

‘More Bikes’: Experiments in Univariate Regression

Tim Lawson

University of Bristol

tim.lawson@bristol.ac.uk

1 Task description

The assignment was organized as a Kaggle competition.¹ The task was to predict the number of available bikes at 75 rental stations in three hours’ time between November 2014 and January 2015, i.e., a supervised univariate regression problem. It is divided into three sub-tasks, which differ in the available information. For sub-task 1, the data from the 75 stations for October 2014 is provided (section 4). This sub-task may be approached by building (a) separate models for each station, and (b) a single model for all stations. For sub-task 2, a set of linear models that were trained on the data from a separate set of 200 stations is provided (section 5). Finally, for sub-task 3, both sources of information may be used. Additionally, the data from the first ten stations between June 2012 and October 2014 is provided for analysis (section 3).

The predictions were evaluated by the mean absolute error (MAE) between the predicted and true numbers of available bikes. Hereafter, the MAE is referred to as the ‘score’.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

The evaluation data was not made available to the competition participants, but the score achieved on a held-out test set was reported on the task leaderboard.² I give the score achieved on this test set by the best estimators for each model class alongside the mean scores on cross-validation folds of the data provided for sub-task 1 in tables 4 and 7.

2 Approach

I used the `scikit-learn` Python package (Pedregosa et al. 2011) throughout this report.³ Preprocessing and feature selection were performed by *estimators* that implemented the *transformer* interface, prediction was performed by estimators that implemented the *predictor* interface, and estimators were composed into *Pipeline* objects over which hyperparameter search was performed (Buitinck et al. 2013, pp. 4–9).

¹See <https://www.kaggle.com/c/morebikes2023/overview>.

²See <https://www.kaggle.com/c/morebikes2023/leaderboard>.

³The underlying code is available at [tslwn/more-bikes](https://github.com/tslwn/more-bikes).

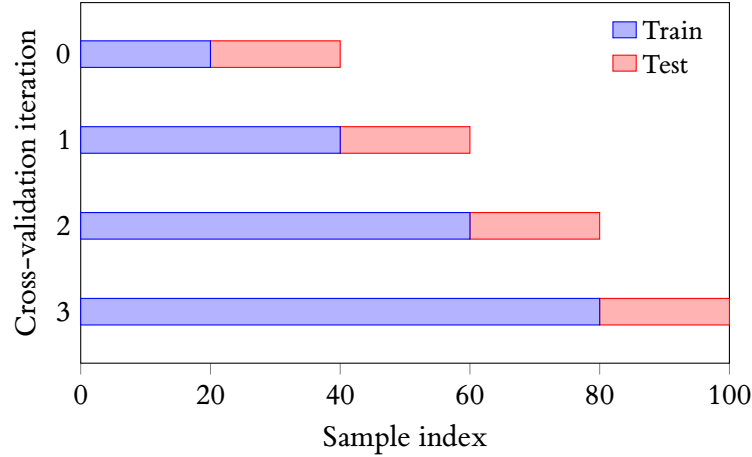


Figure 1: A visualization of the nested time-series cross-validation behaviour that I used, after the documentation for `sklearn.model_selection.TimeSeriesSplit`.

Generally, standard k -fold cross-validation is disfavoured for time-series data due to the inherent correlation between successive folds (Bergmeir et al. 2018). Instead, I used nested time-series cross-validation⁴ with ten folds, which is illustrated in fig. 1. I determined the best estimator for each model class by grid search⁵ with the mean absolute error as the scoring function, following the task description. Finally, I assessed the statistical significances of the differences between the mean scores of the best estimators by paired t -tests and Nemenyi tests on the scores on the cross-validation folds (Flach 2012, pp. 353–354). I describe these methods in more detail in sections 4 and 5.

3 Data analysis

The data provided for sub-task 1 is recorded at hourly intervals with $n = 54385$ instances across the 75 stations. Some features have a natural range – in particular, the number of available bikes at a station is bounded by zero and its number of docks (section 3.2). A summary of the features and the numbers of missing values for the non-temporal quantitative features are given in tables 1 and 2.

