FLIGHT DELAY PREDICTION

(SLDM Mini Project)

Ishita Deodhar - 21060641020

Karnali Naik - 21060641023

Sanika Mhamunkar - 21060641042

Sayali Bora - 21060641045

Contents

Executive summary

The airline industry plays a crucial role in the world's transportation sector and a lot of businesses rely on various airlines to connect them with other parts of the world. However, variables like severe weather, air traffic, etc., may have a direct impact on airline services by causing flight delays. Every year approximately 20% of airline flights are canceled or delayed, costing passengers more than 20 billion dollars in money and their time. In order to address this problem, precise forecasting of these aircraft delays helps airlines to respond to likely causes of the delays in advance to lessen the negative effects and lets passengers be well prepared for the interruption caused to their journey.

The purpose of this project is to look at the approaches used to build models to correctly classify cancelled and delayed flights and predicting flight delays taking various factors into account while suggesting suitable recommendations for the same.

Introduction

Passenger airlines, cargo airlines, and air traffic control systems are the main elements of any transportation system in the modern world. Nations all around the world have attempted to develop different methods over time to enhance the airplane transportation system. This has significantly altered how airlines operate. Modern travelers occasionally experience difficulty from flight delays.

In the first part of the project, we try to classify which flights are on time, delayed or cancelled. The models used for classification were Decision Trees algorithm, Multinomial Naive Bayes Classifier and Multinomial Logistic Regression. They were compared based on their accuracy.

Further, in the second part, three models namely Decision Trees Regression, XGBoost Regression and KNN Regression were used to predict the delays. Various error measures were computed for comparison of the above fitted models.

Business Problem Statement

Every year cancellation or delayed flights not only causes inconvenience to passengers but a huge loss to stakeholders as well.

The motive of the project is to propose an approach that improves the operational performance without hampering or affecting the planned cost.

The goal is to use exploratory data analysis and to build machine learning models to predict airline departure and arrival delays.

