Regression: Demand estimation

Statement of Work

Course Name	AI Algorithms I
Course Code	AIDI 11688
Course Facilitator	Marcos Bittencourt
Student	Manu Sihag (100801028)
Date	06/11/2020

DURHAM COLLEGE SCHOOL OF BUSINESS, IT AND MANAGEMENT

Contents

Executive Summary

This project demonstrates the feature engineering process for building a regression model using bike rental demand prediction as an example. We demonstrate that effective feature engineering will lead to a more accurate model. Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

Problem Statement

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city.
Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

How natural and man-made factors are affecting the bike rental demand for Capital Bike share System in Washington DC?
Natural factors include seasons, months, day of week ,peak timings, working and non-working days, temperature, humidity etc. and man made factors consists of location of bike station, characteristics of the area.

Data Requirements

The Bike Rental UCI dataset is used as the input raw data for this experiment. This dataset is based on real data from the Capital Bikeshare company, which operates a bike rental network in Washington DC in the United States.

The dataset contains 17,379 rows and 17 columns, each row representing the number of bike rentals within a specific hour of a day in the years 2011 or 2012. Weather conditions (such as temperature, humidity, and wind speed) were included in this raw feature set, and the dates were categorized as holiday vs. weekday etc.

The field to predict is "cnt", which contain a count value ranging from 1 to 977, representing the number of bike rentals within a specific hour.

Model/Architecture Approach

The goal is to predict a number (the demand for the bikes, represented as the number of bike rentals), so we will choose a regression model. Given that the number of features is relatively small (less than 100) and these features are not sparse, the decision boundary is very likely to be nonlinear. Based on these observations, we decided to use the Boosted Decision Tree Regression algorithm for the experiment

Overall, the experiment had five major steps:

- Step 1: Get data
- Step 2: Data pre-processing
- Step 3: Feature engineering
- Step 4: Train the model
- Step 5: Test, evaluate, and compare the model