Entrega Informe Final

JUAN PABLO ARENAS VÉLEZ

Tutor: RAUL RAMOS POLLAN

Introducción Inteligencia Artificial 2022-1

Universidad de Antioquia

Contenido

1) Planteamiento del problema

Ante el ascenso del comercio electrónico los hoteles ubicados en zonas de alto valor turístico se llenan cada vez más a un simple click de distancia, en muchos casos con reservas que provienen de múltiples agencias de viajes online a lo largo y ancho de cualquier país que cuente con los permisos indicados para ofrecer una gran variedad de ofertas hoteleras, lo cual hace “que la habitación adecuada esté disponible para el cliente adecuado y el precio justo en el momento indicado a través del canal de distribución adecuado” - Mehrotra y Rutley.[1] Una reserva simboliza un contrato entre cliente y el hotel, lo que le da al cliente el derecho a utilizar el servicio en el futuro a un precio fijo. Por lo general, se incluye una opción para cancelar el contrato antes de la prestación del servicio. Sin embargo, la opción de cancelar un servicio antes de su prestación pone todo el riesgo en los hoteles, que deben garantizar habitaciones a los clientes que respetan sus reservas pero también asumen el coste de las habitaciones vacías cuando se cancela una reserva. Las cancelaciones tienen un impacto significativo en decisiones de gestión de la demanda en el contexto de la gestión de ingresos, con lo cual estas cancelaciones pueden representar el 20% del total de reservas recibidas por los hoteles prestadores del servicio. **Por lo cual se requiere un sistema con el objetivo de desarrollar modelos de clasificación de la probabilidad de cancelación de una reserva de hotel (clasificacion biclase)**, con el fin de poder ofertar un valor de demanda preciso, para lo cual el administrador de un hotel puede hacer una gestión de la demanda más sólida y congruente para así poder realizar la toma de decisiones y mejorar las estrategias de overbooking y las políticas de cancelación. Sin embargo, debido a las características de las variables incluidas en este conjunto de datos, su uso puede ir más allá de un problema de predicción de cancelación.

2) Obtención del dataset

El conjunto de datos del problema consiste en información anónima de reservas reales de dos hoteles en Portugal: un hotel urbano (*city hotel*) ubicado en la ciudad de Lisboa, y en un hotel resort (*resort hotel*) ubicado en la región turística del Algarve. Estos datos se tomaron protegiendo la identidad del reservante. El tamaño de la base de datos es de 119,390 muestras, las cuales serán reducidas a un total de 23,842 muestras para efecto de ahorro en el coste computacional. Además, el conjunto de datos cuenta con un total de 32 características (31 variables de entrada y una variable de salida), con algunos datos faltantes los cuales fueron rellenados en base a la media o moda de la variable específica.

Número de variables: 32 (incluyendo variable de salida)
Número de muestras original: 119,390

Número de muestras empleadas: 23,842
Nombre de la variable de salida: *is_cancelled*
Número de clases de la variable de salida: 2 (biclase)

3) Preprocesado y Limpieza de Datos

El estudio requiere clasificar las reservas de un hotel como cancelada y no cancelada. Para ello, Júpiter Notebook fue la herramienta utilizada para construir los modelos, que se ejecutan en Python usando las librerías de *Keras*, *Numpy*, *Pandas*, *Scikit Learn*, *Label Encoder*, *Scikit Plot*, *Seaborn* y *Matplotlib*.

Para este trabajo, se utilizó la validación cruzada, en particular la de *k-fold*, una técnica conocida para la evaluación de modelos (Hastie et al, 2001). El principal objetivo de la validación cruzada de *k-fold* consiste en dividir aleatoriamente los datos de la muestra dada en submuestras de tamaño k. Aunque esta técnica de evaluación fomenta que la estructura de modelos no se encuentren sobre-ajustados y se puede aplicar a datos independientes al mismo tiempo.

