*Aprendizaje de Máquina*

*Caso Machine Learning – Primer Semestre 2025*

Taller 2 grupal

Rafael Eduardo Torregroza Vidal , Andres Felipe Useche

JUAN CARLOS GARCIA DIAZ

*Estudiante de Ciencia de Datos*

*Profesor, Departamento de Ingeniería Industrial*

*Pontificia Universidad Javeriana, Bogotá, Colombia*

**BUSINESS UNDERSTANDING**

**Determine Business objectives**

Background: Tenemos una compañía interesada en conocer qué factores influyen en la deserción de un cliente y de qué manera lo hacen, con el fin de ajustar un modelo capaz de predecir si un cliente va a desertar o no. Además, buscan identificar qué estrategias son más efectivas para retener a los clientes. Esto se debe a que, en el sector de las telecomunicaciones, está comprobado que resulta más fácil y rentable mantener a los clientes actuales que captar nuevos. Por lo tanto, es de suma importancia entender cómo retener a los clientes existentes y por qué algunos deciden abandonar la compañía, para así poder evitarlo.

Business goal: Objetivos de negocio

1. Encontrar una estrategia que permita evitar que los clientes deserten de la operadora
2. Encontrar una estrategia que permita retener los clientes

Business success criteria:

1. Precisión del modelo predictivo ≥ 85 %: en la identificación de clientes que probablemente van a desertar.
2. Reducción del churn (tasa de deserción) en al menos un 15 %: dentro de los primeros seis meses tras implementar las estrategias sugeridas por el modelo.
3. Incremento en la tasa de retención de clientes en un 10 % anual.
4. Identificación de al menos tres factores clave que influyen significativamente en la deserción de clientes, con evidencia cuantitativa.

**Determine Data mining goals**

Data mining goal:

1. Desarrollar un modelo de clasificación capaz de predecir con alta precisión si un cliente va a desertar o no, utilizando variables históricas y del cliente disponibles en los datos.
2. Encontrar patrones en el comportamiento de los clientes
3. Identificar las variables más relevantes que influyen en la deserción de los clientes mediante técnicas de análisis de características (feature importance).

Data mining success criteria:

1. El modelo predictivo alcanza una **AUC ≥ 0.85** y un **F1-score ≥ 0.80** en datos de prueba, lo que indica buena capacidad de discriminación y equilibrio entre precisión y recall.

**DATA UNDERSTANDING**

**Descripción de las variables:**

Campos de la base de datos id - id anónimo que identifica al cliente.

Fecha de nacimiento - Fecha de nacimiento del cliente, dd/mm/yyyy.

tipo cliente - Tipo de cliente: 1 - Hombre, 2 - Mujer, 3 - Cliente empresarial.

Factura online - Binario que indica si el cliente recibe únicamente su factura online.

Antigüedad equipo - Meses de antigüedad del equipo.

Plan de datos - Binario. 1 = el cliente tiene plan premium de datos.

Facturación - Total facturación de los últimos seis meses (suma).

Mora - Días de mora acumulados.

Fecha inicio contrato - Fecha de inicio de su primer contrato de plan pospago en la compañía.

minutos - total minutos consumidos en los últimos seis meses.

resultado - 0 = No hubo deserción; 1 = Hubo deserción (churn)

**Tipos de datos de las variables:**

Double: id, tipo cliente, Factura online, Antigüedad Equipo, Plan de datos, facturación, mora, minutos, resultado

Date: Fecha de nacimiento, Fecha inicio contrato

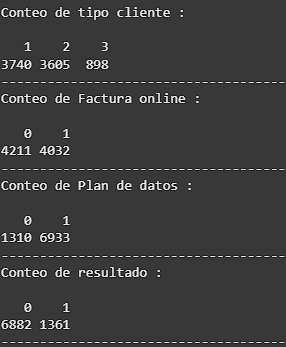
**Dimensiones de las bases de datos:**

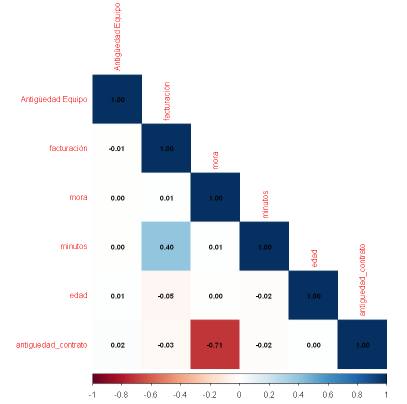
Base de datos Testelco: Cuenta con 12 columnas con 8,243 registros.

