Daqian Dang
Final Report
DATS 6103: Introduction to Data Mining
Dr. Amir Jafari
June 23, 2022

Proposal

Classification problems I selected are supervised learning problems, which the training dataset consists of data related to independent variables and target class. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to help a bank judging credit card applicants as low, medium, or high risk on the approval of issuing credit card.

The Kaggle website I chosen is because there are lots of public classification datasets, and some of the datasets came from the real world. The classification datasets from Kaggle website are the great and useful dataset for data analysis beginner to practice machine learning and improve analysis skill. Python will be primarily utilized for analyzing the large dataset. Python is an interpreted, general purpose, and high-level language with an objected oriented approach and this is a useful tool for analyzing big data.

Since the dataset I selected is a classification problem, the Python libraries could bring me several important analysis tools, such like data mining, data preprocessing, and data modeling along with visualization, which support me to clean, transform, and visualize the data as well as build the model for the data. The resources from Python libraries are Pandas, Numpy, Seaborn, Matplotlib, and Sklearn. In addition, five main model algorithms, logistic regression, naïve bayes, decision tree, random forest, k-nearest neighbors, will be applied to the classification dataset.

To validate and judge the performance of the five models, I will use accuracy score, F1-score, AUC/ROC, and confusion matrix.

Introduction

The dataset, Airline Passenger Satisfaction, was selected and downloaded from Kaggle website, the reason I selected and wanted to analyze the dataset was because I have been interested in understanding what/which important factors could influence an airline passenger/customer's satisfaction level and it would bring the result in how the importance of customer satisfaction could affect airline company business.

In this project, the combination of knowledge, what I learned from data mining algorithms and concept in classes and the code examples and exercises in Python, is implemented to understand and analyze the dataset. The five main phases are outlined in the project …

- Basic understanding about the dataset
- Exploratory data analysis and data visualization
- Data preprocessing
- Data modeling
- Modeling evaluation

Description of the dataset

At first, the dataset from Kaggle website has originally two CSV files with train and test data, and then I merged the two CSV files into one file. The dataset consists of 129,880 observations and 25 columns including '*Unnamed: 0', 'id', 'Gender', 'Customer Type', 'Age', 'Type of Travel', 'Class', 'Flight Distance', 'Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness', 'Departure Delay in Minutes', 'Arrival Delay in Minutes', and 'satisfaction'.*

The dataset includes five columns were objects, such as *'Gender'* includes Male and Female*, 'Customer Type'* includes loyal customer and disloyal customer*, 'Type of Travel'* includes personal travel and business travel*, 'Class'* includes business, Eco, and Eco plus*, 'satisfaction'* includes neutral or dissatisfied and satisfied, and other 20 columns were int64 and float64 as they have numeric values. In addition, the dataset variables divide into two variable groups such as numeric variables and categorical variables. The numeric variables include '*id', 'Age', 'Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes',* and the categorical variables are *'Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness'* as they have five-point satisfaction rating scales, such as 1 represents 'very dissatisfied', 5 represents 'very satisfied', and 0 is for neutral or no opinion.

In particular, the specified feature in the dataset is '*satisfaction*', which is a binary classification. It describes the customer's satisfaction as it has two binary responses named 'neutral or dissatisfied' and 'satisfied'. The binary classification feature *'satisfaction'* is the target variable for the dataset.

Description of algorithms

Outlier detection

An outlier is a point of data that is distant from other cluster of datapoints, and it can skew overall data trends, which will affect the result of data analysis. Therefore, outlier detection is an important method to identity outliers in the dataset which has almost 130K observations.

The interquartile range method I used is to find outliers. First quartile Q1 equals median of the n smallest values, third quartile Q3 equals median of the n largest values, and the interquartile range is IQR = Q3 – Q1. The lower range is Q1 – 1.5*IQR and the upper range is Q3 + 1.5*IQR, which I can find which outliers that are under the lower area or above the upper area and the outliers can be removed. (geeksforgeeks.org., 2020)

Since the dataset is a classification problem, I used five model algorithms to test for finding out the model maximum efficiency, there are …