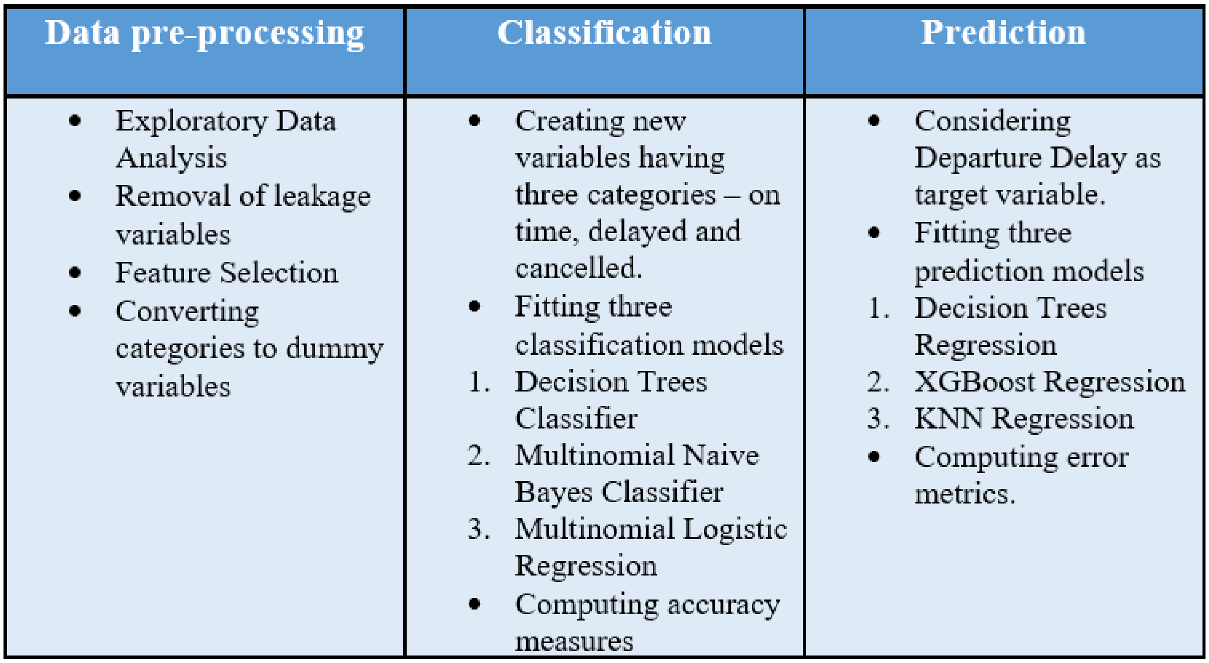
Objectives

1. Predict which flights will be canceled or delayed.
2. Develop a model to predict the delay time of the airlines.
3. Suggest suitable recommendations for stakeholders and customers to lessen the impact of delayed and cancelled flights.

About The Dataset

The data has been extracted from the Marketing Carrier On-Time Performance (Beginning January 2018) data table of the "On-Time" database from the TranStats data library. There were 61 variables with more than 10 lakhs data points. The data included categorical and numerical variables along with identifiers of flights. Some of the important features used in the analysis for cancelled and delayed classification and prediction were - Airline, Origin, Destination, Quarter, departure delay, arrival delay, CRS Elapsed Time, and Day of the week.

Methodology



Data pre-processing	Classification	Prediction
• Exploratory Data Analysis • Removal of leakage variables • Feature Selection • Converting categories to dummy variables	• Creating new variables having three categories – on time, delayed and cancelled. • Fitting three classification models 1. Decision Trees Classifier 2. Multinomial Naive Bayes Classifier 3. Multinomial Logistic Regression • Computing accuracy measures	• Considering Departure Delay as target variable. • Fitting three prediction models 1. Decision Trees Regression 2. XGBoost Regression 3. KNN Regression • Computing error metrics.

Table 1

- **Data preprocessing:**

1. A new variable named Flight Status was formed, which tells if a flight was cancelled, delayed or on time. For calculating this variable, departure delay minutes, arrival delay minutes and cancelled columns were considered.
2. For cancelled flights, the departure delay minutes were imputed as zero.
3. There were 62 values in CRSElapsed time with negative values, so they were removed.
4. The categorical values were converted into dummies.

- **Feature Engineering**

Variable taken into consideration for classification as well as prediction were-

Variables	Description
Airline	Name of the Airline
Cancelled	Cancelled Flight Indicator (1=Yes)
DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
Day of the week	Day of the week
Origin	Origin Airport
Destination	Destination Airport
Quarter	Quarter (1-4)
CRSElapsed Time	CRS Elapsed Time of Flight, in Minutes

- A new variable was introduced, namely “category”.

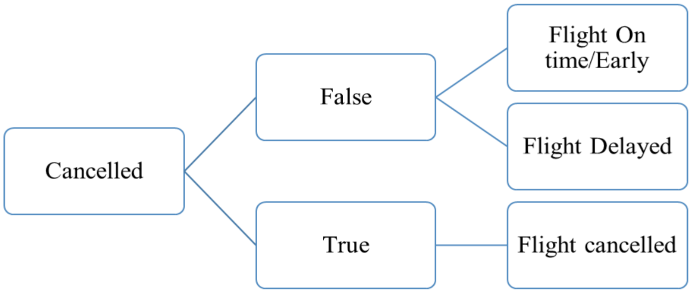


Fig. 1

- **Exploratory Data Analysis:**

Exploratory analysis was performed to observe patterns in data.

- **Classification and prediction Modelling:**

The classification model was built to find if the flight is cancelled or delayed. The prediction model was constructed to predict the delay time. All the leakage variables and identifiers were removed from the data for classification and prediction models. Variables such as Airline, Origin, Destination, CRSTime, CRSElapsedTime, and Day of the week were used as independent variables for modelling.

Classification model:

The dependent variable was Flight Status. Two samples were taken from the data (20% of total data) and three models were fitted using the Decision Trees algorithm, Multinomial Naive Bayes Classifier and Multinomial Logistic Regression. The accuracy of each model was compared.

