AIT – 580

Research Project

Visualization and Prediction of Customer Rating of an Airport

1. Introduction

The current report details the process of answering several research questions regarding the overall satisfaction of the customers of Austin-Bergstrom International Airport. It comprises data sources, exploration of data, analysis of data through visualization, model development, fine-tuning, and performance metrics.

2. Literature Review

Air travel is one of the most vital and frequently used transportation in today's life. It is also one of the world's largest businesses. Maintaining high standards and quality is very important for airports to increase their profits. They should also be able to provide a comfortable and a satisfactory experience to all the travelers who use the airport.

In *'High Enough?: Explaining and Predicting Traveler Satisfaction Using Airline Reviews'* (Lacic et.al., 2016) the author tries to predict the traveler's satisfaction based on the ratings of four to five features. Apart from the rating they also use textual reviews to make the prediction. They obtained the ratings and textual reviews from the internet. Using the best performing features from this data they built classifiers that predict the traveler's satisfaction.

3. Dataset Description

The name of the dataset is Airport Quarterly Passenger Survey**.** The dataset has the results obtained from the customer surveys from Austin-Bergstrom International Airport. These surveys are used to improve the level of customer satisfaction and to improve their performance in specific areas. The selected dataset has 37 columns and 3464 rows. Several features of the dataset are the 'Efficiency of check-in staff', 'Wait time at passport inspection', 'Courtesy of inspection staff', 'Thoroughness of security inspection', 'Feeling of safety and security', 'Ease of finding your way through the airport', 'Overall satisfaction', etc. These features are a few questions in the survey form. The records comprise of each customer rating in a specific quarter in a specific area. The rating ranges from 0 to 5, where 0 is the least and 5 is the best rating possible. The other ratings are 1, 2, 3, and 4.

(i) Source- The dataset used is obtained from the government data catalog (City of Austin 2017). The URL of the data source is https://catalog.data.gov/dataset/airport-quarterly-passenger-survey

(ii) Reason for dataset selection and its Importance:

The dataset has been selected to develop a model that can automate the process of predicting the overall rating of the airport. Most informative features can also be identified so that airports can improve their customer satisfaction by identifying and allocating resources to the areas which affect their ratings highly.

The dataset is important as it details the factors that help improve customer satisfaction. It can help the authorities concentrate more on the important facilities. Through a detailed analysis, we can identify the important and unimportant features. Multi-class classification is performed to predict the overall rating of the Airport.

4. Associated Research Questions

1. Can a multi-class classification model be trained to predict the overall rating of the airport given customer survey data?
2. What are the factors affecting the overall rating?

5. Data types of variables

1. Alpha Numeric – For example Quarter in which the rating is given
2. Date and time – For example date and time of the given rating
3. Int – Rating of Cleanliness of the airport, Courtesy of the airport staff, Overall satisfaction, etc

6. Data Exploration

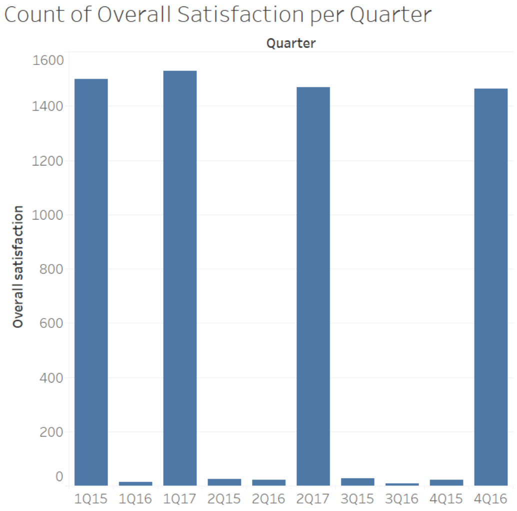


Fig. 1

The entire dataset is divided into ten quarters. The surveys are taken four times in 2015, four times in 2016, and two times in 2017. From fig. 1, we can observe that the first quarter of 2017 has the highest number of survey records. The least survey records are in the third quarter of 2016. The number of surveys taken in the first quarter of 2017, the first quarter of 2015, and the second quarter of 2017 are the top three quarters where there were maximum surveys conducted when compared to the other quarters. The year 2017 has the maximum number of surveys done when compared to 2016 and 2015. The least number of surveys were done in the year 2016.

