FLIGHT DELAY PREDICTION

Vajjala Paidi Vikramaditya | Aditi Bihade
Shashank Thool | Piyush Kulkarni
Rohit Karan

Consulting and Analytics Club

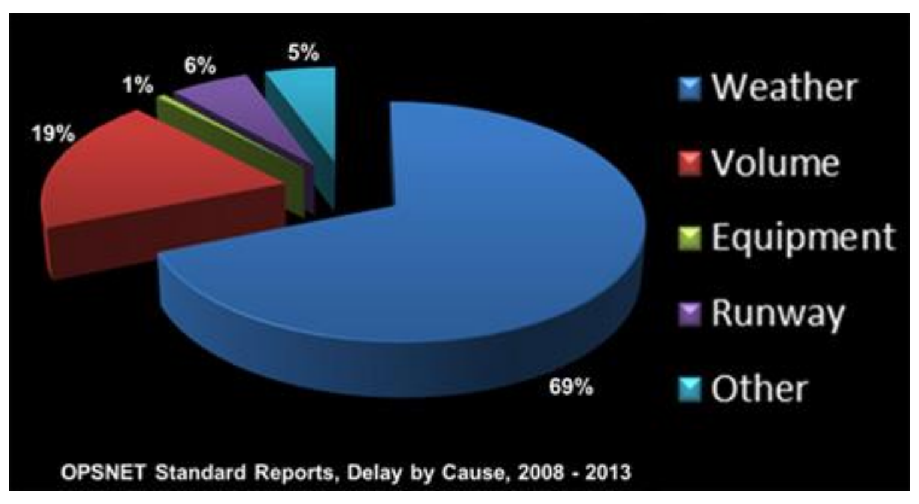
CONSULTING & ANALYTICS CLUB
IIT GUWAHATI

Introduction

Anyone who has ever booked a flight ticket knows how unexpectedly the flight timing varies. Airlines use using sophisticated tactics which they call **"yield management"**.

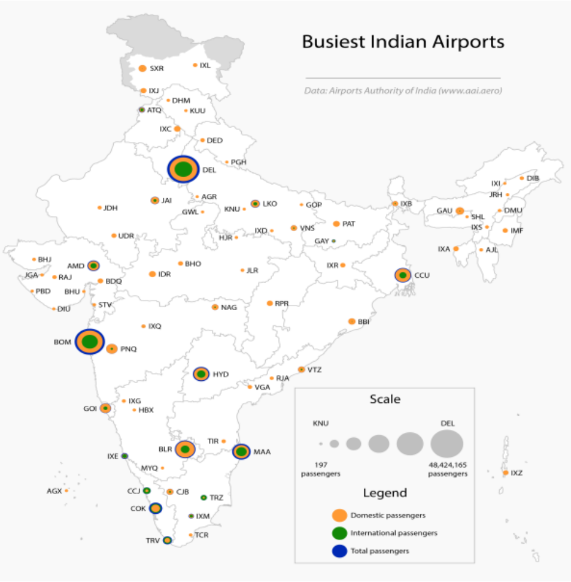
According to a recent study, India is currently the 3rd largest civil aviation market in the world. According to IATA (International Air Transport Association) the number of flyers globally could double to 8.2 billion in 2037.

Nowadays, in this ever advancing world **time management** has become a prominent factor in all business operations. We can't afford to lose time on unnecessary delays. Using this as motivation we have developed a ML Model which predicts arrival flight delay based on various factors that include airport score, general airline trend, air traffic etc.

