[1]

Proyecto análisis de datos Hotel Booking (2022)

Alex Muñoz, *Giselle Acuña*, *Bastián Barrientos*

***Resumen*—Varias organizaciones y empresas buscan sobresalir por sobre las otras en el mercado y muchas de ellas se encuentran en la industria hotelera. Si un hotel quiere competir con los demás hoteles, debe hacer uso de las oportunidades que se presentan, una de ellas es la información de sus clientes respecto a sus reservas.**

El objetivo de este estudio es determinar los beneficios que pueden traer los distintos clientes a los hoteles, de esta manera se tendrá una serie de factores sobre cada cliente que solicita una reserva hotelera. Con este fin, la pregunta es la siguiente: ¿Qué tipo de cliente es el más recomendable o trae más beneficios a los hoteles?

La pregunta de investigación se responde a través de ciertos análisis exploratorios en el conjunto de datos y aplicación de modelos de clasificación. Dichos resultados muestran un cierto una cierta tendencia en el tipo de cliente que es más recomendable para los hoteles, siendo que cuando se encuentran niños en el la reserva hay más beneficios.

Finalmente, cabe destacar que las problemáticas planteadas no fueron tan complicadas, siendo el análisis exploratorio de datos la metodología que respondió la mayoría de preguntas. Además, se pudieron haber evaluado más modelos clasificadores y haber gestionado los hiper parámetros, lo cual no se hizo por una mala gestión del tiempo.

I. Introducción

El equipo de T4 fue contratado para realizar el análisis de un set de datos de reservas hoteleras [3], del cual se espera obtener la mayor información sobre el comportamiento de los huéspedes que reservan y cómo optimizar la ganancia de los hoteles.

La motivación de nuestro equipo detrás de este proyecto, es poder tener una mejor comprensión en cuanto al funcionamiento detrás de la industria hotelera desde la reserva.

En cuanto a la estructuración trabajamos con dos datasets con la misma estructura, el primer dataset corresponde a un resort hotel y el segundo dataset corresponde a un city hotel. Las dimensiones de estos 2 datasets son: Resort Hotel (H1) - 40,060 observaciones City Hotel (H2) - 79,330 observaciones lo que nos da un total de 119,390 observaciones y 36 columnas. Cada observación representa una reserva de hotel realizada entre el 01 de Julio del 2015 y el 31 de Agosto del 2017.

Como es sabido, el análisis descriptivo puede ser aplicado para entender de manera profunda patrones, tendencias y anomalías en los datos. Por lo que el análisis de este dataset en particular nos permite encontrar respuestas a diferentes interrogantes, como la predicción de la cancelación de la reserva, la segmentación de los clientes, la satisfacción de los clientes y la estacionalidad.

[1]

II. Metodología utilizada

En el presente trabajo se plantearon una serie de problemáticas e inquietudes relacionadas a las reservas de hoteles. Si bien todas están relacionadas en torno al tipo de cliente ideal para los hoteles, era necesario separar y especificarlas en distintas preguntas para abordar el problema.

Además, se realizó al principio un preprocesamiento en el conjunto de datos para eliminar o editar columnas que tengan outliers o datos faltantes.

Si bien el conjunto de datos a analizar tenía muchas dimensiones o atributos, a ciertas columnas se les prestó especial atención debido a su relación con las preguntas planteadas. Por lo cual, se establecieron ciertos atributos de interés:

- **adr**: Tarifa media diaria (calculada dividiendo la suma de todas las transacciones de alojamiento por el número total de noches de estancia).
- **adults**: Número de adultos en la reserva.
- children: Número de niños en la reserva.
- **babies**: Número de bebés en la reserva.
- **is_canceled**: Valor que indica si la reserva fue cancelada (1) o no (0).
- **lead_time**: Número de días transcurridos entre la fecha de entrada de la reserva en el PMS y la fecha de llegada.

A. ¿De qué manera se podría predecir la cancelación de una reserva por parte del cliente a partir de los datos disponibles? ¿Qué beneficios aportará a los hoteles esta predicción?

