Data Science W207.2 Final Project

Bike Sharing Demand

Matthew Holmes

Andrew Kabatznick

Grant Whittington

December 19, 2017

**Introduction/Business Case**

The purpose of this assignment is to apply the data science/machine learning process to a competition from Kaggle, a website devoted to machine learning and related public contests in which the public may try their hand at finding the best solution to a given problem. This assignment allows students to demonstrate the knowledge learned over the term in a complete real-world problem.

As beginning practitioners of machine learning, the team chose a problem with a manageable amount of data, limited factors and a simple well-defined goal. That problem is Bike Sharing Demand, defined in the link below:

<https://www.kaggle.com/c/bike-sharing-demand/data>

The goal is to predict the number of bikes which are rented in a given hour, based on data/time information and weather-related effects. This problem is somewhat of a toy case designed for people to cut their teeth on a data set. This aspect comes with benefits in terms of size and simplicity, but is not well defined in terms of goal or scope, and in some ways is ill defined to the type of techniques used in this course. For example, the scoring criterion is given, making model comparison easier and well defined. That equation defines RMSLE (Root Mean Squared Logarithmic Error) and is given below.

Where pi, ai are the predicted and actual values for a test dataset, and n is the size of that dataset. Log refers to the natural log.

To make the problem more well suited to the project, a notional problem definition will be given, and the team will seek to solve that problem. When the actual problem conflicts with the notional one, the team will discuss what route was chosen, why, and what would have been done in a more ideal scenario.

The notional scenario involves a bicycle rental company which has been around for a few years and is looking to optimize their bike rentals over various sites in a city. From day to day, different amounts of bikes are rented based on a number of factors, and accurate prediction of that number will allow for bikes to be redistributed among sites, leading to a higher use rate and more profits. The initial dataset will be that of the rentals per hour at the site, and corresponding date/time information and related weather factors. After the initial model is created, “on-line” features will be implemented, allowing new data to be used in prediction of future data.

This scenario differs from that of the Kaggle competition in a few key ways. One, the Kaggle competition requests that test data only be predicted based on previous training data from a temporal standpoint. As such, it makes its “test set” the times relating to days at the end of the month, and its training set the days at the beginning of the month. Such a time based request is outside the scope of this course and would require techniques beyond those covered. As such, that request will be ignored and the data will all be considered in the past and fair to use as one large dataset. The on-line aspect will be handled by using a subset of previous data as a factor to predict future data. This assumption has one downside in that aspects of the dataset are removed, making some methods or ideas difficult if not impossible to implement. This will be discussed more in future sections, such as factor generation.

**Data Understanding (EDA)**

*Leaving this section to someone else*

**Data Preparation (Cleaning Factor Generation)**

Words

**Modeling**

Linear Model

Bayesian Model

Tree Model

Factor Selection – why we didn’t do it

**Evaluation**

Model quality, error analysis

Deployment (Probably don’t need this)