This analysis revealed that many of the features and the target variable bikes are missing for $n = 73$ instances, which were henceforth excluded. Additionally, 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 October 2014 at each station. Hence, they are missing for approximately $\frac{1}{4}$ of the instances. The meteorological features are constant for all stations at a given timestamp.

3.1 Feature selection

Intuitively, features that have zero variance are not informative, so I automatically excluded them.⁶ For the two cases of sub-task 1, the zero-variance features are indicated in table 1. Notably, the precipitation feature is zero for all instances.

⁴See `sklearn.model_selection.TimeSeriesSplit`.

⁵See `sklearn.model_selection.GridSearchCV` and `HalvingGridSearchCV`.

⁶See `sklearn.feature_selection.VarianceThreshold`.

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

Table 1: A summary of the features and the target variable (bikes), which have been renamed to be easier to read. Except where indicated, the features are quantitative. The zero-variance columns indicate whether the feature has zero variance for the cases of (a) separate models for each station, and (b) a single model for all stations.

Feature	n/a
wind_speed_avg	73
wind_direction	365
temperature	73
humidity	73
pressure	73
bikes	73
bikes_avg_full	12,264
bikes_avg_short	12,264
bikes_3h	292
bikes_3h_diff_avg_full	12,483
bikes_3h_diff_avg_short	12,483

Table 2: The numbers of missing values (n/a) of features and the target variable (bikes). Except where indicated, there are zero missing values.

