Abstract

Team Name: Team Sandwich

Taukir Chowdhury	**Shahriar Tarvir**	**Tauseef Tajwar**
Cogni ID: cogni2003675	**Cogni ID:** cogni2003661	**Cogni ID:** cogni2003662
Institution: Islamic University of Technology, Dhaka, Bangladesh	**Institution:** Islamic University of Technology, Dhaka, Bangladesh	**Institution:** Islamic University of Technology, Dhaka, Bangladesh
Email: taukir.azam16@gmail.com	**Email:** shahriartarvir98@gmail.com	**Email:** tauseef.tajwar36@gmail.com
Contact Number: +8801554356853	**Contact Number:** +8801716456332	**Contact Number:** +8801728569161

The case is about Invistico Airlines which required us to analyse a dataset on their customer base. The dataset containing 129,880 rows and 23 columns was provided in CSV format. All the analysis on this case was done in Jupyter Notebooks using Python. Our approach to solving this case can be divided into 3 separate sections.

Section 1 - Data Cleaning: We started the data cleaning process by first checking for categorical columns and null values in the dataset. The dataset had 393 null values in the 'Arrival Time in Minutes' column and 5 columns with categorical values that needed to be converted to numerical values. One possible way to deal with the null values was to delete all of the 393 rows from the dataset. But that would mean loss of information which is never desirable. So, we calculated the correlation between the features of the dataset and found that 'Arrival Delay in Minutes' also had a 96.5% correlation with another column. This referred to the fact that 'Arrival Delay in Minutes' was not providing us with any new information. So, we could remove the entire column of 'Arrival Time in Minutes' thus removing the null values without the loss of any information. Next we converted the categorical values into numerical values using ScikitLearn's label encoder. The data was then split into training and testing sets with a 85:15 split. The distributions of the features were plotted and found that the features were neither scaled nor normalized. Therefore, we then converted the data into standard normal form (mean = 0, standard deviation = 1) and subsequently scaled the features.

Section 2 - The Models: At first we created a baseline model that only predicts "dissatisfied" for all examples. The accuracy of the baseline model was 45.2%. We selected 7 different classifier models (Gradient Boosting, Logistic Regression, SVM, KNN, Decision Tree, MLP, and Random Forest) and compared them on 6 different metrics (accuracy, precision score, recall score, specificity, F1 score, and confusion matrix). The best performing model among the 7 was the Random Forest Classifier with an accuracy of 96.2%. Logistic Regression was comparatively the least performing model with 83.6% accuracy. One key point to note is that most of the models performed noticeably worse on an uncleaned dataset (accuracy on uncleaned dataset for Logistic Regression - 79%, KNN - 72%, SVM - 61%, MLP - 88%).

	Accuracy	Precision	Recall	Specificity	F1 Score	Training Time
Random Forest	96.21%	97.52%	95.56%	97.02%	96.52%	11.7s
KNN	92.71%	95.02%	91.54%	94.13%	93.25%	0.015s
MLP	95.79%	95.70%	96.18%	94.72%	95.94%	1m 03s
SVM	95.01%	95.70%	95.22%	94.76%	95.46%	4m 08s
Logistic Regression	83.66%	85.17%	85.14%	81.86%	85.16%	0.22s
Gradient Boosting	92.63%	93.13%	93.52%	91.55%	93.32%	22.25s
Decision Tree	94.27%	94.55%	95.06%	93.30%	94.80%	0.73s

*Baseline Accuracy: 45.21%

Fig: Model performance measures

Section 3 - Feature Importance: The third and final section of our methodology was to find the impact of each feature on the satisfaction of the customer. We calculated the feature impact of the dataset using two of our best performing models (Random Forest and Decision Tree) and found that two of the most important features contributing to customer satisfaction are 'Inflight Entertainment' and 'Seat Comfort'. The feature that has the least effect on customer satisfaction the most is 'Inflight WiFi Service'. From this we can conclude that Invistico Airlines has to maintain the quality of the top impacting features to maximize customer satisfaction.

Additional Contents:

Slides:
https://drive.google.com/file/d/1aLVZA63BMfm-hmYZ9an5w5qYTIXiKqOt/view?usp=sharing

Code:
https://drive.google.com/file/d/1qJ19nspnMA9zn61hdwmNtcfHY8T65MaM/view?usp=sharing

If required, download the dataset from this link:
https://drive.google.com/file/d/1jnaVVlh2jNssuzXCjAqGcZrGaE10oZMQ/view?usp=sharing