Logistic regression Algorithm

Logistic regression is a classification algorithm used to predict the probability of the target variable. The target variable 'satisfaction' has its binary response, such as 0 is for 'neutral or dissatisfied' and 1 is for 'satisfied'. The logistic regression form allows me to model a relationship between multiple predicter variables and a target variable. In mathematics, a logistic regression model predicts P(y) as a function of X, where 0.5 is the threshold. The logistic regression equation is $P(y_i)=\frac{1}{1+e^{-(\beta_0+\beta_i X_i)}}$. (Jafari A, 2022)

Naïve Bayes Algorithm

Naïve Bayes is a classification algorithm used to predict the probability of target variable given a set of features. Also, the fundamental Naïve Bayes assumption is that each feature makes 1) independence, 2) the notion of conditional probability, and 3) Bayesian inference. The Naïve Bayes probability equation is $P(C|X) = \frac{[P(x_1|C)*P(x_2|C)...P(x_n|C)]*P(C)}{P(X)}$. (Jafari A, 2022)

Decision Tree Algorithm

Decision Tree is a classification tree defined as a structural mapping of binary decisions that lead to a decision about the class of an object. In the decision tree algorithm for this dataset, a variable can be chosen at each step that best splits the set of items for building the decision tree from a root node that goes from top to down. To determine the best split condition used an impurity measurement, entropy is one of the impurity measurements that can help reduce uncertainty in the dataset. Also, entropy mostly used on categorical dataset. (Jafari A, 2022)

Random Forest Algorithm

Random Forest is a supervised learning algorithm that mostly used on classification and regression problems. It can combine the output of multiple decision trees to reach a single result. The random forest algorithm with a bagging ensemble method has four steps. At first step, numbers of random variables are taken from the dataset having k number of variables. At second step, individual decision tree models are constructed for each sample. At third step, each model will conduct their own output. At the final step, final output is based on majority vote for a classification dataset (Sruthi E. R., 2021, Jafari A, 2022).

In addition, feature importance associated with random forest is utilized in the dataset as it is the ability to evaluate feature importance in which variables are important to training the model.

K-Nearest Neighbors Algorithm

K-Nearest Neighbors is a classification algorithm that used to find K is nearest closed to datapoint for predicting the class value for the new datapoint. For choosing the K value, k =5 as this odd k value is common to be used in the larger training set. (Jafari A, 2022)

Experimental setup

The dataset originally contained 129,880 observations and 25 variables. In order to basically understand the dataset, the following Python codes were used to derive the insights:

```
df_train = pd.read_csv('/Users/daqian.dang/Desktop/DATS 6013/Final
Project/train.csv')
df_test = pd.read_csv('/Users/daqian.dang/Desktop/DATS 6013/Final
Project/test.csv')
df = pd.concat([df_train,df_test])
df.head(10)
df.shape[0]
df.shape[1]
df.info()
df.describe(include='all')
df.columns
```

In order for analyzing the dataset, it is an important step to clean missing values in the dataset. The feature *'Arrival Delay in Minutes'* has a total of 393 empty spaces as they are missing values. I used NumPy to replace empty spaces to NaN and then used median imputation to convert and fill the missing values since the feature *'Arrival Delay in Minutes'* is extremely skewed.

```
df.isnull().sum()
df.replace('', np.NaN, inplace=True)
median_imputer = SimpleImputer(missing_values=np.NaN,
strategy='median')
df[numeric_var] = median_imputer.fit_transform(df[numeric_var])
```

Outlier detection using interquartile algorithm is also another step to remove unnecessary outliers, which will affect results of data analysis. The following Python codes were used to detect and remove outliers in the dataset:

```
def drop_outliers(df_1,var_1):
    iqr = 1.5 * (np.percentile(df_1[var_1],75) -
np.percentile(df_1[var_1],25))
    df_1.drop(df_1[df_1[var_1] > (iqr +
np.percentile(df_1[var_1],75))].index, inplace=True)
    df_1.drop(df_1[df_1[var_1] < (np.percentile(df_1[var_1],25) -
iqr)].index, inplace=True)
drop_outliers(df,'var')
```

In order to understand the correlation between these variables in the dataset, it is important to check out the correlation coefficients, which it can describe the strength and direction of an association between variables. The following Python codes were used to find correlation and plot correlation heatmap:

```
df.corr()
plt.figure(figsize=(20, 15))
sns.heatmap(df.corr(),annot=True, cmap='Oranges_r')
plt.show()
```

