Flight Prediction

**Initial project definition:**

"Predict if a flight will be on-time"

**Defining scope:**

* Define the end goal,
* Define the starting point, and
* Define how the goal will be achieved.

**Solution Statement:**

Define Scope (including data sources):

* US flights only
* Flight between US airports - DOT database - <https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time>

**Key notes:**

A flight is said to be delayed, if he arrives or take off 15 minutes after scheduled time.

**Redefined project definition:**

"Using DOT data, predict with 70+% accuracy if a flight would arrive 15+ minutes after the scheduled arrival time"

**Machine Learning Workflow**

- Process DOT data

- Transform data as required

The Machine Learning Workflow is used to process and transform DOT data to create a prediction model. This model must predict whether a flight would arrive 15+ minutes after the scheduled arrival time with 70+%."

**Process DOT data**

***Data value (parameter)***

* Day of the week
* Day of the month
* Unique Carrier
* Carrier
* TailNum – Tail Number
* FlightNum – Flight Number
* Origin AirportID
* Origin AirportSeqID
* OriginCityMarketID
* Origin
* Dest AirportID – Destination Airport ID
* Dest AirportSeqID – Destination Airport Sequence ID
* Dest - Destination
* DepTime – Departure Time
* DepDel15 – Departure Delayed Time
* DepTimeBLK – Departure Time Block
* ArrTime – Arrival Time
* ArrDel 15 – Arrival Delay Time
* Cancelled
* Diverted

Correlated Columns

* Same information in a different format (ID and value associated with ID)
* Add little information
* Can cause algorithms to get confused

Verifying if two columns is correlated with cor() function.

Tidy Data

* Tidy datasets are easy to manipulate, model, and visualize, and have a specific structure.

Molding Data

* Reviewing the data values in each of the rows and making a decision to whether to keep the rows.
* Adjusting data types to conform to what we need.
* Creating new columns, if required.

Fixing Arr\_Del15

* Arr\_Del15 = 1 if 15 minute delay
* It’s the value we are predicting
* Must be 0 or 1, where 0 is for false and 1 is true
* May contain NA
* May contain “”
* Remove rows with NA or “”
  + Performing similar analysis for Arr\_Del15 and Dep\_Del15

## Dropping some columns to better prep the data for analysis.

Note:

A positive correlated column, explains that two or more column is closely correlated. I.e- Both columns are serving similar purpose and both are independent of another.

**Role of algorithm**

Perform algortihm selection

Compare factor for algorithm function

1. Learning Type

2. Result type the algorithm produces

3. Complexity of the algorithm

4. Basic or enhanced data

Learning type, our prediction is a supervised learning.

Result type, regression and classification. - Classification, discrete values, true or false, small, medium or large

Complexity of the algorithm, keep it simple, eliminate 'ensemble' algorithms

Enhanced or basic,

Candidate algorithms

Naive Bayes, Logictic regression, Decision rule - logistic regression

## Initial selected algorithm - Logistic regression

# 1. Simple - easy to understand

# 2. Fast - up to 100x faster

# 3. Stable to data changes

Next stage

Loading Data

Exploring Data

Cleaning Data

**Training the data**

The Caret Package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

* Data splitting
* Pre-processing
* Feature selection
* Model tuning using resampling
* Variable importance estimation

Registered email address: irorereMet147472@futa.edu.ng

Product key: 1224-D6AA-B8F3-4CA0-30DE-DE76-3D97-2C59