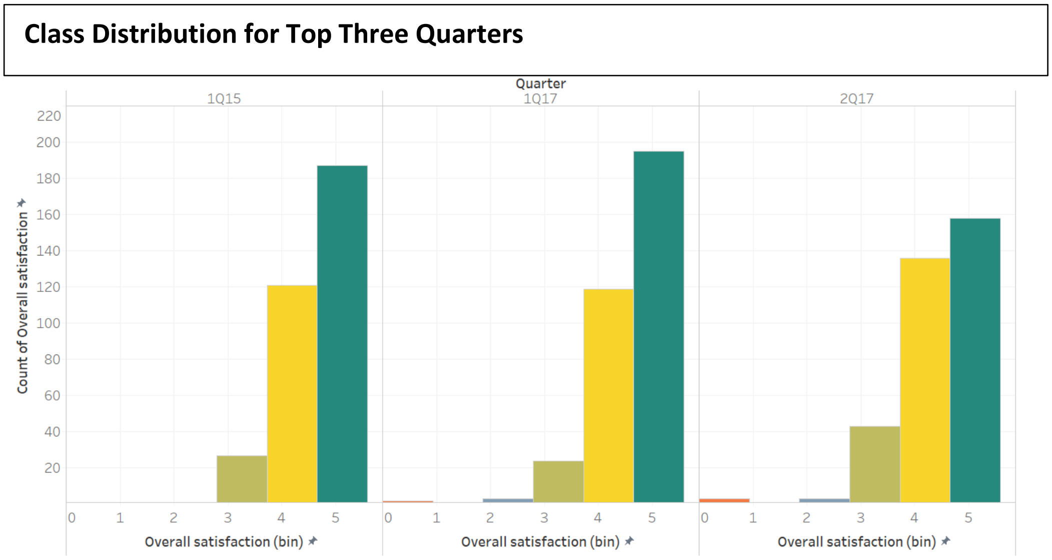


Fig. 2

The class distribution of the top three quarters with the highest number of survey records is visualized in fig. 2. We can notice that the number of records in class five is highest in all three quarters followed by the number of records of class 4 and the number of records of class 3. Keen observation reveals that there are no records for class zero, one and two in the first quarter of 2015.

