CODER HOUSE

TRABAJO FINAL

Arango, Johansen & Piana
Data Science - Comisión 29740

PROBLEMA ESPECÍFICO

Necesidad de predecir retrasos en la aviación comercial de Estados Unidos.

OBJETIVO DE LA INVESTIGACIÓN

Mejorar los servicios prestados a los pasajeros mediante la predicción de retrasos; conociendo la aerolinea que opera el vuelo, el día de la semana en la que está programado, su aeropuerto de origen y aeropuerto destino.

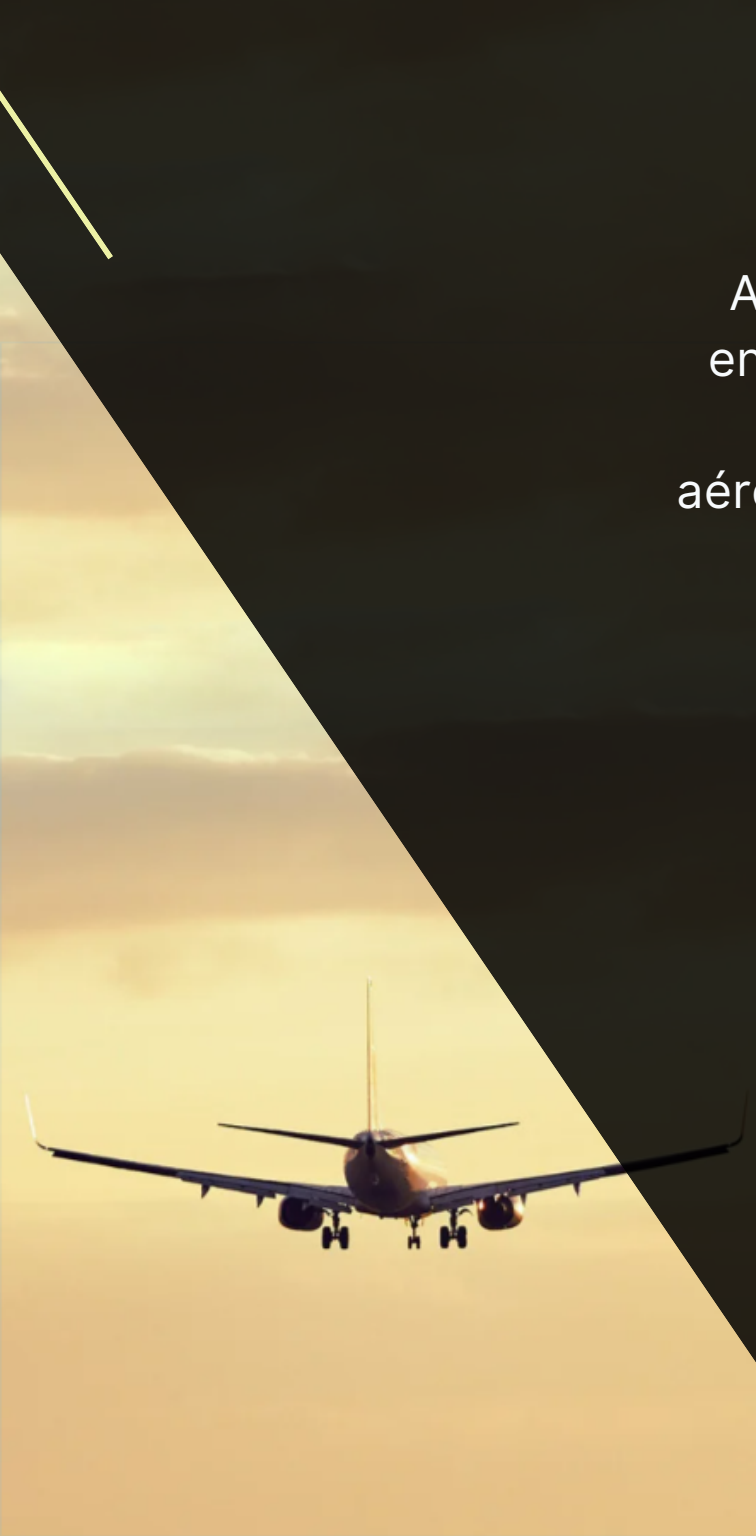
LA FUENTE DE DATOS

El dataset a utilizar se obtuvo de Kaggle.

