FLIGHT DELAY PREDICTION

By Iska Okta Fauziah

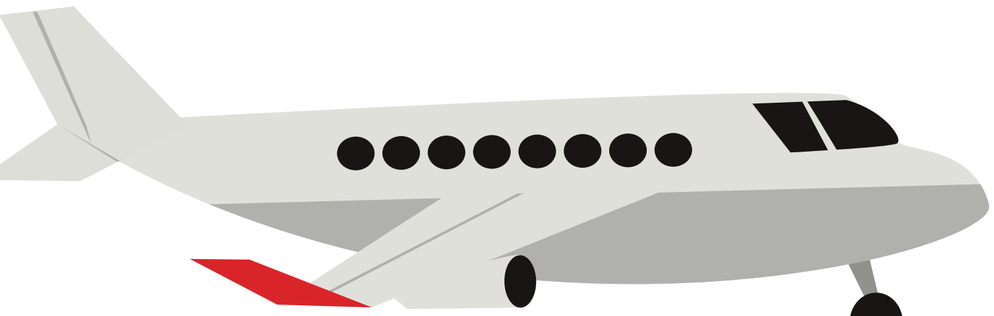


Table of Contents

Background and Objective

Background

Flight delays are gradually increasing and bring more **financial difficulties** and **customer dissatisfaction** to airline companies. This causes **losses for both the airline and passengers.**

To resolve this problem, we can use

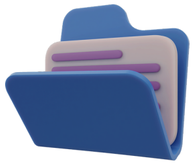
Machine Learning approach

Objective

To build machine learning classification model to predict whether the flight will be delayed or not.

