Project Title

Prediction of hotel reservation cancellation of a customer with reservation details using Machine Learning.

Student Name: Nikhil Reddy Marella

Student ID: 20067093

Supervisor: Dr. John Lones

Aim of the project:

This project aims to identify the reasons behind the cancellations of hotel reservations and develop machine-learning algorithms on oversampled and under-sampled data to predict the customer who might cancel the reservation based on the customer details provided using the machine-learning algorithms with the best performance.

Research Questions

1. What are the columns not important for analysis and why were they dropped from the dataset?

2. What is one hot encoding and why it is crucial for analyzing this dataset?

3. What are the classification algorithms considered and which algorithm has given the best performance on comparing Under-Sampled and Over-Sampled Data?

4. Is the average price change and lead time responsible for predicting hotel reservation cancelations?

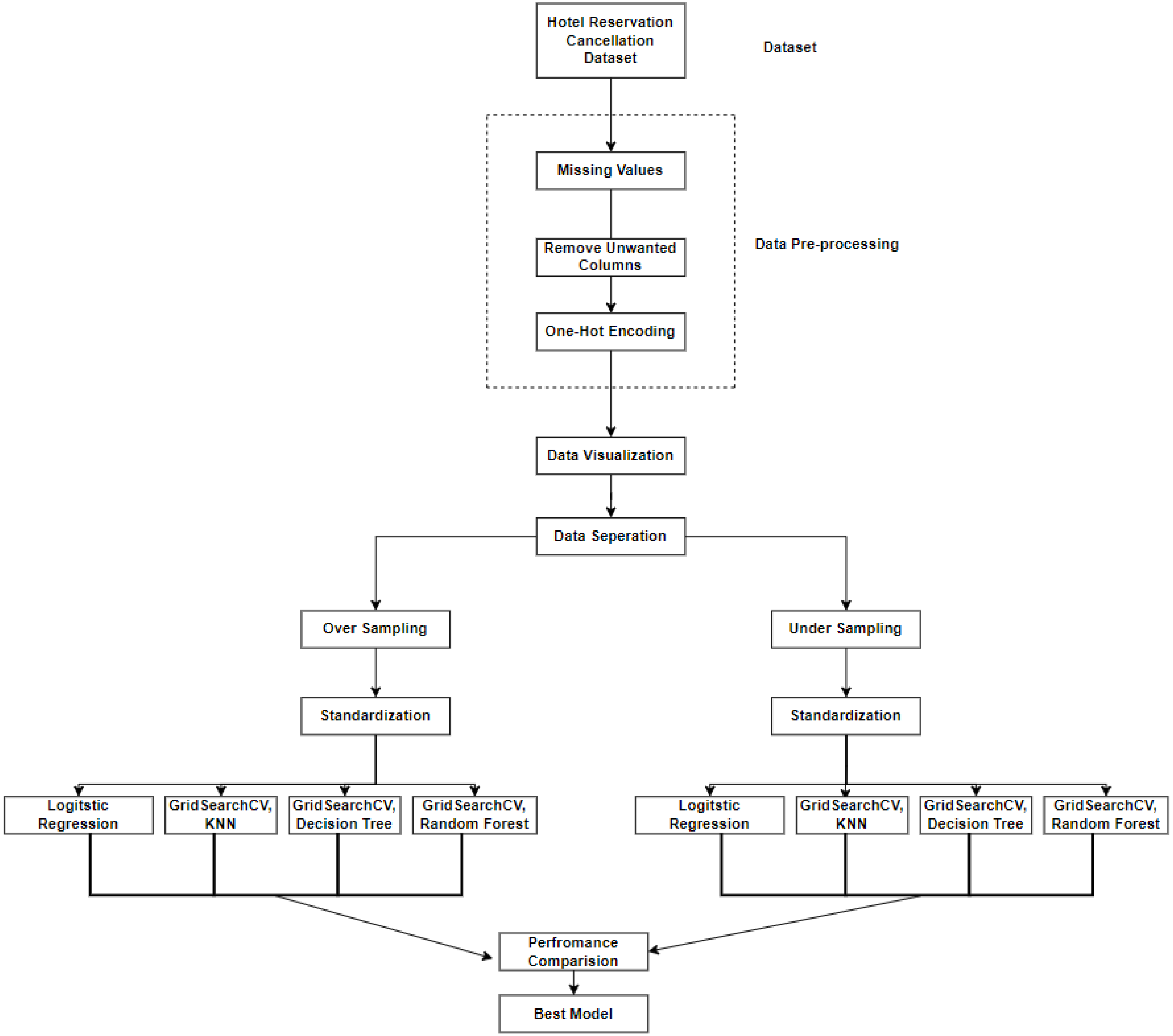
Hypothesis

- **Hypothesis:** will over-sampled data help achieve better performance than under-sampled data on classification algorithms?

- **Null Hypothesis:** There is no significant difference in the performance of classification algorithms trained on over-sampled data and under-sampled data.

- **Alternative Hypothesis:** Classification algorithms trained on over-sampled data will have better performance than classification algorithms trained on under-sampled data.