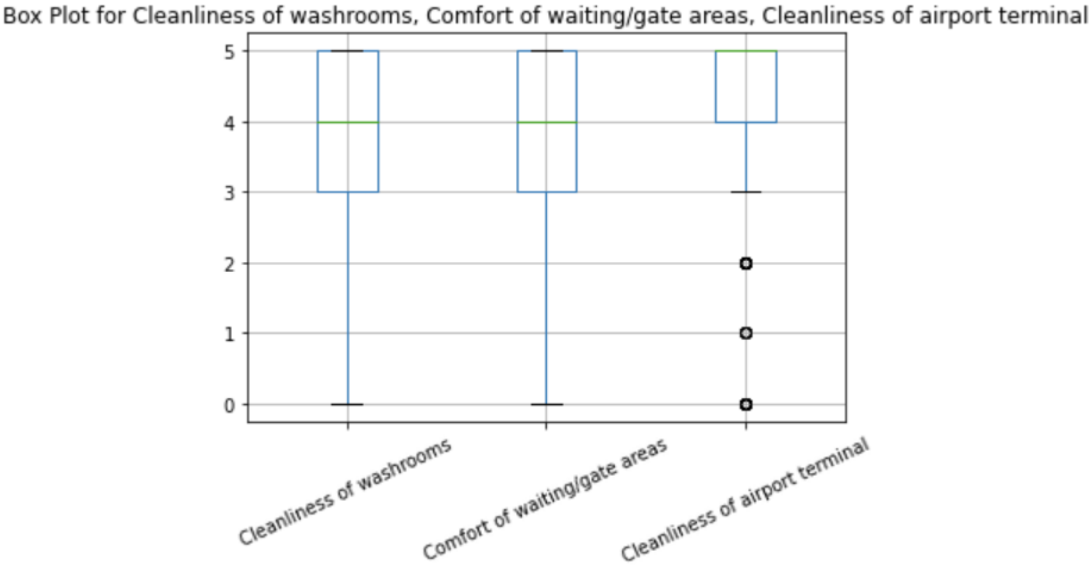


Fig. 3

From among the 37 features 'Cleanliness of washrooms', 'Comfort of waiting/gate areas' and 'Cleanliness of airport terminal’ are the three features for which Box plots are plotted. Box plots are used in exploratory data analysis to show the distribution of the data and skewness by displaying the data quartiles (Mcleod, 2019). Box plots for these columns can be seen in fig. 3. The inter-quartile range is highest for the columns 'Cleanliness of washrooms' and 'Comfort of waiting/gate areas' and least for the column 'Cleanliness of airport terminal’. In the column 'Cleanliness of airport terminal’ we can find several outliers at classes zero, one, and two.

(i) Regression Analysis

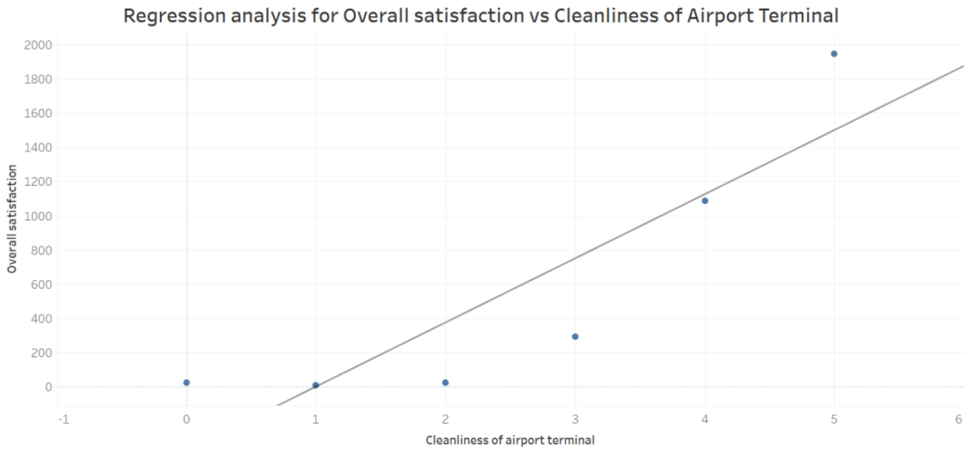


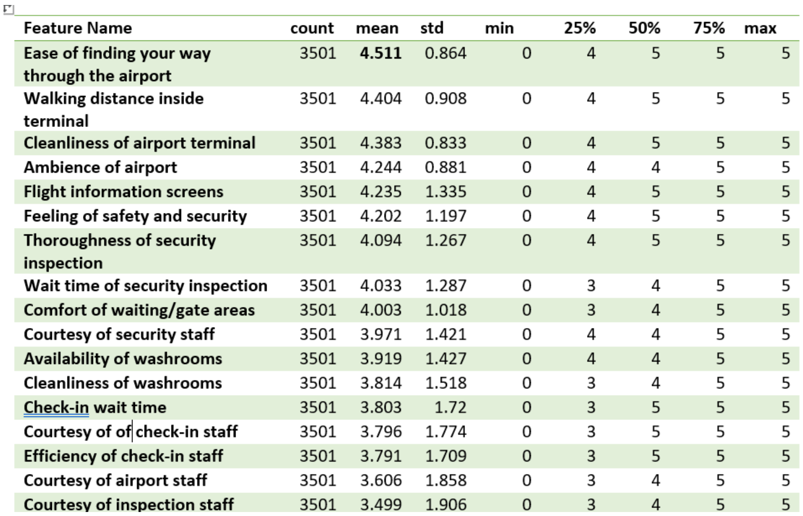
Fig. 4

Regression analysis has been performed on ‘Overall satisfaction’ and ‘Cleanliness of airport terminal’. A trend line has been fitted to understand the relation between both the features. It can be concluded from fig. 4 that both the features are positively correlated. Increase in the ‘Overall satisfaction’ is evident with the increase in ‘Cleanliness of airport terminal’.

