Gresë Berisha
Moritz Baur
Kevin Dietrich

2025-03-17

Problem and Motivation

Problem and Motivation

✈ Problem

- Flight delays are disruptive and costly.
- Passengers lose time ⏱ and confidence.
- Airlines and airports face financial losses 💼 and reduced operational efficiency.

Problem and Motivation

✈ Problem

- Flight delays are disruptive and costly.
- Passengers lose time ⏱ and confidence.
- Airlines and airports face financial losses 💼 and reduced operational efficiency.

📈 Motivation

- Tunisair aims to implement a predictive solution to anticipate delays and mitigate their impact.

🎯 Project Objective

🎯 Project Objective

🥅 Goal

Use machine learning to predict the length of flight delays (in minutes).

🎯 Project Objective

🥅 Goal

Use machine learning to predict the length of flight delays (in minutes).

💥 Impact

- ✔ Better scheduling
- ✔ Reduced operational inefficiencies
- ✔ Improved passenger satisfaction

Dataset and Evaluation

Dataset and Evaluation

Data Source

📁 Flight data provided by Zindi, consisting of a train/test format for model development

Dataset and Evaluation

Data Source

📁 Flight data provided by Zindi, consisting of a train/test format for model development

Prediction Target:

🕒 Delay duration in minutes

Dataset and Evaluation

Data Source

📁 Flight data provided by Zindi, consisting of a train/test format for model development

Prediction Target:

🕒 Delay duration in minutes

Performance Metric:

📉 Root Mean Square Error (RMSE)

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

🔍 Initial Insights from EDA

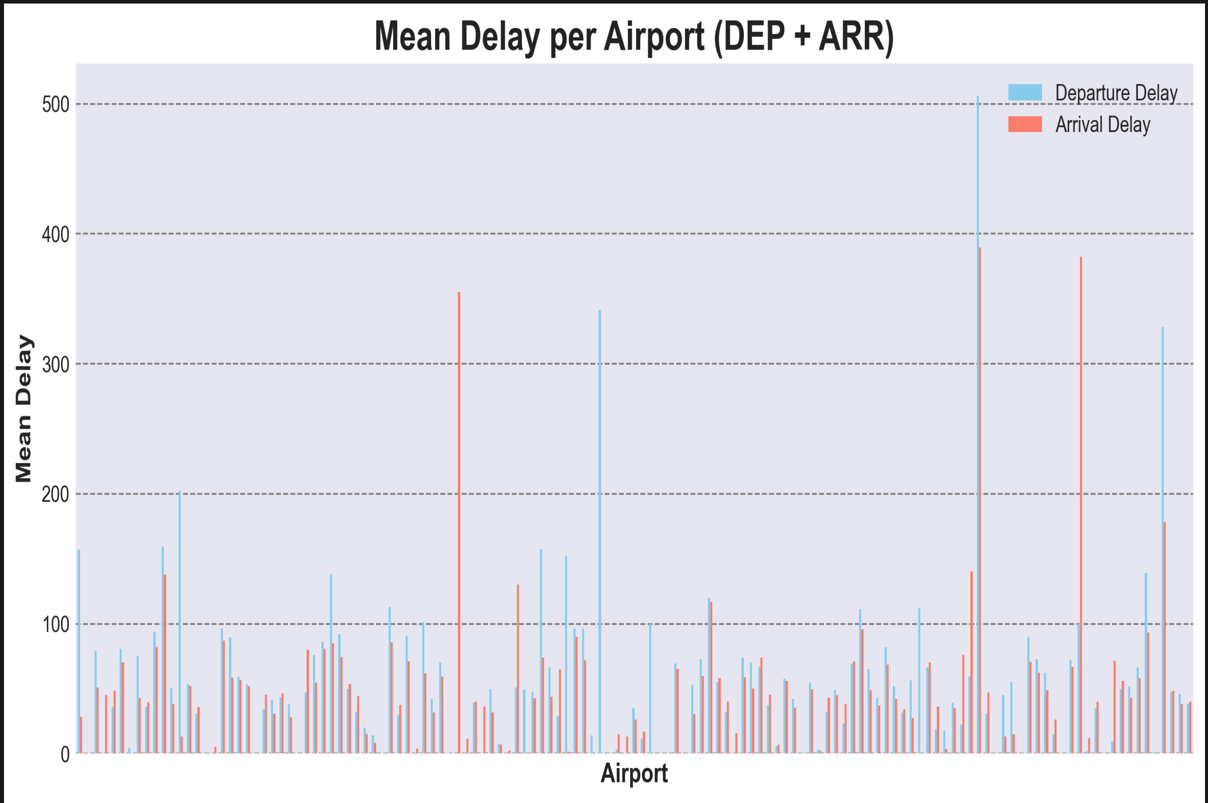
We analysed how delays are distributed based on:

- Departure airports
- Arrival airports
- Temporal trends across the years 2016–2018

Column Description

Column	Description
ID	Unique flight identifier
DATOP	Date of flight
FLTID	Flight number
DEPSTN	Departure point
ARRSTN	Arrival point
STD	Scheduled time of departure
STA	Scheduled time of arrival
STATUS	Flight status
AC	Aircraft code
target	Flight delay (min)

Mean Delay per Airport by Departure and Arrival



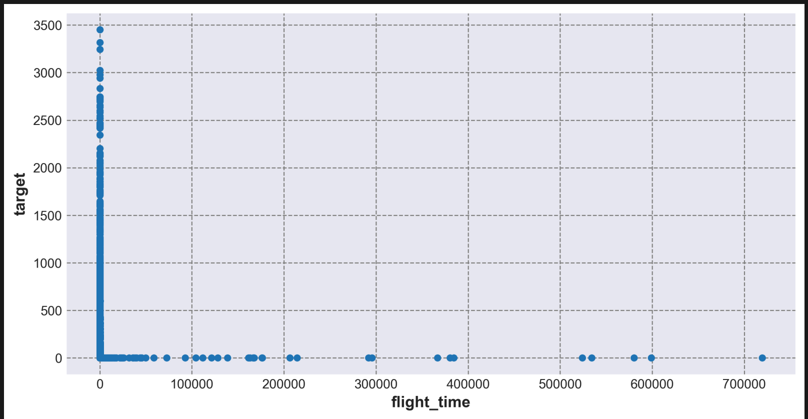
Deriving flight time from STD and STA

Deriving flight time from STD and STA

Removing Service Flights, i.e. flights where departure and arrival airports are the same …

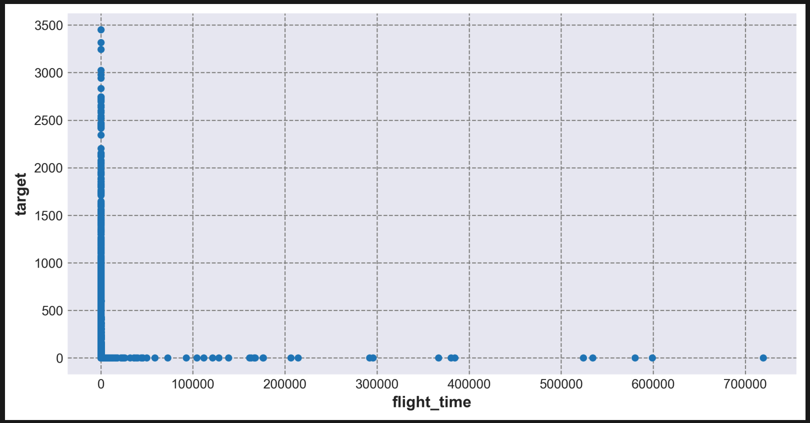
Deriving flight time from STD and STA

Removing Service Flights, i.e. flights where departure and arrival airports are the same …



Deriving flight time from STD and STA

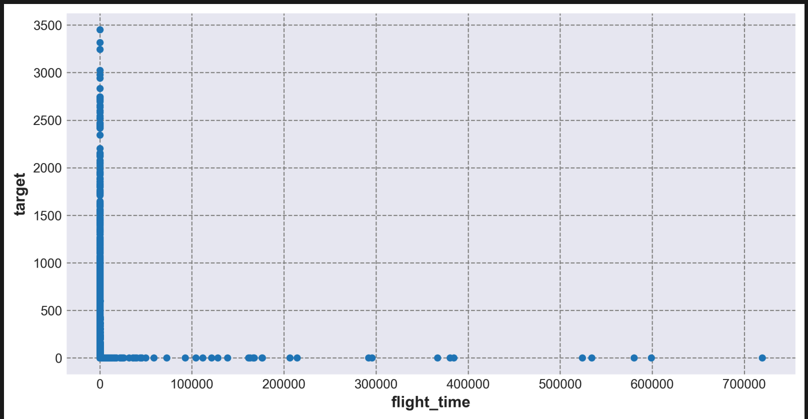
Removing Service Flights, i.e. flights where departure and arrival airports are the same …