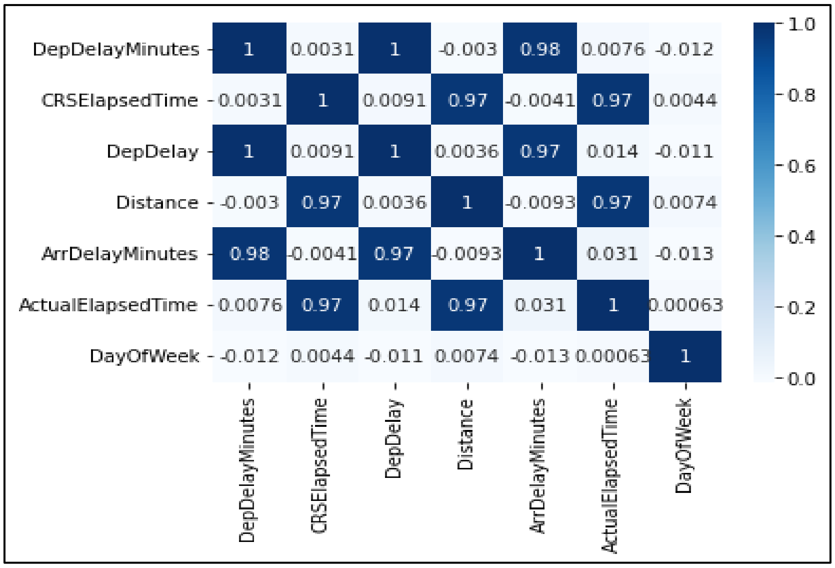
Prediction model:

The dependent variable was departure delay minutes. Two samples were taken from the data (20% of total data) three models were fitted using XGBoost regression, Decision trees regression model and KNN regression model. RMSE, MAE and SMAPE values were computed and compared.

Results

Exploratory Analysis:

- Correlation matrix



	DepDelayMinutes	CRSElapsedTime	DepDelay	Distance	ArrDelayMinutes	ActualElapsedTime	DayOfWeek
DepDelayMinutes	1	0.0031	1	-0.003	0.98	0.0076	-0.012
CRSElapsedTime	0.0031	1	0.0091	0.97	-0.0041	0.97	0.0044
DepDelay	1	0.0091	1	0.0036	0.97	0.014	-0.011
Distance	-0.003	0.97	0.0036	1	-0.0093	0.97	0.0074
ArrDelayMinutes	0.98	-0.0041	0.97	-0.0093	1	0.031	-0.013
ActualElapsedTime	0.0076	0.97	0.014	0.97	0.031	1	0.00063
DayOfWeek	-0.012	0.0044	-0.011	0.0074	-0.013	0.00063	1

Fig. 2

High Correlation was observed between CRS Elapsed Time and Distance. Hence only CRS Elapsed time was used as a predictor in prediction analysis. Departure delay minutes and Arrival delay minutes showed high correlation.

- Proportions of Cancelled flights to total flights for top 10 Airlines

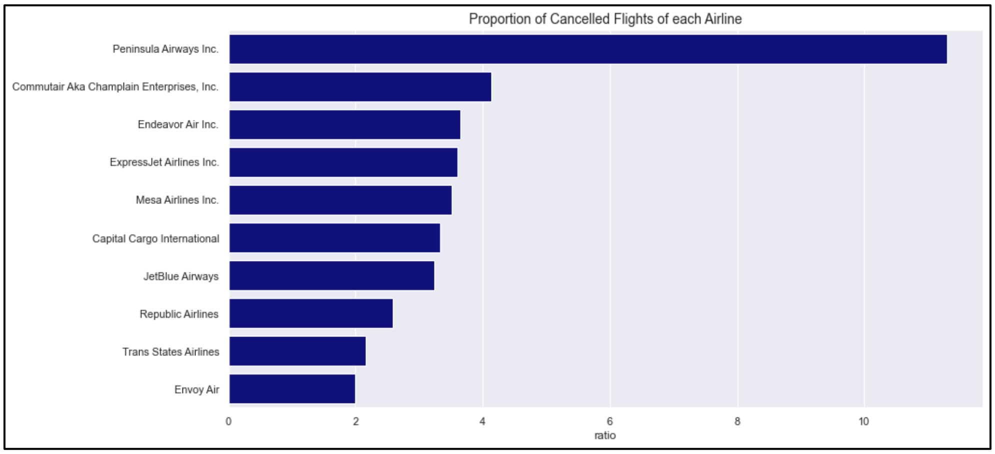


Fig. 3

Peninsula Airways had the most cancelled flights.

