Using smartphone data to predict Beijing bike-sharing demand

By: Terry Liu and Leonhard Fricke
Advisor: Dr. Inma Borrella

Topic Areas: Forecasting, Database Analytics, Tracking & Tracing

Summary: In response to increasing urbanization, China seeks alternative public transportation methods, such as bike-sharing, which has demonstrated social and environmental benefits. Our sponsoring company TalkingData collects bike-sharing usage data via smartphones. We investigated a one-month sample of data TalkingData collected from bike-sharing operators in Beijing merging it with data from online resources. We found that bike-sharing activity varies across the city Beijing throughout the day while time and environmental related factors significantly affect the bike-sharing demand. Our study revealed that some factors stated in literature such as pollution do not affect bike-sharing demand in Beijing significantly. We suggest that drivers of bike-sharing demand differ across cities or countries making it worthwhile to perform location specific analysis. We fitted linear regressions, neural networks and random forests on the compiled dataset and compared their respective performance. We found that, based on the one-month sample, linear regression performs best amongst the three models in predicting hourly bike-sharing demand in Beijing.

Before coming to MIT, Terry worked as a Global Business Development Director at TalkingData Limited. She holds a B.S. in Economics from University of Warwick, UK.

Before coming to MIT, Leonhard worked as a Consultant for BearingPoint GmbH He holds a B.S. in Industrial Engineering from Technical University of Brunswick and a M.S. in Management from European School of Management and Technology, Germany

KEY INSIGHTS

1. Time and environmental related factors significantly affect the bike-sharing demand
2. Drivers of bike-sharing demand differ across cities and countries
3. For a small sample size linear regression performs best amongst random forest and neural networks in predicting hourly bike-sharing demand

Introduction

High population densities observed in many Chinese cities and the growing motorization due to China's economic expansion, both result in high traffic congestion, parking inefficiencies and environmental challenges. This has led to increased interest in sustainable transportation alternatives, such as bike-sharing. The demand of such bike-sharing services, however, is affected by a variety of factors. These factors range from individual characteristics, over societal norms, to physical infrastructure and environmental factors.

Project Context

The sponsoring company of this research project, TalkingData, who is also China's largest independent data platform, collects app-usage of bike-sharing smartphone applications. Working with the data to identify drivers of bike-sharing demand and apply advanced forecasting models to predict demand allows TalkingData serving its clients active in bike-sharing.

We worked with TalkingData to:

- assess the adequacy of mobile data for generating insights into bike-sharing,
- collect additional primary and secondary data and connect it with mobile data,
- determine factors that drive bike-sharing demand in Beijing, and
- build predictive models for bike-sharing in Beijing and evaluate their suitability for mobile data.

We then conducted interviews with bike-sharing stakeholders and investigated a one-month sample of data that TalkingData collected from bike-sharing operators in Beijing merging it with secondary data from online resources.

Bike Sharing Demand Drivers

Overall TalkingData provided us with 1,215,894 single observations over a period of one month (11pm on 05/28/2017 to 9am 06/25/2017), being collected from five different bike-sharing providers from the Beijing area. The level of bike-sharing activity is fluctuating across the area throughout the day (Figure1).

