Informe Final - Grupo Alan Taylor

Introducción

Comenzamos con preprocesamiento de los datos en checkpoint 1, a lo largo del tiempo con cada checkpoint fuimos agregando métodos, entendiendo de mejor forma el dataset.

Para el preprocesamiento de datos decidimos hacer los siguientes cambios:

- **Change Room:** Es la unión de reserved room_type y assigned room_type, que informa si la habitación reservada fue la asignada o no.
- **Company:** Le imputamos 0 a los valores nulos y 1 a los no nulos convirtiéndola en una variable booleana.
- **Agent:** Le imputamos 3 a los valores nulos.
- **Country:** Dividimos los países en 3 categorías, los países con mayor porcentaje de cancelación, menor porcentaje y neutros. A los registros con valor nulo los eliminamos.
- **Nights:** Es el resultado de la suma de stays_in_weekend_nights y stays_in_week_nights
- **Minors:** Es el resultado de la suma de childrens y babies.
- Eliminamos categorías de variables cualitativas que nos causaban problemas al subir las predicciones.

Luego a todas las variables categóricas les hicimos One Hot encoding.

Analizamos y eliminamos outliers que definimos problemáticos para la predicción utilizando técnicas como gráficas de boxplot, z-score modificado para outliers univariados y distancia de mahalanobis para outliers multivariados.

Para el entrenamiento de los modelos siempre antes dividimos el dataset 70% para entrenar y 30% restante para evaluar, después utilizamos random search o grid search para buscar los mejores hiper parámetros. Utilizamos los siguientes modelos (accuracy):

- **Árbol de decisión:** Nos dio un resultado bastante bueno pero no el mejor (%85).
- **KNN:** Nos dio un mal resultado (%79).
- **Support Vector Machine (SVM):** De los peores (%79).
- **XG Boost:** El mejor de los modelos (%88.1).
- **Red neuronal:** Aunque el resultado es mediocre también decepcionante (%84.5)

Luego utilizamos los siguientes ensambles:

- **Random Forest:** Muy bueno, muy cerca del mejor (%88.6).
- **Voting:** El mejor predictor aunque en el dataset de testeo daba peor accuracy en la competencia tenía mayor porcentaje de acierto (%88)
- **Stacking:** Aunque en el dataset de testeo daba mayor accuracy en la competencia daba levemente menor a voting (%88.6).

Cuadro de Resultados

Medidas de rendimiento en el conjunto de TEST:

Para el ensamble voting utilizamos los modelos con mejores resultados de Random Forest, XG Boost y SVM, con un tipo de votación “Hard” que significa que predicen con valores booleanos, a diferencia de “Soft” que predicen con valores de probabilidad entre 0 y 1.

Modelo	CHPN	F1-Test	Precision Test	Recall Test	Accuracy	Kaggle
Arbol de decision	2	0.8533	0.8381	0.8691	0.8515	0.8484
Random Forest	3	0.8857	0.8885	0.8830	0.8853	0.8801
Voting	3	0.8796	0.8851	0.8742	0.8805	0.8815
Red Neuronal	4	0.8517	0.8550	0.8389	0.8542	0.8139

Árbol de decisión: modelo de aprendizaje supervisado que divide el conjunto de datos en nodos y ramas, tomando decisiones basadas en las características de entrada para predecir una variable objetivo.
Random Forest: Conjunto de múltiples árboles de decisión que promedian las predicciones de cada árbol para mejorar la precisión y reducir el sobreajuste.
Voting: método de ensamble que combina las predicciones de múltiples modelos individuales para tomar una decisión final, ya sea por mayoría o promedio.
Red Neuronal: modelo de aprendizaje profundo que consiste en múltiples capas de neuronas interconectadas. Realizan cálculos complejos para aprender representaciones de datos.

Conclusiones generales

Nos dimos cuenta que es necesario el preprocesamiento de los datos pero no excesivo, ya que al principio trabajamos mucho los datos, pero luego a medida que redujimos lo trabajado nos dimos cuenta que los modelos predecían mejor. Al igual que con algunos outliers, nos dimos cuenta que si no los eliminamos los modelos predecían mejor.

De los modelos el que mejor predecía en el test era el Random Forest, por otro lado en Kaggle el que mejor desempeño es Voting, aunque el más rápido y sencillo de entrenar fue el árbol de decisión que para su simpleza dio un muy buen desempeño.

Nos parece que el resultado de nuestro mejor modelo es bastante bueno y aplicable en la realidad.

Creemos que trabajando un poco más el preprocesamiento del dataset, y optimizando hiperparametros de los modelos dentro del ensamble, y probando otras combinaciones de modelos podríamos mejorar los resultados obtenidos.

Tareas Realizadas

Integrante	Promedio Semanal (hs)
Santiago Bautista Trezeguet	7 hs
Estefano Polizzi	8 hs
Ignacio Oviedo	6 hs