- Proportions of Diverted flights to total flights for top 10 Airlines

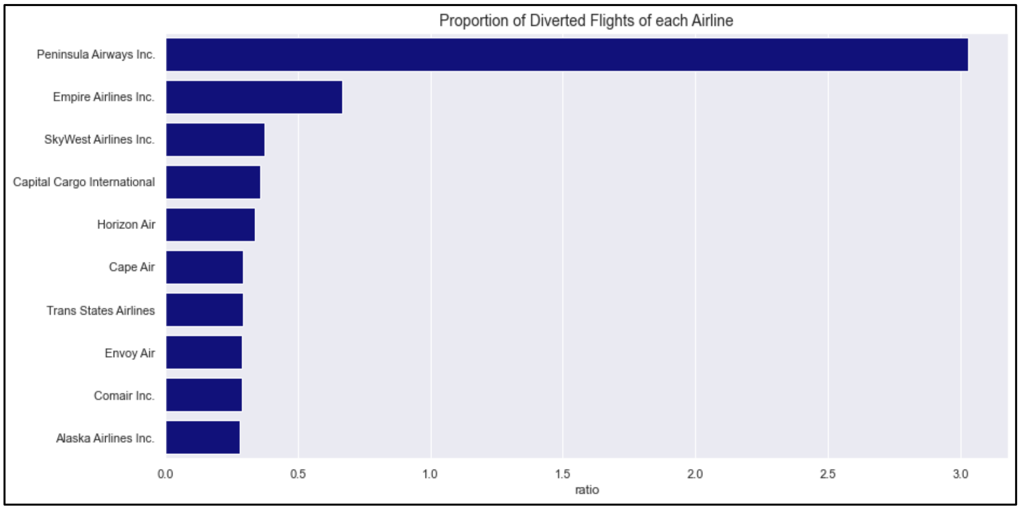


Fig. 4

Peninsula Airlines has the most diverted flight.

- Proportions of cancelled flights in each quarter

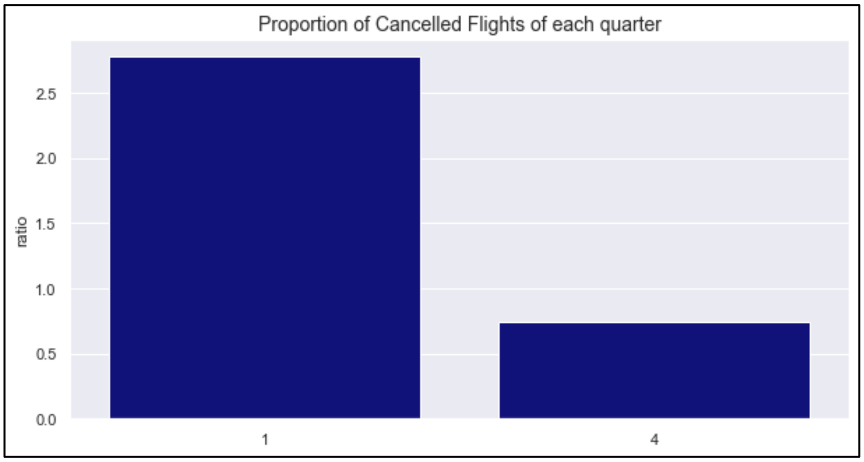


Fig. 5

It is visible that in the 1st quarter around 4% more flights were delayed as compared to 4th quarter.

Proportion of delayed flights to total flights of different Airlines were compared.

