Organización de Datos (75.06)
Trabajo Práctico 1: Reservas de Hotel. Informe Final.
Cuatrimestre, año y grupo: 2ºC 2023
Grupo: 11 - Los Pandas
Corrector: Mateo Suster
Fecha de entrega: 3 de noviembre de 2023
Datos de los integrantes:

Labollita, Francisco	106861	flabollita@fi.uba.ar
Mundani Vegega, Ezequiel	102312	emundani@fi.uba.ar
Otegui, Matías Iñaki	97263	motegui@fi.uba.ar

Introducción

Para la realización del trabajo práctico se utilizaron distintas técnicas, pruebas y modificaciones sobre el dataset de entrenamiento que cambiaron dependiendo del clasificador. Se considera que la más "completa" es la utilizada en el ensamble de stacking, ensamble voting y en las redes neuronales: se creó la columna *room_type_match*, que indica si la habitación asignada coincide con la seleccionada por el cliente. A las columnas *arrival_date_year* y *agent* se las analizó como categóricas. Se eliminaron las filas con datos faltantes en *country*, *market_segment* y *distribution_channel*. Se eliminaron las columnas consideradas irrelevantes, en particular aquellas cuyos porcentajes de cancelación no parecían tener una clara relación con el atributo analizado: *arrival_date_day_of_month*, *stays_in_weeknd_nights*, *stays_in_week_days*, *children*, *babies* , *reserved_room_type*, *deposit_type*, *company*, *adr* e *id*. También se eliminaron los valores considerados outliers de la columna *adults*. Se hizo one-hot encoding de las columnas categóricas y se normalizaron las columnas cuantitativas numéricas. Las siguientes columnas fueron transformadas a variables booleanas (0 si su valor era 0, 1 si su valor era mayor a 0): *required_car_parking_spaces*, *days_in_waiting_list*, *babies*, *previous_cancelations*, *special_requests*, *previous_bookings_not_canceled*, *booking_changes*.

A la hora de trabajar con "Voting" y "Stacking" fue necesario normalizar los valores numéricos cuantitativos dado que se decidió utilizar el mismo dataset para todos los modelos.

Cuadro de Resultados

Modelo	CHPN	F1-Score	Precision	Recall	Accuracy	Kaggle
Arbol de Decisión	2	0,842	0,813	0,873	0,836	0,834
XGBoost	3	0,864	0,859	0,868	0,863	0,859
Random Forest	**3**	**0,858**	**0,868**	**0,848**	**0,859**	**0,866**
Red Neuronal	4	0,855	0,823	0,890	0,849	0,838

El modelo del CHP 1 fue un Árbol de Decisión, optimizado mediante un Grid Search. Es una estructura en forma de árbol donde un nodo interno representa una característica o atributo, una rama representa una regla de decisión y un nodo hoja representa el resultado o etiqueta de clase. Nuestro mejor modelo fue el Random Forest que construye múltiples arboles de decisión durante el entrenamiento y luego combina sus decisiones para alcanzar un solo resultado. El segundo mejor clasificador del CHP 2 fue el XGBoost. Utiliza el método de boosting para construir un modelo predictivo a partir de varios modelos débiles (generalmente árboles de decisión) de manera secuencial, enfocándose en corregir los errores de predicción en cada etapa y logrando una alta precisión en la clasificación y regresión de datos. Finalmente, para el CHP 4, utilizamos una Red Neuronal. Está compuesta por neuronas artificiales interconectadas que procesan y aprenden de datos para realizar tareas como el reconocimiento de patrones y la toma de decisiones para problemas de clasificación y regresión. El resultado no superó a nuestro mejor modelo.

Conclusiones generales

Resultó esencial haber realizado un análisis exploratorio exhaustivo y detallado para poder elegir bien qué variables del dataset considerar y qué modificaciones realizar. Una de las estrategias utilizadas que probó ser muy eficaz fue la de hacer booleanas algunas columnas numéricas (si ≥ 0 es `True`, si $= 0$ es `False`). Esto no solo permitió utilizar filas que tenían valores considerados como outliers, sino también probó mejorar el rendimiento de los modelos. Esto se pudo corroborar por ejemplo con KNN y SVM, cuyos rendimientos fueron mejores al utilizar la "booleanización" para construir ensambles que cuando no se usó para entrenarlos.

De todos los modelos entrenados, el que tuvo un mejor desempeño (mejor F1-score) en TEST fue XGBoost, mientras que en Kaggle fue Random Forest. Ambos tuvieron desempeños similares, el hecho que el mejor varíe dependiendo del conjunto puede deberse a que en Kaggle se está probando con una parte reducida del dataset (solo su parte pública), o sino a que simplemente son sobre datasets diferentes.

El modelo más sencillo de entrenar fue el árbol de decisión, su optimización también fue relativamente rápida. Tuvo una buena relación costo rendimiento. El más costoso de entrenar fue SVM, era tan costoso que incluso se redujo el dataset de entrenamiento y se redujo la cantidad de folds utilizados para la validación cruzada.

El modelo entrenado sería extremadamente útil si se utiliza en un contexto en que las agencias y hoteles, sigan siendo estos. Utilizado en otro contexto seguiría siendo útil, siendo que las variables que más información dieron fueron tipo de depósito, con cuánta anticipación se hizo la reserva y algunos de los países de origen; pero convendría reentrenar utilizando reservas de este nuevo contexto.

Surgieron ideas para mejorar los resultados pero algunas no fueron aplicadas por ser costosas de implementar, porque no se tenía tiempo de probarlas o simplemente porque quedaron así, como ideas... Entre ellas están crear una columna "fecha", que combine el año, mes y día; realizar una optimización mucho más exhaustiva de hiperparámetros para random forest o XGBoost; agregar muchas más neuronas en la capa inicial de la red neuronal; en caso de ser posible, tomar más muestras de los modelos.

Algo que resultó interesante fue la variable que más peso tuvo para entrenar los árboles, que fue *deposit_type*, una variable que previo al análisis exploratorio se pensó que no iba a ser tan importante, y que incluso se creia que la mayoría de las *non_refund* iban a tener una tasa muy baja de cancelación, cuando este tipo tuvo una tasa del 100 %.

Como conclusión final, los integrantes consideran haber logrado entrenar muy buenos modelos predictivos utilizando los conocimientos adquiridos en la materia y creén que fue indispensable entender el funcionamiento de los mismos para haber logrado un buen desempeño de estos.

Tareas Realizadas

Integrante	Promedio Semanal (hs)
Labollita, Francisco	16
Mundani Vegega, Ezequiel	16
Otegui, Matías Iñaki	16