Se seleccionó este dataset en base a la cantidad de registros y variables disponibles, así como la posibilidad de trabajar con un problema de negocio que nos resultara pertinente en la actualidad y acorde a nuestros intereses particulares.

Los principales aeropuertos de todo el mundo se encuentran experimentando una situación cercana al colapso tras la pandemia causada por la COVID-19.

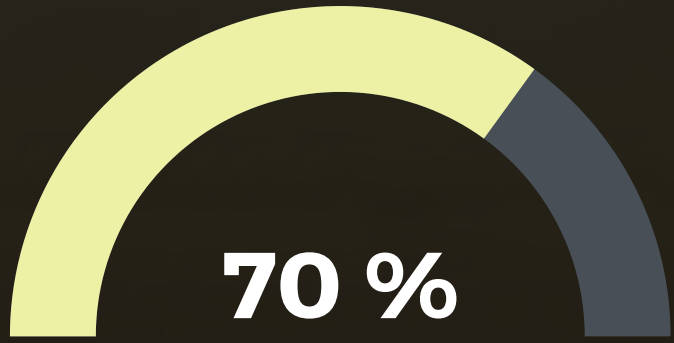
Las consecuencias de esta crisis impactarán de lleno en perjuicio de los usuarios como consecuencia de la reducción de rutas y frecuencias, cancelaciones y un incremento en los precios de los pasajes.



OBJETIVO DE LOS DATOS

Aspiramos a que este conjunto de datos nos permita encontrar correlaciones entre las diferentes variables disponibles para predecir aquellas aerolíneas, rutas aéreas y/o aeropuertos dónde los pasajeros serán más propensos a sufrir retrasos en sus vuelos.

El objetivo que establecimos es utilizar el 70% de los datos para entrenar el modelo, y el 30% restante para testearlo.



CARACTERÍSTICAS DE LOS DATOS

Contamos con **539.383 observaciones**, **8 características o variables** y sin valores faltantes:

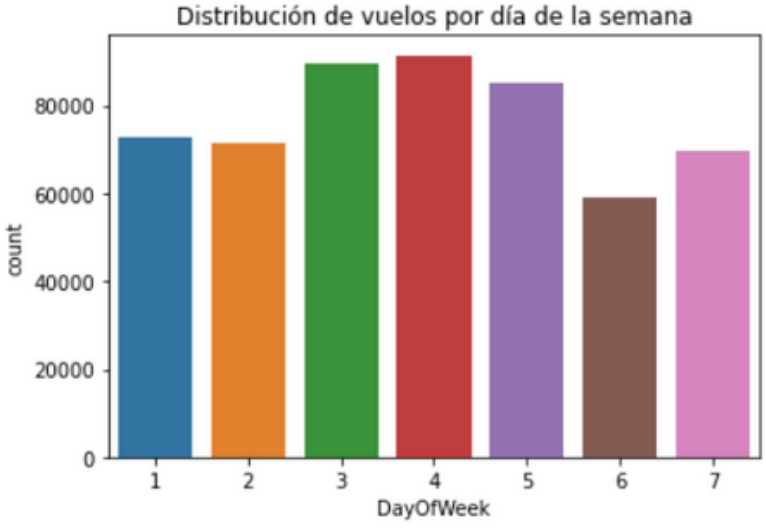
- Airline: Nombre de la aerolínea.
- Flight: Número de vuelo.
- Airport From: Aeropuerto de salida.
- Airport To: Aeropuerto de llegada.
- DayOfWeek: Día de la semana del vuelo.
- Time: Horario de partida del vuelo (en minutos).
- Length: Duración del vuelvo (en minutos).
- Delay: Indica si el vuelo tuvo demoras o no.