Also, the dataset needs to be checked out if this is balanced or imbalanced, I replaced 0 as ‘neutral or dissatisfied’ and 1 as ‘satisfied’ in the target variable ‘*satisfaction*’. And then, the binary responses were counted to compare the numbers.

```
df['satisfaction'].replace({'neutral or dissatisfied': 0, 'satisfied':
1}, inplace=True)
df.satisfaction.value_counts(normalize=True).plot(kind='bar',
```

Before splitting the dataset into train and test, I encoded the data features except the target variable ‘satisfaction’ using LaberEncoder():

```
categ_cols = ['Gender', 'Customer Type', 'Type of Travel',
        'Class', 'Inflight wifi service', 'Departure/Arrival time
         convenient', 'Ease of Online booking', 'Gate location', 'Food
         and drink', 'Online boarding', 'Seat comfort',
         'Inflight entertainment', 'On-board service', 'Leg room
         service', 'Baggage handling', 'Checkin service', 'Inflight
                service', 'Cleanliness']
df[categ_cols]= df[categ_cols].apply(LabelEncoder().fit_transform)
X = df.values[:,:-1]
```

and then encoded the target variable with sklearn’s LabelEncoder and fit and transform the target variable:

```
Y_data = df.values[:, -1]
class_le = LabelEncoder()
y = class_le.fit_transform(Y_data)
```

The dataset was split into the train as 80% and the test as 20%.

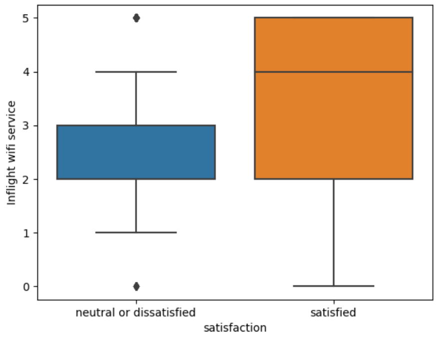
```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=0)
```

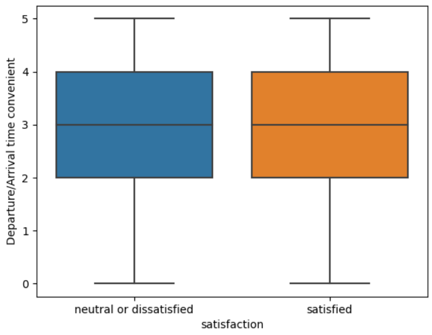
Exploratory Data Analysis Results and Visualization

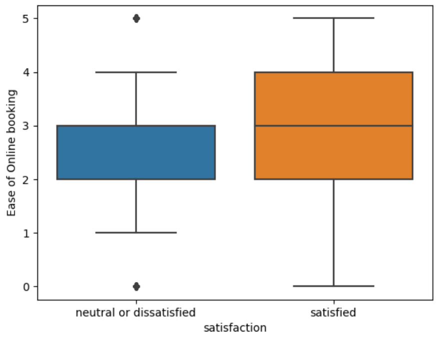
Boxplots

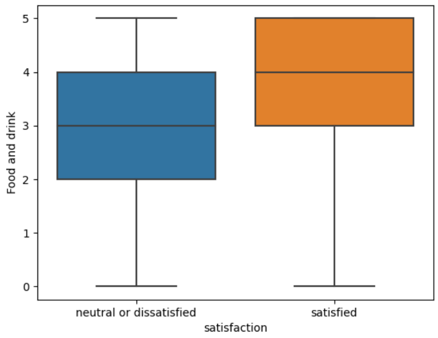
Boxplot is an excellent way to visualize differences among variable groups. Blow are boxplot figures showing different variables with 5-point rating scales based on passengers’ satisfaction. These boxplots help me distinguish significant differences between variable groups.

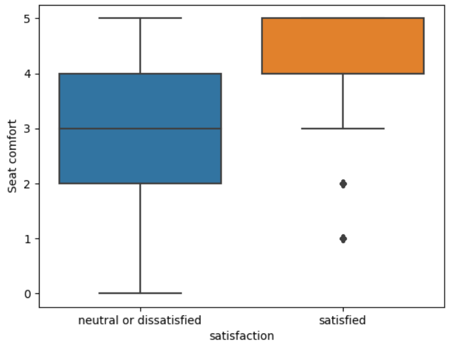
For example, inflight Wi-Fi service, online boarding, inflight entertainment, inflight service, and seat comfort show significantly difference on passenger's satisfaction.

