IME 451

Advanced Statistical Methods

Final project

Prepared by

Mustafa Abd El-Nasser Mahmoud (120170016)

Zaid Mostafa Elashmawy (120170018)

Submission Date

Feb 1, 2021

Submitted to

Dr. Mohamed Ghieth

Introduction

The mobility services booking in the last two decades has become increasingly online. Bike-sharing has become a very attractive mobility service in urban areas in recent years given it is low-priced, flexible, environmentally friendly, healthy, and a fast alternative for covering short distances. As most of the metropolitan areas try to fight against air pollution and strive to reduce car traffic, bike-sharing services are supported or even provided in many cities by local governments.

As of May 2018, more than 1,600 bike-sharing programs were in operation worldwide, providing more than 18 million bicycles for public use. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

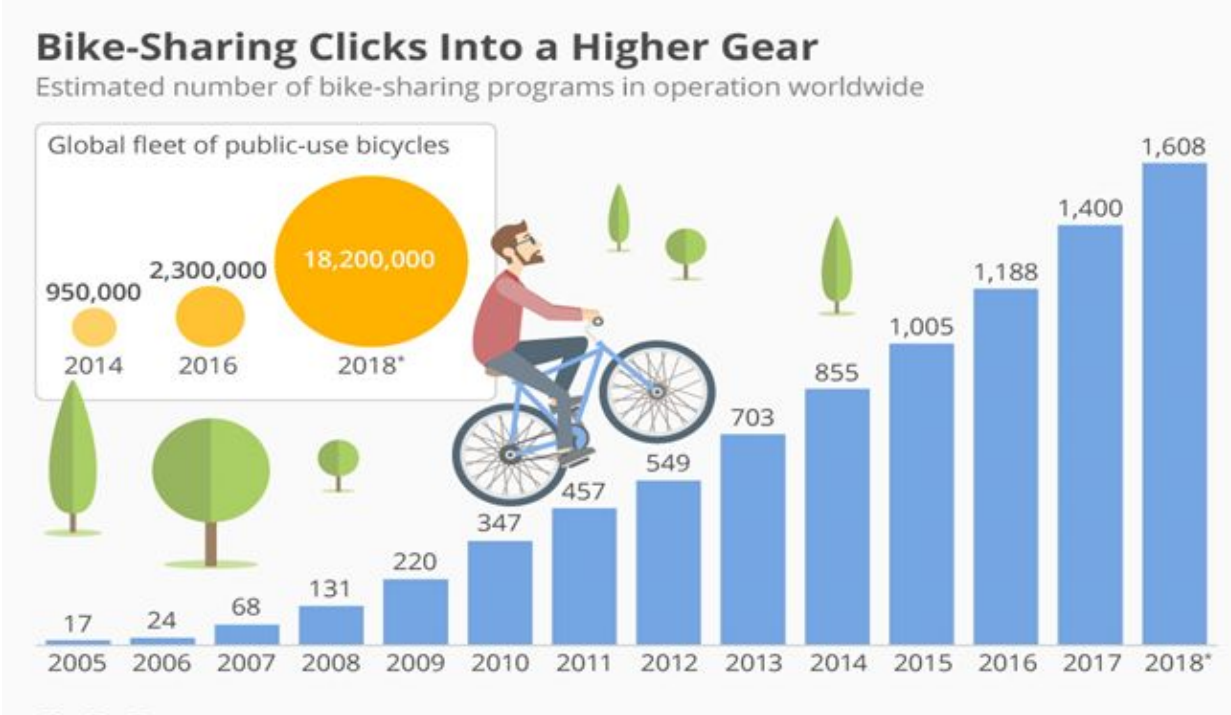


Figure 1. The number of bike-sharing programs in operation worldwide.

Part from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the project. 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 the 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 the important events in the city could be detected via monitoring these data.

Project objective

The project aims to help forecast the demand for the bike-sharing company based on the number of available attributes given in the data. This should help them to create an inventory and workforce-related strategy apart from financial and maintenance estimates. Also, because of the improving bike-sharing culture, the project can have an indirect impact on promoting health, environment, and other social impacts on society.

The Data

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants were asked to combine historical usage patterns with weather data in order to forecast bike rental demand for the data came from a two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which was obtained by the UCI Machine Learning Repository.

Data Understanding and Cleaning

The dimensions of our dataset are 731 observations for 16 variables.

