Organización de datos - TP 1

Grupo 10 - **Reporte final:**

Introducción

En este informe, se presentarán los pasos realizados para analizar el dataset de reservaciones de hotel, preprocesar los datos e implementar y evaluar varios modelos de clasificación. De esta forma se logró predecir si una reserva de hotel será o no cancelada. Se presentarán también los factores que influyen en la cancelación de reservas y los resultados obtenidos.

Exploración de datos

Durante el análisis exploratorio de los datos, se realizaron descubrimientos clave que revelaron información relevante sobre las características que influyen en la cancelación de reservas de hotel. Se observó una correlación lineal positiva entre la variable 'lead_time' (tiempo de anticipación de la reserva) y la cantidad de cancelaciones, lo que indica que a medida que el tiempo de anticipación aumenta, también lo hace la probabilidad de cancelación. Por otro lado, se encontró una correlación negativa entre la variable 'total_of_special_requests' (cantidad de pedidos especiales) y el porcentaje de cancelaciones, sugiriendo que un mayor número de solicitudes especiales está asociado con una menor probabilidad de cancelación. Asimismo, se descubrió que las reservas que requerían al menos un espacio de estacionamiento ('required_car_parking_spaces') no presentaron cancelaciones en su totalidad. Por último, se identificó una tendencia inesperada en la variable 'deposit_type', donde las reservas con un depósito de tipo 'non-refund' mostraron un porcentaje significativamente alto de cancelación. Estos descubrimientos no fueron los únicos pero sí los más relevantes.

Datos faltantes

Se detectaron variables con valores nulos. A continuación las presentamos y sus respectivos porcentajes. También aclaramos al lado cuál fue finalmente la técnica para trabajarlos:

- agent(12.7%): Agregamos la categoría “sin agente” representada por un 0
- company (94.9%): Agregamos la categoría “sin company” representada por un 0
- country(0.35%): Donde eliminamos las filas con estos valores ya que eran muy pocos
- distribution channel(0,006%): Eliminamos las filas con estos valores por la misma razón.
- market segment(0.003%): Eliminamos las filas con estos valores por la misma razón.
- children (0.006%): Eliminamos las filas con estos valores

Transformaciones

A pesar de que para algún modelo en específico pueden haber variado las transformaciones a grandes rasgos las utilizadas fueron:

- OneHotEncoding: Para llevar las variables categóricas a números que puedan utilizar los modelos. Con agent y country en específico se realizó sólo para aquellas categorías cuyas frecuencias representaban mínimo un 1% de las filas.
- StandardScalation: Para llevar los features a una misma escala y que no se les dé más importancia a algunos sobre otros solamente por su magnitud.

Modelos de clasificación entrenados:

Se entrenaron varios modelos de clasificación con el objetivo de predecir la cancelación de reservas. Cada modelo fue entrenado con los datos de entrenamiento y evaluado utilizando los datos de validación. El objetivo principal fue maximizar la métrica f1. Presentamos sus resultados a continuación.

Modelo Entrenado, f1-score validación:

- Árbol de decisión: 0.8561
- KNN: 0.8335
- SVM Lineal: 0.7152
- SVM Polinómico: Sin resultados
- SVM Radial: Sin resultados
- Random forest: 0.877
- XGBoosting: 0.8833
- Stacking: 0.8796
- Voting: 0.8571
- Red neuronal: 0.861

Otras opciones a explorar no utilizadas

A pesar de haber entrenado una variedad amplia de modelos, hubo otras ciertas opciones que se podrían haber empleado. Como cascading dentro de la categoría de ensambles híbridos, regresión logística y bagging (Aunque sí utilizamos random forest).

Conclusiones

En resumen, después de entrenar y optimizar varios modelos de clasificación, obtuvimos resultados prometedores en la predicción de cancelaciones de reservas de hotel. En general, los modelos mostraron un desempeño consistente, con una métrica f1 superior a 0.85. Sin embargo, el modelo XGBoost se destacó, alcanzando un puntaje de 0.883 en el conjunto de validación y 0.882 en el conjunto de prueba en la plataforma Kaggle.