Para responder esta interrogante se utilizaron ciertos modelos de clasificación supervisada. Se seleccionaron 3 modelos que podrían servir basándose en las características del conjunto de datos:

Árbol de decisión: Este modelo tiene la ventaja de que es relativamente fácil de entender e interpretar, además puede analizar atributos cualitativos y cuantitativos para lograr la clasificación. En el presente caso, se analizarán los datos cuantitativos del dataset solamente.

Naïve Bayes: Este modelo suele ser usado para realizar clasificaciones binarias, o sea de 2 clases posibles. Se opta por utilizar este modelo debido a que dentro de sus características se encuentra su robustez ante los valores ruidosos y ante atributos irrelevantes. Siendo estas 2 características presentes en algunos campos del dataset, como el ruido que hay en la columna lead y los valores irrelevantes al ser reemplazados por cero en la limpieza.

[1]

SVM: El modelo Support Vector Machines es utilizado para clasificación, regresión y detección de valores atípicos. Se usará en este dataset porque es eficaz cuando hay grandes dimensiones y es eficiente en memoria.

Aunque se pudo haber probado el algoritmo KNN, se optó por ignorarlo ya que funciona mal cuando hay una alta dimensionalidad en el dataset y es más costoso en potencia y tiempo. Cabe mencionar que se probó el algoritmo KNN y se cayó el entorno Google Colab por falta de memoria RAM.

Además, se gestionó el balance de clases en el conjunto de entrenamiento aplicando oversampling y subsampling cuando fue necesario. De esta manera, se reducirá el posible sesgo del modelo al hacer predicciones en nuevos datos.

Finalmente, para comparar y evaluar estos modelos planteados, se usaron las métricas: Accuracy, Precision, Recall, F1.

B. ¿Qué tipo de cliente (adultos, adultos con niños, adultos con bebés, etc) produce una mayor tarifa diaria para los hoteles (campo adr)? ¿Qué beneficios aportaría a los hoteles esta información?

Esta interrogante se pudo responder en el análisis exploratorio de datos, para ellos se crearon 3 nuevos campos en el conjunto de datos con tal de categorizar a los clientes que hacen reservas en los hoteles:

- Solo adultos:
- Adultos con niños:
- Adultos con bebés:
- Adultos con bebés y niños.

Dichos campos se crearon a través de cálculos hechos con los datos del mismo conjunto de datos.

Luego de eso, se calculó el promedio de tarifa diaria para cada tipo de cliente. De esa manera, se consiguió una muestra representativa de la rentabilidad de cada tipo de cliente para los hoteles.

C. ¿Qué tipo de cliente tiene un tiempo de llegada mayor luego de realizar la reserva a un hotel? ¿Sería de utilidad esta información para los hoteles?

Usando los mismos campos creados anteriormente, se calculó el tiempo de llegada promedio de cada tipo de cliente.

D. ¿A qué tipo de cliente deberían dirigirse los hoteles para mejorar sus ingresos?

Se puede llegar a la respuesta viendo los promedios de tiempo de llegada y tarifa diaria para cada tipo de cliente, el cliente que tenga un menor tiempo de llegada y una mayor tarifa diaria sería el más recomendable.

III. Resultados

A. ¿De qué manera se podría predecir la cancelación de una reserva por parte del cliente a partir de los datos disponibles? ¿Qué beneficios aportará a los hoteles esta predicción?

Los resultados de cada modelo clasificador empleado se midieron a través de métricas en distintos conjuntos de entrenamiento. Esto es debido a que las clases estaban desbalanceadas en el conjunto de entrenamiento resultante. Por lo cual, cada modelo clasificador terminó con 3 conjuntos de entrenamiento:

- Conjunto de entrenamiento original.
- Conjunto de entrenamiento Sobre-muestreado
- Conjunto de entrenamiento Sub-muestreado

Resultados del modelo Árbol de decisión:

```
Datos originales
Puntaje de métricas Árbol de decisión con datos originales:
Accuracy: 0.7932992712957534
Precision: 0.7181960258986381
Recall: 0.7273878437047757
F1: 0.722762711483587

Datos Sobre muestreados
Puntaje de métricas Árbol de decisión con datos sobre muestreados:
Accuracy: 0.7907027389228578
Precision: 0.7144003566651805
Recall: 0.7246743849493488
F1: 0.7195006959723408

Datos sub muestreados
Puntaje de métricas Árbol de decisión con datos sub muestreados:
Accuracy: 0.7686070860206048
Precision: 0.6600208244957773
Recall: 0.7740141099855282
F1: 0.7124867306371376
```

