Hotel
Booking Prediction using ML

Team Members- Group 1:

1. Ketha Tirumuru
2. Chandana Polakonda
3. Arun Thotakuri
4. Numitha Devi Oguri
5. Lalitha Nali

Contents

Problem Statement

Cancellations can have a significant impact on a hotel's revenue and profitability. In fact, studies show that the average hotel loses up to 15% of its revenue due to cancellations. That's why predicting cancellations is so important for hotels.

Objective

The objective of this project is to create supervised learning-based predictive models for hotel booking cancellations.

- Improve revenue management by identifying potential cancellations in advance.

- Optimize resource allocation, such as staffing and inventory, based on cancellation predictions.

- Minimize revenue loss by reducing the impact of last-minute cancellations.

1

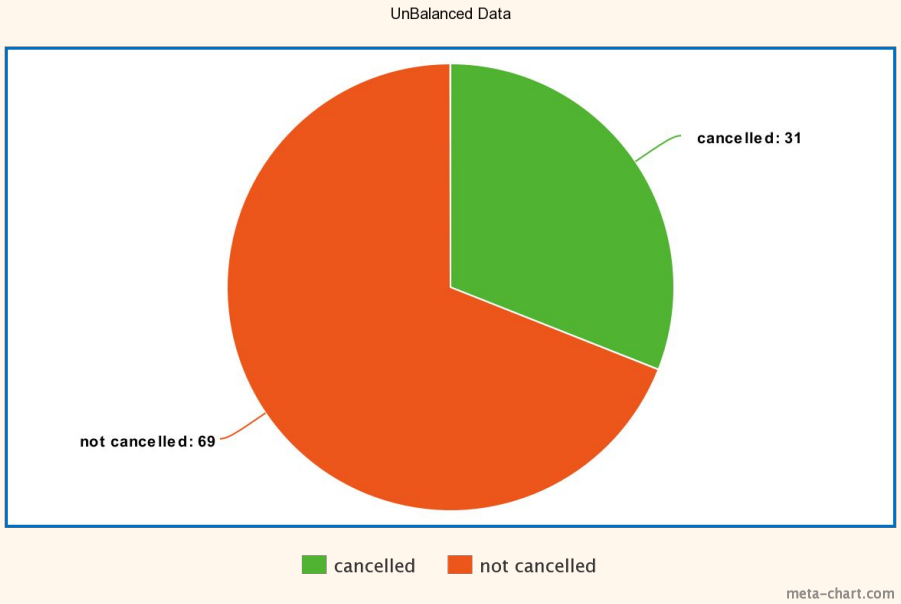
2. Data Description

Data Description

- The dataset consists of **119,390** hotel booking observations with **31 Features** for **2 hotels** (resort and city hotel).
- Here the Target Variable is: **is_cancelled.**
- Each observation represents a hotel booking and contains various features providing relevant details.
- The dataset includes information such as booking cancellations, lead time, arrival dates, guest demographics, room types, and booking channels.
- Features include **12 categorical variables** (hotel type, meal plan, market segment, etc.) and **19 numerical variables** (lead time, number of nights stayed, etc.).
- There are also binary variables indicating repeated guests and deposit types.
- Data exploration and preprocessing are performed to understand the distribution, identify missing values, and handle outliers or inconsistencies.
- The dataset is divided into training and testing sets to build and evaluate the predictive models effectively.

3. Exploratory Data Analysis

Unbalanced data

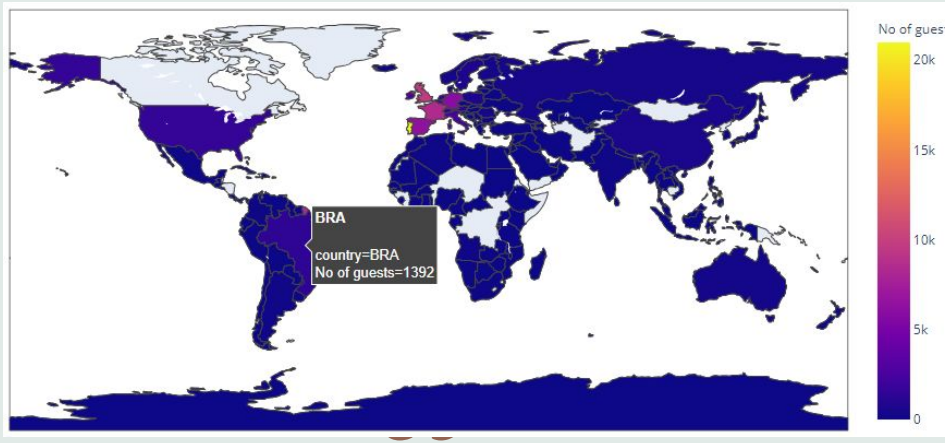


Unbalanced Dataset 31% bookings got cancelled.

