AUGUST 8, 2018

BIKE RENTING DAILY COUNT PREDICTION

PREDICTION AND ANALYSIS DONE IN PYTHON & R

LAKSHVEER SINGH

Table of Contents

1. Introduction........ 2
1.1. Problem Statement 2
1.2. Data 2
2. Data Pre-Processing 3
2.1 Data Size and Structure 3
2.2 Outlier Analysis........ 3
2.2.1. Distribution of numeric variables........ 3
2.3. Feature Selection/ Importance 7
2.3.1. Correlation between continuous variable........ 7
2.3.2. Correlation between Categorical Variable 8
2.3.3. Feature Importance:........ 9
3. Exploratory Data Analysis........ 10
3.1. Univariate Analysis 10
4. Modelling 13
4.1. Visualizations........ 14
4.1.1. Variable Importance:........ 14
4.1.2. Partial dependence Plot 15
4.2. Support and Contradict features........ 17
5. Conclusion 19
5.1. Model Evaluation 19
5.1.1. RSquare on CV fold and Fitted model 19
5.1.2. Root Mean Square Error........ 19
5.1.3. Mean Absolute Error 19
6. References........ 20

Chapter 1

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

The below data represents the first five observations from the raw data and features associated with it.

Table 1.1: Bike renting sample data (Columns: 1-11)

							WORKING	*WEATHER*		
INSTANT	***DATE***	***SEASON***	***YEAR***	***MONTH***	***HOLIDAY***	***WEEKDAY***	***DAY***	***TYPE***	***temp***	***Atemp***
1	1/1/2011	1	0	1	0	6	0	2	*0.344167*	*0.363625*
2	1/2/2011	1	0	1	0	0	0	2	*0.363478*	*0.353739*
3	1/3/2011	1	0	1	0	1	1	1	*0.196364*	*0.189405*
4	1/4/2011	1	0	1	0	2	1	1	*0.2*	*0.212122*
5	1/5/2011	1	0	1	0	3	1	1	*0.226957*	*0.22927*
6	1/6/2011	1	0	1	0	4	1	1	*0.204348*	*0.233209*

Table 1.2: Bike renting sample data (Columns: 12-16)

				TOTAL
HUMIDITY	***WINDSPEED***	***CASUAL***	***REGISTERED***	***COUNT***
0.805833	*0.160446*	*331*	*654*	*985*
0.696087	*0.248539*	*131*	*670*	*801*
0.437273	*0.248309*	*120*	*1229*	*1349*
0.590435	*0.160296*	*108*	*1454*	*1562*
0.436957	*0.1869*	*82*	*1518*	*1600*
0.518261	*0.0895652*	*88*	*1518*	*1606*

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

Table 1.3: Predictor variables

S.No.	**Predictor**
1	instant
2	date
3	season
4	year
5	month
6	Holiday
7	weekday
8	workingday

Table 1.4: Predictor variables

S.No.	**Predictor**
9	weather_type
10	temp
11	atemp
12	humidity
13	windspeed
14	casual
15	registered

Chapter 2

2. Data Pre-Processing

*Data Pre-Processing refers to cleaning the data for missing values, outliers, knowing the relationship within predictors and their relationship with target variable. We have used outlier analysis, co-relation plots, probability distribution and feature importance using three techniques: **Selection Stability method with Randomized Lasso, recursive feature elimination using linear regression and random forest**.*

2.1 Data Size and Structure

Table 2.1: Data size and structure

'DATA.FRAME'	***731 OBS. OF 16 VARIABLES***
$ INSTANT	int 1 2 3 4 5 6 7 8 9 10
$ DTEDAY	chr "2011-01-01" "2011-01-02" "2011-01-03" "2011-01-04"
$ SEASON	int 1 1 1 1 1 1 1 1 1 1
$ YR	int 0 0 0 0 0 0 0 0 0 0
$ MNTH	int 1 1 1 1 1 1 1 1 1 1
$ HOLIDAY	int 0 0 0 0 0 0 0 0 0 0
$ WEEKDAY	int 6 0 1 2 3 4 5 6 0 1
$ WORKINGDAY	int 0 0 1 1 1 1 1 0 0 1
$ WEATHERSIT	int 2 2 1 1 1 1 2 2 1 1
$ TEMP	num 0.344 0.363 0.196 0.2 0.227
$ ATEMP	num 0.364 0.354 0.189 0.212 0.229
$ HUM	num 0.806 0.696 0.437 0.59 0.437
$ WINDSPEED	num 0.16 0.249 0.248 0.16 0.187
$ CASUAL	int 331 131 120 108 82 88 148 68 54 41
$ REGISTERED	int 654 670 1229 1454 1518 1518 1362 891 768 1280
$ CNT	int 985 801 1349 1562 1600 1606 1510 959 822 1321

2.2 Outlier Analysis

2.2.1. Distribution of numeric variables

We can see from these probability distribution that there is not much skewness in the data. Having said that casual and *windspeed* are among the top with skewness, both are right skewed. There is bit of left skewness in the *humidity*, rest all looks good enough. The skewness in these distribution can be most likely explained by the presence of outliers and extreme values in the data.

In Fig 2.1 we have plotted a box plot for these all numeric variables, which confirms the presence of outliers in *casual*, *windspeed* and *humidity* (descending order). We will not remove the outlier in casual as it will not be in the list of predictor variables in the model generation, we need to adjust our target variable which is *total count* if we are treating outliers in the *casual* variable as total count is the sum of casual and registered. The total count box doesn't have any outlier so we don't want to change anything there as it's our point of study.

