PROYECTO DE ANALITICA

Julio-24

Rivero (277134)
Pérez (172866)
Maceiras (315647)

Contenido

Resumen ejecutivo

El objetivo del proyecto es realizar un modelo de Machine Learning para predecir incumplimiento crediticio en American Express utilizando datos financieros históricos.

El tamaño del dataset original presentó desafíos significativos en términos de capacidad de procesamiento y manejo de memoria. Para abordar estas limitaciones, se convirtió el formato .ftr conocido por su eficiencia en la gestión de grandes volúmenes de datos. Esta conversión redujo significativamente el tamaño del archivo a 1.73 GB, lo cual facilitó la carga y manipulación de datos. Adicionalmente para reducir aún más la complejidad del dataset, inicialmente se redujo de 189 variables a 10 mediante la técnica de modelado lineal con Recursive Feature Elimination (RFE) con validación cruzada. De este proceso se obtuvieron las variables más importantes para el análisis posterior.

Las variables se encuentran anonimizadas y normalizadas, entre las categorías disponibles (Morosidad, Gasto, Pago, Balance y Riesgo) se trabaja con las 10 variables clave de las categorías Morosidad (D), Gasto (S) y Balance (B) que se consideraron relevantes para el modelo predictivo basado en su impacto potencial en el incumplimiento crediticio.

Se realizaron resúmenes de los datos, análisis de datos faltantes, outliers y análisis univariados, dada la falta de información sobre las variables no se realiza tratamiento de datos, sin embargo, debido a la gran disponibilidad de datos se eliminan de la base clientes con datos faltantes en cualquiera de las variables.

Se balancean los datos antes de modelar. En primera instancia se entrenó un árbol de decisión con limitaciones de profundidad para la interpretación y se podó para optimizar su rendimiento. Además, se implementó un modelo de XGBoost y se optimizaron los hiperparámetros para mejorar la precisión y la generalización del modelo. Se evaluaron varios modelos según métricas clave como sensibilidad, especificidad, precisión, AUC, F1-score y Gini para determinar el modelo óptimo.

El modelo XGBoost optimizado (pred_boost_opt) mostró el mejor rendimiento en términos de métricas evaluadas, como AUC más alto, mayor precisión, y mejor capacidad para capturar incumplimientos al 4%. Se seleccionó pred_boost_opt como el modelo más adecuado para la predicción de incumplimientos debido a su robustez y estabilidad en los conjuntos de entrenamiento y prueba. Adicionalmente se evalúa la importancia de cada variable para darle interpretabilidad al modelo.

La implementación del modelo permite a American Express mejorar la agilidad en la aprobación de créditos, reducir errores, optimizar la asignación de recursos y personal, y personalizar ofertas para diferentes segmentos de clientes. El modelo también proporciona oportunidades para la innovación y el desarrollo de nuevos productos financieros basados en predicciones precisas de incumplimiento.

1 Entendimiento del Caso de Negocio y Planificación del Trabajo

1.1 Contexto, Problema, Proyecto

El objetivo del proyecto es realizar un modelo de Machine Learning para predecir el default en American Express.

El alcance del proyecto incluye todas las fases necesarias para desarrollar, evaluar, implementar y realizar seguimiento a un modelo de Machine Learning para predecir el default en American Express. Este alcance asegura que el proyecto no solo entrega un modelo preciso, sino también un sistema que puede ser mantenido y actualizado para asegurar su relevancia y precisión a lo largo del tiempo.

American Express (Sponsor) es una compañía global integral de pagos. Es el emisor de tarjetas más grande del mundo, provee a sus clientes la posibilidad de acceder a productos, servicios y experiencias que les aporten valor tanto a nivel personal como para el éxito en los negocios.

El default (incumplimiento crediticio) es la situación en la que el prestatario no cumple con sus obligaciones de pago según los términos acordados para un préstamo o deuda. Predecir el default es fundamental para American Express para poder mitigar el riesgo de incumplimiento.

Se considera para la evaluación la métrica M que se calcula como la media del Coeficiente Gini Normalizado (G) y la tasa de incumplimiento capturada al 4% (D). Esta métrica combina tanto la capacidad predictiva del modelo (Gini) como su capacidad para capturar incumplimientos en las predicciones más importantes (Sensibilidad/Recall).

$M=0.5\cdot(G+D)$

La métrica M es solicitada por el Sponsor, orientada a identificar correctamente el incumplimiento y sobre todo el segmento más crítico. [1]

1.1.1 Importancia del desafío en el sector y en American Express

Predecir el default es de suma importancia para el sector financiero tanto como para American Express. Si el default pudiera predecirse de una manera más precisa y fácil obtendrían los siguientes beneficios:

Para el sector empresarial en general:

- **Estabilidad financiera:** La falta de capacidad para prever incumplimientos puede llevar a una mayor volatilidad en los mercados financieros y a un aumento del riesgo sistémico.
- **Confianza en el sector:**
    - Regulaciones y cumplimiento: En algunos países, las instituciones financieras están sujetas a regulaciones que requieren que mantengan ciertos niveles de capital en reserva para cubrir posibles pérdidas por incumplimientos. La capacidad de predecir el default mitiga el riesgo de incumplir estas regulaciones y evitar sanciones regulatorias.
    - Gestión del riesgo: Una empresa que demuestra una capacidad sólida para gestionar el riesgo de incumplimiento puede atraer inversiones y mantener la confianza del mercado.

  Si American Express se ve afectado todo el sector financiero puede sufrir pérdida de confianza.

Para American Express:

- **Gestión del riesgo crediticio:** American Express, como emisor de tarjetas de crédito, depende en gran medida de su capacidad para evaluar y gestionar el riesgo crediticio de sus clientes. Las políticas de riesgo se ven directamente afectadas por la certeza en la predicción del default, tomando decisiones informadas.
- **Recursos financieros:** El default afecta la rentabilidad de American Express al influir en la cantidad de ingresos generados por las tarifas e intereses asociados con sus productos crediticios. Una mejor capacidad para predecir el default puede mejorar la rentabilidad de la empresa y su competitividad en el mercado. Se minimizan las pérdidas y se optimiza la asignación de recursos.
- **Experiencia del cliente:** Al facilitar la aprobación de tarjetas de crédito para clientes solventes y reducir el riesgo de incumplimiento, American Express puede ofrecer una experiencia más fluida y satisfactoria en la aprobación de tarjetas de crédito.

1.1.2 Impacto en la operativa

Operaciones actuales:

- **Toma de decisiones más ágil:** Al contar con modelos de Machine Learning precisos para predecir el default, American Express puede automatizar y agilizar el proceso de evaluación de crédito. Esto significa que las decisiones sobre la aprobación de tarjetas de crédito o la extensión de líneas de crédito pueden tomarse de manera más rápida y eficiente.
- **Reducción de la carga de trabajo:** Los procesos manuales de evaluación de crédito pueden consumir muchas horas de trabajo del personal y son propensos a errores humanos. Al automatizar parte de este proceso mediante modelos predictivos, American Express puede reducir la carga de trabajo de sus empleados y liberar recursos para otras tareas críticas.
- **Minimización del error:** Al utilizar un programa para la predicción del incumplimiento se minimiza el error. Además, el error del modelo es conocido por lo cual se pueden tomar decisiones considerándolo.
- **Mejora en la asignación de recursos:** Al comprender mejor el riesgo de incumplimiento de los clientes, American Express puede asignar de manera más eficiente sus recursos, esto incluye la optimización de líneas de crédito, la gestión de personal de atención al cliente y la implementación de estrategias de marketing dirigidas a segmentos de clientes que representen un menor riesgo.

Operaciones futuras:

- **Innovación y desarrollo de productos:** Con la capacidad de predecir el default con mayor precisión, American Express puede sentirse más seguro para explorar nuevas oportunidades de mercado y desarrollar productos financieros innovadores. Esto puede incluir la introducción de nuevos programas de recompensas, ofertas personalizadas para clientes de bajo riesgo y servicios adicionales que generen valor tanto para la empresa como para sus clientes.
- **Adaptación a cambios en el mercado:** Los modelos predictivos pueden ayudar a American Express a identificar tendencias emergentes en el comportamiento del cliente y en el riesgo de incumplimiento. Esto permite a la empresa adaptarse rápidamente a los cambios en el mercado y ajustar sus estrategias comerciales y políticas de gestión de riesgos según sea necesario.

1.2 Identificación de Datos Necesarios

Se debe disponer de los datos en todas las categorías disponibles (Delinquency, Spend, Payment, Balance, Risk) para las 189 variables y el dato de default o no default de cada cliente para un momento dado.

Se cuenta con una única base de datos.

Debemos asegurarnos que existan suficientes datos para el modelado, lo cual se confirma con las bases:

Base: 458.913 clientes

Datos: información mensual para cada cliente para un período de 13 meses.

Se considera que el caso tiene toda la información relevante y no se incorporan datos externos.

1.3 Selección de Modelos

La predicción del default es un caso adecuado para el Machine Learning Supervisado dado que tenemos una variable de respuesta a predecir en base a las variables que se incluyan en el modelo.

Se puede predecir la pertenencia de una observación a la categoría default o no default directamente. Alternativamente se puede predecir un valor numérico continuo en función de una o más variables, en el caso del default, el modelo podría predecir la probabilidad de que un cliente incumpla con sus pagos.

Regresión logística:

Aunque la regresión logística es rápida y eficiente para conjuntos de datos pequeños a medianos, puede enfrentar problemas de escalabilidad con grandes conjuntos de datos, especialmente si el número de características es muy alto. No maneja bien las relaciones no lineales entre las variables independientes y la variable dependiente sin la adición de términos polinómicos o transformaciones, lo que puede complicar el modelo y hacerlo menos interpretable.

Árbol:

Los árboles de decisión son fáciles de interpretar y explicar. Se pueden visualizar, lo cual facilita la comprensión de cómo se están tomando las decisiones. Aunque no son tan eficientes como XGBoost, los árboles de decisión pueden manejar grandes cantidades de datos y características. No requiere que los datos sean escalados o normalizados y pueden trabajar con datos faltantes. Adicionalmente puede capturar interacciones no lineales entre las variables.

Boosting:

Generalmente proporciona alta precisión debido a su enfoque iterativo y optimizado. Es extremadamente rápido y eficiente, tanto en entrenamiento como en predicción, gracias a su optimización y capacidad de paralelización. Maneja bien el sobreajuste mediante técnicas de regularización. Puede manejar una gran cantidad de datos y no requiere que los datos sean escalados o normalizados y pueden trabajar con datos faltantes.

La elección del árbol de decisión y XGBoost para predecir defaults se justifica por su capacidad de manejar grandes volúmenes de datos, alta precisión, y robustez contra el sobreajuste. La regresión logística, aunque útil en ciertos contextos, se descartó debido a sus limitaciones en la capacidad de manejo de grandes datos y menor capacidad para capturar relaciones no lineales complejas en comparación con los modelos seleccionados.

1.4 Formulación de Hipótesis

Para plantear hipótesis sobre las dimensiones de análisis y las variables clave que podrían influir significativamente en el logro de los objetivos del proyecto, podemos considerar algunas suposiciones basadas en las categorías de variables proporcionadas.

Hipótesis generales:

H1: Los clientes con altos valores en las variables de morosidad (D_*) tienen una mayor probabilidad de incumplimiento.

H2: Los clientes con valores bajos en las variables de pago (P_*) tienen una mayor probabilidad de incumplimiento.

H3: La interacción entre variables de morosidad (D_*) y gasto (S_*) incrementa la probabilidad de incumplimiento.

H4: Las variables de morosidad (D_*) son las más significativas respecto a otras categorías para predecir el default.

Hipótesis dentro de las categorías:

Delinquency variables (D_*):

H5: A mayor frecuencia de los pagos atrasados mayor probabilidad de default.

H6: A mayores períodos de morosidad mayor probabilidad de default.

Spend variables (S_*):

H7: A mayor nivel de gasto mensual mayor probabilidad de default.

H8: A mayor gasto en comparación con el límite de crédito mayor probabilidad de default.

Payment variables (P_*):

H9: A menor puntualidad de los pagos mayor probabilidad de default.

H10: A menor proporción de pagos realizados en comparación con el saldo pendiente mayor probabilidad de default.

Balance variables (B_*):

H11: A menor saldo total de la cuenta mayor probabilidad de default.

H12: A mayor deuda en comparación con el límite de crédito mayor probabilidad de default.

H13: A mayor relación deuda-ingresos mayor probabilidad de default.

Risk variables (R_*):

H14: A menor puntuación crediticia mayor probabilidad de default.