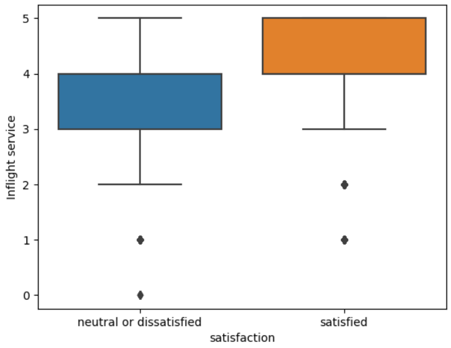


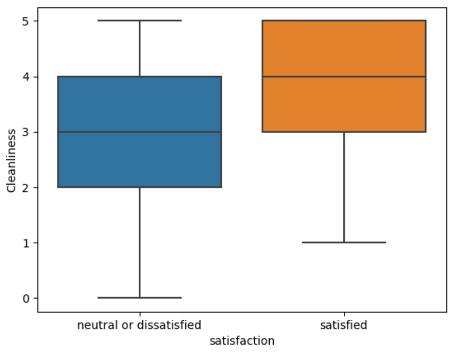


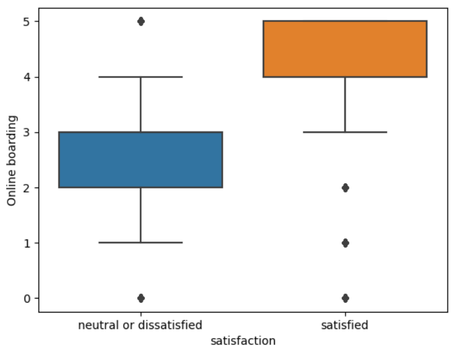


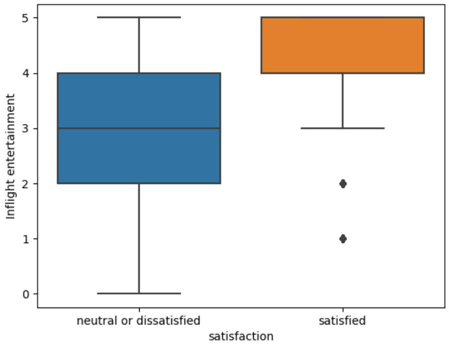






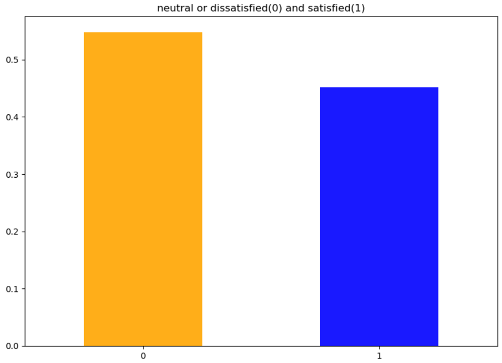






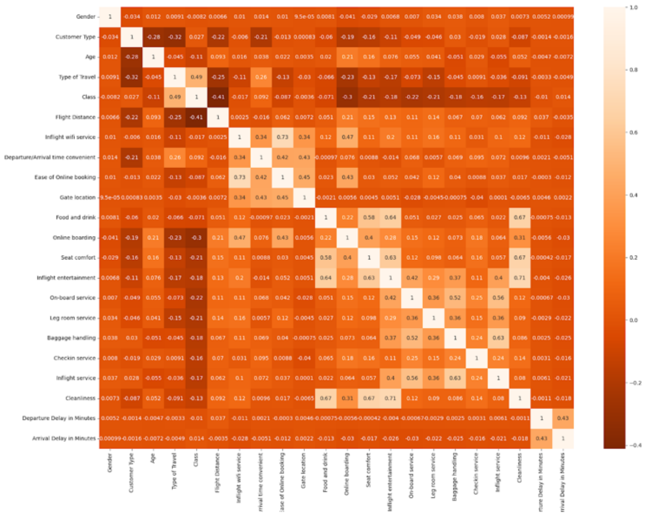
Dataset Balance

It is important to check if the dataset is balanced or imbalanced. As the figure shows below, the dataset is balanced well as neutral or dissatisfied passengers are slightly more than 50% and satisfied passengers are between 40% to 50%. Thus, this does not require to balance the dataset.