Workflow



Hotel Reservation Cancellation Dataset

Dataset

Missing Values

Remove Unwanted Columns

One-Hot Encoding

Data Pre-processing

Data Visualization

Data Seperation

Over Sampling

Under Sampling

Standardization

Standardization

Logitstic Regression

GridSearchCV, KNN

GridSearchCV, Decision Tree

GridSearchCV, Random Forest

Logitstic Regression

GridSearchCV, KNN

GridSearchCV, Decision Tree

GridSearchCV, Random Forest

Perfromance Comparision

Best Model

Data Loading and Understanding

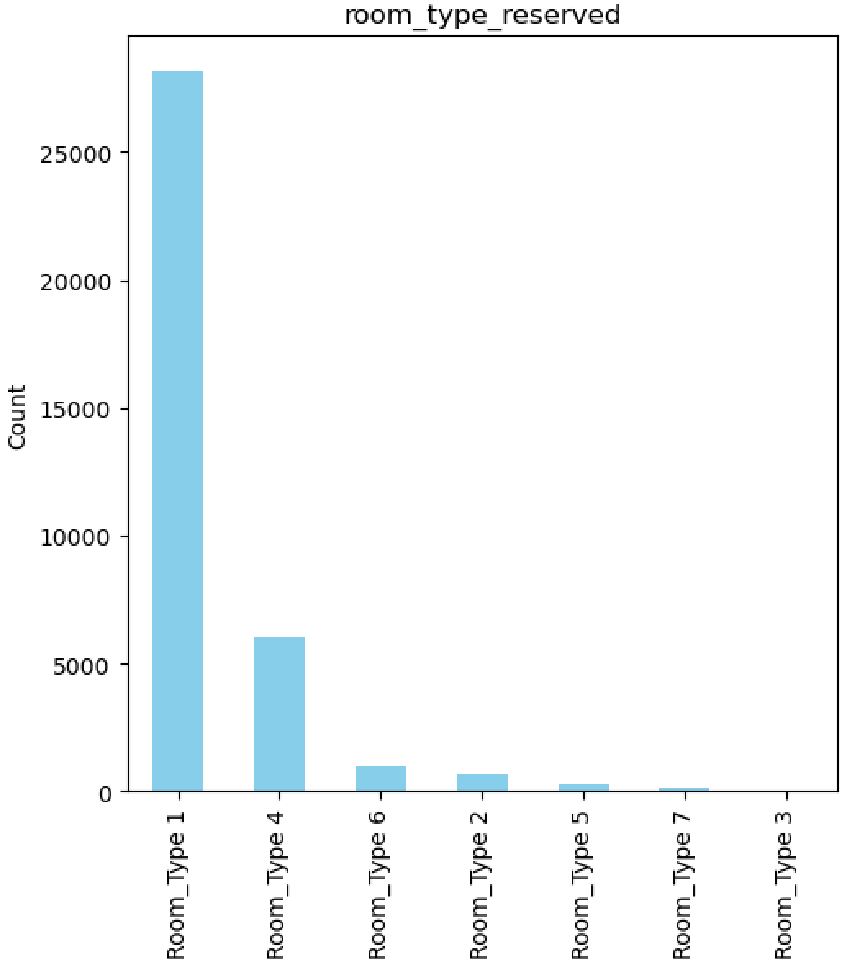
Hotel Reservations Data from Kaggle is considered for this project.

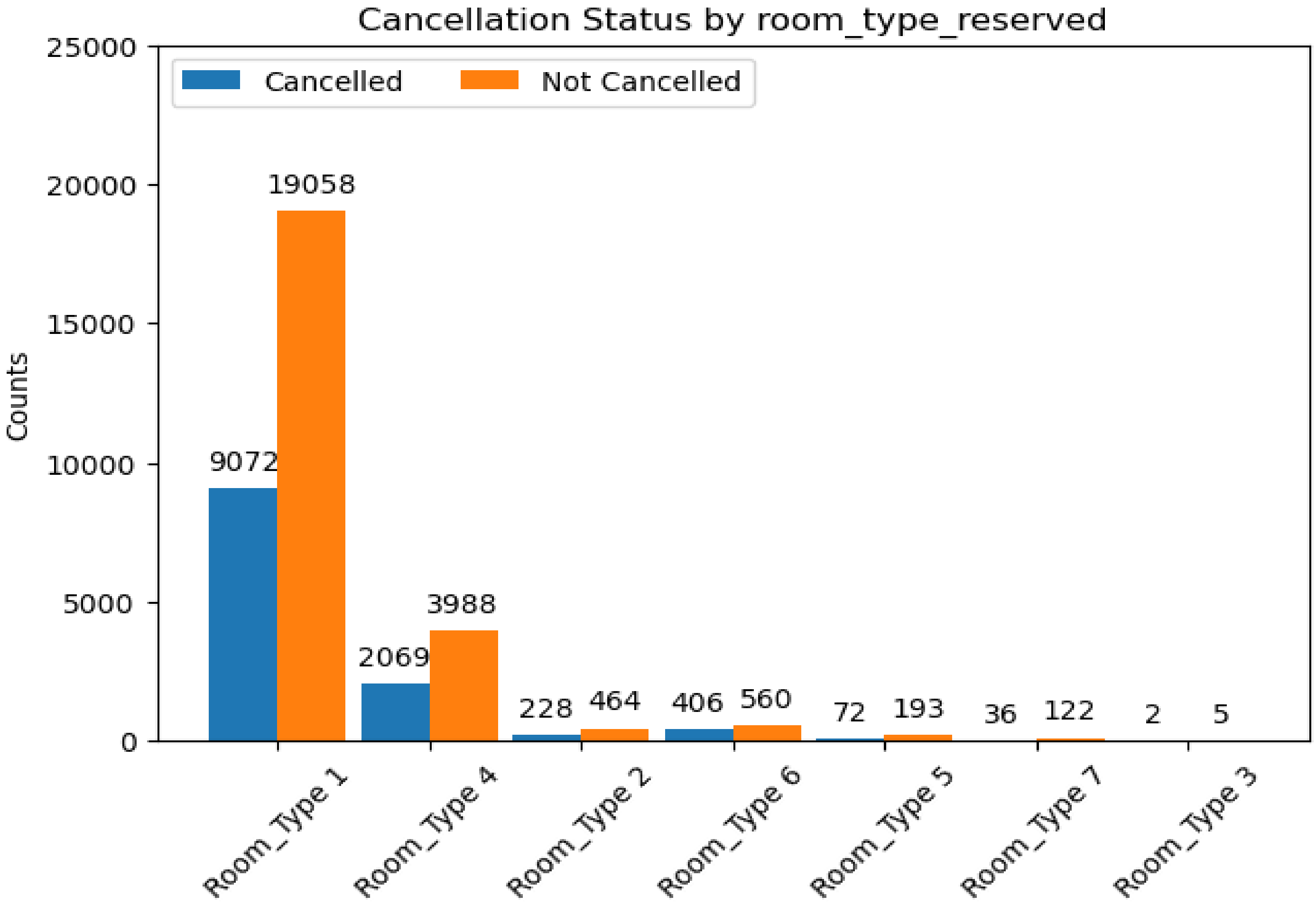
Data consists of 36275 rows and 19 columns.

