Analysis and Modelling of TfL Bikes Hires in London

Heiko Maerz, City, University of London

Jupyter Notebook — https://smcse.city.ac.uk/student/aczf462/INM430_Heiko_Maerz.html

1 Analysis Domain, Questions, Plan

1.1 Domain Overview

This paper aims to analyse bike trip data from the Transport for London (TfL) bike hire scheme (Santander Cycle Hire) in regard to time and weather and build a model to predict trip volumes. This data may be used for efficient planning of bike management.

Bikes can be used for commuting, leisure, or utility. For London O'Brien et al detect one weekend peak and two weekday commuter peaks [7].

Multiple papers explore the effects of temperature and time of day on the bike use. General findings report that rain, wind, and low temperatures in winter or high temperatures in summer decrease the use of bikes [1], [2]. Findings suggest a lower impact of weather for commute than for leisure use [4].

Kaggle hosts a similar analysis which models bike use for Montréal in 2015 depending on weather and day but does it per day and bike lane [5].

1.2 Data Sources

This analysis uses data from April 2017 to March 2018.

1.2.1 Bike Trip Data

Bike trip data for the Transport for London "Santander Cycle Hire" is published on the TfL data store starting in September 2015 [11]. The data is available for download as weekly .csv files containing unique bike rentals.

1.2.2 Weather Data

High frequency weather data is provided by the Met Office Weather Observations Website for a Hammersmith weather station and can be downloaded as monthly .csv files [6]. Additional weather observations were provided by Aidan Slingsby from City, University of London as a .csv file [10].

1.3 Analytical Questions

The main analytical question investigated in this paper is if bike trip volume for a given time period can be predicted using weather, date derived information, and time of day.

I will explore the accuracy of a number of machine learning algorithms for which I will use grid search for hyper-parameter optimization and cross-fold validation to prevent overfitting. The best model for each algorithm will be evaluated against test data hold-out. I will explore weights or feature importance of the attributes for each model.

I will use clustering to see if I can identify the patterns found by O'Brien and investigate how the prediction accuracy changes when applied to each cluster.

1.4 Analysis Strategy

This is the proposed process:

1. Data acquisition: download, transform, and aggregate bike trip and weather data.
2. Data wrangling: derive date and time dependent features [3], deal with missing values of the individual data sets, and merge all datasets.
3. Data exploration: visual analysis of bike trip volumes and attributes, correlation, and clusters.
4. Modelling.
5. Presentation of the results.

The algorithms for modelling were chosen based on the scikit learn flowchart [8].

For this report I will use

- Standard linear regression, and the generalised linear models lasso and elastic net.
- Support vector regression.
- Random forest regression as an example of ensemble methods.

2 Analysis

The details for each process step are documented in the linked Jupyter Notebook.

3 Findings and Reflections

3.1 Data

After loading, transformation, aggregation, and handling of missing data I have a merged dataset with 17486 observations which show total bike trip volume per 30-minute time periods with attributes temperature, rain (categorical), hour (decimal), month (numerical), workday (categorical), and peak travel (categorical) [Figure 1].

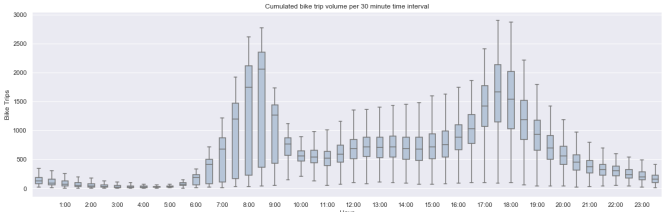

Figure 1 Bike Trip Volume per Day

Visual analysis confirms the influence of workday/ weekend [Figure 2], rain, temperature (see Notebook).

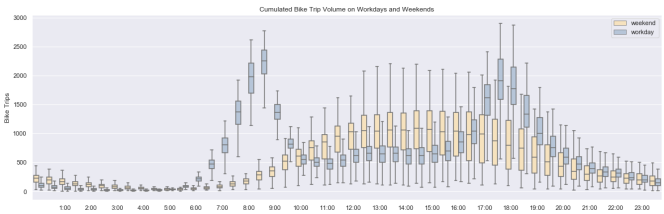

Figure 2 Weekend, Workday

feature, which reflects the non-linear distribution, but a very low p-value.

