*****

Predicting Hotel Reservation Cancellations

Team : Anita Gjurchinovska, Kirby Judd, David Jolia, Chad Richards

A Slidesgo presentation

Contents:

Introduction

- Dataset of hotel reservations from Portugal
- Predict hotel cancellations.
- Objective was to gain insight into factors influencing cancellations.
- Hotel Booking Demand Datasets
  - Nuno Antonio, Ana Almeida, and Luis Nunes, for Data in Brief, Volume 22, February 2019.
  - The dataset is publicly available on Kaggle:
  - https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Methodology:

- Pandas
- Matplotlib
- Scikit-Learn
- SQL Database Postgres
- SQLAlchemy
- Psycopg2

Data Cleaning and Preprocessing

Database Creation, Imports, and Jupyter Notebook:

- Created “hotel_bookings_db” using pgAdmin.
- Imported data into tables from CSV files.
- Established a connection to the local Postgres database using psycopg2 in Jupyter Notebook.

```python
# Replace 'table_name' with the actual table name containing the CSV data
bookings_query = 'SELECT * FROM hotel_bookings'
bookings_df = pd.read_sql_query(bookings_query, engine)
bookings_df.head()
```

✓ 1.1s Python

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	room_type_fulfilled	b
0	1	0	342	2015	7	27	1	0	0	2	...	0	
1	1	0	737	2015	7	27	1	0	0	2	...	0	
2	1	0	7	2015	7	27	1	0	1	1	...	1	
3	1	0	13	2015	7	27	1	0	1	1	...	0	
4	1	0	14	2015	7	27	1	0	2	2	...	0	

5 rows × 31 columns

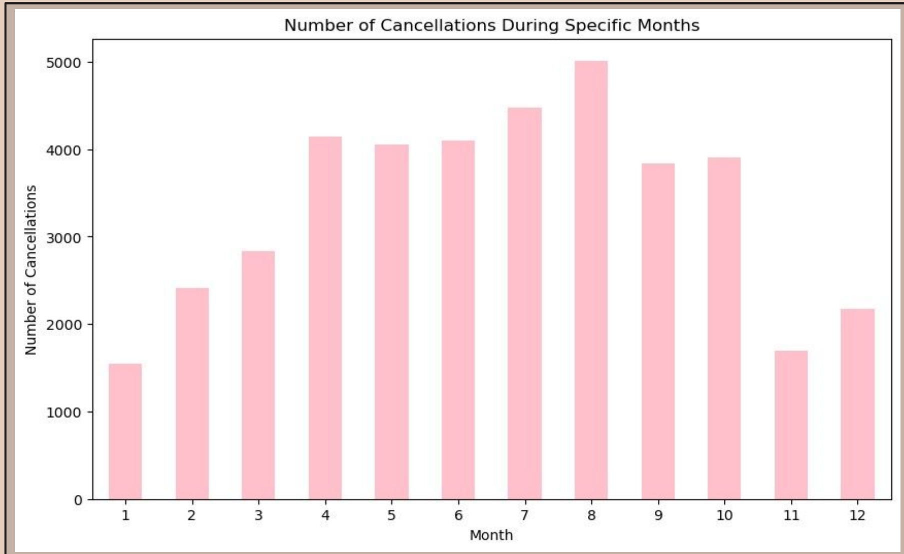
Data Cleaning and Preprocessing

- Reviewed the original dataset for model development.
- Created database schema and imported data.
- Several columns dropped.
- Text data replaced with integer id values.
- Created two new columns to identify international origin and room type fulfillment.
- Dropped NA’s.

