Data R Us

Presented by: Rohan Benhal, Tanmayee Parbat Jeewan Singh

Introduction, Data Cleaning and Preprocessing

- Background and motivation for the project
- Problem statement and research questions
- Steps taken to clean and preprocess the data

Exploratory Data Analysis

Summary statistics and visualization of the data

Analysis of trends and patterns in the data

Modeling and Evaluation

- Selection of appropriate modeling techniques
- Splitting of data into training, validation, and test sets
- Evaluation of model performance and interpretation of results

Conclusion and Future Work

Summary of key findings

Limitations and potential areas for improvement

Contributions

Introduction, Data Cleaning and Preprocessing

- The Expedia dataset is a rich and complex dataset that captures millions of hotel bookings made by users on the popular travel website, Expedia.com. With detailed information about the destinations, travel dates, booking channels, and user behavior, this dataset provides an unparalleled opportunity to explore patterns and trends in the travel industry.

Research Questions

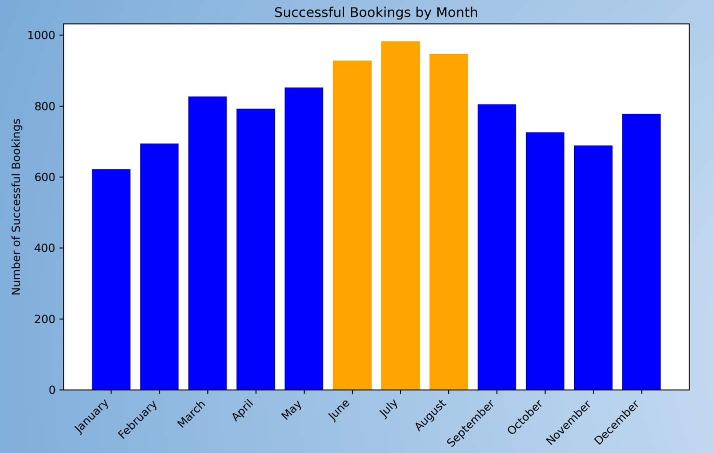
- The goal of this project is to analyze the Expedia dataset to uncover insights that can inform business decisions, such as marketing strategies and product offerings.
- The most popular destinations and travel patterns
- The factors that influence customer booking behaviour
- The demographic characteristics of Expedia customers

Data Cleaning and Preprocessing

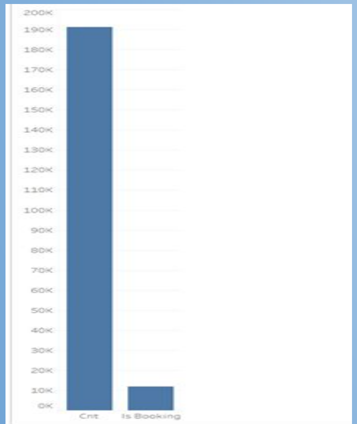
- Merged the data.csv and dest.csv datasets using a common column.
- Removed missing values from the merged dataset.
- Converted the date_time column to a Pandas datetime object for easier manipulation.
- Extracted the year, month, and day as separate numerical features for further analysis.

Exploratory Data Analysis

Year in Review: Successful Bookings and Trends

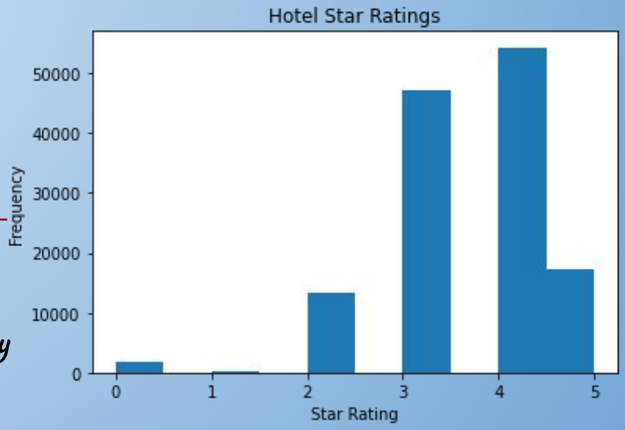


- Bookings steadily increased throughout the year, with the busiest months being June, July, and August
- The slowest months for bookings were October, November, and January
- There were almost 8,500 successful bookings over the course of the year, with July being the most successful month

