Universidad de Buenos Aires
Facultad de Ingeniería

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

75.06/95.58 – Organización de Datos

Curso 02
1C 2023

Trabajo Práctico 1: Reservas de Hotel
Conclusiones finales

Grupo 14

Integrantes:

- *BOTOSHANSKY, ALAN* *108933*
- *BOTOSHANSKY, IVAN* *108897*
- *NIEVA, ULISES TADEO* *107227*

Corrector: Nacho

Fecha de entrega: *25* de *mayo* del 2023

Universidad de
Buenos Aires
Facultad de
Ingeniería

Resumen

Durante el desarrollo de este trabajo práctico, llevamos a cabo diversas tareas con el objetivo de explorar y analizar un conjunto de datos proporcionado, así como entrenar modelos predictivos y realizar predicciones. En el reporte presentaremos las principales conclusiones obtenidas a lo largo del proceso, destacando los hallazgos más relevantes y las decisiones tomadas en cada etapa.

Desarrollo

En primer lugar, realizamos una exploración inicial del conjunto de datos llamado "hotels_train.csv". Identificamos las variables presentes en el dataset y las categorizamos en cuantitativas y cualitativas. Calculamos medidas de resumen para las variables cuantitativas y describimos brevemente las variables cualitativas.

Posteriormente, realizamos análisis gráficos para examinar las distribuciones de las variables. Observamos que la mayoría de las variables presentaban distribuciones desequilibradas y asimétricas. Además, exploramos las correlaciones entre las variables cuantitativas graficando una matriz de correlación mediante un heatmap. Encontramos una falta de correlación significativa entre las variables cuantitativas, con solo algunos casos de correlación analizados en detalle.

Continuamos analizando la relación de las variables con el objetivo principal del trabajo práctico, predecir el valor de la variable "is_canceled". Utilizamos visualizaciones como barplots, boxplots y heatmaps para examinar estas relaciones. Identificamos distintas variables como "hotel", "lead_time", "stays_in_week_nights", "adults", "country" y "assigned_room_type" (entre otras) que parecían tener una influencia significativa en la predicción del valor del target.

Luego, elaboramos más visualizaciones para resaltar observaciones relevantes derivadas de la exploración inicial de las variables. Realizamos pairplots y violinplots para anticipar posibles outliers y la relación entre las distribuciones de las variables "adults", "children" y "babies". Además, utilizamos un heatmap para examinar la relación entre el tipo de habitación reservada y el tipo de habitación asignada, y cómo esto afectaba a la cancelación de las reservas. También combinamos variables significativas en términos de su relación con el objetivo, y usamos piecharts para mostrar el porcentaje de reservas canceladas y no canceladas en el conjunto de datos.

Otro aspecto importante abordado fue el tratamiento de los datos faltantes en el dataset. Se identificaron variables con datos faltantes, como "children", "country", "agent" y "company". Para las variables con un porcentaje de faltantes bajo, se imputaron los valores utilizando la mediana o un valor específico ("Undefined"). Para las variables con un porcentaje más significativo de datos faltantes, también utilizamos el valor "Undefined" para la imputación. Dado el alto porcentaje de faltantes en la variable “company”, podríamos haberla eliminado del dataset. Sin embargo, decidimos no hacerlo ya que el hecho de que hubiera campos vacíos en dicha variable podría significar que la reserva fue realizada sin una compañía de viaje (tipo de dato faltante Missing Not At Random). Por lo tanto, podríamos estar perdiendo información valiosa si eliminamos la variable.

Una vez que realizamos las tareas de exploración y preprocesamiento de los datos, procedimos al entrenamiento de los modelos predictivos. Seleccionamos varios modelos de clasificación, como

Árboles de Decisión, KNN, SVM, RF, XGBoost, Stacking, Voting y Redes Neuronales, con el objetivo de comparar su rendimiento en la tarea de predicción de cancelación de reservas, observando todas las métricas. Dividimos el conjunto de datos en conjuntos de entrenamiento y validación, para poder evidenciar dicho rendimiento.
Además, aplicamos técnicas de validación cruzada para hallar la mejor configuración de hiperparámetros para cada tipo de modelo. Al optimizar dichos parámetros, optimizamos la métrica F1 Score. Observamos que el modelo de tipo Voting presentaba el mejor desempeño, logrando una mayor precisión, recall y F1 Score en la predicción de la cancelación de reservas sobre el set de validación.

Finalmente, utilizamos los modelos entrenados para realizar predicciones en un conjunto de datos de evaluación (no etiquetado). Se aplicaron las mismas técnicas de preprocesamiento utilizadas en el conjunto de entrenamiento y utilizamos los distintos modelos para predecir la cancelación de las reservas. Las predicciones obtenidas fueron registradas y subidas a la competencia de Kaggle.

Conclusión

A lo largo de este trabajo práctico llevamos a cabo un análisis exhaustivo del conjunto de datos, exploramos las relaciones entre las variables y aplicamos técnicas de visualización y procesamiento de datos para obtener una comprensión más profunda del mismo. Entrenamos varios modelos predictivos y evaluamos su rendimiento, destacando el modelo de Voting (compuesto por tres clasificadores XGBoost) como el que tuvo la mejor performance en la tarea de predicción de cancelación de reservas. Las predicciones obtenidas fueron registradas y podrían ser utilizadas por los hoteles que tomaron dichas reservas para la toma de decisiones.