Data Preprocessing

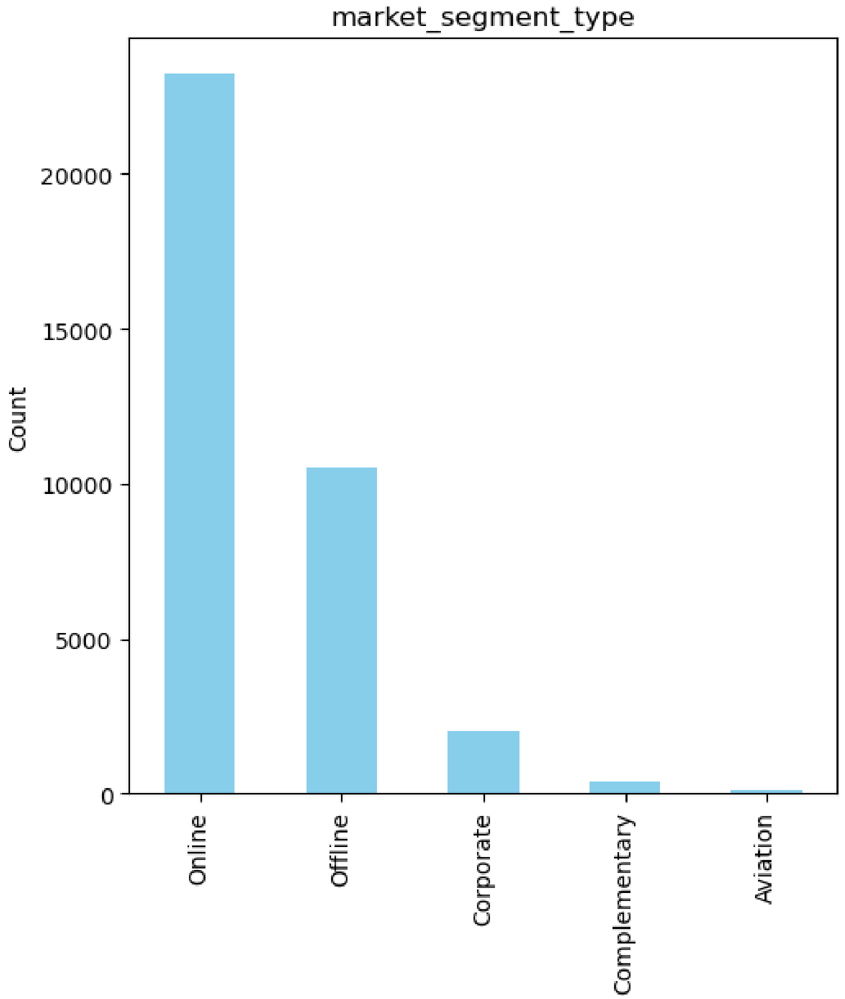
- Checking for missing values in the data.
- Removing unwanted columns from the data.
- One-hot encoding for categorical variables of independent data.

