More Bikes

Machine Learning Paradigms Coursework

Jack Burnett

Interactive AI CDT

2023

Contents

1 Introduction

This paper accompanies a submission to the 'Machine Learning Paradigms' Kaggle Competition[1], which replicates ECML/PKDD 2015's Model Reuse with Bike rental Station data (MoReBikeS) Discovery Challenge (Kull, Lachiche, and Martínez-Usó 2015). In the competition, participants are provided with information on 275 bike rental stations and must use this to predict the future availability of said stations. The competition is divided into three phases. Phase 1 focuses on developing models to predict the availability of 75 stations over three months; the previous month's data is used to train these models. Phase 2 focuses on the adaptive reuse of learnt knowledge by implementing linear models, trained on a year's data for 200 stations, to predict the stations in phase 1. Phase 3 focuses on participants combining the approaches of Phases 1 and 2 to achieve better performance within the task. Methods are evaluated using the Mean Absolute Error of the model on a test data set.

2 Previous Approaches

The top three submissions for the ECML / PKDD 2015 competition (Kull, Lachiche, and Martínez-Usó 2015) were submitted by H. Song and Flach (2015), Pan and Subramaniyan (2015), and Y. Chen and Flach (2015). Kashyap and Swastik's (2021) paper, which implemented linear regression for a similar task, is also of note.

H. Song and Flach (2015) used Model-Based Subgroup Discovery to complete phase 2. This approach reduces the model training time and improves the prediction performance of stations with limited data. It was found that using the quality measure Weighted Relative Negative Mean Absolute Error to find subgroups resulted, in most cases, in the best performance.

Pan and Subramaniyan (2015) approached the competition by implementing the ordinary least squares method, Poisson regression, and zero-inflated Poisson regression models for phase 1; they then reused the provided pre-trained linear models for phase 2, selecting appropriate models using Mean Absolute Error. Pan and Subramaniyan (2015) found that the linear model reuse strategy resulted in the best performance when tested.

Y. Chen and Flach (2015) implemented support vector regression with feature selection for phase 1. Y. Chen and Flach (2015) found that vectorisation of time features and normalisation of bike profiles resulted in the best regression models, stating that vectorisation is the key to increasing performance.

Kashyap and Swastik's (2021) approach to a similar problem, where bike rental demand needed to be predicted using time and environmental features, implemented linear regression with good performance. In this approach, categorical data was vectorised, and features were selected based on correlation with the target variable. The paper suggests implementing classification trees, K-Nearest-Neighbours, and random forest algorithms in future approaches.

[1] https://www.kaggle.com/competitions/morebikes2023

3 Chosen Approach and Motivations

Linear regression allowed for a simplistic and practical approach to this task (Kashyap and Swastik 2021; Pan and Subramaniyan 2015); therefore, the chosen strategy for phases 1 and 2 is to replicate the successes of previous attempts. Phase 1 is approached by implementing linear regression using feature selection; feature selection was implemented using a filter, wrapper, and embedded methods (Karagiannopoulos et al. 2004), with the results being compared. Phase 2 was implemented using the provided linear regression models and a selection method to identify which model to implement for each station; this method was compared against using cumulative models from the pre-trained models. Phase 3 was used to explore further models, evaluate the approaches from Phases 1 and 2, and identify additional models to be implemented. Python 3.12[2] was chosen as the development language for this task, primarily selected as it enables access to the scikit-learn[3] and pandas[4] libraries.

3.1 Strengths

As this approach initially builds on the successes of Kashyap and Swastik (2021) and Pan and Subramaniyan (2015), it allows the creation of effective baseline models in phases 1 and 2, to compare exploratory models with those in phase 3. Research (Bartol et al. 2022; Jiao et al. 2020) has found that linear regression can outperform state-of-the-art deep learning methods; therefore, linear regression should serve as a suitable baseline for comparing models.

3.2 Weaknesses

This approach does not build on the success of H. Song and Flach (2015); instead, this approach implements a less effective algorithm for phase 2. This approach is not novel and does not implement alternative algorithms suggested by previous research; as such, this research does not provide new information on model reuse or algorithm implementation within this domain.