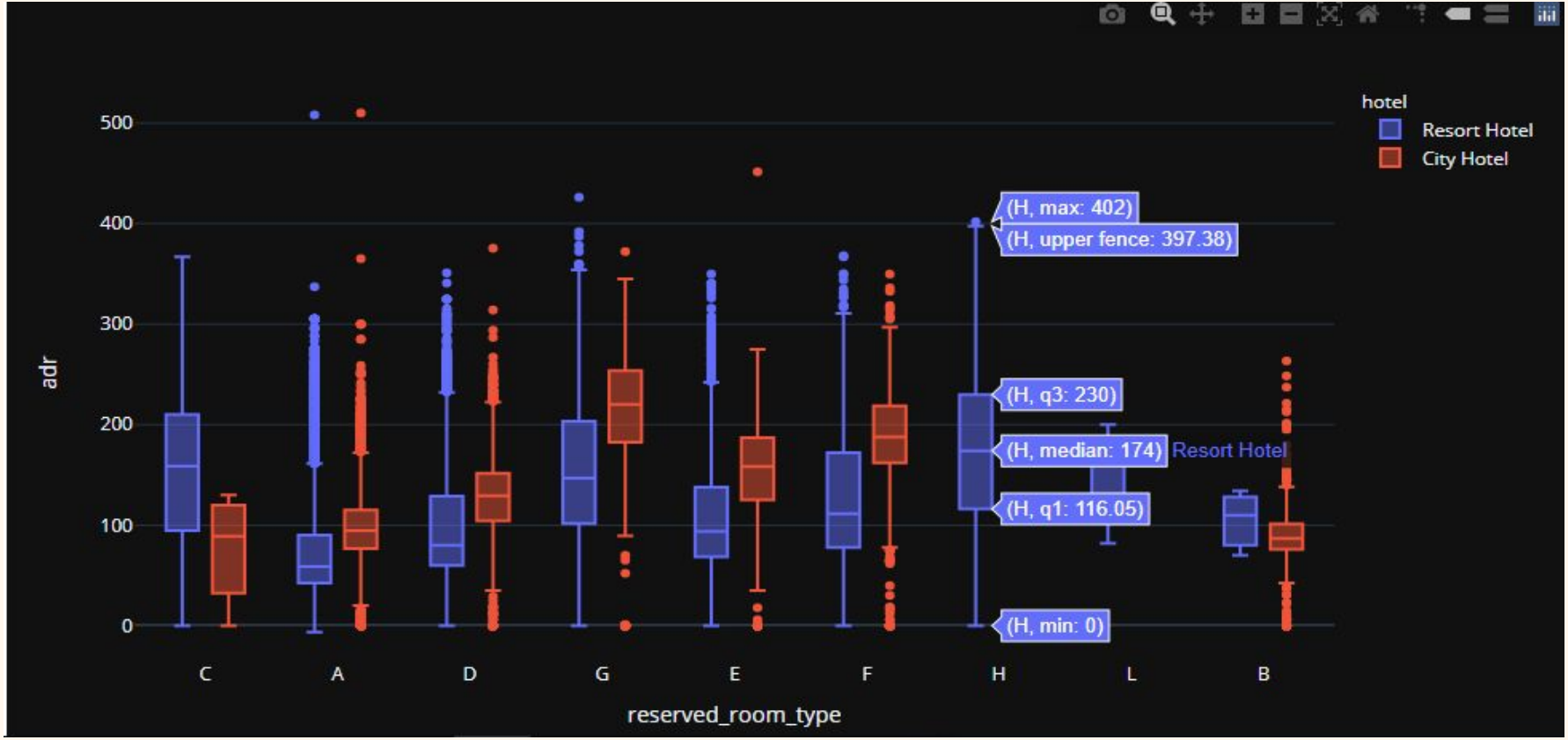
	country	No of guests
0	PRT	20977
1	GBR	9668
2	FRA	8468
3	ESP	6383
4	DEU	6067
...	...	...
161	NPL	1
162	GUY	1
163	MRT	1
164	ATF	1
165	NAM	1

