Autoritas participation at MoReBikeS: Model Reuse with Bike rental Station data Technical Report

Andrés Ramos[1], Francisco Rangel[1]

[1]Autoritas Consulting, S.A., Spain
[andres.ramos,francisco.rangel]@autoritas.es

Abstract. Bicycle rental is a service that besides providing people with the possibility of improving their mobility, it may have a positive impact to current issues such as traffic jams and pollution. In this vein the MoReBikeS challenge was organised. In this paper we explain the participation of Autoritas in this challenge, where we obtained the 8th. position in the ranking with a MAE equal to 2.556.

1 Introduction

Bicycle rental is a service that is being implanted all over the world. Besides the public service given to people to improve their mobility, it may have a positive impact in current issues such as traffic jams or pollution. But bicycles are continuously taken from and returned to rental stations across the city. To predict the availability of bicycles in the different stations is a key objective. In this line, the "MoReBikeS: Model Reuse with Bike rental Station data" challenge has been organised. The challenge is carried out in the framework of historical bicycle rental data obtained from Valencia, Spain. The data consists of time series describing hourly availability of bikes at each station; information on weather and (local) holidays is also provided. The challenge is to, given models trained with historical data, to make predictions with 3 hours a head with regard to the number of bikes available for the new stations. We have approached the task with a distance-based algorithm obtaining the 8th position in the ranking with a mean absolute error of 2.556.

The rest of the paper is structured as follows. In Section 2 the dataset and provided models are described. In Section 3 we describe our approach and results are presented in Section 4. Finally, in Section 5 some conclusions are drawn.

2 Evaluation Framework

In the following sections, the dataset provided by the task organisation is described and the evaluation measure employed is presented.

2.1 Dataset and Models

The organisation provides with models learned from 200 training stations in the period between June 2012 and September 2014. A set of 75 stations is provided to test the predictions made with the previous models. This set of 275 stations covers October 2014. The test data contains 75 stations from the period between November 2014 and January 2015. The data contain 4 station features, 8 time features, 7 weather features, 1 task-specific feature and 4 profile features. The target is the median number of available bikes during the corresponding hour in this rental station. A depth explanation can be seen in the download website[1].

2.2 Evaluation Measures

The measure used to evaluate the quality of predictions is the mean absolute error (MAE). MAE measures how close predictions are from the actual values. Concretely, MAE is an average of the absolute errors between predictions and real values. The MAE is calculated with Equation 1.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\widehat{p_i} - x_i| \qquad (1)$$

Where $\widehat{p_i}$ is the predicted value and x_i is the real value, both for the station i, and n is the total number of stations.

3 Methodology

The method we have proposed for prediction is based on distances among bike stations. Our intuition was that nearby stations may have similar occupation rates and thus it may help predicting their availability. The provided data contains the longitude and latitude for each station. With this information, we generate a table with distances among them. Stations with ID between 1 and 200 are the most reliable because they have been modelled on a set of historical data belonging to two years. We use these stations to predict the availability of bikes in new stations provided with the test set. Our method uses information about the station, hour and weekday. In the following subsections the algorithm is described and an improvement based on holidays information is presented. We developed the algorithm in R[2].

3.1 A Distance-based Approach

For each station in the test set, stations located in a distance below 700 meters with respect to it are selected (e.g. to predict the station number 210, the stations 3, 6 and 8 fit the below 700 meters distance condition. Hence they are selected to make the prediction). Whether there is no station that satisfies the previous condition, we expand the distance to 1000 meters.

Once the closest stations are identified, we select the data that match both hour and weekday. All the provided models are applied to this data for the selected stations and the models with lowest MAEs are selected (e.g. to predict availability in

[1] http://reframe-d2k.org/Challenge_Download

[2] https://github.com/autoritas/RD-Lab/tree/master/doc/projects/MoReBikeS

station 210, all the models are applied to stations 3, 6 and 8, for example in a subset from Wednesdays at 12.30p.m. The best model for each station in terms of MAE is selected. In this case, concretely they are `model_base_ago`; `model_base_fprof` and; `model_base_sprof` respectively). Finally, the prediction for the given station is calculated as the average prediction for each nearby station.

3.2 Holidays-based Tunning

We think that the previous algorithm may be improved by taking into account special days such as Sundays and holidays. In this vein we changed the algorithm to obtain predictions for Sundays and holidays of the nearest stations when the date to predict coincides with some of this special days. The rest of the conditions remain unaltered.

