Universidad de
Buenos Aires

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

75.06/95.58 – Organización de Datos

Curso 02
1C 2023

Trabajo Práctico 1: Reservas de Hotel
Checkpoint 1

Grupo 14

Integrantes:

- *BOTOSHANSKY, ALAN* *108933*
- *BOTOSHANSKY, IVAN* *108897*
- *NIEVA, ULISES TADEO* *107227*

Corrector: Nacho

Fecha de entrega: *13* de *abril* del 2023

Universidad de
Buenos Aires

Resumen

El siguiente reporte tiene por objetivo describir las tareas realizadas, la información relevante sobre el dataset e incluir visualizaciones para mostrar algunos de nuestros hallazgos a lo largo de la realización del checkpoint 1 del trabajo práctico.

Desarrollo

Comenzamos el análisis a partir de la exploración inicial de los datos pertenecientes al dataset proporcionado llamado “hotels_train.csv”.

En primer lugar, analizamos los distintos tipos de variable existentes en el set de datos. Como vimos en clase, los tipos posibles son cuantitativos y cualitativos. En el set original había 32 variables. Primero que todo, eliminamos la variable “reservation_status_date” ya que “reservation_status” había sido removida por la cátedra, por lo que ésta ya no tenía sentido. Luego, nos quedó un total de 31 variables, 16 cuantitativas y 15 cualitativas. Para las cuantitativas, calculamos las medidas de resumen para cada una, y para las cualitativas, hicimos una breve descripción de qué es lo que representa cada una y qué valores pueden tomar.
Además, pensamos qué variables podrían ser irrelevantes respecto a nuestro objetivo, que va a ser predecir el valor de la variable “is_canceled” (target) en función de los valores de las demás variables y concluimos que una de ellas es la variable “id”.
Luego, realizamos un análisis gráfico de las distribuciones de las variables, es decir, graficamos las distintas observaciones de las variables de distintas formas, con el fin de ver cómo éstas se comportan. Algunas, presentan una distribución más equilibrada, por ejemplo “arrival_date_day_of_month”, “is_canceled”, mientras que otras (la gran mayoría) presentan una más desequilibrada y asimétrica, entre ellas “country”, “adults”, “children”, “meal”, etc.
Después, analizamos las correlaciones entre las variables cuantitativas. Esto lo hicimos a través de una matriz de correlación (de Pearson), y luego lo mostramos mediante un heatmap. En general, casi no hay correlación entre las variables, salvo dos casos que analizamos con detalle en la notebook.
Asimismo, continuamos con el desarrollo de la relación de las variables con el target, “is_canceled”, mediante más visualizaciones (barplots, boxplots, heatmaps). A partir de ellas sacamos conclusiones acerca de qué variables podrían ser las más influyentes a la hora de predecir el valor de nuestro target, como por ejemplo “hotel”, “lead_time”, “stays_in_week_nights”, “adults”, “country”, “assigned_room_type”, entre otros.

En segundo lugar, elaboramos más visualizaciones para hacer énfasis en las observaciones más significativas a partir del desarrollo de la exploración inicial de las variables. Llevamos a cabo un pairplot y un violinplot de las variables “adults”, “children” y “babies” con el fin de anticiparnos a posibles outliers y analizar la relación entre sus distribuciones. Además, hicimos un heatmap para ver la relación entre el tipo de habitación reservada y el tipo de habitación asignada con el fin de observar cómo ésta afecta a la cancelación o no de una reserva cuando éstas coinciden o difieren. Luego, seguimos combinando variables cuya tendencia hacia alguno de los dos valores que puede tomar nuestro target fue significativa (por ejemplo: “country”, “deposit_type”, “total_of_special_requests”, entre otras). Esto lo hicimos para poder interpretar mediante piecharts los porcentajes de reservas canceladas y no

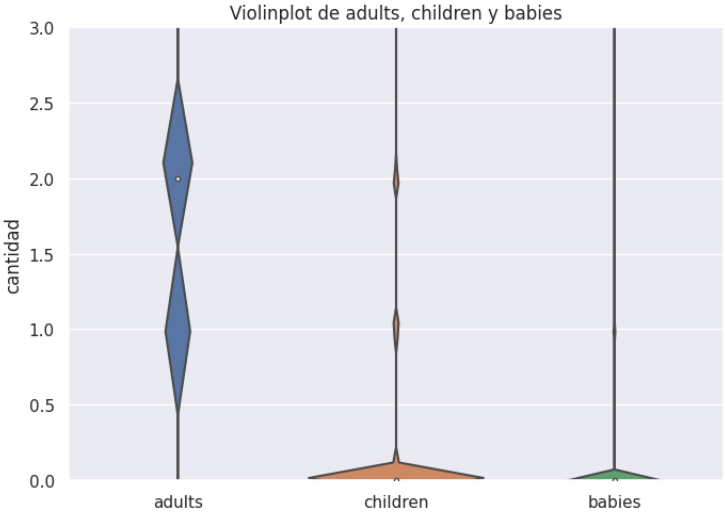
canceladas y cuántos casos abarcan respecto a la cantidad total de reservas del dataset.

