# DIPLOMATURA DATA SCIENCE

GRUPO 6: VELAZQUEZ, GARCIA, PIZARRO

MODULO 1: EDA

# INTRODUCCIÓN

- En esta presentación analizaremos un caso de uso de Machine Learning implementándolo en un Modelo de Churn (tasa de abandono de los clientes).

# SITUACIÓN

- Una empresa de telefonía prepaga necesita predecir, dentro de el conjunto de sus clientes, cuál de ellos dejará de realizar recargas a su línea telefónica en las próximas 4 semanas.

# OBJETIVO

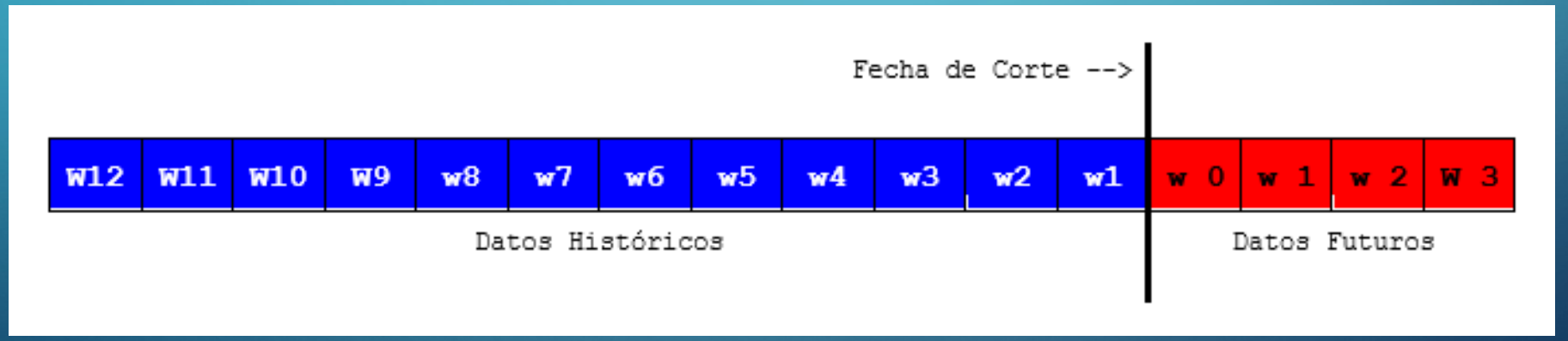
- Lo que se pretende es saber qué clientes no realizarán una recarga en el período de tiempo definido anteriormente para saber cuales son los que tienen mayor porcentaje de dejar de utilizar el servicio de la empresa telefónica.

- Es de utilidad para aplicar diferentes políticas de retención y de marketing para que los clientes permanezcan con la empresa.

# DATOS

Datos 12 semanas

Datos 4 semanas posteriores



# DATOS DISPONIBLES

**Dataset**
226 CAMPOS

- Datos del cliente
- Tiempos
- Montos de paquetes mensuales
- Cantidades de paquetes mensuales
- Montos de paquetes semanales


# LIMPIEZA DE DATOS

En el proceso de limpieza se realizo:

- En cuanto a nulos: Eliminamos datos nulos que formaban parte de los datos categóricos network tech, device model name y device vendor name. Estos datos contenían valores NaN y “NOT_IDENTIFIED”.
- En cuanto a incorrectos:

  Se limpiaron datos sobre las cantidades semanales y acumulados mensuales de pack de datos, sms y voz. Estos datos eran incorrectos debido a una mala recolección de los datos por parte del área de DB.