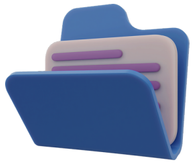
ABOUT DATASET

Link dataset



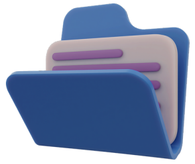
6
Numerical Column

3
Categorical Column



0%
Missing Values

0%
Duplicated Data



539.383
Rows

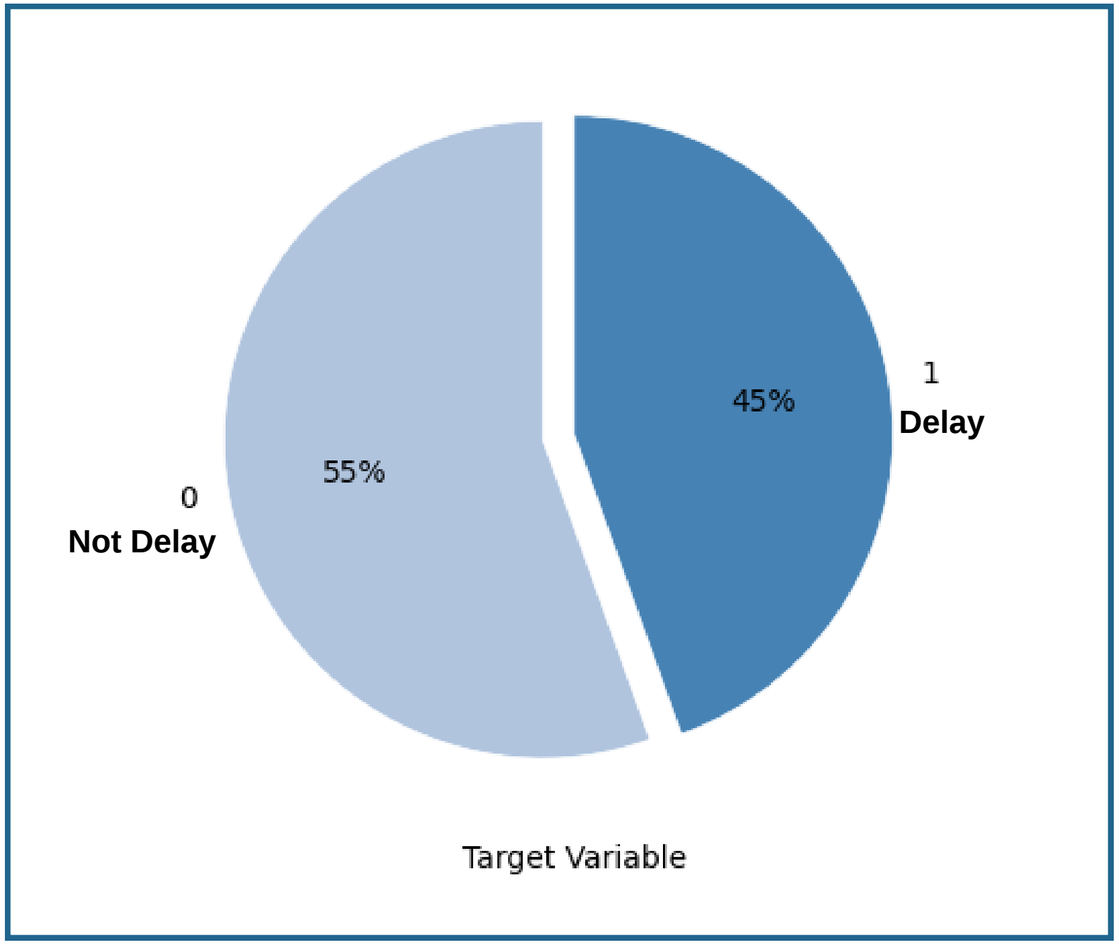
9
Columns

- **id**
- **Airline**
- **Flight**
- **AiportFrom**
- **AirportTo**
- **DayOfWeek**
- **Time**
- **Length**
- **Delay**