DATA WRANGLING

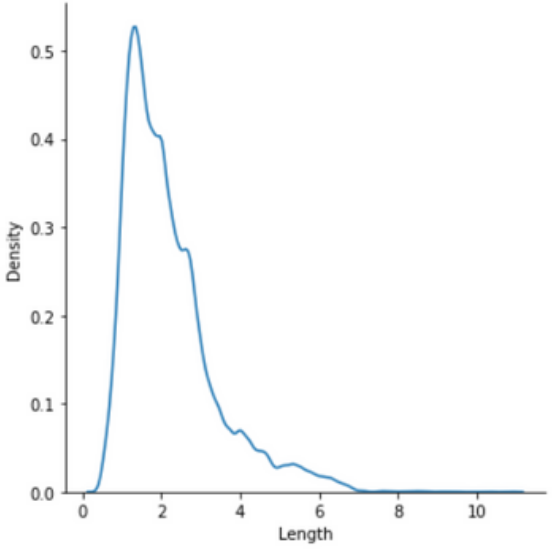
- Eliminar registros con distancia 0 (Length = 0).
- Convertir los registros de las columnas 'Time' y 'Length' a horas.
- Crear agrupación para la columna 'Time':
  - Mañana
  - Mediodía
  - Tarde
  - Noche
- Crear agrupación para la columna 'Length':
  - '<= 1.35'
  - '1.36 - 1.92'
  - '1.93 - 2.70'
  - '>= 2.71
- Incorporar al dataset información sobre los aeropuertos.
- Reemplazar los valores "NaN" de aeropuertos por "Other".

CONCLUSIONES

ANÁLISIS UNIVARIADO



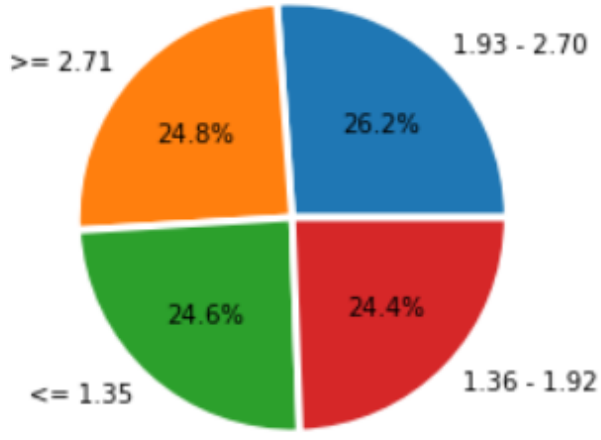
Los días con mayor cantidad de vuelos corresponden a los miércoles, jueves y viernes. En segundo lugar, los días lunes y martes. Finalmente, los sábados y domingos son los días con menos cantidad de vuelos registrados.



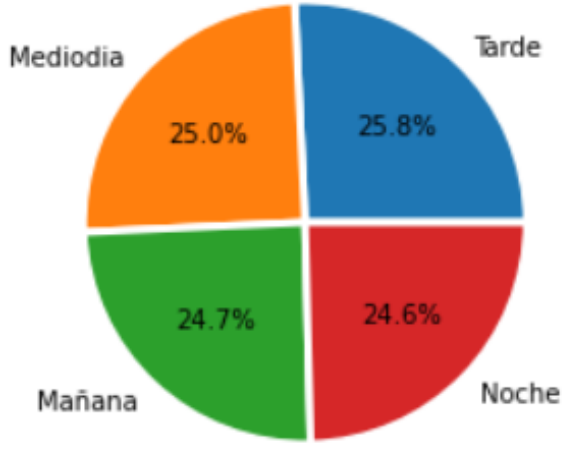
La variable 'Length' cuenta con una distribución asimétrica izquierda. Se evidencia que la duración de los vuelos se concentraron entre 0.5 y 3 horas. Con un valor mínimo de 0 y un máximo de 10.92. Un total de 4 registros con una duración igual a cero fueron excluidos del análisis.

CONCLUSIONES

ANÁLISIS UNIVARIADO



La distribución de la cantidad de vuelos, en base a la duración, es bastante equitativa a lo largo de las 4 agrupaciones realizadas.



La distribución de la cantidad de vuelos, en base al horario de partida, es bastante equitativa a lo largo de las 4 agrupaciones realizadas.