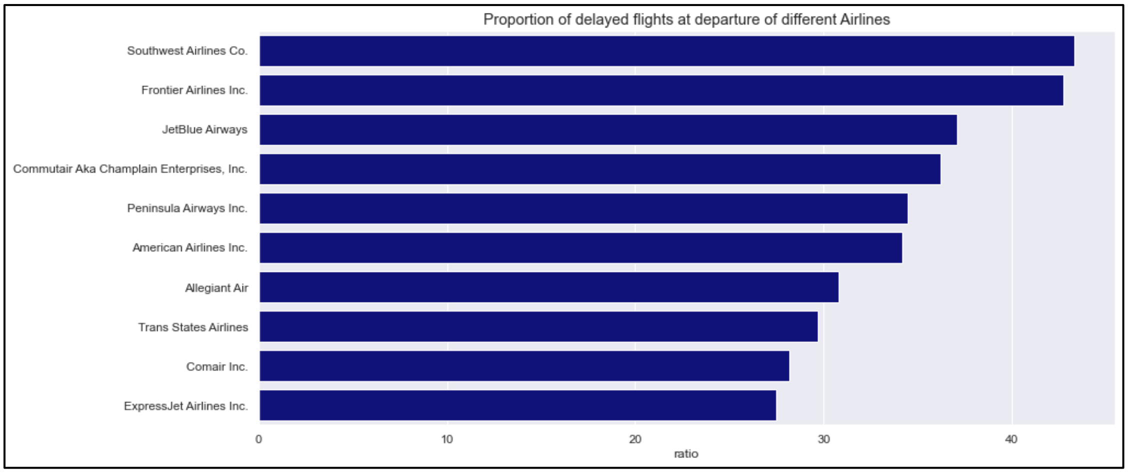
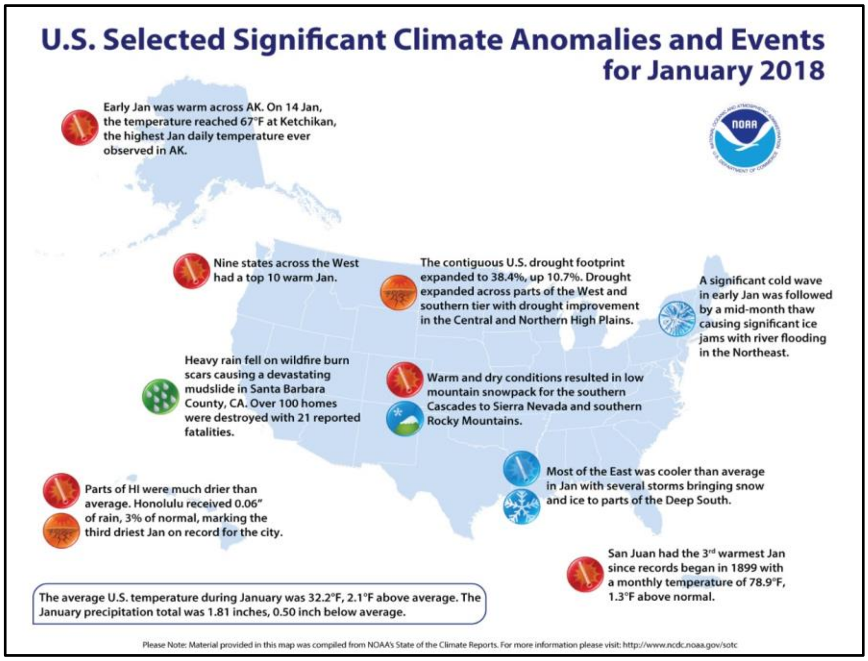


Fig. 6

It was found that more than 40 percent of the flights were delayed by Southwest Airlines Co. and Frontier Airlines Inc.

The data was recorded only for the month of January and October, since the data on weather conditions wasn't available the Climate Anomalies for January and October 2018 were found and are shown below:



U.S. Selected Significant Climate Anomalies and Events for January 2018

Early Jan was warm across AK. On 14 Jan, the temperature reached 67°F at Ketchikan, the highest Jan daily temperature ever observed in AK.

Nine states across the West had a top 10 warm Jan.

The contiguous U.S. drought footprint expanded to 38.4%, up 10.7%. Drought expanded across parts of the West and southern tier with drought improvement in the Central and Northern High Plains.

