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

Get Data

The dataset "Bike-Sharing-Dataset" was obtained by the UCI Machine Learning Repository. This is a collection of databases, domain theories and data generators which are used by the machine learning community for empirical analyses. The UCI Machine Learning Repository is based on donations of researchers, mostly outside of UCI. We found the dataset "Bike Sharing Dataset" under the index "regression" and chose the sub-dataset "day".

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

The dataset contains following columns in name. Below there is a short description:

- weathersit:
    - 1: Clear, Few clouds, Partly cloudy,
    - 2: Mist and Cloudy, Mist and Broken clouds, Mist and Few clouds, Mist
    - 3: Light Snow, Light Rain and Thunderstorm and Scattered clouds, Light Rain and Scattered clouds
    - 4: Heavy Rain and Ice Pallets and Thunderstorm and Mist, Snow and Fog
- instant: record index
- dteday: date
- season: season (1: spring, 2: summer, 3: fall, 4: winter)

- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- holiday: weather day is holiday or not
- weekday: day of the week
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
- temp: Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Data Pre-processing

Data pre-processing is an important step in most real-world analytical applications. The major tasks include data cleaning, data integration, data transformation and data reduction.

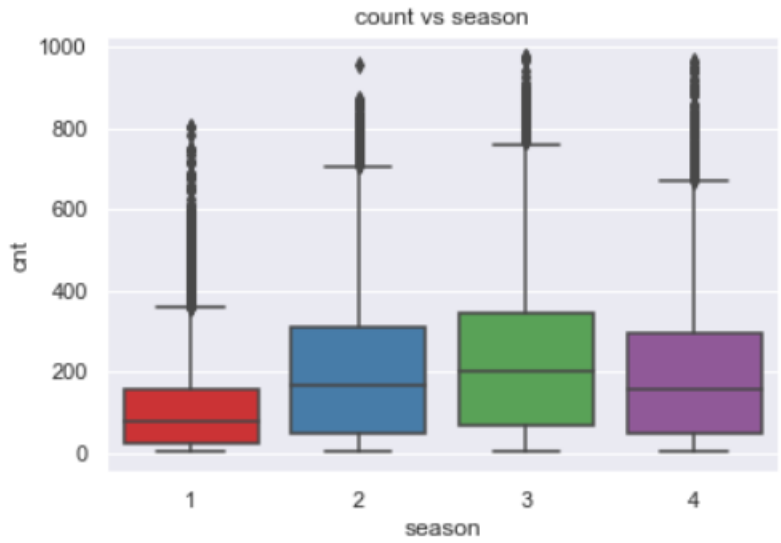
Missing Value Analysis

```
df.isnull().sum()
```

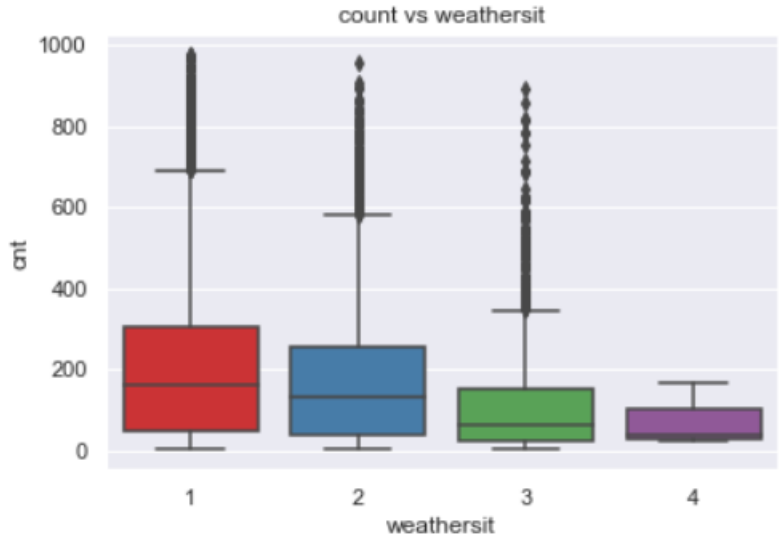
```
instant       0
dteday        0
season        0
yr            0
mnth          0
hr            0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
cnt           0
dtype: int64
```

During our analysis we checked the dataset for any sort of missing data but found that no missing values are present in the dataset.