Variables Information

Dependent Variable: cnt, which represents the count of total rental bikes including both casual and registered. Independent Variables:

- 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.
- weathersit:
    - Nice: Clear, Few clouds, Partly cloudy, Partly cloudy
    - Cloudy: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - Wet: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- Temp: Temperature in Celsius.
- atemp: “feels like” temperature in Celsius
- hum: Relative humidity
- windspeed: wind speed
- casual: count of casual users
- registered: count of registered users

The first step in preparing the data is by removing the unnecessary variables. Thus, we will remove “dteday”,”workingday”, “atemp” for redundancy. We will not remove the “instant” variable as we will need it in our visualizations.

Second: most of the data columns are not in the right format, thus, we will change all the categorical variables to “factor” format. Variables are (“yr”,”mnth”, “season”, “weathersit”, “weekday”,”holiday”)

Third: To make our data more readable and also so we can make our analysis, we will add new variables to the data. It will be generated from the old ones, however, it’s better to use than the old variables. For example, we will create a new variable called “weather” which is the same as the old variable “weathersit”, but instead of (1, 2, 3), the values will be (“nice”, “cloudy”, “wet”). Another example, we will add a “season” variable, but instead of values (1, 2, 3, 4), the values will be (“spring”, “summer”, “fall”, “winder”). Following the same strategy, we will add (“Year”, “month”, “Day”)

Fourth: After creating these steps, we are not in need of the old variables, thus, we will remove them.

Data Analysis

Descriptive Analysis

Descriptive analysis is an important first step for conducting statistical analysis. It gives you an idea of the distribution of your data, helps you detect outliers and typos, and enables identifying associations (correlations) among variables, thus making you ready to conduct further statistical analysis.

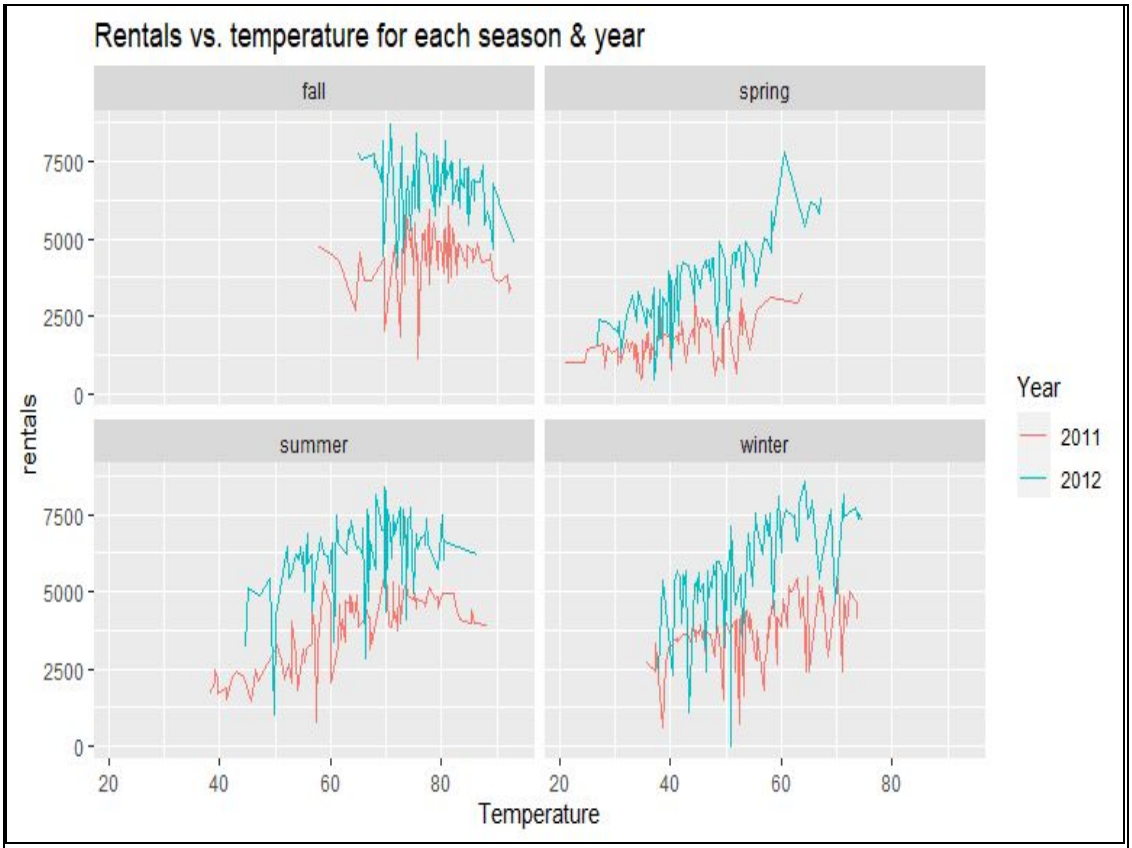


Figure.2: Total rentals vs. the temperature categorized by year and season

Figure 2 describes the behaviour of the Rentals for the company's customers, whether being registered or casual with the changing in temperature with different seasons (Summer, Winter, Fall, and Spring) and years (2011,2012). We can conclude that the number of the total rentals increases with the increasing of the temperature. Where, "fall" and "summer" achieved the highest rentals. And the number of rentals in 2012 is more than the number of rentals in 2011.

