Foundations of Data Science

Christine Schlenker

June 24, 2016

Capstone Project Milestone 1: Bike Sharing Demand Prediction

Introduction

A bicycle sharing program is a specific type of a short-term equipment rental designed to enable accessible intercity transportation. A common program consists of multiple stations (electronically-locking bicycle docks with a payment kiosk), a fleet of compatible bicycles, and a payment scheme that allows for long-term, short-term, and one-time usage. The operating company controls station placement, manages bicycle and station maintenance, and sets usage and payment rules. Users of the system receive the benefits of having easy access to intercity transportation between major public transportation stations without the burdens of initial capital investment, maintenance, storage, or risk of theft.

As with other on-demand rental systems, the operating company is concerned with providing an optimized quantity of bicycles at its stations. Usage demand may be influenced by a variety of factors, including station location, proximity to work or entertainment destinations, weather, and population demographics.

This project will analyze bike sharing demand for Divvy Bike Share in Chicago, Illinois, which is a program of the Chicago Department of Transportation (CDOT). CDOT owns all physical assets of the program and the bike sharing firm Motivate is the operator. Divvy has made public an anonymized version of its trip and station data, and this project uses data from the third quarter of 2013 to the end of 2015.

The Data

Multiple data sets will be used for this analysis.

* Divvy Bike’s trip data, which comprises over 3.2 million bike share trips from 2013Q3 to 2015Q4, found at <https://www.divvybikes.com/data>.
  + Fields:
    - Trip start day and time
    - Trip end day and time
    - Trip start station
    - Trip end station
    - Rider type (Subscriber, aka Member, or Customer, aka 24-Hour Pass User)
    - If a Subscriber trip, it will also include Subscriber’s gender and year of birth
* Divvy Bike’s station data, describes the station attributes across the same time period, found at <https://www.divvybikes.com/data>.
  + Fields:
    - Station ID
    - Station Name
    - Latitude
    - Longitude
    - Capacity – number of docks as of December 31
    - Landmark
    - Online Date
    - Note: 2015 data does not include online date
* Weather data obtained from Weather Underground <https://www.wunderground.com>.
  + Fields:
    - Temp. (°F) – mean, max, min
    - Dew Point (°F) – mean, max, min
    - Humidity (%)–mean, max, min
    - Sea Level Press. (in) – mean, max, min
    - Visibility (mi) – mean, max, min
    - Wind (mph) – mean, max, min
    - Max wind gust speed (mph)
    - Precipitation (in) – sum
    - Cloud Cover – describes the daily average ratio of the sky covered by clouds in eighths from 0 to 8. A score of 1 indicates 1/8 or 12.5% of the sky is covered in clouds.
    - Events – fog, rain, snow, and thunderstorms. This can be null or have multiple hyphen delimited events.
    - Wind direction degrees
* Bike station information compiled by Open Gov Hack Night (linked from the Divvy site), found at, for the purpose of having the starting number of bikes at a dock: <https://github.com/stevevance/divvy-munging/blob/master/data_from_others/station_launch_dates_from_viddy.csv>.
  + Fields:
    - ID
    - Station Name
    - Status
    - Available Bikes
    - Available Docks
    - Time Added
* Potential additional data may include a set of federal holidays and school calendars.

Data Limitations

* Non-subscriber demographics are not included, and subscriber demographics are limited only to year of birth and gender. Further information regarding employment/student status, income, home zip code, etc. would enable greater usage profiling.
* Divvy trip data from 2013 includes only Customer profile types. No demographic data is available.
* Subscription types are not included (Annual, Corporate, or Student)
* While starting number of bikes may be derived from using the Open Gov Hack night data, there is no way to account for terminal closures, relocating bicycles to other stations, bikes out of service, etc. Therefore, no instantaneous dock capacity can be determined.
* This also prevents knowing when a station is full. A user may arrive to his/her intended destination station only to find that the station has no open docks. If the user then goes to a nearby dock instead, showing that dock was in demand when in actually the original dock was preferred provides misleading information for the operating company.
* Note, this information is probably available to the operator. A live JSON feed is available for dock utilization and service status, however, publically-available historical information is not.
* No individual rider ID. Therefore, it’s not possible to create user tracks, e.g., being able to identify users that only two stations (likely commuters), including identifying when a user checks in a bike to avoid a usage fee for trips over 30 minutes.
* Weather data is presented in min, mean, max format, with no times associated with when the minimum and maximum occur. I will look for a more robust data set to try to provide better information on the influence of temperature on demand.

Data Wrangling and Cleaning

* Divvy’s trip and station data sets are both anonymized and cleaned for public use. When the record is a Customer and no demographic information is included, the value is null. This will not present a challenge.
* Divvy only included trip data if both a start and end date were recorded, so that cleaning has already been done.
* As previously mentioned, 2015 Divvy station does not include station online date. Though not yet confirmed, this is probably because no stations came online in 2015 and the field was not needed. If so, this is not a problem.
* As previously mentioned, Divvy’s trip data does not differentiate between Subscriber and Customer for 2013, and no demographics are available. There is no way to replace or impute this missing data. Therefore, analysis on demographics will have to exclude this time period.
* Weather Underground’s (WU) weather data set has missing data when a field is not applicable. This can be fixed.
* One of the WU fields, Events, needs to be separated into multiple binary columns. The formatting within the column is consistent, so the output should be clean. The new columns will be Fog, Rain, Thunderstorm, Snow.
* Potential data reshaping (not yet tested):
  + Two binary columns for Divvy trips Rider Type: Customer and Subscriber
  + Binary column for overage charges (if trip time is greater than 30 minutes)
  + Convert gender to binary isMale to reduce file size (note, this works because the gender is only two options, male or female)

Moving Forward

Based on my study of previous forays into bike share demand, I plan to use a random forest algorithm to predict bike share demand.

I have started with initial visualization of the data using Tableau.

The below graph shows peak trip start times are during the hour blocks of 8 am and 5 pm, suggesting that these trips are used for commuting.

