2022

CODER HOUSE

TRABAJO FINAL

CURSO DATA SCIENCE – CODER HOUSE

IVAN ARANGO, NATALIA JOHANSEN & IGNACIO PIANA

COMISIÓN 29740

Tabla de Contenidos

Descripción del caso de negocio

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 aerolínea que opera el vuelo, el día de la semana en la que está programado, su aeropuerto de origen y aeropuerto destino.

Descripción de los datos

El dataset a utilizar se obtuvo de Kaggle.

ENLACE: https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay

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.

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.

CARACTERÍSTICAS DE LOS DATOS

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

- Airlines: 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

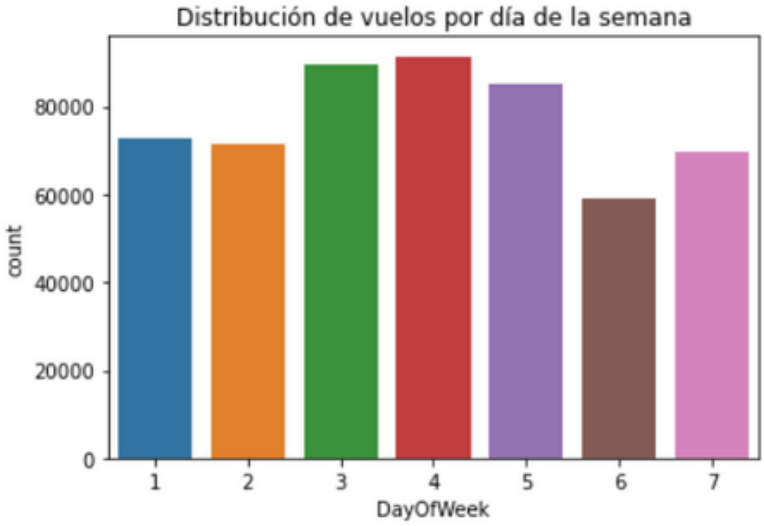
Dentro de las transformaciones que se realizaron para obtener el dataset final se puede detallar:

- Eliminación de registros con distancia 0 (Length = 0).
- Conversión de las columnas 'Time' y 'Length' a horas.
- Creación de agrupación para la columna 'Time':
    - Mañana
    - Mediodía
    - Tarde
    - Noche
- Creación agrupación para la columna 'Length':
    - <= 1.35
    - 1.36 - 1.92
    - 1.93 - 2.70
    - >= 2.71
- Incorporación al dataset de información sobre los aeropuertos. El mismo es un archivo cvs que cuenta con los siguientes atributos:
    1. Airport: Nombre completo del aeropuerto
    2. Cod_Airport: Código de identificación del aeropuerto
    3. Desc_Airport: Nombre corto del Aeropuerto
    4. Loc: Localización del Aeropuerto
- Reemplazo de los valores "NaN" de aeropuertos por "Other".

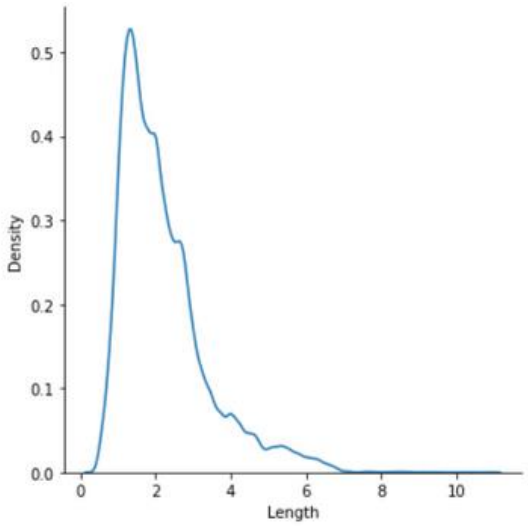
Hallazgos encontrados por el EDA

A continuación de mostraran los resultados obtenidos de los análisis univariados y bivariados realizados:

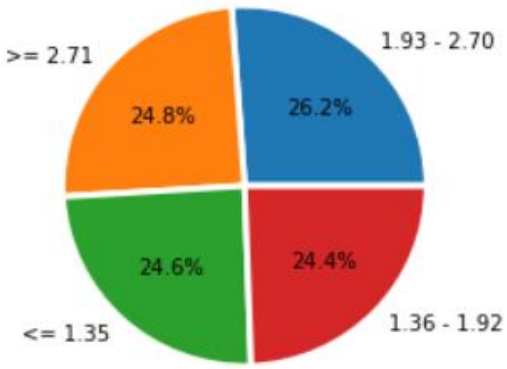
ANÁLISIS UNIVARIADO

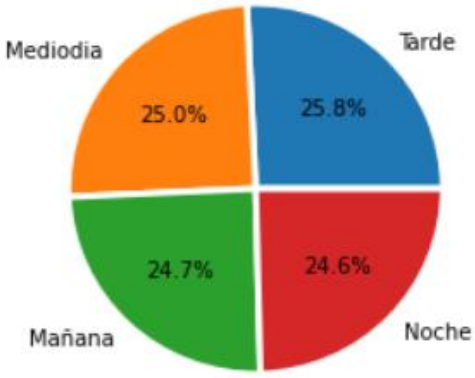


Los días con mayor cantidad de vuelos corresponden a los miércoles, jueves y viernes. En segundo lugar, los días lunes y martes. Finalmente, los sábados y domingos son los días con menos cantidad de vuelos registrados.



La variable 'Length' cuenta con una distribución asimétrica izquierda. Se evidencia que la duración de los vuelos se concentró 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

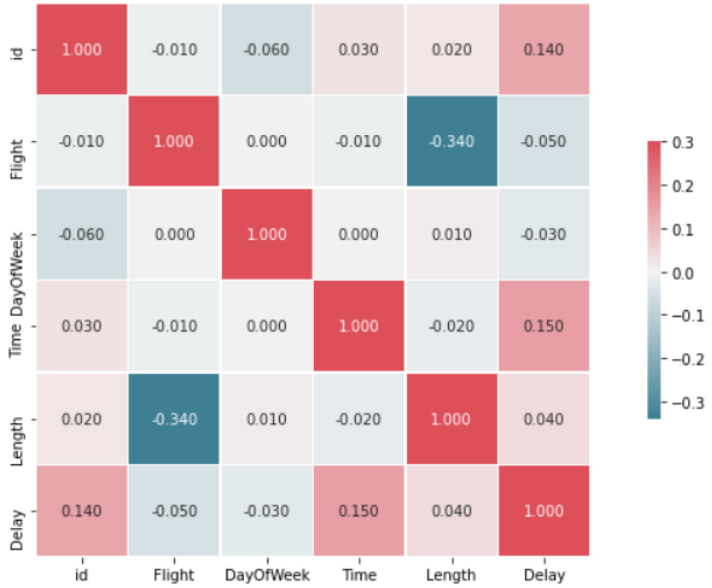




La distribución de la cantidad de vuelos, en base a la duración, es bastante equitativa a lo largo de las 4 agrupaciones realizadas.

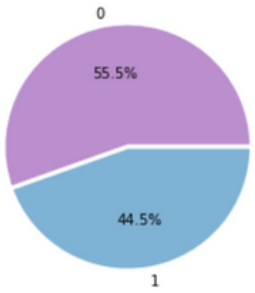
La distribución de la cantidad de vuelos, en base al horario de partida, es bastante equitativa a lo largo de las 4 agrupaciones realizadas.