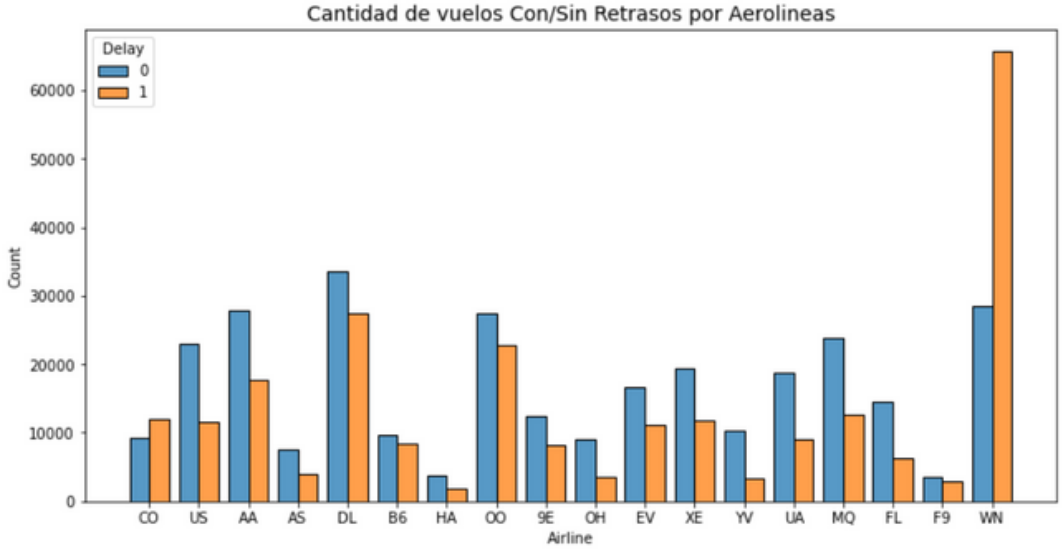
CONCLUSIONES

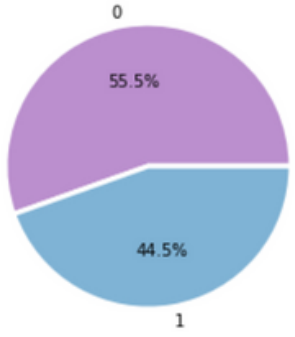
MATRIZ DE CORRELACIÓN DE VARIABLES

Podemos observar que no hay fuertes correlaciones entre las variables.

CONCLUSIONES

ESTUDIO DE LA VARIABLE TARGET: DELAY





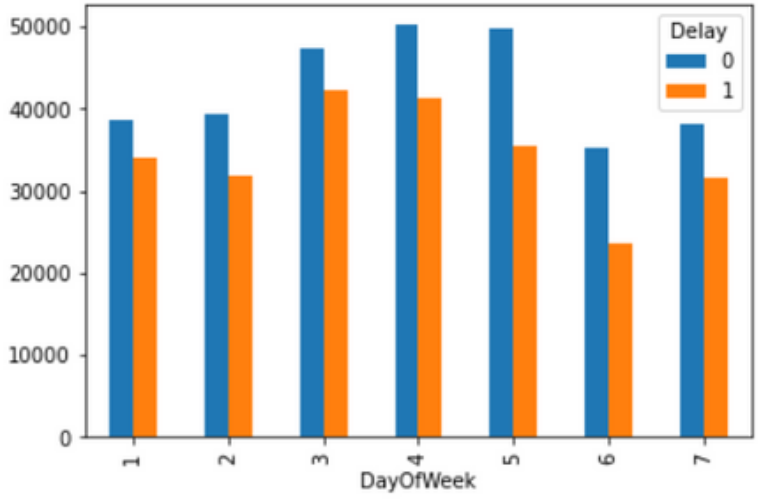
El 44.5% de los vuelos salieron retrasados.

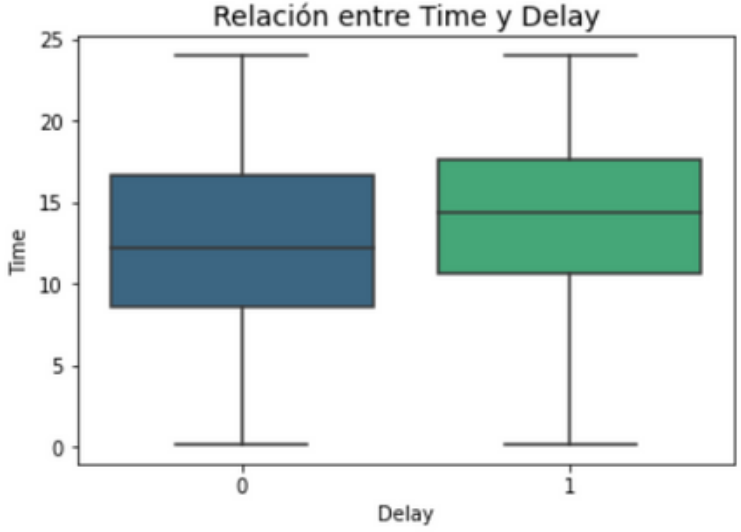
La aerolínea WN si bien es la que mayor cantidad de vuelos realiza, también es la que peor servicio ofrece en cuanto a la puntualidad de los horarios de partida, teniendo más del doble de vuelos retrasados que vuelos puntuales. A su vez, junto con WN, CO es la única otra compañía que tiene más vuelos retrasados que puntuales.