Figure 2.1: Probability distribution of continuous variable

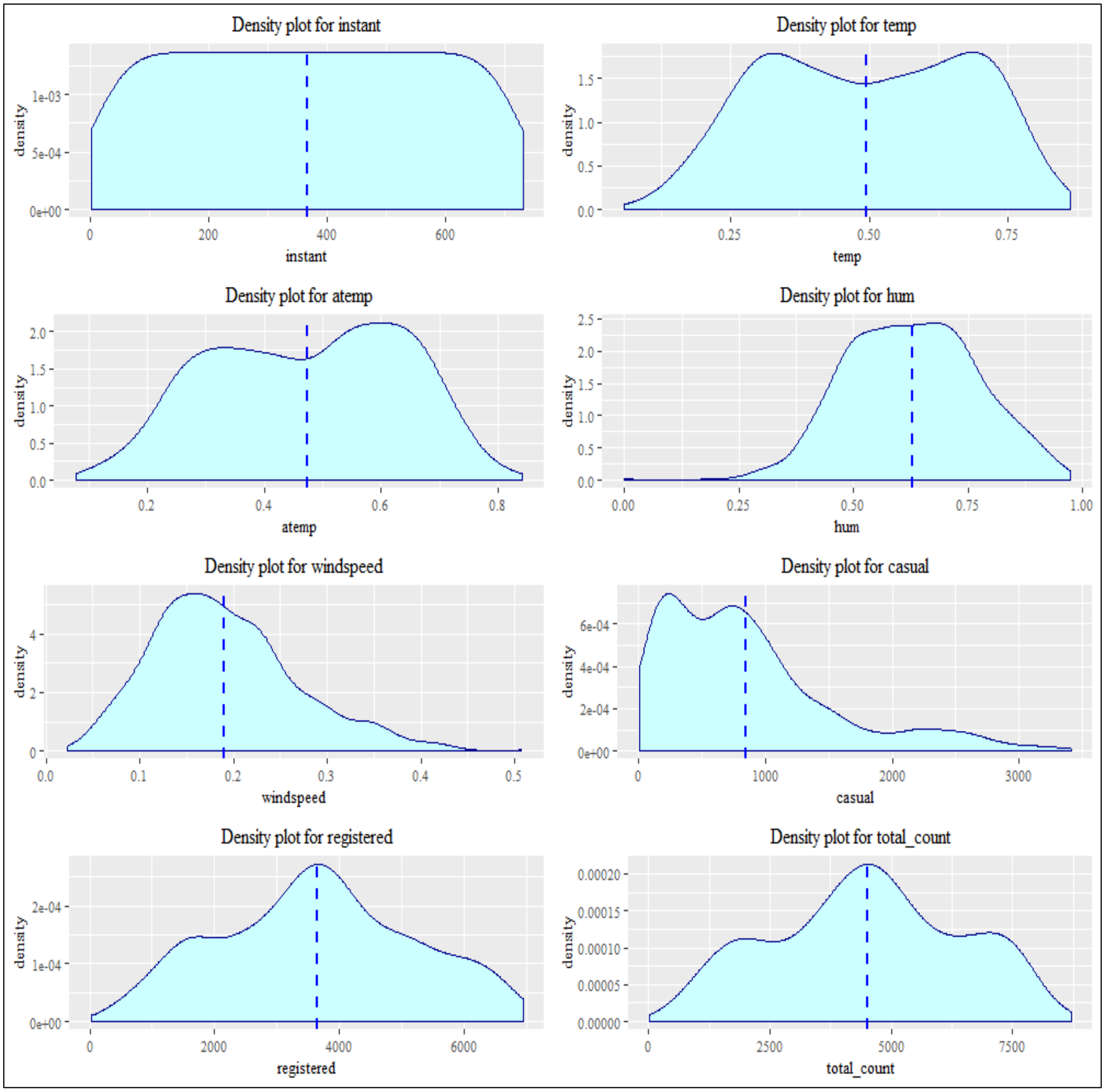
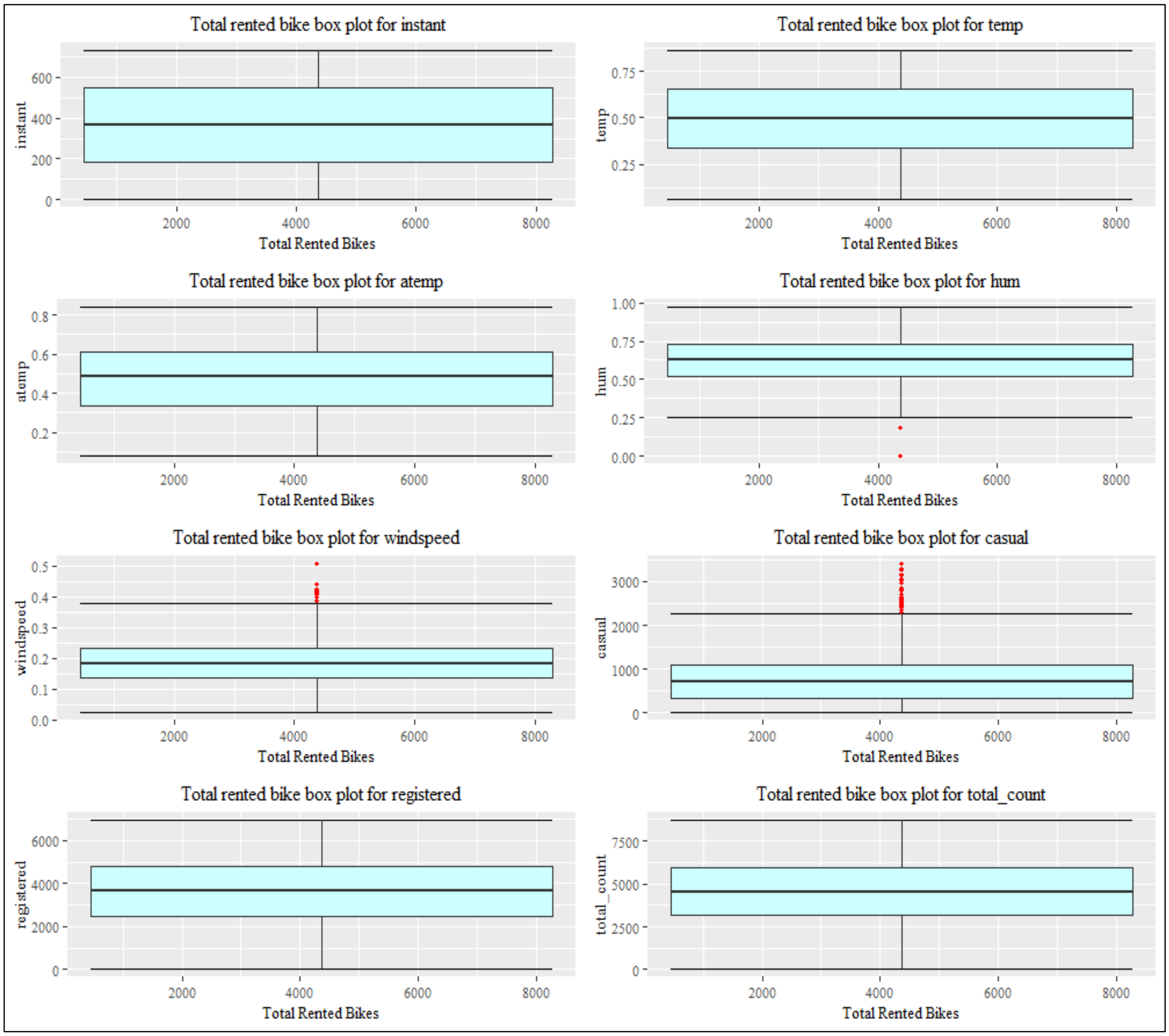


