TfL Cycling Analysis

Manuel Urbano Rodriguez

12 June 2022

Introduction

Business Problem

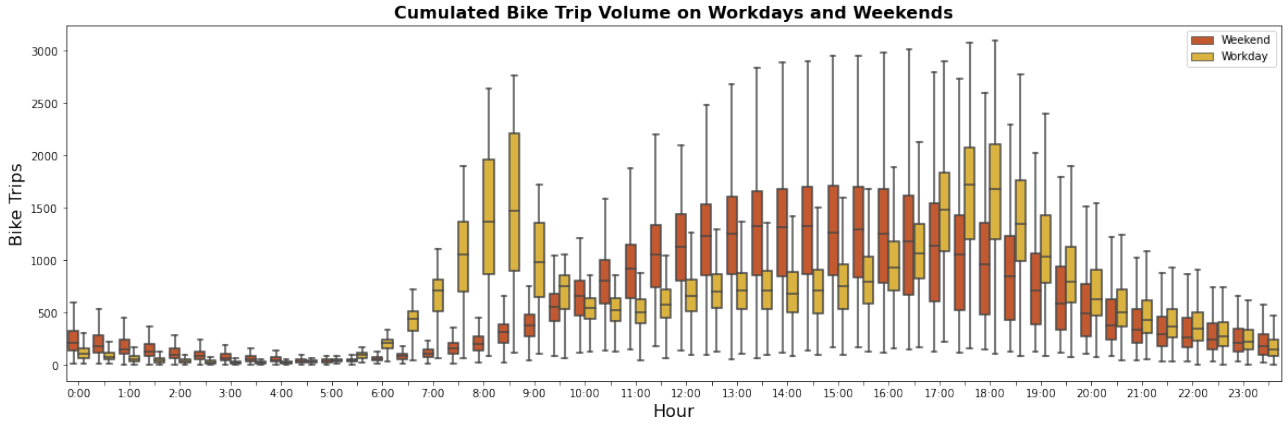
- A client is considering expanding their business into short-term rental cycles.
- To develop a strategy, it is essential to understand how people in London use already-existing cycling services.
- They are interested in identifying customer profiles to attract as well as operational concerns in terms of reliability and supply chain management.
- Use data from the TfL cycling public datasets from 2019 and 2021.

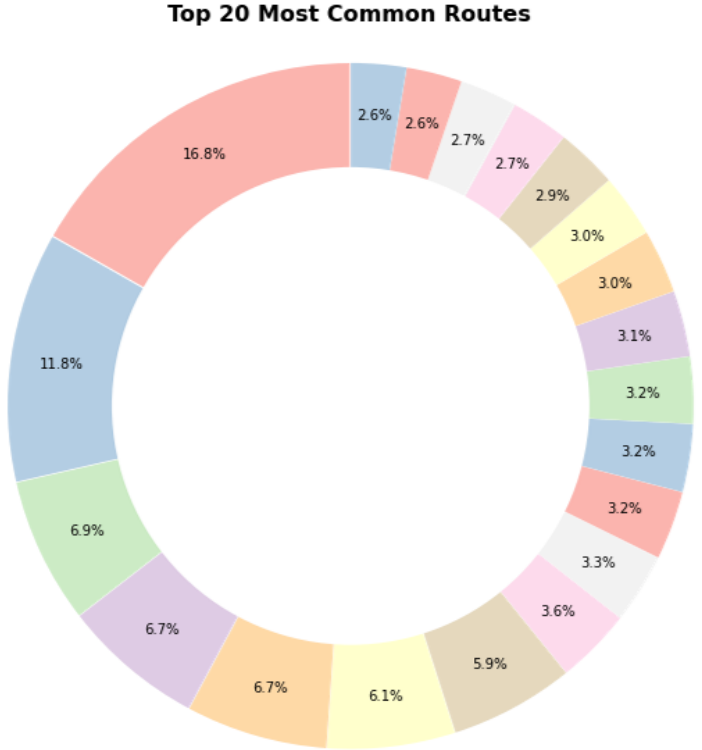
Methodology

- Using web scraping, we gathered a total of 156 CSV files from the TfL website. Data was parsed and concatenated in a single Dataframe (2.3+ GB).
- Weather Data and additional data from stations will be also scraped to complete the analysis.
- An initial Exploratory Data Analysis (EDA) will be performed to determine the best approach to identify different Data Science use cases.
- A usage trip volume prediction will be performed for the customer to anticipate demand in advance for efficient bike management.
- Results will then be presented with the key features that will drive the number of bikes being rented.

Exploratory Data Analysis

What insights can we obtain from our data?