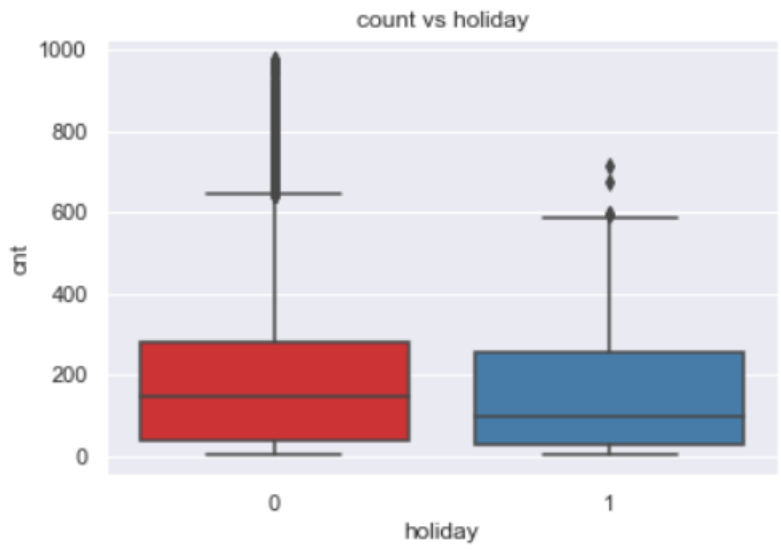
Next, we looked at the relationship between the response variable and each explanatory variable. We selected some plots with obvious patterns as shown below.



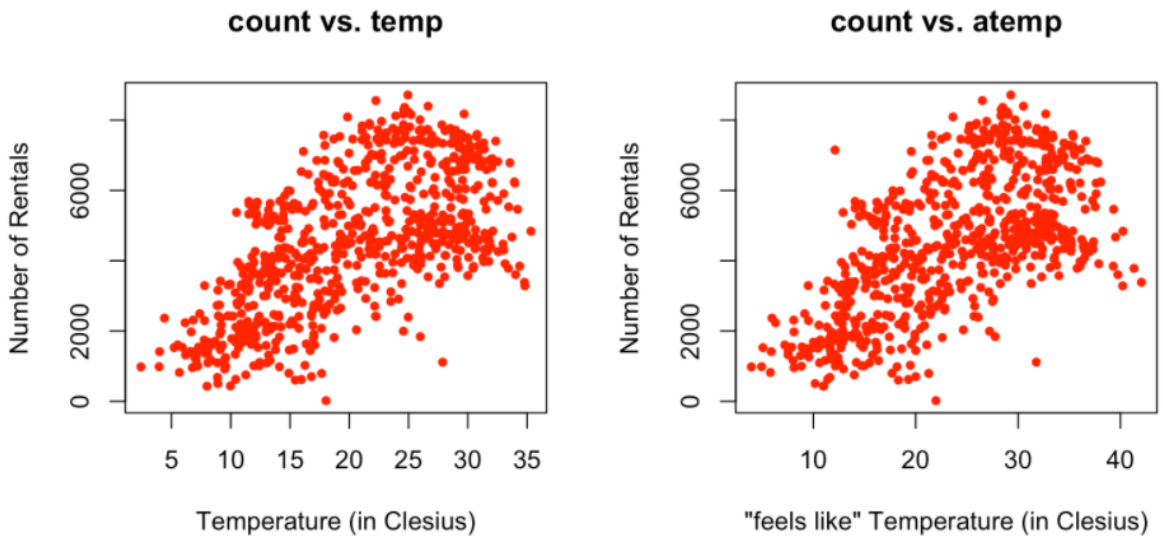
The first plot shows the relationship between cnt variable and season, which confirms the average numbers of bike rentals are the highest during summer and fall.