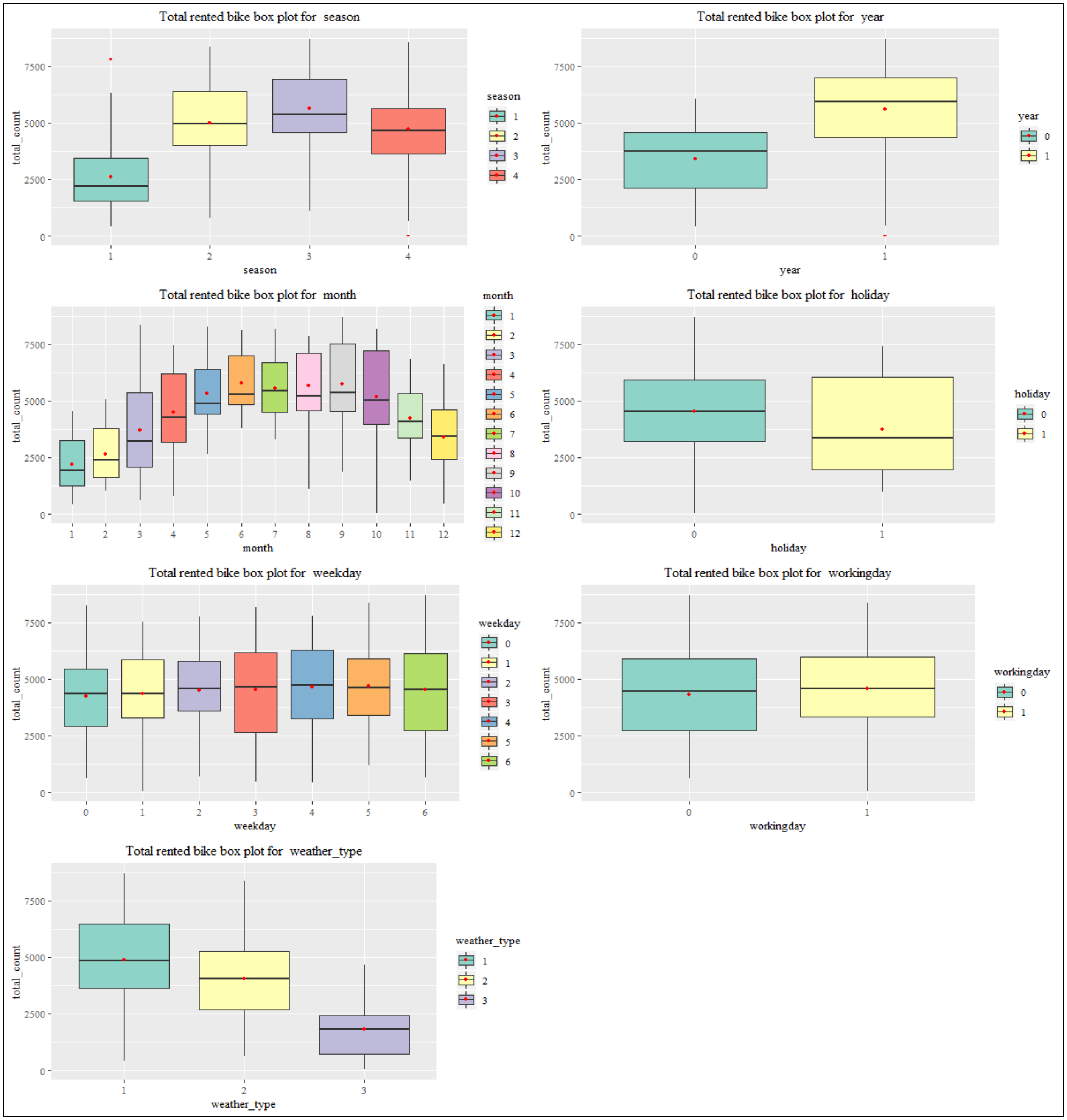
Figure 2.2: Box Plot of continuous variable



In fig 1.2 we have plotted boxplots of all the 7 categorical variable with respect to total count of rented bikes. Following inferences can be made for the box plot apart for the presence of outlier in fig

- Spring season has got relatively lower count. The dip in the median and mean value gives evidence for it.
- The month plot show that people tend to rent bike during summer season since it is really conductive to ride bike at that season. Therefore June, July, August has got relatively higher demand for bike.
- The demand of bike is much higher when weather is clear, few clouds , partly cloudy(i.e. weather type 1) as compared to weather type 3 which is light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds and relatively higher than weather 2 which is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.
- There is an increasing trend as 2012 has much higher demand for renting a bike then year 2011.