3.3 Technical Background

All approaches implemented are linear models (Flach 2012); as such, the formula of each model can be generalised as $f(\mathbf{x}) = a + b_1x_1 + \ldots + b_dx_d = a + \mathbf{b} \cdot \mathbf{x}$, where a is the intercept, b is the slope, $\mathbf{x}$ is a vector and f is a scalar. Regression expresses the relationship between predictor and target variables using an equation, allowing approximation of the proper relationship between them (Montgomery, Peck, and Vining 2021). Feature selection improves the performance and cost-effectiveness of algorithms by selecting features that are most relevant in predicting a target variable (Guyon and Elisseeff 2003).

[2] https://www.python.org/downloads/release/python-3120/
[3] https://scikit-learn.org/stable/
[4] https://pandas.pydata.org/

4 Phase 1

4.1 Data Analysis

Phase 1 starts with data analysis and pre-processing. A summary of the dataframe shows that all initial values are numeric, aside from the weekday column; to resolve this, the weekday values were vectorised, and each vector value was stored in a new column within the dataframe. Following the findings of Y. Chen and Flach (2015), the values of the bike profiles were normalised. The bikes column was also normalised, using the number of docks so that the column represents the percentage of capacity filled rather than the raw values; this allows the relationship between numDocks and bikes to be better analysed; the bike values will need to be returned to true values after any predictions occur.

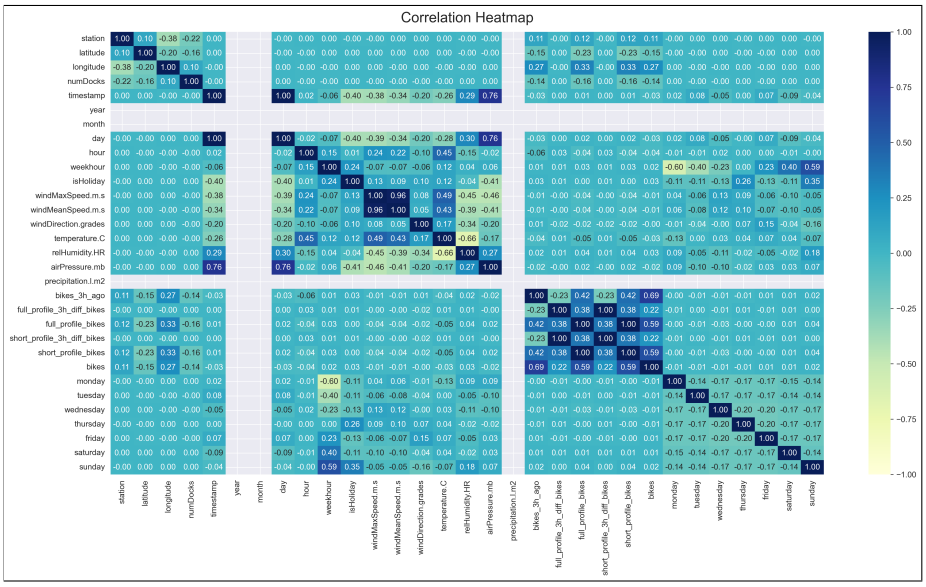

Figure 1: Correlation Matrix

After pre-processing, a correlation matrix, seen in figure 1, was produced to allow for data analysis. The correlation matrix shows that the features most correlated with bikes are the bike profiles and longitude. As latitude was associated with bikes, geographical data was explored through an interactive map; however, no clear pattern was visible; this is revisited in Phase 3. The correlation matrix aligns with previous findings of bike profiles being the best indicators of bikes (Pan and Subramaniyan 2015). Station and numDocks were weakly correlated with the bike column; as the numDock value is caused by the station it belongs to, this will be explored further by creating individual models for each station. A full review of the data analysis stage allowed for implementing the feature selection methods.

