CSUN | CALIFORNIA
STATE UNIVERSITY
NORTHRIDGE

COMP 542 Machine Learning Project

Group 12 (Fall 2022)

Student1: Sree Divya Sudagoni

Student2: Prathyusha Diwakarla

Submitted to: Prof.Mansoureh Lord

Department of Computer Science

California State University, Northridge.

INTRODUCTION –

1. PROBLEM STATEMENT:

One of the tourism industry's fastest-growing industries is the hotel industry. Although the hotel sector is expanding, there are certain advantages for hotels as well as some drawbacks. One issue is the increasing number of hotel reservations that are canceled.

The motivation for this project is to get a clearer picture of hotel booking demand. As a consumer, there are many aspects to consider when choosing a hotel. Predicting cancellations is useful not only for vacationers but also for hotels. The main idea of this project was to find the best classification model for predicting booking cancellations and the best explanatory variables for customer cancellations.

2. DATA DESCRIPTION:

The data for this project is from Hotel Booking Demand Dataset https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand.

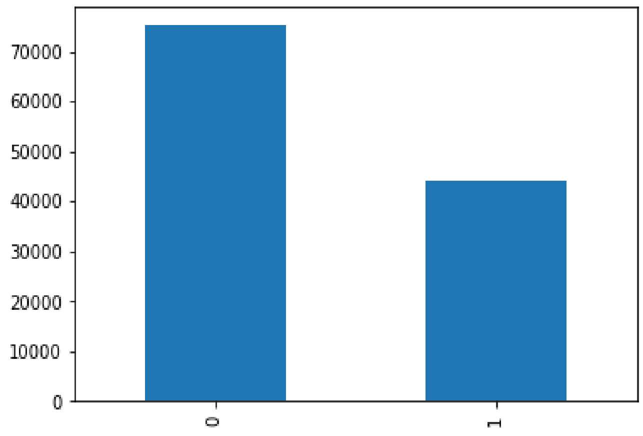
This data is extracted from two hotels in Portugal from July 2015 to August 2017. The data contains 119390 bookings with 32 features of which 20 are categorical and 12 are numerical columns. The target variable for this classification model is “is_cancelled”.

3. DATA PREPARATION:

Before building the model, checked if there are any duplicate rows the data consists of 26.8% of instances as duplicates and 73.2% of the instances are not duplicates. So, dropped all the duplicate rows and the resultant number of instances is “87351”. Checking for missing values in the data set is an important step, the majority of the missing values in the hotel data set are only in four features among which two features have the highest number of missing values. So, two of the features “company” and “agent” are dropped. Replaced the rest of the two features missing values with the mode of that column.

An outlier is a value that lies at an abnormal distance from other values. There were some outliers in the data, but they are not handled as they might be useful in future prediction. There can be a few incorrect values in the data like the people value is 0, but no booking is possible if 0 persons want to stay at the hotel. Dropped the instances where there are incorrect values.

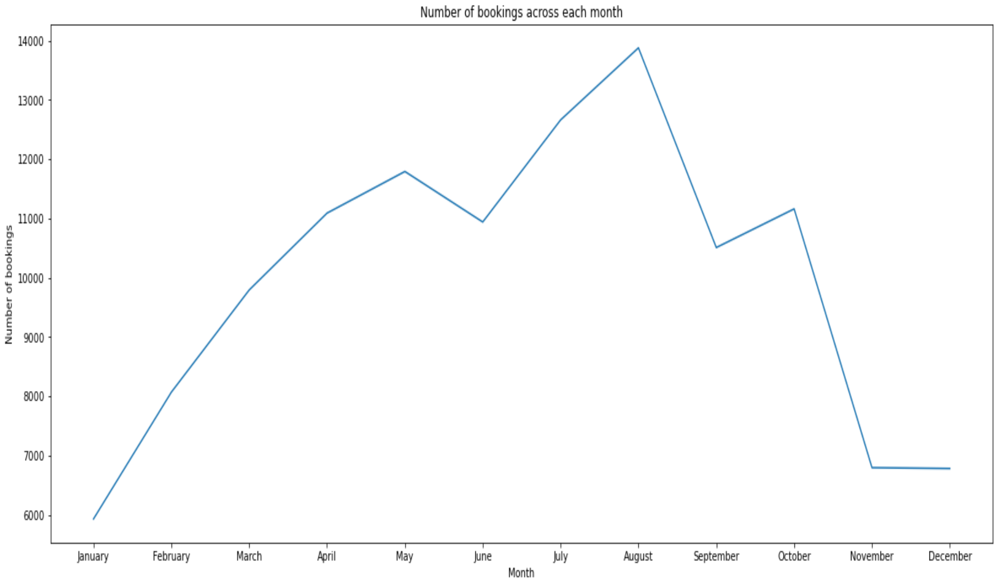
The majority class of this data set is “0” in the target variable “is_cancelled”. If we look at the image below, we can see that there is a difference between the majority class and the minority class. This way the model can be biased towards the majority class. SMOTE technique is used to handle the imbalance present in the data.