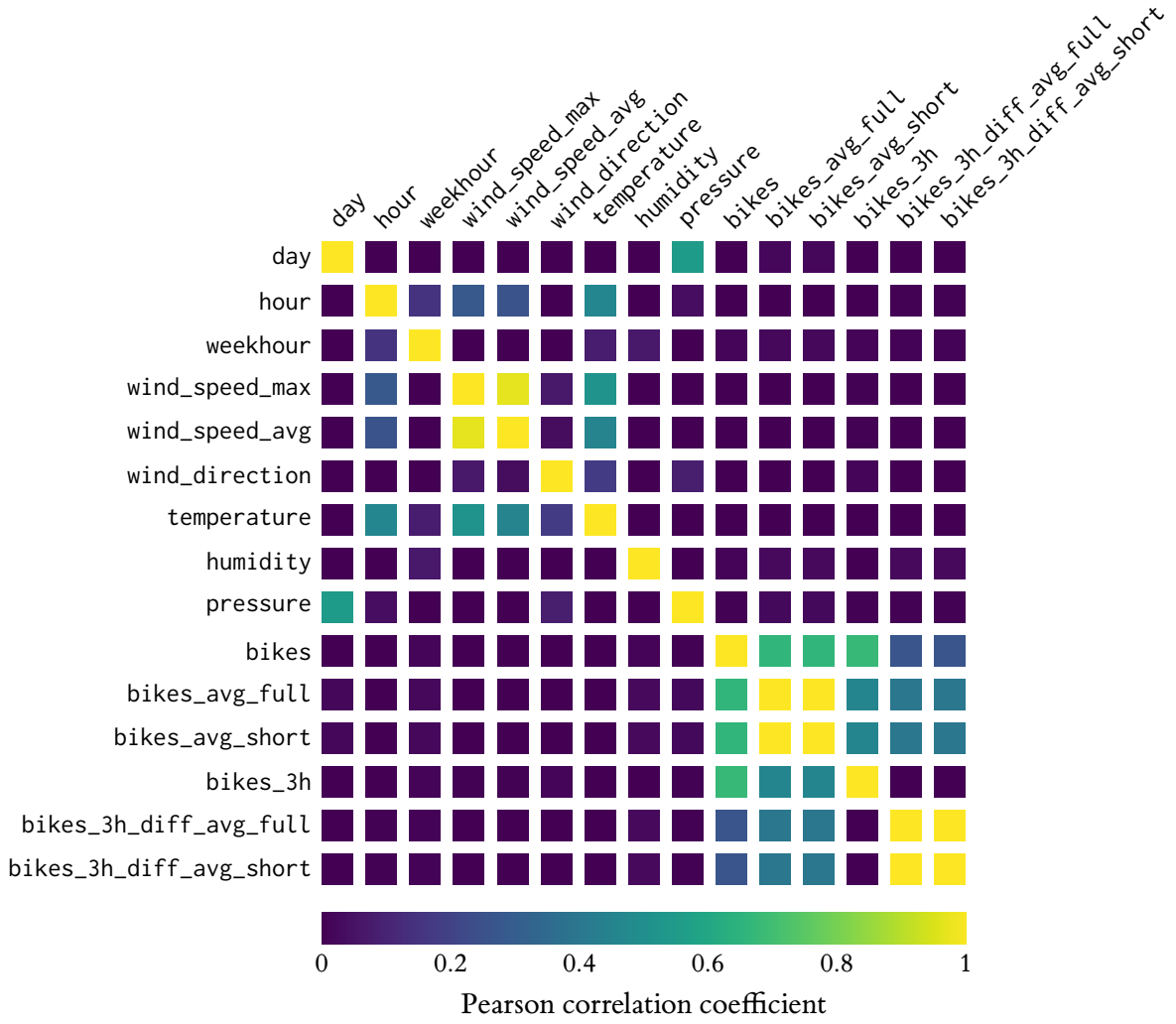


Figure 2: The Pearson correlation coefficients between pairs of the quantitative features and the target variable (bikes). The ordering follows table 1. Features that have zero variance for the case of (a) separate models for each station are excluded (table 1). The timestamp feature is also excluded because it is naturally correlated with the other temporal features.

Correlations between explanatory variables can cause a regression model to make unreliable predictions (Alin 2010). I identified these by computing the Pearson correlation coefficients between pairs of quantitative features (fig. 2), which yielded the following observations:

- bikes_avg_full and bikes_avg_short are perfectly correlated ($r = 1.00$).
- bikes_3h_diff_avg_full and bikes_3h_diff_avg_short are perfectly correlated ($r = 1.00$).
- wind_speed_max and wind_speed_avg are highly correlated ($r = 0.96$).

Hence, for sub-task 1, the second of each of these pairs of features was also excluded. For sub-task 2, the ‘profile’ features were retained because they are inputs to the pre-trained linear models (table 6).

3.2 Feature engineering

For sub-task 1, the bounds on the number of available bikes at each station were enforced by predicting the *fraction* of bikes, i.e., the number of bikes divided by the number of docks. I implemented this by extending the `TransformedTargetRegressor` meta-estimator to allow data-dependent transformations of the target variable.⁷ Another potential transformation of the target variable would have been to predict the *change* in the number of available bikes, but I did not explore this option.

