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.

Next stage

Loading Data

Exploring Data

Cleaning Data

Define target performance

Define context for usage

Define how the solution will be created.

Registered email address: irorereMet147472@futa.edu.ng

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

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.

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