CS260: Project Proposal

Peipei Zhou 204176631, Manna Lin 704434171

November 2015

1 Motivation

Bike sharing is a new form of sustainable public mobility. There are more than 500 bike sharing systems around the world currently. A common issue in bike sharing systems falls on strategy improving and operation planing to optimize the uses of bike sharing. A recent competition on Kaggle "Bike Sharing Demand" provides us valuable data for predicting and data mining. In this project, our goal is to predict the hourly bike rental demand in the Capital Bikeshare program in Washington, D.C., as shown in Fig. 1. We will utilize multiple machine learning techniques to construct effective models for predicting bike sharing demand. We can also compare different machine learning models in this case and analyze which one is more powerful.

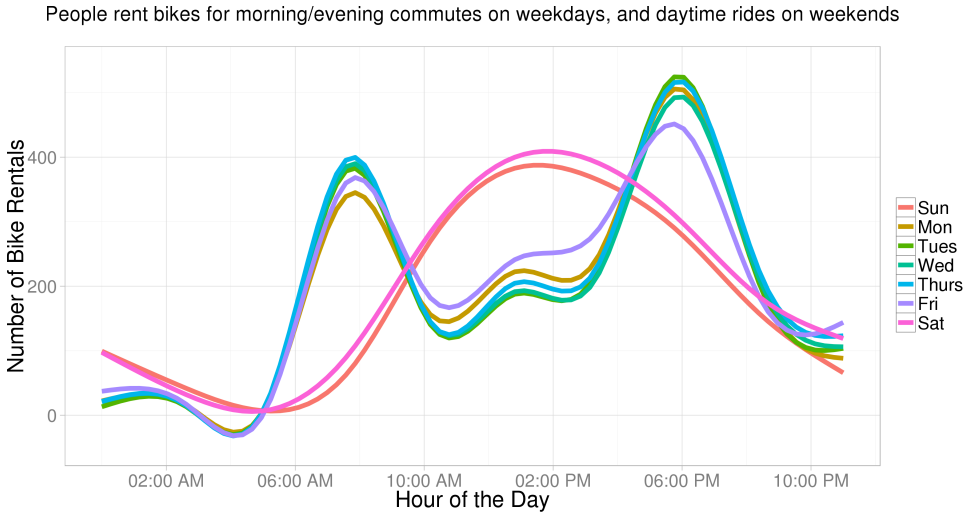


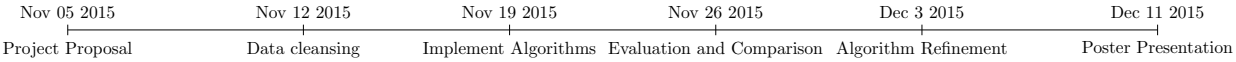
Figure 1: The hourly rental data predicted in a week.

Hopefully, we can find out the causes of the fluctuation of bike demand and give out a strategy to predict it to ensure high bike availability and to maximize the bike uses.

2 Proposed Work & Timeline

There are mainly three procedures in our project. (1) Data cleansing on the raw dataset and feature engineering. (2) Predict model construction with different machine learning algorithms. (3) Model evaluation. To begin with, we need to know the correlation of the features and extract these features from the raw data as the input training data. Then we can train models with various machine learning algorithms such

Figure 2: Time Line



as decision tree, linear regression, SVM, random forest and so on. Lastly, we will compare, analyze and optimize these models, summarize and provide solutions for this bike sharing problem.

The project spans for 5 weeks and the timeline is listed in Fig. 2.

3 Deliverables & Evaluations

The output of our work is the prediction of bike demand. Since this project comes from Kaggle, we should submit our project to Kaggle to compare our solution with other team and compare our different models with separate submission. Kaggle will return a ranking of the submission. Through the ranking, we can evaluate our models and compare different models.

4 Data

The data is the hourly bike rental data that spans for two years. The training data is the rental data from the first 19 days of each month while the test data is the rest days of the month. The task is to predict the total rental bikes in each hour in the test set, based on the prior information. The data features include binary, categorical and quantitative data, which are listed in Table 1.

Table 1: Data Fields

	data fields	description
	datetime	hourly date + timestamp
binary	holiday	whether the day is considered a holiday
	workingday	whether the day is neither a weekend nor holiday
categorical	season	1 = spring, 2 = summer, 3 = fall, 4 = winter
	weather	1: Clear,2: Mist Cloudy,3: Light Snow, 4: Heavy Rain
quantitative	temp	temperature in Celsius
	atemp	"feels like" temperature in Celsius
	humidity	relative humidity
	windspeed	wind speed
	casual	number of non-registered user rentals initiated
	registered	number of registered user rentals initiated
	count	number of total rentals

5 Software Tools & Prior Discussion

In this project, we will take advantage of multiple python modules such as pandas, numpy, sklearn, math, ggplot and so on. As we know, a lot range of powerful machine learning algorithms are included in these modules.

We discussed with Nikolaos for several time on email and office hour. We originally choose online product sales but there are problems with the test data. Then we found this interesting competition. After several discussion, we have more expectations on applying what we have learned into this real-life problem.