(ii) Data Preprocessing

All the columns in the dataset are not being used in the analysis. The columns 'Quarter', 'Date recorded', 'Departure time' are removed. The data set had 2239 NaN values that are replaced with the most repeated value in the column (mode). The description of the columns can be seen below. Only a few columns are in fig. 5. The description of the remaining features is present in the attached file.

Description of Features of the dataset



Feature Name	count	mean	std	min	25%	50%	75%	max
Ease of finding your way through the airport	3501	**4.511**	0.864	0	4	5	5	5
Walking distance inside terminal	3501	4.404	0.908	0	4	5	5	5
Cleanliness of airport terminal	3501	4.383	0.833	0	4	5	5	5
Ambience of airport	3501	4.244	0.881	0	4	4	5	5
Flight information screens	3501	4.235	1.335	0	4	5	5	5
Feeling of safety and security	3501	4.202	1.197	0	4	5	5	5
Thoroughness of security inspection	3501	4.094	1.267	0	4	5	5	5
Wait time of security inspection	3501	4.033	1.287	0	3	4	5	5
Comfort of waiting/gate areas	3501	4.003	1.018	0	3	4	5	5
Courtesy of security staff	3501	3.971	1.421	0	4	4	5	5
Availability of washrooms	3501	3.919	1.427	0	4	4	5	5
Cleanliness of washrooms	3501	3.814	1.518	0	3	4	5	5
Check-in wait time	3501	3.803	1.72	0	3	5	5	5
Courtesy of of check-in staff	3501	3.796	1.774	0	3	5	5	5
Efficiency of check-in staff	3501	3.791	1.709	0	3	5	5	5
Courtesy of airport staff	3501	3.606	1.858	0	3	4	5	5
Courtesy of inspection staff	3501	3.499	1.906	0	3	4	5	5

Fig. 5

(iv) Class Distribution

The dataset has 5 classes, ranging from 0 to 5. Each value represents a rating that determines 'Overall Satisfaction'. The class distribution is as follows.