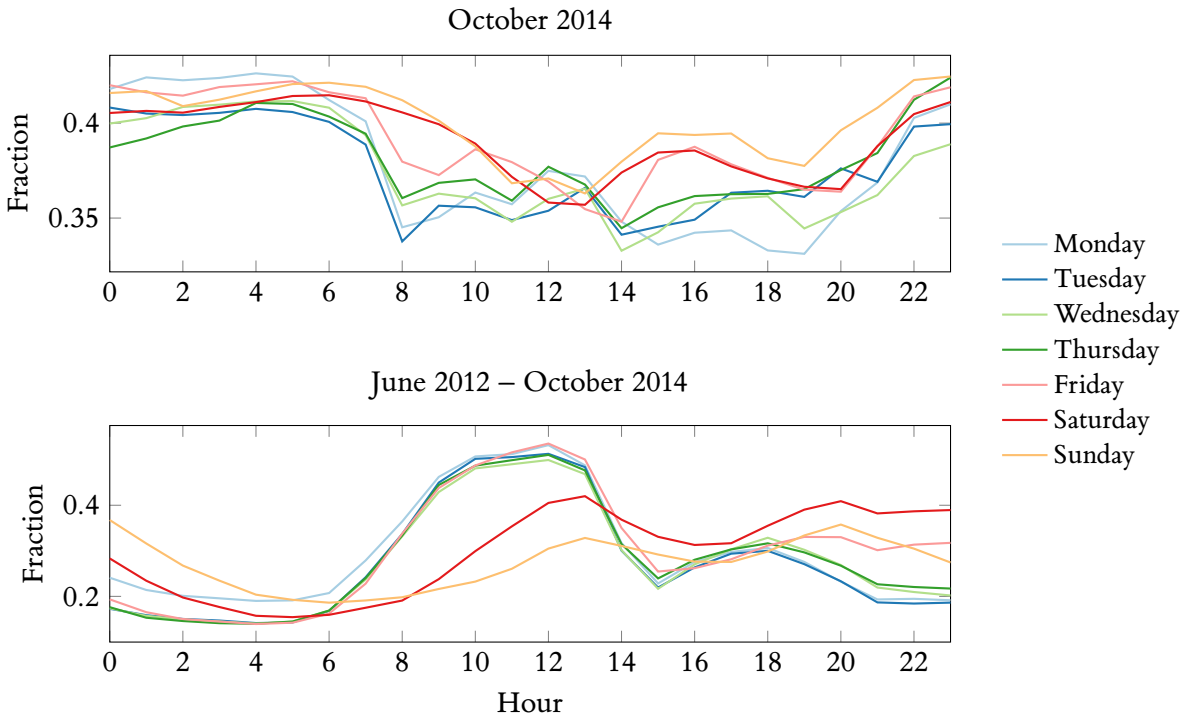


Figure 3: The average fraction of available bikes at each hour of the day, separated by the day of the week. In the top chart, the data is limited to the month of October 2014 for the 75 stations to predict. In the bottom chart, the data from the first ten stations is also included (section 1).

Initially, I investigated introducing derived temporal features, e.g., by discretizing the hour of the day into intervals. However, as shown in fig. 3, the average fraction of available bikes throughout the day differs substantially between the data from the 75 stations for October 2014 and the data from the first ten stations between June 2012 and October 2014 (section 1). Hence, I considered that a feature of this kind would be unlikely to generalize well to the test data. Additionally, because I chose to investigate tree models (sections 4.1 and 5.1), which partition the spaces of quantitative features (Flach 2012, p. 155), discretization may be irrelevant.

4 Sub-task 1

4.1 Model classes

Gradient-boosted decision trees are a popular choice of model class for time-series forecasting problems (Bojer and Meldgaard 2021). In particular, `scikit-learn` provides an optimized implementation of

⁷See `sklearn.compose.TransformedTargetRegressor`.

Parameter	Values
criterion	<u>absolute_error</u>
max_depth	<u>None</u> , 1, 2, 5, 10, 20, 50
max_leaf_nodes	None, <u>7</u> , 15, 31, 63
min_samples_leaf	1, 2, 5, 10, <u>20</u> , 50, 100

(a) Decision tree

Parameter	Values
loss	<u>absolute_error</u>
scoring	<u>neg_mean_absolute_error</u>
l2_regularization	0.1, <u>0.2</u> , 0.5, 1.0
learning_rate	0.02, 0.05, <u>0.1</u> , 0.2, 0.5
max_depth	None, 1, 2, 5, 10, 20, <u>50</u> , 100
max_iter	20, 50, <u>100</u> , 200, 500
max_leaf_nodes	None, 7, <u>15</u> , 31, 63
min_samples_leaf	1, <u>2</u> , 5, 10, 20, 50, 100

(b) Gradient-boosted decision tree

Table 3: The parameter values over which I performed grid search. Except where stated, the default values were used, and the parameters of the best estimator are underlined. For a description of the parameters and their default values, see `sklearn.tree.DecisionTreeRegressor` and `sklearn.ensemble.HistGradientBoostingRegressor`.

gradient-boosted decision trees inspired by LightGBM (Ke et al. 2017). An advantage of this model class is that it supports missing values, which are evident in the data (table 2). For both cases of sub-task 1, I elected to compare the performance of individual decision trees and gradient-boosted decision trees. As a baseline, I predicted the arithmetic mean of the fraction of available bikes.

