Entrega Final

LAURA CECILIA TOBÓN OSPINA

JUAN DAVID ARISMENDY PULGARÍN

Tutor: RAUL RAMOS POLLAN

Introducción Inteligencia Artificial 2023-1

Universidad de Antioquia

Medellìn

2023

1. Planteamiento del problema

En la actualidad el comercio hotelero se ha venido manteniendo pese a las adversidades vividas por este sector en el período del COVID-19, implementando nuevas estrategia de servicio al cliente e innovando en tipos de negocios aleatorios relacionados como catering y planes cortos de fin de semana para locales[1]. Una de las estrategias que genera confianza ha sido permitirle al cliente cancelar una reserva con antelación a su visita sin tener que justificarlo y con un tiempo prudente de anticipación, lo cual dependerá del hotel segun sus politicas ya que pueden ser cancelaciones flexibles, moderadas o firmes, esta última puede incluir hasta reembolsos para el cliente [3]. Todos los metodos para conquistar y atraer huéspedes tienen un impacto económico en los hoteles, pero las cancelaciones y habitaciones que no son ocupadas puede acarrear costes que deben ser compensadas con las reservas que si son efectivas, razón por la cual tratar de implementar un modelo que ayude a la predicción de las posibles cancelaciones ayudaría a diseñar estrategias de amortización en estas circunstancias.

2. Dataset

El dataset a utilizar proviene de una competencia de kaggle en la cual se proporcionan datos que comparan información de varios booking comparada entre dos hoteles “ a city hotel and a resort hotel.” [1]

Un 34% de la información proveniente del resort y un 66% proveniente del hotel en la ciudad

- hotel - (resort o ciudad)
- is_canceled (indica si la reserva fue cancelada o no)
- lead_time (dían transcurridos entre la reserva y la fecha de llegada al hotrl)
- arrival_date_year
- arrival_date_month
- arrival_date_week_number
- arrival_date_day_of_month
- stays_in_weekend_nights
- stays_in_week_nights
- adults
- children
- babies
- meal
- country
- market_segment
- distribution_channel
- is_repeated_guest

previous_cancellations

previous_bookings_not_canceled

reserved_room_type

assigned_room_type

booking_changes

deposit_type

agent

company

days_in_waiting_list

customer_type

adr

required_car_parking_spaces

total_of_special_requests

reservation_status

reservation_status_date

3. Métricas

Accuracy

Representa el número de instancias de datos clasificados correctamente sobre el número total de instancias de datos.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

En cuanto a la métrica de negocio sería de gran utilidad conocer la exactitud del modelo para que la industria hotelera, o un hotel en particular pueda definir estrategias para apaciguar las pérdidas por cancelaciones, con nuevas ideas, ya sea como la mencionada anteriormente en la que un hotel puede ofrecer más habitaciones de las que realmente tiene confiando en la predicción de cancelaciones

Recall (Exhaustividad)

Idealmente debería ser 1 (alto) para un buen clasificador. Recall se convierte en 1 solo cuando el numerador y el denominador son iguales, es decir, TP = TP + FN, esto también significa que FN es cero. A medida que FN aumenta, el valor del denominador se vuelve mayor que el numerador y el valor de recuperación disminuye (lo que no queremos).

$$recall = \frac{TP}{TP + FN}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Exhaustividad (recall)

[4]

Precisión

Idealmente, la precisión debería ser 1 (alta) para un buen clasificador. La precisión se convierte en 1 solo cuando el numerador y el denominador son iguales, es decir, TP = TP + FP, esto también significa que FP es cero. A medida que FP aumenta, el valor del denominador se vuelve mayor que el numerador y el valor de precisión disminuye (lo que no queremos).

$$precision = \frac{TP}{TP + FP}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Precisión (precision)

[4]

Desempeño

La idea es poder **obtener la predicción de una cancelación de reserva**, sería de gran ayuda para las cadenas hoteleras saber si un cliente realmente va a llegar, esto podría ayudar al comercio para planear la gestión de recursos ya sea de personal o de alimentos.

Conociendo el promedio de cancelaciones la industria hotelera podría sobrevender habitaciones y así mitigar el impacto de las mismas y costear el valor del proyecto de predicción

Segun el dataset a usar, las cancelaciones están alrededor del 30%

5. Procesamiento de datos

Inicialmente pudimos realizar la carga del dataset y lo visualizamos en el codelab, pudimos crear un nuevo API token para descargar el dataset.