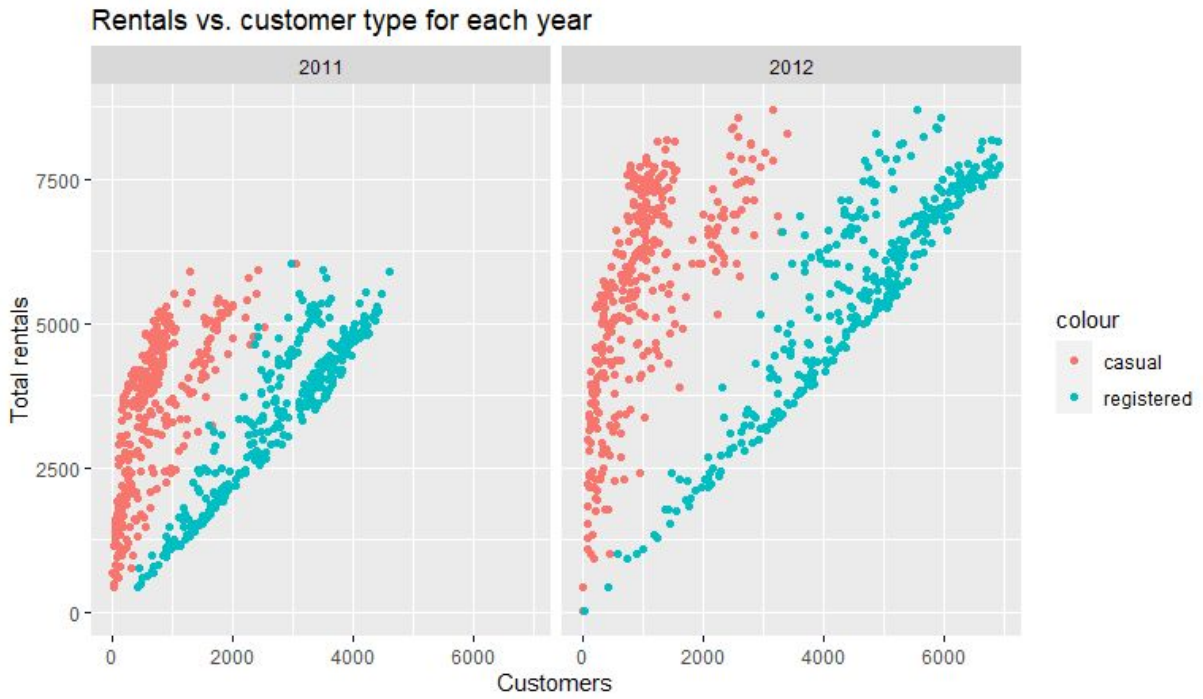


Figure.3: Rentals vs. customer type for each year

We can understand the relationship between the customer type and total rentals form figure 3. These two graphs describe the relation between the total Rentals and the two different types of customer rentals (Registered and Casual) for each year (2011 , 2012).

As shown in the two graphs we see that the registered ridership is high in comparison to the casual ridership. And also there is a separation between the data points for each type, this is due to the demand changing according to the seasons of the year.