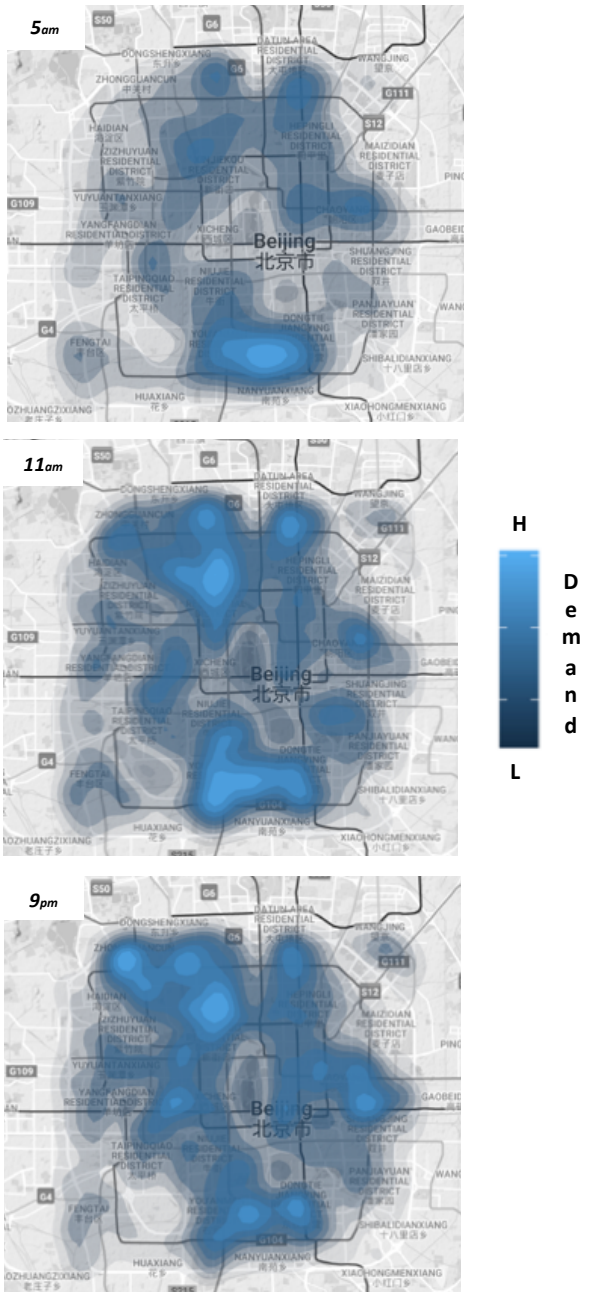

Figure 1: Demand at 5am, 11am and 9pm

While it seems that in the early morning a great demand is located in the south of Beijing, it transitions towards the north just before noon while two main demand locations (north and south) co-exist in the evening. We analyze the relation between time and demand in greater detail and found that the bike-sharing service is greatly used as part of the transportation going to and leaving from work during the rush hours in the morning and evening. We found that the demand pattern, including rush hour, is reflected throughout all seven days of the week. However, after exploring location and time as potential drivers of bike-sharing demand, we find that there is a lot of variability still unexplained, meaning that other factors beyond location and time explain other parts of the bike-sharing demand variability.

To identify the main drivers of bike-sharing demand, given our dataset, we use multi linear regression models with the demand being the dependent variable and several independent variables (such as temperature, pollution, etc.). We chose the linear model, in order to identify the dependent variables linear dependency on the independent variables (i.e. the direct impact). To identify which factors statistically significantly influence the demand, the p-value (in different levels) is taken into account (*p<0.1, **p<0.05, ***p<0.01).

From both the interviews and the bike-sharing demand data set we find that extreme temperatures have an impact of bike-sharing demand, so that a very high temperature lowers the bike-sharing demand. We obtained similar findings for wind and rain. Interviewees as well as the data show that with increasing wind speed and rain (humidity) the demand significantly decreases. Interestingly, we do not obtain consistent results for air quality (pollution). While employees of the bike-sharing service providers and the urban researcher state that the level of pollution would not affect bike-sharing demand. In contrast, bike-sharing users indicate that they would not use the service in case of bad air quality. From the data, we retain the hypothesis that pollution is not affecting bike-sharing demand. Hence, either the interviewed bike-sharing users are not representative for the entire set of bike-sharing users, or people believe that they would not use the service in case of a bad air quality but are unaware when the pollution level rises.

After better understanding the drivers of bike-sharing demand in Beijing by comparing the insights generated from both, interviews and quantitative analysis, we now focus on forecasting models as tools to predict bike-sharing demand in Beijing.

Forecasting Hourly Bike-Sharing Demand

We fit basic models of regression, neural network and random forest to the dataset provided by TalkingData. All of these represent supervised learning algorithms and allow for predicting a numerical value. This means that we fit a function based on pairs consisting of input (independent variables) and an output variable (dependent variable). The inferred function is then applied on new independent variables in order to predict the dependent variable.

To get a more reliable comparison, we fit the models using two different procedures. On the one hand, we fit the models on all observations, forecast the demand and compare it with actual numbers. On the other hand, we divide the area in the seven city districts Chaoyang, Chongwen, Dongcheng, Fengtai, Haidian, Xicheng and Xuanwu. We then fit the models to each respective district, forecast the demand in each, merge the results and compare it with actual numbers. The reasoning is that districts in Beijing differ by certain characteristics such as industry focus, economic activity and demographics, so that a more specific model could perform better. We fit the models and compare their accuracy using data set partitioning (i.e. training and validation set, both on district and city level) and k-fold cross validation (on city level).

For the regression model, we use bike sharing demand as dependent variable and hour, weekday, temperature, temperature^2, humidity, pressure, wind speed, wind speed^2 and air quality as independent variables. For the neural network, we pre-process the data and scale the numerical predictors and the outcome variable to a 0-1 scale. We also convert the categorical predictors to dummies. We then use bike-sharing demand as outcome variable and 34 inputs as a result of our pre-processing (factors, etc.). We fitted several different neural networks with the varying layers and nodes to the data, ultimately ending up using 2 nodes and 1 hidden layer as best fit. We build a basic random forest model using the same dependent and independent variables as in the regression model. We see that when the tree number is greater than 75, the error becomes more stable, and when bigger than 350, the error is very stable. Thus, we set the tree number of the random forest to 350 in this project.

To better be able to compare the models, we plot the predicted values of the linear regression, neural network and random forest model in comparison to each other in Figure 2. Observations on (or close to) the line indicate a good prediction.