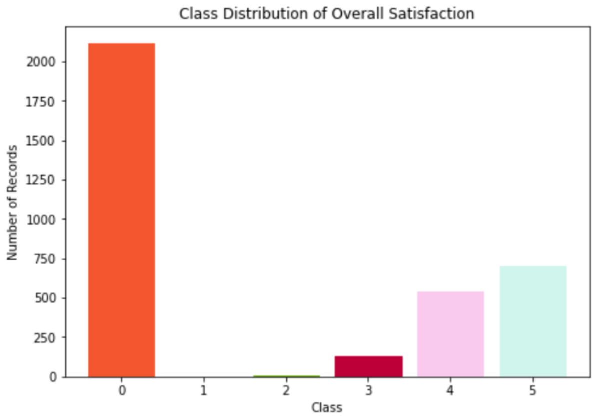


Fig. 6

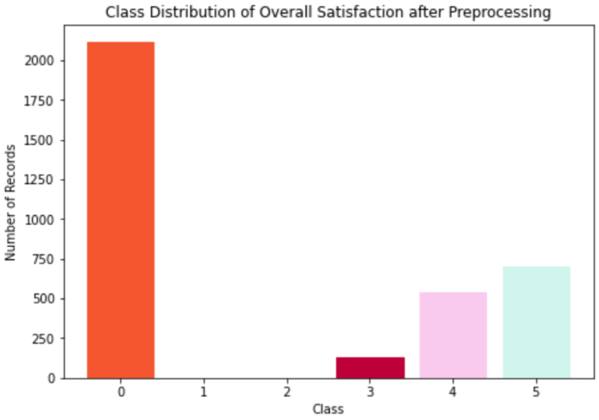


Fig. 7

From figure 6, we can notice that the number of records for class 1 is very low when compared to the other classes. The number of records for class 2 is also very low. It is tough to identify any kind of pattern from the few records present. Their presence may result in inconsistent and inaccurate results. Hence, it is better to remove the records having classes 1 and 2 to obtain reliable and consistent predictions. A total of 11 records were removed from the dataset making the total number of records in the dataset to 3490. The class distribution after preprocessing is as shown in fig. 7.

(v) Using Correlation to reduce features

A correlation matrix has been constructed for all the 35 columns present after preprocessing in the dataset to remove any similar features.

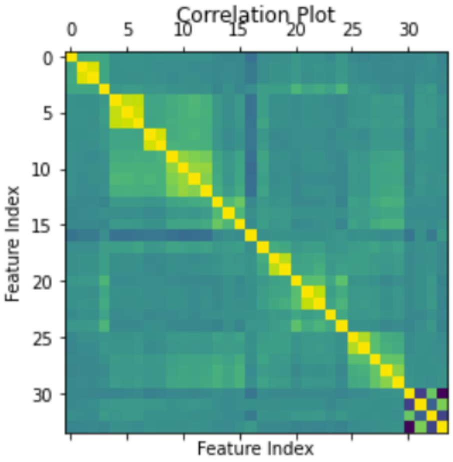


Fig. 8

Hence, highly correlated features whose correlation value is greater than 0.9 are removed from the dataset. The columns Parking facilities and Parking facilities (value for money) have a correlation value of 0.9. Hence, the column Parking facilities (value for money) has been removed. The purpose of constructing a correlation plot is to remove highly correlated columns from the data as they do not add any additional information.

7. Split the data into Train and Test Set

The entire data is split into a training set and testing set. The training set comprises 80% of the records and the test set comprises 20% of the records.

8. Data Modeling

(i) Modeling using Random Forest

Random forest is a supervised learning algorithm that randomly generates and merges several decision trees. The aim of Random Forest is not to rely on one model and instead collectively depend on several models to obtain better accuracy. Random forest is used for classification or regression tasks. Here, we are using a random forest to predict the Overall satisfaction rating of a customer. This task is known as multi-class classification (DeepAI, 2019).

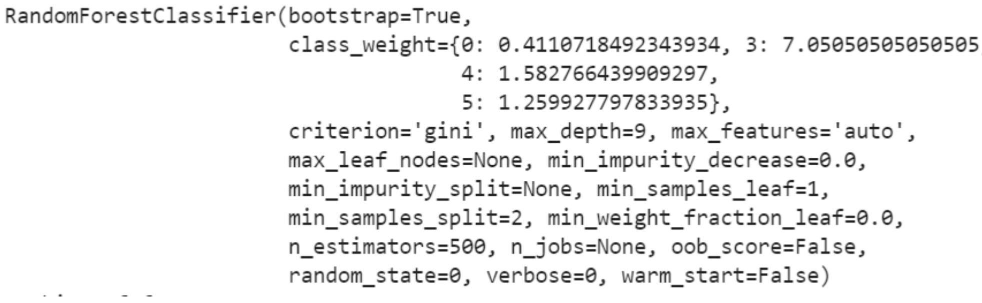
(ii) K-fold cross-validation

K-fold cross-validation of 4 splits is performed on the training data. K-fold cross-validation is a method to measure the performance of a model on a new dataset (Brownlee, 2019).

(iii) Hyper-parameter tuning

In a machine learning model, we always prefer to obtain the most feasible model. To do so, we need to explore several possibilities. Instead of manually exploring the possibilities of model architecture, we can ask the machine to perform this for us. Hyper-parameters are the parameters that define a model's architecture and the process of finding the best model is known as hyper-parameter tuning (Jordan, 2018).