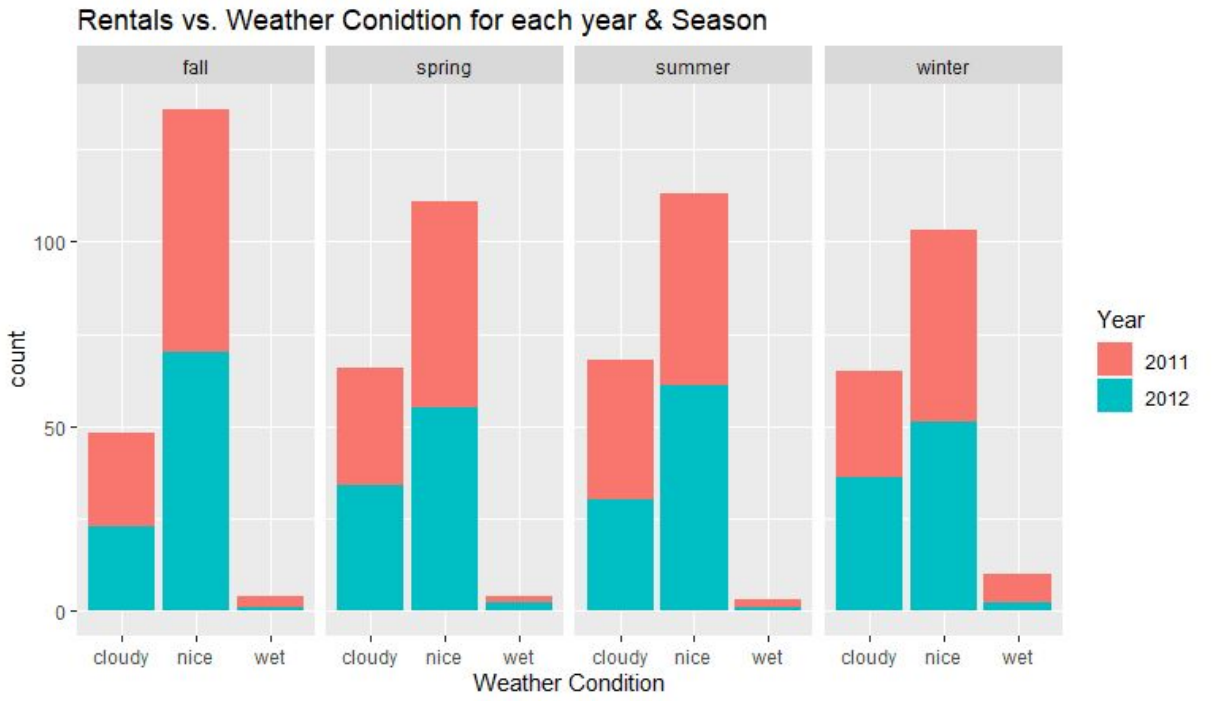


Figure.4: Rentals vs. Weather conditions for each year & season

The weather changes relative to the temperature and the season of the year, thus, affecting the demand on rentals. We can understand this relationship from figure 4. These histograms show the count of rentals in different seasons and different years.The conclusion of these histograms is that the number of rentals increase when the weather condition is “nice” in all four seasons of the year. And generally, the demand increases with the increasing of the temperature. Hence, the demand is high at “fall” and “summer” seasons for each year.

Predictive Analysis

Predictive analytics is the use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future. Though predictive analytics has been around for

decades, it's a technology whose time has come. More and more organizations are turning to predictive analytics to increase their bottom line and competitive advantage.

Why is predictive analytics important? Organizations are turning to predictive analytics to help solve difficult problems and uncover new opportunities. Common uses include:

Detecting fraud. Combining multiple analytics methods can improve pattern detection and prevent criminal behavior. **Optimizing marketing campaigns.** Predictive analytics are used to determine customer responses or purchases, as well as promote cross-sell opportunities. Predictive models help businesses attract, retain and grow their most profitable customers.

Improving operations. Many companies use predictive models to forecast inventory and manage resources. Airlines use predictive analytics to set ticket prices. Hotels try to predict the number of guests for any given night to maximize occupancy and increase revenue. Predictive analytics enables organizations to function more efficiently.

Reducing risk. Credit scores are used to assess a buyer's likelihood of default for purchases and are a well-known example of predictive analytics. A credit score is a number generated by a predictive model that incorporates all data relevant to a person's creditworthiness.(https://www.sas.com/en_us/insights/analytics/predictive-analytics.html).

We will build a model using a training set and applying this model on a Test set. We will compare the forecasted values of the rentals generated from our models with the real observations to measure the performance of the model.

Seasonal naive technique

Any forecasting method should be evaluated by being compared to a baseline model. In our case, we will use a seasonal naive method as our baseline. This helps ensure that the efforts we will put on the upcoming models are worth it in terms of performance. The simplest of all methods is called simple naive. Extremely simple: the forecast for tomorrow is what we are observing today.

Another approach, seasonal naive, is a little more reliable. The forecast for tomorrow is what we observed the week/month/year (depending what horizon we are working with) before.