En 16 de las 18 aerolíneas la cantidad de vuelos sin retrasos es superior a la cantidad de vuelos con retrasos. Llamativamente, encontramos la situación inversa en WN, la aerolínea con mayor cantidad de vuelos, y en CO, una de las aerolíneas con menor cantidad de vuelos registrados.

CONCLUSIONES

ESTUDIO DE LA VARIABLE TARGET: DELAY





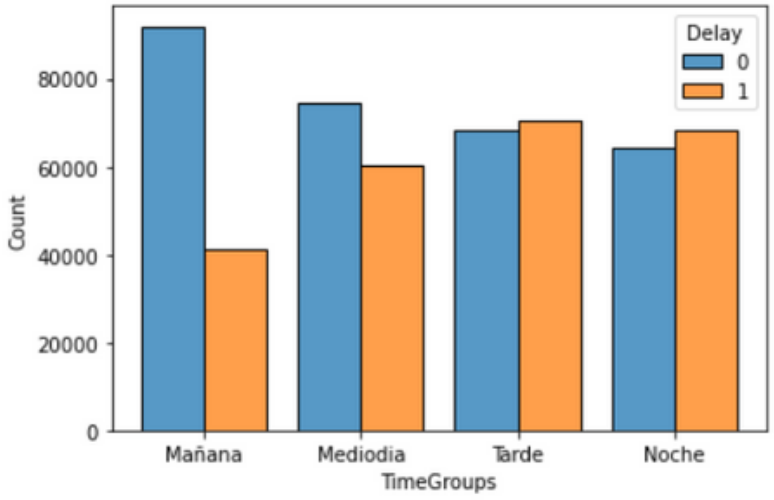
El día de la semana no tiene una gran influencia sobre los retrasos en los vuelos. Analizando los retrasos por día de la semana no se ha identificado ninguna diferencia significativa. Todos cuentan con aproximadamente entre un 40 y 47% de vuelos retrasados.

Advertimos una ligera tendencia de que los vuelos con retrasos suelen tener un horario de partida un poco más tarde que los vuelos sin retrasos.

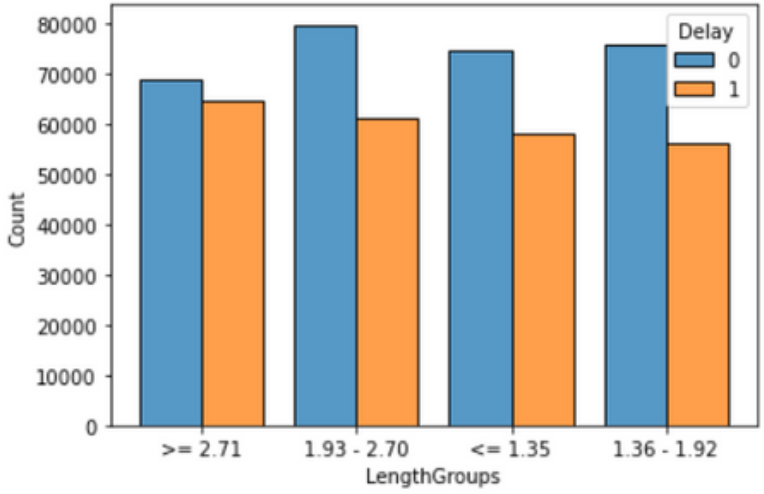
Sin embargo, no es evidencia suficiente para tal afirmación y se debería realizar un test de hipotesis para su confirmación.

CONCLUSIONES

ESTUDIO DE LA VARIABLE TARGET: DELAY



Mientras más tarde es el horario de salida, mayor es el porcentaje de vuelos retrasados. A partir del horario tarde, hay más vuelos retrasados que puntuales.