Según las respuestas la data quedó correctamente subida, pero nos dimos cuenta de algunos datos que no eran útiles y muchos nulos que había que procesar.

```
#checking for null values
dict_={}
for feature in data.columns:
    dict_[feature]=data[feature].isnull().sum()
pd.DataFrame(dict_,index=['null_values']).transpose()
```

Tomamos decisiones como eliminar los valores de “company”, ya que habían más de 100.000 que tenían resultados nulos después de la revisión.

Posterior a ello pudimos crear una copia de la tabla de datos para poder trabajar en ella

Valores faltantes

Se evidencia que también tenemos valores faltantes, una cantidad no tan importante como lo sucedido en “company”, pero si para poner especial atención en ellos.

```
data_set.isnull().sum().sort_values(ascending=False)[:8]
```

```
agent                16340
country                488
Total_guests             4
children                 4
reserved_room_type       0
assigned_room_type       0
booking_changes          0
deposit_type             0
dtype: int64
```

agent: en este caso se identifican y se cargan con ceros los que aparecen nulos, una decisión temporal, mientras analizamos las mejores opciones para nuestro dataset

country: Para estos valores faltantes, se procede a reemplazarlos por su moda

Children: Para estos valores faltantes, se procede a reemplazarlos por su número de promedio redondeado.

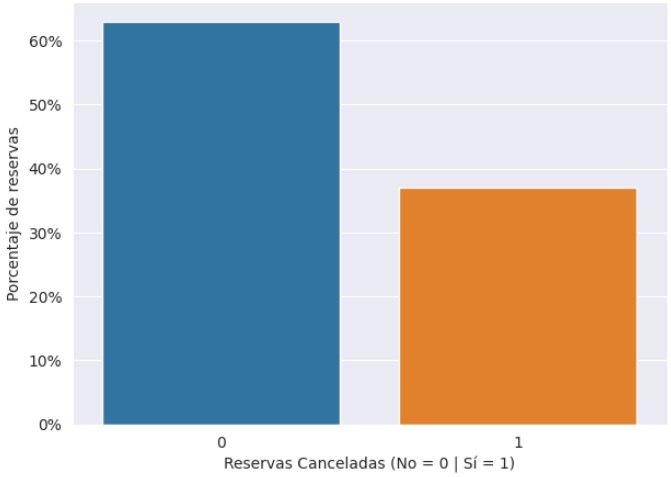
Se realizan verificaciones para comprobar que no se tengan 0 invitados

```
#double check of there are not 0 guesses in the dataset
((data_set.adults + data_set.babies + data_set.children)==0).sum()
```

```
0
```

Ahora, algunos de los tipos de datos parecen ser inadecuados por lo que tuvimos que realizar cambios a flotante y a entero en las columnas “children”, “company”, y “Agent”

Con muchos de esos datos útiles pudimos realizar graficaciones para visualizar de mejor manera la información representativa, como una comparativa de reservas y cancelaciones

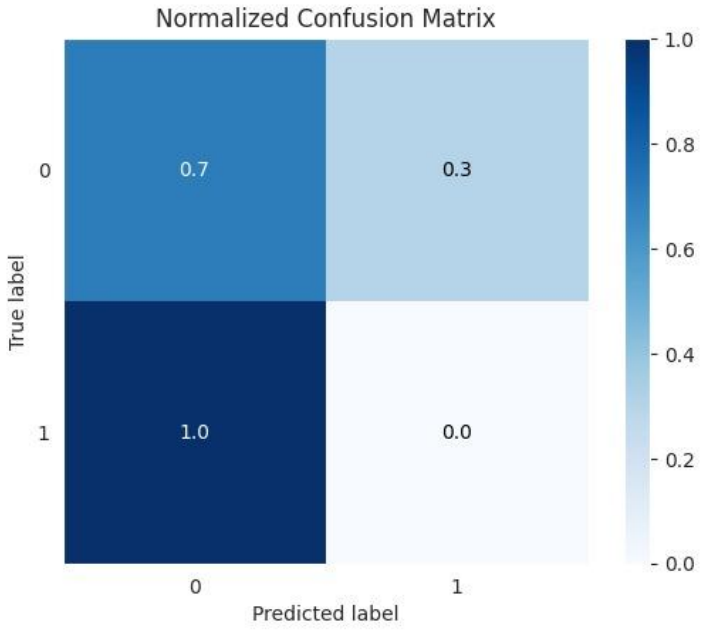


También se pudo observar que la proporción de cancelaciones entre el hotel de ciudad es más alta que en el resort, que la proporción de reservas canceladas en espera parece ser relativamente estable a lo largo del mes y que las reservas con el tipo de comida seleccionado, es decir completa, tienen más probabilidades de ser canceladas y cosas como que los clientes frecuentes tienden a mantener las reservas, más que los clientes nuevos, entre otras.