Figure 2.3: Box Plot of categorical variables with target variable

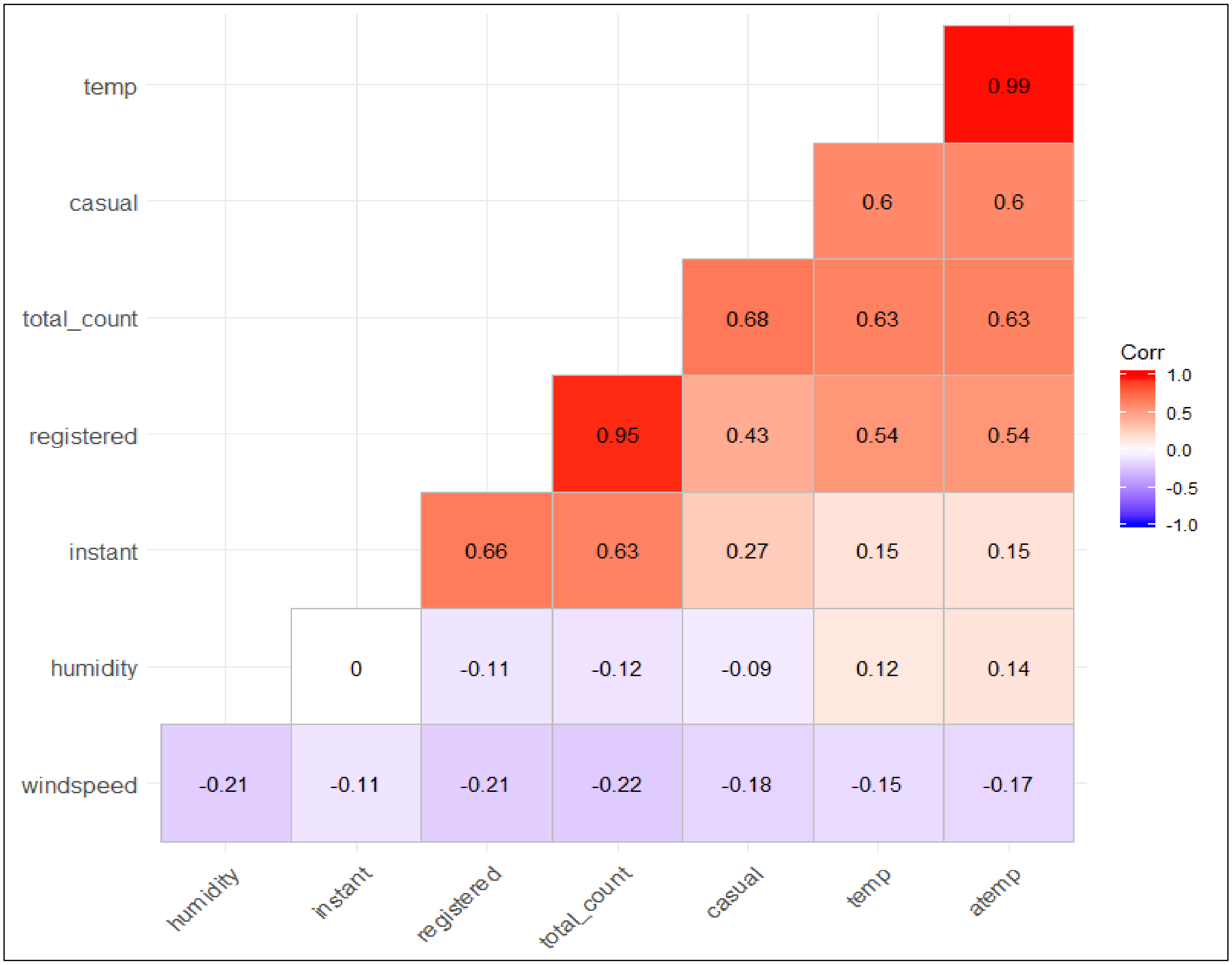


2.3. Feature Selection/ Importance

2.3.1. Correlation between continuous variable

One common to understand how a dependent variable is influenced by features (numerical) is to find a correlation matrix between them. Let's plot a correlation plot between count and numerical variables.

Figure 2.4: Pearson co-relation plot



- "humidity" is not gonna be really useful numerical feature and it is visible from it correlation value with "total_count"
- "atemp" is variable is not taken into since "atemp" and "temp" has got strong correlation with each other and both have almost same correlation with total count. During model building any one of the variable has to be dropped since they will exhibit multicollinearity in the data.
- "windspeed" has got negative correlation with count. Although the correlation between them are not very prominent still the total count variable has got little dependency "humidity".
- "Casual" and "Registered" are also not taken into account since they are leakage variables in nature and need to dropped during model building.