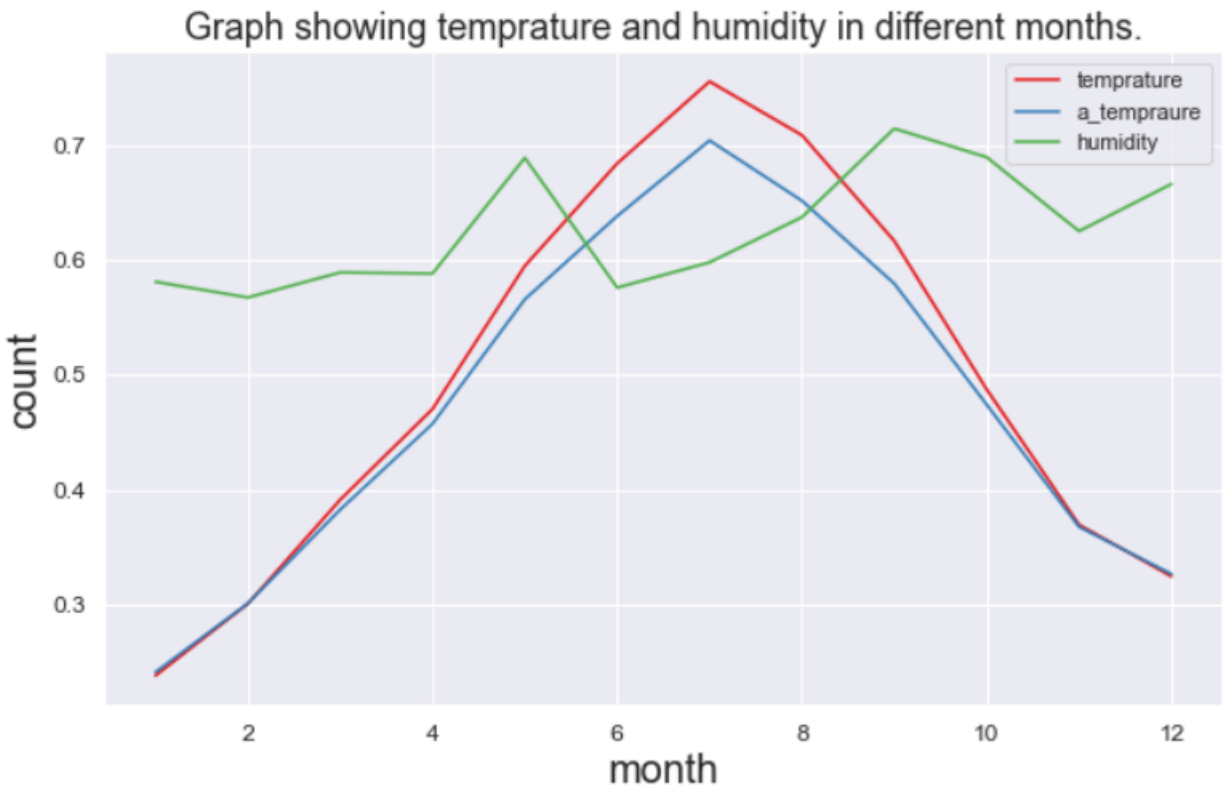
The second plot shows the relationship between cnt variable and holiday. We can see that the average number of bike rentals on non-holiday is higher than holiday, but has more variability as well.



The third plot shows the relationship between cnt variable and weather. There is a clearly decreasing trend of bike rentals when weather conditions grow worse.



These two plots show the relationship between cnt variable and temperature. However, the data seems to be scattered with a lot of variability, so the linear relationship might be weak if there is any.



There seems to be a linear relationship between them, which means more people will rent bikes when it gets warmer.

Next, we converted the date, year, day and hour columns to a single datetime column and removed them from the dataset.

```
df['dteday'] = df['dteday'] + df['hr']
```

```
df.dtypes
```

```
instant                int64
dteday        datetime64[ns]
```

The new dataset looks like this:

```
df.head()
```

	instant	dteday	season	holiday	weekday	workingday	weathersit	temp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01 00:00:00	1	0	6	0	1	0.24	0.81	0.0	3	13	16
1	2	2011-01-01 01:00:00	1	0	6	0	1	0.22	0.80	0.0	8	32	40
2	3	2011-01-01 02:00:00	1	0	6	0	1	0.22	0.80	0.0	5	27	32
3	4	2011-01-01 03:00:00	1	0	6	0	1	0.24	0.75	0.0	3	10	13
4	5	2011-01-01 04:00:00	1	0	6	0	1	0.24	0.75	0.0	0	1	1