6. Modelos y resultados

- **Randomforest:** es un modelo formado por un conjunto de árboles de decisión entrenados con muestras distintas a las usadas con bootstrapping. la predicción de forma agregando las predicciones de todos los árboles individuales, este modelo es de gran utilidad ya que reduce significativamente el riesgo de overfitting y suele ser muy estable al probarse con nuevas muestras. Se afirma que no funciona bien con datasets pequeños pero en nuestro caso los datos son significativos.

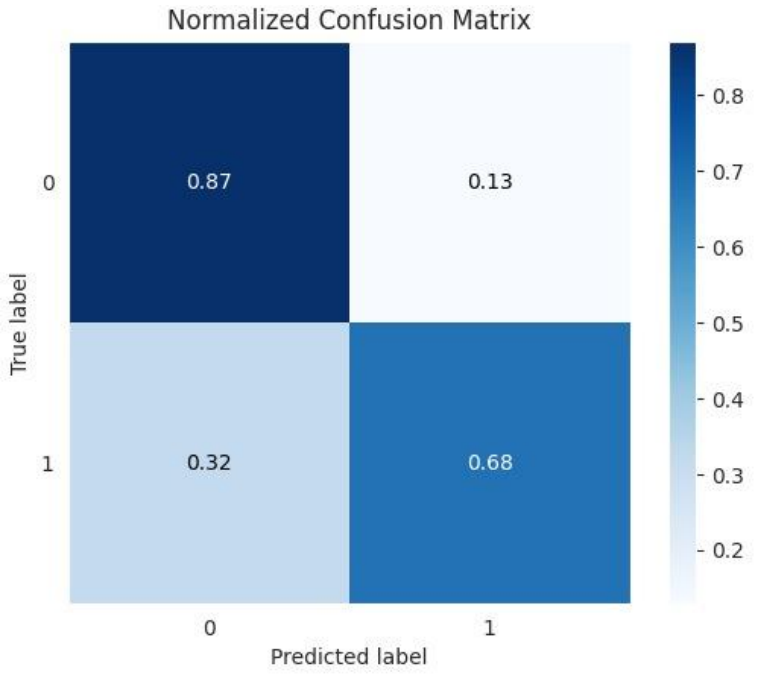
Resultados de la clasificación con 100 árboles




```
==========================

[[1661  721]
 [   2    0]]
              precision    recall  f1-score   support

           0       1.00      0.70      0.82      2382
           1       0.00      0.00      0.00         2

    accuracy                           0.70      2384
   macro avg       0.50      0.35      0.41      2384
weighted avg       1.00      0.70      0.82      2384


Resultado de clasificación con  100 árboles:

Error en la clasificación:  0.273  +/-  0.13

Tiempo total de ejecución:  23.8  segundos.
0.0
0.0
0.0
0.6967281879194631
accuracy: 0.727
```

- **Redes neuronales:**Consiste en encontrar esa relación de pesos a través de un proceso iterativo en el que, secuencialmente, se va analizando cada uno de los patrones de entrada a la red, reajustando en cada iteración la relación de pesos. Es en este punto cuando se introducirá una función de error que irá midiendo el rendimiento de la red en un momento dado, donde el objetivo será, obviamente, minimizar dicha función de error. [5]

```
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 2ms/step
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 2ms/step
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 1ms/step
75/75 [==============================] - 0s 2ms/step
[[1303  196]
 [ 282  603]]
              precision    recall  f1-score   support

           0       0.82      0.87      0.85      1499
           1       0.75      0.68      0.72       885

    accuracy                           0.80      2384
   macro avg       0.79      0.78      0.78      2384
weighted avg       0.80      0.80      0.80      2384


Epocas:  50     Neuronas Capa Oculta:  30

Error en la clasificación:  0.198  +/-  0.015

Tiempo total de ejecución:  199.0  segundos.

--------------------------
```