4.2 Implementation

Three approaches to feature selection were implemented: a naive approach that filtered using correlation values (Ibrahim, Nazir, and Velastin 2021), a feature wrapping method using forward elimination (Chowdhury and Turin 2020), and an embedded approach that implements LASSO (Gauraha 2018). A hybrid approach was also implemented, where a cut-off value can be selected for correlation and utilised alongside feature wrapping; this approach could either output the union of the two approaches or the intersection. These methods were each tested in the modelling phase. Using a cut-off of 0.5 correlation resulted in the features 'bikes 3h ago', 'full profile bikes', and 'short profile bikes' being selected; using feature wrapping resulted in the features 'bikes 3h ago', 'full profile 3h diff bikes', 'full profile bikes' being selected. These two methods show that 'bikes 3h ago' and 'full profile bikes' have the most significant impacts on the bike value. The feature selection method implemented a conservative selection of values by only accepting features with a performance gain of more than 0.01.

A simple linear regression model (Flach 2012) was fit to the data and used for predictions; how this model was fit varied based on the parameters passed to the modelling function. The modelling function allowed for the data approach type, being holistic or per-station, and the filtering method, using the feature selection methods, to be input; in total, there are ten possible variations for modelling, with the results of each approach shown in table 1.

	Filtering	**Wrapping**	**Combination**	Combination-intersect	LASSO
single	*2.65600*	***2.33777***	*2.36533*	*2.61244*	*5.37866*
separate	*5.45422*	*1286.77866*	*1294.01244*	*NaN*	*8.62933*

Table 1: Modelling results

4.3 Evaluation

From table 1, it can be seen that implementing feature wrapping and training on the entire data set had the best results. In all methods, using all stations for training resulted in better MAE of predictions than per-station training. Feature wrapping and the hybrid feature selection model performed best in the all-station approach but had significantly poor performance in the per-station approach. From the code outputs analysis, this was due to the feature wrapping method either selecting only one feature or the majority of features when applied to the per-station approach. Due to the lack of selected features, the hybrid intersect method could not be used in per-station modelling. The feature filtering method performed consistently across the two approaches. The consistency of feature filtering is likely because the technique does not rely on a large amount of training data to calculate features for selection; instead, it applies a simple cut-off point for correlation with the target value.

5 Phase 2

5.1 Implementation

Phase 2 required the reuse of pre-trained models to predict unseen stations. To implement this, a model loader converted the pre-trained models into a suitable format, and a custom model creator implemented the pre-trained linear regression models. Two approaches for model reuse were implemented: a model selection method that iterates over every pre-trained model and every station to identify the best fit for each station and a cumulative model that utilises the average of all model values. Both approaches were implemented using each pre-trained model type, as there were five variations of pre-trained models; the results of this can be seen in table 2.

	Cumulative full	**Cumulative full-temp**	**Cumulative short**	Cumulative short-temp	Cumulative short-full	Cumulative short-full-temp	Cumulative specific
Results	*3.52444*	*4.54844*	*12.27466*	*13.72177*	*3.77777*	4.62222	**2.23377**

Table 2: Phase 2 results

5.2 Evaluation

Table 2 shows that when using a cumulative model, the full profile of a bike has the best impact on performance; this aligns with the findings of the correlation matrix and feature selection of phase 1. Selecting a model for each station had the best performance in this phase, while creating a model for each station had the worst performance in phase 1; this is likely due to overfitting caused by the creation of station-specific models, which selection can mitigate. Utilising the temperature within the models hurt performance, as did implementing the short-term profile of bikes. The findings of this phase show that the full-time profile of bikes is the best predictor of future usage; this aligns with research that identifies seasonal changes as a critical indicator for bike usage (P. Chen, Zhou, and Sun 2017). Using these findings, an improved model may be possible using model selection and feature wrapping.

6 Phase 3

Phase 3 was an exploratory phase in which the researcher attempted to merge the best methods from phases 1 and 2, further explored location data, analysed subgroups, and implemented further models for testing. Within this phase, no performance improvements were made during the exploratory sections; however, models with better performance were identified. This phase serves primarily as a basis for further work and alternative approaches to the task were it to be re-attempted.