H15: A menor edad, mayor probabilidad de default.

H16: A menor antigüedad de la cuenta mayor probabilidad de default.

H17: A mayor inestabilidad laboral (frecuencia de cambio de trabajo) mayor probabilidad de default.

Estas hipótesis se basan en suposiciones comunes sobre los factores que pueden influir en la capacidad de los clientes para cumplir con sus obligaciones financieras. Sin embargo, una vez que se entregue el modelo a la empresa la misma deberá validar a que corresponden las variables identificadas como significativas y constatar que el resultado hace sentido.

1.5 Plan de Trabajo

Se plantea el siguiente plan de trabajo:

Fase 1: Preparación y Recopilación de Datos

- Definir los objetivos y el alcance del proyecto, en su contexto.
- Identificar las fuentes de datos disponibles, proporcionadas por American Express las cuales pueden contar con datos internos de la empresa como adicionales del mercado.
- Formular hipótesis y analizar posibles modelos a aplicar.
- Definir métrica de evaluación.

Fase 2: Carga, exploración y análisis descriptivo de datos

- Carga de las bases.
- Explorar y visualizar los datos para comprender la distribución, la correlación y las relaciones entre las variables.
- Evaluar la calidad y la integridad de los datos disponibles y determinar si se necesitan limpieza, transformación o enriquecimiento de datos.
- Identificar posibles patrones, tendencias o anomalías en los datos que puedan ser relevantes para el análisis.
- Realizar análisis estadísticos descriptivos para resumir las características clave de los datos.

Fase 3: Preprocesamiento de Datos

- Llevar a cabo la limpieza de datos para eliminar valores atípicos, datos faltantes o inconsistentes.
- Transformar y normalizar los datos según sea necesario para cumplir con los requisitos de los modelos seleccionados.
- Manejar variables categóricas.

Fase 4: Construcción de Modelos

- Dividir los datos en conjuntos train y test.
- Entrenar varios modelos utilizando técnicas como la validación cruzada y la optimización de hiperparámetros.
- Evaluar el rendimiento de los modelos utilizando la métrica definida.

Fase 5: Evaluación y Validación

- Evaluar el rendimiento de los modelos utilizando los datos de test.
- Comparar el rendimiento de los modelos entre sí para identificar el mejor modelo predictivo.
- Interpretar los resultados y validar la relevancia de las características seleccionadas.
- Utilizar métricas clásicas de sensitividad y especificidad para evaluar la estabilidad y la generalización de los modelos.
- Evaluar la métrica M definida en el objetivo.

Fase 6: Documentación y traspaso

- Desarrollar documentación detallada y procedimientos para la operación y el mantenimiento continuo del modelo.
- Documentar todos los aspectos del proyecto hasta el momento, incluidos los datos utilizados, los métodos empleados.
- Envío del modelo y documentación al equipo de American Express.

Las siguientes fases serán realizadas por el equipo de American Express y no por el equipo externo que realizó este informe.

Fase 7: Implementación y Despliegue

- Implementar el modelo finalmente seleccionado en un entorno de producción.
- Capacitar al personal para su correcta utilización.
- Integrar el modelo en los sistemas existentes de American Express para su uso en la toma de decisiones en tiempo real.

Fase 8: Monitoreo y Mantenimiento

- Establecer un sistema de monitoreo continuo para evaluar el rendimiento del modelo en producción y detectar posibles problemas o cambios en el comportamiento del cliente.
- Realizar actualizaciones y ajustes periódicos al modelo según sea necesario en función de los cambios en los datos o en el entorno comercial.
- Mantener un proceso de retroalimentación con los interesados y el equipo para garantizar la eficacia continua del modelo en la gestión del riesgo crediticio.

Fase 9: Evaluación Final y Documentación

- Realizar una evaluación final del proyecto para revisar los resultados, los aprendizajes y las lecciones aprendidas.
- Sumar a la documentación del proyecto, los resultados obtenidos y las recomendaciones para futuros análisis.
- Presentar los hallazgos y las conclusiones a los interesados y al equipo de liderazgo de American Express.

1.6 Utilización de los Resultados

Luego de que el proyecto se encuentre implementado impactará en las siguientes decisiones y áreas de la empresa:

- **Toma de decisiones de aprobación de crédito:** Los modelos desarrollados pueden proporcionar información valiosa para optimizar el proceso de aprobación de créditos. Al predecir con mayor precisión el riesgo de incumplimiento, American Express puede tomar decisiones más informadas y eficientes sobre la aprobación de nuevos créditos y la asignación de límites de crédito.
- **Área de mitigación de riesgo:** Los resultados del proyecto pueden ser utilizados para implementar estrategias proactivas que ayuden a reducir el impacto de posibles incumplimientos. Al identificar a los clientes con mayor riesgo de incumplimiento, American Express puede implementar medidas preventivas para minimizar las pérdidas financieras y proteger su cartera de créditos.

- **Áreas de presupuestación anual:** Los resultados del proyecto pueden proporcionar información clave para la planificación y presupuestación anual de American Express. Al predecir el riesgo de incumplimiento con mayor precisión, la empresa puede asignar recursos de manera más eficiente y tomar decisiones estratégicas informadas sobre la gestión de riesgos y la asignación de capital.
- **Marketing:** Los resultados del proyecto también pueden ser utilizados por el equipo de marketing para personalizar las ofertas y campañas dirigidas a clientes potenciales y existentes. Al comprender mejor el riesgo de incumplimiento de los clientes, American Express puede adaptar sus mensajes y ofertas para maximizar el retorno de inversión y mejorar la efectividad de sus estrategias de marketing.
- **Desarrollo de Productos y Servicios:** Los modelos desarrollados pueden proporcionar información útil para el desarrollo de nuevos productos y servicios financieros. Al comprender mejor el comportamiento financiero de los clientes y sus patrones de gasto, American Express puede identificar oportunidades para diseñar productos y servicios que satisfagan las necesidades específicas de diferentes segmentos de clientes y mejoren su experiencia general.
- **Servicio al Cliente:** Los resultados del proyecto pueden ser utilizados por el equipo para ofrecer un servicio más personalizado y adaptado a las necesidades individuales de los clientes. Al tener en cuenta el riesgo de incumplimiento de los clientes, American Express puede anticipar sus necesidades y ofrecer soluciones proactivas para ayudarles a administrar sus cuentas de manera más efectiva y evitar el incumplimiento.
- **Legal y cumplimiento:** Los resultados del proyecto pueden ser utilizados por el área de cumplimiento y legal para garantizar el cumplimiento de las normativas y regulaciones financieras. Al predecir el riesgo de incumplimiento con mayor precisión, American Express puede implementar políticas y procedimientos que cumplan con los estándares regulatorios y protejan los intereses de los clientes y la empresa.

1.7 Integración Interdepartamental

Para poder desarrollar el proyecto se requiere el apoyo de las siguientes áreas de American Express:

- **Departamento de Gestión de Riesgos Crediticios:** Su importancia radica en su experiencia en la evaluación y gestión del riesgo crediticio. Su papel sería proporcionar conocimientos especializados sobre los datos históricos de incumplimiento y colaborar en la definición de las variables relevantes para la predicción del default. Además, este departamento podría ayudar en la validación y la interpretación de los resultados del modelo.
- **Departamento de Tecnología de la Información (TI):** El departamento de TI desempeñaría un papel crucial en la implementación técnica del proyecto. Serían responsables de desarrollar y mantener la infraestructura necesaria para recopilar, almacenar y procesar los datos, así como implementar y desplegar los modelos de aprendizaje automático en un entorno de producción. Además, este departamento colaboraría estrechamente con otros equipos para garantizar la integridad y la seguridad de los datos utilizados en el proyecto.

Si bien estas son las áreas fundamentales para el desarrollo del proyecto, el resultado del proyecto impactará muchas áreas más (punto 1.6) por lo que a medida que se avance se involucrarán a dichas áreas para analizar resultados y obtener mayor valor agregado.

2 Extracción, Transformación y Carga de Datos

2.1 Diseño de la Estructura de la Tabla Analítica

La base de datos original de American Express, denominada train data, se compone de 5.531.451 observaciones (filas) con información en 190 columnas.

Por otro lado, se posee otra base independiente de train labels las cual se compone de 458.913 (filas) (como clientes únicos existen) con los Customer ID y la columna y target como variable integer. La variable binaria objetivo se calcula considerando si el cliente no paga el monto adeudado 120 días después de la fecha de su último estado de cuenta: evento de incumplimiento (defaulf=1).

2.1.1 Información en columnas

- *Variable de Identificación Única:*

  La columna Customer_ID contiene identificadores únicos de clientes, representados como caracteres (chr). Se tiene un total de 458.913 clientes únicos

Las variables (exceptuando al identificador = Customer ID) se dividen en las categorías presentadas a continuación y las mismas están anonimizadas y normalizadas.

*D_ * = Variables de morosidad*
*S_ *= Variables de gasto*
*P_ *= Variables de pago*
*B_ * = Variables de balance*
*R_ * = Variables de riesgo*

- *Variables Categóricas:*

  La base cuenta con un total de 11 variables categóricas que corresponden a las columnas: D_68, D_66, D_64, D_63, D_126, D_120, D_117, D_116 representadas como factor.

- *Variables Temporales:*

  La columnas S_2 representa Fecha de estado de cuenta y se encuentra en formato "POSIXct" "POSIXt"

- *Variables numéricas:*

  Las columnas $ B_7, $ B_23, ..., $ D_77 son variables numéricas (de tipo num) que representan diferentes atributos o características relacionadas con cada cliente.

2.1.2 Información en filas

Cada fila en la tabla analítica representa el cierre de estado de mes de cada cliente, por lo cual se puede observar un mes por cliente con variación en la fecha de cierre ya que la misma puede ser seleccionada según preferencias por cada cliente. Se confirma que no hay meses repetidos por cliente y todos los clientes cuentan con 13 meses de cierres disponibles.

2.1.3 Dataset final

Debido al tamaño del dataset, no fué posible realizar un análisis con todas las variables ya que los archivos .csv originales consumen casi toda la memoria por lo que se procede a realizar un proceso de reducción de las mismas llegando finalmente a una Tabla con 12 columnas, incluyendo Customer ID, target, 10 variables y 458.913 filas (clientes). Ver punto 2.2 y 2.3

2.2 Construcción de Variables

Debido a las limitaciones de capacidad de procesamiento para una base de datos tan amplia en formato .csv, se investigaron las soluciones propuestas por otros participantes. Como resultado, se optó por utilizar un formato de archivo .ftr [2].

El formato .ftr, conocido por su eficiencia en el manejo de grandes volúmenes de datos, permitió al equipo gestionar y procesar la información de manera más efectiva. Esta elección no solo facilitó la carga y manipulación de datos complejos, sino que también optimizó significativamente los tiempos de procesamiento, cumpliendo con los requisitos analíticos del proyecto en curso.

La base de datos train en formato .ftr cuenta con 5.531.451 filas y 191 columnas, con un tamaño de 1.73 GB, en comparación con la base train en formato .csv de 16.39 GB. En esta base de datos train .ftr, la columna target ya se encuentra incluida entre las 191 columnas mencionadas anteriormente, por lo cual no fue necesario realizar la unión de la base train con la base target.

Para la carga del archivo .ftr se utilizó la librería arrow. Se utilizó la función read_feather() para leer el archivo .ftr. y a lo largo del procesamiento de la base guarda la base de datos transformada en un nuevo archivo .ftr utilizando write_feather().

2.3 Construcción de la Tabla de Datos Analítica

Debido al gran tamaño de la base de datos y a la imposibilidad de manejarla directamente en R, se inició un proceso de reducción de la misma.

Inicialmente, se redujeron las 190 variables originales a 75 mediante un estudio previo de selección basado en criterios de relevancia y viabilidad, según estudios previos [2]. Este proceso involucró la aplicación de Recursive Feature Elimination (RFE) con Cross Validation para seleccionar las variables más importantes. Funciona eliminando de forma iterativa las características menos importantes y construyendo un modelo con las características restantes. El proceso se repite hasta que se alcanza un criterio de parada, como un número predeterminado de características o la mejora en la métrica de rendimiento del modelo.

A continuación, se describen los pasos resumidos de análisis previo en Python [2]:

I. *Configuración Inicial:*
   Se define un control para RFE con métodos específicos (en este caso lmFuncs para modelos lineales y validación cruzada (cv) con 5 folds.