Country

Most of the Guests are coming from Portugal

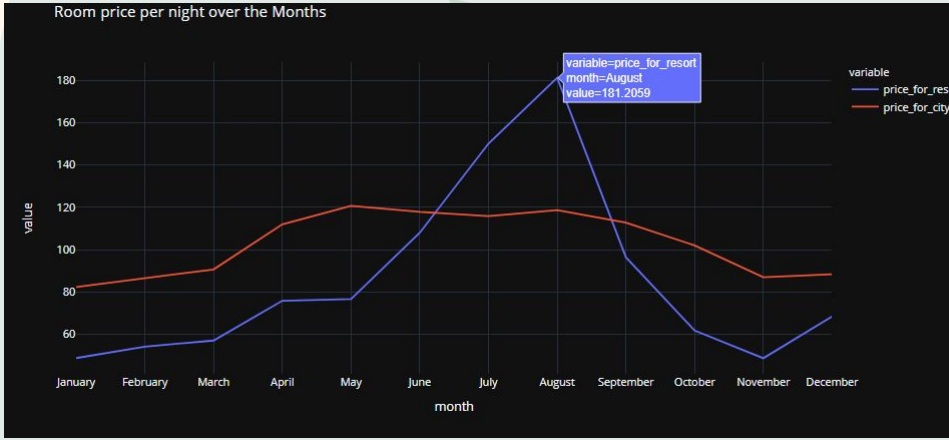


People from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe.



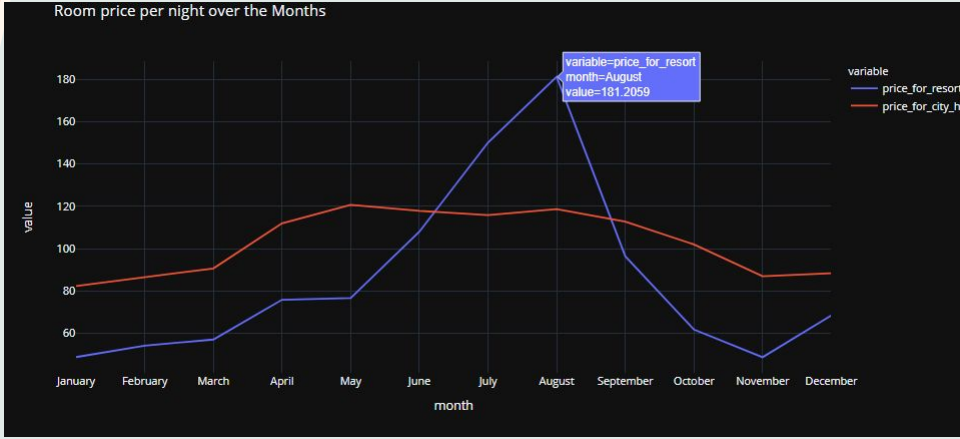
The figure shows that the average price per room depends on its type and the standard deviation.

	month	price_for_resort	price_for_city_hotel
0	January	48.761125	82.330983
1	February	54.147478	86.520062
2	March	57.056838	90.658533
3	April	75.867816	111.962267
4	May	76.657558	120.669827
5	June	107.974850	117.874360
6	July	150.122528	115.818019
7	August	181.205892	118.674598
8	September	96.416860	112.776582
9	October	61.775449	102.004672
10	November	48.706289	86.946592
11	December	68.410104	88.401855