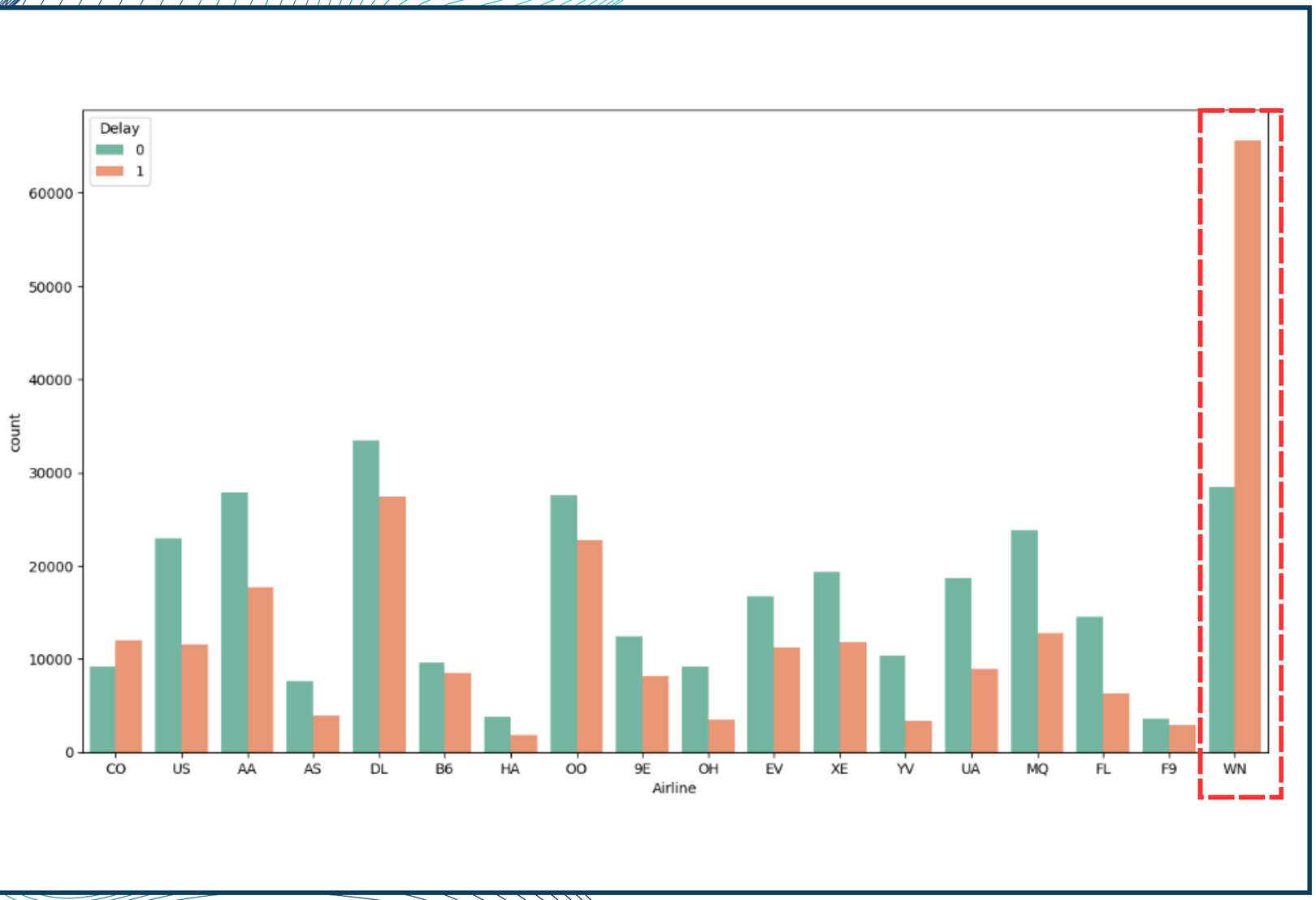
Target Variable

Exploratory Data Analysis

Percentage of Target Variable

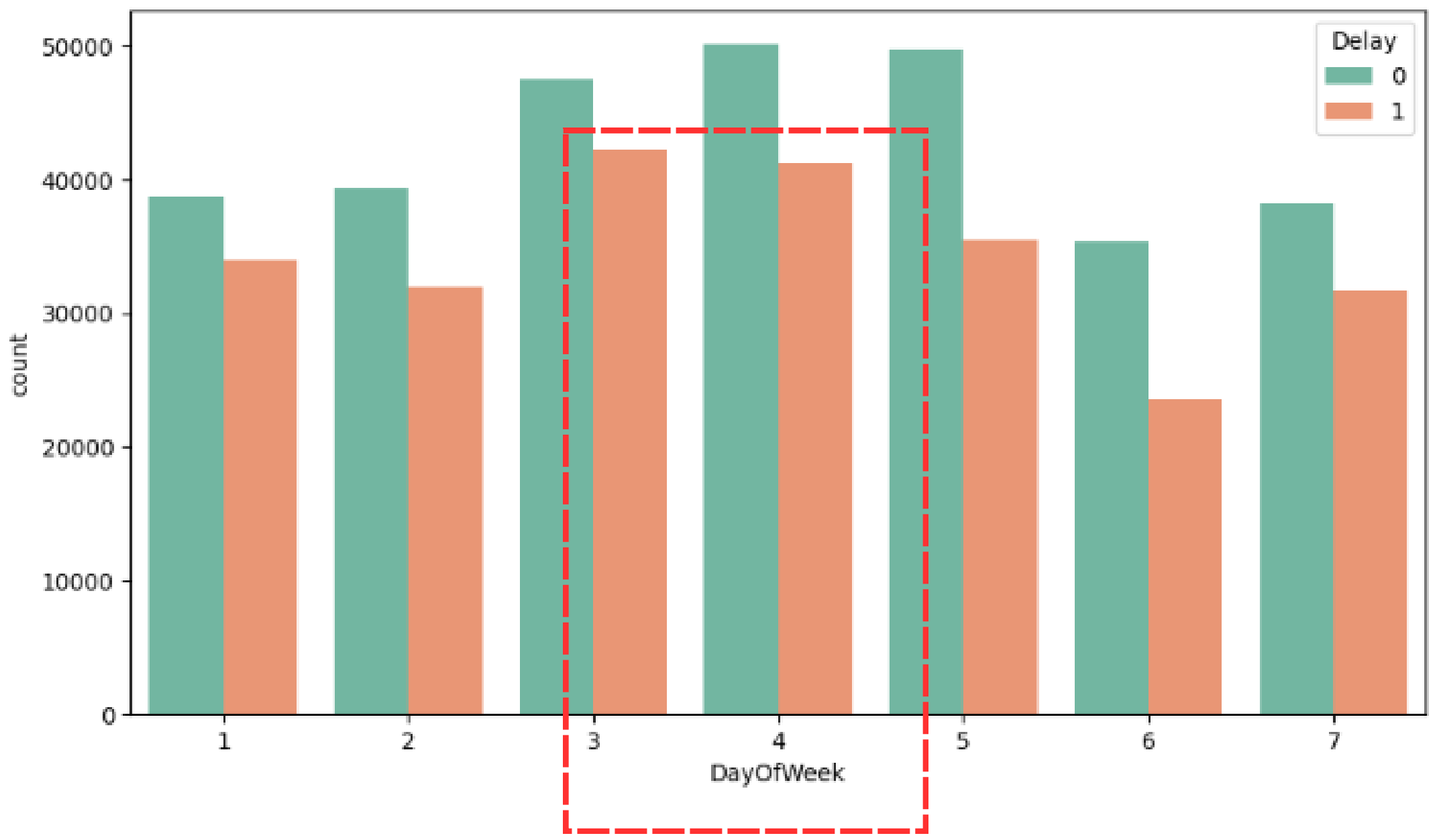


Based on the pie chart, it can