STAT 685: Dr. Suojin Wang's Group

Modeling Seoul Bike Sharing Demand

Nam Tran, Bai Zou

Fall 2020

1. Intruduction

1.1 Background

Our data set is the "Seoul Bike Sharing Demand Data Set", which on a high level contains hourly data for bike usage as well as various covariates that might be useful, e.g., temperature. Further, it contains around one year of data.

The data set has been aggregated and been uploaded to the UCI Machine Learning Repository, located here: http://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

At first glance, relevant pieces are:

- Contains 8760 observations
- There are 14 columns

Regarding motivation for the data set and its potential use, the following is taken from the UCI website and was attached by the team that donated the data: "

"Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information."

1.2 Previous Work

There are two papers dealing with this data set explicitly. We're including both papers and in addition providing high level commentary on the paper and our concerns.

1.2.1 Paper 1: VE, Park, and Cho (2020, Computer Communications), Using Data Mining Techniques

High Level:

- Attempted to predict the hourly demand.
- Claims .96 R^2 and .92 R^2 for train and test, respectively.

- Uses “Boruta” to identify important features, which underneath the hood uses random forest as part of the algorithm.
- Considered linear regression, GBM, SVM’s (radial basis function), boosted trees, and xgboost; had a nice mathematical summary of each.

Thoughts/Concerns

- Didn’t consider L1 Regularization, i.e., LASSO, which seems the most obvious for feature importance/selection/ranking. For example, do the full path as we go over λ and see the order they get dropped, where the sooner they drop the less important they are.
- Make no reference to how they split up into 75% train and 25% test, e.g., is it interleave of a specific calendar day and everything after is post? If interleaved, the the 75% train’s distribution and 25% test’s distribution are effectively identical and learning on the train portion is considered “cheating” since data has leaked.
- Didn’t consider non-linear transforms of the data, which may not be that important given the usage of decision trees, but could have allowed plain linear regression to perform better.

1.2.2 Paper 2: VE and Cho (2020, European Journal of Remote Sensing), A Rule-Based Model

High Level

- Consider CUBIST, regularized random forest, CART, KNN, Conditional Inference Tree.
- Almost identical to previous paper.
- They had an additional data source, the “Capital Bikeshare program,” for which the dataset came from Kaggle. It didn’t seem that they used this in an “interesting” way (interleaving the data, using it as a validation) but instead just considered it as another data source for which they ran their same methodology and if they got the same Rˆ2 and other metrics, then they’d consider that would be successful.

Thoughts/Concerns

- Still don’t talk about train/test split methodology.
- The last bullet point of high level thoughts.

1.3 Scope and Goal

Based on the description above, we break it into two potential business requirements here:

- Predict next day hourly demand based on historical data until the current day.
- Real-time prediction for next hour demand based on historical data until the current hour.

The scope in this study is to:

- Re-define training and testing data with anchor time.
- Re-evaluate estimation methods with data splitting by anchor time.
- Build forecasting models to predict the next hour demand and compare the models in terms of prediction accuracy, prediction variance, running time, etc.

2. Data Exploratory

3. Estimation Methods Comparison

TODO: Nam - Please feel free to put the methods introduction comparison here.

4. Forecasting Model Comparison

There are three models considered to predict the next hour demand.(Assume the current day is X, and current hour is Y).

- Model 1: One-time prediction for day X hourly demand by end of day X-1 with data updated to day X-1.
- Model 2: Real-time model training to predict next hour (Y+1) demand with data updated to day X hour Y.
- Model 3: Continuous model training based on Model 1 and updated data from hour 1 to Y on day X.

4.1 Model 1: One-time Prediction

4.2 Model 2: Real-time Model Training

4.3 Model 3: Continuous Model Training

4.4 Model Comparison

5. Result and Conclusion