   ```
   ctrl <- rfeControl(functions = lmFuncs, method = "cv", number = 5)
   ```

II. *Ejecución de RFE*

Se utiliza la función rfe para realizar la selección recursiva de características. Se especifican los conjuntos de datos de entrenamiento (train_X, train_Y) y se proporciona una lista de tamaños de subconjunto de características (sizes) sobre los cuales evaluar el rendimiento.

```
rfe_result <- rfe(train_X, train_Y,
                  sizes = c(25, 50, 75, 100),
                  rfeControl = ctrl)
```

III. Resultados de RFE

Después de ejecutar rfe, se obtienen varios resultados que incluyen el rendimiento del modelo (como RMSE, R-squared, MAE) para cada tamaño de subconjunto de características evaluado durante la validación cruzada. También se identifican las variables más importantes seleccionadas por RFE.

Por ejemplo, se muestran estadísticas como RMSE, R-squared y MAE para diferentes tamaños de subconjunto de características, y se destacan las variables más importantes seleccionadas.

Variables	RMSE	R2	MAE	RMSE SD	R2 SD	MAE SD
25	0.277	0.564	0.192	0.007	0.008	0.007
50	0.270	0.588	0.182	0.002	0.007	0.002
75	0.270	0.597	0.181	0.002	0.007	0.002
100	0.270	0.599	0.182	0.002	0.007	0.002
158	0.267	0.601	0.179	0.002	0.007	0.002

Tabla 1 – Resultados RFE

Considerando estos resultados se debe tener en cuenta que al eliminar variables por restricciones de procesamiento se está sacrificando rendimiento del modelo.

IV. Interpretación de Resultados

Además de los resultados de rendimiento del modelo, se obtiene una lista de las variables más importantes para el modelo.

```
# The top 75 features
predictors(rfe_result)[1:75]
```

'B_7''B_23''B_1''B_11''B_36''B_15''S_19''B_14''D_104''D_77''D_62''D_103''P_2''B_25''B_28''B_27''D_48''B_3''R_2''R_23''D_133''D_58''S_7''D_131''R_1''D_43''B_9''D_139''S_23''B_2''D_118''R_24''D_121''D_55''B_18''D_143''S_25''R_15''D_141''R_27''D_47''D_119''B_37''D_140''D_116''R_22''D_52''D_112''R_4''R_12''D_127''D_54''D_120''S_3''B_8''R_28''D_102''B_22''R_21''D_44''D_41''P_3''R_19''B_31''S_12''D_46''D_92''D_109''D_78''R_25''R_18''S_6''D_72''R_10''P_4'

Siguen existiendo problemas de procesamiento de datos con 75 variables, se prosigue la reducción de las mismas a 40 según orden de relevancia presentado anteriormente. Se obtiene la base *Train3* con 5.531.451 observaciones y 42 columnas.

Se procede a eliminar el efecto de la variable de tiempo (S_2 de Fecha) del estado de cuenta ya que no se trabajará con series temporales.

Se cuenta con dos alternativas para trabajar con los datos: tomar el último cierre de cada cliente o promediar los cierres de cada mes disponible para cada cliente. Dado que el rango temporal (13 meses) no tiene una extensión demasiado grande que pueda incluir patrones del cliente muy diferentes se opta por promediar los datos por cliente. El promedio reduce la variabilidad, es representativo y robusto, reflejando el comportamiento del cliente.

Se avanza con el cálculo de promedio por Customer ID. Para esto se divide la base en 8 secciones ya que la base de datos continúa con un tamaño importante.

Se realiza un cálculo de promedio para cada subconjunto de datos (*Train3.1, Train3.2, ..., Train3.8*) agrupando por customer_ID.

Una vez calculados los promedios, se procedió a unir los ocho subconjuntos en un único dataframe denominado *Train4*.

Conteniendo 458.913 filas (así como clientes únicos) y 42 variables.

A pesar de haber disminuido considerablemente el tamaño de la base, se decide realizar una última reducción ya que la capacidad de procesamiento de nuestras computadoras no era suficiente para el cálculo de los modelos en tiempos considerables. Por consiguiente, se llevó a cabo una selección adicional de las variables más relevantes para el análisis y modelado final. Se identificaron las 10 variables más significativas dentro del dataframe *Train4*, en función de su impacto y relevancia para los objetivos del estudio. Estas variables se consolidaron en un nuevo dataframe denominado *Train5,* que se utilizó como base para análisis adicionales y el desarrollo de los modelos predictivos.

A continuación, se muestran las primeras 6 filas de la base *Train5*

```
head(Train5)
# A tibble: 6 × 12
  customer_ID                  B_7   B_23    B_1  B_11  B_36  B_15  S_19  B_14  D_104   D_77 target
  <chr>                      <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1 0000099d6bd597052cdcda90ffabf565… 9858   9350.  8466. 7847. 7201. 9646. 7344. 9516. 15370  14006.     0
2 00000fd6641609c6ece5454664794f03… 9914.  8525.  9343. 7815. 7537. 7126. 7040  8556.  7023. 13098.     0
3 00001b22f846c82c51f6e3958ccd8197… 10220.  9604.  7010. 6938. 6786. 6603  7207. 7005.  6874. 13838.     0
4 000041bdba6ecadd89a52d11886e8eaa… 10303   9512. 10174. 9201. 6773. 6735. 7535. 9843.  6995. 13943   0
5 00007889e4fcd2614b6cbe7f8f3d2e5c… 11774. 11414.  7525. 6940. 6533. 7583. 6960. 6595. 15279. 13353      0
6 000084e5023181993c2e1b665ac88dbb… 9393.  7677.  9757. 8307. 7287. 7110. 7049  8157.  6605.   NaN      0
```

3 Exploración y análisis descriptivo

3.1 Análisis Descriptivo de la Tabla de Datos

En esta sección se presenta una descripción general de las categorías de variables contenidas en la tabla de datos. Es importante destacar que el análisis se fundamenta exclusivamente en las categorías generales proporcionadas, ya que se carece de información detallada sobre las variables individuales. Por lo tanto, no se llevó a cabo un análisis exhaustivo de cada variable debido a esta limitante en los datos.

Como se describió en 2.3 *Train5* se compone de 458.913 observaciones (filas) y 10 variables (12 columnas considerando Customer ID y target). Se analizó la estructura general y los datos contenidos en la base *Train5*. Después de la preparación y análisis de los datos de *Train5*, se procedió a realizar un resumen estadístico y cálculos adicionales para comprender mejor la distribución de los datos y la variable objetivo "target".

3.1.1 Resumen Estadístico

Se utilizó la función summary() para obtener estadísticas descriptivas de las variables numéricas contenidas en *Train5*, incluyendo B_7, B_23, B_1, B_11, B_36, B_15, S_19, B_14, D_104, D_77, y target.

customer_ID		target		B_7		B_23		B_1		B_11	
Length	458.913	Min	0	Min	2.874	Min	73	Min	960	Min	215
Class	character	1st Qu	0.0000	1st Qu	10.115	1st Qu	9.403	1st Qu	8.666	1st Qu	7.804
Mode	character	Median	0.0000	Median	11.457	Median	11.007	Median	10.167	Median	9.405
		Mean	0.2589	Mean	11.648	Mean	11.197	Mean	10.429	Mean	9.851
		3rd Qu	10.000	3rd Qu	13.295	3rd Qu	13.120	3rd Qu	12.120	3rd Qu	11.582
		Max	10.000	Max	44.515	Max	15.844	Max	44.498	Max	16.146
B_36		**B_15**		**S_19**		**B_14**		**D_104**		**D_77**	
Min	83	Min	9	Min	11	Min	425	Min	16	Min	25
1st Qu	6.770	1st Qu	6.870	1st Qu	6.770	1st Qu	8.146	1st Qu	7.090	1st Qu	10.718
Median	7.072	Median	7.222	Median	7.071	Median	9.989	Median	8.808	Median	12.502
Mean	7.060	Mean	7.955	Mean	7.034	Mean	10.082	Mean	10.873	Mean	12.029
3rd Qu	7.344	3rd Qu	7.708	3rd Qu	7.340	3rd Qu	11.716	3rd Qu	15.283	3rd Qu	13.537
Max	14.793	Max	46.549	Max	12.027	Max	46.306	Max	15.518	Max	18.689
		NA's	604					NA's	2.532	NA's	90.688

Tabla 2 – Resumen de datos

Al observar el resumen, se destacan los siguientes aspectos:

Valores Faltantes (NA's):
Algunas variables tienen valores faltantes significativos (B_15, D_104, D_77) que ascienden a 93.824 (604 + 2.532 + 90.688). Esto podría requerir técnicas de imputación o manejo de valores faltantes en el análisis. Dado el gran tamaño inicial de la base de datos, que constaba de 458.913 registros, se eliminan los clientes (filas) con valores faltantes (NA) del

conjunto *Train5* para obtener un conjunto limpio denominado *Train_clean*. En total, se eliminaron 179.772 registros, resultando en un conjunto limpio llamado *Train_clean* con 279.141 datos.

No se realizaron imputaciones de datos perdidos debido al volumen sustancial del conjunto inicial y a las limitaciones prácticas en su gestión eficiente. Esta decisión se tomó para asegurar que el análisis se basará en una muestra completa y representativa, evitando la complejidad adicional. Si el contexto fuera distinto, es decir, si se contará con menos datos, se hubiera considerado la imputación de valores perdidos.

Variabilidad *y Rango:*
Hay una gran variabilidad en las variables numéricas, con valores mínimos y máximos que difieren significativamen

3.1.2 Resumen Estadístico con Outliers

El siguiente diagnóstico muestra la cantidad de outliers según la regla de Tukey en cada variable. Se destaca que la variable B_15 tiene 85.598 outliers, seguida por B_36 con 12.141, y S_19 con 9.342. Las demás variables tienen menos de 5.000 outliers cada una.

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
B_7	2,874	10,114.692	11,648.1671247	11,457.308	13,295.231	44,515.00	0	0	222
B_23	73	9,403.154	11,197.0471553	11,007.385	13,120.154	15,843.69	0	0	23
B_1	960	8,665.769	10,429.3012027	10,167.385	12,119.692	44,498.00	0	0	902
B_11	215	7,804.000	9,851.3197675	9,405.154	11,581.769	16,146.00	0	0	24
B_36	83	6,770.462	7,060.0595400	7,071.923	7,344.000	14,793.31	0	0	12,141
B_15	9	6,869.615	7,954.6399820	7,222.154	7,708.385	46,549.00	0	0	84,598
S_19	11	6,770.385	7,033.7084754	7,071.077	7,339.769	12,027.08	0	0	9,342
B_14	425	8,145.846	10,081.8767317	9,989.231	11,716.154	46,306.00	0	0	1,164
D_104	16	7,090.385	10,873.1712048	8,807.846	15,282.615	15,518.46	0	0	0
D_77	25	10,718.077	12,029.2775917	12,502.000	13,537.333	18,688.77	0	0	3,709
target	0	0.000	0.2589336	0.000	1.000	1.00	340,085	0	0

Tabla 3 – Resumen de datos con outliers

Outliers*:*
En el conjunto de datos *Train_clean*, los valores atípicos identificados (variables B_15, B_36, entre otras) no fueron ajustados ni excluidos. Esta decisión se fundamenta en el hecho de que no se cuenta con información detallada sobre las variables, más allá de las categorías a las que pertenecen. Por lo tanto, no se dispone de suficiente contexto para determinar si un valor atípico es simplemente un dato incorrectamente cargado o representa información válida pero inusual. En consecuencia, todos los valores atípicos se mantuvieron en el análisis sin aplicar métodos de transformación o exclusión.

El proceso de preparación de datos se centró en garantizar la integridad y la calidad de los datos limpiando los NA's, lo que resultó en un conjunto de datos limpio y consistente para el análisis posterior. La decisión de mantener los valores atípicos sin ajustes adicionales se basó en la falta de información detallada sobre las variables subyacentes más allá de sus categorías.

3.1.3 Análisis del Porcentaje de Default

La base de Amex original cuenta aproximadamente con 1,72% de clientes en default, sin embargo, la base disponible ya se encuentra submuestreada (*Train5*) al 5% de manera que se llega a un porcentaje de default de 25,89%. La base *Train5* sigue desbalanceada por lo que se decide avanzar con un segundo submuestreo (Sección 3.1.6) llegando a una base 50/50.