MATRIZ DE CORRELACION DE VARIABLES

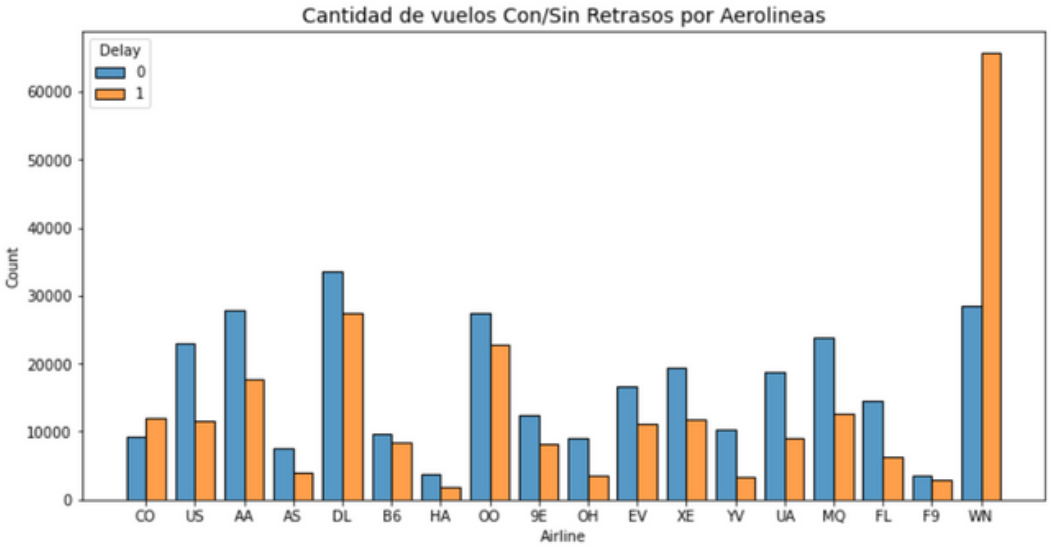


Podemos observar que no hay fuertes correlaciones entre las variables.

ANÁLISIS BIVARIADO – Estudio de la variable “DELAY”

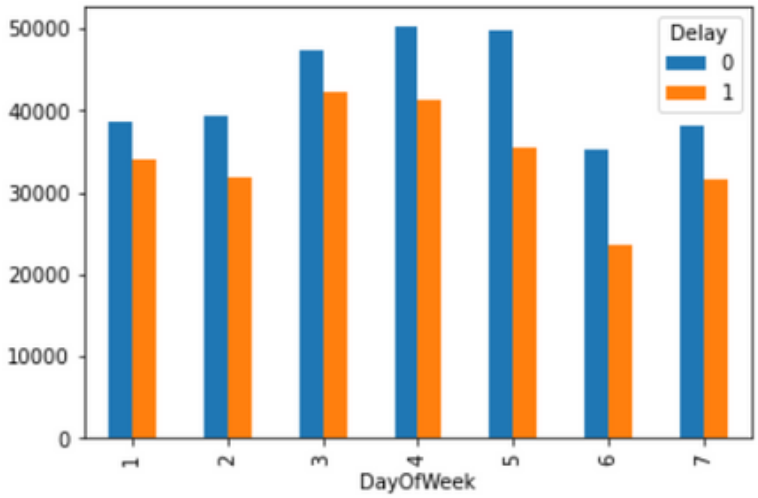


El 44.5% de los vuelos salieron retrasados.

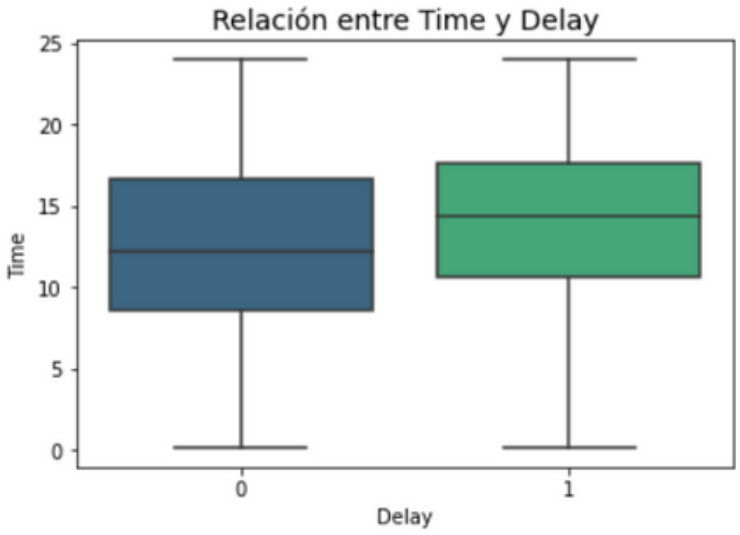


La aerolínea WN si bien es la que mayor cantidad de vuelos realiza, también es la que peor servicio ofrece en cuanto a la puntualidad de los horarios de partida, teniendo más del doble de vuelos retrasados que vuelos puntuales. A su vez, junto con WN, CO es la única otra compañía que tiene más vuelos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