Fig. 1: Resultados de predicción en el modelo Árbol de decisión

Resultados del modelo Naïve Bayes:

```
Datos originales
Puntaje de métricas modelo Naïve Bayes con datos originales:
Accuracy: 0.4786330513443337
Precision: 0.4125807640815693
Recall: 0.9616497829232996
F1: 0.577425968418623

Datos Sobre muestreados
Puntaje de métricas modelo Naïve Bayes con datos sobre muestreados:
Accuracy: 0.47623754083256553
Precision: 0.4116136258303723
Recall: 0.9639562228654125
F1: 0.5768918480025982

Datos sub muestreados
Puntaje de métricas modelo Naïve Bayes con datos sub muestreados:
Accuracy: 0.475081665131083
Precision: 0.4111926358156643
Recall: 0.965629522431259
F1: 0.5767771039587244
```

Fig. 2: Resultados de predicción en el modelo Naïve Bayes

Resultados del modelo Support Vector Machines:

```
Datos originales
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWar
  ConvergenceWarning,
Puntaje de métricas modelo Support Vector Machines con datos originales:
Accuracy: 0.4103023703827791
Precision: 0.384803576457336
Recall: 0.9887391461649783
F1: 0.5539985809852017

Datos Sobre muestreados
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWar
  ConvergenceWarning,
Puntaje de métricas modelo Support Vector Machines con datos sobre muestreados:
Accuracy: 0.43442499371806687
Precision: 0.3942562538576045
Recall: 0.9821816208393632
F1: 0.5626570636544989

Datos sub muestreados
Puntaje de métricas modelo Support Vector Machines con datos sub muestreados:
Accuracy: 0.6365022196163833
Precision: 0.9578713968957872
Recall: 0.019536903039073805
F1: 0.038292780215396886
```

Fig. 3: Resultados de predicción en el modelo Support Vector Machines

B. ¿Qué tipo de cliente (adultos, adultos con niños, adultos con bebés, etc) produce una mayor tarifa diaria para los hoteles (campo adr)? ¿Qué beneficios aportaría a los hoteles esta información?

```
Promedio de tarifa diaria solo adultos: 97.37077704483092
Promedio de tarifa diaria adultos con bebes: 112.22223719676549
Promedio de tarifa diaria adultos con niños: 158.20546405228757
Promedio de tarifa diaria adultos con bebes y niños: 152.09565714285714
```

Fig. 4: Promedio de tarifa diaria por tipo de cliente

C. ¿Qué tipo de cliente tiene un tiempo de llegada mayor luego de realizar la reserva a un hotel? ¿Sería de utilidad esta información para los hoteles?

```
Promedio de dias de llegada solo adultos: 105.3226026277054
Promedio de dias de llegada adultos con bebes: 79.75471698113208
Promedio de dias de llegada adultos con niños: 89.73761140819964
Promedio de dias de llegada adultos con bebes y niños: 68.61714285714285
```

Fig. 5: Promedio de días de llegada por tipo de cliente.

IV. Análisis de resultados

A. ¿De qué manera se podría predecir la cancelación de una reserva por parte del cliente a partir de los datos disponibles? ¿Qué beneficios aportará a los hoteles esta predicción?

Resultados del modelo Árbol de decisión:

El modelo Árbol de clasificación en cada conjunto de entrenamiento tiene resultados parecidos en las métricas, por un margen muy pequeño dió mejores resultados en los datos de entrenamiento originales. Por lo cual, se elige el modelo entrenado con los datos originales.

Resultados del modelo Naïve Bayes:

El modelo Naïve Bayes tiene resultados muy similares en cada conjunto de entrenamiento, al parecer entre los modelos entrenados en datos originales y sub-muestreados hay una calificación casi igual a simple vista. En ese caso se opta por usar el modelo entrenado con datos sub-muestreados, ya que puede tener un menor sesgo al predecir nuevos datos.