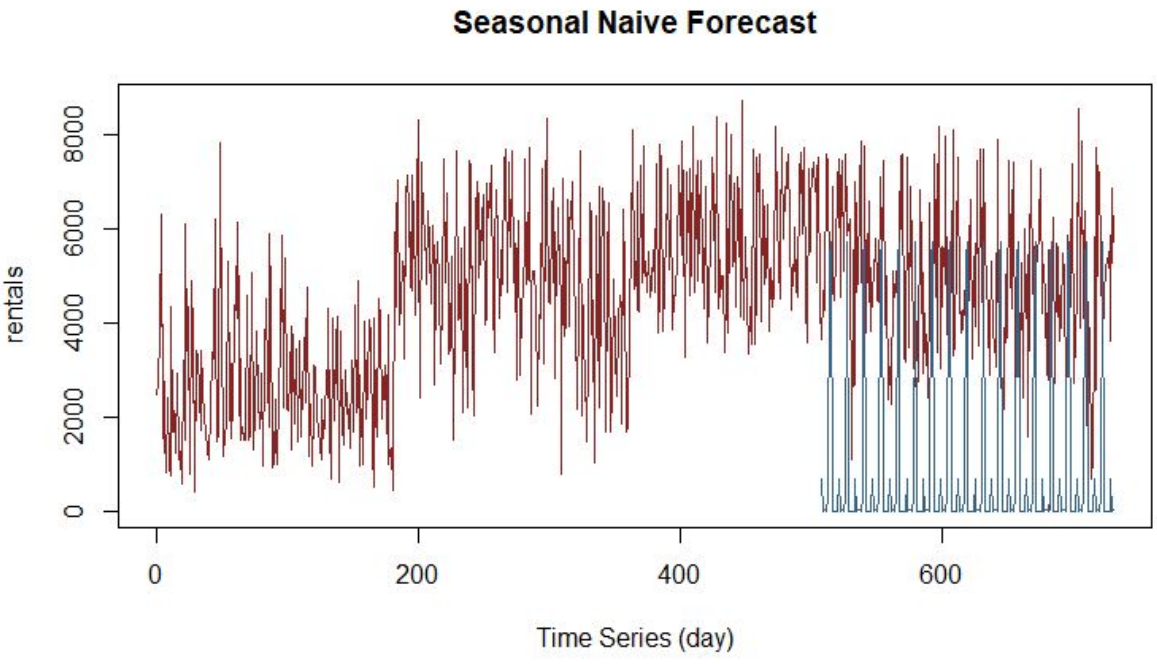


Figure.5: Seasonal Naive forecasting technique

Figure.5 illustrates the fitting of the seasonal naive model on the test data. The accuracy of this prediction is low and the values of R2 = 51% and MAPE = 10608.

Demand forecasting by multilinear regression model

Linear regression is a basic and commonly used type of 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 (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

The base model in which we used all the numeric variables to predict the rental amount, showed that there is a multicollinearity between the independent variables. The correlation matrix explain the correlation between the variables as shown

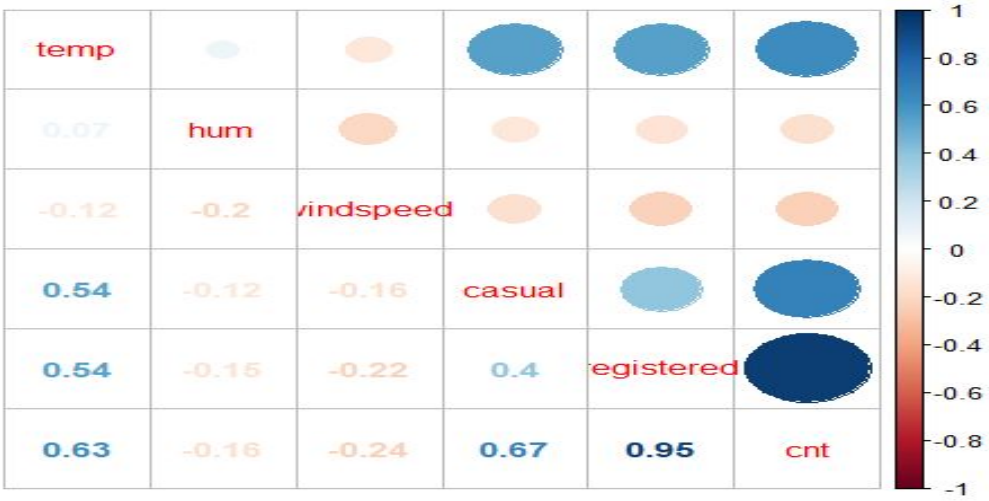


Figure.6: correlation matrix between the dependent and independent variables

The correlation matrix in figure 6 indicates that there is a high positive correlation between the dependent variable (cnt) and the independent variables (registered, casual, temp) and natural negative correlation between cnt and (hum, windspeed). In addition, there is a high positive correlation between temperature and (casual, registered). Thus, we need to consider all these correlation relationships in our model.