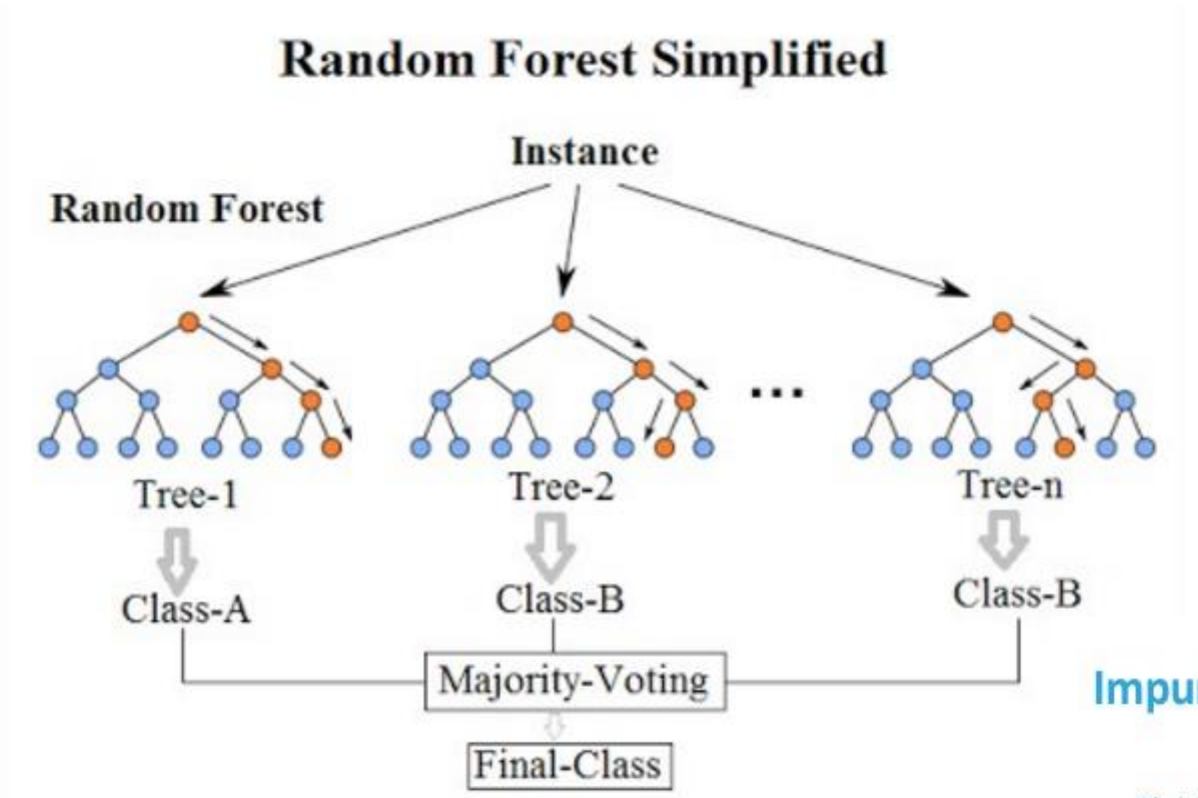


OPSNET Standard Reports, Delay by Cause, 2008 - 2013

It is clearly visible from the pie chart that **weather** is one of the most prominent factors causing delay, hence we have extracted weather features separately.



The Busiest Airports are the ones where the probability of flight delay is high, hence we chose Bombay, Delhi, Bengaluru, Chennai, Hyderabad



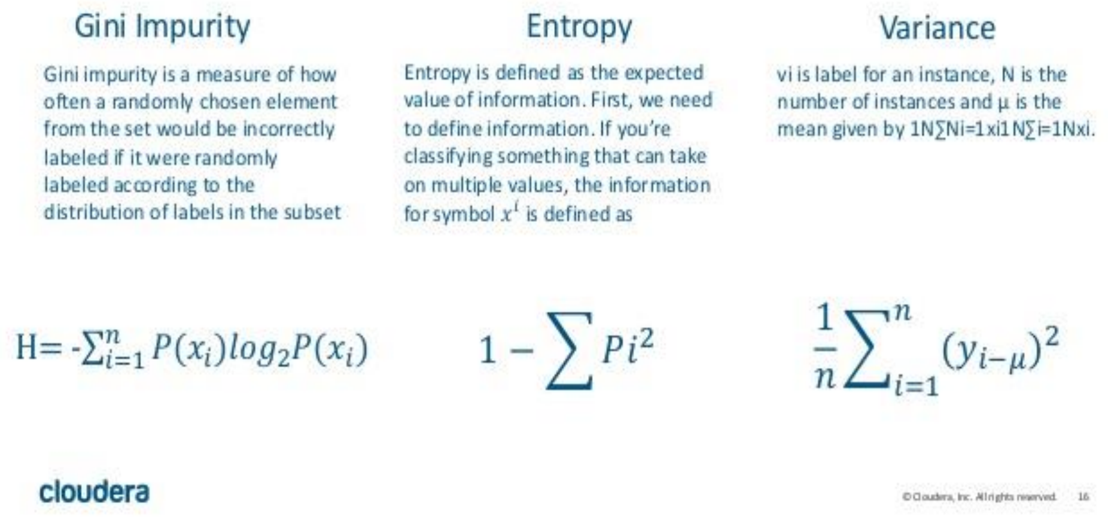
What is Random Forest?