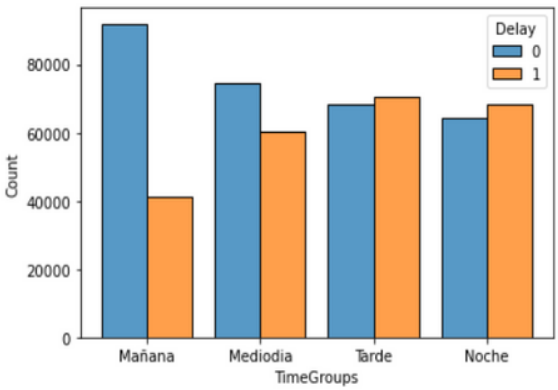
En 16 de las 18 aerolíneas la cantidad de vuelos sin retrasos es superior a la cantidad de vuelos con retrasos. Llamativamente, encontramos la situación inversa en WN, la aerolínea con mayor cantidad de vuelos, y en CO, una de las aerolíneas con menor cantidad de vuelos registrados.

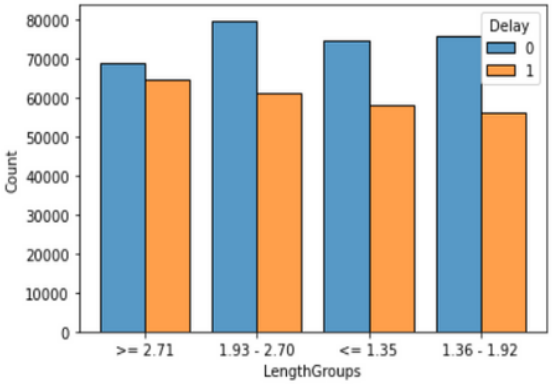


El día de la semana no tiene una gran influencia sobre los retrasos en los vuelos. Analizando los retrasos por día de la semana no se ha identificado ninguna diferencia significativa. Todos cuentan con aproximadamente entre un 40 y 47% de vuelos retrasados.



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.

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.

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.

Algoritmo elegido

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.

Al contar con muchos atributos categóricos y una variable target del tipo booleana, nos hemos inclinado por elegir un algoritmo de Clasificación sobre otros tipos, como por ejemplo de Regresión.

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.

Métricas de desempeño del modelo

Antes de implementar el modelo aplicamos:

- Preparación de datos: convertimos aquellas variables categóricas en numéricas para poder aplicar el algoritmo de predicción. (Función get_dummies)
- Análisis de Componentes Principales (PCA): se normalizaron las columnas y se ejecutó el algoritmo con todos los componentes.

El resultado de la corrida fue:

- El porcentaje de aciertos sobre el set de evaluación es: **59,89%**
- El porcentaje de precisión sobre la evaluación es: **75,89%**
- Recall o la Sensibilidad del algoritmo es: **14,86%**
- La especificidad del modelo es: **24,86%**

Iteraciones de optimización

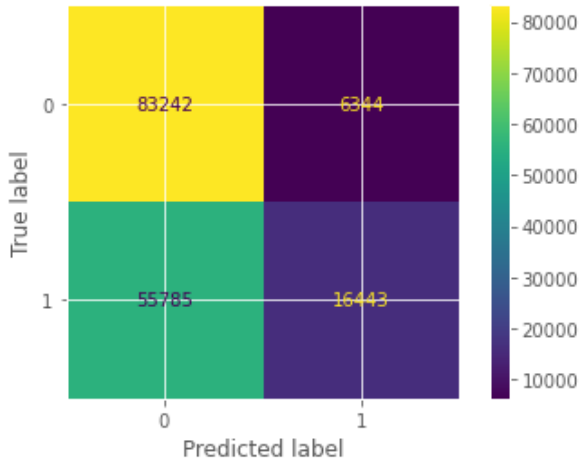
A continuación, se numeran todas las iteraciones que se fueron realizaron para la optimización del modelo:

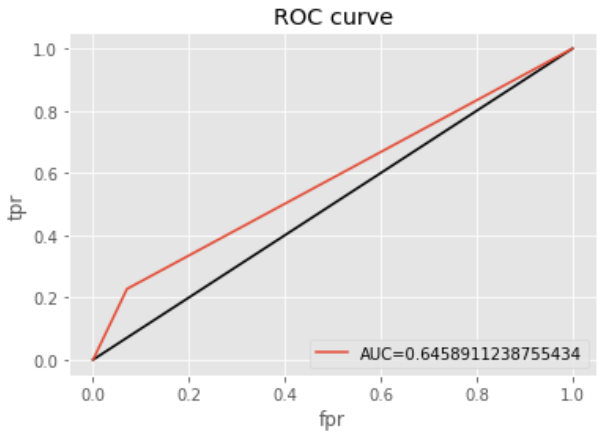
1. PCA – 400 componentes

Se seleccionan los primeros 400 componentes (de un total de 606) que corresponden al 75% del peso del modelo aproximadamente.

Los resultados fueron:

- El porcentaje de aciertos sobre el set de evaluación es: **61,60%**
- El porcentaje de precisión sobre la evaluación es: **72,19%**
- Recall o la Sensibilidad del algoritmo es: **22,76%**
- La especificidad del modelo es: **34,61%**

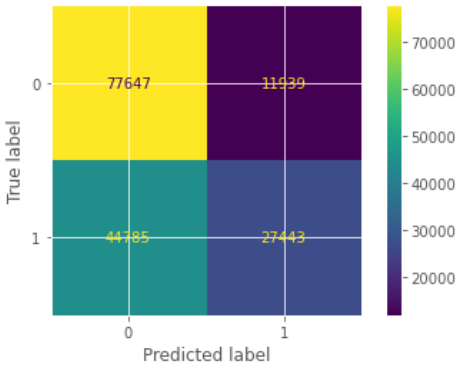




2. PARAMETRO max_depth = 10

Modificamos el parámetro "max_depth" que define la profundidad de los árboles del modelo. En este caso lo aumentamos de 2 a 10.

- El porcentaje de aciertos sobre el set de evaluación es: **64,94%**
- El porcentaje de precisión sobre la evaluación es: **69,68%**
- Recall o la Sensibilidad del algoritmo es: **37,99%**
- La especificidad del modelo es: **49,17%**



3. Modificación de Hyperparametros con **RandomizedSearchCV**

En este caso realizamos varias ejecuciones buscando mejores resultados

OPT 1: Primero corrimos el RandomizedSearchCV con un "n_iter" y "cv" bajo para obtener los primeros resultados rápido y continuar ajustando el modelo.

n_iter = 1 y cv = 2

Parámetros obtenidos:

```
{'n_estimators': 20,
 'min_samples_split': 2,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 10,
 'bootstrap': False}
```

Resultado:

- El porcentaje de aciertos sobre el set de evaluación es: **63,59%**
- El porcentaje de precisión sobre la evaluación es: **74,66%**
- Recall o la Sensibilidad del algoritmo es: **27,91%**
- La especificidad del modelo es: **40,64%**

OPT 2:

n_iter = 5 y cv = 2

Parámetros obtenidos:

```
{'n_estimators': 100,
 'min_samples_split': 6,
 'min_samples_leaf': 3,
 'max_features': 'auto',
 'max_depth': 78,
 'bootstrap': False}
```

Resultado:

- El porcentaje de aciertos sobre el set de evaluación es: **70,04%**
- El porcentaje de precisión sobre la evaluación es: **69,89%**
- Recall o la Sensibilidad del algoritmo es: **57,76%**
- La especificidad del modelo es: **63,25%**

OPT 3:

En base a los mejores hyperparámetros obtenidos previamente, realizamos una nueva lista de estos con valores próximos a los obtenidos inicialmente para poder perfeccionar el modelo.

Parámetros obtenidos:

```
{'n_estimators': 90,
 'min_samples_split': 8,
 'min_samples_leaf': 2,
 'max_features': 'sqrt',
 'max_depth': 109,
 'bootstrap': True}
```

Resultado:

- El porcentaje de aciertos sobre el set de evaluación es: **70,22%**
- El porcentaje de precisión sobre la evaluación es: **70,28%**
- Recall o la Sensibilidad del algoritmo es: **57,65%**
- La especificidad del modelo es: **63,35%**

OPT 4:

Por último, realizamos una última corrida con los parámetros recomendados por nuestro tutor.