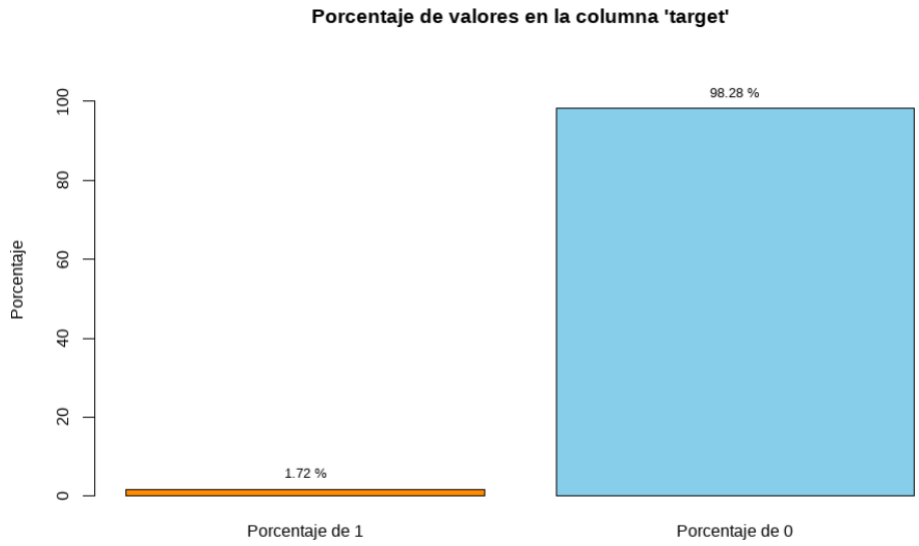

Figura 1 – Default en base original de Amex

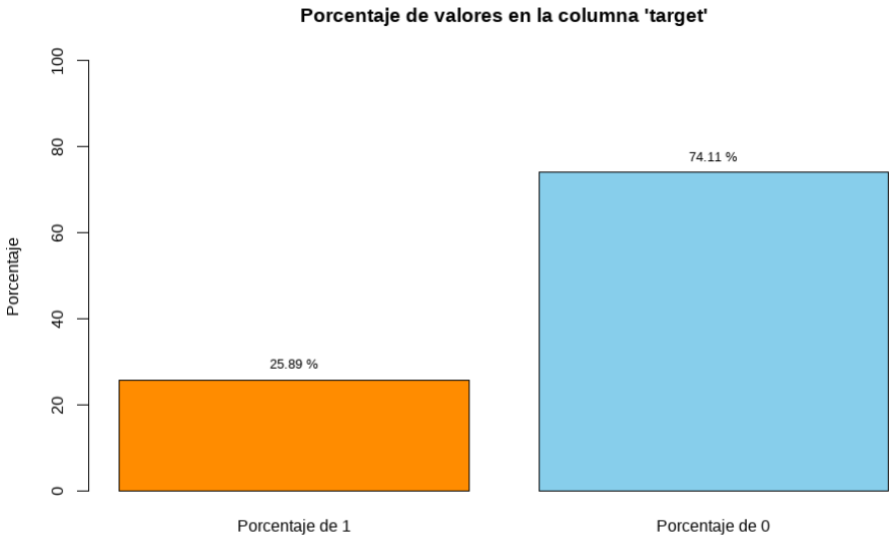

Figura 2 – Default en base de trabajo (Train5)

3.1.4 Desbalance de datos

Para abordar el desbalance en la variable objetivo (target), se implementó un proceso de submuestreo, con el fin de equilibrar los datos 50/50. A continuación se detallan los pasos realizados:

Submuestreo para equilibrar categorías:
Se optó por el submuestreo para mitigar el efecto del desbalance inicial en la variable objetivo (target). Esto implica reducir la cantidad de muestras de la categoría mayoritaria (no default, target = 0) para igualar el número de muestras de la categoría minoritaria (default, target = 1). Este enfoque ayudó a mejorar la capacidad del modelo para predecir ambas clases de manera equitativa.

Selección aleatoria de muestras:
Utilizando la función sample(), se seleccionaron aleatoriamente 50.000 filas de cada grupo (sampled_0 para target =0 y sampled_1 para target = 1). Esta selección aleatoria aseguró que la muestra conservará la representatividad de los datos originales, reduciendo el riesgo de sesgo. Se obtuvieron así los conjuntos de datos *Train_clean_0* y *Train_clean_1*

División en conjuntos de Train y Test:
Posteriormente, cada categoría se dividió aleatoriamente en conjuntos de Train (train_0, train_1) y Test (test_0, test_1). Se asignó el 70% de las muestras para Train y el 30% restante para Test.

Combinación de conjuntos balanceados:
Los conjuntos de Train resultantes *Train_sub* y Test *Test_sub* fueron combinados para formar conjuntos de datos completamente balanceados.

3.2 Evaluación de la Pertinencia de las Variables

3.2.1 Análisis de Correlaciones

Se llevó a cabo un análisis de correlación para identificar las relaciones entre las variables del conjunto de datos. Para este análisis se utilizó el coeficiente de correlación de Spearman, adecuado para evaluar relaciones entre variables. Las correlaciones fueron visualizadas y calculadas con las funciones rcorr y corrplot.

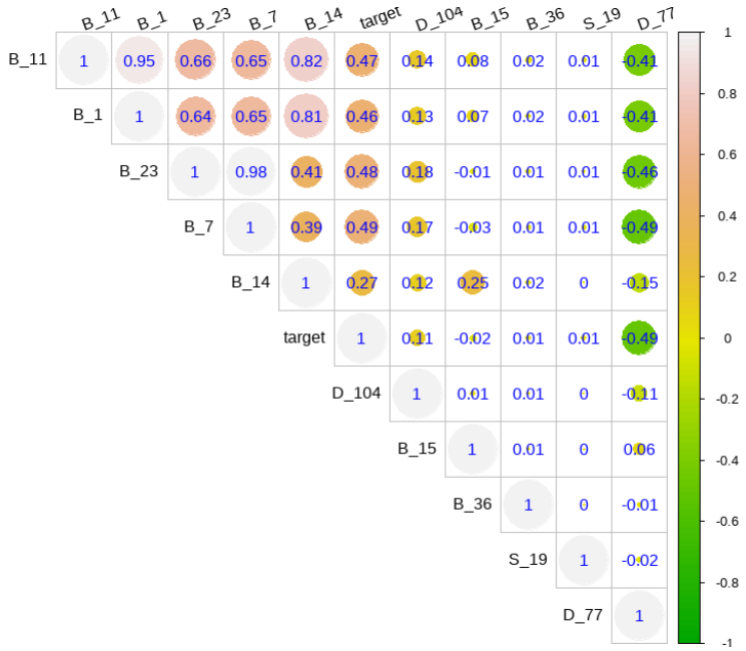

Figura 3 – Análisis de correlación

Las siguientes variables mostraron correlaciones significativas entre sí:

- **B_7 y B_23:** 0,98
- **B_11 y B_14**: 0,82
- **B_1 y B_14**: 0,81
- **B_1 y B_11**: 0,95

A pesar de las correlaciones observadas entre las variables del conjunto de datos, no se eliminan automáticamente aquellas que muestran una alta correlación entre sí. Esta decisión se basa en la implementación de un método avanzado para la selección de características conocido como el método de eliminación recursiva de características (RFE) con validación cruzada, como se describe en el módulo 2.

El método RFE con validación cruzada evaluó exhaustivamente las variables en términos de su capacidad para mejorar la precisión del modelo predictivo a través de múltiples iteraciones de entrenamiento y validación.

Por lo tanto, aunque el análisis de correlación proporcionó información valiosa sobre las relaciones entre las variables, la decisión final de inclusión se guio por un enfoque riguroso y orientado al rendimiento del modelo.

3.2.2 Análisis Univariado

En este capítulo, se emplearon histogramas y diagramas de cajas (boxplots), con el fin de poder visualizar la distribución y las características de las variables de interés, ajustando los rangos para resaltar mejor las características específicas de cada variable.

Se presentan los análisis de las variables más significativas:

B_7

El análisis del histograma de la variable B_7 revela una dispersión de datos entre 6.000 y 16.000 en el eje x. Se observa una concentración notable alrededor de 11.000, con un pico significativo en la frecuencia, indicando que este valor es uno de los más comunes en la distribución. Esta configuración sugiere una distribución normal.

Al examinar el boxplot, se nota que la caja está posicionada más hacia abajo en el gráfico, indicando una mediana que está ligeramente desplazada hacia el primer cuartil. Se observan más outliers por encima del bigote superior que por debajo del inferior, lo cual sugiere la presencia de valores atípicos en el extremo superior de la distribución.

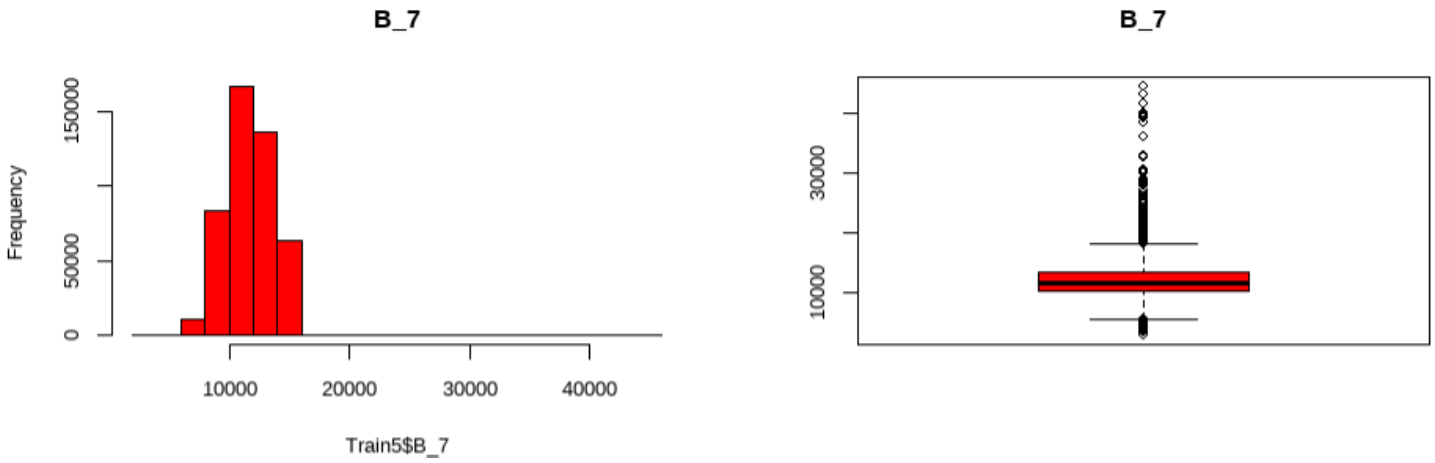

Figura 4 – Análisis univariado B_7

B_23

El análisis del histograma de la variable B_23 muestra picos notables en torno a 9.000 y 10.000, indicando que estos son los valores más comunes en la distribución bimodal. La dispersión de los datos se extiende desde 5.500 hasta 16.000, lo que sugiere una variabilidad considerable dentro de este rango.

El boxplot muestra un bigote inferior más largo que el superior, con outliers en el extremo inferior de la distribución.

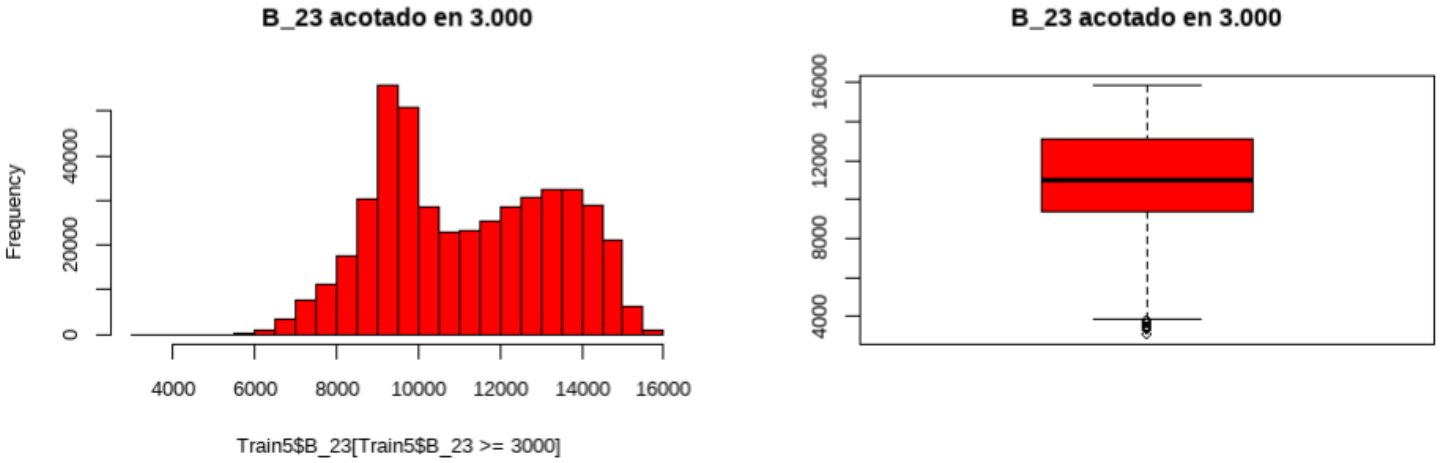

Figura 5 – Análisis univariado B_23

B_11

El histograma de la variable B_11 revela una distribución con sesgo a la izquierda con picos importantes alrededor de los 7.000 y 9.000 en el eje x, la distribución de la variable B_11 presenta una amplia dispersión de valores a lo largo del rango de 5.000 a 16.000 en el eje x. La frecuencia de los valores disminuye gradualmente fuera del rango 7.000 - 9.000 indicando una variabilidad considerable.