be seen that **45% of flight in this dataset are labelled as flight delay** and **55% of flight are not delay.**



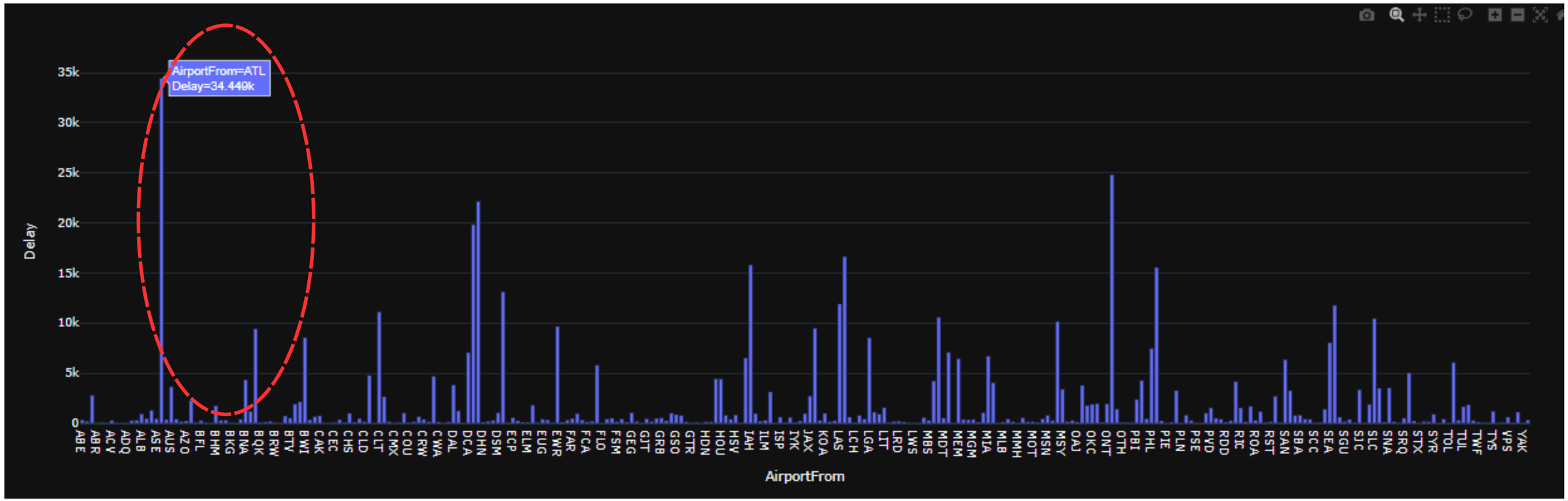
Which airline has the most delays?