Random Forest is an ensemble algorithm, which **randomly** creates a set of **decision** trees, which then aggregates the votes from all **decision** trees to decide the final class of the test object.

What is Impurity in Random Forest ?

The **impurity** of a node is the probability that a randomly chosen sample in a node would be **incorrectly** labeled if it was labeled by the distribution of samples in the node.

Impurity: Gini, Impurity, Variance



Gini Impurity

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset

Entropy

Entropy is defined as the expected value of information. First, we need to define information. If you're classifying something that can take on multiple values, the information for symbol x^i is defined as

Variance

vi is label for an instance, N is the number of instances and μ is the mean given by 1N∑Ni=1xi1N∑i=1Nxi.

$$H = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

$$1 - \sum Pi^2$$

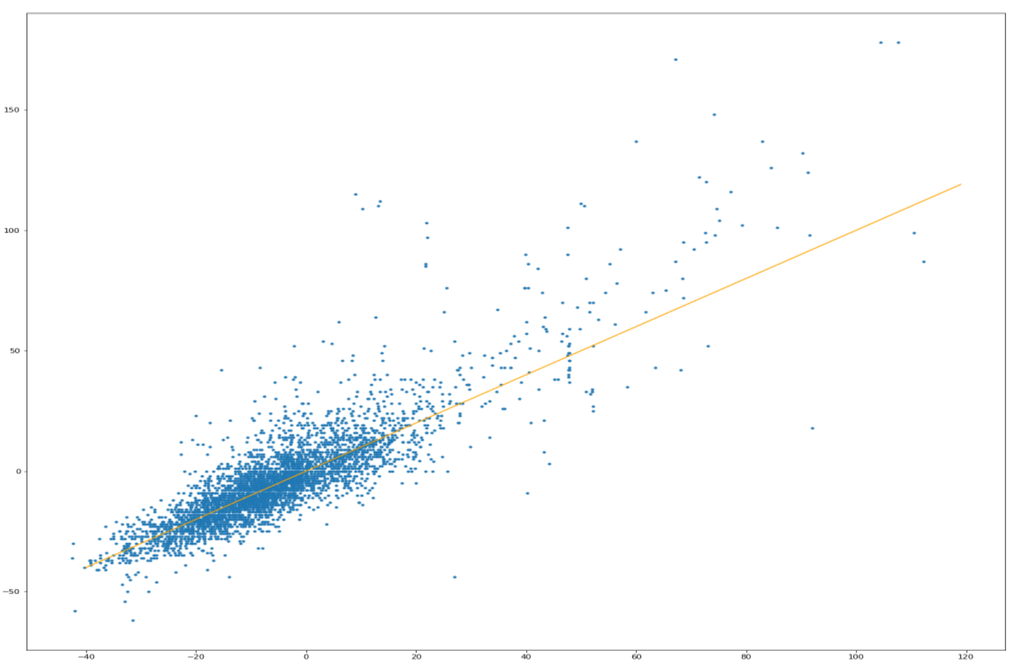
$$\frac{1}{n}\sum_{i=1}^{n}(y_{i-\mu})^2$$

cloudera

Results

After careful hyperparameter tuning by Randomized Search Cross Validation and application of various ML models

The training mean square error is obtained to be 0.13 and the test set mean square error is obtained to be 0.25.