Building a model without multicollinearity (model_3)

After a process of Permutations and combinations between the independent variables to find the best model with the lowest errors, we concluded that the best model that can predict the amount of rentals with the least errors is the one in which we used (temp, hum, windspeed, registered) independent variables in it. The model resulted in Adjusted R2 = .91 And the prediction model has a good performance, R2 = .901, RMSE=614

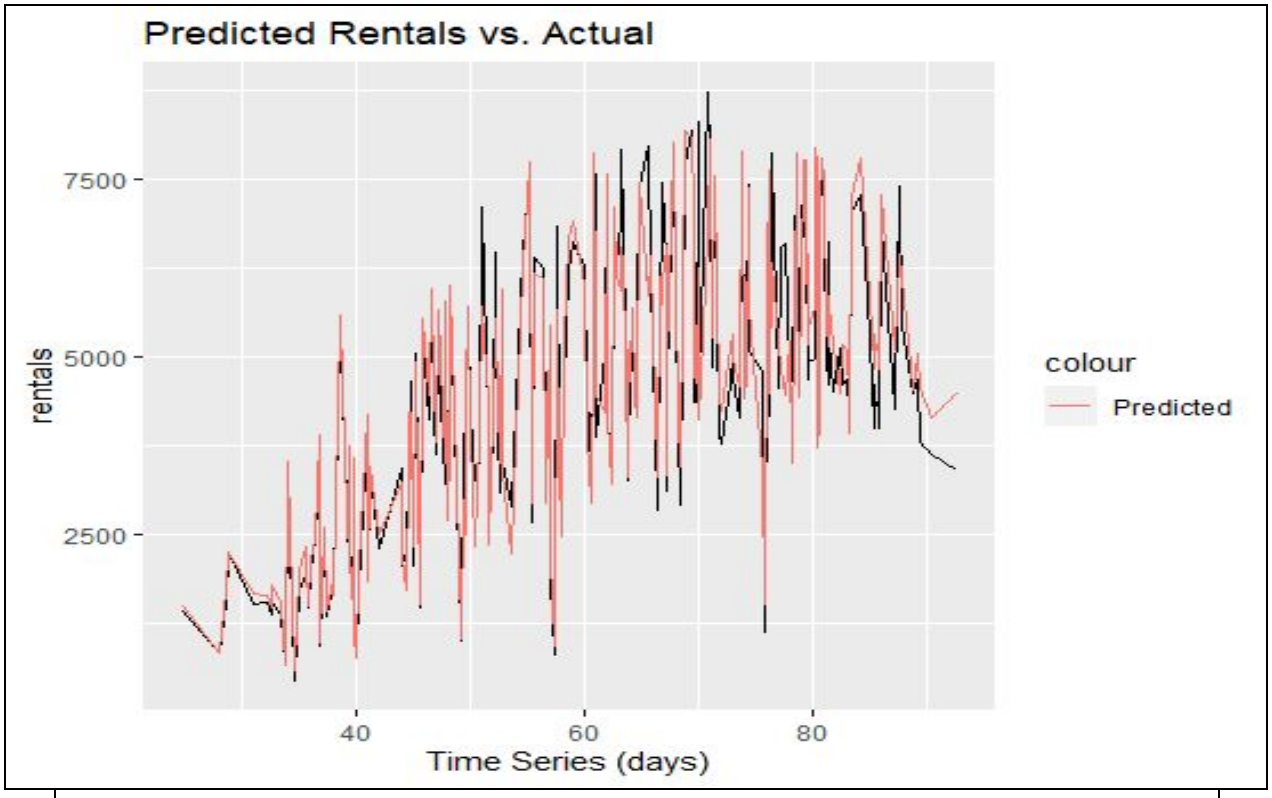


Figure.7: predicted amount of rentals vs. actual data using a multilinear model

Modelling using Logistic Model (model_4)

The data contained a number of categorical independent variables, what can we get from a model dependent on such variables? First, we tried all the categorical variables at once, however, we observed multicollinearity. So we used the certain categorical variables, which are (Weather, Season, Year, month). We have got an accepted logistic model with a prediction accuracy of R2= 0.78, and RMSE= 934.

