Segunda entrega - Introducción a la Inteligencia Artificial: ¿Se cancelará la reserva del hotel?

Paola Andrea Posada Restrepo <paola.posada1@udea.edu.co>

Stiven Guerra Chaverra<stiven.guerra@udea.edu.co>

Sebastián Gómez Ramírez <sebastian.gomez35@udea.edu.co>

Descripción del progreso alcanzado

Se hizo una exploración del dataset y se buscó entender el problema. En esta exploración se vio con qué tipos de variables contaba el dataset, qué datos faltantes tenía y posterior al análisis, se limpió el dataset al decidir qué columnas se eliminarían. Finalmente, se le hizo un escalamiento estándar a los datos para iniciar con las pruebas del modelo Random Forest con el cual se obtuvieron muy buenos resultados.

Exploración descriptiva del Dataset

Debido a la gran cantidad de datos con los que contaba el dataset original y al alto costo computacional que supone el procesamiento de la totalidad de datos, se seleccionó el conjunto de datos pertenecientes al año 2015, reduciendo así la cantidad de datos total, contando finalmente con 21996 registros.

El dataset cuenta en un inicio con un total de 36 características, de las cuales se logran identificar 16 de tipo entero, 4 de tipo float y 16 de tipo objeto, estas últimas siendo candidatas a pasar por un proceso de categorización.

La variable de clasificación es *is_canceled*, donde 0 significa que la reservación NO fue cancelada (con 13854 datos) y 1 que la reservación fue cancelada (con 8142 datos).

Preprocesamiento del Dataset

La primera característica electa a eliminar, derivada de la decisión inicial de tomar datos que solo comprenden fechas dentro del año 2015, es en efecto la columna *arrival_year*, ya que todos los registros contarán con el mismo dato en esta columna.

Como se mencionó previamente, al empezar con el análisis de los datos de tipo objeto, se deciden eliminar un total de 5 características debido a su variedad de datos, impidiendo así la categorización de las mismas. Las características mencionadas son: *name*, *email*, *card_number*, *phone_number* y *reservation_status_date*. Además, debido a la alta correlación de la columna reservation_status con nuestra columna is_cancel, se elimina esta primera puesto que significan casi lo mismo.

El siguiente paso dado correspondía a identificar datos nulos en el dataset y durante el proceso decidir si hay alguna columna que deba ser eliminada debido a una alta cantidad de datos faltantes. Para nuestra fortuna no fue necesario eliminar columnas por esta causa. Sin embargo, se identificaron datos nulos en las columnas *country, children, agent, company.* Cabe destacar que más del 90% de datos de la columna *company* correspondían

a datos vacíos, no obstante, debido a la definición de esta columna, se decidió conservarla ya que el dato vacío se puede interpretar como que no se realizó reserva por medio de una compañía.

Volviendo al procedimiento realizado al resto de columnas con datos faltantes, se les realizó un tratamiento en el cual se reemplazaron los datos faltantes por 0, indicando que no tenían niños en la columna *children*, no tenía compañía o agencia en las columnas *company, agent,* respectivamente o NC, que significa *No Country* para columna *country*.

Finalmente, las columnas candidatas para categorizar fueron las siguientes:

- hotel
- arrival_month_date
- meal
- company
- agent
- country
- market_segment
- distribution_channel
- is_repeated_guest
- reserved_room_type
- assigned_room_type
- deposit_type
- customer_type

Para todas las columnas categóricas que seguían siendo de tipo objeto, se les realizó un encoding para permitir un escalado posterior.

Se finaliza el proceso de preprocesamiento con un total de 28 características, donde 13 de ellas son categóricas.

Modelos Supervisados

Se realiza una búsqueda de hiper-parámetros mediante Grid Search, inicialmente para el modelo Random Forest usando las métricas de validación: área bajo la curva y F1 Score. También se generaron un conjunto de curvas ROC y matrices de confusión pertenecientes a los resultados.

Inicialmente se varían los parámetros *criterion* con los valores gini y entropy. También se varía el parámetro *n_estimators* con los valores: 5, 10, 20, 50, 100, 200, 250. Y finalmente el parámetro *max_features* con las opciones 4, 8, 12, auto.

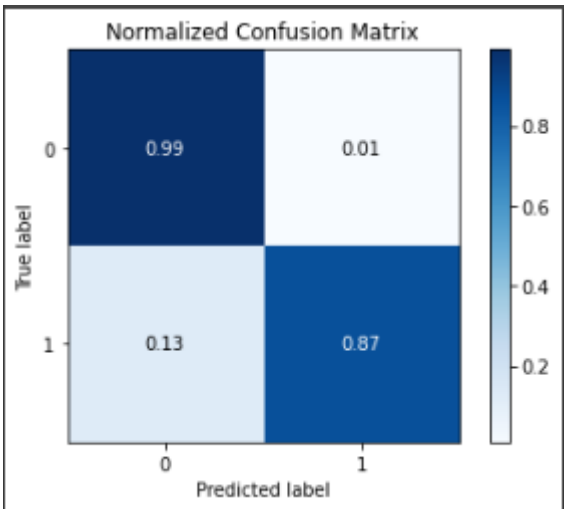
Nuestra función Grid Search arroja la siguiente combinación de hiper-parámetros como la más apropiada para nuestro modelo: class_weight: balanced, criterion: entropy y n_estimators: 200. Luego se hace uso de la función experimentar, en la cual se crean modelos usando los mejores hiperparametros, haciendo uso de validación cruzada; teniendo en cuenta que nuestro dataset tiene un ligero desbalance con relación a la variable respuesta se utiliza una validación cruzada estratificada, donde se busca que en cada fold

conserve la relación de la variable respuesta. Finalmente se obtienen cuatros métricas de validación con los siguientes resultados

	precision	**recall**	**f1-score**
0	0.93	0.99	0.96
1	0.97	0.87	0.92
accuracy			0.94

	AUC medio
0	0.868829

Finalmente obtenemos la matriz de confusión



Con lo anterior vemos reflejados muy buenos resultados pero se espera utilizar otros modelos como el RNA, SVM y Análisis discriminante cuadrático.