El boxplot muestra que la mediana está ligeramente inclinada hacia el primer cuartil. Los bigotes tienen un tamaño similar, pero existen outliers en el extremo inferior.

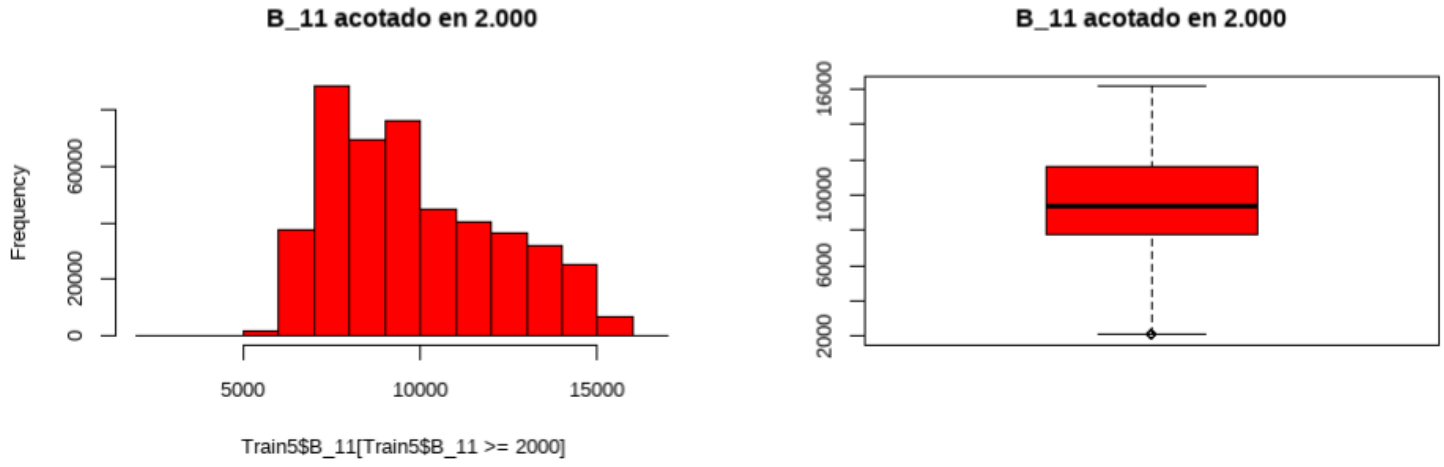

Figura 6 – Análisis univariado B_11

D_77

El histograma de la variable D_77 revela una distribución caracterizada por una concentración notable de valores entre 11.000 y 15.000, indicando una tendencia central hacia valores más altos. La dispersión de los datos abarca un rango más amplio, desde 4.000 hasta 16.000, lo que sugiere una variabilidad significativa en los datos observados. Además, se observa la presencia de posibles valores atípicos en el extremo superior del rango, superiores a 16.000.

En el diagrama de caja correspondiente, la mediana está más cercana al tercer cuartil que al primero, lo que indica una distribución sesgada hacia valores más altos. Se identifican puntos que se extienden más allá del bigote inferior, indicando la presencia de valores atípicos en el extremo inferior, y un punto por encima del bigote superior, señalando posibles valores atípicos en el extremo superior del conjunto de datos.

Estos hallazgos sugieren una distribución no normal de la variable D_77, con una tendencia central hacia valores más elevados y la presencia de valores atípicos.

.

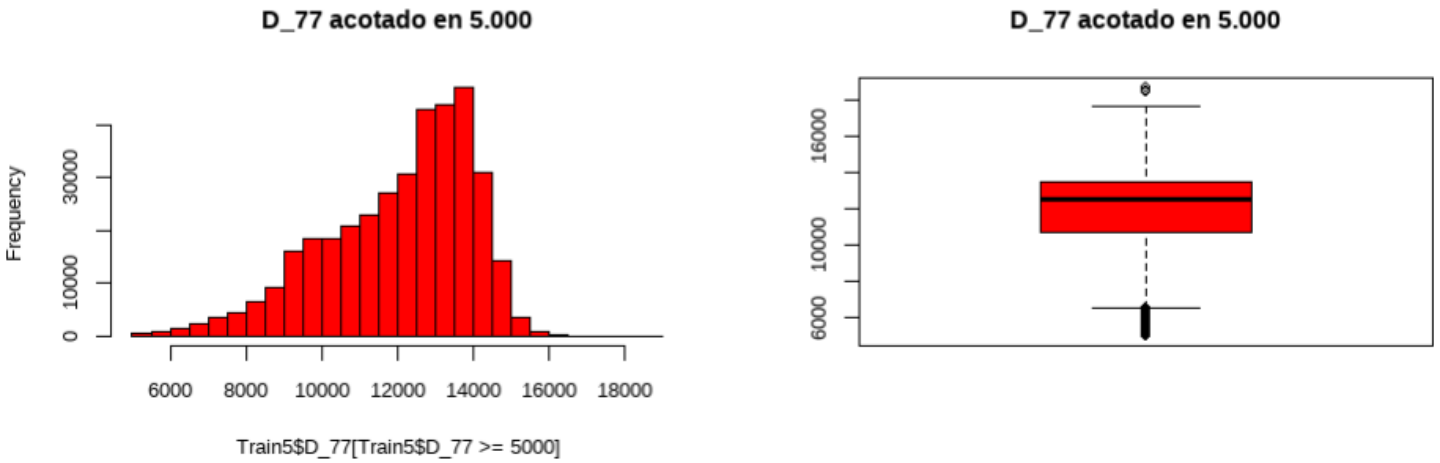

Figura 7 – Análisis univariado D_77

El resto de los análisis univariados se encuentran en el Anexo 1.

3.3 Identificación de Variables Clave

Según las hipótesis planteadas en el punto 1.3 las variables que podrían ser significativas pertenecen a las siguientes categorías: Morosidad (D_*), Gasto (S_*) y Pago (P_*).

Adicionalmente luego de realizar el análisis RFE [2] se confirma que las variables más importantes pertenecen a los grupos: Morosidad (D_*), Gasto (S_*) y Balance (B_*).

4 Modelado y Evaluación

4.1 Selección de Variables

Entre estas variables seleccionadas se encuentran B_7, B_23, B_1, B_11, B_36, B_15, S_19, B_14, D_104, y D_77. Cada una representa diferentes aspectos o atributos que pueden ser relevantes para el modelo predictivo. Por ejemplo, las variables que comienzan con "B_" de Saldo podrían estar relacionadas con balances o montos financieros específicos, mientras que las que comienzan con "D_" de Morosidad podrían estar asociadas con métricas de impago o delincuencia crediticia. Identificar y entender el impacto relativo de cada una de estas variables en la predicción del objetivo es fundamental para construir un modelo robusto y efectivo.

4.2 Estimación y Comparación de Modelos

4.2.1 Árbol

Para el modelado de los árboles, se limitó (por capacidad de procesamiento) maxdepth a 4. El árbol de decisión solo podrá realizar hasta 4 divisiones para clasificar o predecir la variable objetivo. Esto significa que cada rama del árbol tendrá como máximo 4 niveles antes de alcanzar una decisión final (terminal node).

Se realizó un árbol inicial que luego fue podado. La poda del árbol inicial se realiza por la regla del codo obteniéndose un cp de poda de 0.011 como se ve en la gráfica debajo:

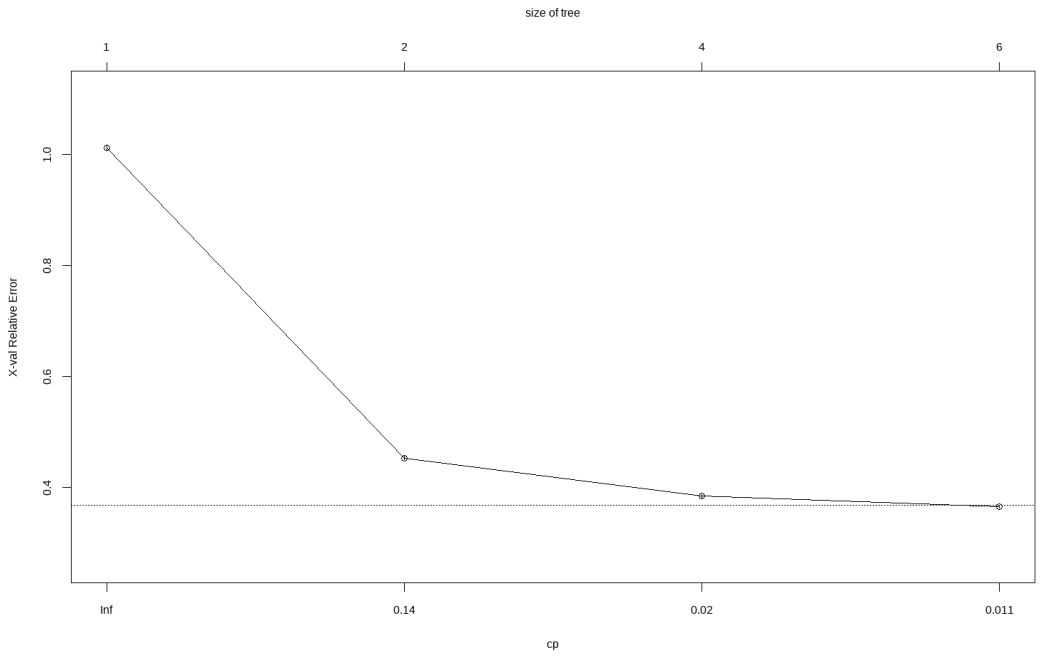

Figura 8 – Regla del codo

Con el cp calculado anteriormente se poda el árbol obteniéndose el *arbol.pr*

```
> arbol.pr
n= 70000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 70000 35000 0 (0.50000000 0.50000000)
   2) B_7< 11259.73 26281  3551 0 (0.86488338 0.13511662)
     4) D_77>=10870.74 22635  1543 0 (0.93183123 0.06816877) *
     5) D_77< 10870.74 3646  1638 1 (0.44925946 0.55074054)
      10) B_7< 10324.19 1661   623 0 (0.62492474 0.37507526) *
      11) B_7>=10324.19 1985   600 1 (0.30226700 0.69773300) *
   3) B_7>=11259.73 43719 12270 1 (0.28065601 0.71934399)
     6) D_77>=11909.95 12118  5836 0 (0.51840238 0.48159762)
      12) B_11< 10164.69 4988  1260 0 (0.74739374 0.25260626) *
      13) B_11>=10164.69 7130  2554 1 (0.35820477 0.64179523) *
     7) D_77< 11909.95 31601  5988 1 (0.18948767 0.81051233) *
```

En el árbol de decisión, las variables importantes se pueden identificar según su aparición en los nodos de división. Las variables que más influyen en las decisiones del modelo se encuentran en los nodos superiores del árbol, ya que son las que proporcionan la mayor ganancia de información en cada división; como puede verse debajo:

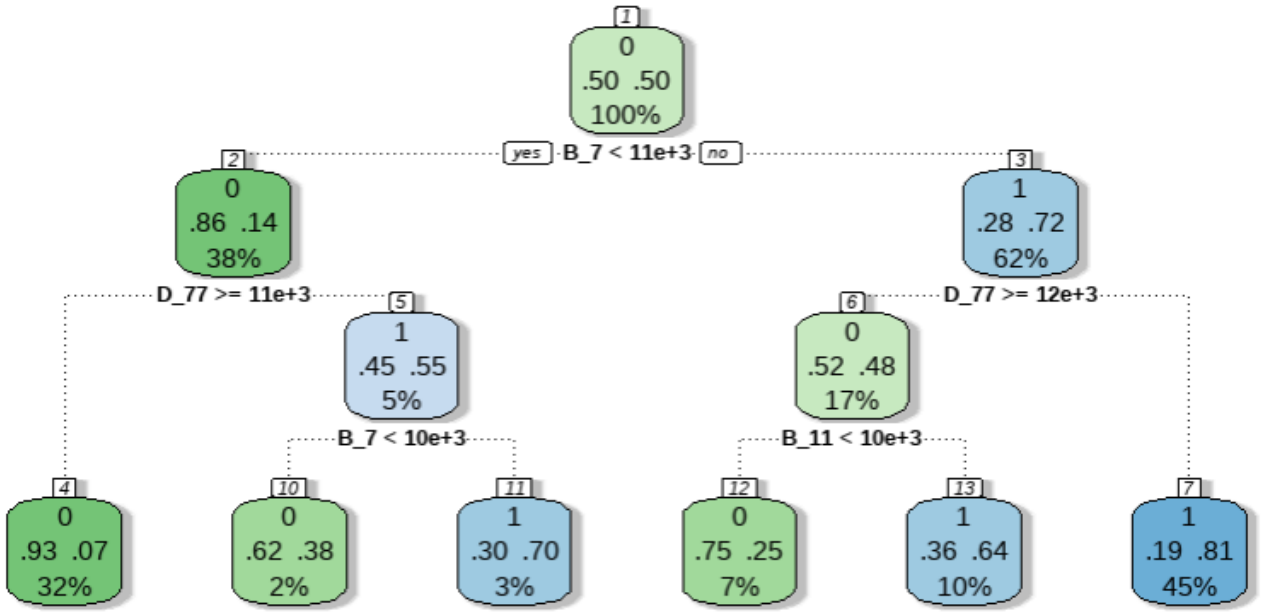

Figura 9 – Árbol Podado

Revisando el resultado del árbol, se obtiene que las variables en orden de importancia son: B_7, D_77 y B_11. Podemos destacar que con solo tres variables se logra un buen modelo (ver Resultados 4.2.3) y que las categorías más significativas son las de saldo (Balance) y luego las de morosidad (Delinquency).

También se utilizó *variable.importance* para revisar la importancia. Este código calcula y muestra la importancia relativa de cada variable en relación con el conjunto completo de variables consideradas por el modelo de árbol de decisión. Obteniéndose el siguiente resultado:

B_7	B_23	D_77	B_1	B_11	B_14	D_104
2.509557e-01	2.379824e-01	1.658158e-01	1.408877e-01	1.366855e-01	6.714166e-02	3.028222e-04
B_15	B_36					
2.110670e-04	1.747855e-05					

Como puede verse, la prioridad calculada con *variable.importance* no es la misma que aparece en los nodos terminales del árbol. Esto se debe a que los árboles de decisión pueden tomar decisiones basadas en la interacción compleja de múltiples variables y la estructura específica de los datos.

4.2.2 XGBoost

Se realiza un XGBoost inicial como punto de partida.

```
##### xgb.Booster
raw: 54.2 Kb
call:
  xgb.train(params = list(eta = 0.05, objective = "binary:logistic",
            max_depth = 4, subsample = 0.8), data = datos_train, nrounds = 30)
params (as set within xgb.train):
  eta = "0.05", objective = "binary:logistic", max_depth = "4", subsample = "0.8", validate_parameters = "TRUE"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
# of features: 10
niter: 30
nfeatures : 10
```

Luego se optimizan los parámetros:

Learning Rate (tasa de aprendizaje):
Parámetro que controla la contribución de cada árbol al modelo. Un valor más bajo hace que el modelo aprenda más lentamente, lo que puede mejorar la precisión final del modelo al permitir una convergencia más suave. Sin embargo, un valor muy bajo puede requerir más árboles para lograr un rendimiento óptimo y aumentar el tiempo de entrenamiento. Por lo general, valores típicos van entre 0.01 y 0.3, aunque la elección depende del problema específico.

Tree Depth (profundidad máxima de cada árbol):
Controla la complejidad del modelo al limitar cuántas divisiones se pueden hacer en cada árbol. Un valor más alto permite árboles más profundos y complejos, lo que puede capturar relaciones más complejas en los datos. Sin embargo, árboles demasiado profundos pueden llevar a sobreajuste (overfitting), especialmente en conjuntos de datos pequeños. Un valor común podría ser entre 3 y 10, dependiendo de la naturaleza del problema y la cantidad de datos disponibles.

Sample Size (tamaño de muestra):
Se refiere al tamaño de muestra utilizado para la construcción de cada árbol.

Óptimos:

```
> optboosting [12,]
  tree_depth learn_rate sample_size
12          6       0.05           1

##### xgb.Booster Optimizado
raw: 456.7 Kb
call:
  xgb.train(params = list(eta = 0.05, objective = "binary:logistic",
            max_depth = 6, subsample = 1), data = datos_train, nrounds = 100)
params (as set within xgb.train):
  eta = "0.05", objective = "binary:logistic", max_depth = "6", subsample = "1", validate_parameters = "TRUE"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
# of features: 10
niter: 100
nfeatures : 10
```

```
> print(importance_matrix)
        Feature        Gain       Cover   Frequency
 1:         B_7 0.501394464  0.10271038  0.07517241
 2:        D_77 0.234363142  0.26743221  0.17241379
 3:        B_23 0.110278586  0.14259843  0.11896552
 4:        B_11 0.081475790  0.13390517  0.11603448
 5:         B_1 0.023037758  0.09193481  0.08500000
 6:        B_14  0.014132134 0.06775838  0.10500000
 7:        B_15 0.012676264  0.06079362  0.10655172
 8:       D_104 0.011901479  0.06902622  0.09413793
 9:        B_36 0.005649712  0.02951791  0.06465517
10:        S_19 0.005090672  0.03432287  0.06206897
```

A continuación, se describe el resultado obtenido y el significado de cada output.

Gain: Es la métrica principal que indica la importancia de cada característica. Un valor alto de Gain para una característica significa que esa característica es más decisiva para las predicciones del modelo.

Cover: Indica la fracción de ejemplos de entrenamiento que pasan por los nodos donde se utiliza esta variable. Es decir, cuánto contribuye la variable a cubrir los datos durante el proceso de entrenamiento.

Frequency: Proporción de veces que la variable se utiliza para dividir los datos a lo largo de todos los árboles en el modelo final. Esto refleja la frecuencia con la que la variable es relevante para tomar decisiones en el proceso de predicción.

B_7 tiene el mayor Gain, lo que sugiere que es la variable más informativa para el modelo, contribuyendo significativamente a la precisión de las predicciones.

D_77 tiene un alto Cover, indicando que se utiliza ampliamente para cubrir los datos durante el entrenamiento, aunque su Gain es menor en comparación con B_7.

B_23, B_11, y B_1 también muestran contribuciones importantes (Gain relativamente altos), aunque con coberturas y frecuencias algo menores en comparación con B_7 y D_77.

El resultado tiene una gran similitud en su interpretabilidad al árbol.

4.2.3 Comparación de modelos

TRAIN	Sensitividad	Especificidad	Precision	AUC	F1	Gini	Tasa inc. 4%	M
Árbol Podado	0.902	0.739	0.820	0.853	0.859	0.706	0.811	0.758
pred_boost_train	0.892	0.771	0.832	0.895	0.861	0.789	0.901	0.845
pred_boost_opt_train	0.906	0.779	0.843	0.912	0.873	0.823	0.959	0.891
TEST	**Sensitividad**	**Especificidad**	**Precision**	**AUC**	**F1**	**Gini**	**Tasa inc. 4%**	**M**
Árbol Podado	0.900	0.739	0.819	0.851	0.858	0.701	0.809	0.755
pred_boost_test	0.889	0.769	0.829	0.892	0.858	0.783	0.896	0.839
pred_boost_opt_test	0.895	0.769	0.832	0.898	0.862	0.796	0.938	0.867

Tabla 4 – Resultados de los modelos

Métricas de rendimiento

Sensitividad y Especificidad: En términos de sensitividad (capacidad para identificar positivos verdaderos) y especificidad (capacidad para identificar negativos verdaderos), los tres modelos muestran valores competitivos.

Precisión: La precisión (proporción de predicciones positivas correctas) varía entre los modelos, pero todos muestran valores relativamente altos, pred_boost_opt tiene la precisión más alta.

AUC: El área bajo la curva ROC (AUC) es una medida integral del rendimiento del modelo, pred_boost_opt tiene el AUC más elevado, indicando que tiene mejor capacidad de discriminación entre clases.

F1 Score: El puntaje F1 combina precisión y sensitividad en una sola métrica, pred_boost_opt también muestra el F1 más alto.

Gini: El índice Gini se relaciona con la capacidad de clasificación del modelo, pred_boost_opt_train tiene el índice Gini más alto.

Tasa de incumplimiento capturada al 4%: La tasa de incumplimiento capturada al 4% mide la capacidad del modelo para identificar correctamente los incumplimientos dentro de las predicciones más importantes (predicciones con y_pred por encima del cuantil 0.96), pred_boost_opt tiene la tasa más alta, lo cual es deseable el contexto del problema.

M: Es una métrica específica que combina el índice de Gini y la Tasa de incumplimiento capturada al 4%, dándole un peso de 50% a cada una, pred_boost_opt también muestra el valor más alto.

Estabilidad de los modelos

Todos los modelos muestran una buena estabilidad general entre los conjuntos de entrenamiento y prueba, con variaciones mínimas en las métricas claves.

pred_boost_opt es el modelo que destaca por tener las métricas más altas y consistentes en ambos conjuntos, lo que sugiere una estabilidad excelente y una capacidad robusta de generalización.

La estabilidad observada en estos modelos es indicativa de que son adecuados para aplicaciones prácticas donde se requiere una buena performance tanto en datos de entrenamiento como en datos nuevos.

Selección del Modelo

Es importante que el modelo tenga buen desempeño y también que sea interpretable.

La importancia del desempeño de un modelo radica en su capacidad para ofrecer predicciones precisas y confiables basadas en los datos disponibles. Evaluar y entender el desempeño de un modelo es crucial para validar su efectividad y garantizar su utilidad en aplicaciones prácticas. Esto no solo implica métricas estadísticas como precisión, sensibilidad y especificidad, sino también consideraciones sobre la estabilidad del modelo a través de diferentes conjuntos de datos, como el conjunto de entrenamiento y prueba.

En contextos empresariales donde la transparencia y la comprensión de las decisiones son fundamentales se requiere interpretabilidad para la toma de decisiones estratégicas. La interpretabilidad facilita la identificación de patrones y reglas claras que pueden utilizarse para optimizar procesos internos o mejorar la eficiencia operativa. En industrias reguladas, donde se requiere explicar y documentar el proceso de toma de decisiones, los modelos interpretables como pueden ser preferidos para garantizar la conformidad con las normativas.

Por lo tanto, basándonos en una evaluación equilibrada de las métricas estadísticas y principalmente en la métrica M solicitada por el Sponsor y la relevancia funcional para el problema de negocio, pred_boost_opt es el modelo seleccionado como el más adecuado para este caso, porque muestra consistentemente mejores resultados en la mayoría de las métricas evaluadas en comparación con los otros modelos (Árbol Podado y pred_boost).

4.3 Descripción de Resultados

El modelo pred_boost_opt ha demostrado un rendimiento robusto en términos de métricas clave como precisión, AUC y F1-score tanto en los datos de entrenamiento como en los de prueba, lo cual indica una buena capacidad para predecir la morosidad en los clientes. Esto es crucial para las instituciones financieras, ya que les permite identificar de manera proactiva a los clientes que tienen mayor riesgo de incumplimiento de pagos.

Las predicciones de este modelo pueden ser utilizadas estratégicamente de varias maneras. Primero, permiten a la empresa focalizar sus recursos y esfuerzos en aquellos clientes identificados como de alto riesgo, implementando medidas preventivas como recordatorios de pagos, ajustes en los términos del préstamo, o incluso ofreciendo opciones de refinanciamiento para mitigar el riesgo de morosidad. Esto no solo puede reducir las pérdidas financieras debido a impagos, sino que también mejora la satisfacción del cliente al evitar cargos por pagos tardíos.

Además, el modelo puede ser integrado en los sistemas de toma de decisiones para automatizar el proceso de evaluación del riesgo crediticio, lo que agiliza los procedimientos internos y optimiza la eficiencia operativa. La capacidad de predecir la morosidad con precisión también permite a la empresa planificar mejor sus flujos de caja y reservas financieras, anticipando posibles impactos negativos en la liquidez.

