BIKE RENTING

By

Seethu Joseph

Contents

Chapter 1

Introduction

1.1 Problem Statement

The objective of this Case is to predication of bike rental count on daily based on the environmental and seasonal settings.

1.2 Data

My task is to build regression models which will predict the count of bike renting on environmental and seasonal settings. Given below is a sample of the data set that we are using to predict the count of bike:

Table 1:1 Bike Renting sample data (Columns 1 to 6)

instant	dteday	season	yr	mnth	holiday
1	2011-01-01	1	0	1	0
2	2011-01-02	1	0	1	0
3	2011-01-03	1	0	1	0

Table 1:2 Bike Renting sample data (Columns 7 to 12)

weekday	workingday	weathersit	temp	atemp	hum
6	0	2	0.3441670	0.3636250	0.805638
0	0	2	0.3634780	0.3537390	0.6960
1	1	1	0.1963640	0.1894050	0.437372

Table 1:3 Bike Renting sample data (columns 13 to 16)

windspeed	casual	registered	cnt
0.1604460	331	654	985
0.2485354	131	670	801
0.2483090	120	1229	1349

As you can see in the table below we have the following 16 variables, using which we have to correctly predict the count of the bikes:

Table 1:4 Predictor Variables

1.season
2.yr
3.mnth
4.holiday
5.weekday
6.workingday
7.weathersit
8.temp
9.atemp
10.hum
11.windspeed

The details of data attributes in the dataset are as follows:

instant: Record index

dteday: Date

season: Season (1:springer, 2:summer, 3:fall, 4:winter)

yr: Year (0: 2011, 1:2012)

mnth: Month (1 to 12)

holiday: weather day is holiday or not (extracted from Holiday Schedule)

weekday: Day of the week working day: If day is neither weekend nor holiday is 1, otherwise is 0.

weathersit: (extracted fromFreemeteo) 1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)

atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_maxt_min), t_min=-16, t_max=+50 (only in hourly scale)

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

Chapter 2

Methodology

2.1 Pre-processing

Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable. My variables are normally distributed here.

2.1.1 Outlier analysis

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. It needs close attention else it can result in wildly wrong estimations. I could not observe the presence of outlier in the data set.

2.1.2 Missing value Analysis

Missing values occur when no data value is stored for the variable in an observation. Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Typically, ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. I could not find missing value in the dataset.

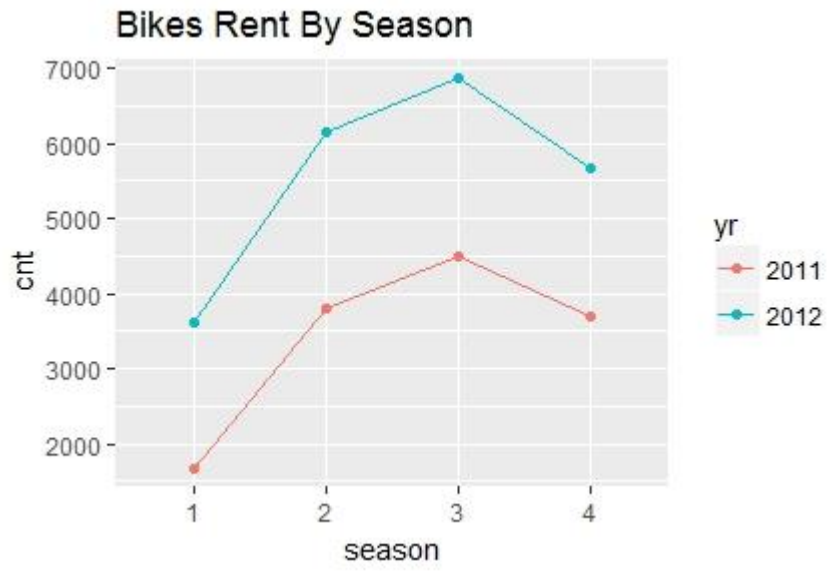
2.2 Feature Engineering

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. Feature engineering means deriving new variables from the raw data.