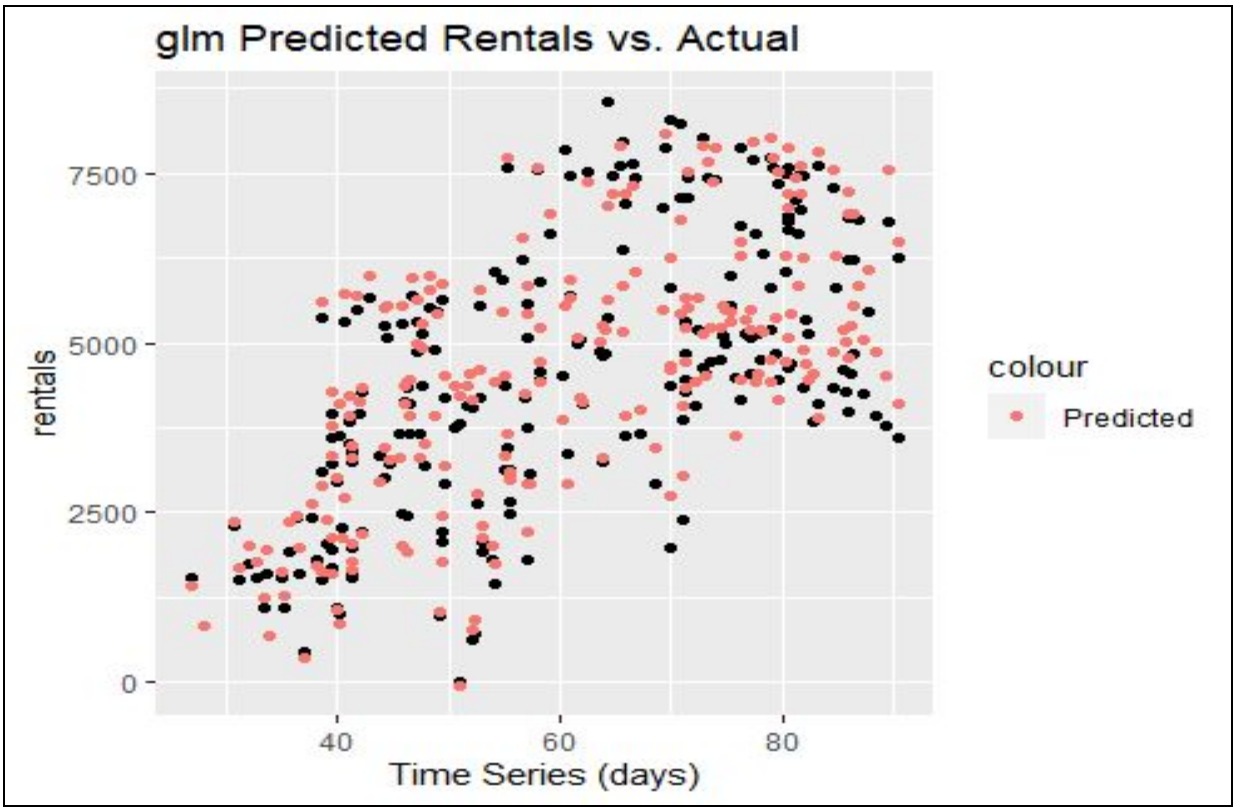
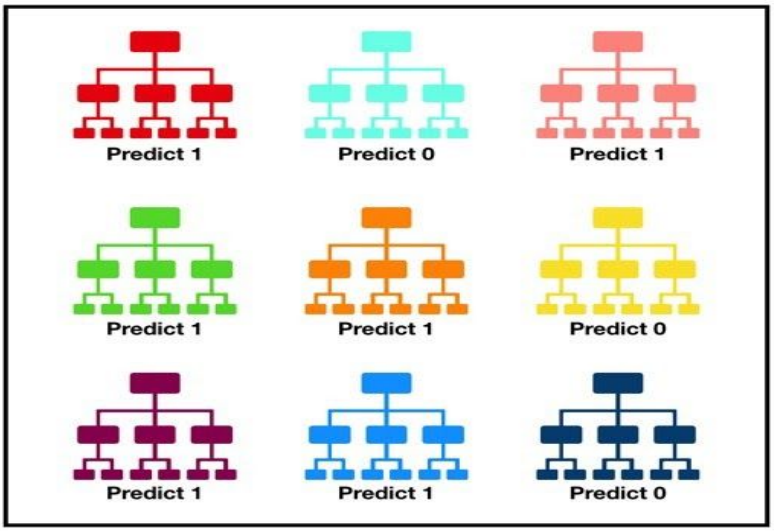


Figure.8: predicted amount of rentals vs. actual data using a logistic model

Random Forest Method

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below). We will apply this concept to our bike sharing demand forecasting so we can get a robust model.



We will use the same independent variables we used in model_3. And apply the model as a predictor to the demand of rentals.