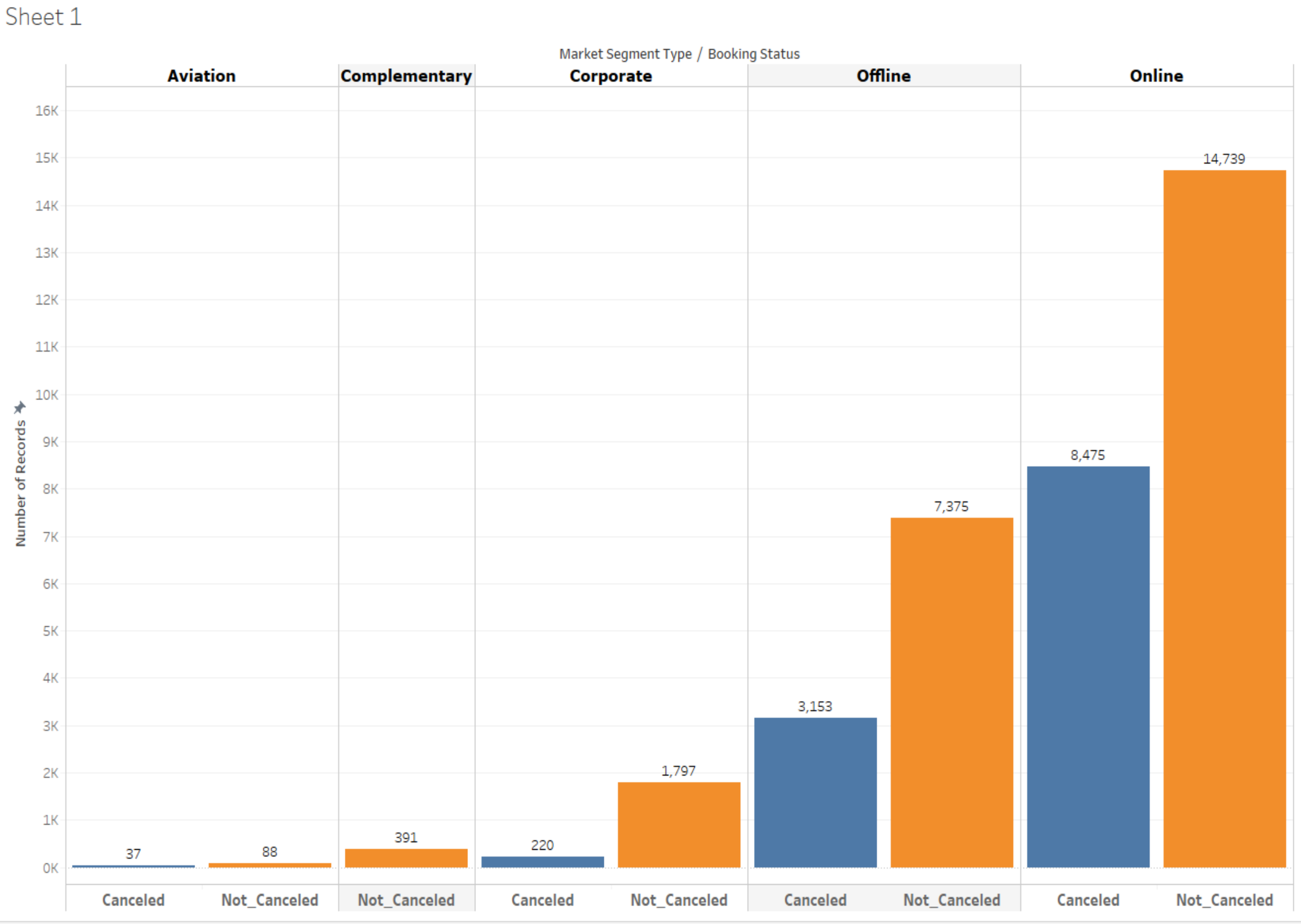
Data Visualizations



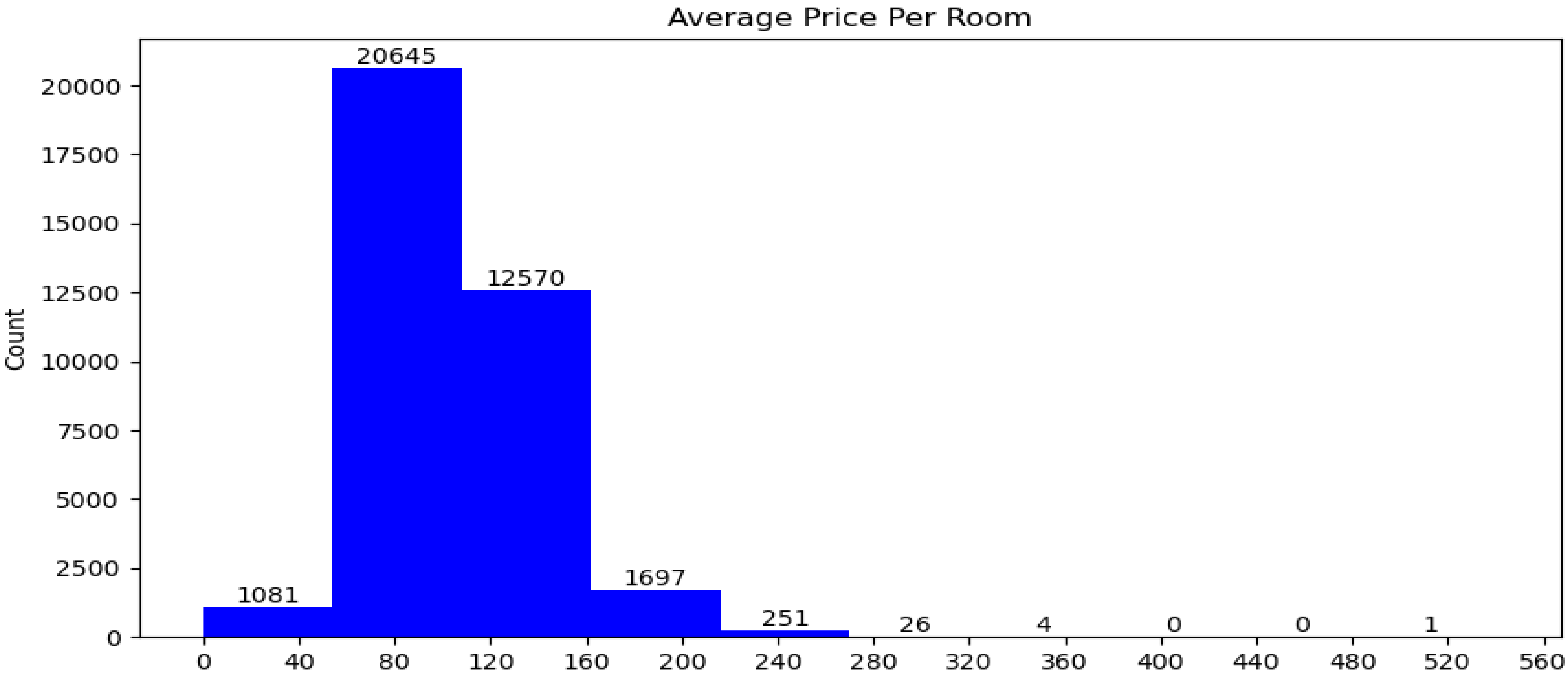


Data Visualizations

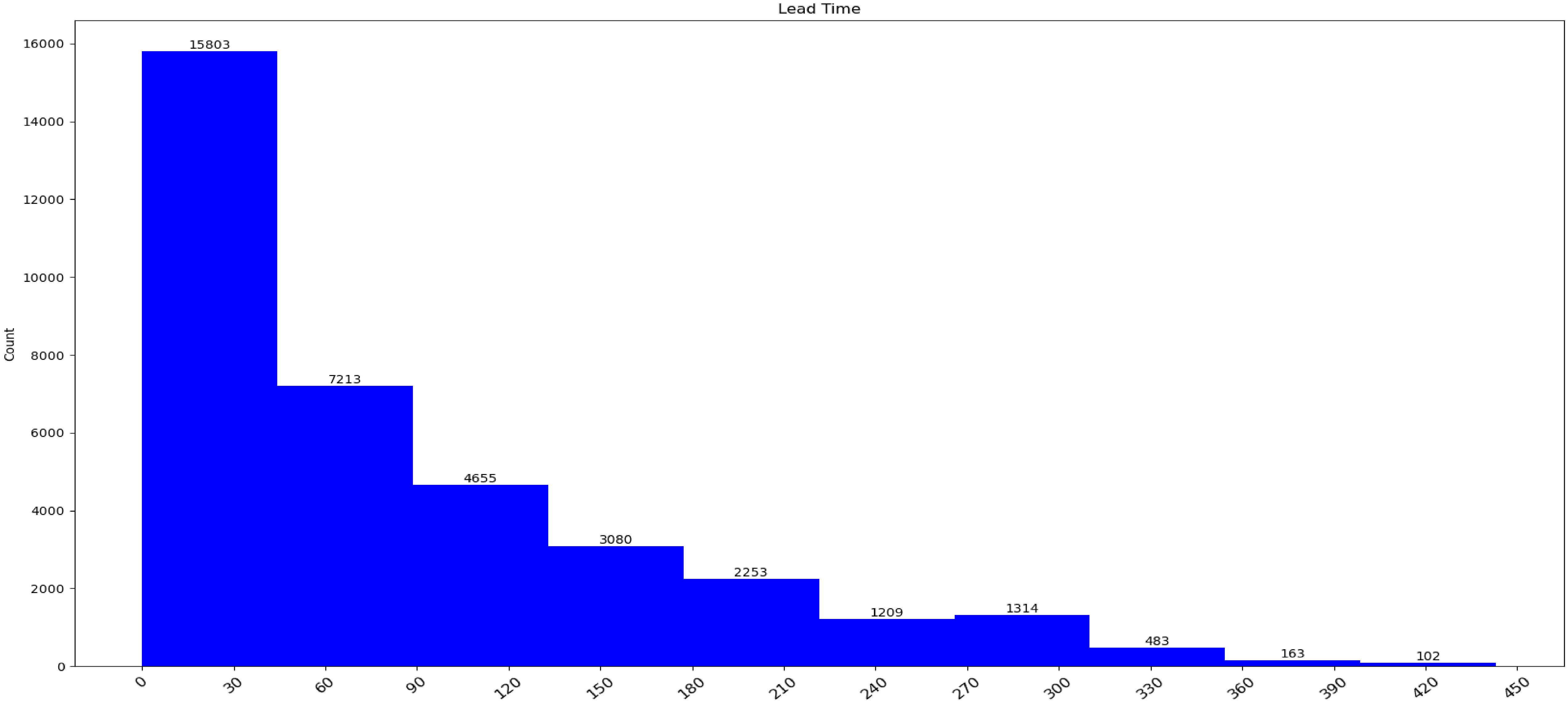




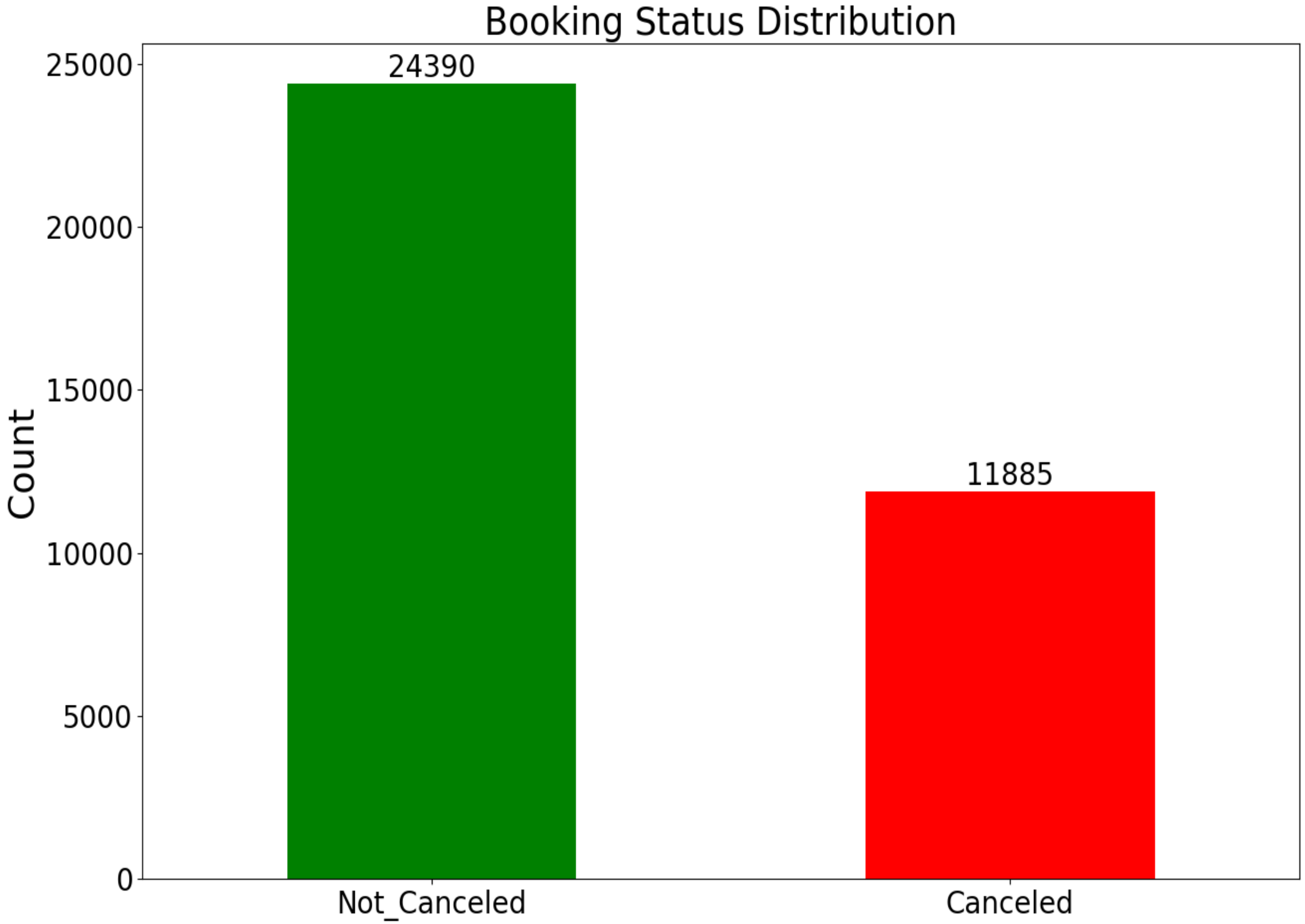
Data Visualizations



Data Visualizations



Data Visualizations



This picture shows the data is imbalanced with 24390 records in the Not Canceled category (Majority Class) and 11885 in the Canceled category (Minority class)