6.1 Merging Best Methods

In an attempt to enhance the performance of the models from phases 1 and 2, a hybrid approach was introduced; this approach used the model selector to identify the best model for each station, as it did in phase 2, then compared this to the cumulative feature-wrapping model from phase 1. All models predicted the bike values for a station and then implemented the model with the best R2 score; this may have resulted in overfitting in some stations. The resulting MAE on the test data was 2.36, slightly worse than the individual methods; this is likely due to the method allowing for overfitting to occur, thus implementing the weaknesses from both methods. This method should be re-explored with a new model evaluation approach.

6.2 Exploring Location

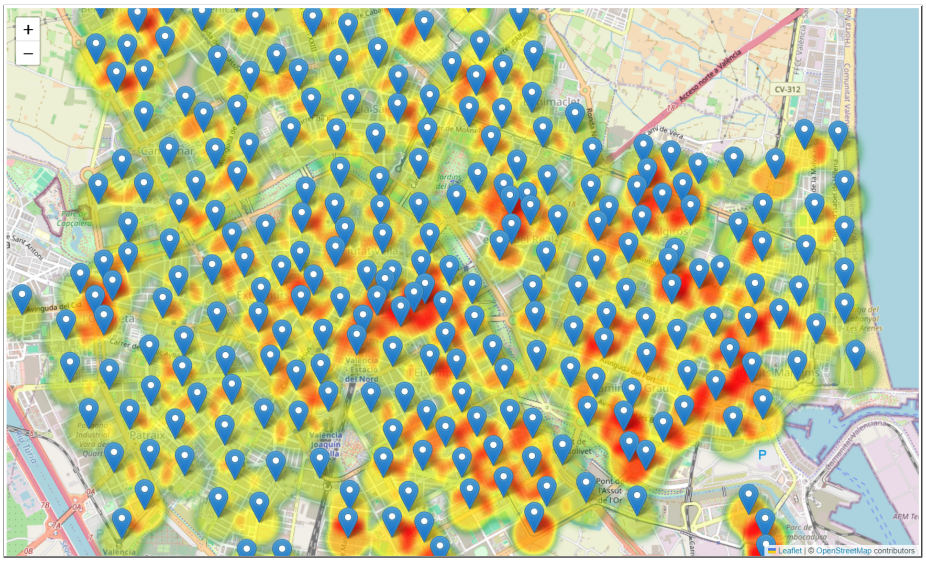

Figure 2: Station Heatmap

Latitude and longitude were weakly correlated with the target variable; as such, this is an area that was explored. Figure 2 shows a heatmap of the stations in Valencia. Visual analysis shows that stations with high bike usage are in dense clusters, and stations with less usage are more distanced. KMeans clustering (Jin and Han 2010) could be implemented to group the stations, which can then be used to identify correlations between bikes and groups; however, the randomness of KMeans would make this metric unreliable. To ensure the reproducibility of outcomes, three methods were developed for location-based metrics: the distance to the centre of all stations, the average distance to the nearest k stations, and stations within x kilometres.

