**FLIGHT CANCELLATION PREDICTION**

Taduka Hemanth Krishna

**Introduction:**

We are a group of three people who have chosen to work on the airline's cancelation dataset for our final project. In this document, we will report our dataset results, the problems we would like to address, and the methodology we would use to create ML models for prediction as part of this project.

Firstly, we have performed the exploratory data analysis based on our business questions. And then, we performed the data cleaning as per the requirements and performed the five predive models on classifying the cancelation of the fights.

We have performed this project on the Azure machine learning studios notebooks.

**Predictive models used in the project:**

Random Forest Classification  
Gradient Boosting Classification  
XGB Classification  
GaussianNB Classification  
KNeighbors Classification

**Dataset:**

The on-time performance of domestic flights run by major airlines is tracked by the U.S. Transportation Department (DOT) and Bureau of Transportation Statistics (BTS). In addition to summary tables, the DOT's monthly Air Travel Consumer Report includes a summary of the number of flights that were on time, delayed, canceled, and diverted. BTS started gathering information on the reasons for aircraft delays in June 2003.

**Source :** <https://www.kaggle.com/datasets/giovamata/airlinedelaycauses>.

**Dataset Name:** DelayedFlights.csv

A picture containing text

Description automatically generated

First, we need to get the API auth token from Kaggle to start with because the datasets in Kaggle cannot be exported without Kaggle's authorization.

We use the commands

Export Kaggle\_User = XXXXXXXXXXXXX

Export Kaggle\_Key = XXXXXXXXXXXXX

Then we can import the dataset's zip file to work with. After extracting the zip file, we have the CSV file to get started.

Table

Description automatically generated

Here, we have a dataset with 1953912 rows and 30 columns, of which we can drop the **Unnamed:0 column** as it has no significant value to work with. We start with our data frame information to check the data types of the columns. As we can see below, we have many numerical columns, and out target, variable is ‘Cancelled’, which is a categorical variable with outputs 0 or 1 mentioning Cancelled or Not Cancelled.

Graphical user interface, text, application, email

Description automatically generated

**Business questions:**

1. Frequency of the Flights being canceled on days, months.
2. Is departure time have an impact on flight cancelations?
3. Is the day of the month having an impact on flight cancelations?
4. Is flight distance having an impact on flight cancelations?

**Data Pre-processing:**

We started with the most basic of steps, cleaning the data of null values. The null values in the data frame were too many and replacing them with their statistical counterparts would mean that the model would get a little troubled because considering the variables like CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay their significance in the model would be disturbed by replacing them with their mean or median, a small set of replacements would prove to be good but a change of 689270 rows would not be a great fit to the model.

Graphical user interface, application

Description automatically generated

Now, we can define the target variable and check the features that would improve the model but wouldn’t perform or improve the model. We choose to remove those leaky features. We have named a list called leaky\_features to pull all the features that are irrelevant but with a seemingly good correlation score. We have considered removing these features by calculating the VIF scores and removing them.

Text

Description automatically generated

Then we create a new df with the name Features that has all the numerical columns with good correlations with the target variation could be used in the model we are trying to build.

Table

Description automatically generated

We found the below 15 features, which are relevant to our target variable cancellation.

Graphical user interface, application

Description automatically generated

We wanted to segregate the data based on the type of the data. So we have

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

We have taken the data frame with numerical values to answer each of our business questions because we need to subset the columns that are only required to answer these questions.

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**Exploratory Data Analysis:**

EDA is an essential step to understanding the variables and how one influences the other. The features selected are to be used to answer the business questions; we have considered the Flights canceled and not canceled separately because we need to see what causes the delay and which days and months the delays are happening frequently. All this analysis can help us see the patterns and trends that are causing the delays.

Also, we have normalized many columns used in the Data visualization as a part of data preprocessing. Our business questions require us to consider times of flights, dates, and months of December, November, and October.

A picture containing text, clipart

Description automatically generated

The above visualization shows the percentage of flights canceled every week. More than fifty-five percent of flights were canceled on Tuesday, and the least percentage of flights were canceled on Thursday.

Chart, line chart

Description automatically generated

The above visualization shows the delays in the months for canceled and all flights. The most common reasons we have identified are the Late Aircraft Delay, Carrier Delay and NAS Delay (Air traffic). Weather delay can be seen mostly in the last three months of the year in the Northeastern parts like JFK and BOS. There were no security Delays in that year. Here, NAS Delay and Carrier delay are controllable and are human intervened delays, they include maintenance, stowing problems, flight load checkup, fueling of flights. These categories both are multicollinear with a good correlation.

