Sustainability Datathon

Team- Hadoop is not Dead

RYANAIR

ie UNIVERSITY

The Problem/Task

- Predict the optimal Ramp Weight of an Aircraft given a set of variables

- According to a McKinsey analysis, carriers reduced their fuel consumption per passenger-kilometer by approximately 39 percent between 2005 and 2019 (pre-COVID-19) by deploying various measures. However, advances in machine learning and Analytics can bring in even further improvements

- Essentially the need is to predict optimal fuel that a Aircraft should carry. The benefits are 2 fold
  - Cost Savings
  - Less Carbon emissions by burning less fuel

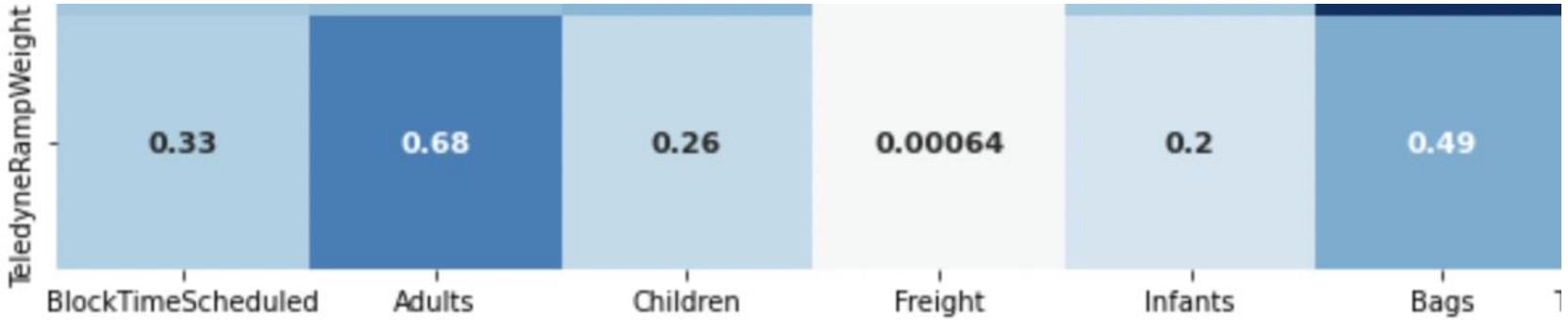
EDA

Starting off with the extended training set, no missing values and outliers in quite a few numerical columns.

In the original train set, we identified the following as Categories and Numericals:

Numericals	BlockTimeScheduled Adults Children Freight Infants Bags TeledyneRampWeight
Categories	FlightID FlightNumber AircraftRegistration AircraftCapacity AircraftTypeGroup ServiceDescription Carrier AOCDescription ScheduledRoute DepartureScheduled ArrivalScheduled

EDA- Correlation to target



As expected- high correlation of the number of passengers with target and the total block time affecting target as well

Also, our EDA revealed quite a few outliers in the target column itself and all outliers over 2 standard deviations were removed later

Assumptions after EDA

After some research and EDA, we decided on the following factors that can affect an Aircraft weight and decided to do feature engineering based on these assumptions and findings

- **Temporal**- Blocktime
- **Distance**- Scheduled Route
- **Payload**- All passengers/Crew + Bags + Cargo
- **Rate of fuel consumption & expected fuel Weight (only available in extended set)**- Burnoff, difference between ZFW and TOW/RampWeight
- **Type and Age of Aircraft**- AircraftRegistration & AircraftTypeGroup
- **Weather**- Not included in training or extended training
- **Delays**- Included only in extended

Features Engineered

- **Temporal**
  - Extracted Day of Month, Day of Week, Month, Year and Shift (day divided into 6 different shifts from ‘Early Morning’ to ‘Late Evening’) from column Departure Scheduled

- **Distance**
  - Extracted latitude and longitude from Scheduled Route Column and computed distances between Airports

- **Weight**
  - Assumed AircraftType Group can affect Weight of the Aircraft.
  - Computed payload assuming Average weights of Adults, Children, Infants and checked in bags + Freight (which was already in kgs)

Features Engineered

- **Age of Aircraft**
  - Assuming Aircraft Registration has this information
- **Weather**
  - External Data was required for this and was not considered.
- **Delays & Deviations**
  - Only available in the extended training set and not considered

Features Engineered

- Ran a model to compute Burnoff and projected it on the actual test set
- Ran a second model to compute Fuel Weight (difference between ZFW and TOW). This was also projected on the actual test set
- Ran a linear Regression Model using a subset of variables (dayofweek, month, shift and Scheduled Route) to compute errors and control overfitting to come extent (done after training the model)

Final Set of Features to predict target

- AircraftRegistration
- AircraftCapacity
- AircraftTypeGroup
- ServiceDescription
- AOCDescription
- ScheduledRoute
- BlockTimeScheduled
- Burnoff
- Fuel
- Payload
- Month
- Dayofweek
- Year
- Shift
- Scheduled_distance

Not Used by the model or for any feature engineering

- Flight ID
- Flight Number
- Carrier
- Arrival Scheduled

Model Selection & Evaluation

Two Models tried with Randomised Search CV

- Random Forest Regressor
- XGBoost Regessor

Our score on train/test split was giving us a score of 785 - 795. There was a small difference between XGBoost and Randomforest with XGBoost giving marginally better MAE.

Hence, our final submission was with XGBoost with a final MAE of 931

Conclusion and Learning

Firstly, we understand that our model was overfitting and was not able to generalise as well on the actual test. Hence we did run a linear model to project errors onto the test set- this helped slightly

Things that we tried but did not help

- Tried to compute delays into our model
- Tried to take care of outliers in Fuel and Payload but it did not help

Things to be explored

- Since XGBoost is sensitive to Hyper-parameters, we would like to do more Grid searches. Given the volume of data, this was challenging
- We would like to explore how to incorporate delays and route deviations and model that information
- Extract further information from the extended dataset
- We would like to enrich our data with weather related information since this plays a crucial role in fuel consumption, cruising speeds, delays etc.

Questions that we would like to have answers of

- How much of an impact does age of aircraft have on Fuel consumption? What is Ryan Air's policy on older Aircrafts?
- Ryan Air has a fairly mature analytics division- Approximately by how much has Ryan Air been able to reduce fuel consumption by using data and analytics/ML?
- Is there a yearly KPI for the above?
- If there is then how is Ryan Air performing in terms of its competitors? How much of a competitive advantage is this?
- Given that 2022 summer had unprecedented travel demand, was there a deviation in fuel consumption efficiency?
- How does RyanAir make use of Analytics/ML when making fuel and Carbon Credit hedging decisions? (from the quarterly earnings report, it seems 80% of RyanAir's fuel requirements in 2023 are hedged)

Thank You

!= DEAD