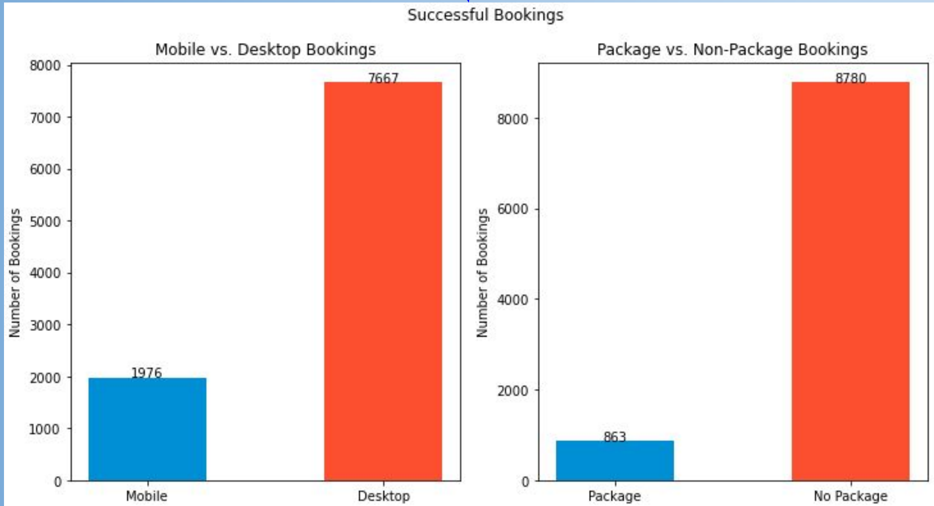


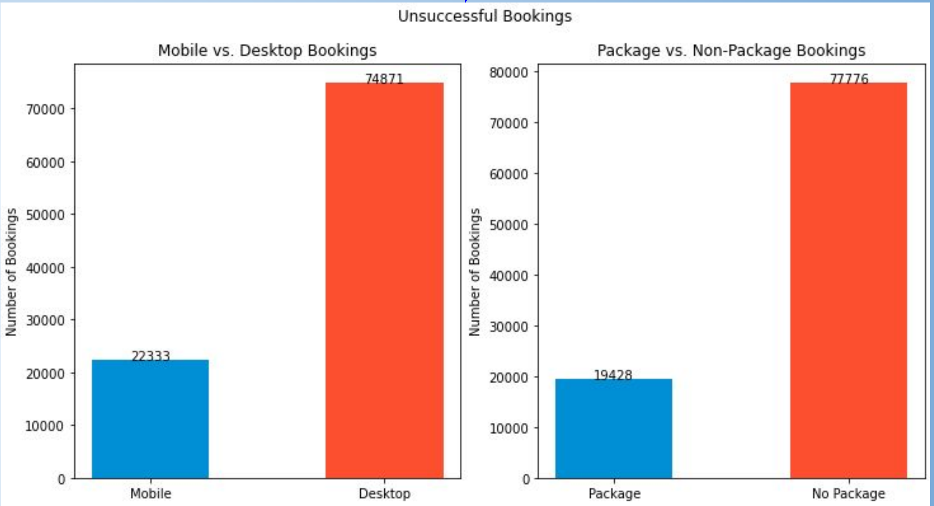
Only around 5% actual bookings done out of all the inquiry calls