Correlation heatmap

Based on the correlation heatmap displays below, it seemed that '*inflight Wi-Fi service'* has positively correlated with '*Ease of Online booking'*. Additionally, '*Food and drink'* is positively dependent on '*Cleanliness*', '*Inflight entertainment*', and '*Seat comfort'.* Also, *'Seat comfort'* is dependent on '*Cleanliness*', '*Inflight entertainment*'. From observing the correlation heatmap, it could help me understand that cleanliness with food and drink, seat comfort, and inflight entertainment that can impact on passengers' overall satisfaction rating during a flight trip.



Model Analysis Protocol

The following model protocol was applied to each of the models:

1) A logistic regression was fit with the selected features and with 80% of training dataset and 20% of testing dataset.
2) A naïve bayes was fit with the dataset split as 80% of train and 20% of test.
3) A decision tree was fit with the Entropy criterion to get view of the data structure.
4) A random forest was fit with all features and K features and with 100 estimators. Also, feature importance was applied to find random forest feature ranking.
5) A K-Nearest Neighbor was fit with 5 of n_neighbors. And StandardScaler was used to scale both train and test.

Model Analysis Results

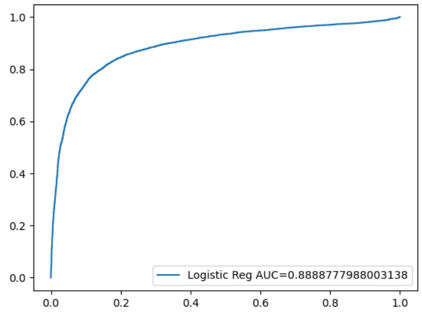
Logistic Regression Model Result

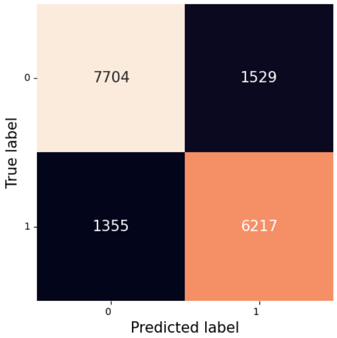
As the tables and figures for logistic regression show below, the percentage of correct prediction is 83%. This tells me that the 16,805 observations used in the model, the model correctly predicted whether or not passengers satisfied 83% of the time. F1-socre shows 81%, which means both precision and recall have higher values. The ROC_AUC and AUC curve shows 89%, which there is a high chance that the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.

Classification Report for Logistic Regression Model:

	precision	recall	f1-score	support
0	0.85	0.83	0.84	9233
1	0.80	0.83	0.81	7572
accuracy			0.83	16805
macro avg	0.83	0.83	0.83	16805
weighted avg	0.83	0.83	0.83	16805

Accuracy	82.84
F1-score	81.17
ROC_AUC	88.89





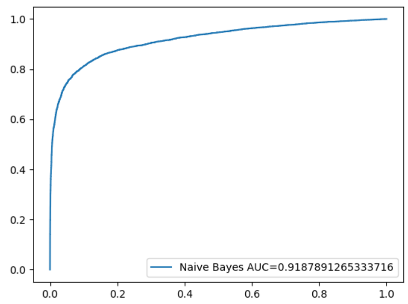
Naïve Bayes Model Result

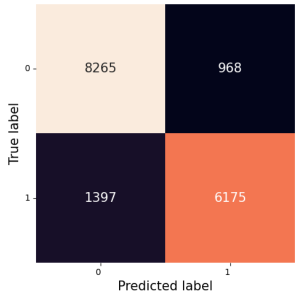
As the tables and figures for Naïve Bayes model show below, the percentage of correct prediction is 84%. The F1 score displays 84%. The ROC_AUC and AUC curve shows 91%.