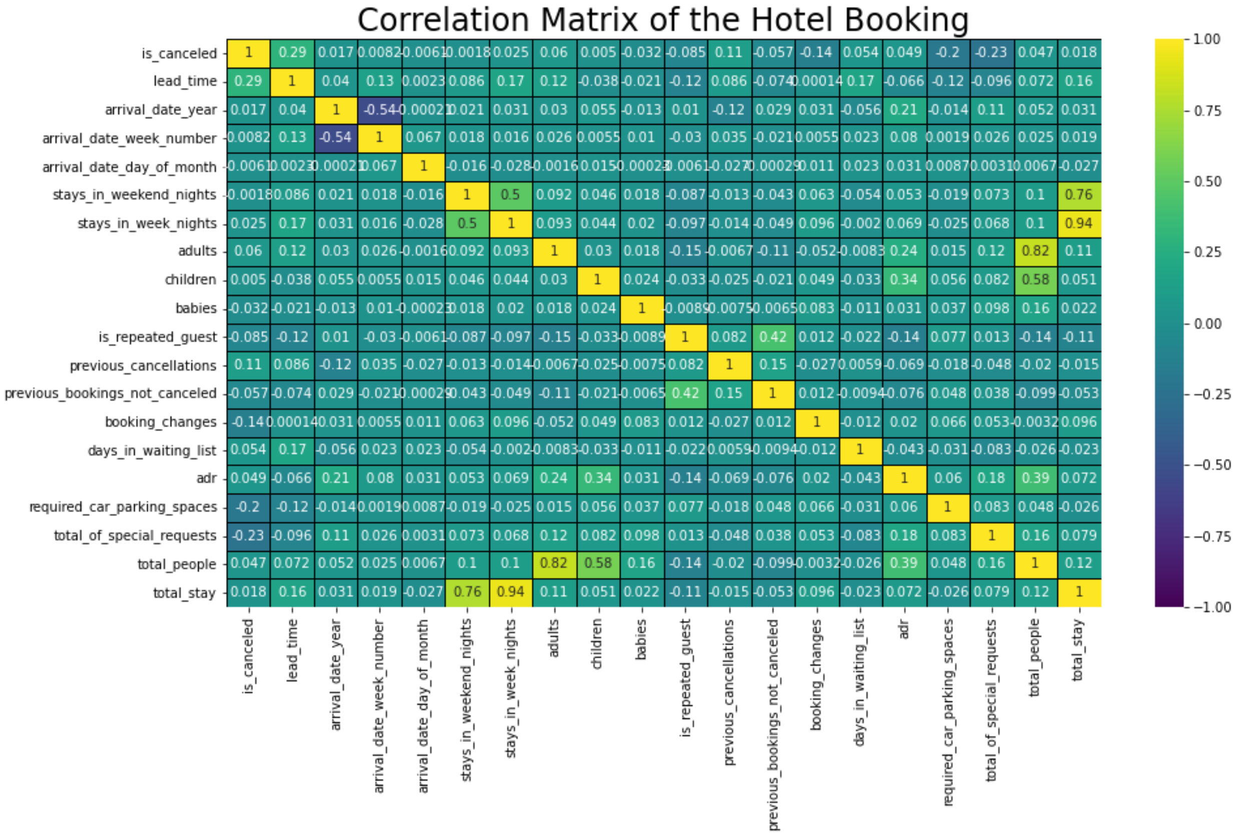
Combined few columns which has the same meaning into a new column and added them to the data set. Thereby reducing the number of features.

4. DATA VISUALIZATION:

Univariate Data visualization summarizes the statistics of one variable in the data set. Visualized the data to understand it better by using Univariate, Bivariate, and Multivariate analysis. Below is the Line graph which shows which month has the highest number of bookings.