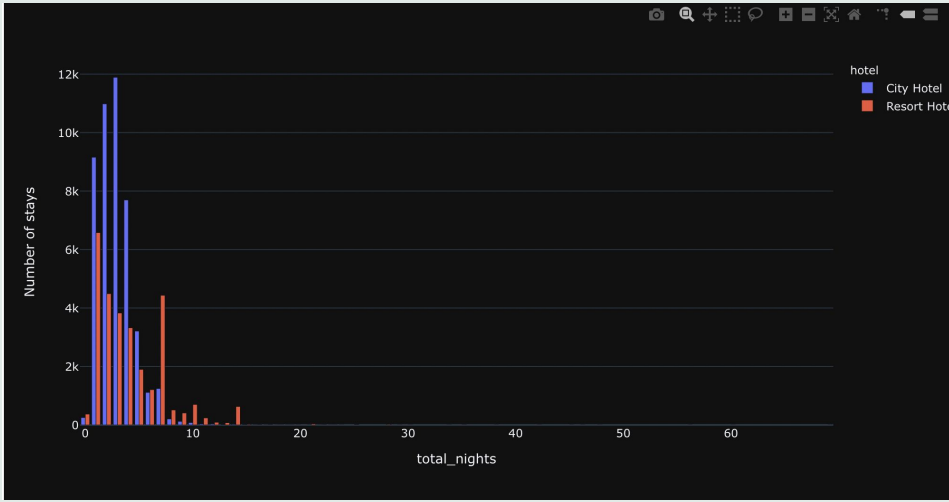
How does the price vary per night over the year?

	month	no of guests in resort	no of guest in city hotel
0	January	1866	2249
1	February	2308	3051
2	March	2571	4049
3	April	2550	4010
4	May	2535	4568
5	June	2037	4358
6	July	3137	4770
7	August	3257	5367
8	September	2102	4283
9	October	2575	4326
10	November	1975	2676
11	December	2014	2377



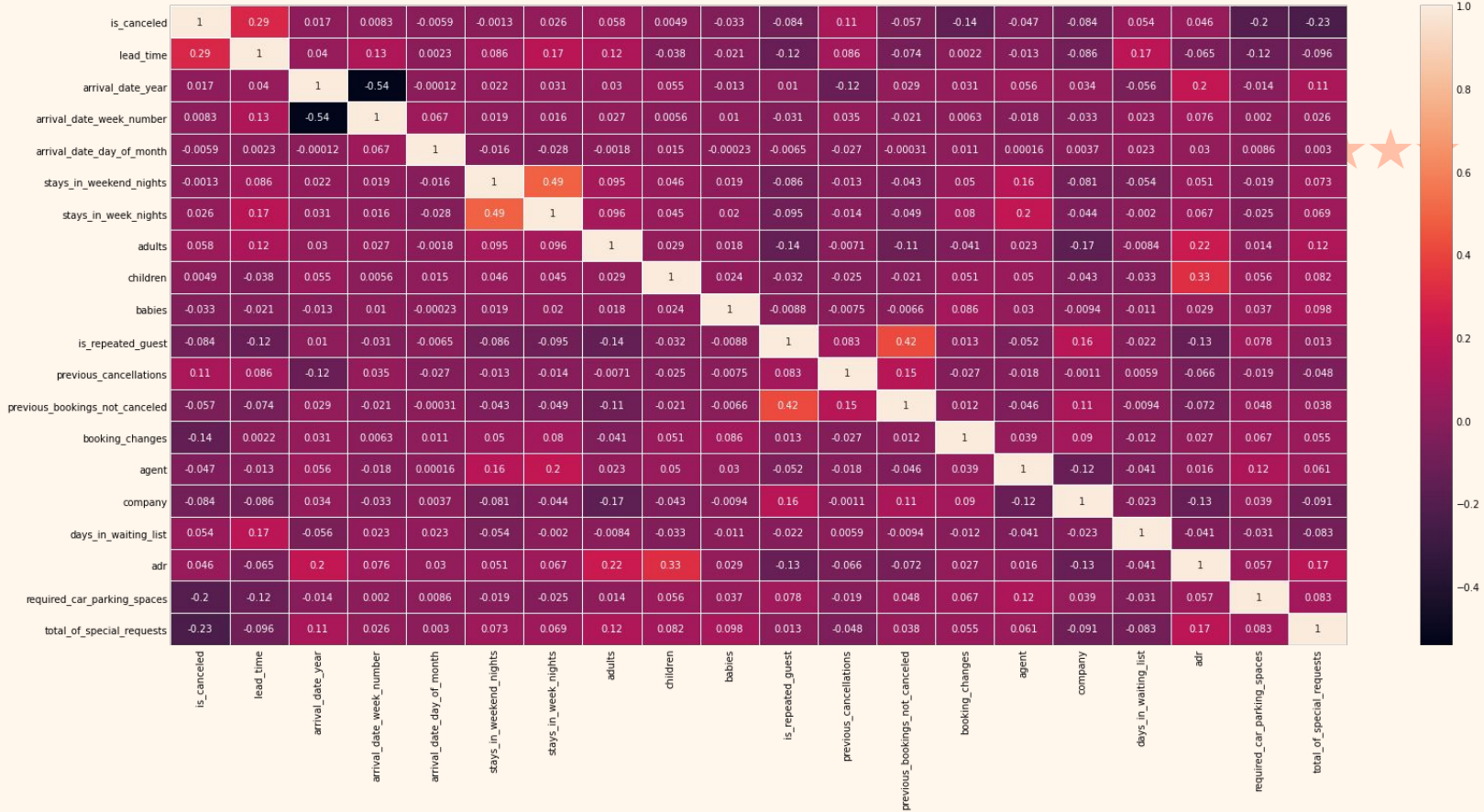
Which are the most busy months?