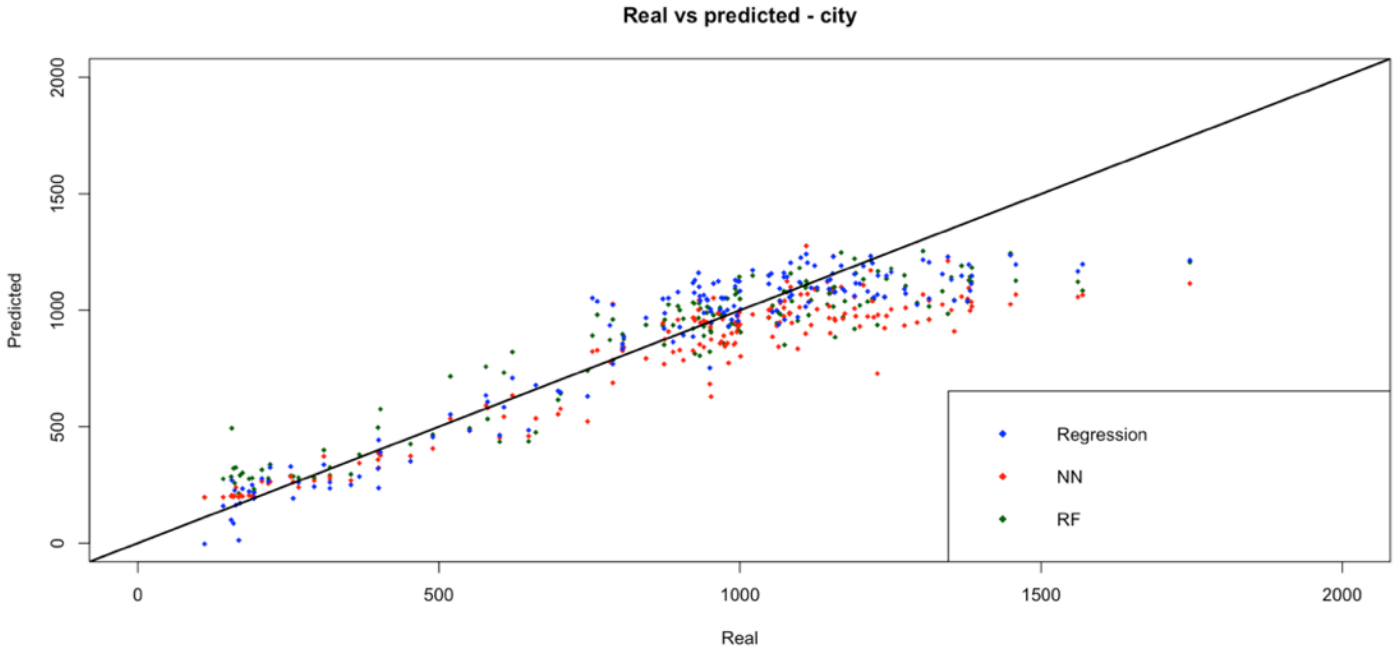

Figure 2: Real vs. predicted demand - city level

We can see that the neural network seems to perform worst amongst the respective models. The performance difference between the linear regression and random forest models seem to be less. Both of them show a distribution of predicted values and observations which are to some extent comparable to a “line”, meaning they tend to be more accurate than the neural network. It is noticeable that the extreme values (i.e. higher demand) are predicted less accurately across all forecasting models.

To further address which model performs most accurate, the RMSE of the partitioning models (both district and city level) and the cross-validation models are taken into account (Figure 3).

	Fitting Approach		
Algorithm	District Level	City Level	Cross Validation
Linear Regression	133.51	133.51	138.93
Neural Network	176.99	153.03	221.38
Random Forest	146.28	138.64	143.52

Figure 3: Forecasting models' RMSE

As already indicated, the RMSE of both the regression and random forest are smaller than the RMSE of the neural networks. We found that the regression model performs best amongst all three models in all three fitting approaches. Hence, the regression model is recommended for further use as a forecasting model based on the data at hand.

However, the models are fitted with a relatively small number of observations in terms of sufficiency for machine learning algorithms. While the regression model works comparably best in the current setting, we expect the performance of random forest and neural networks to increase when fitted on a larger dataset (for instance twelve months of data).

Conclusion

Bike-sharing providers have undoubtedly enhanced user convenience and reduced travel time. Our main objective in this project was to analyze smartphone data to understand the drivers of bike-sharing demand in Beijing and to use advanced forecasting methods as tool to predict demand. We provided TalkingData with a tool to enhance bike sharing's performance by understanding customer needs and behavior leveraging data collected during trips, as well as about external conditions. We recommend using regression models to forecast the demand when only a small data set is available. However, it would be valuable to further analyze the performance of other models such as neural networks and random forest if the data set available is larger.

In order to gain more robust results, the time-frame of the data should be extended (e.g. 1 year or 2 years of data). It would be possible to gain an understanding on the impact of seasonality, growth development, etc. Our analysis on demand drivers also could be run on specific districts or only for certain bike-sharing service providers to see whether they differ from the average Beijing customer, so that both districts and bike-sharing service providers could better understand the demand drivers of their respective area or customer base.

Additionally, to gain further insights on the individual behavior, the dataset could be enriched with personal information, such as age, sex or new vs. old users. This would allow to further segment the customer base and target customers more specifically. Another possibility would be to add further variables to the model that are more holistic, such as gasoline price or government policies. Since the car-sharing follows the free-floating principle, one could map specific locations such as public transport stations, etc. and find their impact on personal behavior.