WN is Airline has the most delay



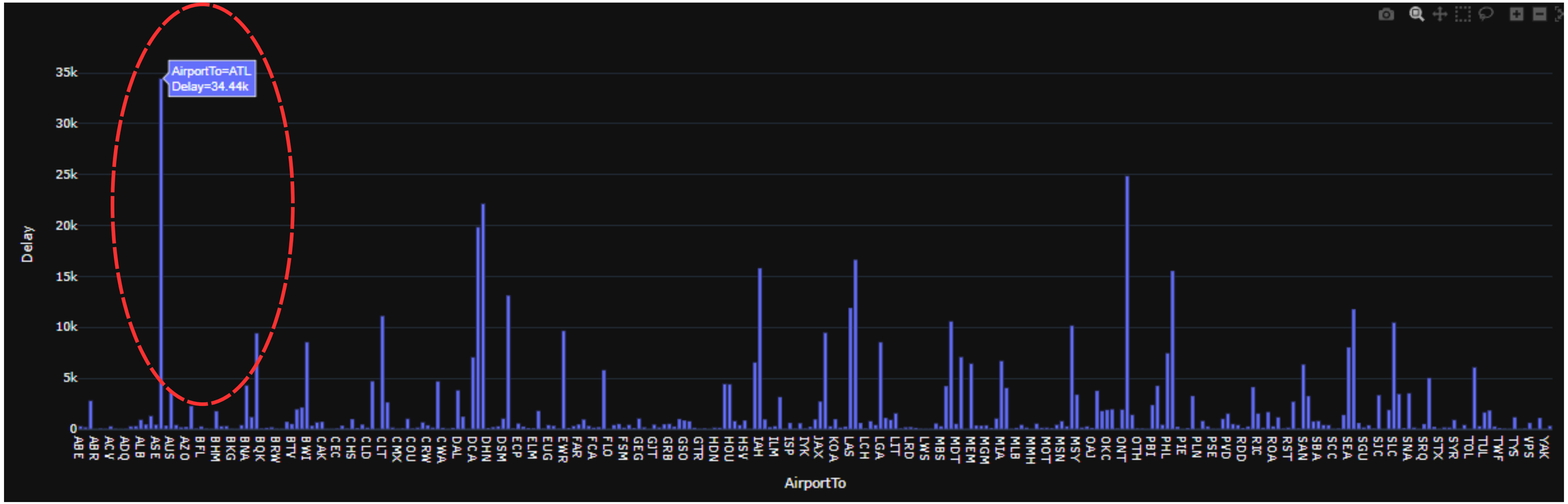
What days are there a lot of flight delays?

On Wednesday and Thursday many flight delay.



Which departure airport has the most flight delays?

Based on graph above, it can be seen that the departure airport 'ATL' has a lot flight delay among other airports.



Which destination airport has the most flight delays?

Based on graph above, it can be seen that the destination airport also 'ATL' airport has a lot flight delay among other airports.

Statistic Test

Is the delay significantly affected by day of week?

p-value

2.6007280556741412e-251

After doing statistic test using Chi-Square statistic, it can be known the p-value above. Because of p-value < 0.05, then 'day of week' affects delay of flight.

Is the delay significantly affected by time, length, and velocity?

P>|z|	
time	**0.000**
length	**0.000**
velocity	**0.000**

After doing statistic test using Logit Regression, it obtained that the three variables above (time, length, and velocity) have P>|z| is 0. Beacuse the value of P>|z| is small, it can be conclude that time, velocity, and length affected the delay.

Data Pre-processing

Data Pre-processing

01 **Handling Outlier**

02 **Adding ‘Velocity’ column (Velocity = Length/Time)**

03 **Mean Encoding for Categorical Variable**

04 **Handling Imbalance Data (Oversampling SMOTE) And drop ‘id’ column**

Modelling

Evaluation - Baseline

	Baseline
Modelling Classification Algorithms	**Metric Evaluation;Recall**
Logistic Regression	0.4273
Decision Tree	0.5250
KNN	0.4931
Random Forest	0.5307
Gaussian Naive Bayes	0.4161
Gradient Boosted Tree	0.4954

Evaluation - Imbalance Data

	Imbalance + Handling Outlier + Frequency Encoding	Imbalance + Mean Encoding	Imbalance + Handling Outlier + Mean Encoding
Modelling Classification Algorithms	**Metric Evaluation;Recall**	**Metric Evaluation;Recall**	**Metric Evaluation;Recall**
Logistic Regression	0.3623	0.4584	0.4785
Decision Tree	0.4703	0.4775	0.4734
KNN	0.5557	0.5583	0.5546
Random Forest	0.5533	0.5584	0.5508
Gaussian Naive Bayes	0.4094	0.3823	0.4125
Gradient Boosted Tree	0.4286	0.4593	0.4599