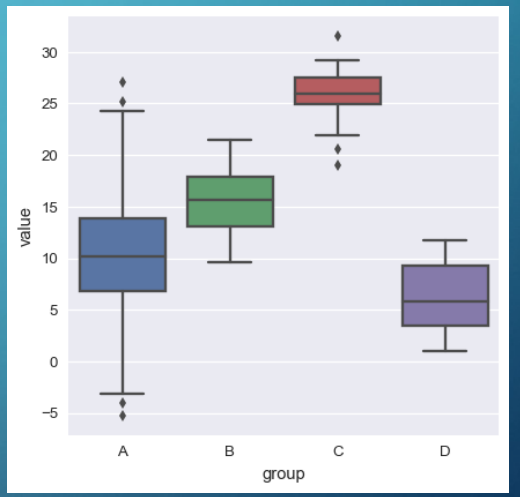
# LIMPIEZA DE DATOS

En el proceso de limpieza se realizo:

- En cuanto a outliers:

Encontramos outliers  para el trafico de datos y de voz.

Para imputar los outliers dependiendo cuan distribuidos uniformente estaban los datos utilizamos la media o la mediana. Para definir el rango interquartil (IQR) usamos el Q1 y Q3, que luego de multiplicarlos por 1.5 nos definirán los valores que se consideran como outliers.



# LIMPIEZA DE DATOS

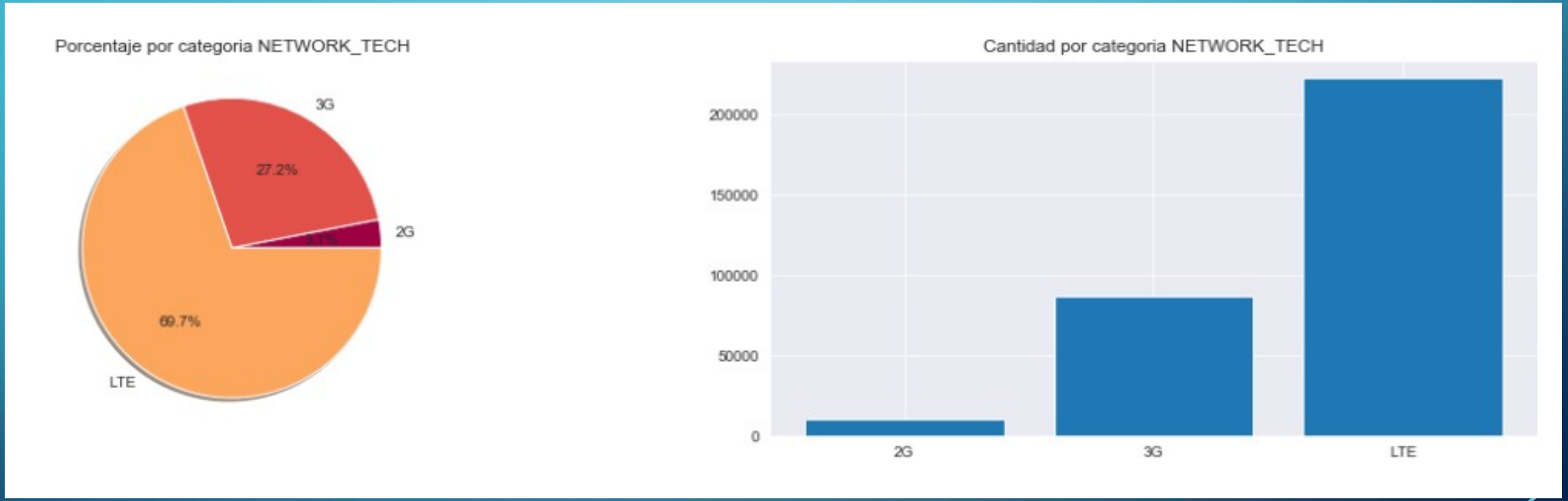
En el proceso de limpieza se realizo:

- En cuanto a duplicados:


Encontramos información que es duplicada que puede ser obtenida haciendo cálculos sobre las columnas, por ejemplo para el caso de los montos de recargas y los packs el desglose a nivel semanal al sumarlo nos daba la columna de su acumulado mensual, por lo que decidimos quitar estas columnas para no repetir datos.