A significant cold wave in early Jan was followed by a mid-month thaw causing significant ice jams with river flooding in the Northeast.

Heavy rain fell on wildfire burn scars causing a devastating mudslide in Santa Barbara County, CA. Over 100 homes were destroyed with 21 reported fatalities.

Warm and dry conditions resulted in low mountain snowpack for the southern Cascades to Sierra Nevada and southern Rocky Mountains.

Most of the East was cooler than average in Jan with several storms bringing snow and ice to parts of the Deep South.

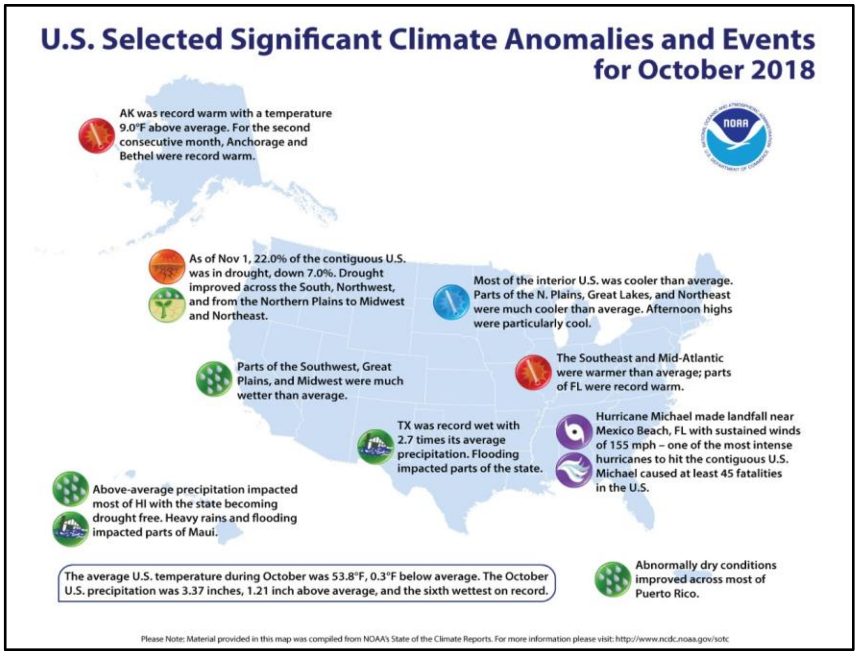
Parts of HI were much drier than average. Honolulu received 0.06" of rain, 3% of normal, marking the third driest Jan on record for the city.

San Juan had the 3rd warmest Jan since records began in 1899 with a monthly temperature of 78.9°F, 1.3°F above normal.

The average U.S. temperature during January was 32.2°F, 2.1°F above average. The January precipitation total was 1.81 inches, 0.50 inch below average.

Please Note: Material provided in this map was compiled from NOAA's State of the Climate Reports. For more information please visit: http://www.ncdc.noaa.gov/sotc

Fig. 7



U.S. Selected Significant Climate Anomalies and Events for October 2018

AK was record warm with a temperature 9.0°F above average. For the second consecutive month, Anchorage and Bethel were record warm.

As of Nov 1, 22.0% of the contiguous U.S. was in drought, down 7.0%. Drought improved across the South, Northwest, and from the Northern Plains to Midwest and Northeast.

Most of the interior U.S. was cooler than average. Parts of the N. Plains, Great Lakes, and Northeast were much cooler than average. Afternoon highs were particularly cool.

Parts of the Southwest, Great Plains, and Midwest were much wetter than average.

The Southeast and Mid-Atlantic were warmer than average; parts of FL were record warm.

TX was record wet with 2.7 times its average precipitation. Flooding impacted parts of the state.

Hurricane Michael made landfall near Mexico Beach, FL with sustained winds of 155 mph – one of the most intense hurricanes to hit the contiguous U.S. Michael caused at least 45 fatalities in the U.S.

Above-average precipitation impacted most of HI with the state becoming drought free. Heavy rains and flooding impacted parts of Maui.