Para trabajar con el conjunto de datos se clasifican las variables de entrada en (categóricas y discretas), siendo las categóricas las que contienen datos como: market_segment, assigned_room_type o la variable is_canceled la cual es la salida del sistema, y las variables discretas aquellas expresadas en números enteros; para convertir estas variables categóricas en discretas se usa de la librería Scikit Learn la función **LabelEncoder**. La base de datos cuenta con valores faltantes en los siguientes campos: company, country, agent, children; para lo cual se procede a rellenar estos campos faltantes en función de la media y de la moda de estas columnas.

Como se trata de un problema de clasificación binaria con un desequilibrio de clases usarán las siguientes métricas para evaluar el desempeño de un modelo.

La base de datos contiene un 62,97% de reservas no canceladas, dejando un 37,04% de reservas que son canceladas.

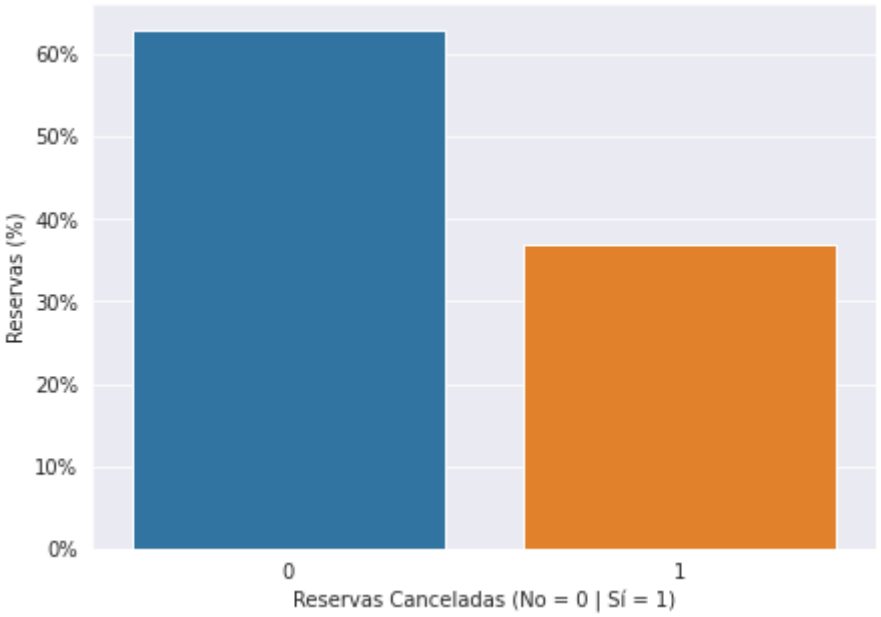


Figura 1. Porcentaje de reservas canceladas (0: No, 1: Sí)

Con respecto a la cantidad de registros de reservas de cada hotel, la base de datos está distribuida de la siguiente forma: 61.24% son reservas al hotel urbano (H2), y 38.76% son del hotel resort (H1).

Figura 2. Porcentaje de reservas entre tipo de hotel.