GRAPH:
x-axis units : minutes
y-axis units : minutes

Orange line equation:
x = y

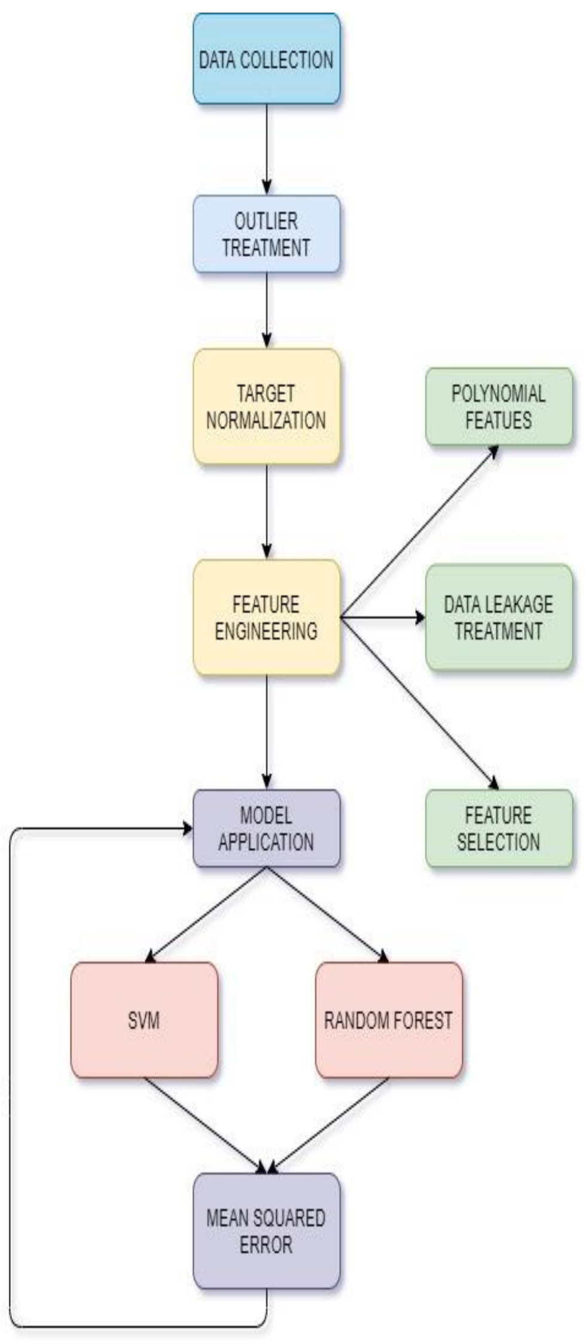
Techniques

Data Collection and Cleaning:

1. Collection of data from flightradar24.com by web scraping using python package **BeautifulSoup** which parsing HTML and XML documents.
2. Overall 23 features collected were collected from the website
3. Weather features were also extracted in a similar way using Beautiful Soup, 21 weather features were extracted
4. After categorical encoding, imputation of missing values and remove of rows with many null values, final dataset obtained had 10313 rows and 44 columns.

Data Pre-processing and Model Application:

1. **Removal** of the Departure Delay feature it caused Data Leakage
2. **Outlier Removal**: We considered only delays from 3 hrs early to 3 hrs late.
3. **Normalized** the target variable so that our model generalises well in real life.
4. Applied **Interaction** features which allowed us to capture better insights about the data.
5. Careful **feature selection** based on their importance. This allowed the model to improve its performance as well as reduce training time.
6. Started with basic linear models but they performed poorly, hence shifted to **tree based ensemble models**, they performed much better than better prediction were obtained.
7. The evaluation metric used was **Mean Squared Error**,
8. Finally inverse transform was applied to target variable to show predicted delays.



DATA COLLECTION
OUTLIER TREATMENT
TARGET NORMALIZATION
POLYNOMIAL FEATUES
FEATURE ENGINEERING
DATA LEAKAGE TREATMENT
FEATURE SELECTION
MODEL APPLICATION
SVM
RANDOM FOREST
MEAN SQUARED ERROR

Conclusion

Go Air was found to be the airline with least delay with an average delay of 11 minutes.

SpiceJet was found to be the airline with most delay with an average delay of 18 minutes.

The Airport which had the most Delay was **Bangalore,** and the airport with the least delay was **Hyderabad.**

Of of the given 44 parameters, **Wind Gust, Cloud Cover** and **Airport Location** were the most important which affected the delay the most.

As expected the Delay is high on Public Holidays and major religious festivals.

References

Airline Data: flightradar24.
Weather Data: accuweather.com
Scikit-Learn :
Numpy : numpy.org
Scipy : scipy.org
Pandas : pandas.org
MatPlotLib : matplotlib.org