Evaluation - Balance Data

	Balance + Handling Outlier + Frequency Encoding	Balance + Mean Encoding	Balance + Handling Outlier + Mean Encoding
Modelling Classification Algorithms	**Metric Evaluation;Recall**	**Metric Evaluation;Recall**	**Metric Evaluation;Recall**
Logistic Regression	0.5550	0.5980	0.5834
Decision Tree	0.5060	0.5074	0.5068
KNN	0.5893	0.5983	0.6022
Random Forest	0.5827	0.5829	0.5841
Gaussian Naive Bayes	0.5182	0.4601	0.4921
Gradient Boosted Tree	0.5822	0.5911	0.5972

Best Model

Best Model:

KNN +
BALANCE DATA
HANDLING OUTLIER
MEAN ENCODING

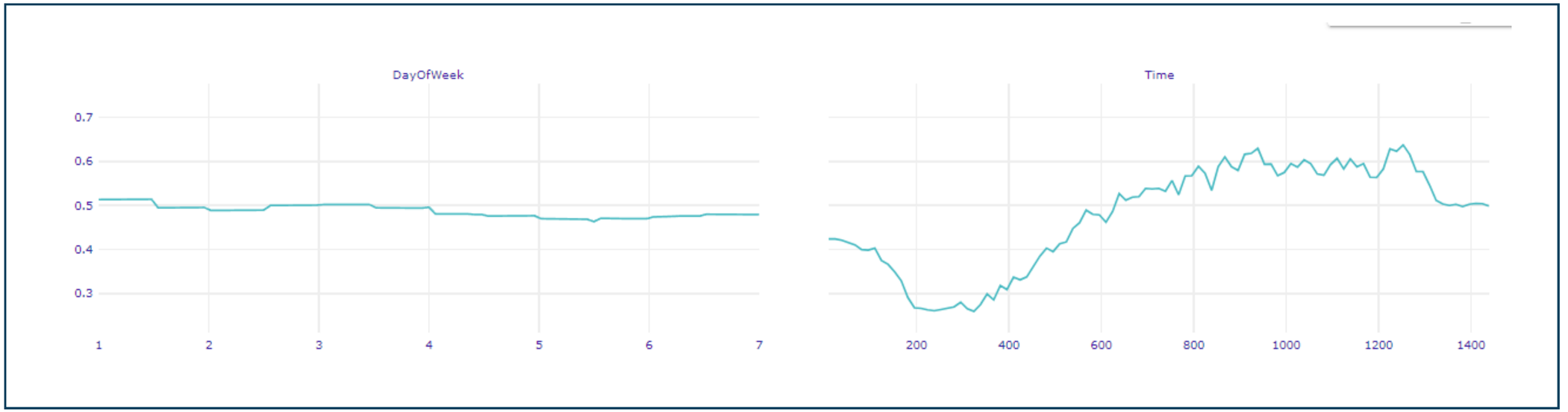
RECALL

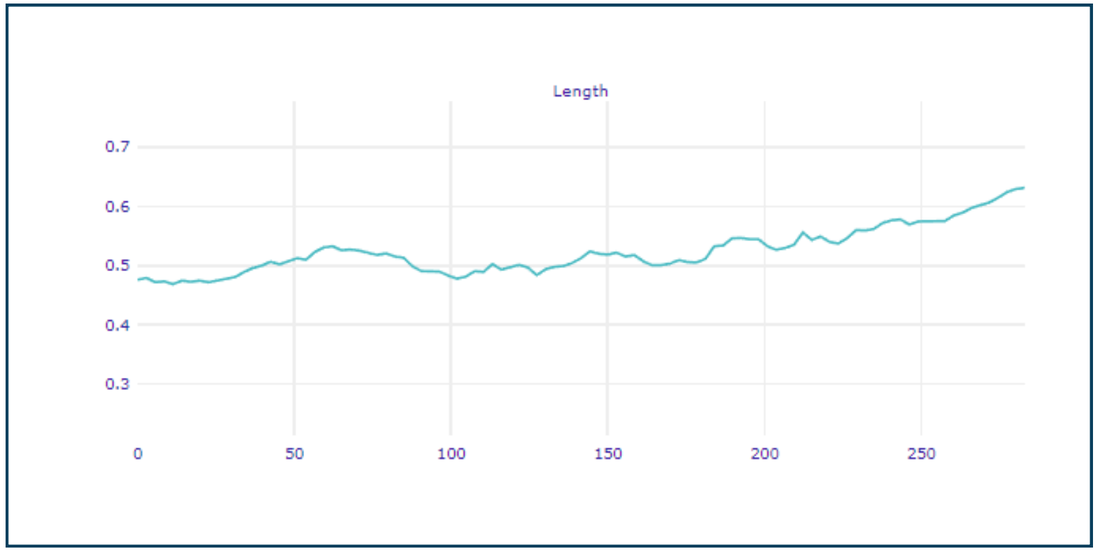
60.22%

Note:
After tunning with several parameters, there is no difference with the recall value of best model, so the best value will still be used for this modeling.

Dalex
(Feature Importance)