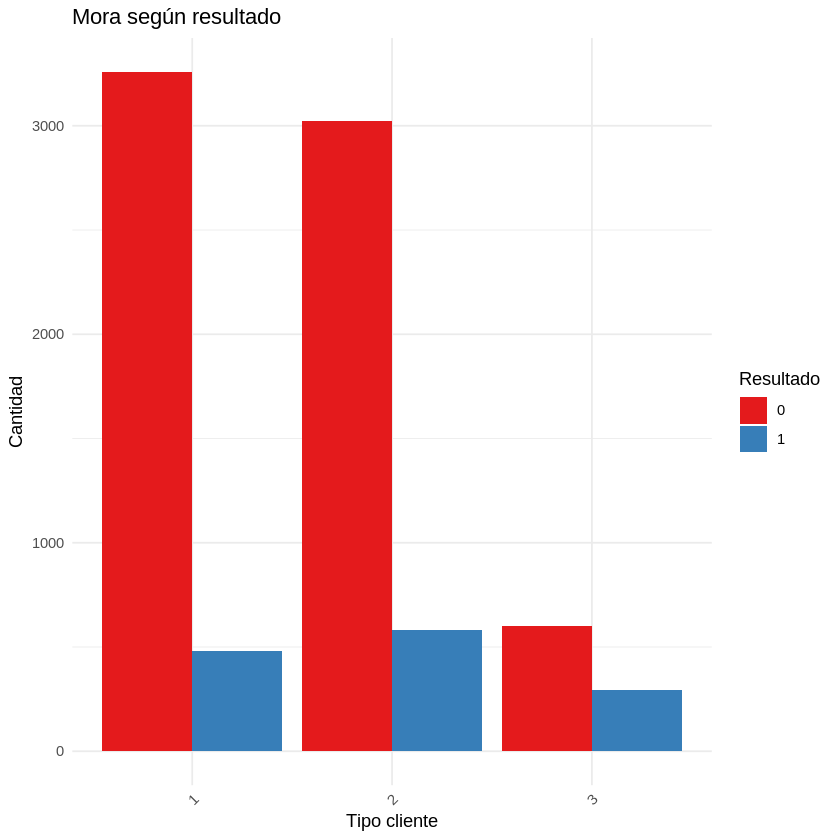
Base de datos Traintelco: Cuenta con 11 columnas(Ya que la columna resultado es la variable a predecir), con 2000 registros.

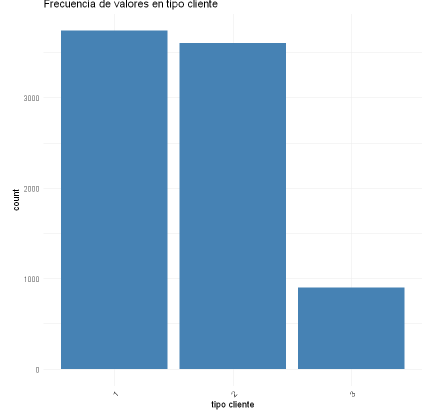
**Algunas Gráficas sobre el análisis descriptivo**

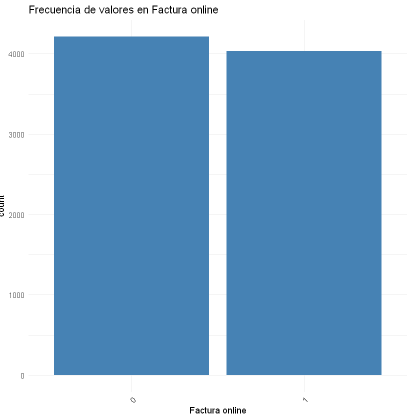
Estas gráficas corresponden únicamente a las variables de la base de datos Traintelco