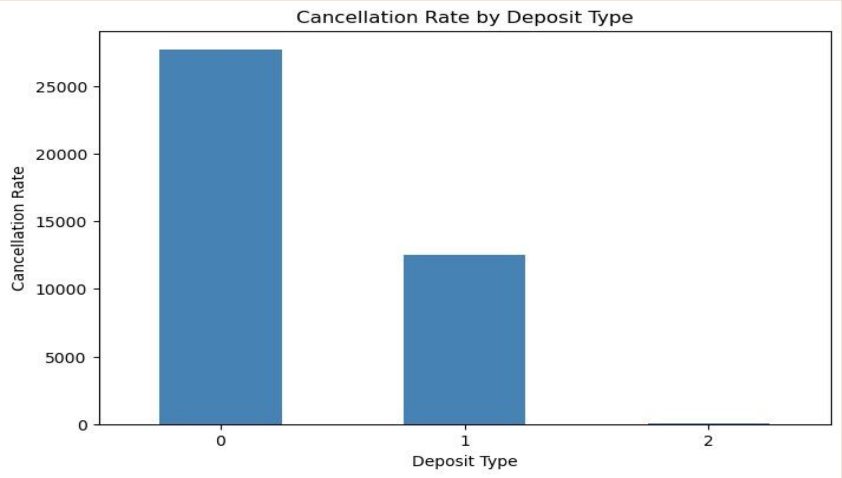
Data Analysis

Questions:

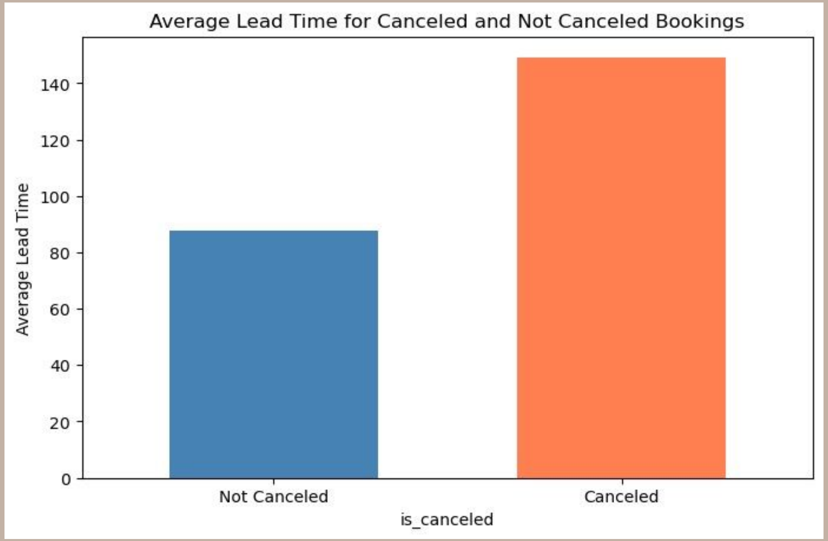
1. Are there more cancellations during specific months?
2. Do different deposit types affect cancellations?
3. Does lead time affect cancellations?
4. Does the origin of tourist (international or local) contribute to more or fewer cancellations?



The data indicates that customers who opted for the “0- No Deposit” option were more likely to cancel their bookings compared to those who choose “1- Non-Refundable” or “2- Refundable” deposit options.

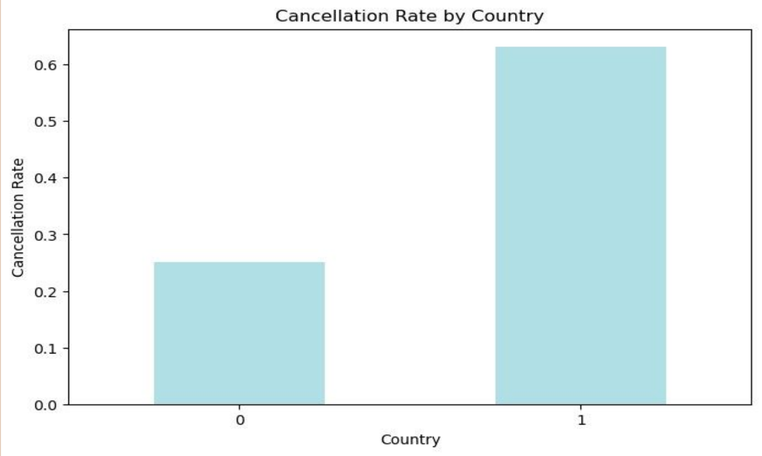


The analysis of the dataset revealed that the month of August experienced the highest number of cancellations compared to other months.