Classification Report for Naïve Bayes Model:

	precision	recall	f1-score	support
0	0.86	0.90	0.87	9233
1	0.86	0.82	0.84	7572
accuracy			0.86	16805
macro avg	0.86	0.86	0.86	16805
weighted avg	0.86	0.86	0.86	16805

Accuracy	85.93
F1-score	83.93
ROC_AUC	91.88





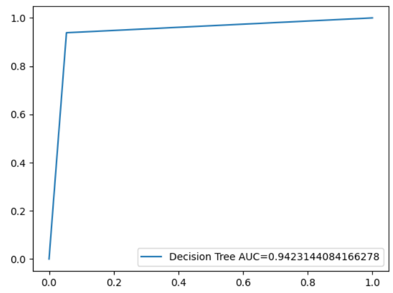
Decision Tree Mode Result

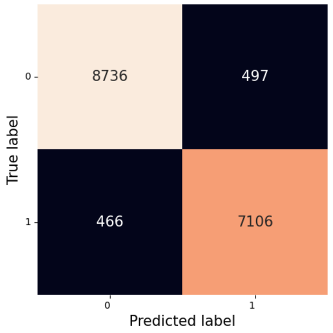
As the tables and figures for Decision Tree with Entropy model show below, the percentage of correct prediction is 94%. F1 score is 93.65. The ROC_AUC and AUC curve shows 94% as well. This model is better than Naïve Bayes model.

Classification Report for Decision Tree with Entropy Model:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	9233
1	0.94	0.94	0.94	7572
accuracy			0.94	16805
macro avg	0.94	0.94	0.94	16805
weighted avg	0.94	0.94	0.94	16805

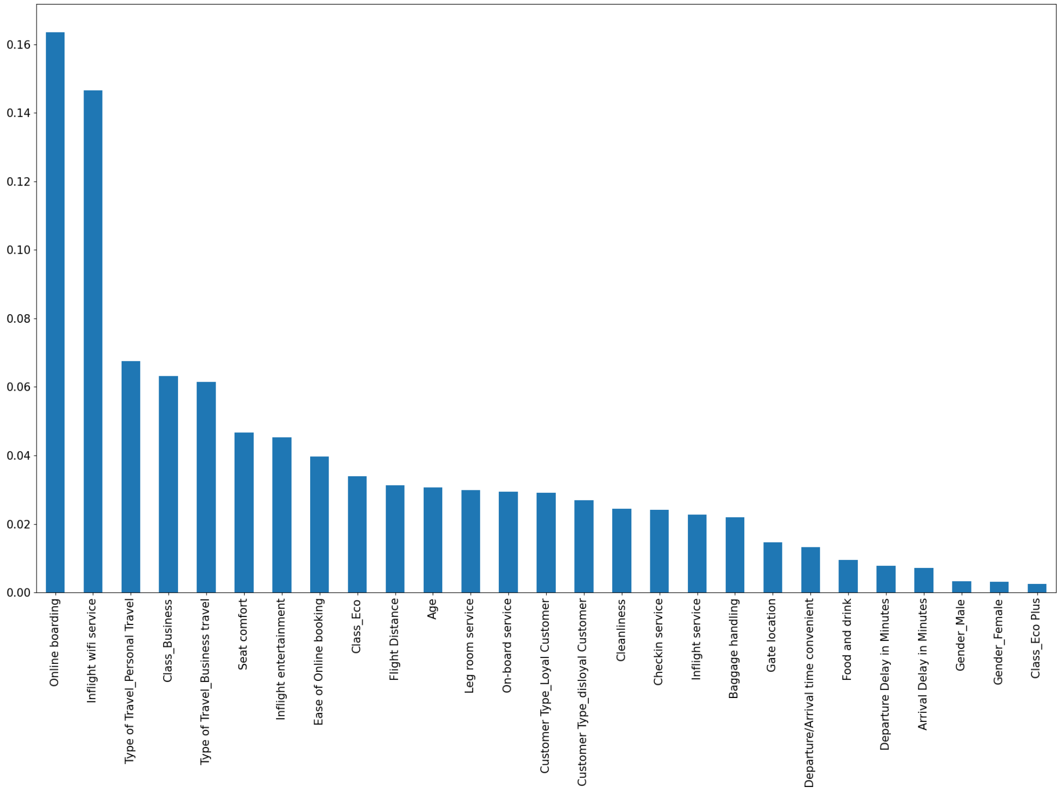
Accuracy	94.27
F1-score	93.65
ROC_AUC	94.23





Random Forest Model Result

Before building the Random Forest model, feature importance is a nice method to figure out which features are important to test the dataset. As the feature importance figure shows below, this clearly tell me that which features are from most important to least important. We can see that ‘*online boarding’* and ‘*Inflight Wi-Fi service’* are the most important features, while ‘*Gender*’ and ‘*Class_Eco Plus’* are the least important features.

