIBLIOGRÁFICAS

BOULEXTEIX A & STROBL C & HOTHORN, Zeileis (2007). Bias in Random Forest.Variable Importance Measures.

BREIMAN, L. (2001). Random forests.Mach Learn. Statistics Department, University of California.

CHAWLA, Nitesh; BOWYER, Kevin; HALL, Lawrence; and KEGELMEYER (2002) Philip. SMOTE: Synthetic Minority Over-sampling Technique.

CHIRAG, Soni (2019). Role of Machine Learning in redefining Retail Banking

GÉRON, Aurélien (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2da edicion

HANLEY J.A., MCNEIL B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve.

HASTIE T, TIBSHIRANI & R, FRIEDMAN, J (2008). Data Mining, Inference and Prediction. Spriger, 2 da edición.

HASTIE, Trevor & TIBSHIRANI, Robert & FRIEDMAN, Jerome (2017). The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2da edición

KELLEHER, John D & MAC NAMEE Brian (2015). Fundamentals of Machine Learning for Predictive Data Analytics.

KUHN, Max & JOHNSON, Kjell (2016). Applied Predictive Modeling.5ta edicion.

KRUPPA, SCHWARZ, ARMINGER & ZIEGLER (2013). Consumer credit risk: Individual probability estimates using machine learning.

POWERS, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

ROKACH , Lior & MAIMON, Oded (2015). Data Mining with Decision Trees. 2da edición.

SHENG, Ling Kock & WAH, Teh Ying. A comparative study of data mining techniques in predicting consumers' credit card risk in banks (2011)

THEOBALD, Oliver (2017). Machine Learning for absolute beginners. 2da edicion

VILLENA, Julio (2016) CRISP-DM: La metodología para poner orden en los proyectos

WAH, Yap Bee & IBRAHIM, Irma Rohaiza (2010). Using Data Mining Predictive Models to Classify Credit Card Applicants.

WILLIAMS, G (2011). Data Mining with Rattle and R.

WITTEN Ian H., EIBE Frank (2011). Data Mining: Practical Machine Learning Tools and Techniques. 3era edición

7. ANEXOS

Anexo 1: Códigos de Procesamiento en R

Prueba de Normalidad

H_0: la variable sigue una distribución normal

H_1: la variable no sigue una distribución normal

```
> ad.test(Datos$AntMeses)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$Rat_SOWBG_T1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$PersonalLoanTotalBalance_T2)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$PersonalLoanTotalBalance_T1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$NRO_PROD_FIN_6UM)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$Mean_Rat_Apalancamiento_T1)[2]
$p.value
[1] 3.7e-24
```

```
> ad.test(Datos$Rat_SaturacionTC_T1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$Rat_Ing_DeudaTC_T1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$PersonalLoanTotalBalance_T2_T1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$NroBanks_01_T1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$NRO_PROD_FIN_1)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$LCreditoUSD_6UM)[2]
$p.value
[1] 3.7e-24

> ad.test(Datos$Rat_Apalancamiento_1)[2]
$p.value
[1] 3.7e-24
```

Pruebas de Independencia de variables

H0: los dos factores son independientes

H1: los dos factores son dependentes

```
> chisq.test(Datos$SegmentoRFMTC,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$SegmentoRFMTC and Datos$FlagVenta
X-squared = 3708.5, df = 4, p-value < 2.2e-16

> chisq.test(Datos$CreditCardType_1,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$CreditCardType_1 and Datos$FlagVenta
X-squared = 3411.7, df = 8, p-value < 2.2e-16

> chisq.test(Datos$Genero,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$Genero and Datos$FlagVenta
X-squared = 1000.1, df = 2, p-value < 2.2e-16
```

```
> chisq.test(Datos$SituacionLaboral,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$SituacionLaboral and Datos$FlagVenta
X-squared = 466.55, df = 3, p-value < 2.2e-16

> chisq.test(Datos$Macrozona,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$Macrozona and Datos$FlagVenta
X-squared = 2066.8, df = 6, p-value < 2.2e-16

> chisq.test(Datos$RangoIngreso,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$RangoIngreso and Datos$FlagVenta
X-squared = 643.83, df = 7, p-value < 2.2e-16

> chisq.test(Datos$RangoEdad,Datos$FlagVenta)

        Pearson's Chi-squared test

data:  Datos$RangoEdad and Datos$FlagVenta
X-squared = 100.01, df = 8, p-value < 2.2e-16
```

Calibración de Hiperpárametros

```
> hiperparametro_mtry <-  tuning_rf_mtry(df = train, y = "FlagVenta")
> hiperparametro_mtry %>% arrange(oob_err_rate)
# A tibble: 11 x 2
   n_predictores oob_err_rate
           <int>        <dbl>
 1             5        0.252
 2             6        0.254
 3             4        0.254
 4            10        0.254
 5             3        0.254
 6             7        0.254
 7             9        0.255
 8             8        0.256
 9            11        0.256
10             2        0.263
11             1        0.304
```

```
> hiperparametro_nodesize %>% arrange(oob_err_rate)
# A tibble: 20 x 2
    size oob_err_rate
   <int>        <dbl>
 1     1        0.252
 2     3        0.253
 3     2        0.254
 4     4        0.254
 5     5        0.255
 6     6        0.256
 7     7        0.256
 8     8        0.258
 9    11        0.259
10     9        0.259
11    12        0.259
12    10        0.260
13    13        0.260
14    15        0.261
15    14        0.262
16    16        0.262
17    17        0.262
18    18        0.263
19    19        0.264
20    20        0.264
>
```