Impact of hotel star rating on enquiry frequency

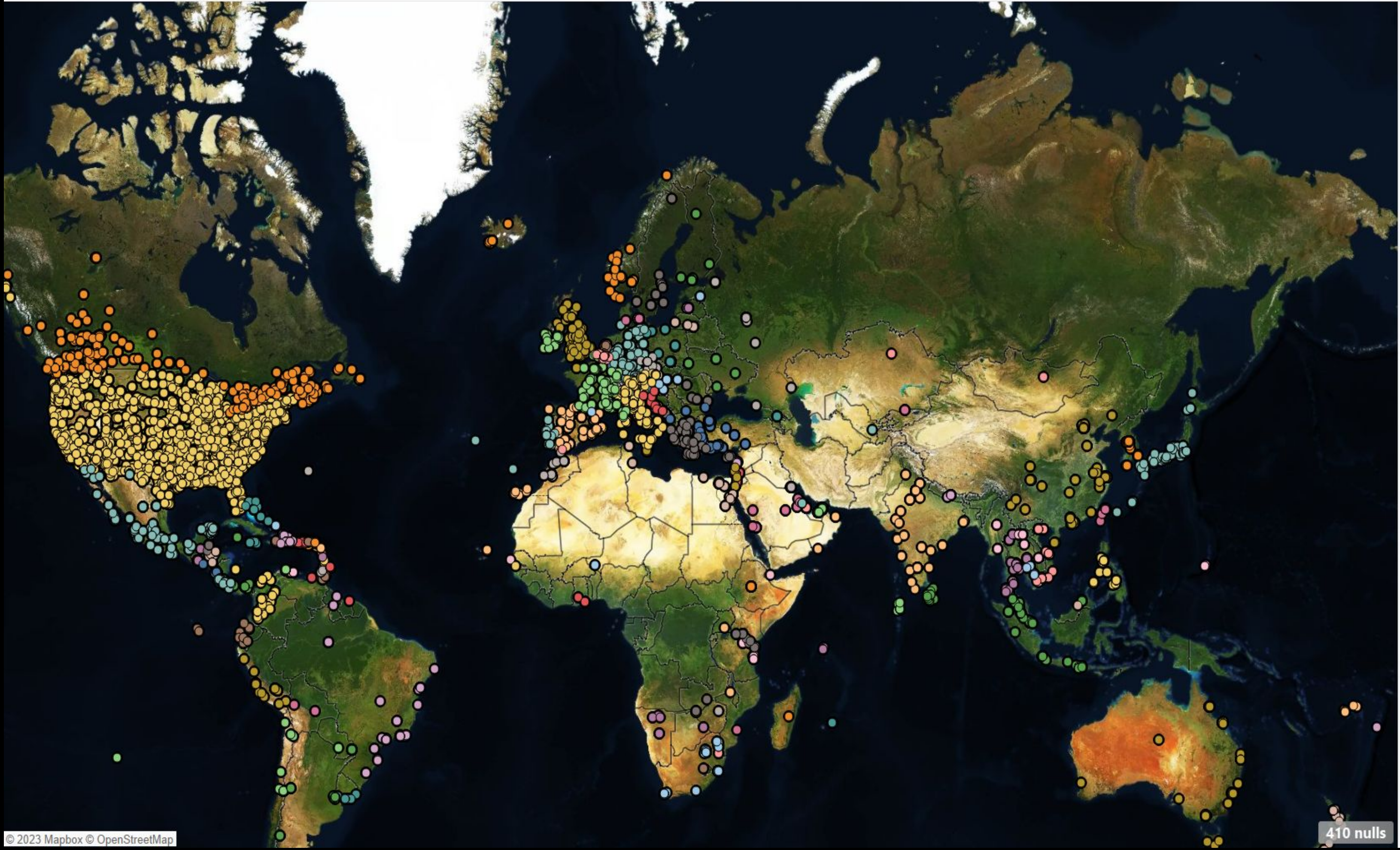


Analyzing Factors Contributing to Booking Success





Bookings around the Globe

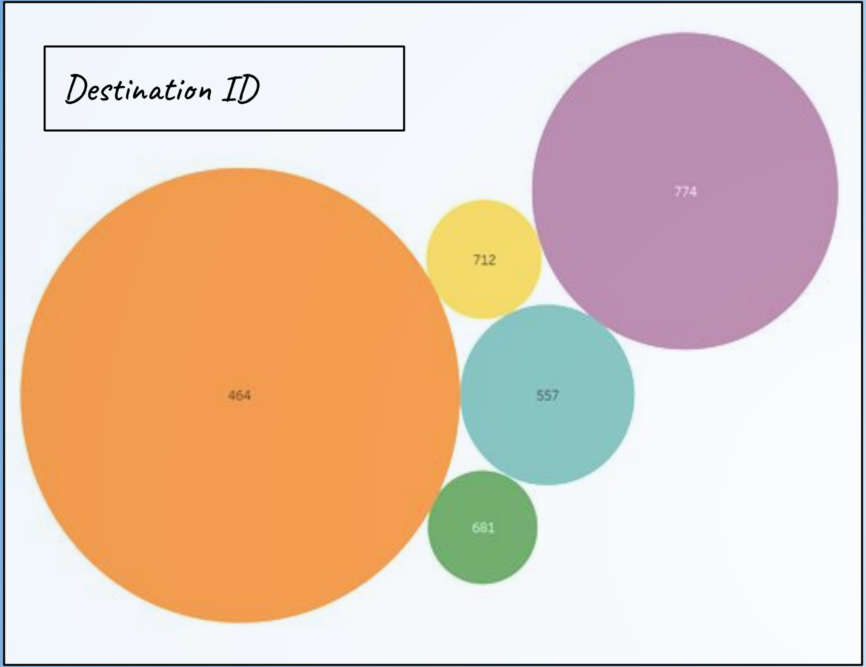


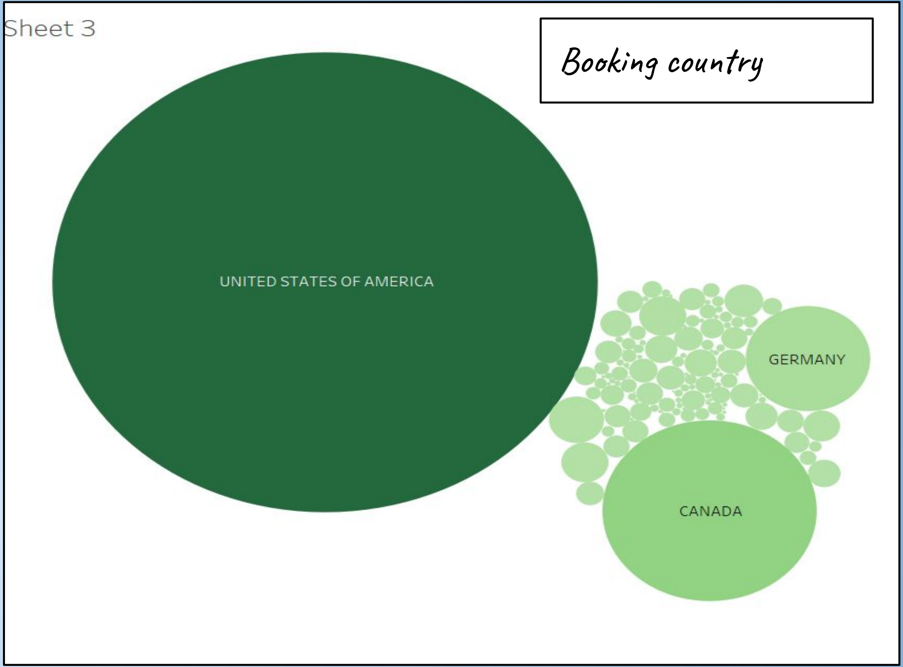
AVG(Is Booking)

0.000 1.000

Hotel Country

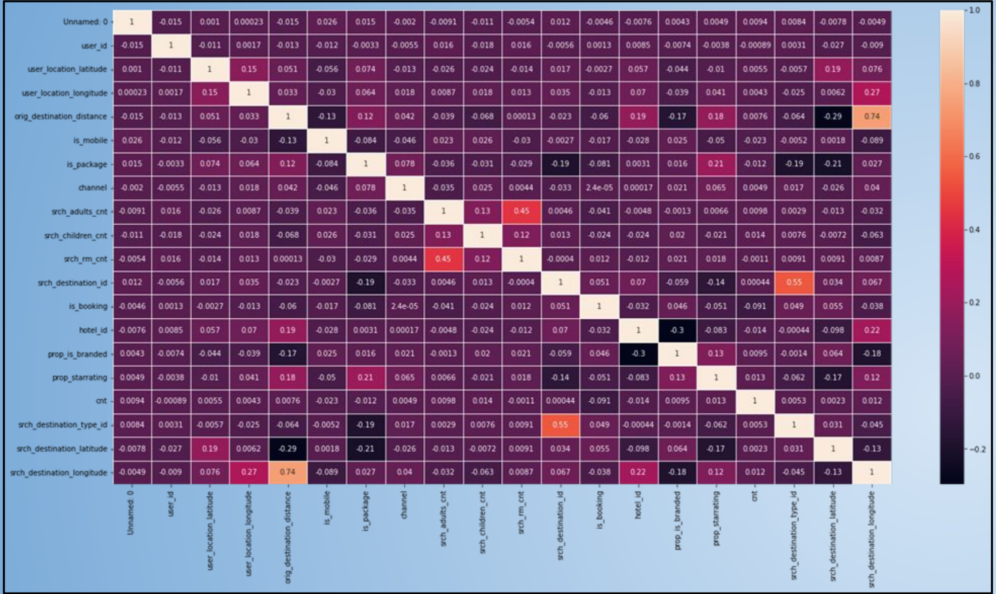
- ALBANIA
- ANDORRA
- ANGUILLA
- ANTIGUA AND BAR..
- ARGENTINA
- ARUBA
- AUSTRALIA
- AUSTRIA
- BAHAMAS
- BAHRAIN
- BARBADOS
- BELGIUM
- BELIZE
- BERMUDA
- BOLIVIA
- BONAIRE, SINT EU..
- BOTSWANA
- BRAZIL
- BRITISH VIRGIN IS..
- BRUNEI
- BULGARIA
- CAMBODIA
- CANADA
- CAPE VERDE
- CAYMAN ISLANDS
- CHILE
- CHINA
- COLOMBIA
- COOK ISLANDS
- COSTA RICA
- CROATIA
- CURACAO
- CYPRUS
- CZECH REPUBLIC
- DENMARK
- DJIBOUTI
- DOMINICA
- DOMINICAN REPU..
- ECUADOR
- EGYPT
- EL SALVADOR
- ESTONIA