Los vuelos de una duración larga (mayor a 2.71 horas) suelen tener mayor probabilidad de retraso que las otras 3 agrupaciones.

APLICACIÓN DEL MODELO PREDICTIVO

Con el fin de poder predecir que vuelos se retrasaran a partir de los datos obtenidos, hemos seleccionado el algoritmo **Random Forest** para tal objetivo.

El algoritmo Random Forest es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento.

Los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado.

Esté será el algoritmo al que recurriremos para las distintas pruebas.

PCA

Aplicamos el modelo de **Análisis de Componentes Principales** para intentar reducir la dimensionalidad, dada la cantidad de variables o dimensiones del dataset (607).

Para ello:

- Definimos la variable *target:* en base al objetivo planteado de poder predecir que vuelos puede presentar demoras, nuestra variable objetivo será "Delay".

- Creamos *dummies* para aplicar los algoritmos.

- Dividimos el dataset:
    - El 70% del mismo para el entrenamiento.
    - El 30% restante como test.

- Realizamos una estandarización de variables.

PCA

Resultados obtenidos:

El % de aciertos sobre el set de evaluación es: **0.5989**

El % de precisión sobre la evaluación es: **0.7589**

Recall o la sensibilidad del algoritmo es: **0.1486**

La especificidad del modelo es: **0.2486**

Probamos aplicar nuevamente el método PCA pero esta vez considerando 400 componentes principales (de un total de 600 aproximadamente).

Estos primeros 400 componentes representan cerca de un 75% del peso del modelo.

Nuevos resultados obtenidos:

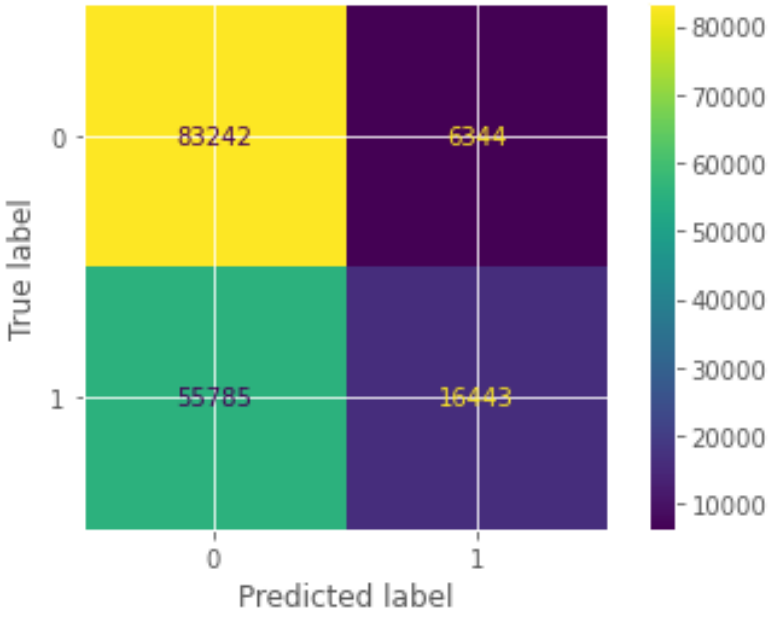
El % de aciertos sobre el set de evaluación es: **0.6161**

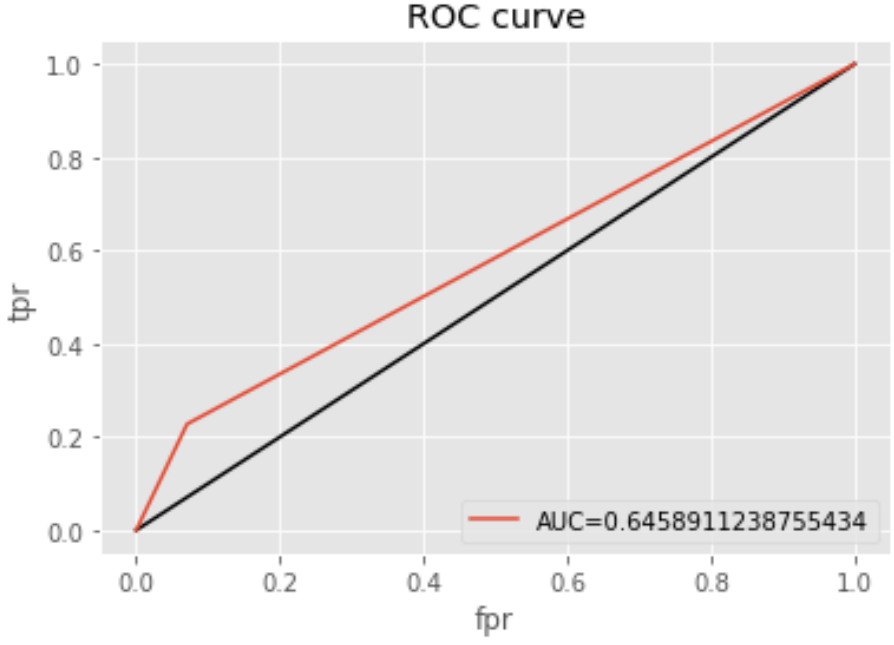
El % de precisión sobre la evaluación es: **0.7216**