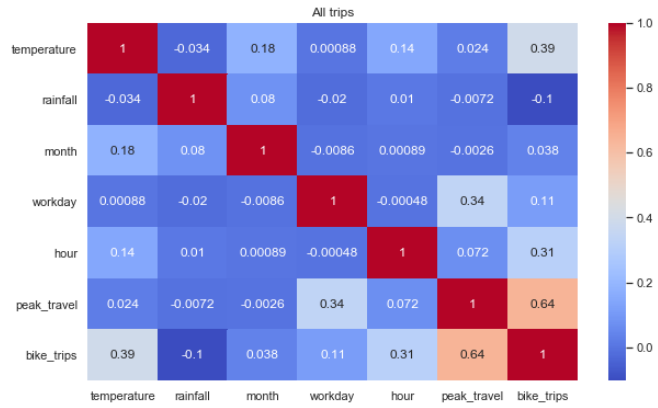

Figure 3 Correlation Matrix

Performing k-means clustering (k=3) resulted in three clusters which seem to reflect weekend use, workday commute and leisure use [Figure 4]:

Figure 4 Clustering

3.2 Model Results

Model / Data	Train	R2 Test	RMSE	MAE
Linear Regression	0.623	0.621	345.64	267.526
Lasso	0.623	0.621	345.642	267.486
Elastic Net	0.623	0.621	345.642	267.486
SVR	0.539	0.555	374.726	269.15
Random Forest	0.944	0.946	130.121	77.975
Lin Reg Cluster 0	0.459	0.472	338.912	278.944
Lasso Cluster 0	0.459	0.472	338.881	278.902
Elastic Net Cluster 0	0.459	0.472	338.881	278.902
SVR Cluster 0	0.428	0.442	348.695	276.335
Rnd Forest Cluster 0	0.942	0.951	103.398	67.774
Lin Reg Cluster 1	0.188	0.201	546.565	464.083
Lasso Cluster 1	0.188	0.201	546.557	464.106
Elastic Net Cluster 1	0.188	0.201	546.557	464.106
SVR Cluster 1	0.134	0.16	560.291	466.993
Rnd Forest Cluster 1	0.852	0.852	235.068	164.898
Lin Reg Cluster 2	0.511	0.501	207.083	173.07
Lasso Cluster 2	0.511	0.501	207.108	173.091
Elastic Net Cluster 2	0.511	0.501	207.108	173.091
SVR Cluster 2	0.495	0.476	212.318	174.21
Rnd Forest Cluster 2	0.93	0.932	76.668	48.373

The models received better scores when trained and tested against the full dataset rather than the individual clusters.

The random forest regressor outperforms all other models by a wide margin.

3.2.1 Linear Models

The linear regressors are not able to model the non-linear distribution of bike trip volumes. Visual analysis clearly shows how the predictions smooth observed data [Figure 5].

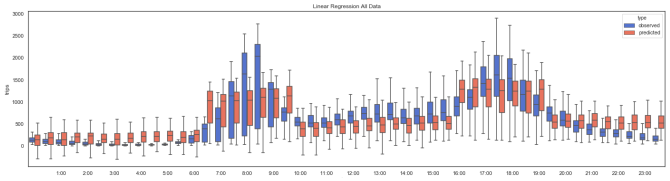

Figure 5 Linear Model Observed vs Predicted

QQ plots for residuals show normally distributed average values but show that the model will underpredict high volumes and overpredict low volumes [Figure 6].

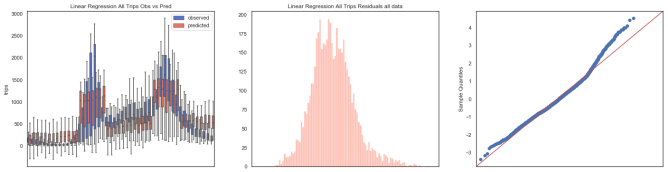

Figure 6 Residuals and QQ Plot

Linear regression, lasso regression, and elastic net have very similar values both for scores and regression coefficients. In fact, lasso and elastic net are the same [Table 2].