- ➔ *Destination ID which is searched the most i.e (ID: 464)*
- ➔ *In the total search 7 times more enquiry were made for just adults than people who have children.*
- ➔ *Only around 22% booking call were made via Mobile Phone.*
- ➔ *Most amount of booking were done in United states, 2nd canada, 3rd germany.*
- ➔ *Around 600% more booking were done in united states than canada*

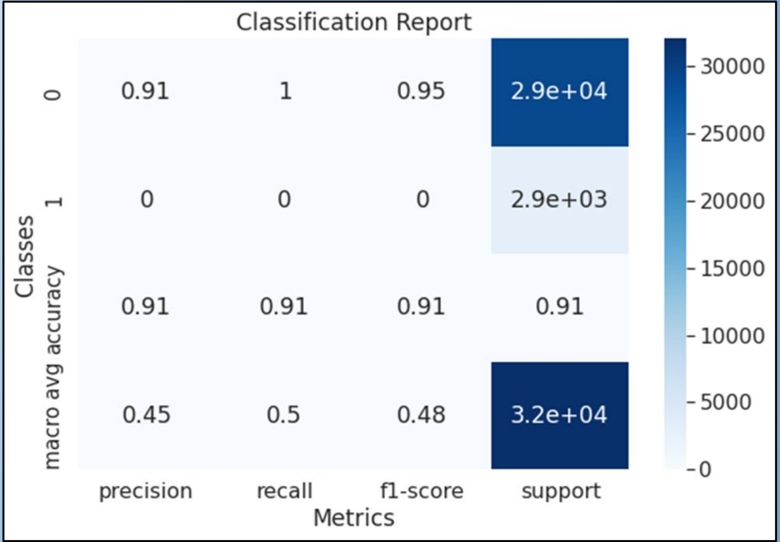
Modeling and Evaluation



Supervised Learning

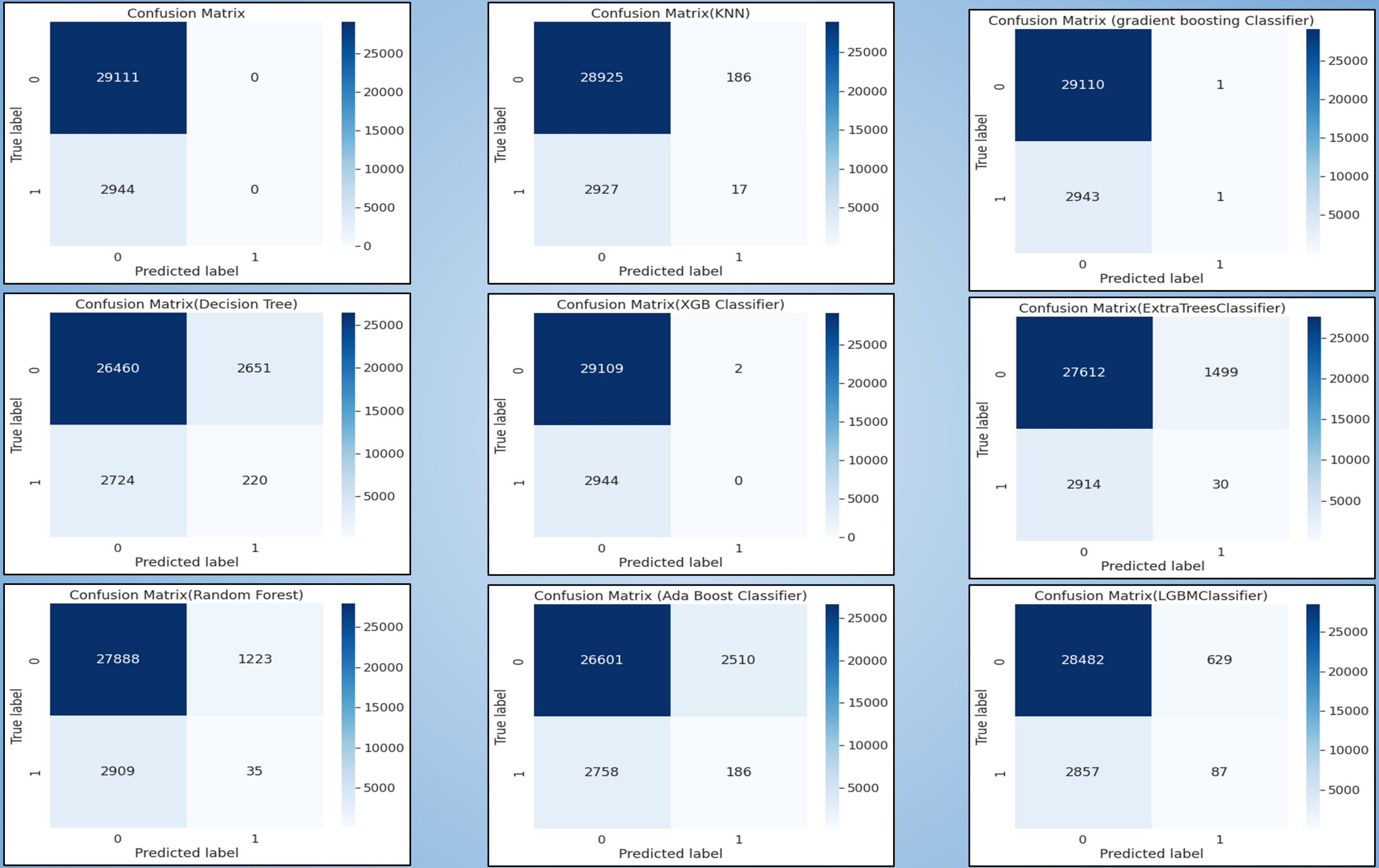
Splitting the Data Training and Testing (70:30)