Chart, bar chart

Description automatically generated

The above visualization shows all canceled flights on the last 3 months where most of the cancellations happened. We can see from the bar graph that the highest number of flights were canceled in December because of the Weather and the Carrier Delays. Here, NAS and carrier delays are both codependents, we need to consider either one of them for the model to predict, We have considered the variable ArrDelay which has all the Delay times combined in it. Just considering this would help us in our model.

Chart, bar chart

Description automatically generated

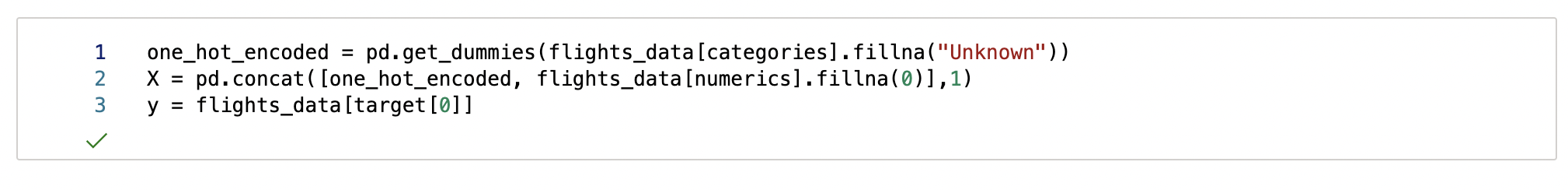
The above visualization shows the Average Delay by Origin Airport whereas we speculated before, the northeastern Airports are mainly affected by NAS Delay. This NAS delay is air traffic that is high for JFK and BOS because of the inflow and outflow of the international flights. If we look at the causes for the delay, there seems to be many reasons that contribute to the Cancellation and mostly they are form NAS or Carrier.

Chart, scatter chart

Description automatically generated

The above visualization shows flight distance in miles for canceled and all flights. We can see a pattern that as the distance increases, the number of flights cancelled decreases. We can also see that for the flight distance, this inverse correlation can be seen here where shorter flights are the ones with most of the delays on arrival.

There are seven categorical variables in the selected features for predictive modeling. So, to proceed future, we have encoded the categorical data using one hot encoding. Once the categorical data is encoded, we can use them in the predictive models.



A picture containing qr code

Description automatically generated

We can see from the below correlation matrix that the cancellation has a high correlation with the month.

**Predictive Models:**

For training and evaluating the models, we have used the cross-validation technique. We implemented the cross-validation by employing the complementary subset of the data set to test our model after it had been trained using the subset of the original data set.

Graphical user interface, text, application

Description automatically generated

Here, we have built a few predictive models for classifying the Cancelled Flights and non-Cancelled Flights. As we had all numerical independent variables and a categorical target variable, we need to consider only classification models for the predictions. We have used the Random Forest classifier, XGB classifier, Gradient Boosting classifier, Naïve Bayes classifier, and K nearest classifier to predict the cancellation of the flights we have in the test section.

Below, we implemented a function to tabulate the classifier models that were thought in the first steps. This function is easy to implement and can show the accuracy and the processing time of the models.

Text

Description automatically generated

Table

Description automatically generated

As we can see here, the Models have performed well with an accuracy of 88-90% range for the Random Forest classifier. We checked for outliers by comparing the mean with the maximum values, and there wasn’t much difference, and nothing seemed odd in the boxplots that indicated many outliers. The processing time of the models is shown above. They performed well and fast with Azure 28GB cloud CPU.

**Conclusion:**

After considering all the analysis, we found the month of the year has a high impact on flight cancelations. Among all the five performed predictive models, Random Forest Classifier got the highest mean accuracy of 89.81 percent compared to the other models. So, Random Forest Classifier is the most performed model among all the classification models.

# References:

*airlinedelaycauses*. (n.d.). Retrieved from www.kaggle.com: https://www.kaggle.com/datasets/giovamata/airlinedelaycauses

*Beginner’s Guide to XGBoost for Classification Problems*. (n.d.). Retrieved from towardsdatascience.com: https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

*Cross Validation in Machine Learning*. (n.d.). Retrieved from www.geeksforgeeks.org: https://www.geeksforgeeks.org/cross-validation-machine-learning/

*Gradient Boosting In Classification: Not a Black Box Anymore!* (n.d.). Retrieved from blog.paperspace.com: https://blog.paperspace.com/gradient-boosting-for-classification/

*sklearn.naive\_bayes.GaussianNB*. (n.d.). Retrieved from scikit-learn.org: https://scikit-learn.org/stable/modules/generated/sklearn.naive\_bayes.GaussianNB.html

*Understanding Random Forest*. (n.d.). Retrieved from towardsdatascience.com: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

*What is Exploratory Data Analysis?* (n.d.). Retrieved from towardsdatascience.com: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15