Trabajo Final Inteligencia de Negocios 2025

Remanente octubre-2025

Maestría en Economía Aplicada, Facultad de Ciencias Económicas de la Universidad de Buenos Aires

Condiciones de Entrega

a) La fecha límite de entrega es el domingo 16 de noviembre a las 23:59:59.
b) Deben enviar una un archivo .ipynb (notebook de jupyter o colab) con el formato tp_final_bi_2025_{apellido}.ipynb. Ejemplo: si el alumno se llamase Juan Pérez, debe enviar el archivo tp_final_bi_2025_perez.ipynb.
c) Este archivo debe ser adjuntado en un correo electrónico a fmastelli@gmail.com con el asunto "TP final REMANENTE OCTUBRE bi 2025 {apellido}". Ejemplo: si la alumna se llamara Marta Calvo, debe enviar el email con el asunto "TP final REMANENTE OCTUBRE bi 2025 Calvo".
d) El archivo que adjuntan en el mail debe estar totalmente ejecutado sin errores. Debe contener tanto el código como las explicaciones.

Conjunto de datos

a) Contarán con 1 dataset de viajes de UBER:
   - uber csv
   - Para mayor información puede recurrir a la fuente

Consignas

1) Lea el archivo "uber.csv". ¿Qué puede decir acerca de la estructura del dataset? Mencione cantidad y tipos de columnas, datos faltantes, de qué va el conjunto de datos en términos generales.
2) Análisis exploratorio
   a) Genere una variable de distancia de recorrido, para ello debe tener en cuenta la geolocalización de partida (pickup_longitude, pickup_latitude) y la geolocalización de llegada (dropoff_longitude, dropoff_latitude). Ver fórmula de semiverseno (vale dialogar y apoyarse en la IA generativa para la implementación o el uso de libreŕias que lo resuelvan de manera directa, por supuesto).
   b) Obtener la matriz de correlaciones para las variables numéricas . ¿Qué puede decir acerca de la correlación entre la tarifa (fare_amount) y la distancia de recorrido ? Por default la correlación se reporta con el método de Pearson, ahora vuelva a generar la matriz pero con el método de spearman. ¿Qué diferencias encuentra? ¿A qué se puede deber?
   c) Obtener estadísticas descriptivas para la variable target (fare_amount) y realizar un histograma de la misma. Comente los resultados obtenidos
   d) Graficar un scatterplot de la variable price y la distancia de recorrido. ¿Detecta alguna anomalía?
   e) Eliminar los outliers univariados de las variables fare_amount y distancia de recorrido. Utilizar y fundamentar el o los criterio/s y métodos que consideren

adecuados. Deberá trabajar con este dataset filtrado en lo que resta del Trabajo Práctico

f) Vuelva a realizar la matriz de correlaciones y el histograma de la variable fare_amount. Comente los cambios observados si los hubiera.

3) Partición de data
   a) Genere un dataset de entrenamiento y otro de test usando split de 80% y su Número de Documento como random_state.
4) Modelado tradicional

   Durante esta consigna, **puede generar las transformaciones que considere conveniente sobre las variables.**

   a) Ajuste un modelo lineal sobre el dataset filtrado de entrenamiento.
      i) ¿Qué puede concluir e interpretar acerca del signo y tamaño de los coeficientes?
      ii) ¿Qué puede decir acerca de la significatividad estadística de los mismos?
   b) Ahora ajuste un modelo lineal LASSO (previa estandarización/normalización de variables), optimizando el parámetro de penalización (lambda en nuestra clase teórica, alpha en sklearn). ¿Alguna variable quedó eliminada? Comente.
5) Modelos de Aprendizaje

   Durante esta consigna, **puede generar las transformaciones que considere conveniente sobre las variables.**

a) Random Forest: Realice búsqueda de hiperparámetros (puede usar conjunto de validación o validación cruzada) sobre el dataset de entrenamiento y obtenga un mejor modelo. Reporte métricas de validación (rmse, mae).
b) Boosting: Realice búsqueda de hiperparámetros (puede usar conjunto de validación o validación cruzada) sobre el dataset de entrenamiento y obtenga un mejor modelo. Reporte métricas de validación (rmse, mae).
c) Redes Neuronales: ídem. Puede probar al menos 3 arquitecturas diferentes variando la cantidad de capas y neuronas de la red densamente conectada (Dense en keras).

6) Performance
   a) Evalúe la performance de todos los modelos (lineal por MCO, LASSO, Random Forest, Boosting y Redes Neuronales) sobre el conjunto de datos de test.. Reporte RMSE y MAE. Concluya