The best estimator for each model class was determined by grid search with ten-fold cross-validation. I fixed scoring criteria to the mean absolute error, as per section 2. With a separate model for each of the 75 stations, I found that it was too computationally expensive to perform grid search over a wide range of hyperparameters. Hence, I performed grid search for the case of (b) a single model for all stations, then chose the values of the best estimator for the case of (a) a separate model for each station. I note that this approach biased the results in favour of case (b), because I did not optimize the hyperparameters for each station individually. The parameter values are listed in table 3.

4.2 Results

The mean scores achieved by the best estimators for each model class on the cross-validation folds and stations are listed in table 4. With a separate model for each station, there is greater variance in the mean score because there are 75 times as many cross-validation folds with commensurately fewer instances per fold. Intuitively, the baseline achieved a lower score with a separate model for each station, i.e., the average fraction of available bikes varies between stations. For both cases of sub-task 1, decision trees and gradient-boosted decision trees achieved a lower score than the baseline.

	Train		Test
	Mean score	Variance	Score
Baseline	4.43	3.82	4.47
Decision tree	3.53	3.52	<u>3.01</u>
GBDT	<u>3.37</u>	2.61	3.23

(a) Separate models for each station. The mean and variance are over the 75 stations and ten cross-validation folds.

	Train		Test
	Mean score	Variance	Score
Baseline	5.45	0.06	5.38
Decision tree	3.51	0.08	3.08
GBDT	<u>2.60</u>	0.08	<u>2.15</u>

(b) Single model for all stations. The mean and variance are over the ten cross-validation folds only.

Table 4: The mean scores and variances of the best estimators for each model class on the data provided for sub-task 1, and the corresponding score on the held-out test set (section 1). The best scores on each dataset are underlined.

I assessed whether the differences between the mean scores achieved by the best estimators were statistically significant by performing paired t -tests and the Nemenyi test on the scores on the cross-validation folds. The null hypothesis was that the differences between the mean scores were due to chance. I did not compare the scores between estimators for the different cases of this sub-task because the samples are different (Flach 2012, p. 354). With separate models for each station, I only performed paired t -tests on the scores, so the p -values are not corrected for multiple comparisons (table 5a).

Regardless, this indicated that the best decision trees and gradient-boosted decision trees performed better than the baseline for more than half of the stations. The best gradient-boosted decision trees only performed better than the best decision trees for seven stations. With a single model for all stations, I also performed the Nemenyi test to correct for multiple comparisons (table 5b). This indicated that only the best gradient-boosted decision tree performed significantly better than the baseline, but it did not perform significantly better than the best decision tree.

5 Sub-task 2

5.1 Model classes

The available information for sub-task 2 is a set of linear models that were trained on the data from a separate set of 200 stations. There are six models for each station, each of which uses a different set of features, listed in table 6. Unlike the gradient-boosted decision tree implementation in `scikit-learn`, these models do not support missing values, so I imputed the missing values of quantitative features with the mean value of the feature over the data provided for sub-task 1.

Model A	Model B	$t > 0$	$p < 0.05$
Baseline	Decision tree	70	39
Baseline	GBDT	70	44
Decision tree	GBDT	44	7

(a) Separate models for each station. The numbers of positive t -statistics and ‘significant’ p -values from paired t -tests on the mean scores of the best estimators for each model class on the ten cross-validation folds for each of the 75 stations. The p -values are not corrected for multiple comparisons, unlike in table 5b.

Model A	Model B	t -statistic	p -value	$p < 0.05$
Baseline	Decision tree	18.63	$6.53 \cdot 10^{-2}$	
Baseline	GBDT	31.63	$1.00 \cdot 10^{-3}$	✓
Decision tree	GBDT	8.19	$6.53 \cdot 10^{-2}$	

(b) Single model for all stations. The t -statistics from paired t -tests, and p -values from the Nemenyi test, on the mean scores of the best estimators for each model class on the ten cross-validation folds for all 75 stations.

Table 5: The results of statistical tests on the scores achieved by the best estimators for each model class. A positive t -statistic indicates that ‘Model B’ achieved a lower mean score than ‘Model A’.