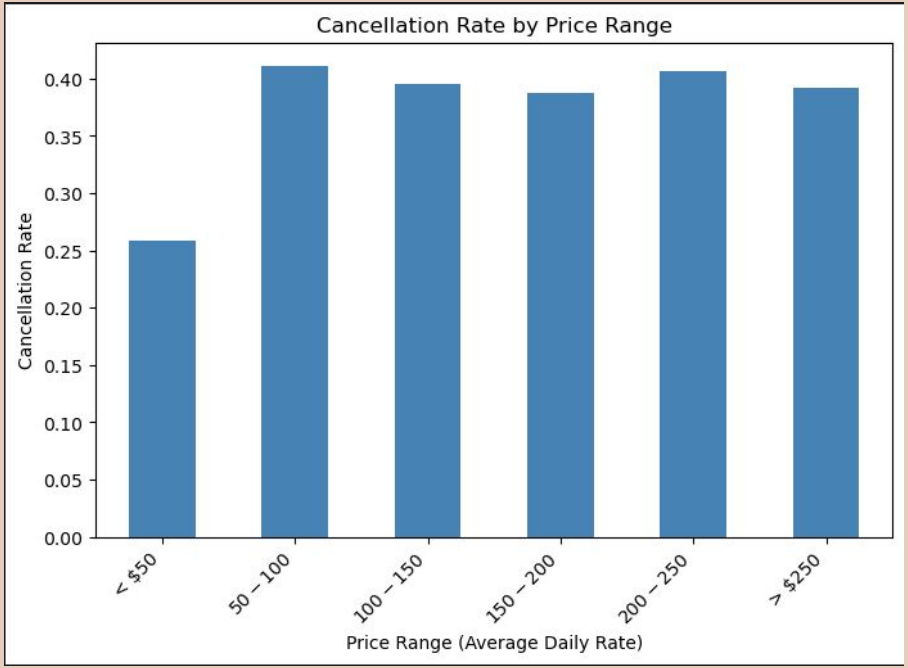


The bar plot compares the average lead time for the "Cancelled" and "Not Cancelled" bookings. The "Not Cancelled" category has an average lead time of 90 days, while the "Cancelled" category has an average lead time of approximately 150 days.

The bar plot shows cancellation rates between the 2 categories: 0- International, and 1- Not International. The "International" bar has height of 0.25, which means that around 25% of the bookings made by international guests were canceled. The "Not International" bar has height of 0.65, meaning that 65% of the bookings made by domestic guests were canceled.



This bar plot shows us the cancellation rates associated with different price ranges of hotel rooms. By analyzing this data, we can gain insights into whether higher prices lead to more cancellations, or if customers are more sensitive to price changes within specific price ranges. This understanding can help us optimize pricing strategies and potentially reduce cancellations by offering competitive prices in the most price-sensitive segments.



Model Building and Evaluation

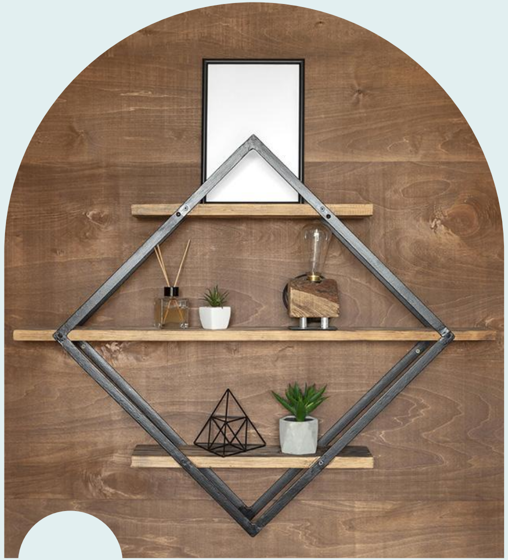
	Expected
0	61.0
1	39.0

	Stratified
0	61.0
1	39.0

```
# Ratio of selected items by is_canceled
stratified_ratio = stratified_sample['is_canceled'].value_counts(normalize=True)

# Convert to percentage
stratified_ratio = stratified_ratio.round(4)*100

# We did stratified sampling. So give it proper name
stratified_ratio.name = 'Stratified'

# Proving the stratified ratio matches the whole dataset ratio (is_canceled_ratios)
stratified_ratio=pd.DataFrame({'Stratified':stratified_ratio})
stratified_ratio
```

Data stratification:

- The dataset was stratified to create manageable testing and training subsets.