Correlation Matrix of the hotel booking (Multivariate analysis):



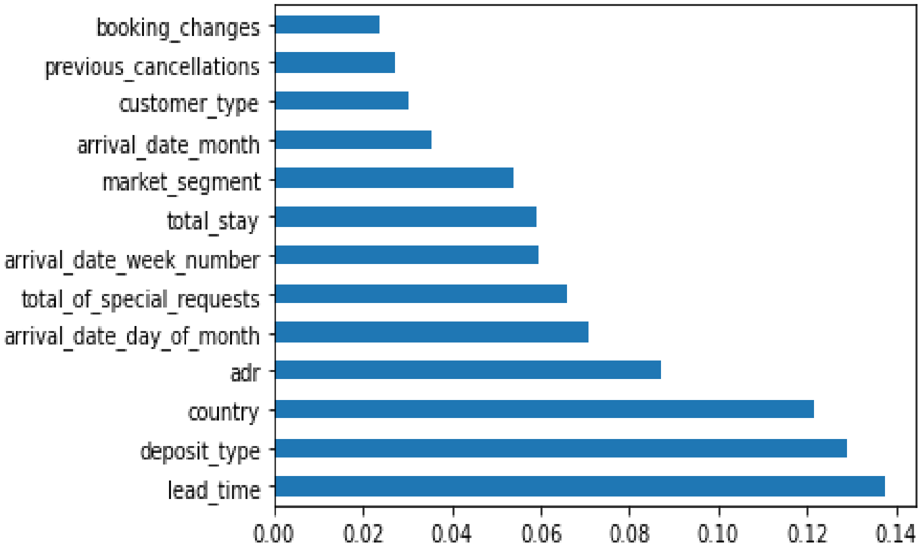
RELATED WORK -

There is a related paper that worked on the same problem “Predicting hotel booking cancellations to decrease uncertainty and increase revenue” which was published at the 2017 16th IEEE Conference on Machine Learning and Applications (ICMLA). Data were collected directly from the Hotel’s PMS databases using Microsoft SQL Server. The classification algorithms used were ‘Boosted Decision Tree’, ‘Decision Forest’, ‘Locally Deep Support Vector Machine’, and ’Neural Network’. Cross-validation was used to evaluate the performance of each algorithm, specifically k-fold cross-validation.

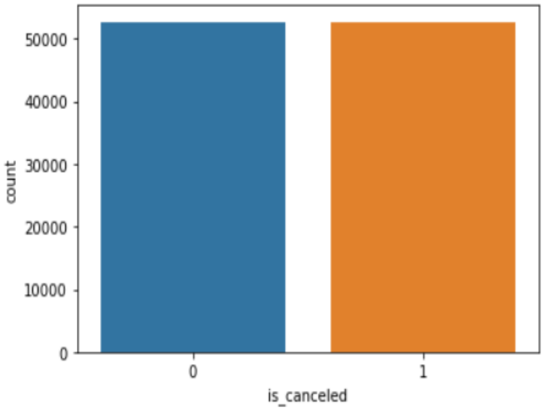
By using PMS data from four hotels in conjunction with the application of data visualization, data mining, and machine learning, was possible to answer the three main objectives of the research: Identity which features contribute to predicting a booking cancellation probability. Build a model to classify bookings likely to be canceled and with that build a better net demand forecast. Understand if one model could be applied to all hotels.

METHODS–

To perform the feature extraction, ExtraTreesClassifier() technique is used. Initially, the data set contained 32 features, and by using the above technique the most important features are extracted and the number of features after the technique is 14. Extra Trees Classifier is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a forest to output its classification result.



SMOTE (Synthetic minority oversampling technique) is an oversampling technique used for imbalanced data problems. It aims to increase the instances of minority classes by adding new instances. These new instances are generated by using the k-nearest neighbor algorithm to find data points between the minority class data point and its nearest neighbor.

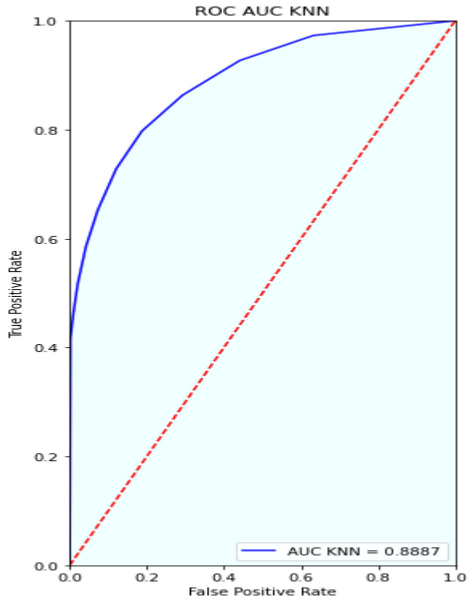


Label Encoding technique is used as the data contains Categorical data. It converts the labels into a numeric form to convert them into machine-readable form. The robust Scaler technique is used to perform the feature scaling to standardize the independent features present in the data in a fixed range.

