Proyecto Segunda Entrega

POR:

Tomas Edil Urango Ruiz

Lucas Bustamante

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

2023

informe

Se tiene interés en averiguar cuántas de sus reservas son canceladas, y poder predecir cuándo se cancelará una reserva, por lo cual con el código desglosamos el porcentaje de reservas canceladas en el hotel, el cual es de un 37%.

La cantidad de reservas que son canceladas es significativa, y por lo tal tener una herramienta con la cual predecir cancelaciones tiene el potencial de ser útil para la hotelera.

Tenemos que para los modelos de predicción de las cancelaciones hoteleras se elaboró la exploración de los datos previamente obtenidos, se hizo el procesamiento y con la ejecución de código procedimos a realizar la revisión de cantidad y porcentaje de valores nulos obteniendo datos relevantes en los factores country, agent y company.

country	488	0.408744
market_segment	0	0.000000
distribution_channel	0	0.000000
is_repeated_guest	0	0.000000
previous_cancellations	0	0.000000
previous_bookings_not_canceled	0	0.000000
reserved_room_type	0	0.000000
assigned_room_type	0	0.000000
booking_changes	0	0.000000
deposit_type	0	0.000000
agent	16340	13.686238
company	112593	94.306893

Revisamos qué clase de variable contienen las columnas de valores nulos, para poder llenarlas adecuadamente

Los valores nulos de la columna "country" pueden reemplazarse arbitrariamente por "desconocido", luego, los demás valores numéricos representan identificaciones, además de hijos, variables que podemos reemplazar con cero razonablemente. usando un heatmap, podemos ver qué columnas presentan correlación con nuestro dato de interés y seguido podemos visualizar la correlación de las variables a continuación

```
text/plain
  is_canceled                      1.000000
  lead_time                        0.293123
  total_of_special_requests        0.234658
  required_car_parking_spaces      0.195498
  booking_changes                  0.144381
  previous_cancellations           0.110133
  is_repeated_guest                0.084793
  company                          0.082995
  adults                           0.060017
  previous_bookings_not_canceled   0.057358
  days_in_waiting_list             0.054186
  adr                              0.047557
  agent                            0.046529
  babies                           0.032491
  stays_in_week_nights             0.024765
  arrival_date_year                0.016660
  arrival_date_week_number         0.008148
  arrival_date_day_of_month        0.006130
  children                         0.004393
  stays_in_weekend_nights          0.001791
  Name: is_canceled, dtype: float64
```

Análogamente, veamos qué columnas numéricas tenemos, la porción categórica del data frame, pasamos las fechas a un formato legible por el modelo y luego, descartamos las columnas redundantes.

Para las columnas de variables numéricas, un método de normalización ayuda a que el modelo no le dé prioridad indebida a ciertas columnas, visualizamos a continuación la varianza de las diferentes columnas sin normalizar, se aplicará un método de normalización logarítmica para aquellas columnas que lo requieren, cambiar el método puede afectar el puntaje del modelo.

```
Resultado de la ejecución
10KB
  text/plain
    hotel  meal  market_segment  distribution_channel  reserved_room_type  \
    0      0     0              0                     0                 0.0
    1      0     0              0                     0                 0.0
    2      0     0              0                     0                 1.0
    3      0     0              1                     1                 1.0
    4      0     0              2                     2                 1.0
```

```
   deposit_type  customer_type  reservation_status_year  \
0             0              0                        0
1             0              0                        0
2             0              0                        0
3             0              0                        0
4             0              0                        0
```

```
   reservation_status_month  reservation_status_day  ...  babies  \
0                         7                       1  ...       0
1                         7                       1  ...       0
2                         7                       2  ...       0
3                         7                       2  ...       0
4                         7                       3  ...       0
```

```
   is_repeated_guest  previous_cancellations  previous_bookings_not_canceled
0                  0                       0                               0
1                  0                       0                               0
2                  0                       0                               0
3                  0                       0                               0
4                  0                       0                               0
```

	agent	company	days_in_waiting_list	adr	\
0	0.000000	0.0	0.0	0.000000	
1	0.000000	0.0	0.0	0.000000	
2	0.000000	0.0	0.0	4.330733	
3	5.720312	0.0	0.0	4.330733	
4	5.484797	0.0	0.0	4.595120	

	required_car_parking_spaces	total_of_special_requests
0	0	0
1	0	0
2	0	0
3	0	0
4	0	1

[5 rows x 27 columns]

Finalmente, obtuvimos nuestro dataframe "X" preprocesado, y hacemos splits de test y de train. Se puede jugar con el test size para probar en el modelo.

En nuestro caso el modelo obtuvo una predicción de hasta un 99.5%

	Model	Score
7	Cat Boost	0.995274
11	ANN	0.993038
6	XgBoost	0.982076
9	LGBM	0.963370
10	Voting Classifier	0.963034
3	Random Forest Classifier	0.953108
2	Decision Tree Classifier	0.951039
4	Ada Boost Classifier	0.950759
8	Extra Trees Classifier	0.949109
5	Gradient Boosting Classifier	0.905181
1	KNN	0.888236
0	Logistic Regression	0.810447