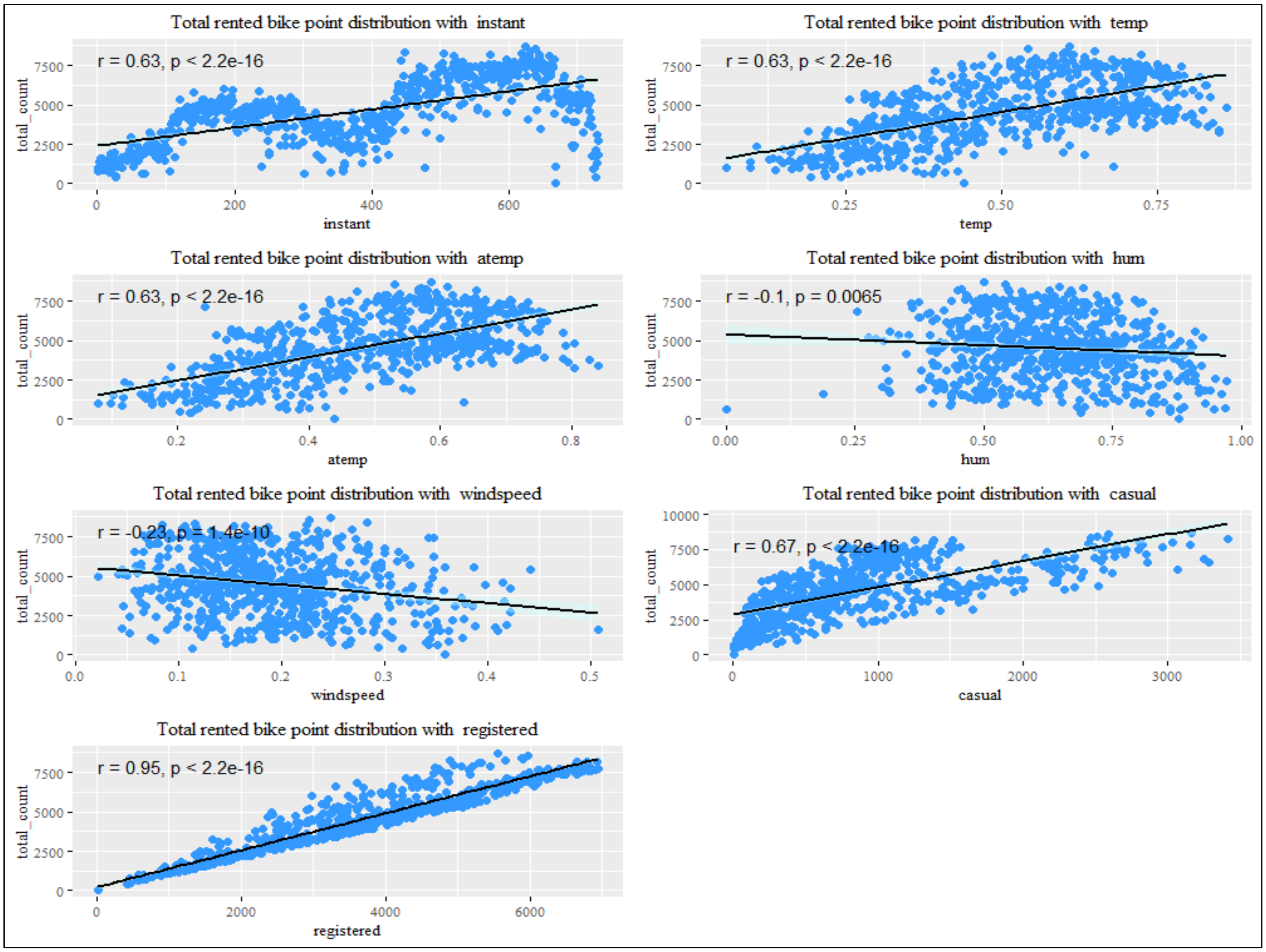


Figure 2.4: scatter plot with regression line

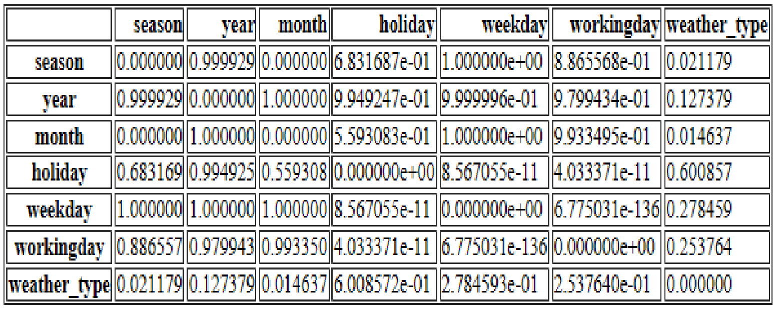
Point plot which regression line is one useful way to depict the relationship between these two features.

2.3.2. Correlation between Categorical Variable

- Some inferences from Table 2.2 chi-square test:
- p value between season and weather type is less than 0.05, which means both features depend on each other which seems to be quite obvious as there is a specific weather in specific season.
- Holiday, weekday and working day has got very less p value which means they are highly related with each other, which again is quite obvious as when they is a holiday; it's not a working day and it's has highly significant with weekday as majority of holiday falls on Monday(1).
- Weekday and working day are very highly associated with each other as weekday (0 and 6) are non-working day all the time.

If p value is less than 0.05 than we reject the null hypothesis saying that these two values depend on each other.

Table 2.2:Chi-square test b/w categorical variables



	season	year	month	holiday	weekday	workingday	weather_type
season	0.000000	0.999929	0.000000	6.831687e-01	1.000000e+00	8.865568e-01	0.021179
year	0.999929	0.000000	1.000000	9.949247e-01	9.999996e-01	9.799434e-01	0.127379
month	0.000000	1.000000	0.000000	5.593083e-01	1.000000e+00	9.933495e-01	0.014637
holiday	0.683169	0.994925	0.559308	0.000000e+00	8.567055e-11	4.033371e-11	0.600857
weekday	1.000000	1.000000	1.000000	8.567055e-11	0.000000e+00	6.775031e-136	0.278459
workingday	0.886557	0.979943	0.993350	4.033371e-11	6.775031e-136	0.000000e+00	0.253764
weather_type	0.021179	0.127379	0.014637	6.008572e-01	2.784593e-01	2.537640e-01	0.000000

2.3.3. Feature Importance:

Knowing the important feature always help in performing the best exploratory analysis and help in selecting the important feature for model building. However, our purpose here is to find the important features and analyzing them first. In Fig we can see that year and *atemp* are among the top predictors who are contributing much information in predicting the target variable. *Humidity* and *windspeed* contributes the same.

The below importance plot has been ranked according to the mean score obtained from ***Stability Selection using randomized Lasso, Recursive feature elimination using linear regression and random forest.***