The next steps deal with the imbalanced data using over-sampling and under-sampling techniques and training the algorithms individually on both over-sampled data and under-sampled data their performances will be compared to answer the hypothesis and also to find out the best algorithm with the best performance.

Data Separation

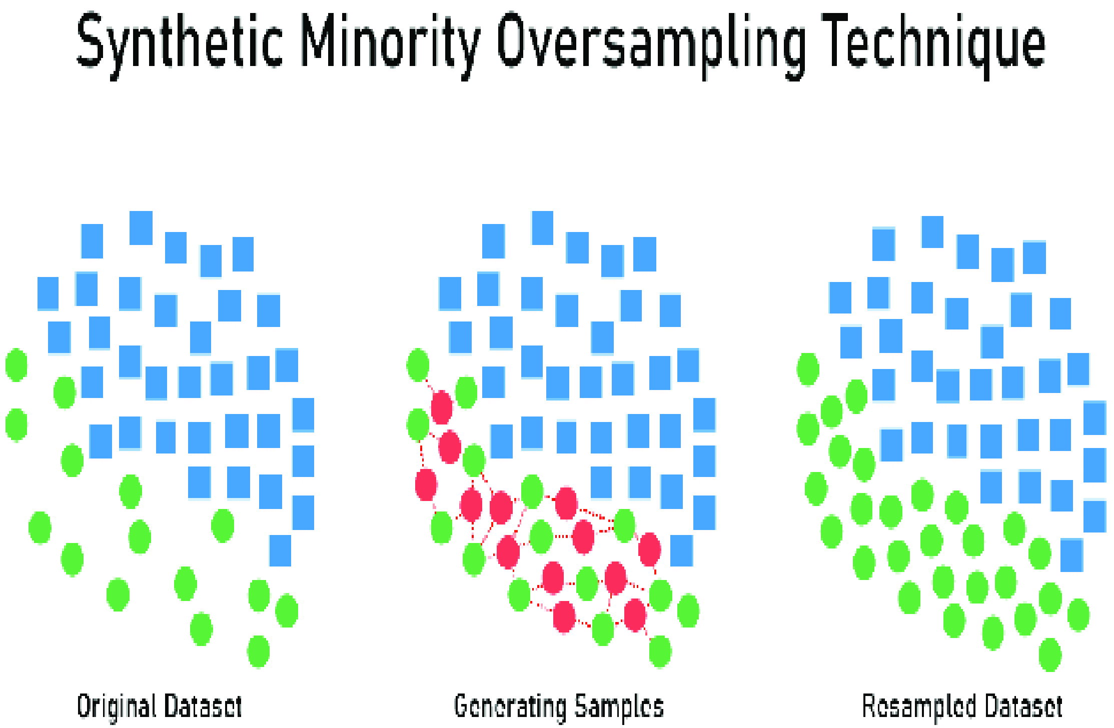
This dataset is split two times. Initially, data is separated into 85% train data with (30833, 27) and 15% into validation data with (5442, 27). Secondly, validation is split into