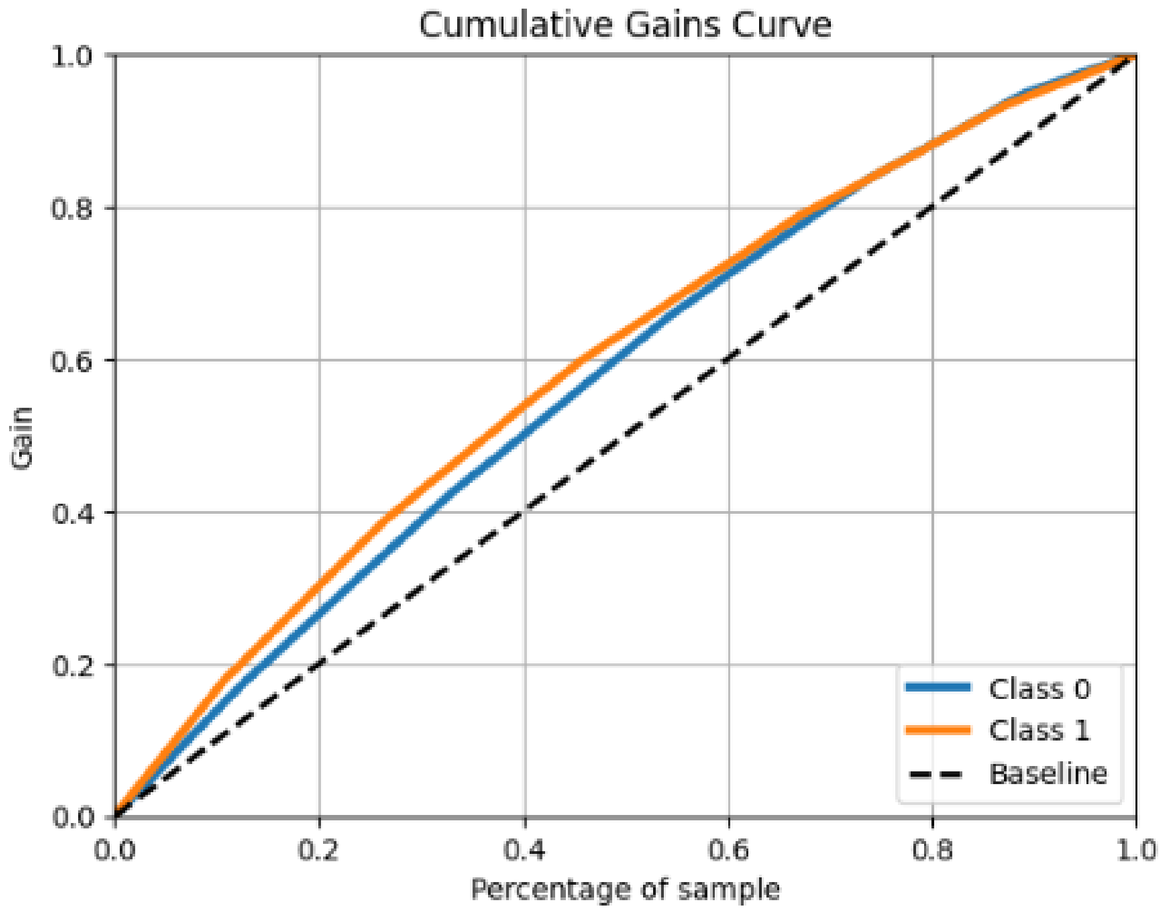
Feature Importance





after analyzing the feature importance, it can be known that **'DayOfWeek, Time, and Length'** are **the most influential predictor variable on delay.**

Business Case



Based on the Curve, it can be seen that there is **an increase in company performance** between using a model and without a model**.**

Baseline in curve means without modelling

Calculating Profit based on True Positive (TP) and True Negative (TN) when Using Modeling

*It used 20% of sample to calculate the profit

Profit Based on TP		
	Model	Baseline (Without Model)
Saved	2019	1153
Failed	3750	4616
Total cost	$17,307	$17,307
Bruto	$2,019,000	$1,153,000
Netto	$2,001,693	$1,135,693
Difference Using Model	$866,000	

Profit Based on TN		
	Model	Baseline (Without Model)
Saved	2599	1575
Failed	5279	6303
Total cost	$23,634	$23,634
Bruto	$2,599,000	$1,575,000
Netto	$2,575,366	$1,551,366
Difference Using Model	$1,024,000	

Profit the company can increase

$1,890,000

when using modeling

Note:

TP and TN are data successfully predicted correctly by our modeling

Conclusion and Recommendation

Conclusions

- Among several modelling methods, ‘Balance data + handling outlier + mean encoding’ is the best one.
- The best kind of model is K-Nearest Neighbour (KNN).
- Some features that most influence to delay is day of week, time, and length of flight.
- After analyzing the business case, it can found that there are increasing of profit of the company when using modeling.