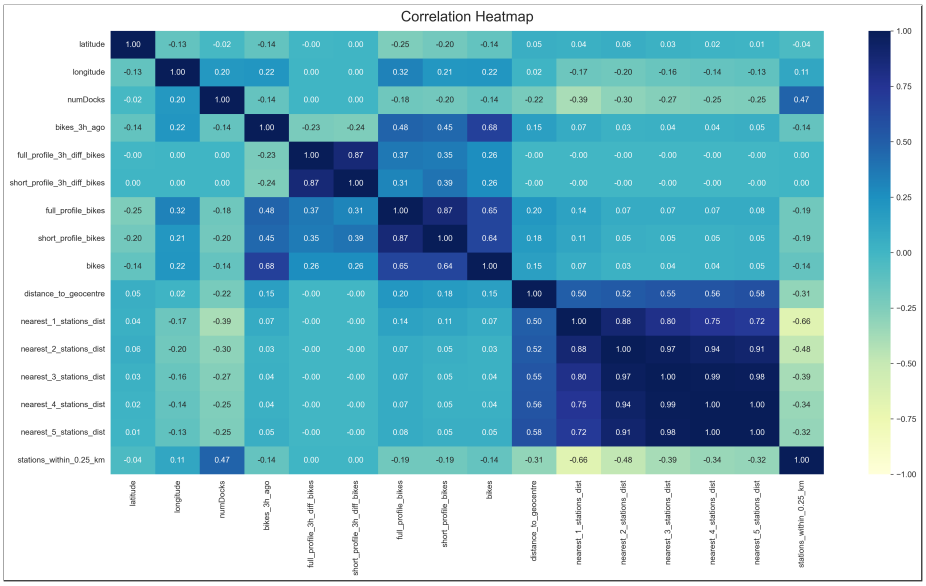

Figure 3: Location Metrics Correlation Matrix

The metrics utilised the World Geodetic System[5] to calculate distances as accurately as possible. Figure 3 depicts the correlation matrix for location metrics; from the correlation matrix, it can be seen that the new metrics are all weakly correlated with bikes. Longitude is still the location feature most correlated with bikes. Executing feature wrapping with the new metrics still results in the features 'bikes 3h ago', 'full profile 3h diff bikes', 'full profile bikes', and 'short profile 3h diff bikes' being selected; therefore, these metrics do not have a significant effect on predicting the number of bikes for a station. Further exploration into the longitude correlation could yield more meaningful results; the longitude correlation is likely due to the Gulf of Valencia being the east of the stations, with stations closer to the coast and centre having higher usage. Location metrics regarding landmarks and amenities may be significant.

6.3 Exploring Subgroups

Subgroups were created using a decision tree regressor (Y.-Y. Song and Lu 2015); these subgroups were then output as a tree diagram; however, due to a large number of features, this was computationally expensive and resulted in a graph that did not allow for easy visual analysis. A method for inferring subgroups from the tree regressor was also implemented but yielded the same outcomes. Subgroups were analysed using Orange[6] but produced no tangible results without further computations.

[5] https://gisgeography.com/wgs84-world-geodetic-system/
[6] https://orangedatamining.com/

6.4 Implementing Further Models

	RandomForest	**GradientBoostingRegressor**	**Ridge**	SVR
Results	*3.14222*	***2.16266***	*75.78044*	*5.35377*

Table 3: Further model exploration results

A method was developed to allow further models to be fit to the data and tested. The results in table 3 show that implementing a Gradient Boosting Regressor yields better performance than all methods implemented within phases 1 and 2. Implementing this model with feature wrapping delivered worse performance, with an MAE of 2.27288. Future attempts at this competition would benefit from Gradient Boosting Regressor implementation and exploration.

7 Conclusion

Within this paper, methods to predict future usage of bikes for rental stations in Valence were analysed. The two most successful methods implemented were a linear regression model trained on short-term data using feature wrapping and a linear regression model using pre-trained models and model selection; both models had an MAE of 2.33777 when tested on unseen data. Fitting a model to each station individually, using short-term data, yielded the worst results. It was found that the profile of a bike station's usage over the long term was the best predictor of future use.

Further metrics were analysed within the paper, focusing on location metrics. While these metrics did not yield significant results, they began to understand the patterns in the data regarding longitude. Further models were also evaluated, with findings indicating that gradient-boosted regression is the most effective model for this task.

This paper identified that model reuse of pre-trained can yield significant performance, parallel with methods of creating models from data while being efficient computationally. Model reuse may reduce the carbon footprint of artificial intelligence (Dhar 2020); as such, it is an area that should be explored with further discovery challenges.

Further studies using this challenge that may give more insight into model reuse would be identifying if models can reused across data sets, for example, through implementing the pre-trained Valencia models to predict the Seoul bike rental dataset Choi et al. 2023.

8 Bibliography

Bartol, Kristijan et al. (2022). "Linear Regression vs. Deep Learning: A Simple Yet Effective Baseline for Human Body Measurement". In: *Sensors* 22.5. ISSN: 1424-8220. DOI: 10.3390/s22051885. URL: https://www.mdpi.com/1424-8220/22/5/1885.