two parts i.e., validation (80% of (5442, 27)) and test data (20% of (5442, 27)). Overall, below is the structure of the data before applying sampling techniques.

- ***Train data:*** (30833, 26)
- ***Train target:*** (30833,)
- ***Validation data:*** (4353, 26)
- ***Validation target:*** (4353,)
- ***Test data:*** (1089, 26)
- ***Test target:*** (1089,)

OVER SAMPLING DATA USING SMOTE

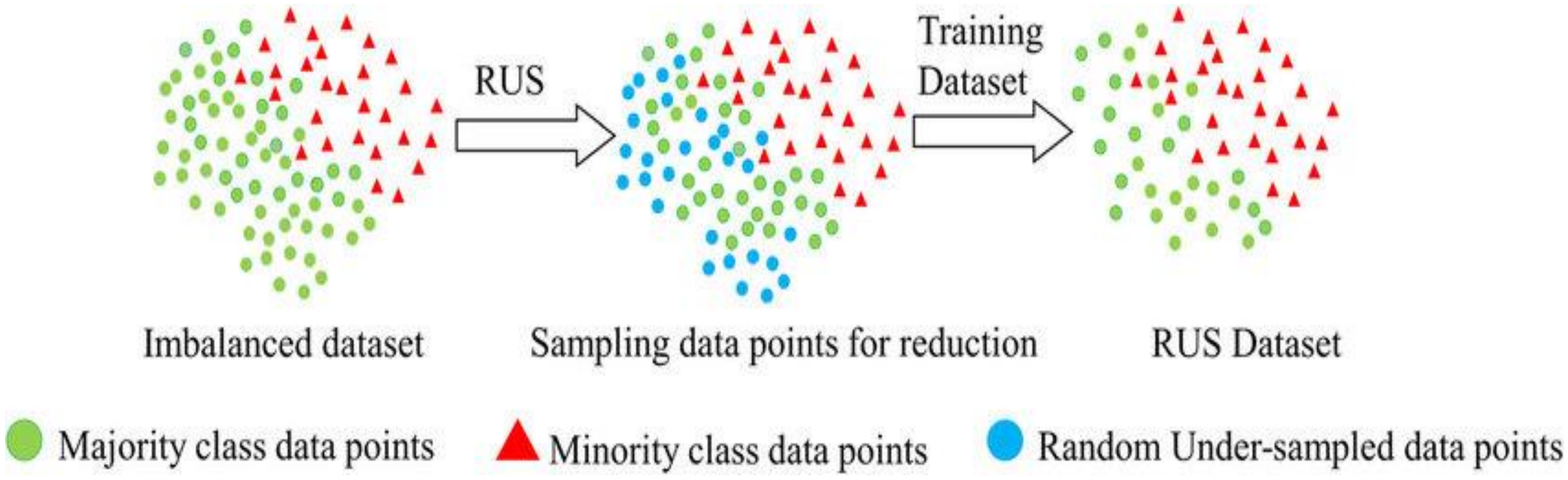
How Does SMOTHE(Synthetic Minority Over-sampling Technique) Work?