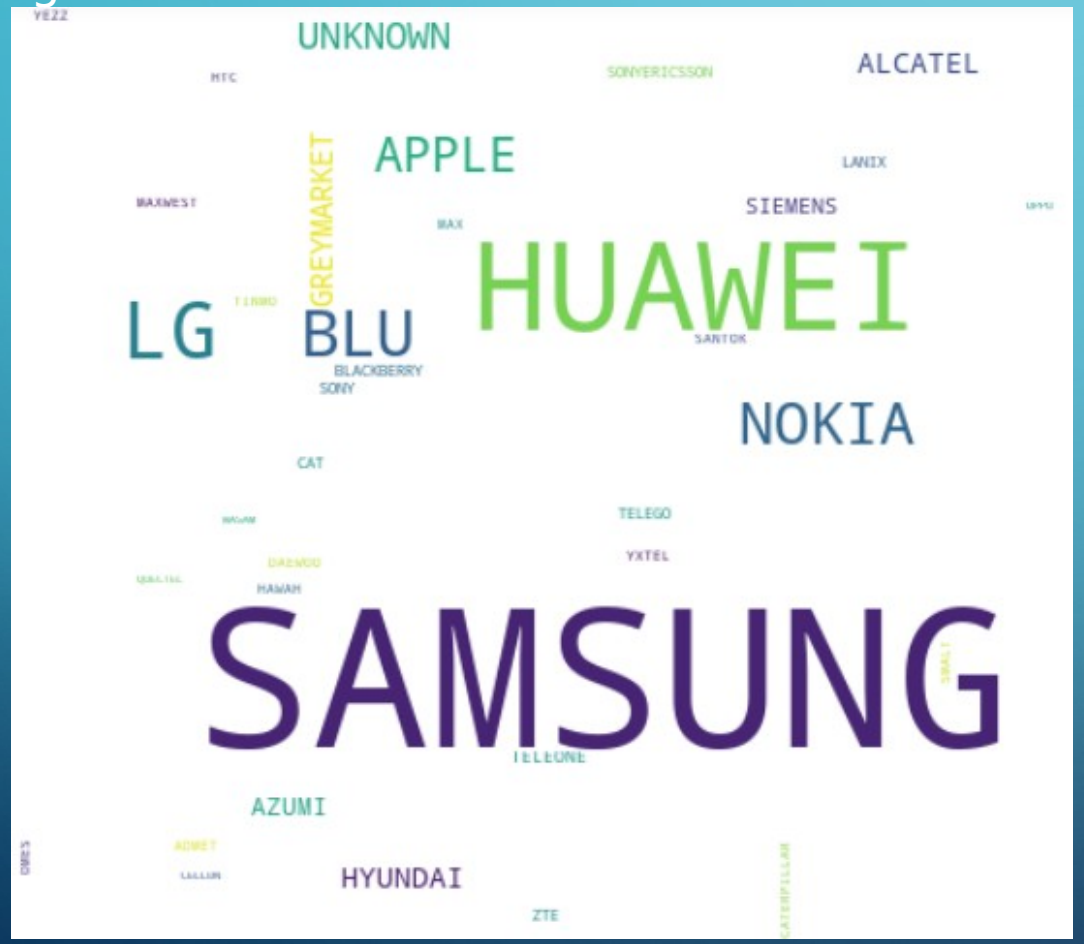
# ANÁLISIS DE LOS DATOS

NETWORK_TECH: para esta categoría vemos que son muchos mas las cantidades de líneas que utilizan la tecnología LTE casi con un 70% del total de las líneas analizadas en esta caso.