Resultados del modelo Support Vector Machines:

El modelo SVM tiene resultados diferentes en cada conjunto de entrenamiento, aún así da la impresión que el promedio de sus métricas es similar. En ese caso, se opta por usar el modelo entrenado con datos sub-muestreados, ya que a pesar de tener un bajo valor en Recall las demás métricas son bastante altas puede haber un menor sesgo.

B. ¿Qué tipo de cliente (adultos, adultos con niños, adultos con bebés, etc) produce una mayor tarifa diaria para los hoteles (campo adr)? ¿Qué beneficios aportaría a los hoteles esta información?

A partir de los resultados, podemos concluir que los adultos acompañados por niños tienen una tarifa diaria superior que cuando van adultos solamente, los hoteles probablemente deberían atraer más adultos acompañados de niños y/o bebés, teniendo en cuenta además que la mayoría de las reservas son solo para adultos.

C. ¿Qué tipo de cliente tiene un tiempo de llegada mayor luego de realizar la reserva a un hotel? ¿Sería de utilidad esta información para los hoteles?

A partir de los resultados, podemos concluir que cuando hay una reserva solamente para adultos suele haber un tiempo de llegada superior a cuando van acompañados de niños y/o bebés. Por lo cual, nuevamente es conveniente para los hoteles atraer adultos acompañados de niños y/o bebés.

V. Conclusiones

A partir de los resultados, se puede concluir que un modelo de clasificación supervisada puede predecir la cancelación de una reserva. Este tipo de información puede ser de utilidad para los hoteles dependiendo de la exactitud del modelo, ya que si se sabe de antemano que un cliente cancelará la reserva se puede empezar a ofrecer la habitación desde antes.

Además, según lo visto en el comportamiento de cada cliente, los hoteles deberían atraer adultos acompañados con niños, bebes o ambos; ya que en general tienen un tiempo menor de llegada y una tarifa media superior a los adultos solitarios en su estadía.

Cabe mencionar que el presente trabajo es muy mejorable, debido a que se pudieron haber planteado interrogantes más complicadas y probar más modelos de clasificación a través de o mejorar los ya existentes con los mejores hiper parámetros.

VI. Referencias

[1] [1] Kumar, B. (2021, 17 octubre). Pandas Replace Nan With 0. Python Guides. https://pythonguides.com/pandas-replace-nan-with-0/

[2] [2] Myrianthous, G. (2022, 26 febrero). How to Count NaN Values in pandas | Towards Data Science. Medium. https://towardsdatascience.com/count-nan-values-pandas-27a50acfc929

[3] [3] Hotel Booking. (2021, 29 junio). Kaggle. https://www.kaggle.com/datasets/mojtaba142/hotel-booking

[4] [4] Petrou, T. (2021, 29 noviembre). Finding the Percentage of Missing Values in a Pandas DataFrame. Medium. https://medium.com/dunder-data/finding-the-percentage-of-missing-values-in-a-pandas-dataframe-a04fa00f84ab

[5] [5] Yan, N. (2022, 4 junio). How to delete a column in pandas. Educative: Interactive Courses for Software Developers. https://www.educative.io/edpresso/how-to-delete-a-column-in-pandas

[6] [6] Custer, C. (2022, 13 abril). Tutorial: Add a Column to a Pandas DataFrame Based on an If-Else Condition. Dataquest. https://www.dataquest.io/blog/tutorial-add-column-pandas-dataframe-based-on-if-else-condition/

[7] [7] Z. (2022, 20 abril). How to Calculate Conditional Mean in Pandas (With Examples). Statology. https://www.statology.org/conditional-mean-pandas/

[8] [8] Joshi, S. (2021, 25 febrero). Función Pandas DataFrame DataFrame.boxplot(). Delft Stack.

https://www.delftstack.com/es/api/python-pandas/pandas-dataframe-dataframe.boxplot-function/

[9] [9] GeeksforGeeks. (2021, 4 marzo). Bar Plot in Matplotlib. https://www.geeksforgeeks.org/bar-plot-in-matplotlib/

[10] [10] sklearn.svm.LinearSVC. (s. f.). Scikit-Learn. Recuperado 28 de junio de 2022, de https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[11] A. (2022, 29 junio). ajeloy/Hito3_G6_repositorio. GitHub. Recuperado 29 de junio de 2022, de https://github.com/ajeloy/Hito3_G6_repositorio