Model	Temp	Rain	Month	WorkD.	Hour	Peak
Linear Reg	29.7	-222.6	-1.8	-131.1	17.3	928.1
Lasso	29.7	-218.8	-1.8	-129.6	17.3	926.3
Elastic Net	29.7	-218.8	-1.8	-129.6	17.3	926.3
SVR	25.1	-118.0	-2.4	-14.7	21.8	571.9
Lin Reg Cl 0	39.7	-151.8	-5.5	0.0	19.2	0.0
Lasso Cl 0	39.7	-148.3	-5.5	0.0	19.2	0.0
El Net Cl 0	39.7	-148.3	-5.5	0.0	19.2	0.0
SVR Cl 0	32.7	-40.9	-6.5	0.0	19.5	0.0
Lin Reg Cl 1	34.3	-469.8	-1.5	0.0	7.3	0.0
Lasso Cl 1	34.3	-465.5	-1.5	0.0	7.3	0.0
El Net Cl 1	34.3	-465.5	-1.5	0.0	7.3	0.0
SVR Cl 1	39.1	-56.0	-4.1	0.0	13.6	0.0
Lin Reg Cl 2	20.7	-154.8	-1.6	0.0	18.4	0.0
Lasso Cl 2	20.7	-150.5	-1.6	0.0	18.4	0.0
El Net Cl 2	20.7	-150.5	-1.6	0.0	18.4	0.0
SVR Cl 2	18.6	-76.0	-2.3	0.0	22.4	0.0

Table 2 Linear Regression Coefficients

A look at the hyperparameters for lasso and elastic net confirms this. The best lasso model has an alpha = .2 (the lowest value passed for hyperparameter optimisation), which approximates linear regression. The best elastic net model equally has alpha = .2 and L1 ratio = 1.0, which approximates lasso regression [9].

3.2.2 Support Vector Regressor SVR

I have included the support vector regressor into the coefficient table because the kernel was passed as for hyperparameter optimisation, and the best model used the linear kernel, which then also returns regression coefficients. The derived coefficients differ from the linear models, and the overall score for SVR is not as good as any of the linear models.

3.2.3 Linear Models and Clusters

No linear model could properly estimate cluster data, specially the bimodal distribution for cluster 1. This is reflected in the poor scores [Figure 6].

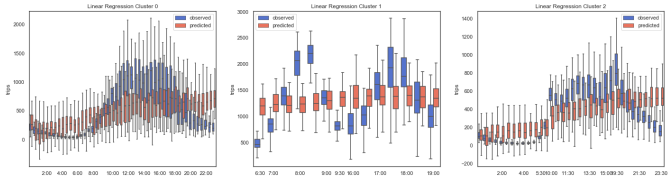

Figure 7 Observed vs. Predicted Clusters

Neither model used the attributes workday and peak travel, as they are the same within each cluster.

3.2.4 Random Forest Regressor

The random forest regressor consistently yielded the highest scores, both for all data [Figure 8] and clusters [Figure 9].

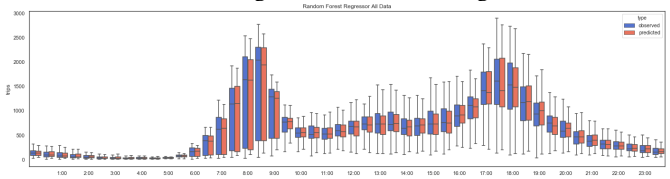

Figure 8 Observed vs Predicted All Data

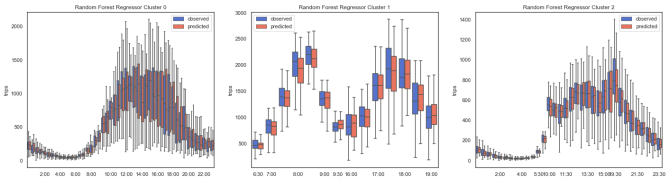

Figure 9 Observed vs. Predicted Clusters

QQ plots however suggest that this model predicts extreme values less well than average values [Figure 10].