SMOTE (Synthetic Minority Over-sampling Technique) is a popular technique for addressing class imbalance by generating synthetic samples for the minority class. It creates new synthetic samples by interpolating between existing minority class samples.



UNDER SAMPLING THE DATA USING RUS

- Random under-sampling (RUS) is based on randomly removing the majority class, but other methods selectively undersample the majority class while keeping the original population of the minority class.
- Tough undersampling can lead to loss of the original data from the majority class but it is preferred if people want to experiment only on the original data instead of the synthetic data generated.


RUS
Training Dataset
Imbalanced dataset
Sampling data points for reduction
RUS Dataset
● Majority class data points
▲ Minority class data points
● Random Under-sampled data points

Standardization

Data standardization is a preprocessing technique that modifies data such that its mean (average) is 0 and its standard deviation is 1. All aspects may be immediately compared by making sure that each characteristic has a comparable scale.

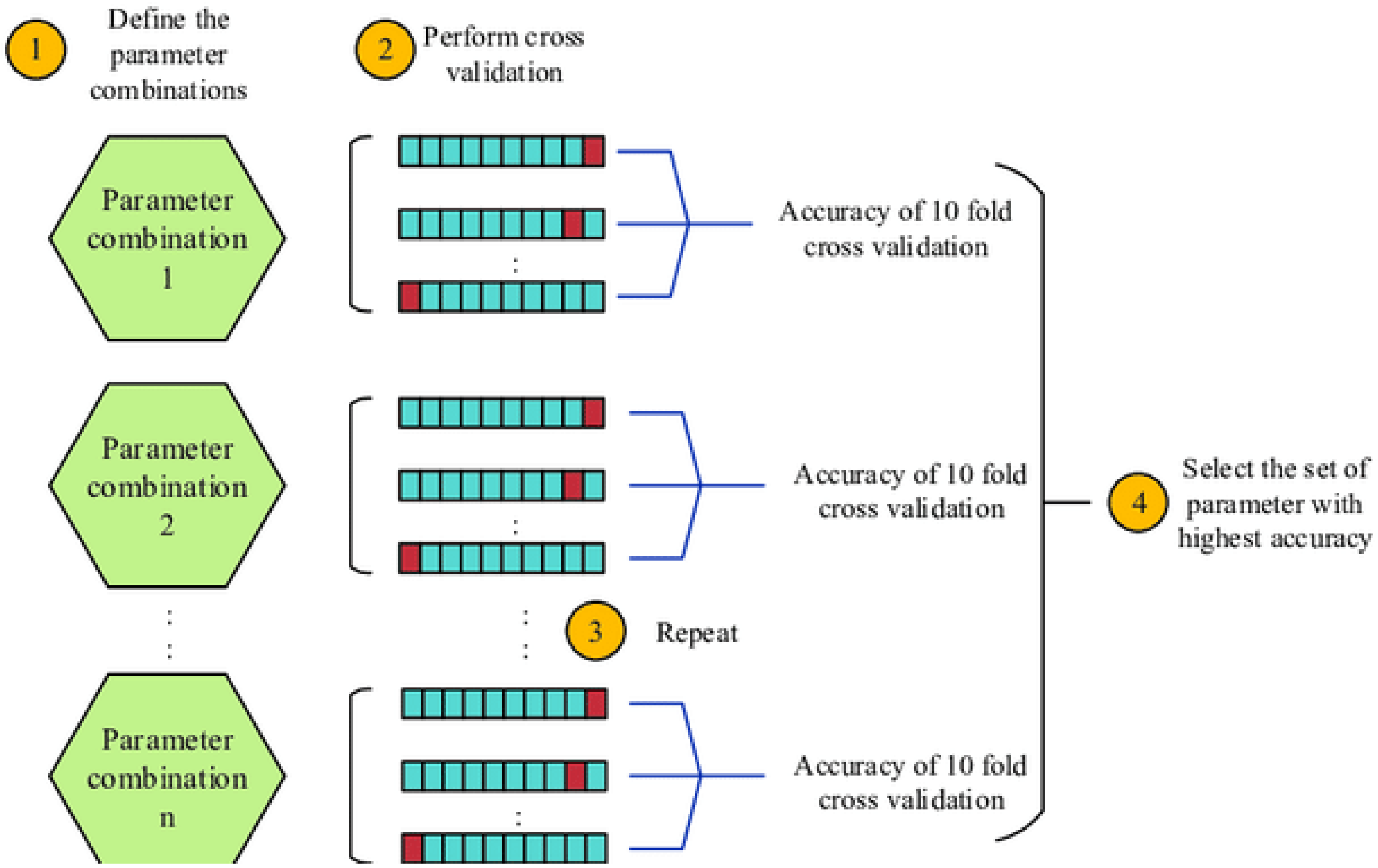
$$X' = \frac{X - \mu}{\sigma}$$

Where μ is the mean of the column values and σ represents the standard deviation of the column values, X is the datapoint from the population.