Grid search is performed to obtain the best-suited hyper-parameters. Grid search is a method to perform hyper-parameter optimization. The best combination of hyper-parameters for any given model and any test set can be obtained [5]. The hyper-parameters used for random forest are



```
RandomForestClassifier(bootstrap=True,
                       class_weight={0: 0.4110718492343934, 3: 7.05050505050505,
                                     4: 1.582766439909297,
                                     5: 1.259927797833935},
                       criterion='gini', max_depth=9, max_features='auto',
                       max_leaf_nodes=None, min_impurity_decrease=0.0,
                       min_impurity_split=None, min_samples_leaf=1,
                       min_samples_split=2, min_weight_fraction_leaf=0.0,
                       n_estimators=500, n_jobs=None, oob_score=False,
                       random_state=0, verbose=0, warm_start=False)
```

Fig. 9

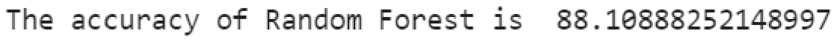
The best hyper-parameters for random forest can be seen in fig. 9. These hyper-parameters can be given as input to the random forest classifier. In k-fold cross-validation, the model is trained on k-1 folds and tested on 1-fold. Average accuracy, average precision, and average recall and average f1 score obtained by the set of hyper-parameters are as follows.

Cross-Validation Metrics for Random Forest

Model	Average Accuracy	Average Precision	Average Recall	Average F1 score
Random Forest	0.8843	0.7423	0.7517	0.7354

Table. 1

The random forest classifier is now trained on the entire training set and the model is saved. The test set is given as input to the saved model to obtain the predictions. Using these predictions, the accuracy of the random forest classifier on the entire data set can be obtained.



```
The accuracy of Random Forest is  88.10888252148997
```

Fig. 10

The classification report of the classifier can be obtained by using the classification report function.

Classification Metrics for Random Forest

Class	Precision	Recall	F1-score	Support
0	**0.98**	**0.95**	**0.96**	**421**
3	0.84	0.52	0.64	31
4	0.66	0.82	0.73	97
5	0.81	0.81	0.81	149

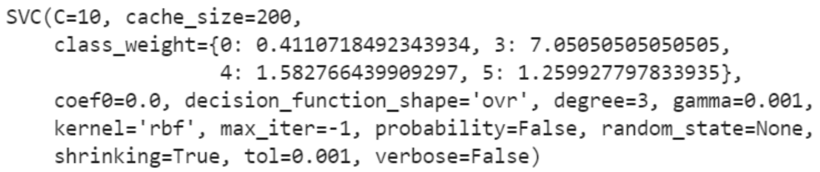
Table. 2

(iv) Modeling using SVM

SVM stands for Support Vector Machines which is a machine learning algorithm used to analyze data to perform classification and regression. It gives an output map of the sorted data with margins between the two. They are used to classify text, images, used to identify handwriting (Techopedia, (n.d.). In this project, we are using support vector machines to classify the Overall satisfaction of a traveler into multiple classes.

(v) K-fold cross-validation and Hyper-parameter tuning

4-fold cross-validation is performed on the training data. A grid search is performed to obtain the hyper-parameters for the SVM classifier. The hyper-parameters obtained for the Support Vector Machine classifier are as follows



```
SVC(C=10, cache_size=200,
    class_weight={0: 0.4110718492343934, 3: 7.05050505050505,
                  4: 1.582766439909297, 5: 1.259927797833935},
    coef0=0.0, decision_function_shape='ovr', degree=3, gamma=0.001,
    kernel='rbf', max_iter=-1, probability=False, random_state=None,
    shrinking=True, tol=0.001, verbose=False)
```

Fig. 11

The best hyper-parameters for SVM can be seen in fig. 11. The obtained hyper-parameters are given as input to the SVM classifier. The model is trained on k-1 folds and tested on 1-fold in k-fold cross-validation.