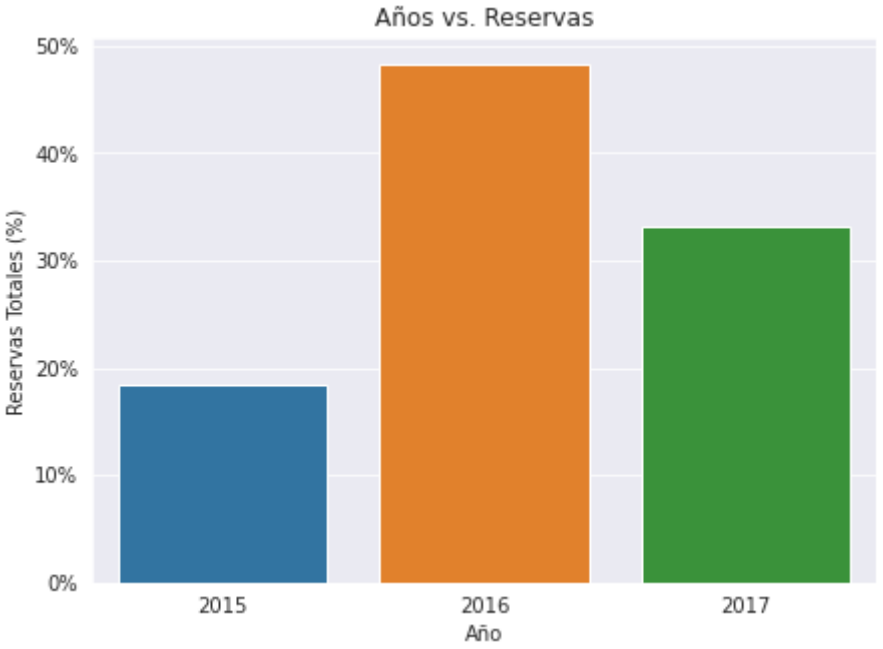
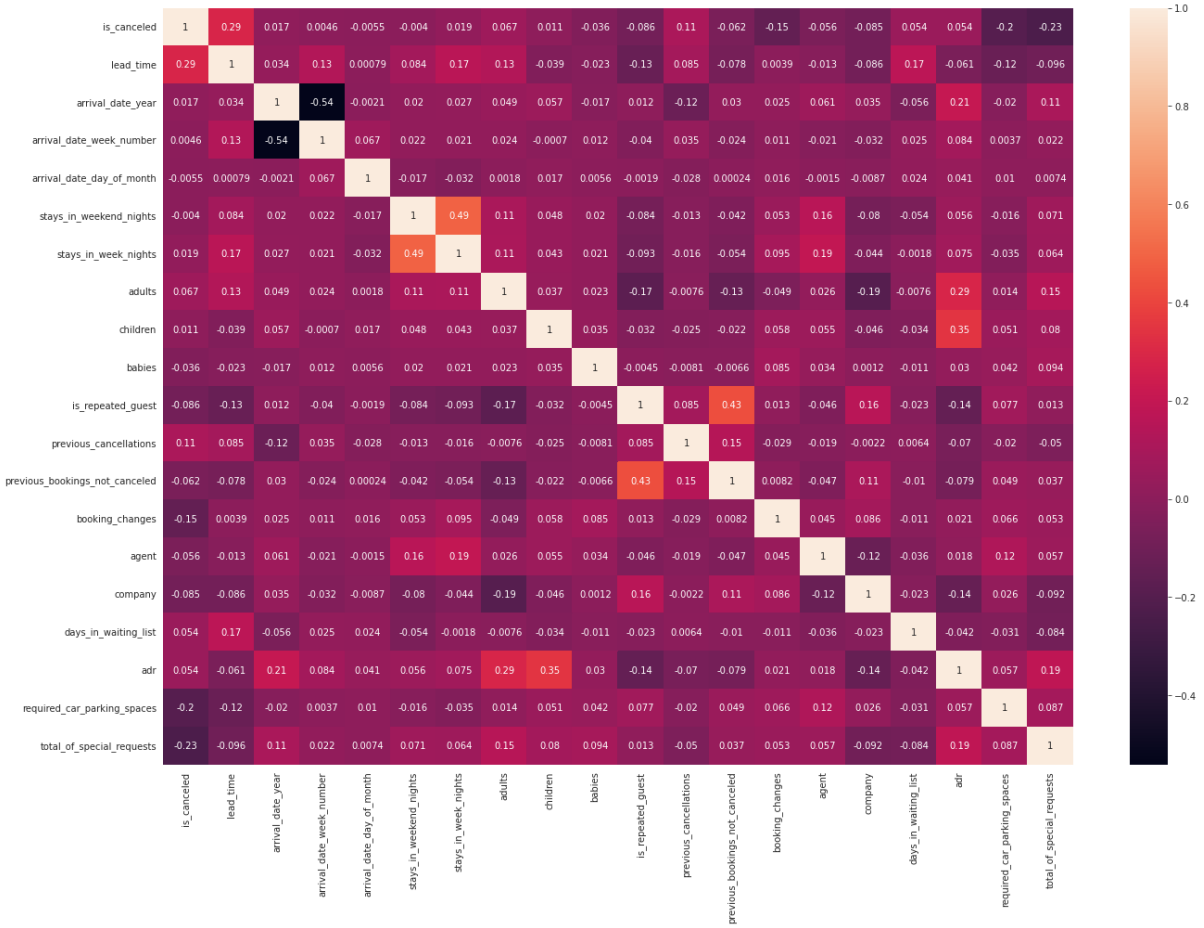