	total_nights	hotel	Number of stays
0	0	City Hotel	251
1	0	Resort Hotel	371
2	1	City Hotel	9155
3	1	Resort Hotel	6579
4	2	City Hotel	10983
...	...	...	...
57	46	Resort Hotel	1
58	48	City Hotel	1
59	56	Resort Hotel	1
60	60	Resort Hotel	1
61	69	Resort Hotel	1



How long people stay at hotels?

4. Data Pre-processing



Correlation Matrix

Data Pre-processing

- Created a correlation Matrix.
- *Dropping columns that are not useful.*
- *Creating numerical and categorical dataframes.*
- *Encoding categorical variables.*
- *Normalizing numerical variables.*
- *Splitting data into training set and test set.*

5. Model Building

Models we Used

- Logistic Regression
- K-Nearest Neighbors (KNN) model
- Decision Tree Classifier model
- Random Forest Classifier model
- XGBoost model
- catBoost Model

Logistic Regression

```
Accuracy Score of Logistic Regression is : 0.8106422839247267
Confusion Matrix :
[[21339  1097]
 [ 5675  7652]]
Classification Report :
              precision    recall  f1-score   support

           0       0.79      0.95      0.86     22436
           1       0.87      0.57      0.69     13327

    accuracy                           0.81     35763
   macro avg       0.83      0.76      0.78     35763
weighted avg       0.82      0.81      0.80     35763
```

KNN

```
Accuracy Score of KNN is : 0.8920951821715181
Confusion Matrix :
[[21692   744]
 [ 3115 10212]]
Classification Report :
              precision    recall  f1-score   support

           0       0.87      0.97      0.92     22436
           1       0.93      0.77      0.84     13327

    accuracy                           0.89     35763
   macro avg       0.90      0.87      0.88     35763
weighted avg       0.90      0.89      0.89     35763
```

Decision Tree

```
Accuracy Score of Decision Tree is : 0.949053491038223 9
Confusion Matrix :
[[21578   858]
 [  964 12363]]
Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.96      0.96     22436
           1       0.94      0.93      0.93     13327

    accuracy                           0.95     35763
   macro avg       0.95      0.94      0.95     35763
weighted avg       0.95      0.95      0.95     35763
```

Random Forest

```
Accuracy Score of Random Forest is : 0.9531638844615944
Confusion Matrix :
[[22287   149]
 [ 1526 11801]]
Classification Report :
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     22436
           1       0.99      0.89      0.93     13327

    accuracy                           0.95     35763
   macro avg       0.96      0.94      0.95     35763
weighted avg       0.96      0.95      0.95     35763
```

XGBoost Model

Accuracy Score of XG Boost Classifier is : 0.9840897016469535

Clear output

executed by Tirumuru Ketha
11:30 PM (5 minutes ago)
executed in 19.078s

:

	on	recall	f1-score	support
0	0.98	1.00	0.99	22612
1	1.00	0.96	0.98	13151
accuracy			0.98	35763
macro avg	0.99	0.98	0.98	35763
weighted avg	0.98	0.98	0.98	35763