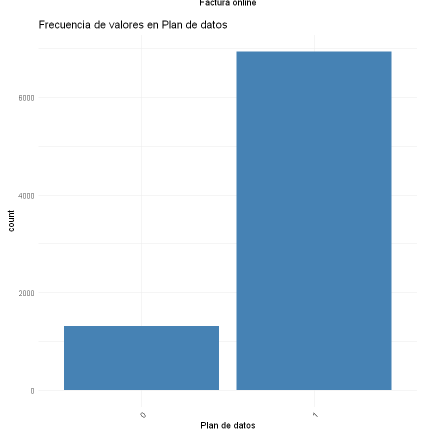


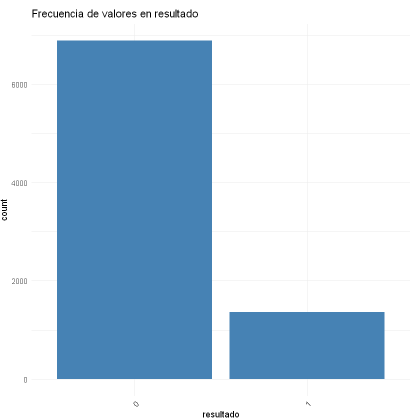


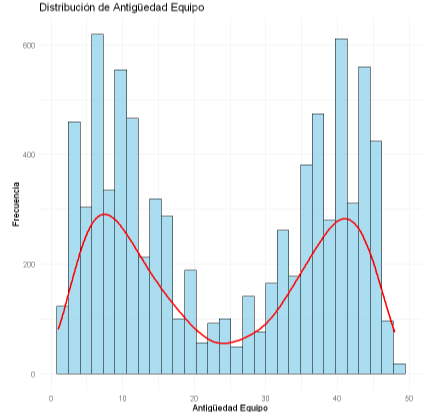


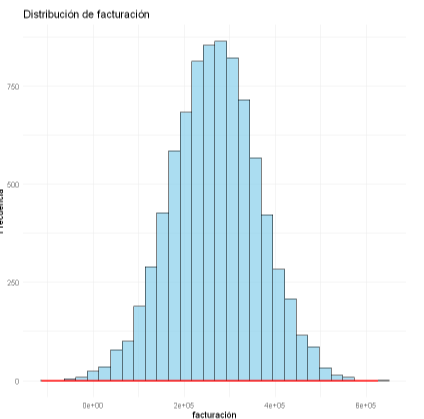












**DATA PREPARATION**

Select data:

**Valores nulos:**

No se encontraron valores nulos

**Outliers**

Antigüedad Equipo → 0 outliers detectados

facturación → 51 outliers detectados

mora → 41 outliers detectados

minutos → 101 outliers detectados

edad → 78 outliers detectados

antiguedad\_contrato → 159 outliers detectados

Construct data: Las variables: tipo cliente, Factura online, Plan de datos, resultado. fueron cambiadas a tipo factor para el análisis dado que son variables categóricas.

La variable Fecha de nacimiento fue tratada para crear una columna de “edad” restando la fecha actual a la de nacimiento,la unidad de medida es en años. Finalmente la columna de Fecha de nacimiento se eliminó.

La variable Fecha inicio contrato se transformó de la misma manera para obtener la cantidad de meses que el cliente lleva en contrato con la operadora(antigüedad contrato), la unidad de medida está en meses. Finalmente fecha inicio contrato fue eliminada.

Se elimina la columna ID dado que no sirve para el análisis..

**MODELING & EVALUATION**

Select modeling techniques:

Para modelar los datos se utilizarán las técnicas de: Bagging, Random Forest, GBM, XG Boost, Light GBM y regresión logística. Teniendo en cuenta que es un problema de clasificación para una variable binaria cualquier de los modelos se puede utilizar.