7. Retos y consideraciones de despliegue

Nuestro principal reto fue entender el funcionamiento de los modelos, el por qué no eran tan precisos como se esperaba y el tiempo que se gastaba en probar y entrenar cada modelo

A parte de eso pudimos identificar otro retos y consideraciones adicionales que se pueden tener en cualquier trabajo con inteligencia artificial y modelos predictivos

Sesgo de datos: uno de los mayores desafíos en este tipo de trabajos es el problema del sesgo de datos ya que los modelos se entrenan con grandes conjuntos de datos, datos que son obtenidos externamente y no se sabe realmente

el origen y confiabilidad y, si estos datos están sesgados, el modelo aprenderá a estar sesgado también. Esto puede conducir a problemas como la discriminación y el trato injusto.

Interpretabilidad: otro desafío que se tiene es el tema de la interpretabilidad. Puede ser difícil entender cómo toman decisiones los modelos de IA, y esto puede hacer que sea difícil confiar en ellos.

A pesar de estos desafíos, la IA tiene el potencial de revolucionar muchas industrias y aspectos de nuestras vidas. Es importante estar al tanto de los desafíos y las consideraciones involucradas en la IA, pero también es importante estar entusiasmado con los beneficios potenciales que la IA puede ofrecer.

El costo de entrenar e implementar modelos de IA: los modelos de IA pueden ser muy costosos de entrenar e implementar, en nuestro caso al momento de hacer las iteraciones incluso cuando se redujo el dataset , este procedimiento tomó muchísimo tiempo Esto puede ser una barrera de entrada para muchas personas.

La necesidad de habilidades especializadas: con este trabajo pudimos notar que la inteligencia artificial es un campo complejo y requiere habilidades especializadas para desarrollar e implementar modelos de IA. Esto puede dificultar la búsqueda de personal calificado.

La necesidad de mejora continua: los modelos de IA se mejoran constantemente. Esto significa que siempre se debe estar preparado para actualizar continuamente sus modelos de IA a fin de mantenerse por delante de la competencia.

A pesar de estos desafíos, la IA es un campo en rápido crecimiento con el potencial de revolucionar muchas industrias (incluso ya está pasando) y aspectos de nuestras vidas. Las empresas y organizaciones que puedan adoptar con éxito la IA tendrán una ventaja significativa sobre sus competidores.

8. Conclusiones

- Pudimos tener pérdidas en accuracy y precisión ya que no se realizó el balancea de la base de datos
- Usando el modelo random forest obtuvimos valores en las métricas bastante alejadas de 1, es decir lejanas al 100% por lo que no vimos que fuese una buena opción para concluir sobre las predicciones objetivo del proyecto
- Usando el modelo de redes neuronales pudimos obtener métricas más altas, como precisión de acierto de cancelaciones del 75% y un accuracy del 80% lo cual parece indicarnos que fue un modelo más acertado
- elegimos la prueba con epoca de 50 y neuronas capa oculta de 30 ya que fue el conjunto de variables que menor error en la clasificación, con valor de 0,198 +/- 0,015

- Entre más grandes fueran los arreglos de época y número de neuronas (arreglos a iterar) mayor costo computacional tenía la reproducción del segmento de código, y la espera llegaba por split en algunos casos de 10 a 15 minutos, a futuro se debe buscar cómo optimizar esto sin que implique mayores capacidades de máquina

link codelab:
https://colab.research.google.com/drive/1Vt2Fw4kWQpVfjMmocI5ombuw9W-djcfB?usp=sharing#scrollTo=JxPH3hSiyViM

link dataset:
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

9. Bibliografía

[1].| Kaggle. https://www.kaggle.com/code/santhosh77/hotel-booking-eda-and-cancellation-prediction/notebook

[2]. https://www.hosteltur.com/comunidad/005020_como-el-covid-19-ha-afectado-al-sector-hotelero-y-perspectiva-para-el-nuevo-ano.html

[3]. https://www.airbnb.com.co/help/article/475/

[4]. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

[5]. https://www.merkle.com/es/es/blog/prediccion-dato-redes-neuronales-artificiales#:~:text=Consiste%20en%20encontrar%20esa%20relaci%C3%B3n,iteraci%C3%B3n%20la%20relaci%C3%B3n%20de%20pesos.