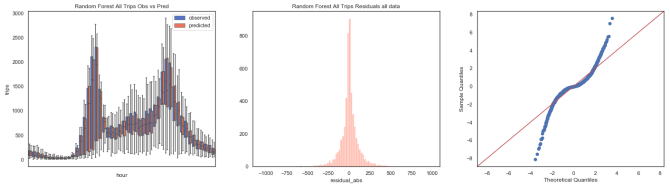

Figure 10 Residuals and QQ Plots

The hyperparameters for optimisation were number of trees (2, 5, 10, 20, 50, 100, 200) and the number of features to consider for best split (1 to 6). Grid search always returned the maximum number of trees for all datasets, and 2 (cluster 0) or 3 (all data, clusters 1 and 2).

Feature importance mirrors the other models with hour, peak travel, and temperature for all data and hour and temperature for the clusters.

Dataset	Feature	Ranking
All Data	hour	0.4071
All Data	peak_travel	0.3593
All Data	temperature	0.1570
All Data	month	0.0397
All Data	workday	0.0232
All Data	rainfall	0.0136
Cluster 0	hour	0.5596
Cluster 0	temperature	0.3376
Cluster 0	month	0.0896
Cluster 0	rainfall	0.0133
Cluster 0	workday	0.0000
Cluster 0	peak_travel	0.0000
Cluster 1	hour	0.6324
Cluster 1	temperature	0.2447
Cluster 1	month	0.0907
Cluster 1	rainfall	0.0321
Cluster 1	workday	0.0000
Cluster 1	peak_travel	0.0000
Cluster 2	hour	0.6941
Cluster 2	temperature	0.2381
Cluster 2	month	0.0504
Cluster 2	rainfall	0.0174
Cluster 2	workday	0.0000
Cluster 2	peak_travel	0.0000

ACKNOWLEDGMENTS

The author would like to thank Abi Sowri and Alex Galking for their feedback and advice, and Roger Brugge from the University of Reading, Mark Hadley, Head of Geography, City of London School, and Aidan Slingsby who assisted in the acquisition of weather data.

REFERENCES

[1] Caulfield, B., O'Mahony, M., Brazil, W., Weldon, P. 2017, "Examining usage patterns of a bike-sharing scheme in a medium sized city", Transportation Research Part A: Policy and Practice. 100, pp. 152-161.

[2] Flynn, B.S., Dana, G.S., Sears, J. & Aultman-Hall, L. 2011;2012;, "Weather factor impacts on commuting to work by bicycle", Preventive Medicine, vol. 54, no. 2, pp. 122-124.

[3] Gov.Uk (2018) "UK bank holidays", available at https://www.gov.uk/bank-holidays (Accessed at 3 November 2018)

[4] Helbich, M., Böcker, L. & Dijst, M. 2014, "Geographic heterogeneity in cycling under various weather conditions: evidence from Greater Rotterdam", Journal of Transport Geography, vol. 38, pp. 38-47.

[5] Kaggle (2018) "Bikes vs weather!", available at https://www.kaggle.com/rosemondeld/bikes-vs-weather (Accessed at 3 November 2018)

[6] Met Office (2018) "WeatherObservationsWebsite Observation Site Hammersmith", Available at

https://wow.metoffice.gov.uk/observations/details/20181101zcpxieg7hoe6tk8myyb96sc7my (Accessed: 1 November 2018)

[7] O'Brien, O., Cheshire, J. & Batty, M. 2014, "Mining bicycle sharing data for generating insights into sustainable transport systems", Journal of Transport Geography, vol. 34, pp. 262-273.

[8] scikit-learn (2018) "Choosing the right estimator", available at https://scikit-learn.org/stable/tutorial/machine_learning_map/ (Accessed at 3 November 2018)

[9] scikit-learn (2018) "Generalized Linear Models", available at https://scikit-learn.org/stable/modules/linear_model.html#elastic-net/ (Accessed at 15 November 2018)

[10] Slingsby, Aidan: Weather .zip file, link for download provided at 20 November 2018

[11] TfL (2018) "cycling.data.tfl.gov.uk", Available at https://cycling.data.tfl.gov.uk/ (Accessed: 1 November 2018)

[12] TfL (2018) "Peak and off-peak times", Available at https://tfl.gov.uk/fares-and-payments/fares/peak-and-off-peak-times (Accessed: 1 November 2018)