Robert H. Smith School of Business
BUDT 758B Big Data: Strategy and Analytics
Professor Anand Gopal,Gordon Gao

Capital Bikeshare

Team 6

December 7, 2016

Team Members:

Aanchal Kanodia
Sahil Bareja
Varsha Purswani
Xiao Lu
Zalak Parekh
Zhe Du

Introduction

Capital Bikeshare is a bike sharing system available in DC, Maryland and Virginia. It has more than 450 stations and 2500 bicycles. Our aim was to analyze their dataset from year 2010 to 2016 and find systematic patterns in bike rental activities to help create business value.

Objective

- Find specific seasonal/day-specific patterns
- Effect of weather on bike rental activities
- Build a prediction model that predicts the demand at a given station at a certain time

Preliminary Analysis

The dataset consisted 7 years of bike rental duration data in 25 files, size of dataset being 2GB.
It contained the following data fields:
Duration, Start Date, End Date, Start Station, End Station, Bike#, Member Type

Challenges in working with the dataset

- Discrepancies in column names and number of columns
- Column sequencing is different
- Station names and codes assigned to them change across the years
- Categories of MemberType column changing over the years
- Formatting of the duration column varied over the years

Data Standardization

- Column names, sequence and units of values in different columns was made consistent across all data files
- Introduced new codes for station names
- Combined Weather Data to see the effect of weather on bike renting activity

Data Storage

Amazon Simple Storage Service (S3) was used for data storage. A bucket was created and made public. Pig and Hive clusters were created in the location US-West (Oregon) for parallel processing of data stored in S3. Spark clusters were created for Predictive analysis. Both “on demand” and “spot” instances were used. 4 instances were running in parallel at a time

Descriptive Analysis

Hive queries were used on the data to answer business questions and output of these queries was visualized in Tableau.

Key Findings

- The data exhibits quarterly seasonal pattern. Demand for bikes is observed to be the highest in Quarter 3 for each year from 2010 to 2016.
- Demand for bikes is highly influenced by variation in weather. As temperature and visibility is high in Quarter 3 there is a rise in demand for renting bikes.
- It is observed that the top 5 (by Bike count) stations exhibit the same pattern in demand for bike rental activities.
- It is seen that for station like **Union Station** the demand is high on weekdays and it falls on weekends. At the same time stations closer to tourist spots like **Lincoln Memorial** shows that demand is low during weekdays and exhibits a sharp increase during weekends.

Network Analysis

Due to the complexity and the scale of the network, we used SparkR along with AWS for making the network analysis. We examined the network on both a global and a local level and selected the most representative nodes for illustrating the network diagram, based on the following findings on key metrics: (Due to lack of information, such as operating cost for each station, profit margin for each station and etc., we assigned the same amount of strength/weight to each edge).

Capital Bikeshare's has a disconnected operating network where some nodes might not be connected to any other node. In such case, we limited the closeness measure to the largest component of nodes (i,e., measured intra-component). Given such nature, we strongly recommend the company to re-strategize such stations with minimal business.

Predictive Analysis

In order to predict the demand for a given station at a certain time, we built two multi-linear regression models:
1) Using processed bike share data – **RMSE 22.25**

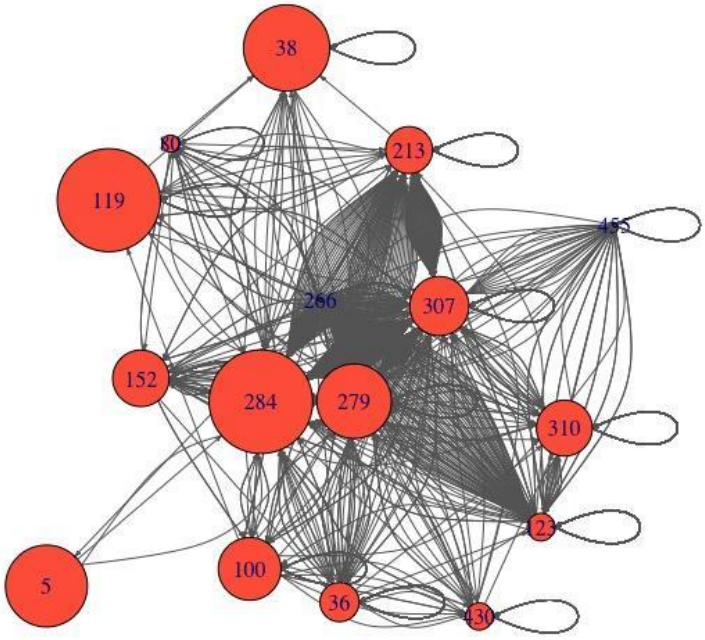
Clearly, the above RMSE isn't ideal and hence we quickly realized the need to change our data set as well as our model. We decided that it would be more meaningful to target a given station at a certain time while utilizing a Spark query to fetch only the records of the station that we need to predict the demand for. We also created a new data set that included the weather and holiday data.

2) Using new data set as described above:
- Without weather and holiday attributes – **RMSE 6.3**
- With weather and holiday attributes - **RMSE 5.74**

Recommendations

- Make week days more attractive by running special offers for weekdays
- Surging/decreasing the price after predicting a demand at a given station
- Close down some bike stations during the low season
- Loyalty privilege to patrons: Customer retention costs are always lower than customer acquisitions

Appendix



Network Diagram

Key Metrics	Value
Number of Total Edges	43468
Avg. Edge Per Node	95
Network Connectance	0.209
Network Density	0.207
Network Diameter	6
Global Cluster Coefficient	0.7117
[1]**Avg. Betweenness**	333.81
Avg. Degree	3165.35

[1] The centrality metrics in bold font was calculated at a local level for each node. 16 nodes were selected for plotting based on their individual centrality score. We set the average duration to be the cutoff for selecting edges.

Heatmap showing duration values between stations (Higher Duration Shown By Darker Color)

Link to the Capital Bikeshare Datasets:

https://www.capitalbikeshare.com/trip-history-data

Source Code for the project:

https://drive.google.com/drive/u/0/folders/0B4Y-6Lo8wTtleDBidk5pZFZfcTQ