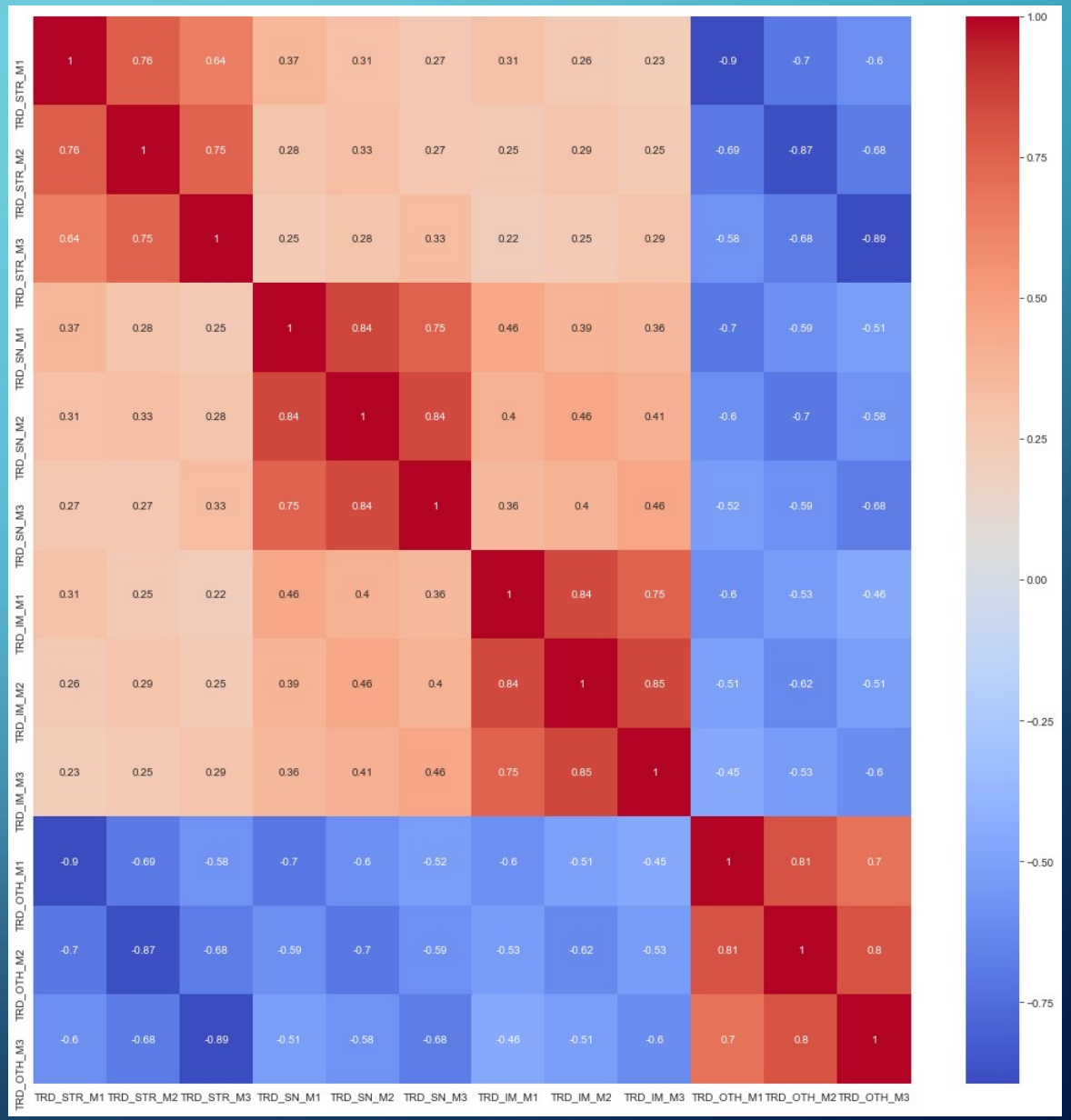
# ANÁLISIS DE LOS DATOS

DEVICE MODEL NAME y DEVICE VENDOR NAME: para esta categoría vemos que las mayoría de las líneas tienen un dispositivo marca Samsung con el modelo Galaxy grand.