Note: Standardization is applied only after the data is completely ready after generation or reduction using various techniques in machine learning.

Grid Search Cross Validation (GridSearchCV)

Grid Search evaluates each hyperparameter combination's performance and chooses the optimal values. However, especially when there are several hyperparameters, it might be computationally costly.


1
Define the parameter combinations
2
Perform cross validation
Parameter combination 1
Accuracy of 10 fold cross validation
Parameter combination 2
Accuracy of 10 fold cross validation
4
Select the set of parameter with highest accuracy
3
Repeat
Parameter combination n
Accuracy of 10 fold cross validation

Training Machine Learning on Over-Sampled and Under-Sampled data separately

Grid search is applied for all the algorithms below on the user-chosen hypermeters and with the best parameters, the algorithm will be trained and tested on validation data and unseen data to track the performance.

Logistic Regression

K-Nearest Neighbours (KNN)

Decision-Tree

Random Forest and its feature importance

Performance Results

	Training Accuracy (Over Sampled)	Training Accuracy (Under Sampled)
Logistic Regression	79.30	77.79
K-Nearest Neighbours	99.38	99.26
Decision Trees	99.46	91.46
Random Forest	**99.46**	**97.92**

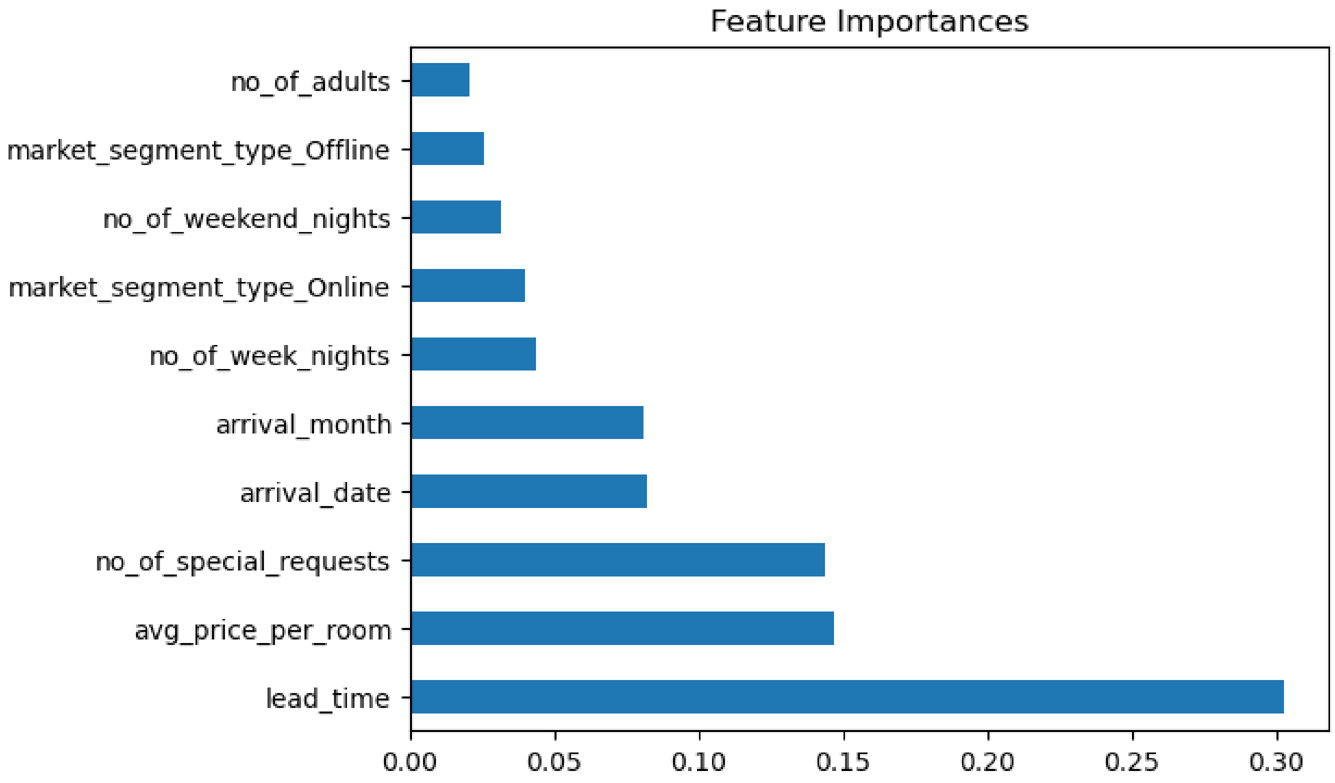
	Validation Accuracy (Over Sampled)	Validation Accuracy (Under Sampled)
Logistic Regression	78.15	77.92
K-Nearest Neighbours	86.72	84.58
Decision Trees	87.08	85.38
Random Forest	**90.55**	**89.34**

	Testing Accuracy (Over Sampled)	Testing Accuracy (Under Sampled)
Logistic Regression	77.68	78.23
K-Nearest Neighbours	87.87	84.38
Decision Trees	88.33	85.21
Random Forest	**89.99**	**89.25**

On comparing the performances of various machine learning algorithms on oversampled and under-sampled data, it is observed that Random Forest has performed the best in both cases. But Random Forest on Oversampled data has given the training accuracy of 99.46% Validation accuracy of 90.55% and Testing Accuracy is 89.99%.

All the algorithms trained on the oversampled data have achieved better performance on validation and test data compared to the algorithms trained on the under-sampled data. Hence, the **“Null Hypothesis has been rejected”.**

Feature Importance's from Random Forest



As expected, lead time and average price per room columns are the top two columns that are very important in predicting hotel reservation cancellation and can be the reasons for cancelations also. Hence, research question 4 is answered with the feature importance.

QUESTIONS?

THE END