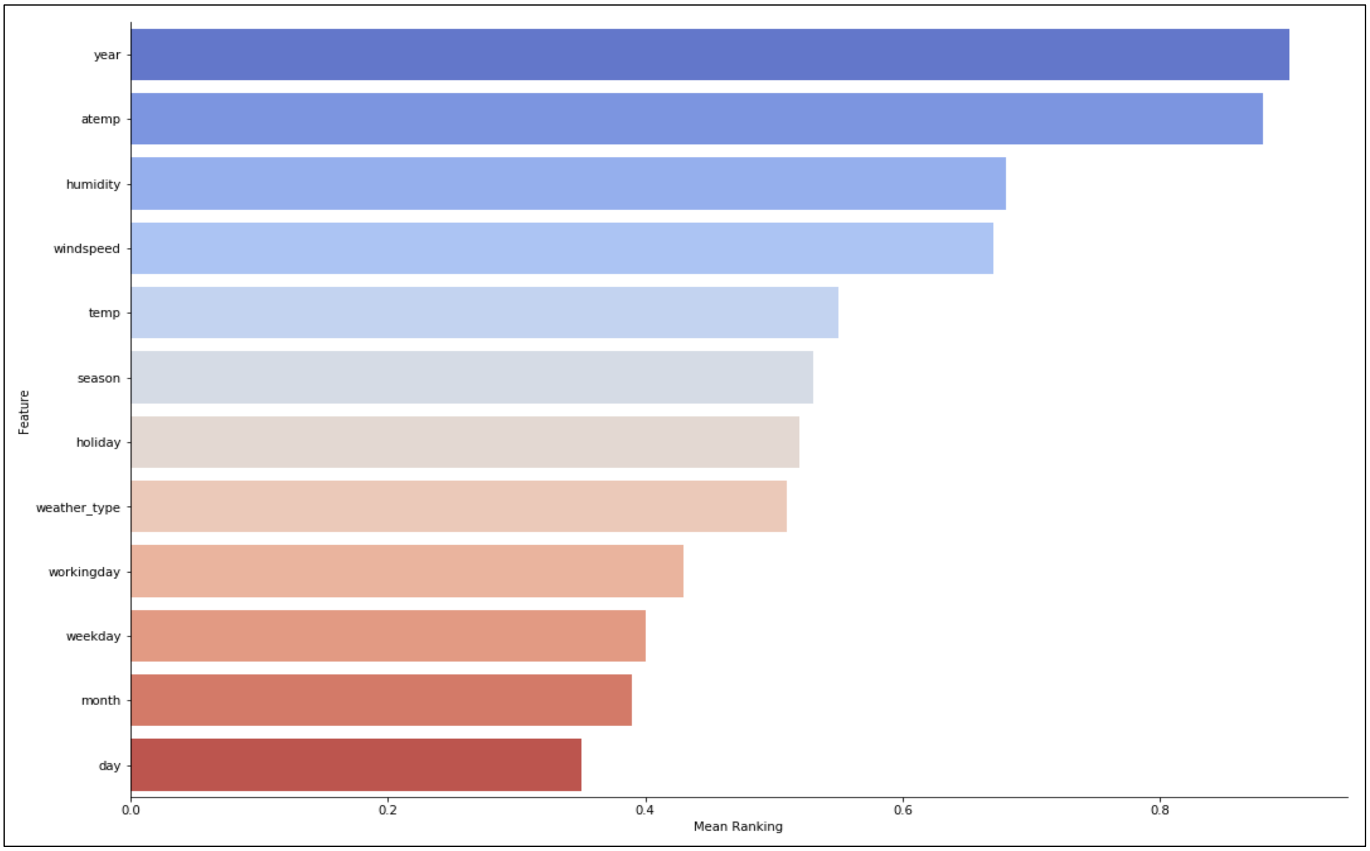


Figure 2.5: Feature importance

Chapter 3

3. Exploratory Data Analysis

3.1. Univariate Analysis

Figure 3.1: Univariate Analysis

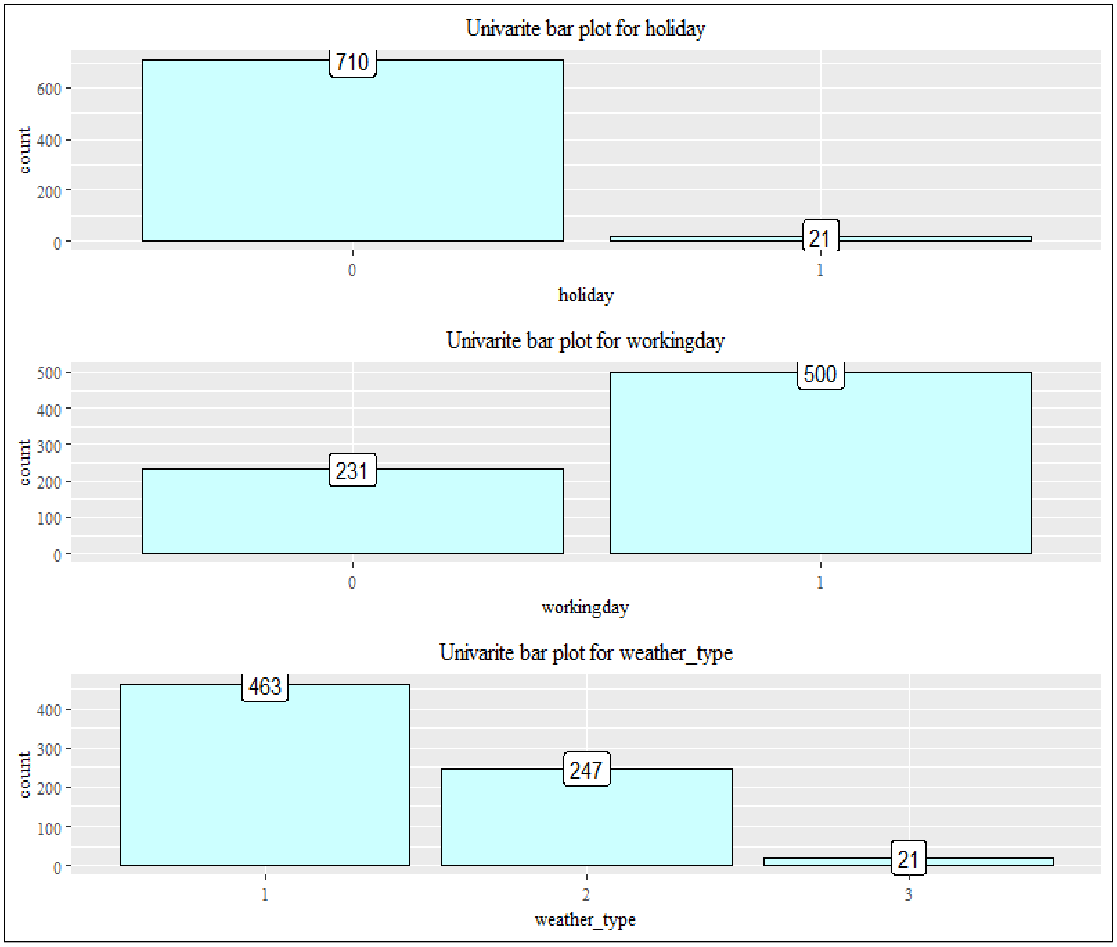
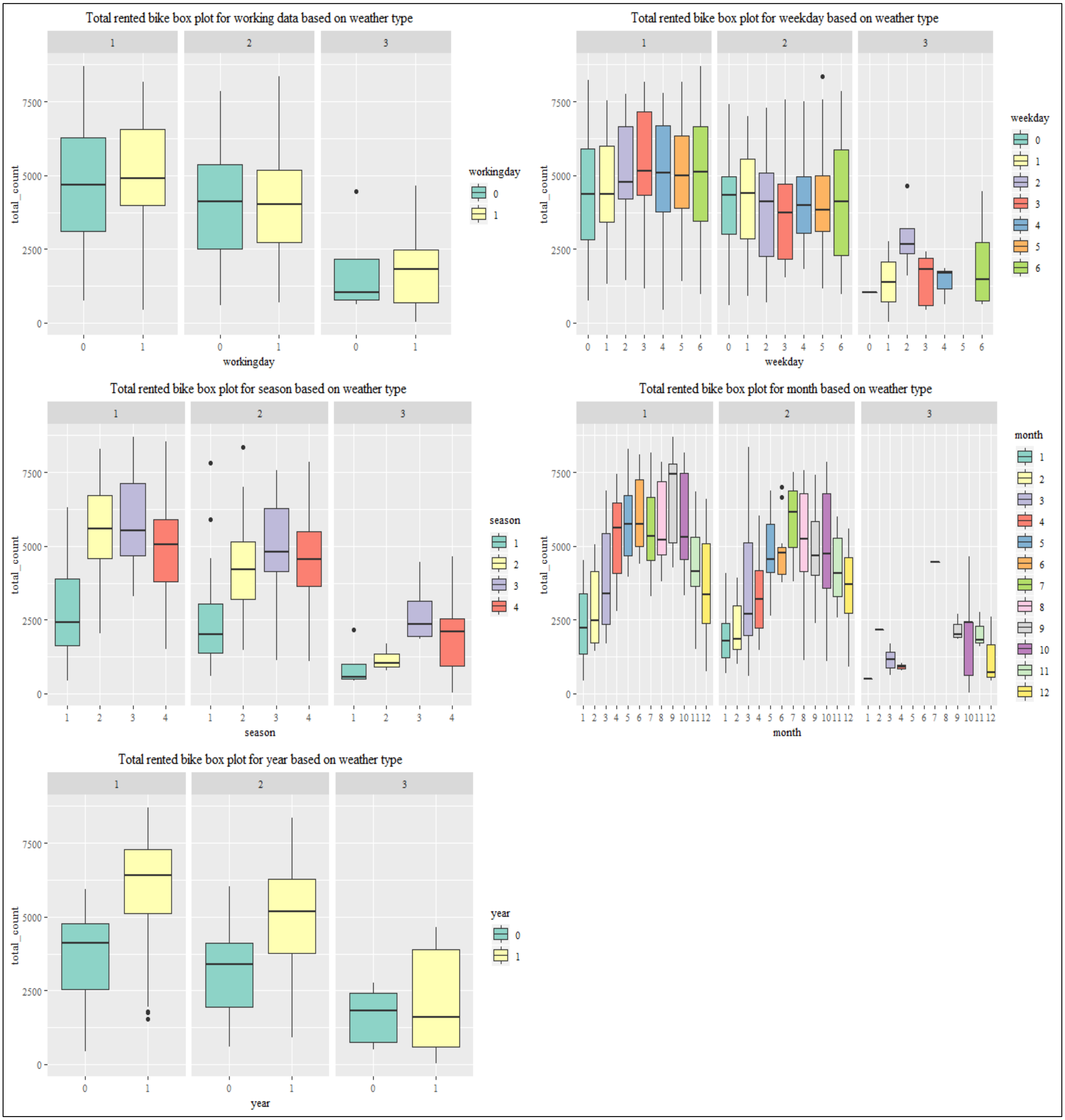