4 Experimental Results

We participated as the *arp* team and achieved the 8th. position in the final ranking[3] obtained on the small test data with a mean absolute error equal to 2.556. In Figure 1, all the results are presented.

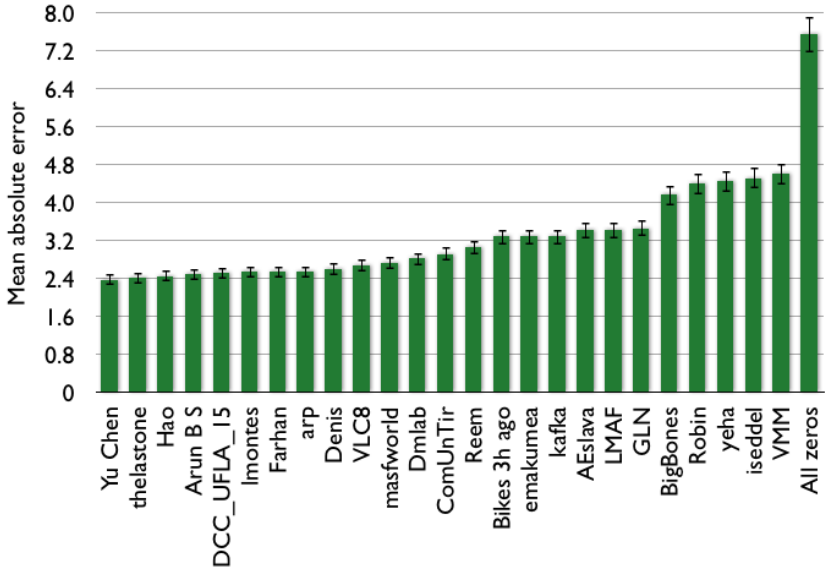

Fig. 1. Final MAEs of all submissions on small test data.

As can be seen there are three groups of results and an extreme outlier: *i)* results below the *Bikes 3h* ago baseline with MAEs below 3.2; *ii)* results between this baseline and *BigBones* with MAEs about 3.2 and; *iii)* results over *BigBones* with MAEs over 4.

[3] http://reframe-d2k.org/Challenge_Leaderboards

The outlier is the *All zeros* baseline which obtained a MAE over 7.5. It is noteworthy that, given the percentage of error (see error lines in Figure 1), the proposed approach might obtain best results when testing with a larger test dataset. In Figure 2, the distribution of MAEs of all the submissions on the small test is shown. As can be seen, there are three outliers in the lowest bound of the distribution. Concretely, the three worst results obtained MAEs over 6.5. The mean MAE is 3.3137 with a standard deviation of 0.9815. The median is equal to 2.9170 and the first quantile is 2.5585. Our MAE is below the first quantile, hence our approach is statistically among the best approaches.

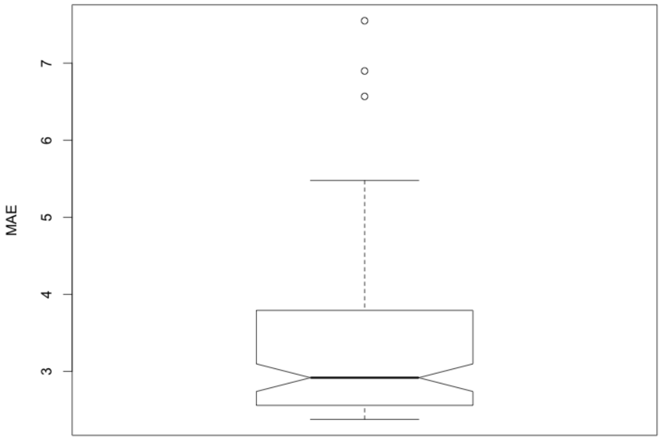

Fig. 2. Distribution of MAEs in the final leaderboard of all submissions on small test data.

The prediction with the modified algorithm which takes into account holidays was sent in the third submission. But contrary to what we may expect, the MAE obtained was a little worse (2.558). We consider that this algorithm may obtain better predictions when executed with the full dataset.

5 Conclusions

We have participated in the MoReBikeS challenge obtaining competitive results with a distance-based approach. We tried to improve results taking into account holidays, but we obtained a little worse results. We think this may be due to the fact that the test data only contained information from October 2014. We believe the improvement will be higher with a higher dataset. In future work we plan to apply machine learning to obtain the distance threshold and the weights for nearby stations.

ACKNOWLEDGEMENTS

This work was partially funded by Autoritas Consulting SA and by Ministerio de Economia de España under grant ECOPORTUNITY IPT-2012-1220-n430000.