"Is_booking" is the dependent variable.

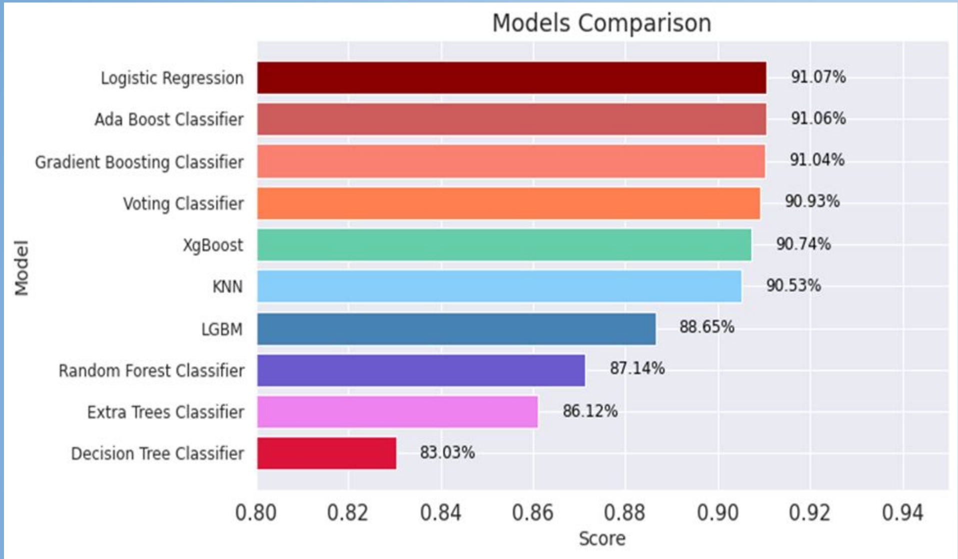


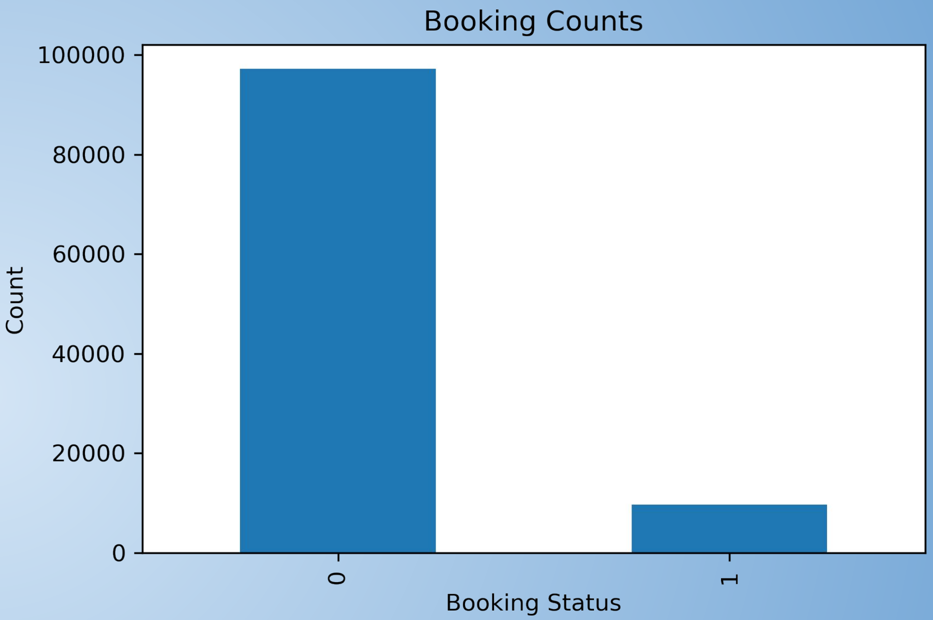
Positive correlation (with is_booking)