Fig 3.2 shows the difference in the mean of total count based type of working day, weather type, season, month and year.

Figure 3.2: Box plot for three variables



Total rented bike box plot for working data based on weather type

Total rented bike box plot for weekday based on weather type

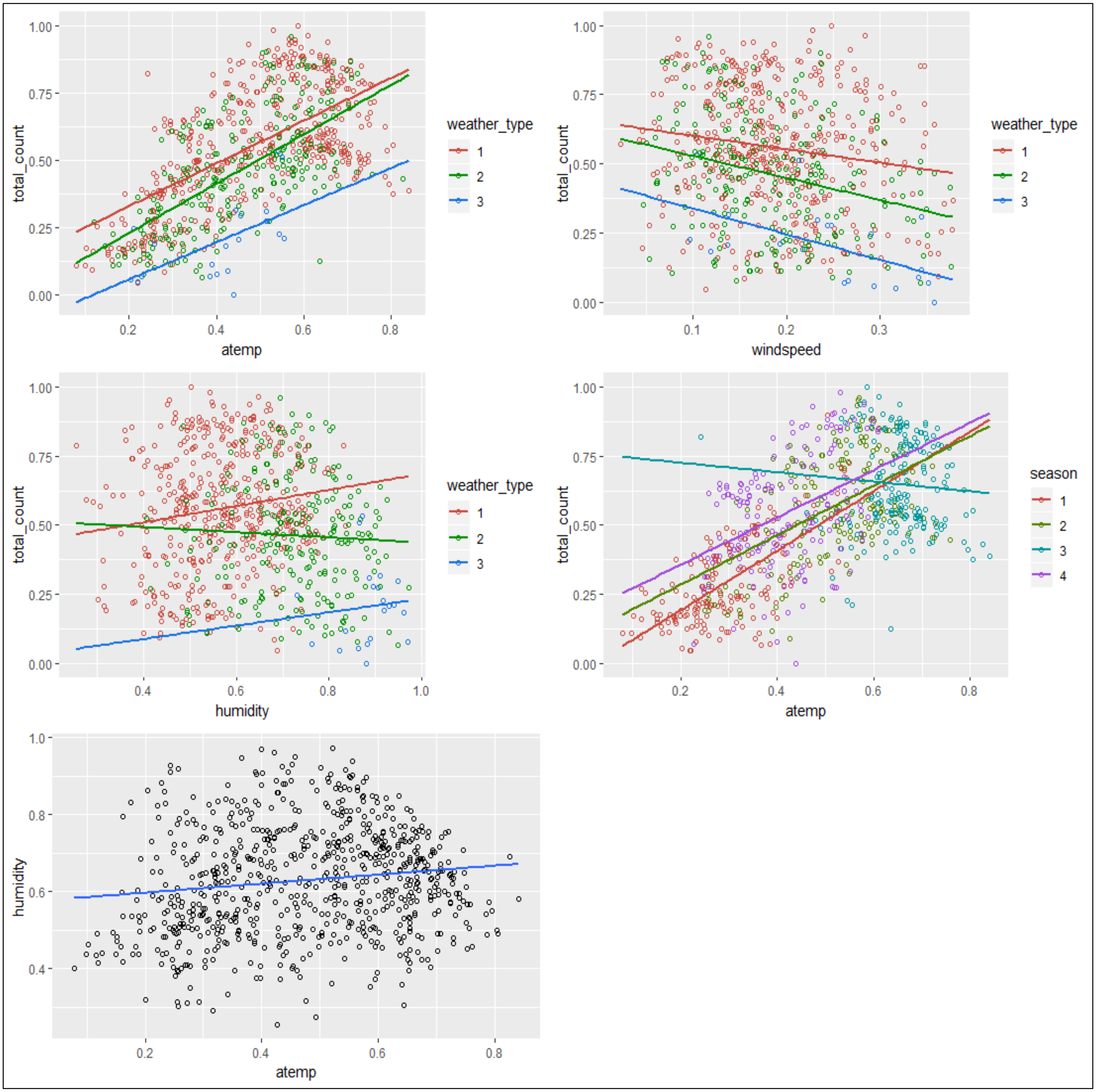
Total rented bike box plot for season based on weather type