Sin embargo, es importante reconocer algunas limitaciones y consideraciones prácticas del modelo. Por ejemplo, aunque el modelo puede predecir con alta precisión en el contexto actual de datos, su rendimiento puede verse afectado por cambios en las condiciones económicas, comportamientos del mercado o factores externos imprevistos. Además, la interpretación de los resultados del modelo XGBoost puede ser más compleja debido a su naturaleza lo que puede limitar la capacidad de explicar las decisiones de predicción de una manera fácilmente comprensible para los usuarios no técnicos.

En resumen, el modelo pred_boost_opt ofrece una herramienta poderosa para la gestión del riesgo de morosidad, permitiendo a la empresa tomar decisiones informadas y estratégicas que no solo protegen los activos financieros, sino que también mejoran la experiencia del cliente y la eficiencia operativa. Sin embargo, es esencial complementar su implementación con un monitoreo continuo y una evaluación de las condiciones del mercado para asegurar su relevancia y efectividad a lo largo del tiempo.

5 Distribución

5.1 Aplicación del Análisis en la Empresa

American Express, como la compañía de pagos global más grande y emisora de tarjetas de crédito, enfrenta el desafío de gestionar el riesgo de incumplimiento crediticio. Predecir el default es crucial para mitigar riesgos financieros y asegurar la estabilidad y rentabilidad de la empresa. Algunas de las aplicaciones que puede resultar de la integración del modelo son las siguientes:

- **Agilidad en la Aprobación de Créditos:** Utilizando el modelo, American Express puede automatizar el proceso de evaluación de crédito, permitiendo decisiones más rápidas y precisas en la aprobación de tarjetas y extensión de líneas de crédito. Utilizando el modelo seleccionado podrán tener una evaluación rápida con tan solo 10 variables informativas del cliente.
- **Reducción de Errores:** La automatización basada en modelos predictivos reduce errores humanos en la evaluación de créditos, mejorando la fiabilidad del proceso. Al utilizar un número reducido de variables, errores de imputación se minimizan, dando prioridad a la información del modelo.
  Adicional el modelo, realiza el cálculo de manera automática evitando procesamiento manual.
- **Asignación Eficiente de Recursos:** Con una mejor comprensión del riesgo de incumplimiento, American Express puede optimizar la asignación de límites de crédito, gestión del personal y estrategias de marketing hacia clientes con menor riesgo.
- **Reducción de la Carga de Trabajo:** Al disminuir los procesos manuales de evaluación de crédito, el personal puede enfocarse en otras tareas críticas, aumentando la eficiencia operativa.
- **Aprobaciones Más Rápidas:** Los clientes solventes pueden recibir aprobaciones de crédito más rápidas, mejorando su experiencia con la empresa.
- **Ofertas Personalizadas:** Con una mejor predicción del riesgo, American Express puede personalizar ofertas y programas de recompensas para diferentes segmentos de clientes, aumentando la satisfacción y lealtad.

5.1.1 Impacto en la Operativa Futura

Innovación y Desarrollo de Productos:

- **Nuevas Oportunidades de Mercado:** La capacidad de predecir el default con precisión permite a American Express explorar nuevas oportunidades y desarrollar productos financieros innovadores. Como por ejemplo bots automatizados que recopilen la información y den una respuesta preliminar sobre la aprobación del crédito.
- **Adaptación a Cambios en el Mercado:** Los modelos predictivos pueden identificar tendencias emergentes, permitiendo una adaptación rápida a los cambios y ajustes en las estrategias comerciales y de riesgo.

5.2 Acciones Basadas en los Resultados

5.2.1 Implementación de Mejoras en la Infraestructura de TI

Responsable: Equipo de TI de Amex con Soporte del Equipo Externo.

Acciones:

- Colaboración estrecha con el equipo interno de TI de AMEX para garantizar la adecuada implementación y mantenimiento de la infraestructura de soporte para el modelo.
- Asistencia en la optimización de la infraestructura de almacenamiento y procesamiento de datos para manejar eficientemente las operaciones del modelo.
- Implementación de medidas de seguridad y cumplimiento de normativas bajo la guía del equipo de TI de AMEX.

5.2.2 Comunicación y Formación de Equipos Interdepartamentales

Responsable: Equipo de Amex.

Acciones:

- Organización de reuniones periódicas entre el equipo de consultores externos, analistas de datos, equipos de TI y líderes de AMEX para alinear estrategias y revisar progresos.
- Facilitación de sesiones de formación y capacitación para el personal interno sobre el uso adecuado de los resultados del modelo predictivo.
- Establecimiento de canales de comunicación efectivos para la retroalimentación continua y la resolución proactiva de problemas.

5.2.3 Ajuste de Políticas de Crédito

Responsable: Departamento de Gestión de Riesgos Crediticios

Acciones:

- Ajustar las políticas de crédito en función de las predicciones del modelo, incluyendo la modificación de límites de crédito y la implementación de programas de asistencia para clientes de alto riesgo.

5.2.3 Monitoreo y Reporte de Resultados:

Responsable: Equipo de Analistas de Datos de Amex.

Acciones:

- Desarrollo de dashboards interactivos para el monitoreo en tiempo real de las predicciones del modelo y métricas claves.
- Generación de informes periódicos sobre el rendimiento del modelo, recomendaciones de ajuste y próximos pasos.

- Implementación de sistemas de alerta temprana para identificar posibles desviaciones y acciones correctivas oportunas.

5.3 Implementación del Modelo

Se propone una estrategia de implementación que combine tanto el procesamiento en tiempo real, adaptándose a las necesidades específicas de AMEX y las características del problema de predicción de incumplimientos en tarjetas de crédito.

5.3.1 Implementación en Tiempo Real

Esta fase implica la integración directa del modelo en los sistemas operativos del cliente, permitiendo decisiones instantáneas basadas en las predicciones del modelo.

Ventajas:

- **Respuesta ágil**: Permite abordar rápidamente eventos críticos como clientes muy riesgosos.
- **Gestión proactiva de riesgos**: Facilita la identificación y mitigación temprana de comportamientos financieros riesgosos.
- **Optimización de decisiones financieras**: Mejora la capacidad del cliente para tomar decisiones informadas en tiempo real, reduciendo potenciales pérdidas por incumplimientos.

Requisitos:

- Infraestructura tecnológica robusta para el procesamiento y almacenamiento en tiempo real.
- Implementación de sistemas de monitoreo continuo y alertas tempranas.

5.3.2 Monitoreo y actualización del modelo

Se continuará con la ejecución análisis periódicos, sobre la base de datos, facilitando análisis retrospectivos y la generación de informes programados sobre riesgo crediticio para ver el impacto del modelo.

De manera trimestral, se ajustará nuevamente el modelo con la nueva base para asegurar que refleja la realidad del mercado.

5.4 Evaluación del Éxito del Proyecto

Para evaluar el éxito del proyecto se establecieron métricas de rendimiento que se detallan a continuación:

5.4.1 Exactitud del Modelo Predictivo

Métricas de rendimiento específicas:

- **Tasa de precisión (Accuracy)**: La proporción de predicciones correctas sobre el total de predicciones realizadas.

- **AUC-ROC (Área bajo la curva ROC)**: Una medida del rendimiento del modelo, evaluando la capacidad de distinguir entre clientes que incumplen y los que no.
- **Métrica M:** Media del Coeficiente Gini Normalizado (G) y la tasa de incumplimiento capturada al 4% (D). Combina tanto la capacidad predictiva del modelo (Gini) como su capacidad para capturar incumplimientos en las predicciones más importantes (Sensibilidad/Recall).

  $M=0.5 \cdot (G+D)$

5.4.2 Impacto Financiero

Métricas de evaluación:

- **Reducción de costos**: Evaluar la disminución de costos asociados con la gestión de incumplimientos y la recuperación de deudas.
- **Aumento de ingresos**: Medir el incremento de ingresos debido a la mejora en la gestión de riesgos y la optimización de las políticas de crédito.
- **Análisis de retorno de inversión (ROI)**: Comparar los beneficios financieros obtenidos con la inversión realizada en el desarrollo e implementación del algoritmo.

5.4.3 Mejora en la Eficiencia Operativa

Métricas de evaluación:

- **Reducción del tiempo de procesamiento**: Medir la disminución en el tiempo necesario para evaluar el riesgo de incumplimiento de los clientes.
- **Tiempo de respuesta**: Medir la mejora en el tiempo de respuesta a las solicitudes de crédito.

5.4.4 Mejora en la Satisfacción del Cliente

Criterios de evaluación:

- **Encuestas de satisfacción**: Realizar encuestas a los clientes para evaluar la percepción de la gestión del crédito y la comunicación relacionada con el riesgo de incumplimiento.

5.4.5 Validación y Mejora Continua

Criterios de evaluación:

- **Feedback y mejora continua**: Establecer mecanismos para recopilar feedback de los usuarios del modelo y realizar mejoras continúas basadas en el mismo.

5.5 Reflexión sobre Debilidades y Mejoras

5.5.1 Debilidades

Falta de Conocimiento Específico de las Variables:

Una de las principales debilidades es que, aunque se conozcan las categorías a las que pertenecen las variables (D_, *S_*, P_, *B_*, R_*), no se cuenta con información detallada sobre cada variable individual. Esta falta de detalle limita la capacidad para interpretar adecuadamente los resultados y entender los factores específicos que contribuyen al incumplimiento.

El desconocimiento de las características exactas de las variables, impide la realización de análisis detallados sobre el impacto individual en el riesgo de incumplimiento. Esto también dificulta la identificación de variables altamente predictivas y la optimización del modelo.

Falta de Hardware Adecuado:
El equipo externo no dispone de hardware adecuado para procesar grandes bases de datos, lo que puede afectar la eficiencia y la efectividad del análisis.

Reducción Inicial de la Base de Datos:
La reducción inicial de la base de datos puede generar pérdida de información valiosa, afectando la calidad y precisión del análisis posterior.

Conocimiento Limitado del Negocio:
El equipo externo no tiene acceso directo a la empresa, lo que conlleva una falta de conocimiento profundo del negocio e impide su participación directa en las últimas fases del proyecto.

Falta de Comunicación con el Sponsor:
No existe comunicación fluida durante el proyecto con el sponsor, lo cual puede llevar a una falta de alineación con los objetivos y expectativas del mismo.

Ajuste del modelo
El modelo requiere ajustes periódicos para mantenerse preciso y relevante a medida que cambian los comportamientos de los clientes y las condiciones del mercado.

5.5.2 Propuestas de Mejora

Obtener Información Detallada de las Variables:
Trabajar para obtener una descripción más detallada de cada variable en el conjunto de datos. Esto podría implicar colaboraciones con los proveedores de datos o la realización de investigaciones adicionales para entender mejor cada categoría.

Con un conocimiento más profundo de las variables, se pueden realizar análisis más específicos y precisos, identificar factores clave de riesgo y mejorar la interpretación de los resultados.

Comunicación con el sponsor:
Generar espacios de revisión periódicos con el sponsor para aclarar dudas y alinear el trabajo.

Hardware:
Replicar el análisis con un Hardware con mayor poder de procesamiento para validar los resultados o ampliar el modelo.