- *Is_package* 0.080633
- *orig_destination_distance* 0.059763
- *srch_destination_latitude* 0.054978
- *prop_starrating* 0.051226
- *srch_destination_id* 0.050617


Confusion Matrix
29111
0
2944
0
Confusion Matrix(KNN)
28925
186
2927
17
Confusion Matrix (gradient boosting Classifier)
29110
1
2943
1
Confusion Matrix(Decision Tree)
26460
2651
2724
220
Confusion Matrix(XGB Classifier)
29109
2
2944
0
Confusion Matrix(ExtraTreesClassifier)
27612
1499
2914
30
Confusion Matrix(Random Forest)
27888
1223
2909
35
Confusion Matrix (Ada Boost Classifier)
26601
2510
2758
186
Confusion Matrix(LGBMClassifier)
28482
629
2857
87
True label
Predicted label
0
1
25000
20000
15000
10000
5000
0

Model Scores





➔ *The models achieved a high accuracy score, but it should be noted that the data was imbalanced*

➔ *This indicates that the model is good at predicting the negative class (unsuccessful bookings) but is very poor at detecting positive cases, which is the main problem that needs to be solved in this scenario.*

Logistic Regression

```
Undersampled dataset accuracy:
0.5393986521513737
              precision    recall  f1-score   support

           0       0.55      0.44      0.49      1941
           1       0.53      0.64      0.58      1917

    accuracy                           0.54      3858
   macro avg       0.54      0.54      0.54      3858
weighted avg       0.54      0.54      0.53      3858

[[ 853 1088]
 [ 689 1228]]
Oversampled dataset accuracy:
0.534720436191554
              precision    recall  f1-score   support

           0       0.54      0.45      0.49     19446
           1       0.53      0.62      0.57     19436

    accuracy                           0.53     38882
   macro avg       0.54      0.53      0.53     38882
weighted avg       0.54      0.53      0.53     38882

[[ 8721 10725]
 [ 7366 12070]]
```

Decision Tree

```
Undersampled dataset accuracy:
0.5780196993260757
              precision    recall  f1-score   support

           0       0.58      0.60      0.59      1941
           1       0.58      0.56      0.57      1917

    accuracy                           0.58      3858
   macro avg       0.58      0.58      0.58      3858
weighted avg       0.58      0.58      0.58      3858

[[1159  782]
 [ 846 1071]]
Oversampled dataset accuracy:
0.8820276734735868
              precision    recall  f1-score   support

           0       0.88      0.89      0.88     19446
           1       0.88      0.88      0.88     19436

    accuracy                           0.88     38882
   macro avg       0.88      0.88      0.88     38882
weighted avg       0.88      0.88      0.88     38882

[[17211  2235]
 [ 2352 17084]]
```