Figura 3. comparación de años y reservas

Figura 5. Matriz de correlaciones



4) Métricas

Precisión: Es la capacidad de un clasificador de no etiquetar una instancia positiva que en realidad es negativa. Para cada clase, se define como la relación entre verdaderos positivos y la suma de verdaderos y falsos positivos.

$$\frac{(True\ Positive)}{(True\ Positive + False\ Positive)}$$

Recall: Es la capacidad de un clasificador para encontrar todas las instancias positivas. Para cada clase, se define como la relación entre verdaderos positivos y la suma de verdaderos positivos y falsos negativos.

$$\frac{(True\ Positive)}{(True\ Positive + False\ Negative)}$$

F1 Score: Es una media armónica ponderada de la Precisión y de la Exhaustividad de manera que la mejor puntuación es 1.0 y la peor es 0.0. En términos generales, los F1 Score son más bajos que las medidas de precisión, ya que incorporan precisión y recuerdan su cálculo. Como regla general, el promedio ponderado de F1 debe usarse para comparar modelos de clasificadores, no la precisión global.

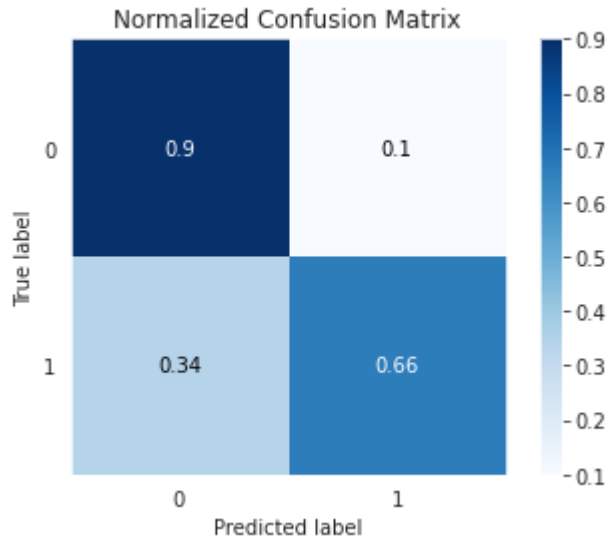
F1 Score = 2 ∗ (Recall ∗ Precision) / (Recall + Precision)

$$\frac{2\ (Recall * Precision)}{(Recall + Precision)}$$

5) Proceso y resultados de los modelos a ejecutar.

Redes Neuronales Artificiales

Épocas	Capas Ocultas	Error Clasificación	Intervalo de confianza
10	10	0.224	0.018
10	20	0.227	0.015
10	30	0.236	0.016
10	50	0.216	0.014
30	10	0.213	0.014
30	20	0.216	0.02
30	30	0.215	0.012
30	**50**	**0.196**	**0.012**
50	10	0.211	0.012
50	20	0.219	0.02
50	30	0.205	0.0177
50	50	0.199	0.016