Parámetros obtenidos:

```
{'n_estimators': 142,
 'min_samples_split': 2,
 'min_samples_leaf': 3,
 'max_features': 'sqrt',
 'max_depth': 79,
 'criterion': 'entropy',
 'bootstrap': False}
```

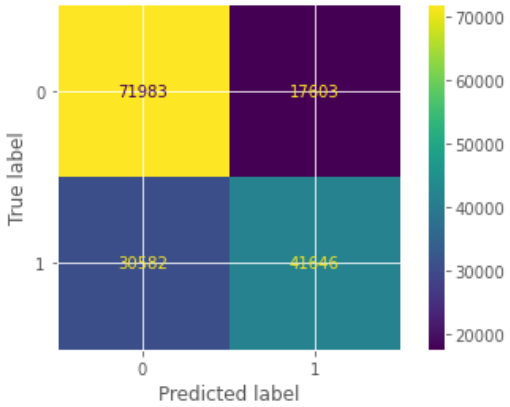
Resultado:

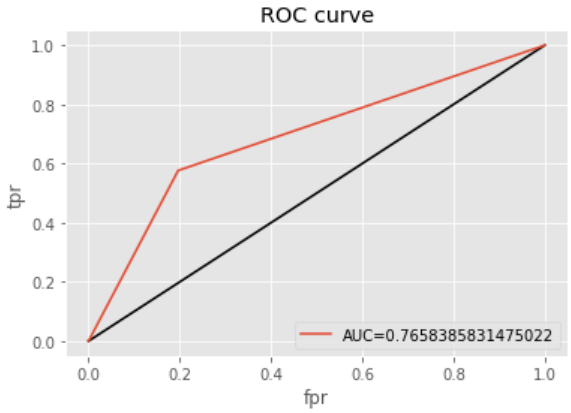
- El porcentaje de aciertos sobre el set de evaluación es: **70,06%**
- El porcentaje de precisión sobre la evaluación es: **69,90%**
- Recall o la Sensibilidad del algoritmo es: **57,84%**
- La especificidad del modelo es: **63,30%**

Métricas finales del modelo optimizado

Como modelo optimizado hemos seleccionado el obtenido en la **optimización 3**

- ✓ El porcentaje de aciertos es: 59,89% vs 70,23% = **+10,34%**
- ✓ El porcentaje de precisión es: 75,89% vs 70,26% = **-5,56%**
- ✓ Recall o la Sensibilidad es: 14,86% vs 57,77% = **+42,91%**
- ✓ La especificidad del modelo es: 24,86% vs 63,41% = **+38,55%**





Futuras líneas

Para complementar el proyecto se podría comenzar a trackear más información (tamaño del avión, clima, entre otros). Esto sumado a la recolección de más meses de información podrían ayudar a aumentar significativamente la precisión del modelo.

También consideramos que poder contar con la fecha de cada vuelo nos permitirá diferenciar (en el caso que exista) la época donde mas retrasos se producen: por ejemplo, en invierno, en cierto mes, con ciertas condiciones meteorológicas según la ubicación o el destino, contemplar feriados, etc.

Otro posible enfoque sería probar con otros tipos de algoritmos de clasificación, como, por ejemplo, arboles de decisión, K-Means, etc.

Por una cuestión de limitación de recursos no hemos podido aprovechar al máximo los métodos de optimización debido a problemas de performances y elevados tiempos de ejecución. De esta manera, si logramos contar con un entorno de procesamiento óptimo creemos que podríamos mejorar sustancialmente nuestro modelo.

Conclusiones

El objetivo era lograr obtener un modelo que nos permita saber si un vuelo va a sufrir un retraso o no. Logramos un modelo que predice correctamente 7/10 casos, lo que permite comenzar a tomar medidas para mejorar la experiencia de los usuarios cuyos vuelos saldrán con retraso, y a su vez, comenzar a investigar causales para reducir los mismos.

A medida que se recolecte más información y se agreguen variables, esperamos poder aumentar la precisión del modelo para que esta sea >90%.