CatBoost Model

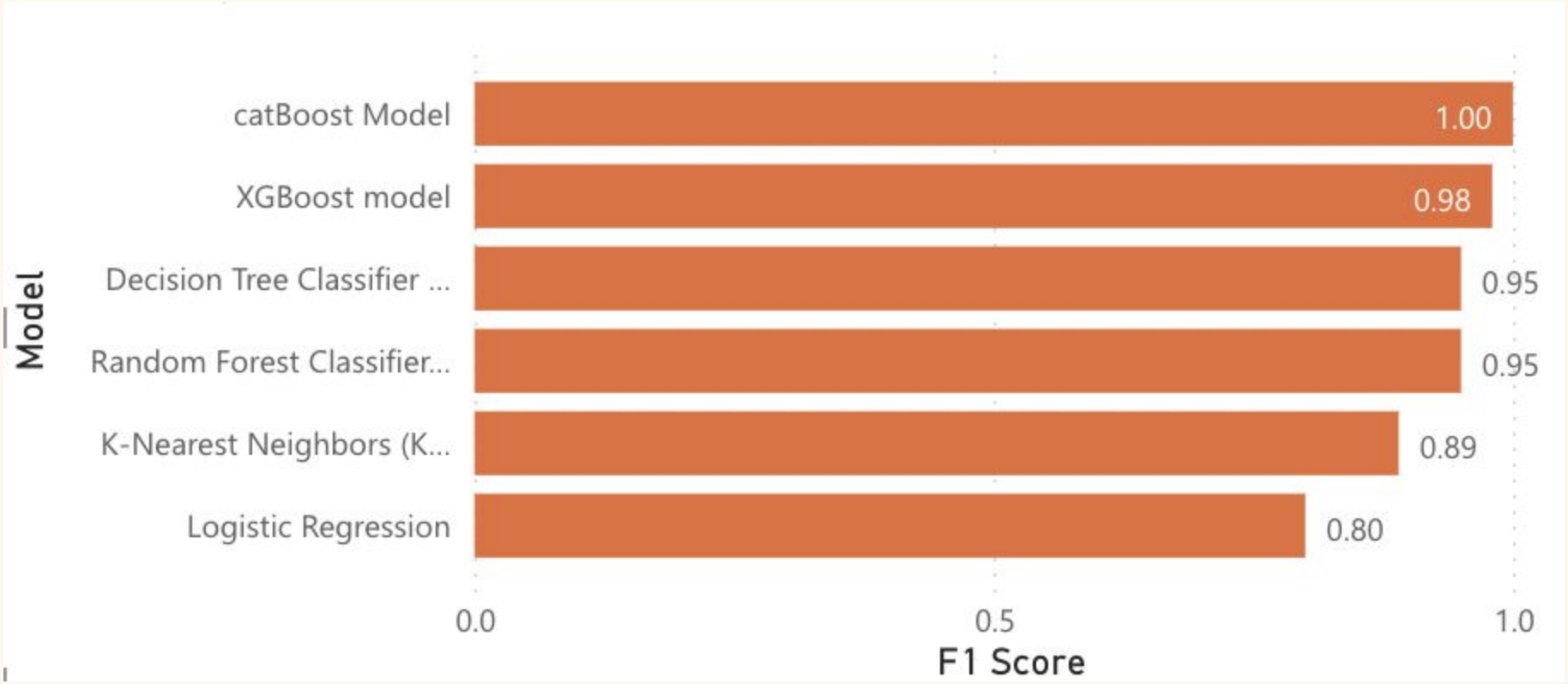
```
Accuracy Score of  CatBoost Classifier is : 0.9954142549562397
Confusion Matrix :
[[22602    10]
 [  154 12997]]
Classification Report :
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     22612
           1       1.00      0.99      0.99     13151

    accuracy                           1.00     35763
   macro avg       1.00      0.99      1.00     35763
weighted avg       1.00      1.00      1.00     35763
```

