Predicting Passenger Satisfactory

Using 3 different methods Jiankun Dong

# Abstract

This project uses a dataset that contains information about airline passengers and their satisfactory level. All passengers are either ‘satisfied’ or ‘Neutral or unsatisfied’. I use logistic regression, naive Bayes and random forest methods to predict the passenger’s satisfactory level, and each method is then evaluated based on their accuracy and confusion matrix.

# Dataset

Link to dataset: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/>

The dataset contains null values in the ‘Arrival Delay in Minutes’ column, and the corresponding rows are discarded.

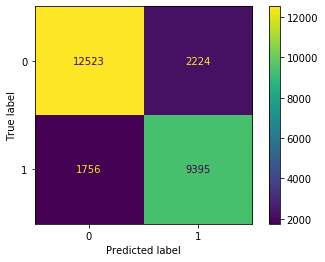
The dataset was split between test.csv and train.csv. However to avoid overfitting, I merged the 2 dataframes and randomly selected a training and testing set. The split is 4:1.

The data is only scaled for the logistic regression model using MinMaxScaler (because it’s not normally distributed) There are only 4 continuous variables: age, flight distance, departure delay in minutes and arrival delay in minutes. The scaler should only be applied to those columns. Then, I created dummy variables for the categorical variables following one hot encoding.

# Naive Bayes

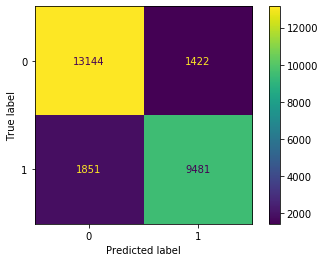
The naive Bayes has an accuracy of 84.632%. This accuracy slightly varies each run based on the random state of the train test dataset, but it’s overall between 84-85%

Confusion matrix:

 TPR: 0.8425 TNR: 0.8492

# Logistic Regression

Logistic regression yields a 87.36% accuracy rate (again, changing each time due to random train test split), with TPR 0.8367 and TNR 0.9



# Random Forest

For the random forest, I found that D=5 and N=10 yields the best result. The plot of error rate vs N for each D is on the left below. The confusion matrix of (D=5,N=10) is on the right.

Accuracy is about 92.83%, TPR is about 0.91 and TNR is 0.94. Note that TP is 10238.

# 

# 

# Conclusion

| Model | TP | FP | TN | FN | Accuracy% | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Naive Bayes | 9395 | 2224 | 12523 | 1756 | 84.632 | 0.8425 | 0.8492 |
| Logistic Regression | 9481 | 1422 | 13144 | 1851 | 87.36 | 0.8367 | 0.9 |
| Random Forest | 10238 | 866 | 13793 | 1001 | 92.83 | 0.91 | 0.94 |

From this we can see that random forest has the best accuracy, TPR and TNR result, where we can be 92.83% accurate of our prediction.