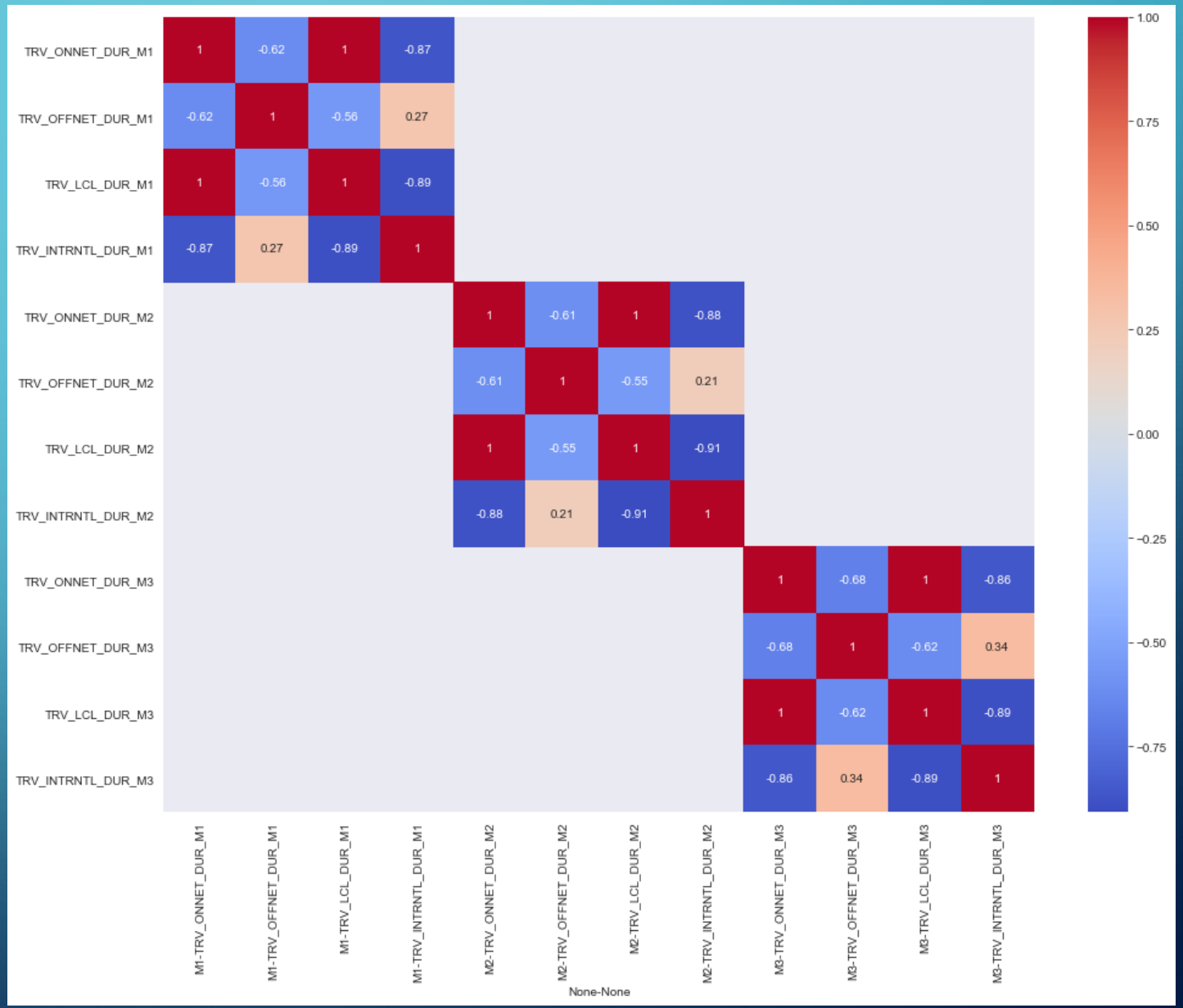

# ANÁLISIS DE LOS DATOS

Trafico de datos: para el trafico de datos analizamos la correlación que tienen agrupados por tipo a través de las semanas en la que tuvimos los datos para analizarlos.