Chen, Peng, Jiangping Zhou, and Feiyang Sun (2017). "Built environment determinants of bicycle volume: A longitudinal analysis". In: *Journal of Transport and Land Use* 10.1, pp. 655–674. ISSN: 19387849. URL: http://www.jstor.org/stable/26211749 (visited on 01/10/2024).

Chen, Yu and Peter Flach (2015). "SVR-based Modelling for the MoReBikeS Challenge: Analysis, Visualisation and Prediction". In: *Proceedings of the ECML/PKDD 2015 Discovery Challenges*.

Choi, Seung Jun et al. (2023). "Combatting the mismatch: Modeling bike-sharing rental and return machine learning classification forecast in Seoul, South Korea". In: *Journal of Transport Geography* 109, p. 103587. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2023.103587. URL: https://www.sciencedirect.com/science/article/pii/S0966692323000595.

Chowdhury, Mohammad Ziaul Islam and Tanvir C Turin (Feb. 2020). "Variable selection strategies and its importance in clinical prediction modelling". en. In: *Fam Med Community Health* 8.1, e000262.

Dhar, Payal (Aug. 2020). "The carbon impact of artificial intelligence". In: *Nature Machine Intelligence* 2.8, pp. 423–425. ISSN: 2522-5839. DOI: 10.1038/s42256-020-0219-9. URL: https://doi.org/10.1038/s42256-020-0219-9.

Flach, Peter (2012). "Linear models". In: *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, pp. 194–230.

Gauraha, Niharika (Apr. 2018). "Introduction to the LASSO". In: *Resonance* 23.4, pp. 439–464. ISSN: 0973-712X. DOI: 10.1007/s12045-018-0635-x. URL: https://doi.org/10.1007/s12045-018-0635-x.

Guyon, Isabelle and André Elisseeff (2003). "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar, pp. 1157–1182.

Ibrahim, Sara, Saima Nazir, and Sergio A Velastin (Oct. 2021). "Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis". en. In: *J Imaging* 7.11.

Jiao, Shuming et al. (Feb. 2020). "Does deep learning always outperform simple linear regression in optical imaging?" In: *Opt. Express* 28.3, pp. 3717–3731. DOI: 10.1364/OE.382319. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-28-3-3717.

Jin, Xin and Jiawei Han (2010). "K-Means Clustering". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, pp. 563–564. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_425. URL: https://doi.org/10.1007/978-0-387-30164-8_425.

Karagiannopoulos, M et al. (2004). "Feature selection for regression problems". In: *Educational Software Development Laboratory, Department of Mathematics, University of Patras, Greece.*

Kashyap, Aditya Singh and Swastika Swastik (2021). "Regression Model to Predict Bike Sharing Demand". In: *Int. J. Innov. Sci. Res. Technol.,* 6.3, pp. 1024–1028.

Kull, Meelis, Nicolas Lachiche, and Adolfo Martínez-Usó (2015). "MoReBikeS - Model reuse with bike rental station data". In: *Proceedings of the ECML/PKDD 2015 Discovery Challenges.*

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Pan, Rong and Arun Bala Subramaniyan (2015). "Prediction of Bike Rental using Model Reuse Strategy". In: *Proceedings of the ECML/PKDD 2015 Discovery Challenges.*

Song, Hao and Peter Flach (2015). "Model Reuse with Subgroup Discovery". In: *Proceedings of the ECML/PKDD 2015 Discovery Challenges.*

Song, Yan-Yan and Ying Lu (Apr. 2015). "Decision tree methods: applications for classification and prediction". en. In: *Shanghai Arch Psychiatry* 27.2, pp. 130–135.

9 Appendix

The code for this paper can be accessed via DataLore[7]. Jupyter Notebook[8] was used to develop the methods discussed in this paper, as it allows for shareable and reproducible results. The notebook was executed on a cloud-based virtual machine with 4GB RAM.

[7] datalore.jetbrains.com/report/static/xpFJ0NI0hLRUqjTY7AiDhW/gJfzOP9X0rRh0OeSLzSlyr
[8] https://jupyter.org/