2.3 Exploratory Data Analysis

First step in any analytical project is to understand data using exploratory data analysis. Once you understand data then you can drive data according to the problem statement. Exploratory data analysis is an approach for summarizing and visualizing the important characteristics of a data set. Exploratory data analysis focuses on exploring data to understand the data's underlying structure and variables, to develop intuition about the data set, to consider how that data set came into existence, and to decide how it can be investigated with more formal statistical methods. To accomplish this task (Understand the data), we have graphical and non-graphical tools. Graphical tools includes all visualizations like scatter plot for correlation, boxplot for outlier analysis, histogram for distribution, bar plot for missing value analysis, heat map for locations. Non-Graphical tools include all central statistics and statistical techniques/test.

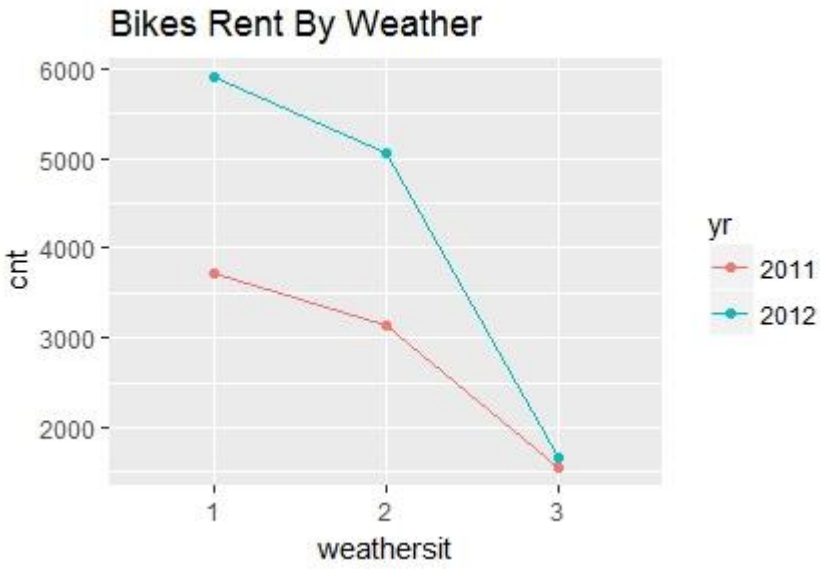
Fig 1: Bikes Rent By Season



Season: 1-Spring, 2- summer, 3-fall, 4-winter

The above figure shows the plot of bikes rented by season. The count of bike rented is double times in 2012 than 2011.The more bikes are rented on fall season and less bikes rented on spring.

Fig 2: Bikes Rent By Season



weathersit: 1: Clear, Few clouds, Partly cloudy

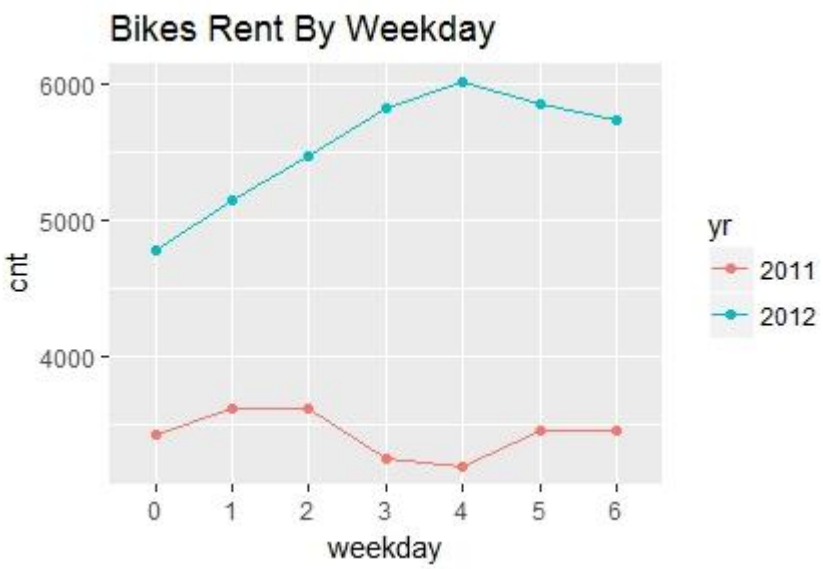
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Plot shows that more bikes rented on clear, few clouds and partly cloudy and no bikes are rented during heavy rain, ice pallets, thunderstorm, mist snow and fog

Fig 3: Bikes rent by weekday



weekday: 0-Sunday, 1-Monday, 2-Tuesday,3-wednessday,4-Thursday,5-Friday,6-Saturday

In 2011 more bikes are rented on Tuesday and less on Thursday. The count of bike rented on Monday, Tuesday, Friday and Saturday is almost equal. While we are analysing 2012, more bikes are rented on Thursday and less on Sunday. Each day, the count is varying.