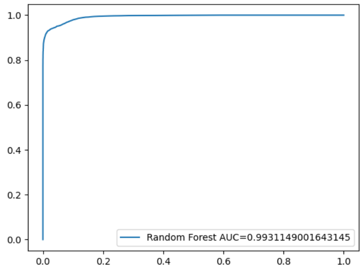


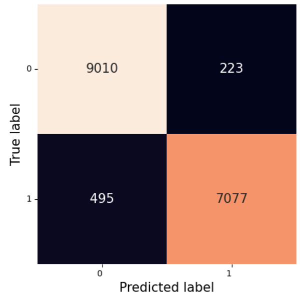
As the tables and figures for Random Forest model using k features show below, the percentage of correct prediction is 96%. The percentage of F1 score is 95%. The ROC_AUC and AUC curve shows 99%. This model seems to be the best of all model comparisons.

Classification Report for Random Forest Model:

	precision	recall	f1-score	support
0	0.95	0.98	0.96	9233
1	0.97	0.94	0.95	7572
accuracy			0.96	16805
macro avg	0.96	0.96	0.96	16805
weighted avg	0.96	0.96	0.96	16805

Accuracy	95.77
F1-score	95.21
ROC_AUC	99.34



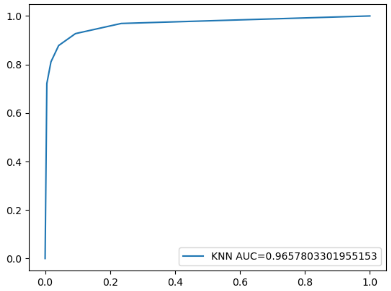


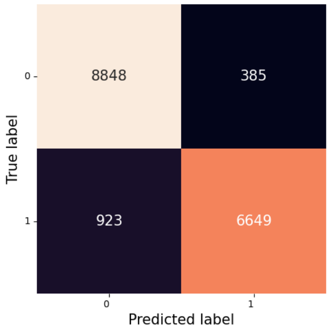
KNN Model Result

After scaling the dataset for KNN model, the tables and figures for KNN model show below, the percentage of accuracy is 92%. The F1 score shows 91%. The ROC_AUC score displays 96%.

Classification Report for KNN Model:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	9233
1	0.94	0.87	0.91	7572
accuracy			0.92	16805
macro avg	0.92	0.91	0.92	16805
weighted avg	0.92	0.92	0.92	16805





Summary and conclusions

As the model comparison table shows below, random forest model has the highest accuracy score 95%, F1 score 94%, and ROC_AUC score 99% when compared with other models. In result, random forest model preforms the best on classification accuracy on the dataset.

Model	**Accuracy**	**F1 score**	**ROC_AUC**
Logistic Regression	82.84	81.17	88.89
Naïve Bayes	85.93	83.93	91.88
Decision Tree with Entropy	94.27	93.65	94.23
Random Forest	95.78	95.21	99.34

KNN	92.22	91.04	96.58

What I learned from this project is data preprocessing, which is important part of the whole steps for doing this project. Without doing properly data preprocessing, it is impossible to make it easy to analyze and interpret data. The process includes cleaning or imputing missing values and removing outliers, which can otherwise negatively affect model's accuracy. However, it is important to notice that data preprocessing could remove many of potential problems that can lead to erroneous assumption while training model with the resulting dataset.

The signification improvement in this project I would like is the exploratory data analysis and data visualization. Because EDA can help people understand the basic data analysis, and data visualization can help visualize and interpret data to communicate information in effective and understandable manner. Also, data visualization positively impacts on a company or an organization's important decision-making process with visual representations of data.

References

Geeksforgeeks.org. (2020, June 3rd). *Interquartile Ranger to Detect Outliers in Data.*
https://www.geeksforgeeks.org/interquartile-range-to-detect-outliers-in-data/

Sruthi E.R., (2021, June 17th). *Understanding Random Forest*
https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Jafari. A., (2022, June). *Introduction to Data Mining, DATS 6103.*