Predicting Bike Share Demand in Washington D.C.

CS221 Final Project

Beri Kohen, Ayush Singla

Mentor: Cindy Jiang

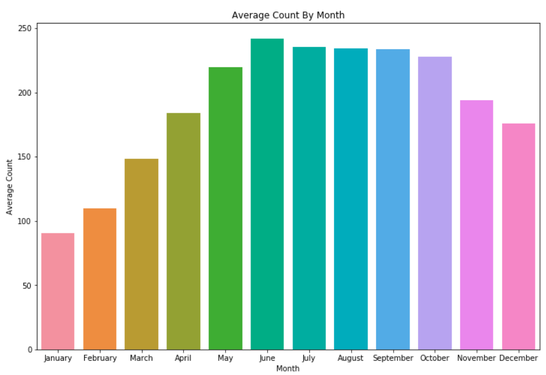
Abstract

A bike-sharing system is a service in which a fleet of bicycles is made available to the public on a short-term basis through self-served docking stations. These stations are limited in capacity and are often depleted or saturated with bikes due to sudden spikes in demand. These spikes are hard to avoid and are both detrimental to the user experience and the effectiveness of the system. The aim of this project is to predict the total count of bikes rented during each hour by combining historical usage patterns with the related information of users, weather, and other factors available prior to the rental period. Thus, we hope that by answering the question of how many number of bikes would meet user demand for a given future time, we can help provide a solution to the problem of bike re-balancing, which affects lower-income households detrimentally since these are a cheap mode of transportation.

Data

• The dataset used is hosted on UCI machine learning repository, which itself collected from the Capital Bikeshare program in two years (2011 - 12) around Washington D.C.
• The dataset contains 17379 rows of hourly count of rental bikes with the corresponding features, which are shown in the table below:

Features	Value
Date	MM/DD/YYYY
Time	HH:MM
Season	Takes 4 values: 1 = spring, 2 = summer, 3 = fall, 4 = winter
Holiday	1 = yes, 0 = no
Working Day	1 = yes, 0 = no
Weather	1: Clear, few clouds, partly cloudy. 2: Mist + cloudy, mist + broken clouds, mist + few cloud 3: Light snow, light rain + thunderstorm + Scattered clouds, light rain + scattered clouds. 4: Heavy Rain + Ice Pallets+ thunderstorm + mist, snow + fog
Temp	Temperature in celsius
Atemp	“feels like” temperature
Humidity	Relative humidity
Count [LABEL]	Number of user rentals



Methodology

Feature Extraction

Our model includes non-continuous (categorical) features, which are made fit using dummy coding, where in general, a variable with k possible values will be transformed into k binary variables each which can be included in the regression model.

Linear Regression

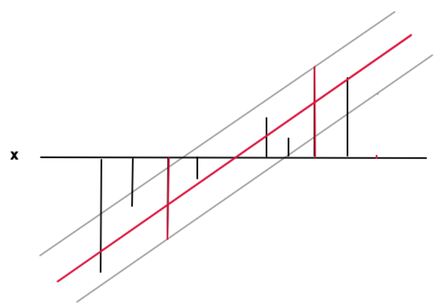
Linear regression tries to find the “best” line model to fit the distribution of target variable with the variables which influence it, and use this line to predict the target variable. Formally, using L2 regularization, it is:

$$\ell(\theta) = \frac{1}{m} \sum_{i=1}^{m} (\theta x^{(i)} - y^{(i)})^2 + \alpha \|\theta\|^2$$

Support Vector Regression

In this case, our goal is to find a function f(x) under the condition that f(x) is within a required accuracy ε from the labels of every data point. Thus, we can formalize this as:

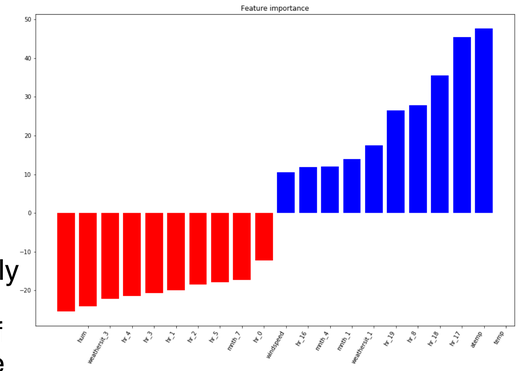
$$|y(x) - f(x)| \leq \epsilon$$



where ε is the distance between the red and the grey line.

Analysis

Linear Regression



• Our multiple linear regression on our processed training set gave an MAPE of around 65 percent.

• We then split the training and test datasets chronologically and implemented a feature d_7, the count of bikes rented during the same hour 7 days prior. Using these optimizations, the MAPE was reduced to 55 percent.

Support Vector Regression

• Using SVR reduced the MAPE value to 44 percent, as expected since it is able to capture non-linearities.

• After analyzing the distribution of absolute error, we found that very few predictions were incorrect by extreme margins.

- In real life, we would ideally want to predict the range of bike demand, so a classification would make more sense and be more accurate.
- Therefore, our final step will be to try SVM classification, where we will create buckets (0-25,25-50,50-100, etc.) that contain around the same number of bikes.