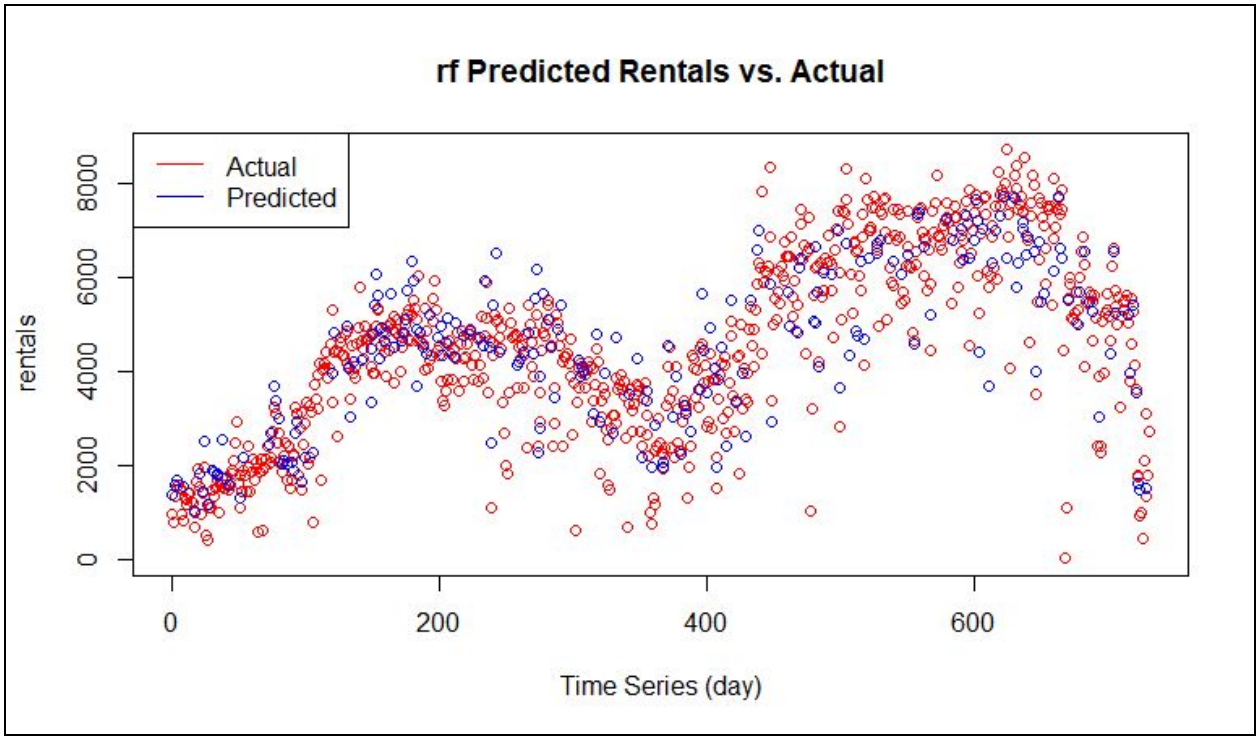


Figure.9: Predicted vs. Actual rentals using rf model

The plot in figure 9 shows the prediction of the demand for rentals for bike-sharing according to the factors (temperature, humidity, wind speed and the registered customers). the amount of rentals get affected negatively by bad weather or low temperature. The model indicates high accuracy and the performance relative to value of R2 = 89.6%

Finally, we have created a custom function based on the criteria for appropriate weather for biking where you input the temperature, wind speed, and the weather condition of today = the function will output the probability of renting a bike for that day. In other words, it will output the probability of were appropriate for biking that day.

Conclusion

Conducting a range of different statistical test and plotting the data with variety of plots on the dataset comprising two-year historical log on bike rentals in Washington D.C. allow to make the following conclusions:

1. The mean temperatures vary significantly over the seasons
2. Figures of total bike rents changes depending on weather condition and vary regarding their means. The most significant pairwise mean difference is typical for spring, summer, and fall seasons, while the most insignificant for winter.
3. Weather condition and total number of bike rentals also seemed to be significantly correlated. The two popular weather conditions for bike rentals are nice and Cloudy weather.
4. There exists a significant correlation between the number of total bike rentals and (casual, registered, temp) variables.

To sum it up, we can say, that the total amount of bike rentals is dependent on the weather and also on the users' status (registered vs. casual).

Future Scope

1. Currently, the demand forecasting was done on a daily basis. As our hourly demand data is also available on UCI ML data repository, we can perform hourly demand prediction as well.
2. Attributes having similar coefficients can be clubbed together and considered as a single feature. This helps in reducing the number of features and decreasing model complexity.
3. Given the right data, the scope of the study can be increased to other regions to gather more data points or to test model accuracy.
4. Exploration of other feature selection methods like forward and backward stepwise can also be incorporated.