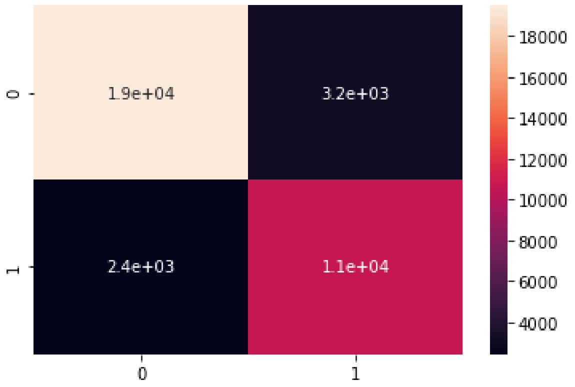
The algorithms used to solve the problem are K-Nearest Neighbor and Decision Tree. Also, after oversampling the data using SMOTE the model is trained again using the above two algorithms to compare the results.

Using KNN the hyperparameters used are different K values ranging from 5-40, weight distribution is set to uniform and distance. The neighbor distance is calculated using Euclidian and Manhattan distance techniques. GridSearchCV method is used to use all the different available combinations and perform the cross-validation according to the 'cv' value. The accuracy score achieved using KNN is 85% on test data. After using the SMOTE technique, the training data is fit into the model, and prediction is performed. The accuracy score after using SMOTE on KNN resulted in 84%. There is no drastic difference between the accuracy score, but the recall value is generalized when SMOTE technique was used.

The ROC(Receiver Operating Characteristic curve) displays the performance of the model based on the True Positive rate and False positive rate.



Heat map of KNN using SMOTE:



Decision tree algorithm :

Decision tree is a tree shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, or an occurrence of an event. The interpretability of decision trees is likely their most significant advantage, and they are quite helpful. Since they are simple to understand and very quick to train, their potential extends far beyond the confines of traditional science. Entropy and information gain are the two important terminologies of DTs. Entropy is the measure of randomness in the data while information gain is the decrease in entropy after the dataset is split. Nodes, branches, and leaves make up DTs. Each leaf represents an outcome, each branch represents a rule, and each node represents an attribute (or feature). The number of levels, excluding the root node, determines the depth of a tree.

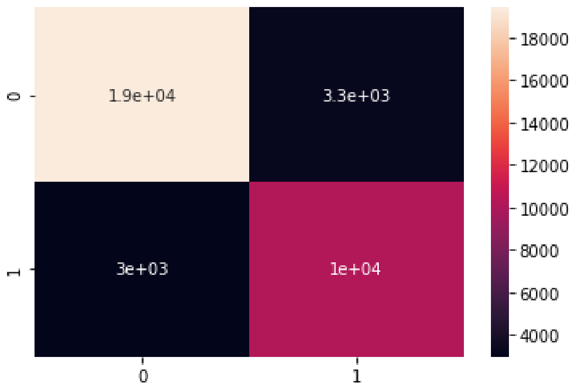
Parameters:

1. **criterion*{"Gini", "entropy", "log_loss"}, default="Gini"***
2. **min_samples_split*int or float, default=2***
   The minimum number of samples required to split an internal node:
3. **min_samples_leaf*int or float, default=1***
   The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches.
4. **max_depth*int, default=None***
   The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

I have used the default criterion, the Gini impurity measure. It is used to decide the optimal split from a root node, and subsequent splits.

Heat map of Decision Tree using SMOTE:

```
<matplotlib.axes._subplots.AxesSubplot at 0x218305bf160>
```

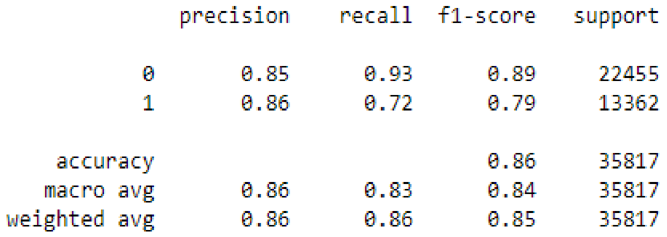


EVALUATION -

The accuracy of decision tree algorithm after hyperparameter tuning is 83% and the accuracy of K-Nearest Neighbors algorithm after hyperparameter tuning is 85%.

