Informe del Trabajo Practico 1

Facultad de Ingeniería de la Universidad de Buenos Aires

Grupo 09

Castro Martinez, Jose Ignacio
Padrón: 106957
email: jacastrom@fi.uba.ar

Douce, German Alejandro
Padrón: 106001
email: gdouce@fi.uba.ar

Orsi, Tomas Fabrizio
Padrón: 109735
email: torsi@fi.uba.ar

Reporte final

En este trabajo, se consiguió estudiar, analizar y construir diversos modelos que permitieron llevar a cabo la tarea de predecir, con una seguridad relativamente alta, el valor de una variable basado en el estudio de un conjunto de datos. Particularmente, la cancelación de una reserva de hotel. El trabajo estuvo dividido en las siguientes 4 partes:

- Análisis Exploratorio + Ingeniería de Features
- Árboles de decisión
- Knn, SVM y Ensamble de modelos
  - KNN
  - SVM
  - Random Forest
  - XGBoost
  - Voting
  - Stacking
- Redes neuronales.

En la primera entrega se realizó un primer acercamiento a los datos con los que trabajamos. Graficamos las variables para ver su comportamiento y características generales y utilizamos métodos univariados y multivariados para la detección y tratamiento de outliers tales como z-score, gráficos en 2D y 3D, imputación de valores por la media y eliminación de entradas. De esta forma buscamos tener un dataset lo más limpio posible para obtener la máxima precisión y efectividad al predecir con los modelos. Previo a la construcción y predicción realizada con los modelos debimos homologar los datasets de entrenamiento y testeo. Lo cual implicó, aplicarle al dataset de testeo las modificaciones realizadas sobre las columnas del dataset de entrenamiento.

En las siguientes etapas se construyeron varios modelos para realizar una predicción sobre los datos. Todos con características muy variadas, de manera que, se pueda detectar cual de ellos es el que mejor se adapta a la hora de realizar una predicción sobre los datos previamente tratados.

Para medir la precisión y capacidad de generalización del modelo utilizamos la métrica F1 debido a que tiene en cuenta tanto el recall como la precisión. El modelo con el que obtuvimos el mejor valor de esta métrica fue XGBoost con un valor de 0,8423 en la competencia de Kaggle seguido por voting que dio prácticamente lo mismo (0,84007) y Random Forest que dio 0,819. En el caso de los clasificadores SVM Y KNN no se obtuvieron muy buenos resultados (0,60 y 0,76 respectivamente).

Finalmente, la red neuronal construida nos dio valores de la métrica F1 muy inferiores a los otros modelos (0.67 en este caso) Por otro lado, se confirma una observación anteriormente planteada en la cual, se consideraba mejor a los modelos basados en árboles de decisión. A pesar de que la gran mayoría de los modelos consiguen tener buenas métricas luego de haber sido optimizado por hiperparametros.

Como detalle final, pudimos observar que, a pesar de su fama, las redes neuronales no implicaron una mejora sustancial en comparación con los otros algoritmos. A pesar, de que estas últimas requirieron una complejidad computacional mucho mayor.