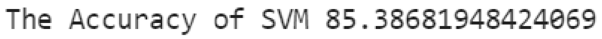
By using the obtained hyper-parameters, average accuracy, average precision, average recall, and average f1 score can be obtained.

Cross-Validation Metrics for Support Vector Machines

Model	Average Accuracy	Average Precision	Average Recall	Average F1 score
Support Vector Machines	0.8492	0.7423	0.7517	0.7354

Table 3

The entire training set can be used to train the model which is saved. The test set is given as input to the saved model to get the predictions. The accuracy of the SVM can be obtained by using the predicted values.



```
The Accuracy of SVM 85.38681948424069
```

Fig. 12

The classification function can be used to obtain the classification report of the Support Vector Machine.

Classification Metrics for Support Vector Machine

Class	Precision	Recall	F1-score	Support
0	**0.98**	**0.93**	**0.95**	**421**
3	0.47	0.68	0.55	31
4	0.62	0.70	0.66	97
5	0.8	0.77	0.79	149

Table 4

9. Results and Evaluation

Model Name	Precision	Recall	F1-Score	Accuracy
Random Forest	**0.8225**	**0.775**	**0.785**	**88.1**
Support Vector Machine	0.7175	0.77	0.737	85.3

Table 5

The precision, recall, f1 score, and accuracy for the Random Forest is greater than the accuracy of SVM. Hence, it can be concluded that Random Forest can model the data better than SVM.

10. Answers to Research Questions

Answer for Research Question 1: Two models namely Random Forest classifier and SVM classifier have been constructed to predict the overall satisfaction of the travelers. Good performance was observed across different classification metrics by both the classifiers. The metrics also prove that it is possible to build a machine learning model on this dataset. The performance of the model can be improved if more data could be made available.

Answer for Research Question 2: The feature importance for each column in the dataset is obtained during the training process. The decision trees inside random forest calculate the information gain and split the feature having the highest information gain first. The feature importance is high for a feature if its information gain value is high. Thus, the feature importance for each column is determined. The top three features which have the highest feature importance are 'Cleanliness of washrooms', 'Comfort of waiting/gate areas', and 'Cleanliness of airport terminal'. It can be implied that Overall satisfaction is highly dependent on these features.

11. Limitations

The dataset has only the ratings of various attributes. Also, the size of the data set is not that large. As it is a survey dataset there is no guarantee that the people will always give honest reviews. The answers to the questions may be interpreted by different people differently. The respondents may not be aware of their answers due to a lack of memory.

12. Future recommendations

Better metrics can be obtained if the data set has different data types. For example, the number of travelers that use the airport every day, the number of arrivals and departures, etc. Also, the presence of textual data would provide great scope to build a multiclass classifier using Natural Language Processing. The presence of more records will help in obtaining better classification metrics.

References

Brownlee, J. (2019, August 08). A Gentle Introduction to k-fold Cross-Validation. Retrieved May 10, 2020, from https://machinelearningmastery.com/k-fold-cross-validation/

City of Austin. (2017). Airport Quarterly Passenger Survey [Data file]. Retrieved on March 29th,2020 from https://catalog.data.gov/dataset/airport-quarterly-passenger-survey

DeepAI. Random Forests. (2019, May 17). Retrieved May 10, 2020, from https://deepai.org/machine-learning-glossary-and-terms/random-forest

Jeremy Jordan. (2018, December 05). Hyperparameter tuning for machine learning models. Retrieved May 10, 2020, from https://www.jeremyjordan.me/hyperparameter-tuning/

Lacic, E., Kowald, D., & Lex, E. (2016, July). High enough? explaining and predicting traveler satisfaction using airline reviews. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (pp. 249-254).

Mcleod, S. (2019, July 19). Box plots (also known as box and whister plots). Retrieved May 10, 2020, from https://www.simplypsychology.org/boxplots.html

Techopedia. (n.d.). What is a Support Vector Machine (SVM)? Retrieved May 10, 2020, from https://www.techopedia.com/definition/30364/support-vector-machine-svm