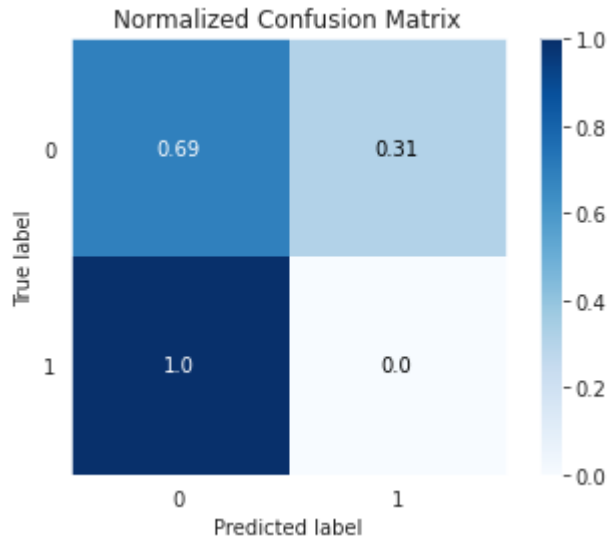
Matriz de confusión - RNA (*30 épocas, 50 neuronas*)

Clase	Precision	Recall	F1-Score	Support
0	0.83	0.88	0.86	1519
1	0.77	0.69	0.73	865

Se usa una combinación de tres diferentes épocas (10, 30 y 50) y cuatro cantidades diferentes de neuronas en una capa oculta (10, 20, 30 y 50) obteniéndose 12 configuraciones distintas.

Random Forest

Árboles	Error Clasificación	Intervalo de confianza
10	0.274	0.137
20	0.276	0.131
30	0.273	0.13
40	0.271	0.128
50	0.273	0.131
60	0.272	0.133
70	0.272	0.131
80	0.271	0.13
90	**0.269**	**0.132**
100	0.27	0.129



Matriz de confusión - Random Forest (100 árboles)

Clase	Precision	Recall	F1-Score	Support
0	1.00	0.69	0.82	2382
1	0.00	0.00	0.00	2

Se generaron diez configuraciones con distinta cantidad de árboles, observándose poca diferencia en el error de clasificación de ellas. De la tabla anterior se puede observar que las mejores configuraciones fueron las de 90 y 100 árboles, presentando errores muy parecidos.

6) Diagnóstico

En general, los resultados fueron un poco inferiores a los que se esperaba en comparación con algunos ejemplos observados en la página de Kaggle, solo que al no haber hecho uso de un balanceo de la base de datos, la precisión o la eficacia de los modelos se puede haber perdido. Se aplicaron varios modelos de clasificación binaria. Los modelos construidos alcanzaron una precisión general superior al 74,1%. Esto muestra que en nuestra situación. El algoritmo de Redes Neuronales, el cual obtuvo una precisión de 80.4% eficiencia fue el que mejor se comportó antes de hacerles reducción de dimensiones, ya que fue el que menor error obtuvo, es un gran técnica para crear modelos predictivos para cancelaciones de reservas.

Para trabajo a futuro se puede aplicar en las cadenas hoteleras más grandes del mundo, como Marriot. Internacional, que utiliza plataformas de big data para recopilar datos de una variedad de sus operaciones. Uno de los principales objetivos de estas plataformas puede ser la fijación de precios dinámicos. automatización que contribuye a optimizar los precios de las habitaciones, que aprovechan las ventajas globales y los factores económicos.
Ahora para la parte del despliegue del proyecto se usó la herramienta Google Colab para correr los algoritmos de Machine Learning en un entorno de Python 3, donde se visualizan su ejecución paso a paso. Ahora para su despliegue a producción hoy en día se cuenta con tecnologías integrables a python como lo son Django la cual facilita la creación de sitios web complejos y que pone énfasis en el re-uso, la conectividad y extensibilidad de componentes, aparte de esto, si es necesario se puede hacer uso de una plataforma como servicio (PAAS) para desplegar el modelo como lo serían los servicios de AWS para implementar y ejecutar aplicaciones de aprendizaje automático en la nube.

7) Referencias

[1] Mehrotra, R., Ruttley, J., & American Hotel & Lodging Association. (2006). “Revenue Management”. Washington, DC: American Hotel and Lodging Association.
[2]https://medium.com/analytics-vidhya/exploratory-data-analysis-of-the-hotel-booking-demand-with-python-200925230106