Recommendation

- **Flight Planning and Scheduling**

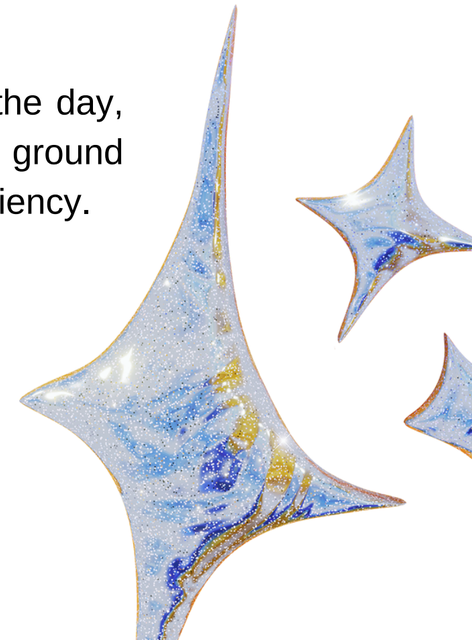
Airlines can utilize the insights from our model to optimize their flight planning and scheduling processes. By considering the impact of the day, duration, and length of the flight on delays, airlines can adjust their schedules to minimize the likelihood of delays. This can result in improved operational efficiency, better on-time performance, and increased customer satisfaction.

- **Passenger Communication:**

Airlines can proactively communicate potential delays to passengers based on the predicted. By providing passengers with advance notice and managing their expectations, airlines can enhance the overall travell experience. Passengers can make informed decisions and adjust their plans accordingly, reducing frustration and improving customer satisfaction.

- **Crew and Resource Allocation:**

Flight delay predictions can assist airlines in optimizing crew and resource allocation. By considering the impact of the day, duration, and length of the flight, airlines can allocate resources more effectively, ensuring that the necessary crew, ground staff, and facilities are available at the right time and place. This can minimize disruptions and improve operational efficiency.



Thank You!

Feel Free to Contact me!

https://wa.me/082393617436

iskaoktafauziah03@gmail.com

www.linkedin.com/in/iska-okta-fauziah

https://medium.com/@iskaoktafauziah5

https://github.com/IskaOktaFauziah03