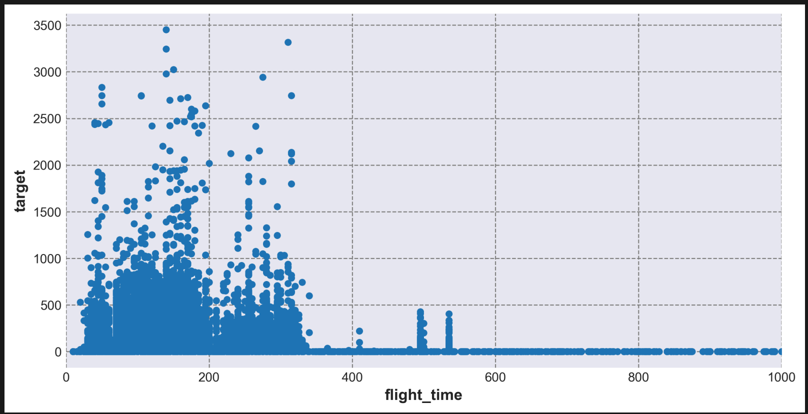
$\Rightarrow$

DERIVING FLIGHT TIME FROM STD AND STA

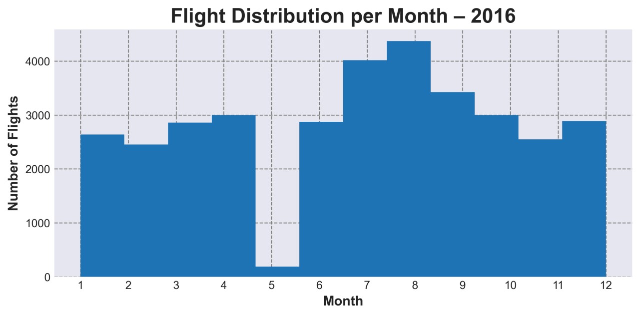
Removing Service Flights, i.e. flights where departure and arrival airports are the same …

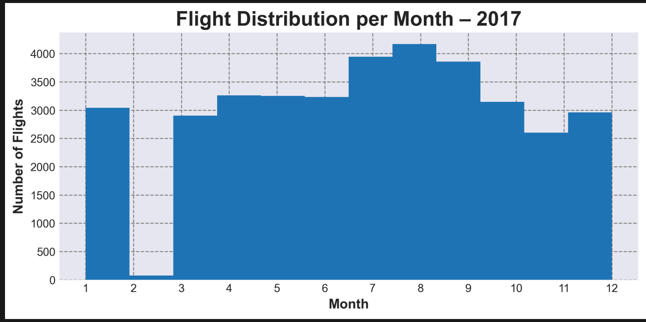


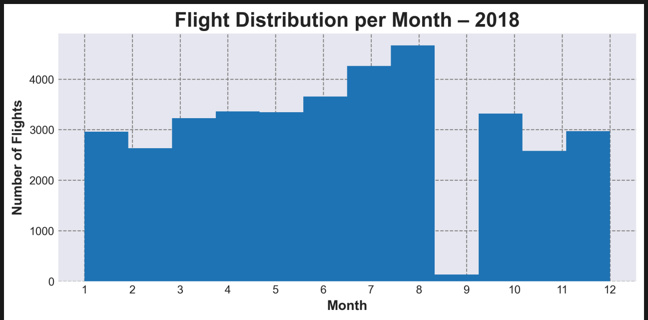
⇒

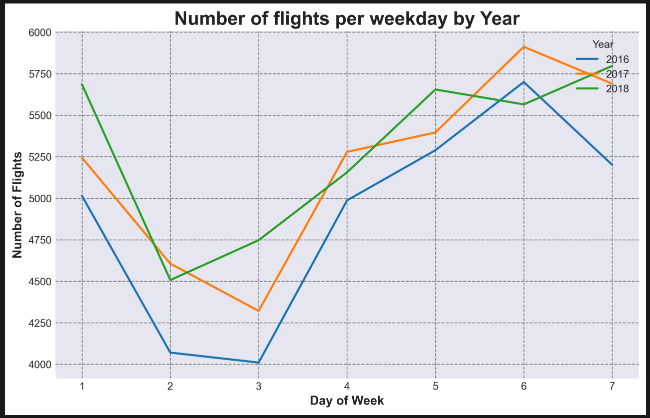


DISSECTING DATOP INTO YEAR, MONTH AND DAY OF THE WEEK









Baseline Model

Baseline Model

Approach

A simple linear regression model using only the day of the week (or aircraft code) as the predictor.

