Trabajo Practico 1

Checkpoint 1

Exploracion incial

A continuacion se hara una exposicion resumida del proceso de la exploracion inicial. Se podra acceder al notebook con la informacion completade esta seccion y las demas a traves del siguiente link:
https://colab.research.google.com/drive/1TPtbpLvU7TZHx7fK5QtKd-Jlt_MLBXth#scrollTo=4OKYYzPcvuF0

Tipo de variable

Analisis en cuanto a como estan definidas las variables en el dataset sin preprocesar.

Variable	Tipo
hotel	string
lead_time	int64
arrival_date_year	int64
…	…
reservation_status	object
reservation_status_date	object
id	object
is_canceled	int64

Variables cuantitativas

Calculo de medidas de resumen para las variables cuantitativas

Variable	mean	std	min	25%	50%	75%	max
lead_time	112.2	110.7	0.0	23.0	78.0	172.0	629.0
…	…	…	…	…	…	…	…
adr	102.4	47.8	-6.4	70.0	95.0	126.0	510

Variables cualitativas

Posibles valores y frecuencia

Variable	Valores	Frecuencia	Frecuencia Relativa
hotel			
	City Hotel	42129	0.68
	Resort Hotel	19784	0.32
meal			
	BB (solo desayuno)	47837	0.77
	HB (desayuno y cena)	7452	0.12
	FB (desayuno almuerzo y cena)	5556	0.09
	SC (ninguna comida)	591	0.0095
	Undefined	477	0.0077
country			
	PRT	27950	
	GBR	5733	
	FRA	4809	
	ESP	4210	

	...	...	

Variables irrelevantes para el analisis

- **Id**: La unica informacion que otorga es para distinguir una fila de otra. Esto se soluciona elminando los valores duplicados del dataframe.
- **reservation_status**: La variable enfoque “is_cancelled” nos otorga informacion mas precisan sobre lo que queremos analizar.
- **reservation_status_date**: Porque acompania la variable anterior
- **company**: Porque no aporta suficiente informacion al estudio debido a que la mayoria de las reservas no tiene compania asociada.

Datos faltantes

El analisis mostro que las variables Company, Agents y Country presentaban datos faltantes.

Company: Se tomo la decision de eliminar la columna entera para el analisis ya que solo un 5% de las filas del data set tiene informacion sobre esta variable.

Agents: Se tomo la decision de transformar la variable en categorica para que no colisione al momento de entrenar un modelo usando un arbol de decision.

Country: Se tomaron todos los datos faltantes y todos los paises que tenian una unica aparicion en el dataset y se los reemplazo por un “pais generico”.

Children: Al ser solo 4 casos los que estaban indefinidos se tomo la decision de reemplazarlos por el valor con mas apariciones en este caso 0.

Outliers

Adr: Utilizando Z-score modificado detectamos 1087 outliers y basándonos la regla de oro, decidimos eliminar estas reservas ya que modificaban ligeramente los cuartiles de la variable, incluimos un gráfico antes y después de la limpieza en la notebook.

Lead_time: Utilizando Z-score modificado detectamos 1079 outliers y basándonos en la regla de oro, decidimos eliminar estas reservas ya que modifican ligeramente los cuartiles de la variable.

Adults: Utilizando Z-score modificado detectamos 2968 outliers, basándonos en la regla de oro y con cosas que pudimos comprobar en el checkpoint 2 (genera una mejora en la precisión al predecir), decidimos eliminar estas reservas.

Reservas sin personas: Para las reservas que no tienen ingresados adultos, niños y bebes, decidimos borrarlas ya que representan menos del 1% del total de los datos.

Children: No tiene outliers buscados con z-score modificado.