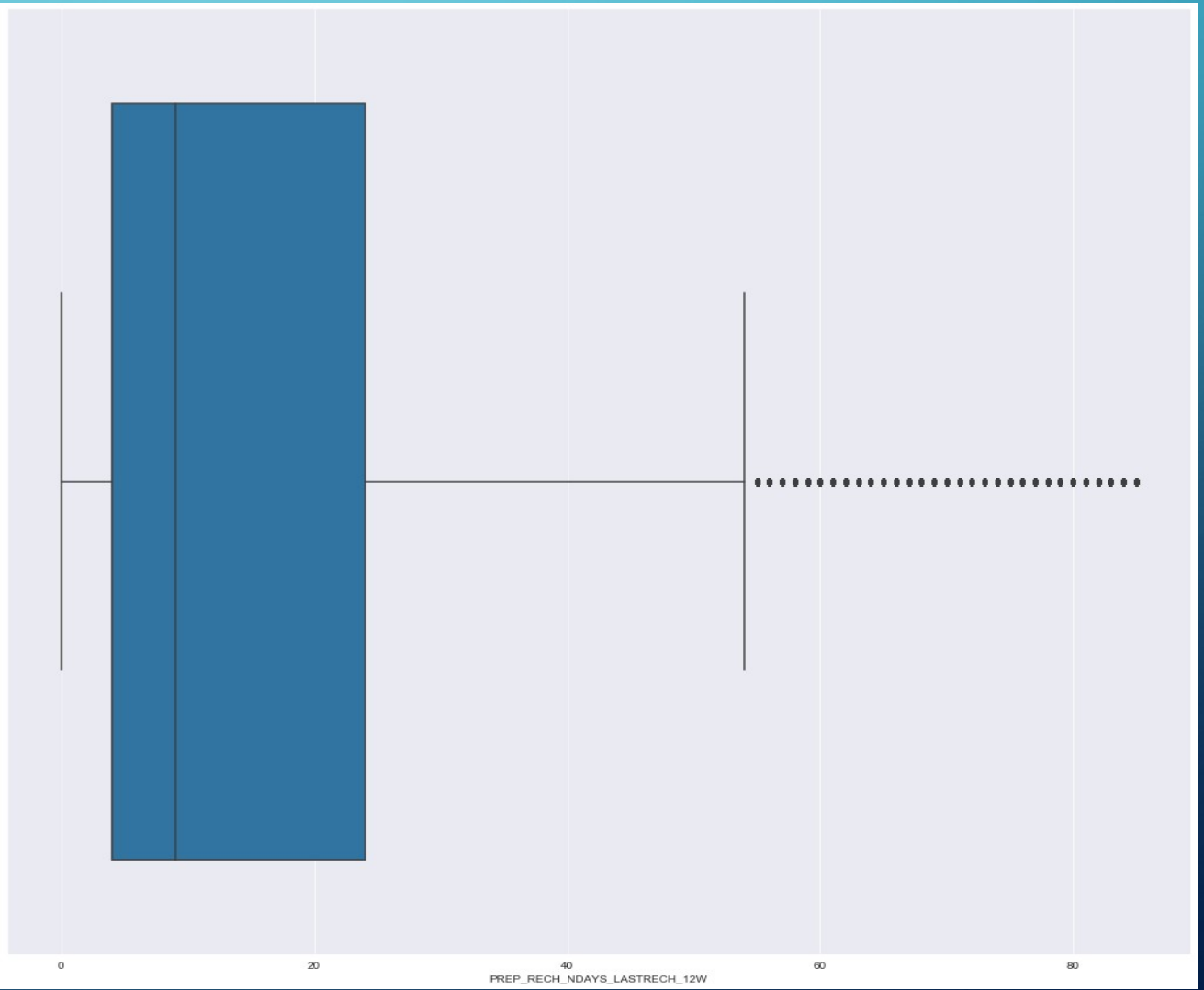
# ANÁLISIS DE LOS DATOS

Llamadas de voz: Hay una fuerte correlación entre las llamadas de voz por mes, esto nos llevo a poder agrupar las llamadas en diferentes columnas mensuales.