Baseline Model

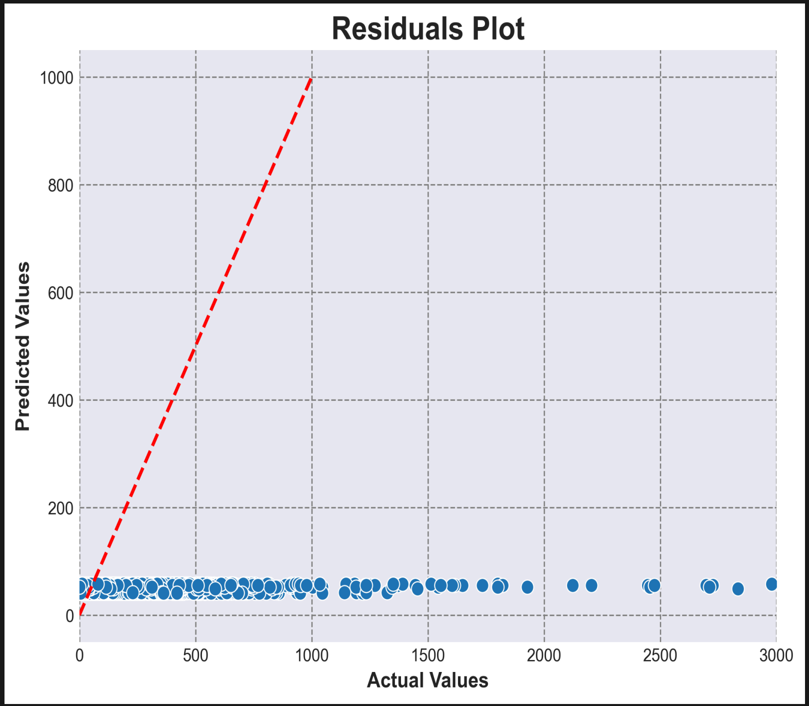
Approach

A simple linear regression model using only the day of the week (or aircraft code) as the predictor.

Performance

📊 RMSE ≈ 114.69

🤖 R_2 ≈ 3.01 %



ML Model

ML Model

Many categorical variables …

ML Model

Many categorical variables …

… but there is:

ML Model

Many categorical variables …

… but there is:

CatBoost

Machine-generated image—no animals were harmed

ML Model

ML Model

Approach

A regression model using CatBoost with the following predictors: *Flight Status; Aircraft Code; Departure and Arrival Point; Year, Month and Weekday of Departure Time*

ML Model

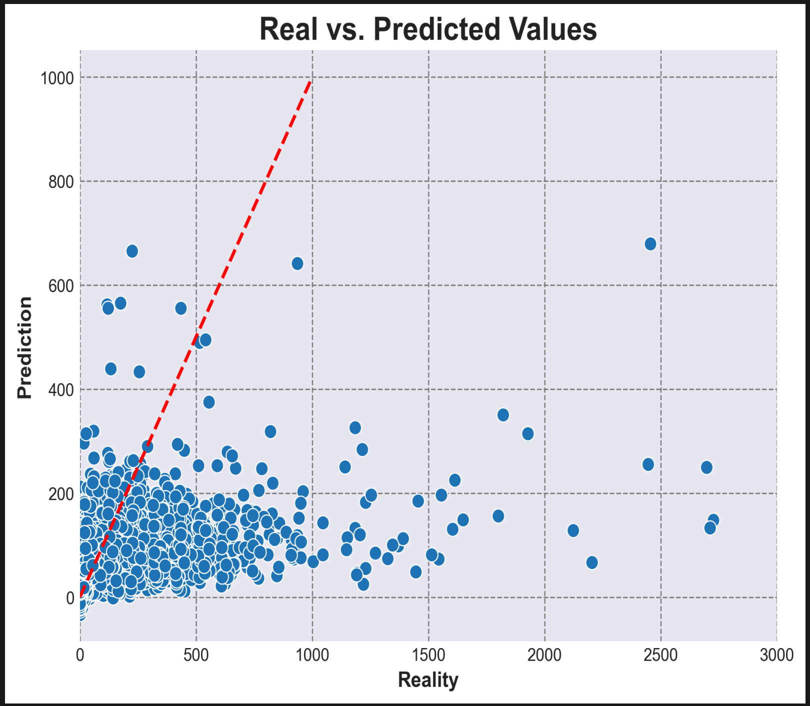
Approach

A regression model using CatBoost with the following predictors: *Flight Status; Aircraft Code; Departure and Arrival Point; Year, Month and Weekday of Departure Time*

Performance

📊 RMSE ≈ 96.14 (< 100)

🤖 R_2 ≈ 30.48 %



💬 Questions and Feedback

We're happy to answer your questions and look forward to your feedback.