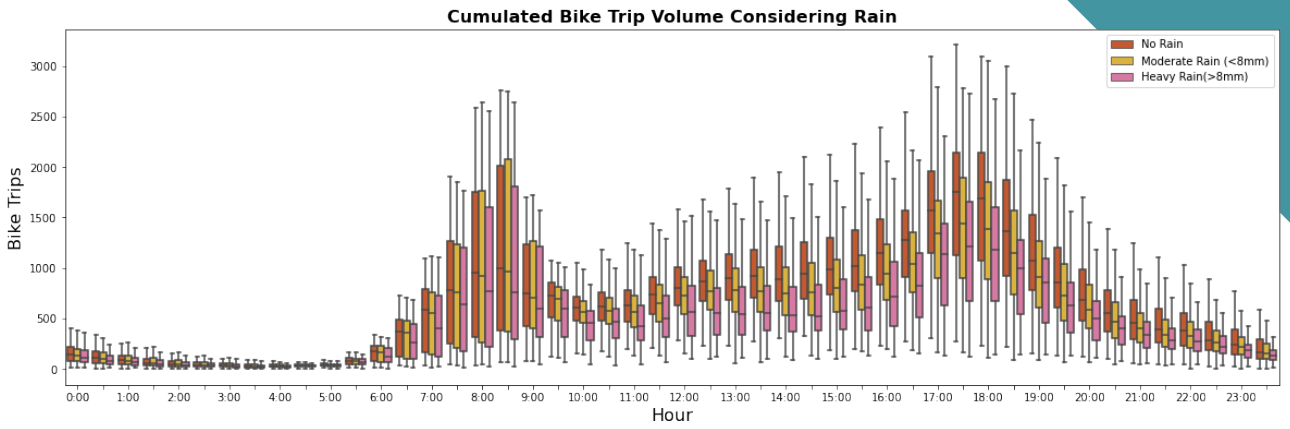


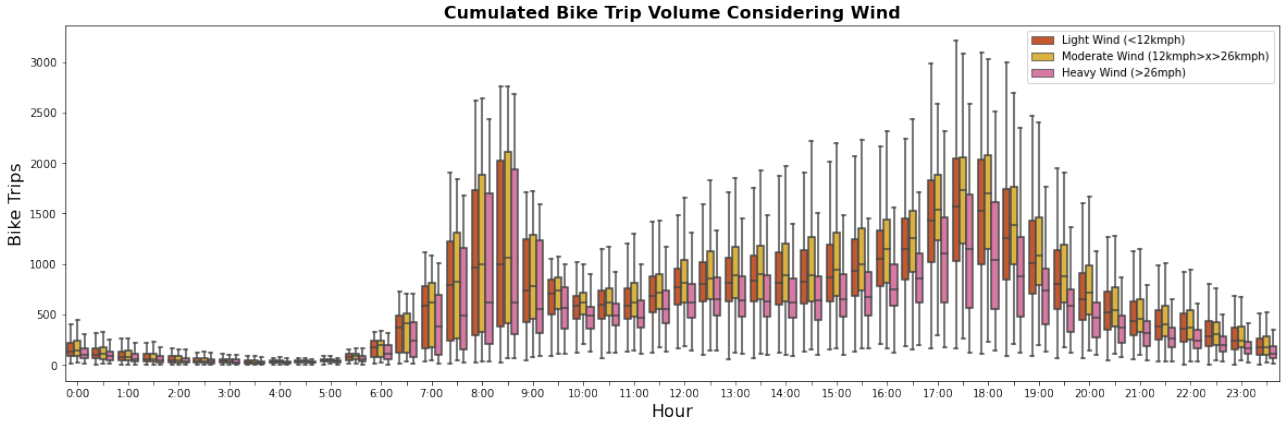
From: Aquatic Centre, Queen Elizabeth Olympic Park; To: Aquatic Centre, Queen Elizabeth Olympic Park
From: Hyde Park Corner, Hyde Park; To: Hyde Park Corner, Hyde Park
From: Black Lion Gate, Kensington Gardens; To: Black Lion Gate, Kensington Gardens
From: Park Lane , Hyde Park; To: Park Lane , Hyde Park
From: Albert Gate, Hyde Park; To: Albert Gate, Hyde Park
From: Triangle Car Park, Hyde Park; To: Triangle Car Park, Hyde Park
From: Podium, Queen Elizabeth Olympic Park; To: Podium, Queen Elizabeth Olympic Park
From: Timber Lodge, Queen Elizabeth Olympic Park; To: Timber Lodge, Queen Elizabeth Olympic Park
From: Serpentine Car Park, Hyde Park; To: Serpentine Car Park, Hyde Park
From: Queen's Circus, Battersea Park; To: Queen's Circus, Battersea Park
From: Wellington Arch, Hyde Park; To: Wellington Arch, Hyde Park
From: Palace Gate, Kensington Gardens; To: Palace Gate, Kensington Gardens
From: Aquatic Centre, Queen Elizabeth Olympic Park; To: Podium, Queen Elizabeth Olympic Park
From: Podium, Queen Elizabeth Olympic Park; To: Aquatic Centre, Queen Elizabeth Olympic Park
From: Lee Valley VeloPark, Queen Elizabeth Olympic Park; To: Lee Valley VeloPark, Queen Elizabeth Olympic Park
From: East Village, Queen Elizabeth Olympic Park; To: East Village, Queen Elizabeth Olympic Park
From: Hyde Park Corner, Hyde Park; To: Triangle Car Park, Hyde Park
From: Hyde Park Corner, Hyde Park; To: Albert Gate, Hyde Park
From: Hyde Park Corner, Hyde Park; To: Serpentine Car Park, Hyde Park
From: Triangle Car Park, Hyde Park; To: Hyde Park Corner, Hyde Park