Fig 4: Bikes Rent By Month

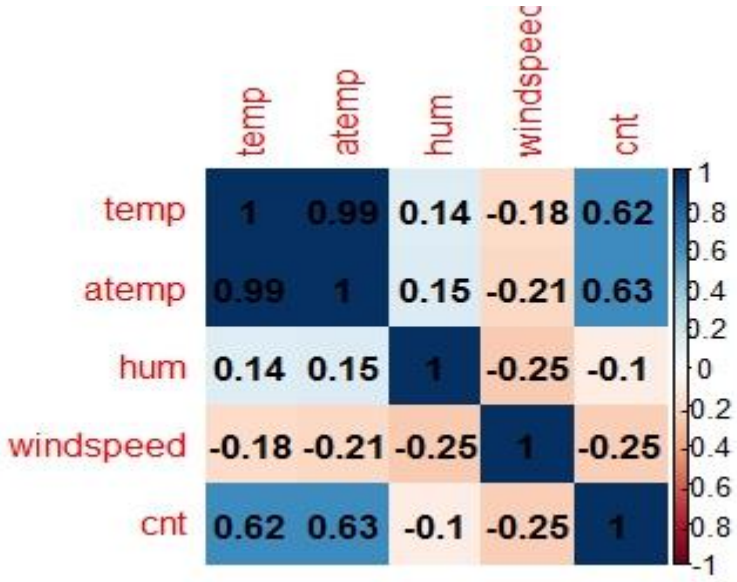
mnth: 1-Jan, 2-Feb, 3-Mar, 4-Apr, 5-May, 6-Jun, 7-Jul, 8-Aug, 9- Sep,10- Oct, 11-Nov, 12-Dec

In 2011, more bikes are rented in June and less in January. Count is increasing from January to June and it is decreasing from July to December and it is mainly shows the season. In 2012, the pattern is almost same. The count of bike mainly depends on weather situation.

2.4 Correlation Analysis

Correlation refers to the extent to which two variables have a linear relationship with each other. It is a statistical technique that can show whether and how strongly variables are related. It is a scaled version of covariance and values ranges from -1 to +1.

Fig 5: Correlation plot between fields



It shows that temp and atemp has much more correlation. The correlation between temp and atemp are o.99 so they are highly positive correlated variables. The features temp and atemp are strongly correlated. If both features are included in the model, this will cause the issue of Multicollinearity (a given feature in the model can be approximated by a linear combination of the other features in the model). Hence I include only one temperature feature

into the model. The features casual and registered are omitted because that is what we are going to predict.The feature dteday is omitted.

2.5 Modelling

2.5.1 Linear Regression

Linear regression is the most basic type of regression and commonly used predictive analysis. The overall idea of regression is to examine two things:

(1) Does a set of predictor variables do a good job in predicting an outcome variable? Is the model using the predictors accounting for the variability in the changes in the dependent variable?

(2) Which variables in particular are significant predictors of the dependent variable?

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

2.5.2 Random Forest Regression

Random Forest or decision tree forestsisan ensemble learning method for classification, regression and other tasks.This method was championed by Leo Breimanand Adele Cutler, and combines the base principles of bagging with random featureselection to add additional diversity to the decision tree models.It consists of an arbitrary number of simple trees, which are used to determine the final outcome. For classification problems, the ensemble of simple trees vote for the most popular class. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases).

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART model with different sample and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

In particular, trees that are grown very deep tend to learn highly irregular patterns. They over fit their training sets, i.e. have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance of the final model.

Chapter 3

Conclusions

3.1 Model Evaluation

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models.

We can compare the models using any of the following criteria:

1. Predictive Performance

2. Interpretability

3. Computational Efficiency

3.1.1 MAPE

Regression error metrics**:** We want to know how well the model predicts new data, not how well it fits the data it was trained with. Key component of most regression measures are difference between actual y and predicted y ("error")

MAPE**:** Stands for Mean absolute percentage error. Measures accuracy as a percentage of error.

3.2 Model Selection

While analysing both the models, I got MAPE of Linear regression model is 36% and Random forest regression is 28%.Both model is having error even though random forest algorithm is little bit better than linear regression.

References

1: https://www.kaggle.com/c/bike-sharing-demand/kernels