The average U.S. temperature during October was 53.8°F, 0.3°F below average. The October U.S. precipitation was 3.37 inches, 1.21 inch above average, and the sixth wettest on record.

Abnormally dry conditions improved across most of Puerto Rico.

Please Note: Material provided in this map was compiled from NOAA's State of the Climate Reports. For more information please visit: http://www.ncdc.noaa.gov/sotc

Fig. 8

Classification

Two samples were taken and three classification Models were built on these samples. The accuracy of each model was calculated.

	Sample 1	Sample 2
Decision Tree Classifier	0.56	0.56
Multinomial Naive Bayes Classifier	0.66	0.66
Multinomial Logistic Regression	0.66	0.66

Table 2

For both the samples the accuracy was the same.

Prediction

Two Samples were taken and three prediction models were built on these samples. Error metrics were compared for both the samples.

Sample 1

	RMSE	**MAE**	**SMAPE**
XGBoost Regression	43.71	17.24	165.50
Decision Tree Regression	56.75	19.10	91.24
KNN Regression	47.83	18.35	135.79

Table 3

Sample 2

	RMSE	**MAE**	**SMAPE**
XGBoost Regression	40.22	17.02	165.96
Decision Tree Regression	55.65	18.92	90.33
KNN Regression	43.63	17.64	136.37

Table 4

Conclusion

- According to their relative frequencies, Peninsula Airlines has the most cancelled and diverted flights when compared to other airlines.
- It was observed that more flights were cancelled in Quarter 1 as compared to Quarter 4. Since the data provided in quarter 1 was only for January month, it can be concluded that the more flights were cancelled due to weather conditions as December, January or February are the snowiest month for a majority of the U.S.
- More than 40% of the flights by SouthWest Airlines and Co were delayed. Suggestions can be made to the airline company (I don't know what).
- In this project, we use flight data, to predict flight departure delay and find its status. Our result shows that the for classifying if the flight is on time or delayed or cancelled Naive Bayes Classifier as well as Multinomial Logistic Regression can be used since they both had same accuracy of 66% but it can be suggested that Naive Bayes Classifier can be used since it has lesser computational time.
- It was seen that the XGBoost regression method yields the best performance compared to the Decision Tree regression model and KNN regression model.
- However, the delayed flights are only correctly predicted approximately 40% of the time. As a result, there can be additional features related to the causes of flight delay that are not yet discovered using our existing data sources.

Limitations

1. Delay/ cancellations due to weather conditions cannot be predicted or classified as it was not provided.
2. Data is available only for the month of January and October, resulting in limited available information.
3. According to US Bureau of transportation Statistics-

Causes	Air Carrier Delay	Aircraft Arriving Late	National Aviation System Delay	Security Delay	Extreme weather
% Delays in flights	30.1	39.6	24.9	0.1	5.6

Table 5

But none of the variables were included in the dataset, so predictions might have more error rate.

4. Factors like visibility index, Crew problems etc. should also be provided.
5. Also, the flight occupancy can be a major factor for cancellation or delay in flights.
6. If Global time were to be instead of local time, the comparisons would have been possible.

Business Proposal / Recommendations

1. Factors like visibility index, Crew problems etc. should also be provided. Also, the flight occupancy can be a major factor for cancellation or delay of flights.
2. Well-executed operational procedures, predictive maintenance, and anticipatory adjustments on days with a lot of disruption can minimize delays.
3. Airlines can also build up resource flexibility through employee contracts, but this requires time and advance planning.
4. Buffers may include added block time or turn time to create more flexibility in the event of non-flight–related delays, such as those caused by unexpected maintenance.
5. When it comes to human power, some training can be provided in order to improve the reliability of any such personnel.

Future Scope

Further analysis can be conducted if data on variables like weather conditions, visibility index and passenger occupancy of flights are available.
Also, the scope of this project is very much confined to the flight of United States, but we can include more countries like China, India, and Russia.

References

- Assessing the U.S. Climate in January 2018 | News | National Centers for Environmental Information (NCEI).
- Assessing the U.S. Climate in October 2018 | News | National Centers for Environmental Information (NCEI).