More Bikes: Experiments in Univariate Regression

Tim Lawson

December 21, 2023

1 Task description

The assignment is to predict the number of available bikes at 75 rental stations in three hours' time for a period of three months, beginning in November 2014, i.e., a supervised univariate regression problem. It is divided into three sub-tasks, which differ in the information that is available:

- **Sub-task 1.** The number of available bikes at each of the 75 stations for the month of October 2014. This sub-task may be approached by building a separate model for each station or a single model for all 75 stations.
- **Sub-task 2.** A set of linear models that were trained on the number of available bikes at each of a separate set of 200 stations for a year. For the first ten stations, this data is available for analysis but not training.
- **Sub-task 3.** Both of the above.

Sub-tasks 2 and, optionally, 3 require the use of ensemble methods. The predictions are evaluated by the mean absolute error (MAE) between the predicted and true numbers of available bikes over the period of three months, beginning in November 2014. The evaluation data is not available to participants but the score achieved on a held-out test set is reported on the task leaderboard. This report begins with a preliminary analysis of the data and then describes the approach taken to each sub-task and the cross-validation results obtained.

2 Data analysis

The data is recorded at hourly intervals; a summary of its features is given in fig. 1. The 'station' features are constant for all instances at a given station and the meteorological features are constant for all instances at a given timestamp. Hence, the variances of the 'station' features are zero for the first case of sub-task 1. The 'profile' features, i.e., the features derived from the numbers of available bikes at preceding times are not defined for the first week of instances at each station. Naturally, the number of bikes available at a given station is bounded by zero and the number of docks at that station. Additionally, the variance of the `precipitation` feature is zero for both cases of sub-task 1.

The following analysis is based on the combination of the available data for sub-task 1 and the first ten stations in sub-task 2. It is assumed that the distributions of the feature values are representative of those in the evaluation data.

2.1 Feature selection

Variance Zero-variance features were removed automatically from the data.

- Which features are most informative?
- Describe the distributions of the features and their correlations.

Category	Feature	Data type	Kind
Station	station	int	ordinal
	latitude	float	
	longitude	float	
	docks	int	
Temporal	timestamp	int	
	year	int	
	month	int	
	day	int	
	hour	int	
	weekday	str	ordinal
	weekhour	int	
	is_holiday	bool	categorical
Meteorological	wind_speed_max	float	
	wind_speed_avg	float	
	wind_direction	float	
	temperature	float	
	humidity	float	
	pressure	float	
	precipitation	float	
Bikes	wind_speed_max	int	
	bikes_3h	float	
	bikes_3h_diff_avg_full	float	
	bikes_avg_full	float	
	bikes_3h_diff_avg_short	float	
	bikes_avg_short	float	
	bikes	int	

Figure 1: A summary of the features of the data.

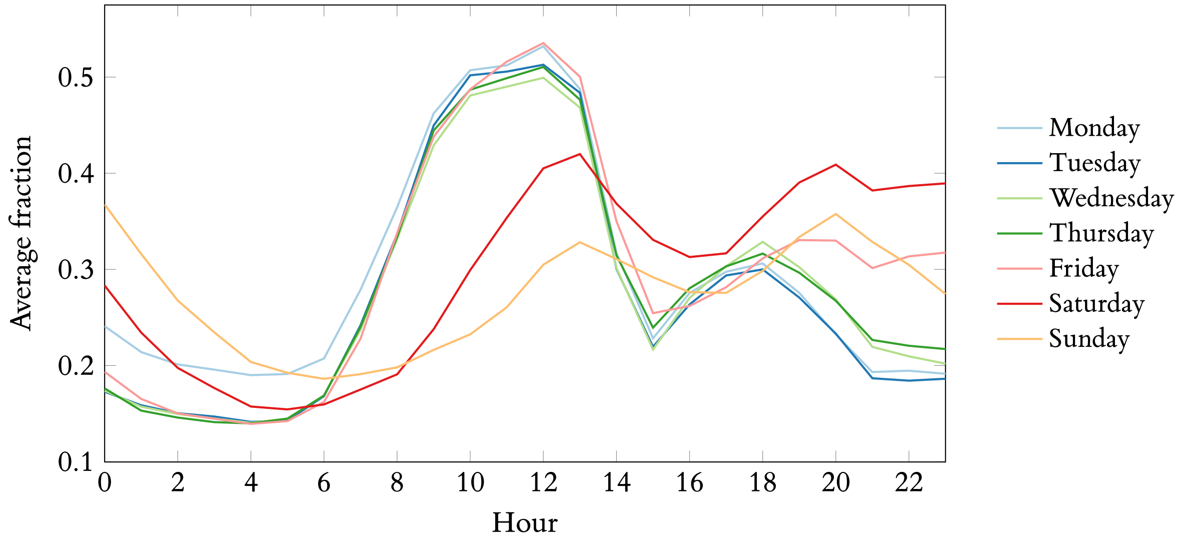

Figure 2: The average hourly fraction of available bikes on each day of the week. The distributions are generally bimodal, with peaks in the middle of the day and in the evening. A distinction between weekdays and weekends is also evident, with the exception of Friday evenings, where the distribution is more similar to that of a weekend day.

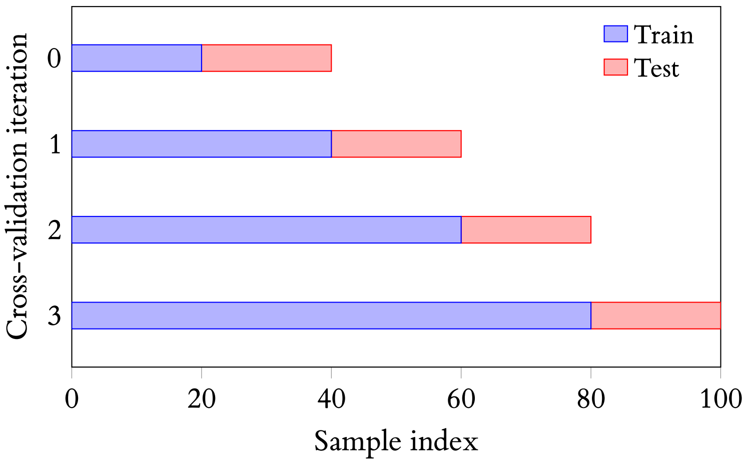

Figure 3: A visualisation of the cross-validation behaviour used throughout this report, after that of the `TimeSeriesSplit` function from `scikit-learn` (2023).

- Describe the distributions of the fraction of available bikes in terms of temporal features and the possible derived features.

3 Methodology

Throughout this report, ten-fold ‘forward chaining’ or nested cross-validation is performed, which is illustrated in fig. 3. Generally, standard k-fold cross-validation is disfavoured for time-series data due to the inherent correlation between successive folds (Bergmeir et al. 2018).

- Hyperparameter search.
- Evaluation metric.
- Statistical significance tests (paired t-tests).

References

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo (2018). “A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction”. In: *Computational Statistics & Data Analysis* 120, pp. 70–83.

Sklearn.Model_selection.TimeSeriesSplit (2023). URL: `https://scikit-learn/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html`.