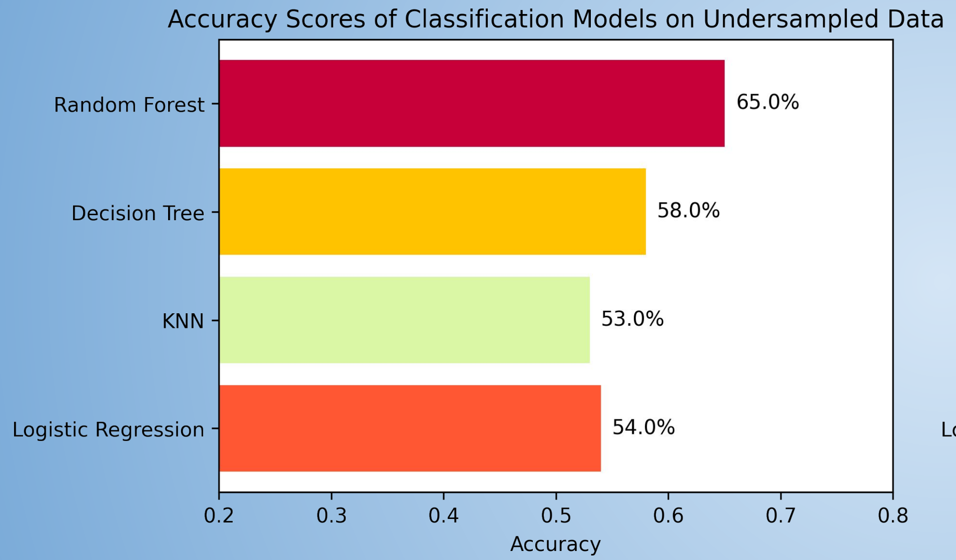
Random Forest

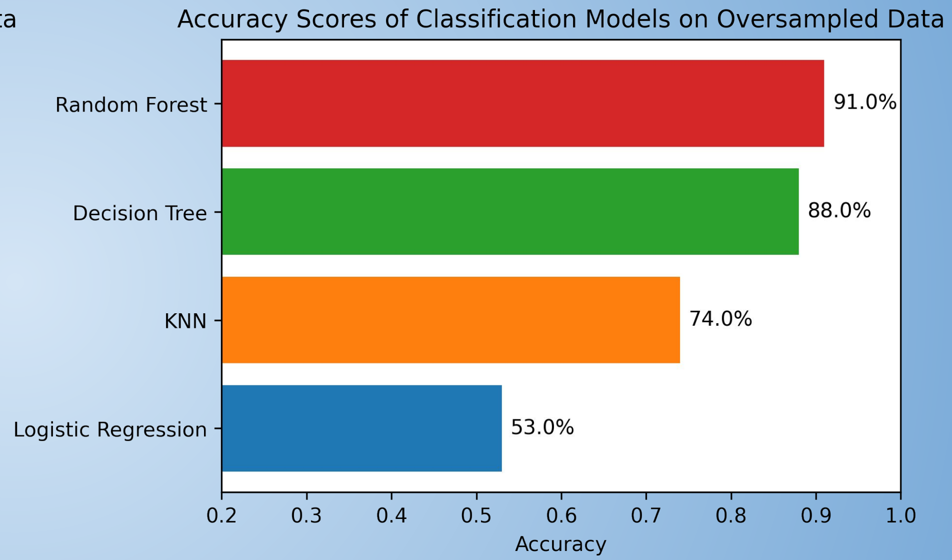
```
Undersampled dataset accuracy:
0.6498185588387766
              precision    recall  f1-score   support

           0       0.67      0.60      0.63      1941
           1       0.63      0.70      0.66      1917

    accuracy                           0.65      3858
   macro avg       0.65      0.65      0.65      3858
weighted avg       0.65      0.65      0.65      3858

[[1170  771]
 [ 580 1337]]
Oversampled dataset accuracy:
0.9145105704439072
              precision    recall  f1-score   support

           0       0.90      0.93      0.92     19446
           1       0.93      0.90      0.91     19436

    accuracy                           0.91     38882
   macro avg       0.91      0.91      0.91     38882
weighted avg       0.91      0.91      0.91     38882

[[18061  1385]
 [ 1939 17497]]
```

KNN

```
Undersampled dataset accuracy:
0.5274753758424053
              precision    recall  f1-score   support

           0       0.53      0.54      0.53      1941
           1       0.52      0.52      0.52      1917

    accuracy                           0.53      3858
   macro avg       0.53      0.53      0.53      3858
weighted avg       0.53      0.53      0.53      3858

[[1046  895]
 [ 928  989]]
Oversampled dataset accuracy:
0.7419885808343192
              precision    recall  f1-score   support

           0       0.76      0.71      0.73     19446
           1       0.73      0.77      0.75     19436

    accuracy                           0.74     38882
   macro avg       0.74      0.74      0.74     38882
weighted avg       0.74      0.74      0.74     38882

[[13847  5599]
 [ 4433 15003]]
```

Comparison of Accuracy Scores for Undersampled and Oversampled Data





SMOTE has successfully balanced the classes and improved model performance

Conclusion and Future Work

- Focus on improving the user experience on mobile devices, as they had fewer successful bookings compared to desktop.
- The oversampled dataset produced significantly better results than the undersampled dataset, with the random forest model achieving the highest accuracy score of 0.91. This suggests that oversampling is a promising approach to address class imbalance in the Expedia dataset.
- Among the four models tested, the random forest model consistently outperformed the other models in both the oversampled and undersampled datasets. This may be due to the ability of random forests to handle high-dimensional data and capture complex interactions between features.
- In future work, it may be worth exploring more advanced oversampling techniques such as adaptive synthetic sampling (ADASYN) to further improve classification performance.
- It may also be useful to investigate feature engineering techniques to identify the most important features for classification and potentially improve model performance.


Thank You