I used `sklearn.ensemble.StackingRegressor` to combine the predictions of the models by stacked generalization (Wolpert 1992). This model class is a meta-estimator that trains a second-level regressor on the predictions of a set of first-level regressors – in this case, the pre-trained models.

Model	temperature	bikes_3h	bikes_avg_full	bikes_avg_short	bikes_3h_diff_avg_full	bikes_3h_diff_avg_short
r1m_full		✓	✓		✓	
r1m_full_temp	✓	✓	✓		✓	
r1m_short		✓		✓		✓
r1m_short_full		✓	✓	✓	✓	✓
r1m_short_full_temp	✓	✓	✓	✓	✓	✓
r1m_short_temp	✓	✓		✓		✓

Table 6: The features used by the different kinds of pre-trained linear models for sub-task 2. The features are described in table 1 and follow the same ordering.

For this sub-task, I excluded zero-variance and highly correlated features, but retained the ‘profile’ features (section 3.1). The `full` and `short` features are perfectly correlated, and the latter were excluded for sub-task 1, so I first compared the different kinds of pre-trained models. Figure 4 indicates that individual `r1m_short` and `r1m_short_temp` models generally achieved lower mean scores. However, with the default second-level regressor (ridge regression), stacked generalization performed very similarly with each kind of model, so I included all the models in the final analysis. For the second-level regressor, I evaluated the default, decision trees, and gradient-boosted decision trees, as in section 4.1.

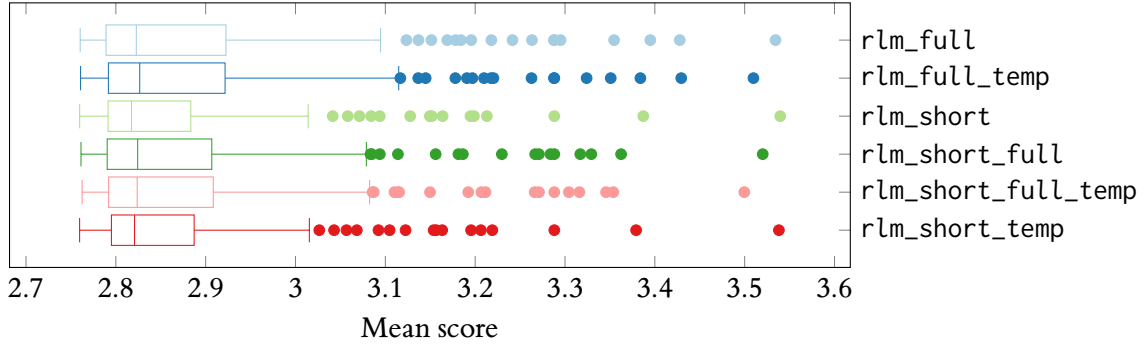


Figure 4: A box plot of the mean scores achieved by individual pre-trained models for sub-task 2 on the ten cross-validation folds of the data from the 75 stations for the month of October 2014, i.e., the same data as for sub-task 1 (b). The kinds of models are listed in table 6.

	Train		Test
	Mean score	Variance	Score
Ridge	2.85	0.13	2.35
Decision tree	3.79	0.10	4.53
GBDT	<u>2.84</u>	0.19	<u>2.25</u>

Table 7: The mean scores and variances achieved by the best estimators for each model class for sub-task 2 on the data provided for sub-task 1, and the corresponding score on the held-out test set (section 1). The best scores on each dataset are underlined.

5.2 Results

The mean scores achieved by the best estimators for each model class on the cross-validation folds are listed in table 7. All the model classes achieved a lower score than the baselines in sub-task 1. As in section 4, I performed paired t -tests and the Nemenyi test on the scores on the cross-validation folds to determine whether the differences between the mean scores achieved by the best estimators were statistically significant (table 8). Stacked generalization with ridge regression or a gradient-boosted decision tree as the second-level regressor performed significantly better than with a decision tree, but they did not perform significantly better than each other.