Exploratory Data Analysis

What insights can we obtain from our data?

- In the top 20 most common routes, all of them have as origin or destination a park in London.
- During workdays, bikes are mostly rented at peak hours. At the weekend, bikes are rented following a pseudo-normal distribution with the centre at 14:30.
- During the Covid-19 lockdowns, bikes were no longer used for commuting, they were mostly used after 16:00.
- Bikes are more often used during days with no rain.
- Bikes are less often used during days with heavy wind (above 26 kmph).



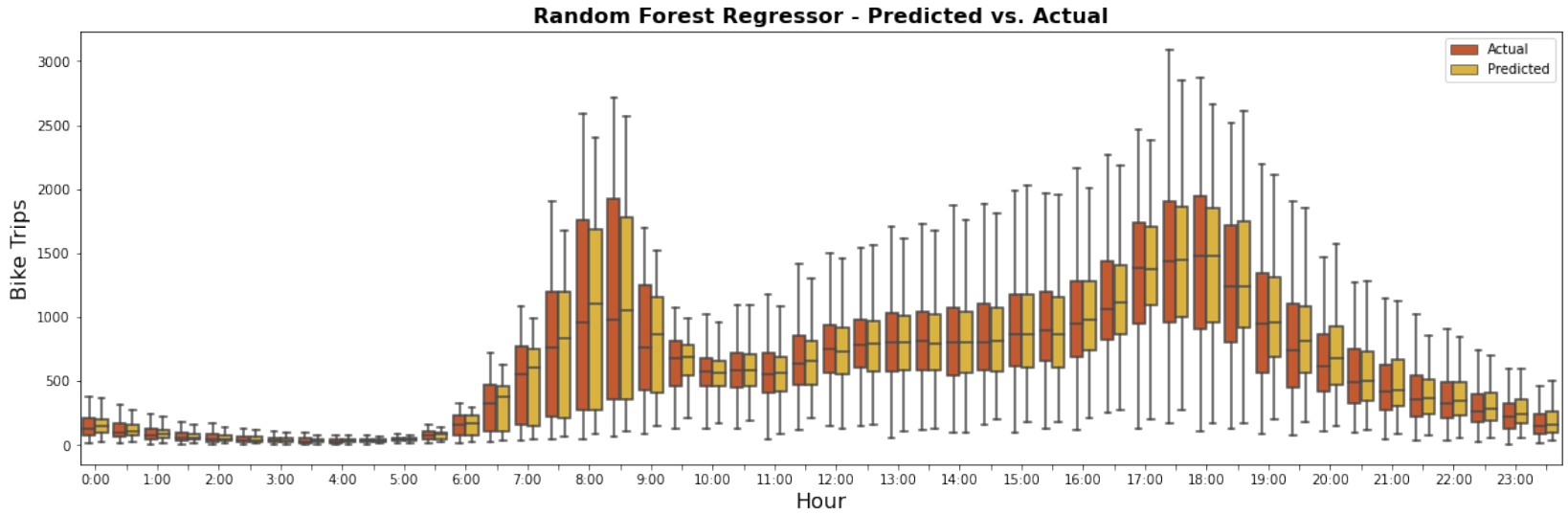


Modelling – Predicting Usage Volume

Methodology and Deployment

- Pipeline with two columns transformer as pre-processor. One-Hot Encoding for categorical data and Standard Scaler to standardise features removing mean and scaling to unit variance.
- Three different algorithms are trained and tested to select the best one:
  - Linear Regression
  - Support Vector Regressor
  - Random Forest Regressor

Model Type	Mean Absolute Error	Accuracy Score
Linear Regression	316.43	0.45523
Support Vector Regressor	273.83	0.46091
Random Forest Regressor	73.93	0.94616