Recall o la Sensibilidad del algoritmo es: **0.2277**

La especificidad del modelo es: **0.3461**

RESULTADOS OBTENIDOS

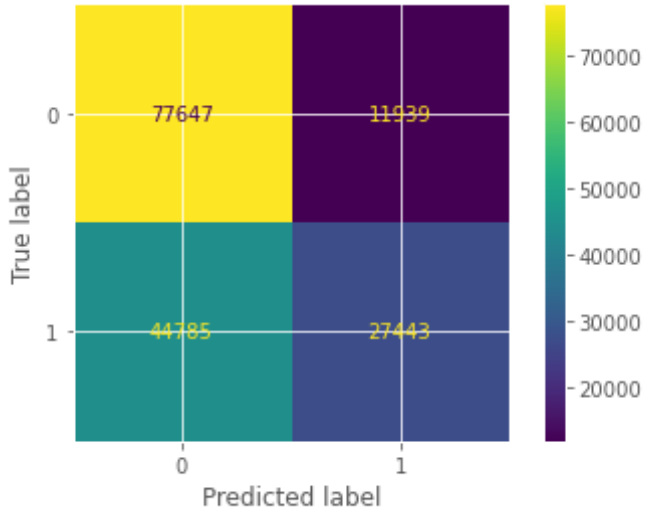


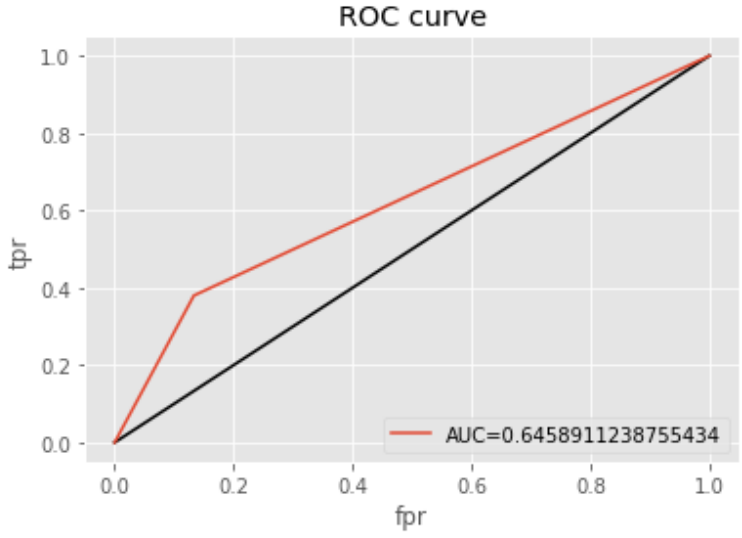


En la matriz de confusión podemos observar:

- Verdaderos Positivos: 83338
- Verdaderos Negativos: 16273
- Falsos Positivos: 55955
- Falsos Negativos: 6248

NUEVA PARAMETRIZACIÓN





El % de aciertos sobre el set de evaluación es: **0.6494**
El % de precisión sobre la evaluación es: **0.6968**
Recall o la sensibilidad del algoritmo es: **0.3799**
La especificidad del modelo es: **0.4918**

En esta nueva parametrización se modificó la profundidad cada árbol de 2 a 10. Sin embargo, observamos que los modelos no son muy efectivos ya que sólo tenemos un poco más de probabilidad de acierto que tirando una moneda.