Model Selection:

Three model were build and trained on the training data:

- Logistic Regression
- Random Forest
- Decision Tree

Performance Metrics:

To evaluate the model performance on the test data were generated:

- Confusion Matrix
- Accuracy Score
- Classification Reports

Hyperparameter Tuning

```python
# Creating three logistic regression models
# Testing three different logistic regression solvers: lbfgs, liblinear, and newton-cg
# The hyperparameters were tuned to find the best C value to control the regularization strength
# For the max_iter we wanted to make sure it wasn't too high to avoid overfitting
model_lr1 = LogisticRegression(solver='lbfgs', random_state=78, max_iter=6000)
model_lr2 = LogisticRegression(solver='liblinear', random_state=78, C=100, max_iter=2000)
model_lr3 = LogisticRegression(solver='newton-cg', random_state=78, C=4, max_iter=2000)

# Train the data
model_lr1.fit(X_train_scaled, y_train)
model_lr2.fit(X_train_scaled, y_train)
model_lr3.fit(X_train_scaled, y_train)

```

Logistic Regression

```
Using model_lr3: newton-cg
----------------------------
Confusion Matrix
```

	Predicted 0	Predicted 1
Actual 0	14192	1522
Actual 1	3752	6297

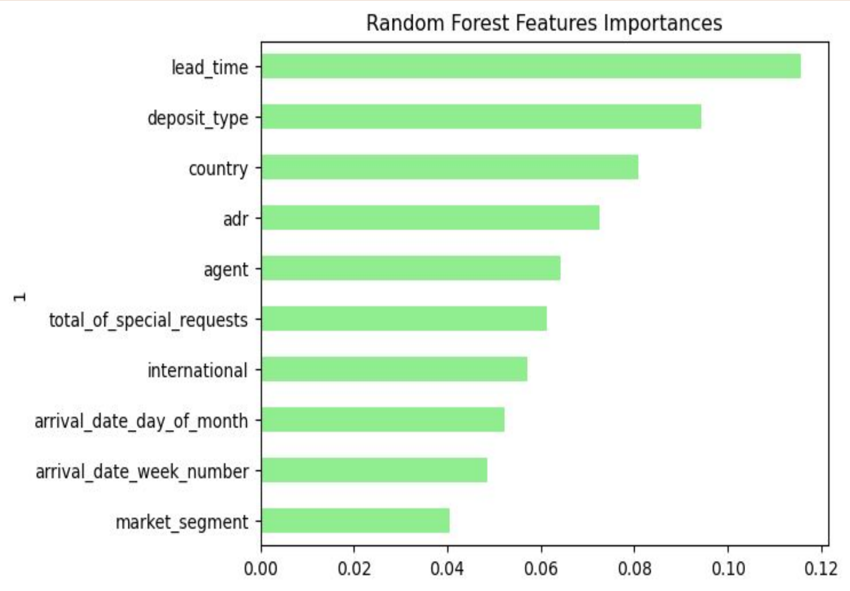
```
Accuracy Score : 0.7952878158599542
Classification Report
              precision    recall  f1-score   support

           0       0.79      0.90      0.84     15714
           1       0.81      0.63      0.70     10049

    accuracy                           0.80     25763
   macro avg       0.80      0.76      0.77     25763
weighted avg       0.80      0.80      0.79     25763
```

Random Forest

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	14575	1139
Actual 1	1587	8462

```
Accuracy Score : 0.8941893413034196
Classification Report
              precision    recall  f1-score   support

           0       0.90      0.93      0.91     15714
           1       0.88      0.84      0.86     10049

    accuracy                           0.89     25763
   macro avg       0.89      0.88      0.89     25763
weighted avg       0.89      0.89      0.89     25763
```