6. Model Comparision

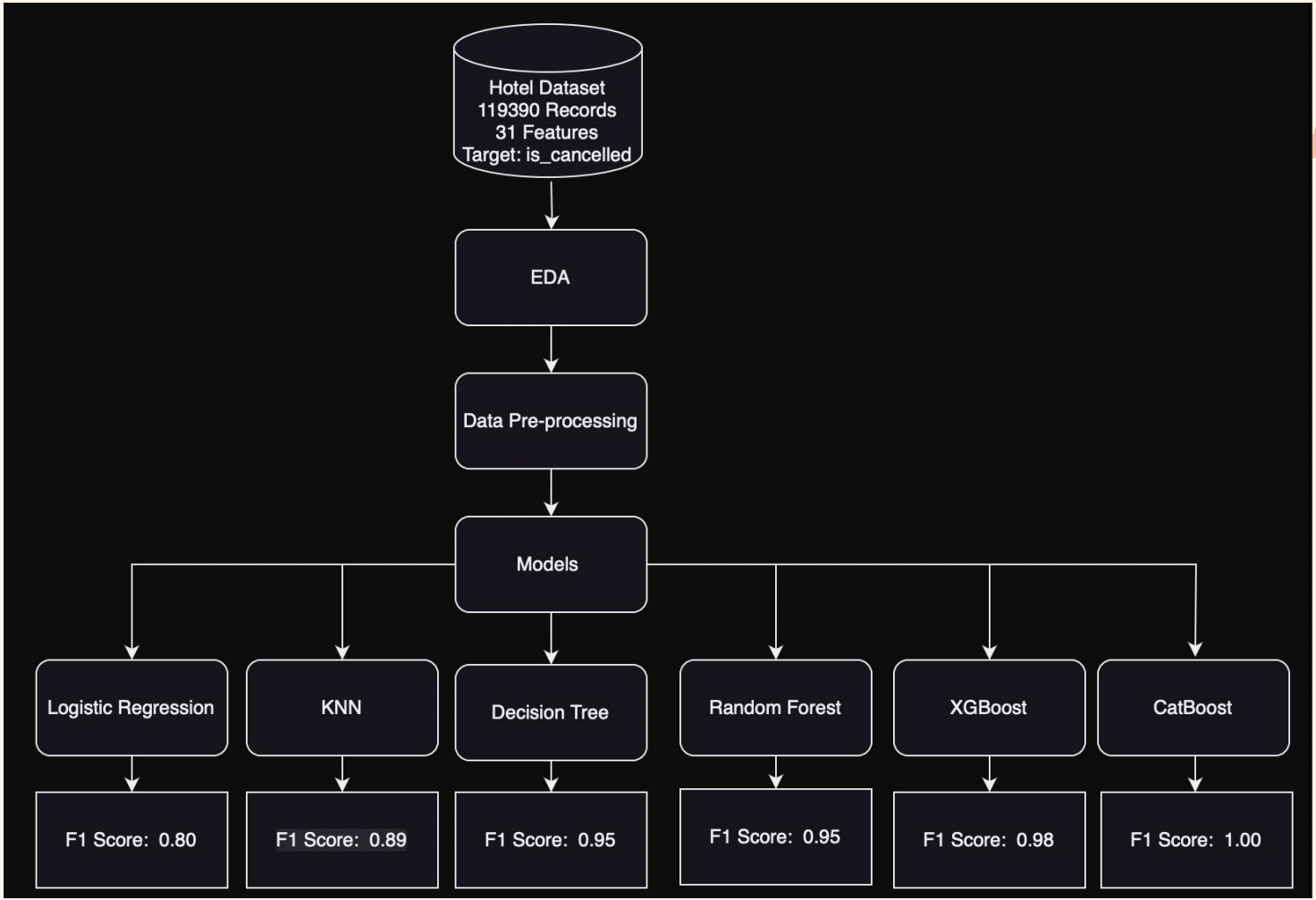
Due to unbalanced class we consider Model Metric as F1-Score

Model	F1 Score
catBoost Model	1.00
XGBoost model	0.98
Decision Tree Classifier model	0.95
Random Forest Classifier model	0.95
K-Nearest Neighbors (KNN) model	0.89
Logistic Regression	0.80

Model Comparison Graph



7. Overall Project Design


Hotel Dataset
119390 Records
31 Features
Target: is_cancelled
EDA
Data Pre-processing
Models
Logistic Regression
KNN
Decision Tree
Random Forest
XGBoost
CatBoost
F1 Score: 0.80
F1 Score: 0.89
F1 Score: 0.95
F1 Score: 0.95
F1 Score: 0.98
F1 Score: 1.00

8. Repo Link

Repo Link:

https://github.com/tketha/SD-for-AI---Group1/tree/main

9. References

1. https://www.kaggle.com/code/niteshyadav3103/hotel-booking-prediction-99-5-acc/input
2. https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand
3. https://www.sciencedirect.com/science/article/pii/S2352340918315191

Thanks

Any Questions?