Generate test design:

Para entrenar los modelos, los datos de entrenamiento se dividirán en un 80-20 % para entrenar y validar el entrenamiento, y se planea utilizar grid search en la medida de lo posible para probar que parámetros dan mejores resultados para cada modelo, de esta manera podemos a su vez evitar el sobreajuste. El desempeño de las técnicas se va a comprar mediante la curva ROC y el AUC. De cada grid search en los modelos se tomará aquellos parámetros con el mejor resultado, este modelo luego será comparado con los mejores resultados de los demás modelos.

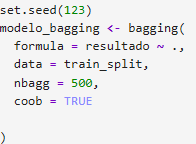
Build model:

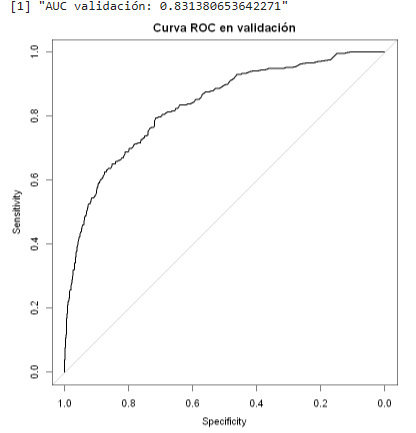
Para el modelo de Bagging se hará grid con el número de árboles. Para el modelos de random forest se considerará la cantidad de Mtry (Número de variables que se seleccionarán en cada partición de cada árbol). Para el caso de GBM se tendrá en cuenta el número de árboles, la profundidad y el ratio de aprendizaje. Para XG Boost se considerara: número de árboles, profundidad, ratio de aprendizaje, penalización por dividir nodos, Fracción de columnas (features) para cada árbol, Mínimo de observaciones en un nodo hoja y Fracción de observaciones para cada árbol. Para LightGBM se cambia la clase objetivo a numeric y se tendrá en cuenta: el número de hojas, el ratio de aprendizaje y la profundidad máxima de cada árbol. Finalmente para la regresión logística no hay forma de hacer grid search.

Assess model:

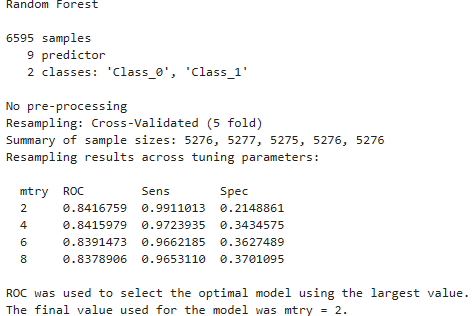
En dos de los siguientes modelos probamos el comportamiento de estos al entrenarlos con el data frame con y sin outliers y los modelos obtuvieron mejores métricas con outliers, lo cual significa que estos datos son importantes y todos los modelos se entrenaron incluyendo estos valores.

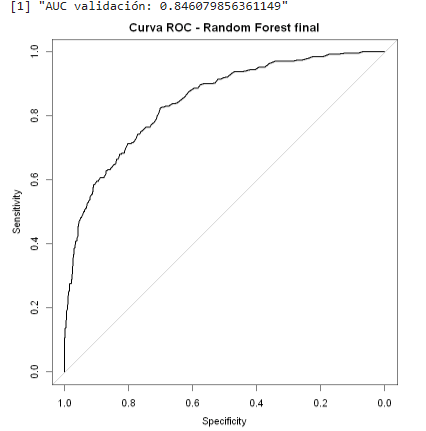
Mejor modelo Bagging:



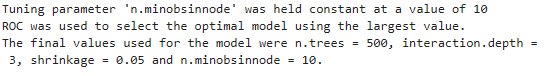


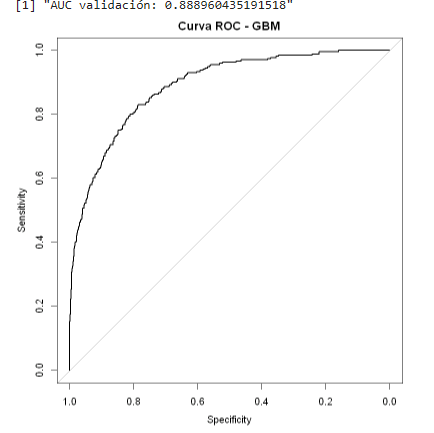
Mejor modelo de RF:



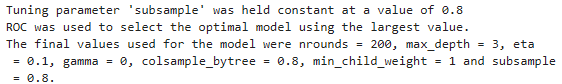


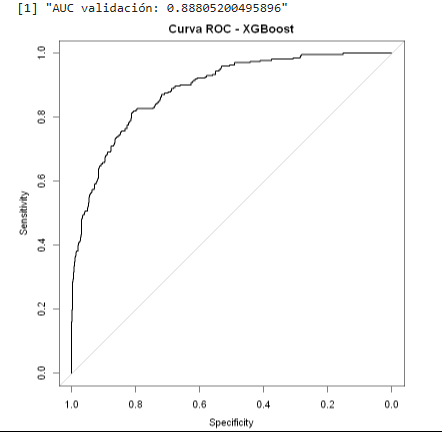
Mejor modelo GBM:



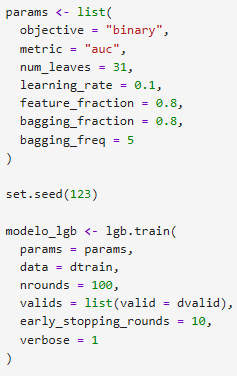


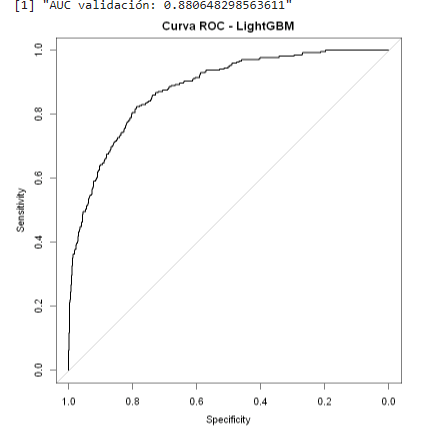
Mejor modelo de XG Boost:



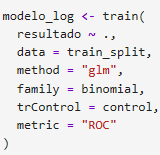


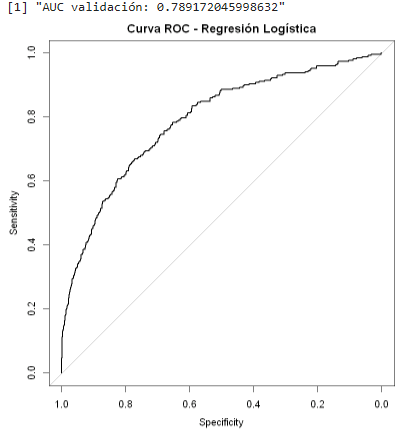
Mejor modelo LightGBM:





Regresión logística:





**DEPLOYMENT**

### **Estrategias Sugeridas:**

#### **Objetivo 1: Evitar desertores**

* Implementando el modelo de predicción de GBM el cual fue el modelo con los mejores resultados, predecir qué clientes están en riesgo de desertar para poder darles atención inmediata y evitar que abandonen la operadora, en tiempo real

#### **Objetivo 2: Retener a los clientes actuales**

* Crear campañas personalizadas a los clientes según el uso de su plan en cuanto a costos y minutos utilizados, para poder mantenerlos fieles a la operadora

### 

**Evaluación del impacto:**

* Cada estrategia será evaluada según cuanto afecten estas en la reducción de clientes que se retiran, o que tanto mantienen la los clientes tasa de retención.
* Las simulaciones pueden actualizarse en función de la tasa de conversión esperada en campañas de retención.

### **Siguientes pasos de análisis recomendados**

1. **Monitoreo continuo del modelo en producción**
   * Validar que el modelo siga teniendo buen desempeño (AUC > 0.85) al recibir nuevos datos.
   * Detectar cambios en el comportamiento de los clientes que puedan llegar a afectar las predicciones del modelo.
2. **Análisis del ciclo de vida del cliente**
   * Identificar patrones según el tiempo de permanencia para desarrollar estrategias por fase del ciclo de vida de los clientes.