# ANÁLISIS DE LOS DATOS

Días desde la ultima recarga:
Observamos que los días desde la ultima recarga tienen sus valores concentrados entre 0 y 25, estos datos nos sirven además para conocer sobre los clientes en cuanto a su comportamiento.



# FEATURES PARA EL MODELO


# ELECCIÓN DEL MODELO

- Se aprobaron 4 modelos de Machine Learning para la clasificacion:

El modelo elegido es  xgboost, y las variables utilizadas son las referidas a los datos de clientes, datos sobre packs, sobre trafico de datos y de recargas.

El modelo muestra diferentes métricas que nos indican cuan preciso es, además de que luego de ser entrenado nos muestra el porcentaje de aciertos y desaciertos con su respectivo error.

- En cada caso se obtuvieron las métricas:

Matriz de confusión, Curva ROC –AUC, Accuracy, Precision – Recall – F1

# ELECCIÓN DEL MODELO

- Decision XGBoost:

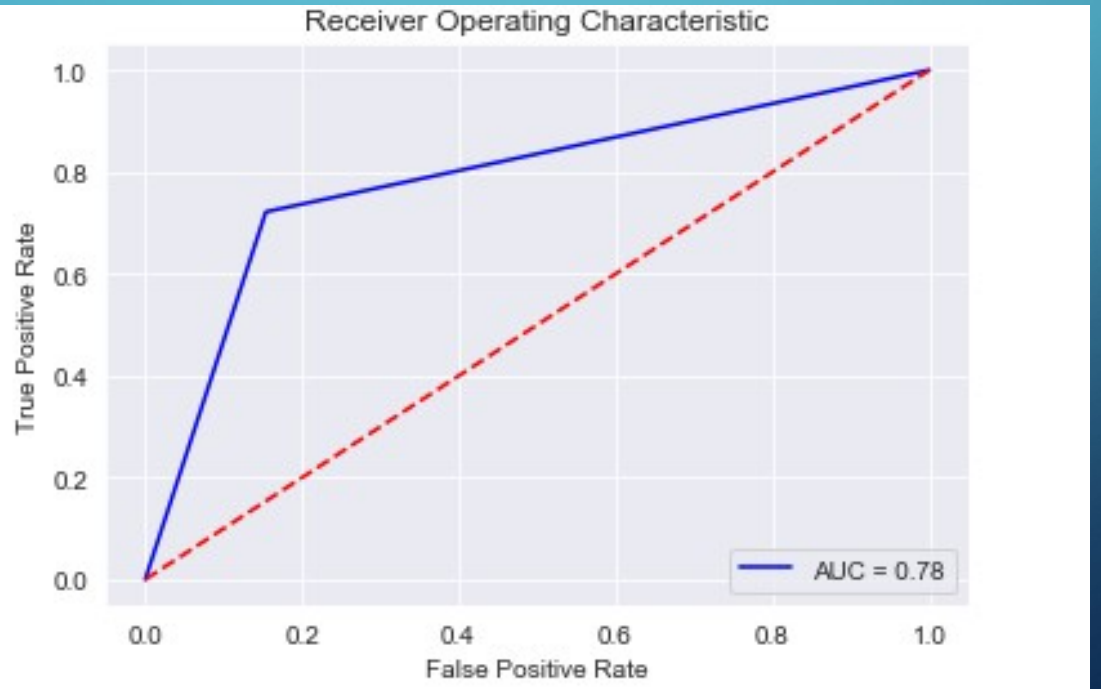
Reporte de Clasificac

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.85 | 0.85 | 37790 |
| 1 | 0.71 | 0.72 | 0.71 | 19539 |
| accuracy | | | 0.80 | 57329 |
| macro avg | 0.78 | 0.78 | 0.78 | 57329 |
| weighted avg | 0.80 | 0.80 | 0.80 | 57329 |

Matriz de confusión

```
[[31974  5816]
 [ 5438 14101]]
```

ROC



FIN

# MUCHAS GRACIAS !