Hence, tuned K-Nearest Algorithm has the best accuracy when compared with Decision Tree Algorithm Furthermore, after using Smote, K-Nearest algorithm has the best accuracy among the two.

Classification report of K-Nearest algorithm–

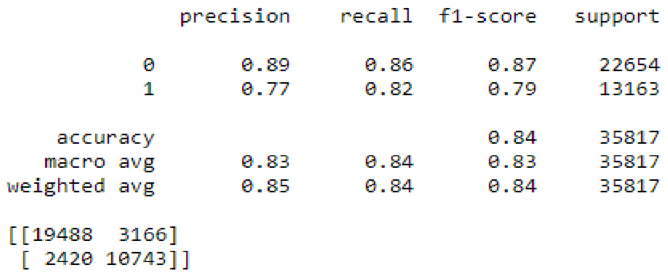


```
              precision    recall  f1-score   support

           0       0.85      0.93      0.89     22455
           1       0.86      0.72      0.79     13362

    accuracy                           0.86     35817
   macro avg       0.86      0.83      0.84     35817
weighted avg       0.86      0.86      0.85     35817
```

Classification report of K-Nearest algorithm Using SMOTE –



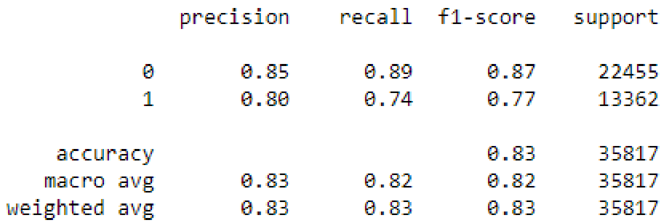
```
              precision    recall  f1-score   support

           0       0.89      0.86      0.87     22654
           1       0.77      0.82      0.79     13163

    accuracy                           0.84     35817
   macro avg       0.83      0.84      0.83     35817
weighted avg       0.85      0.84      0.84     35817

[[19488  3166]
 [ 2420 10743]]
```

Classification report of Decision Tree –

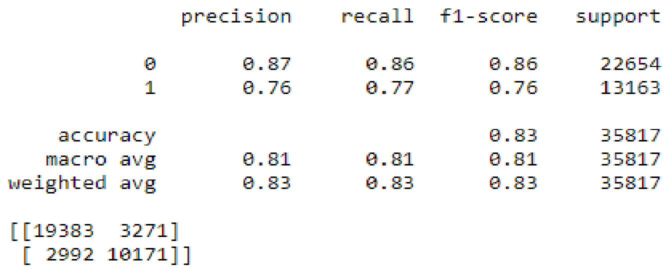


```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87     22455
           1       0.80      0.74      0.77     13362

    accuracy                           0.83     35817
   macro avg       0.83      0.82      0.82     35817
weighted avg       0.83      0.83      0.83     35817
```

Classification report of Decision Tree Using SMOTE –



```
              precision    recall  f1-score   support

           0       0.87      0.86      0.86     22654
           1       0.76      0.77      0.76     13163

    accuracy                           0.83     35817
   macro avg       0.81      0.81      0.81     35817
weighted avg       0.83      0.83      0.83     35817

[[19383  3271]
 [ 2992 10171]]
```

CONCLUSION –

The Machine learning models created using K-Nearest Neighbors and Decision Tree can predict whether a person is going to cancel his hotel booking or not. The above has been concluded by using some instances of the data set as test data and the Models were able to predict the target variable value successfully as expected.

FUTURE WORK:

This project employed data from two hotels in Portugal which raise some questions that further research could help explain:

Can similar results be obtained given any location?

Can we train the model with more hotels integrated into the model?

Further research could also make use of features of additional data sources such as weather information, the social reputation of the hotel, and many more to measure the influence of these features in booking cancellations.

REFERENCES –

1. https://towardsdatascience.com/principal-component-analysis-vs-extratreesclassifier-d0217bbcc9a8
2. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/
3. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
5. https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/
6. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/
7. https://www.youtube.com/watch?v=4jRBRDbJemM
8. https://scikit-learn.org/stable/modules/grid_search.html
9. https://proclusacademy.com/blog/robust-scaler-outliers/