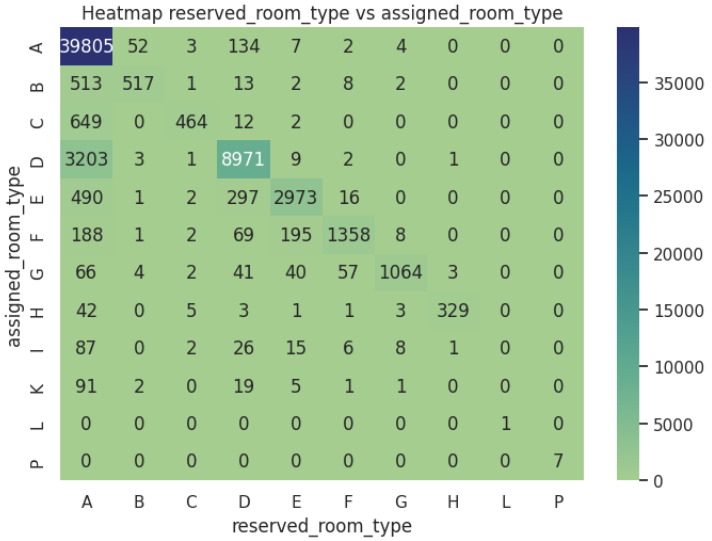
En tercer lugar, analizamos los datos faltantes del dataset. Observamos que faltaban datos en las variables “children”, “country”, “agent” y “company”. Respecto a “children” y “country”, el porcentaje de faltantes era casi nulo respecto al tamaño del dataset, por lo que a la primera variable la imputamos por su mediana y a la segunda por el valor “Undefined”. Respecto, a “agent” y “company”, que tenían un porcentaje de faltantes más significativo, también los imputamos por “Undefined”.

Por último, utilizamos distintos métodos para identificar valores atípicos tanto de forma univariada como multivariada. Para la primera, elegimos usar el método IQR, clasificamos las observaciones según si no eran valores atípicos, o si lo eran de forma moderada o severa. Además, graficamos boxplots para poder visualizarlos en cada variable. Para la segunda, nos inclinamos por la Distancia de Mahalanobis. La calculamos para cada observación del dataset y determinamos un umbral para filtrar los valores atípicos. Además, también detectamos dichos valores mediante el método de Isolation Forest. Comparamos los resultados de ambos métodos y decidimos quedarnos con el de Mahalanobis ya que creemos que es más coherente.

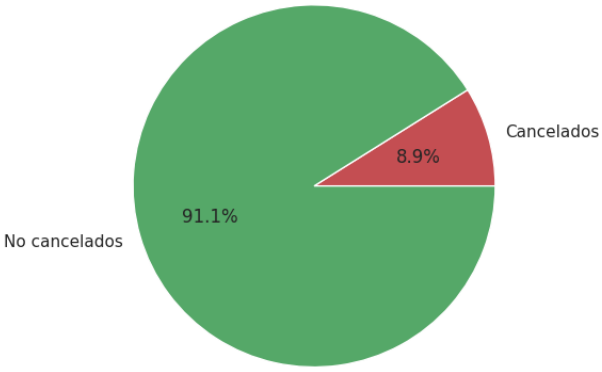
Universidad de
Buenos Aires

Visualizaciones

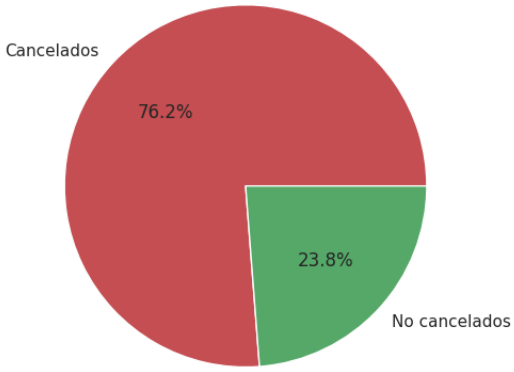




Porcentaje de cancelaciones cuando el tipo de habitación asignado y reservado difieren



Porcentaje de cancelaciones PRT + 0 special_requests



Estas son algunas de las visualizaciones que acompañan nuestros hallazgos. Las descripciones de las mismas están incluidas en la notebook.

Conclusión

El dataset original contiene una gran cantidad de variables y sucede que algunas van a ser más influyentes que otras a la hora de hacer las predicciones. Además, es importante saber cómo lidiar con los datos faltantes y los valores atípicos para que nuestras predicciones estén lo más cerca posible de lo que sucede en la realidad. Es fundamental llevar a cabo una exploración de las variables y analizarlas en profundidad porque necesitamos conocerlas al detalle para luego poder realizar correctamente las predicciones del valor del target, en este caso el valor de “is_canceled”.