

Master of Science in Data Science

Métodos de aprendizaje de máquinas en Data Science

Tarea 3 modelos de clasificación

Las aerolíneas venden un servicio de transporte entre un punto A y un punto B. Una vez que el vuelo despegó, un asiento vacío ya no se puede vender, por lo que podría decirse que es un producto perecible. Sumado a esto, existe un número significativo de pasajeros que no se presentan en la puerta del vuelo, debido a diferentes razones como: atrasos en las conexiones de otros vuelos, cambio de fecha de pasajes, clima, atrasos en llegar al aeropuerto, cancelación de la reserva, entre otros. Este fenómeno se denomina no-show, y provoca que algunos asientos de cada vuelo salgan vacíos, lo que les genera a las aerolíneas un alto costo de oportunidad.

Para mitigar estos costos, las aerolíneas suelen permitir la sobre-reserva de algunos asientos en sus vuelos (overbooking en inglés). Si la estimación del no-show es correcta, la aerolínea puede salir con todos sus asientos ocupados; si lo subestiman, pueden salir con asientos vacíos, y si la sobreestiman pueden incluso tener que dejar a pasajeros fuera del vuelo (se denomina denied boarding).

El objetivo principal de este proyecto es crear un programa computacional que permita a la aerolínea PANAM estimar si un vuelo tendrá más de 4 no show en su vuelo. Por ejemplo, si un vuelo de Santiago a Concepción tiene 3 no show entonces el vuelo se cataloga como 0, ya que el número de no show fue bajo. Caso contrario, si 4 o más personas no se presentaron, entonces el vuelo se cataloga como 1, ya que la aerolínea pudo haber realizado overbooking.

La base de datos incluye 1,000,000 de vuelos de PANAM entre 2009 y 2012 y 21 variables, cuyo diccionario esta al final de este documento. Además, tiene una segunda base de datos de prueba con 248,880 vuelos y 20 variables, excluyendo la variable de no-show que es la que debe predecir (recuerde, su modelo final tiene que predecir 0 ó 1).

Para la evaluación de la segunda base de datos, usted deberá generar un archivo CSV con una sola columna y 248,880 filas con valores 0 o 1 correspondiente a la evaluación de la segunda base de datos. Para evaluar, se considerará el F1-score entre todos los grupos, utilizando la clase 0 como TP. Es decir, el grupo con mejor F1-score obtendrá 1.0 punto, a partir de ese valor se realizará una regresión hacia los puntajes más bajos. En caso de que el archivo tenga menos filas se agregará el valor 1 hasta cumplir el número de filas requeridas.

Datos de entrenamiento: <https://www.dropbox.com/s/0r1no40e5nj33q0/ALUMNOS-trainData.csv>

Datos de prueba: <https://www.dropbox.com/s/vtmvg4tiuu1dhfl/ALUMNOS-evalData.csv>

Para esta entrega usted deberá entregar 2 archivos

1. Un archivo Rmd que muestre todo el proceso de selección de variables y limpieza de datos aplicados. Además, deberá mostrar la búsqueda de hiperparametros. Este archivo ya deberá haber sido ejecutado y cuando se cargué uno deberá ver todo el proceso de ejecución.
2. Un archivo csv con las 248,880 estimaciones realizadas para el modelo.

Los puntajes asignados a cada tarea corresponden a:

- Limpieza de datos y selección de variables (0.5 puntos)
- Modelo básico (a elección suya entre SVM, naive bayes, y/o árbol de decision), (1 punto)
- Modelo de ensamblado (bagging, boosting, o gradient boosting), con la restricción que utilice un modelo base distinto al anterior, por ejemplo, si uso árbol de decisión en el punto anterior, no puede usar un modelo de ensamblado con árboles. (1 punto)
- Modelo de stack generalization (aprender por su cuenta) (2.0 puntos)
- Comparación de modelos en forma correcta (0.5 puntos)
- Competencia (1 punto)

La fecha de entrega es el 27 de Noviembre a las 23:59 horas.

La tarea se puede realizar hasta en grupos de 3 personas.

¡Mucha suerte!

Diccionario de variables:

- date: Fecha del vuelo
- departure_time: Hora programada para el despegue
- capacity: Capacidad física del avión (# de asientos)
- revenues_usd: Suma de los ingresos totales de un avión (en dólares)
- bookings: Total de reservas en el vuelo al inicio del día de vuelo
- flight_number: Numero de vuelo (indica la ruta)
- origin: Aeropuerto de origen
- destination: Aeropuerto de destino
- distance: Distancia entre origen y destino
- no-show: Número de pasajeros que no se presentaron al vuelo
- denied_boarding: Número de pasajeros que no pudieron abordar por vuelo sobre reservado
- pax_high: Número de pasajeros que compran la tarifa más alta
- pax_midhigh: Número de pasajeros que compran la segunda tarifa más alta
- pax_midlow: Número de pasajeros que compran la segunda tarifa más baja
- pax_low: Número de pasajeros que compran la tarifa más baja
- pax_freqflyer: Número de pasajeros que canjearon el pasaje con kilómetros PANAM
- group_bookings: Número de pasajeros que van con un grupo de turismo

- out_of_stock: Número de días en la historia del vuelo donde no hubo venta por capacidad completada (es un indicador de la demanda del vuelo)
- dom_cnx: Número de pasajeros que provienen o continúan a una conexión domestica (dentro del mismo país de origen)
- int_cnx: Número de pasajeros que provienen o continúan a una conexión internacional
- p2p: Número de pasajeros punto a punto, no conectan ni en origen ni en destino