Model A	Model B	t -statistic	p -value	$p < 0.05$
Ridge	Decision tree	-20.7	$1.02 \cdot 10^{-2}$	✓
Ridge	GBDT	0.46	$6.32 \cdot 10^{-1}$	
Decision tree	GBDT	17.18	$1.00 \cdot 10^{-3}$	✓

Table 8: The t -statistics from paired t -tests, and p -values from the Nemenyi test, on the mean scores of the best estimators for sub-task 2. As in table 5, a positive t -statistic indicates that ‘Model B’ achieved a lower mean score than ‘Model A’.

Model A		Model B		t -statistic	p -value	$p < 0.05$
Sub-task	Model class	Sub-task	Model class			
1 (b)	Baseline	1 (b)	Decision tree	18.63	$1.60 \cdot 10^{-1}$	
1 (b)	Baseline	1 (b)	GBDT	31.63	$1.00 \cdot 10^{-3}$	✓
1 (b)	Baseline	2	Ridge	25.83	$1.82 \cdot 10^{-3}$	✓
1 (b)	Baseline	2	Decision tree	21.11	$6.81 \cdot 10^{-1}$	
1 (b)	Baseline	2	GBDT	21.38	$1.00 \cdot 10^{-3}$	✓
1 (b)	Decision tree	1 (b)	GBDT	8.19	$4.53 \cdot 10^{-3}$	✓
1 (b)	Decision tree	2	Ridge	5.41	$6.81 \cdot 10^{-1}$	
1 (b)	Decision tree	2	Decision tree	-2.14	$9.00 \cdot 10^{-1}$	
1 (b)	Decision tree	2	GBDT	4.54	$3.96 \cdot 10^{-1}$	
1 (b)	GBDT	2	Ridge	-7.29	$2.61 \cdot 10^{-1}$	
1 (b)	GBDT	2	Decision tree	-25.02	$1.00 \cdot 10^{-3}$	✓
1 (b)	GBDT	2	GBDT	-4.4	$5.43 \cdot 10^{-1}$	
2	Ridge	2	Decision tree	-20.7	$1.60 \cdot 10^{-1}$	
2	Ridge	2	GBDT	0.46	$9.00 \cdot 10^{-1}$	
2	Decision tree	2	GBDT	17.18	$4.74 \cdot 10^{-2}$	✓

Table 9: The t -statistics from paired t -tests, and p -values from the Nemenyi test, on the mean scores of the best estimators for sub-tasks 1 (b) and 2. As elsewhere, a positive t -statistic indicates that ‘Model B’ achieved a lower mean score than ‘Model A’.

6 Conclusion

Finally, I performed the same comparison for the best estimators for sub-task 1 (b) and sub-task 2 (table 9). The results were broadly consistent with those in tables 5b and 8, i.e., the following model classes performed significantly better than the baseline but not significantly better than each other:

- sub-task 1 (b), gradient-boosted decision tree;
- sub-task 2, ridge regression; and
- sub-task 2, gradient-boosted decision tree.

Among these, the first performed best on the held-out test set (tables 4b and 7). I investigated whether combining the best estimators would improve the score on the held-out test set, the results failed to improve on the best score of 2.15, which thus formed the final submission.

References

- Alin, Aylin (2010). “Multicollinearity”. In: *WIREs Computational Statistics* 2.3, pp. 370–374.
- 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.
- Bojer, Casper Solheim and Jens Peder Meldgaard (2021). “Kaggle Forecasting Competitions: An Overlooked Learning Opportunity”. In: *International Journal of Forecasting* 37.2, pp. 587–603.

- Buitinck, Lars, Gilles Louppe, and Mathieu Blondel (2013). “API Design for Machine Learning Software: Experiences from the scikit-learn Project”. In: *ECML/PKDD 2013 Workshop: Languages for Data Mining and Machine Learning*.
- Flach, Peter (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. 1st ed. Cambridge, UK: Cambridge University Press.
- Ke, Guolin et al. (2017). “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Ed. by I. Guyon et al. Vol. 30. Curran Associates.
- Pedregosa, Fabian et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *The Journal of Machine Learning Research* 12, pp. 2825–2830.
- Wolpert, David H. (1992). “Stacked Generalization”. In: *Neural Networks* 5.2, pp. 241–259.