6 Bibliografía

[1] American Express - Default Prediction | Kaggle. Disponible en: https://www.kaggle.com/competitions/amex-default-prediction/overview

[2] AMEX EDA which makes sense. Disponible en: https://www.kaggle.com/code/ambrosm/amex-eda-which-makes-sense/input

[3] AMEX - final project EDA&features selection (RFE). Disponible en: https://www.kaggle.com/code/yuhsinchu/amex-final-project-eda-features-selection-rfe

Anexo 1

Análisis univariado de variables restantes:

B_1

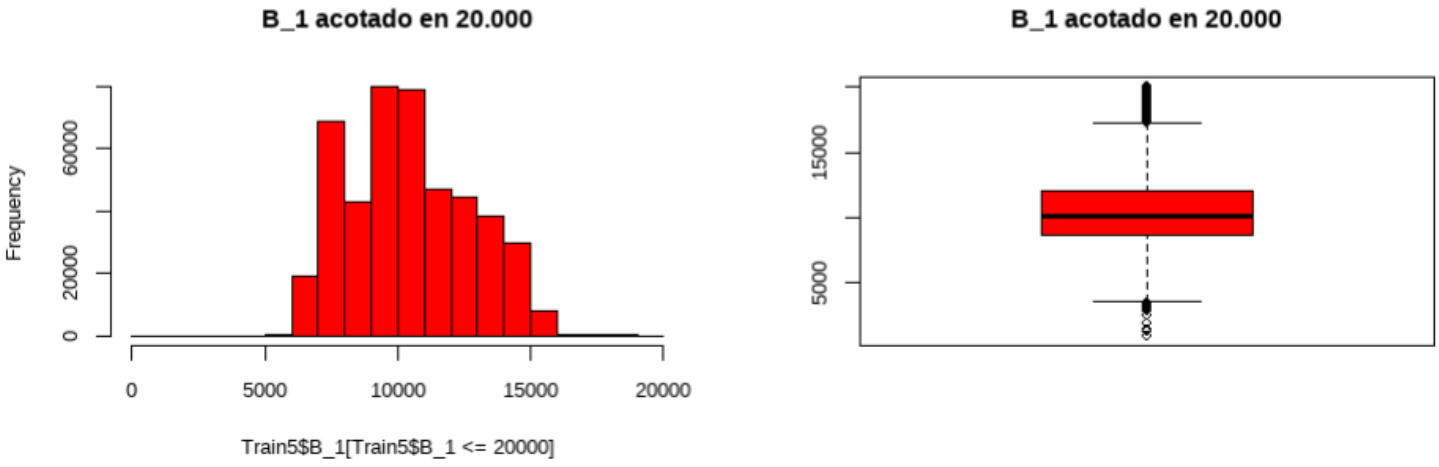

Figura 10 – Análisis univariado B_1

B_36

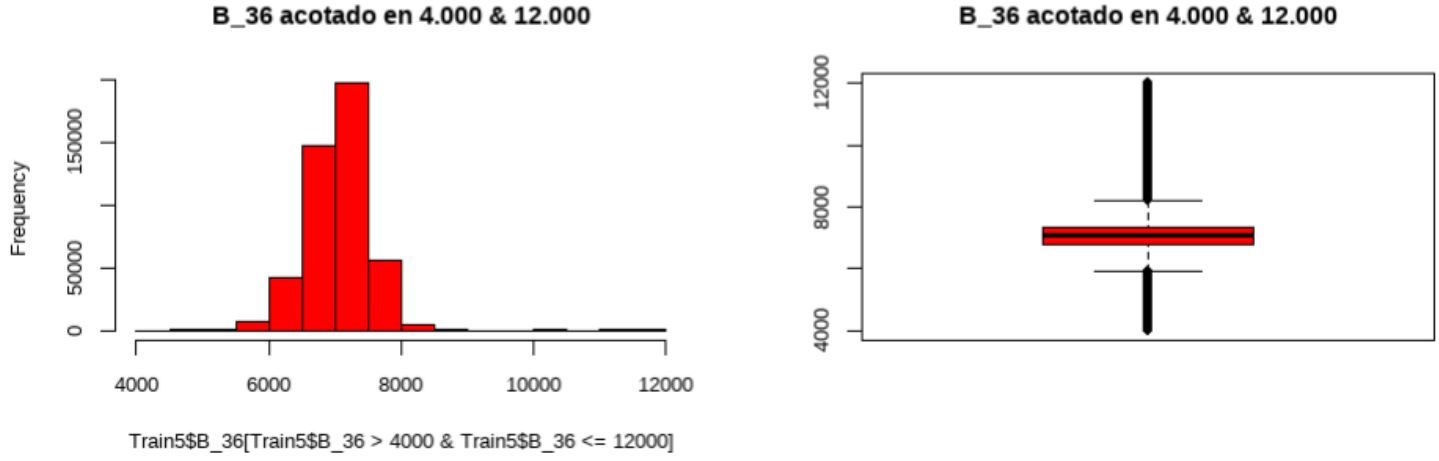

Figura 11 – Análisis univariado B_36

B_15

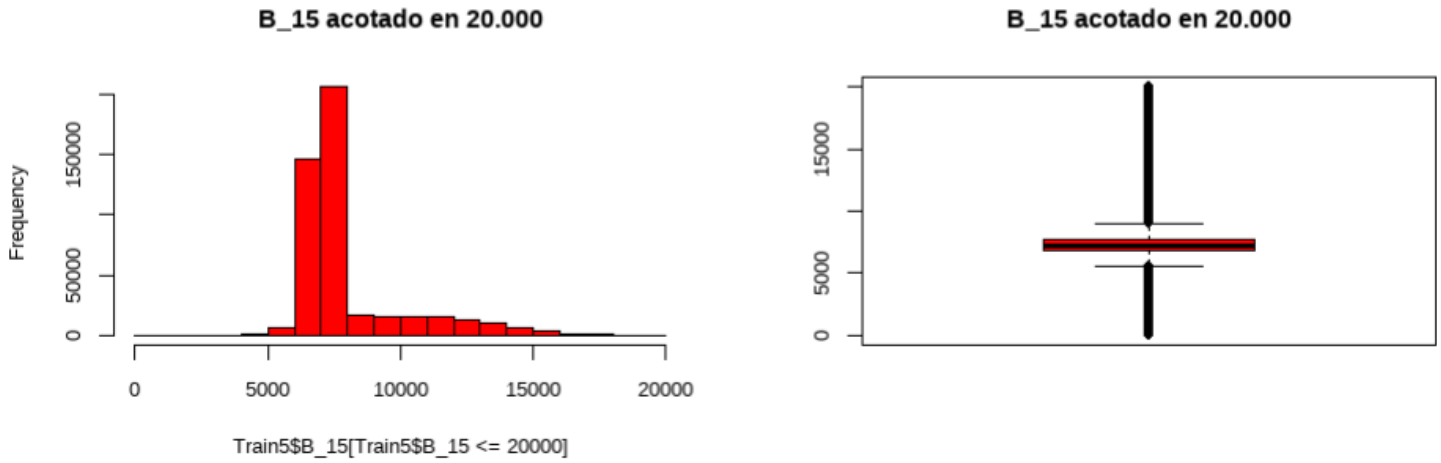

Figura 12 – Análisis univariado B_15

S_19

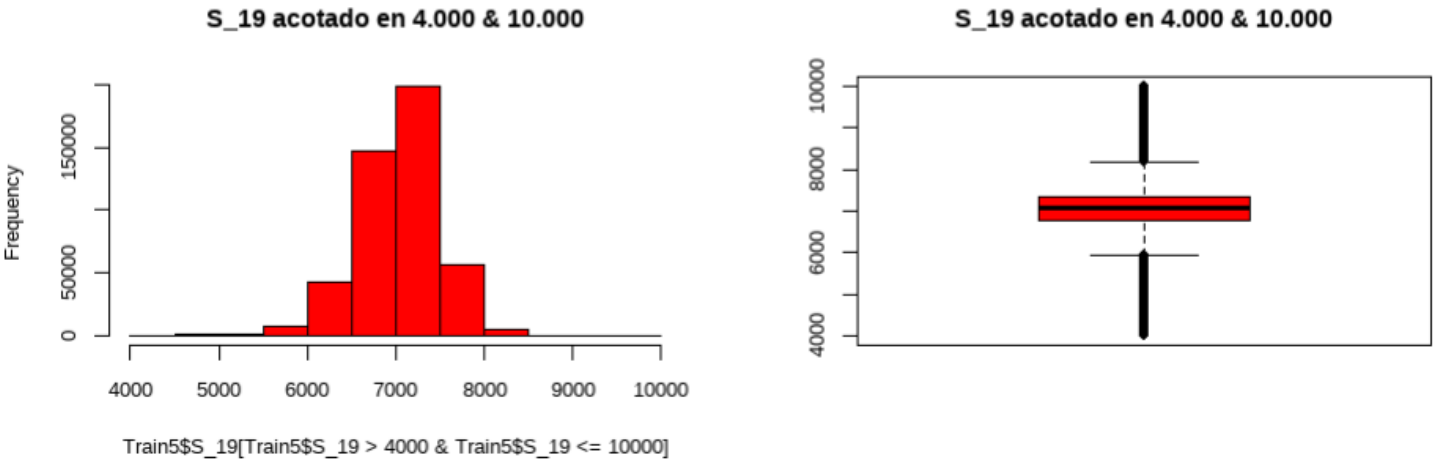

Figura 13 – Análisis univariado S_19

B_14

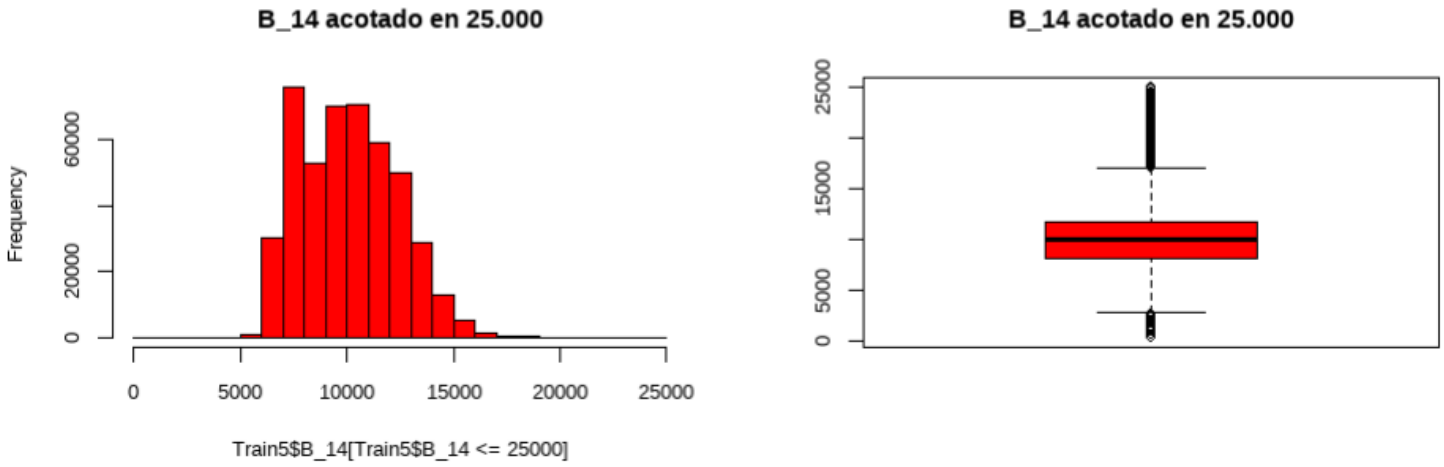

Figura 14 – Análisis univariado B_14

D_104

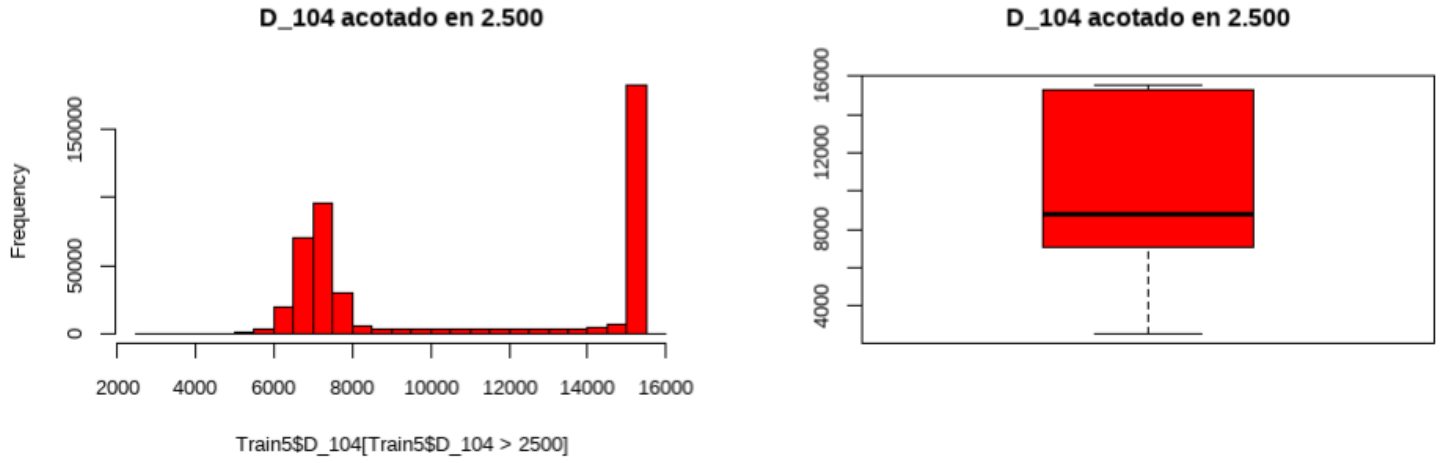

Figura 15 – Análisis univariado D_104