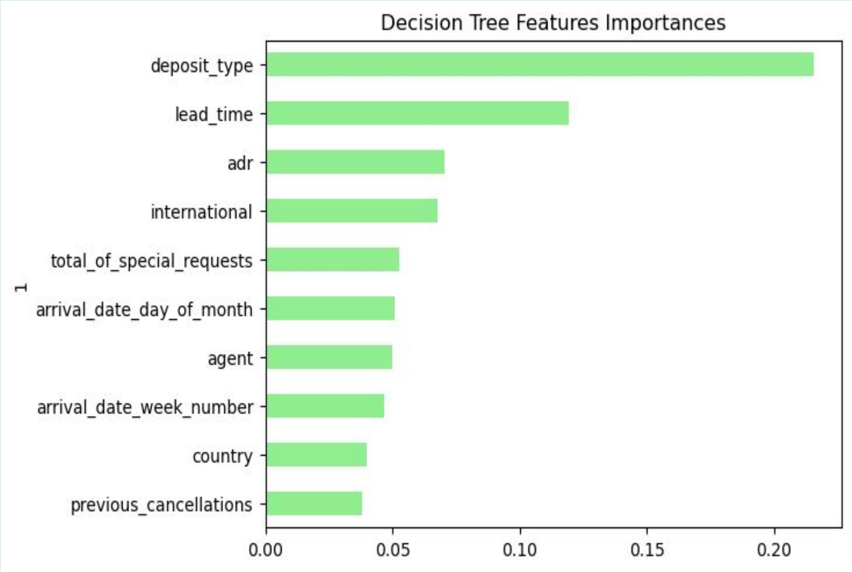


Decision Tree

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	13817	1897
Actual 1	1878	8171

```
Accuracy Score : 0.8534720335364671
Classification Report
              precision    recall  f1-score   support

           0       0.88      0.88      0.88     15714
           1       0.81      0.81      0.81     10049

    accuracy                           0.85     25763
   macro avg       0.85      0.85      0.85     25763
weighted avg       0.85      0.85      0.85     25763
```



Results:

Model	Accuracy Score	F1 score	Precision	Recall
Random Forest	0.894189	0.861272	0.881367	0.842074
Desicion Tree	0.853472	0.812348	0.811581	0.813116
Logistic regression	0.795288	0.704835	0.805346	0.626630

Conclusion:

The models are effective in identifying features contributing to cancellations.

Three features most important to predict cancellations:

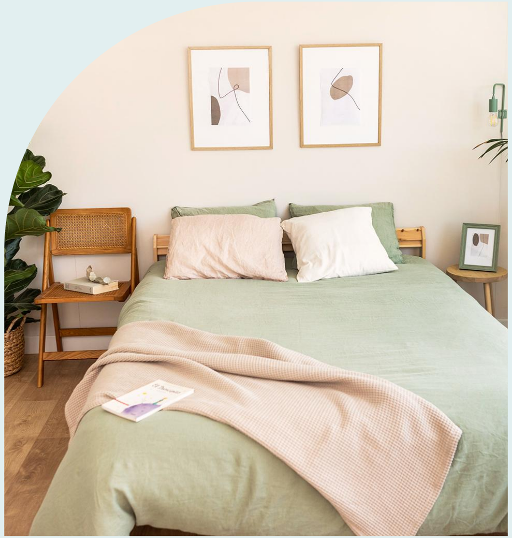
- -Deposit Type
- -Country of Origin (Domestic vs International)
- -Lead Time of Reservation

-Recommendations to Reduce Cancellations:

- -Always charge a deposit
- -Cater to international clients
- -Reduce early booking window
- -Optimize price strategies

Limitations:

1. **Dataset from single country “Portugal”:** The analysis is based on hotel reservation data from Portugal. As a result, the findings may not be directly applicable to hotels in other countries with different travel patterns and preferences.
2. **Data Age (2015-2017):** The data used in the analysis spans from 2015 to 2017, which may not fully capture the current trends and dynamics in the hospitality industry.
3. **Pre-COVID Era:** The dataset predates the COVID-19 pandemic, which significantly impacted the travel and hospitality industry. As a result, the predictive models may not accurately account for the post-COVID travel economy and uncertainties.



Thank you!

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon** and infographics & images by **Freepik**