Modelling – Predicting Usage Volume

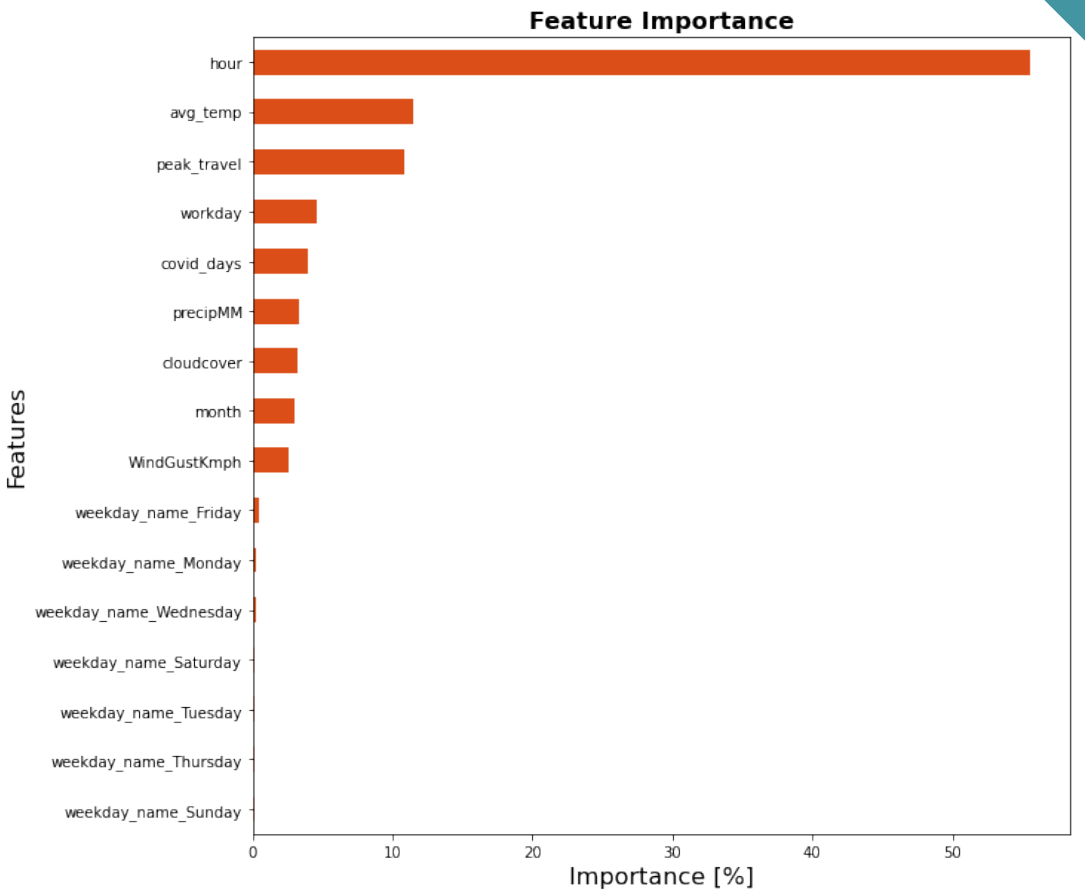
Performance and Features Importance

- Performing a Randomized Search, the best model parameters are:

  Accuracy of RFR= 0.94616 MSE = 17237.42

  Pipeline_MAE = 73.93

- The most relevant features to determine the number of bikes being rented:
  - Hour - 55.60%
  - Average Temperature - 11.50%
  - Peak Travel Hour - 10.85%
  - Workday - 4.55%
  - Covid Lockdown - 3.99%



Conclusions and Future Work

Recommendations

Rental Bikes Help Commuters

During workdays, bikes are mostly rented at peak hours. During weekends, however, they are mostly rented at central day hours.

Bad Weather equals Less Use

Bikes are mostly rented when there is no heavy wind or moderate/heavy rain.

Rental Bikes for Leisure around Parks

The top 20 most common routes all have as origin or destination a park/garden in London. During the Covid-19 lockdowns, bikes were mostly used after 16:00, after work as a means of leisure.

Other Data Science Use Cases

Predictive Bike Maintenance

Using historic data for every Bike ID to identify gaps where they were not used, possibly for being in maintenance.

Creation of an Interactive Map

Use the data to develop an interactive map where the user would be able to set routes selecting multiple stations.

Route Prediction for Journeys using OSRM

To explore the most common routes with the Open Source Routing Machine. Perform an analysis of the traffic depending on the hours.

Creation of a Dashboard

Use Apache Airflow to gather data from the different sources, create a Data Warehouse (Amazon EMR or Azure Blob). Use parquet to communicate with a Dashboard.