Total rented bike box plot for month based on weather type

Total rented bike box plot for year based on weather type

We have plotted scatterplot based and the regression line based on the different weather type and season. We can see that for atemp the regression line is always increasing one but for windspeed it's decreasing one, which means as the atemp increases in any weather type our total_count decreases whereas as the windspeed increases our total_count decreases on any weather type.

Fig3.3 Point plot for three variables



Chapter 4

4. Modelling

We have a combination of categorical and continuous data in our predictors. Both contributing in it's own way in defining the total count. We have fixed Gradient Boosting Regression for our model as it gives us the best accuracy that no other model can beat. Gradient boosted machines (GBMs) are an extremely popular machine learning algorithm that have proven successful across many domains. Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful "committee" that are often hard to beat with other algorithms.

Table 4.1: gbm model parameters

```
gbm(formula = train$total_count ~ ., distribution = "gaussian",
    data = train, n.trees = 100, interaction.depth = 5, shrinkage = 0.1,
    cv.folds = 10, verbose = TRUE, n.cores = NULL)
A gradient boosted model with gaussian loss function.
100 iterations were performed.
The best cross-validation iteration was 86.
There were 10 predictors of which 10 had non-zero influence.
```

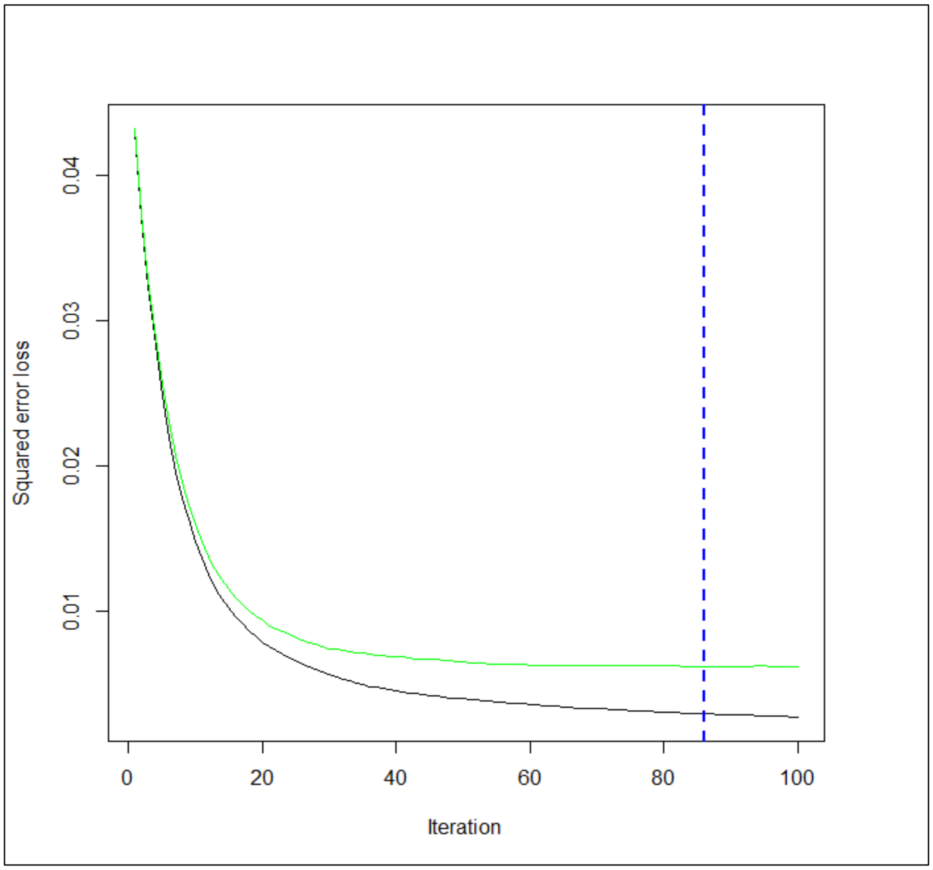
Table 4.2:

Iter	TrainDeviance	ValidDeviance	StepSize	Improve
1	0.0431	nan	0.1	0.0067
2	0.0371	nan	0.1	0.0058
3	0.0324	nan	0.1	0.0047
4	0.0283	nan	0.1	0.004
5	0.0249	nan	0.1	0.0032
6	0.022	nan	0.1	0.0028
7	0.0196	nan	0.1	0.0021
8	0.0178	nan	0.1	0.0019
9	0.0163	nan	0.1	0.0014
10	0.0148	nan	0.1	0.0013
20	0.0079	nan	0.1	0.0003
40	0.0045	nan	0.1	0
60	0.0036	nan	0.1	0
80	0.0031	nan	0.1	0
100	0.0027	nan	0.1	0

The minimum cross-validation error on the train data: 0.07867435

Loss function as the result of 100 trees added to the ensemble.

Figure 4.1:Gbm Iteration Vs Squared error loss



4.1. Visualizations

4.1.1. Variable Importance:

After running the model we likely want to understand the variables that have the largest influence on total_count. The default method used by gbm for computing variable importance is with relative influence. At each split in each tree, gbm computes the improvement in the split-criterion (MSE for regression). gbm then averages the improvement made by each variable across all the trees that the variable is used. The variables with the largest average decrease in MSE are considered most important.

Figure 4.2: gbm relative influence plot

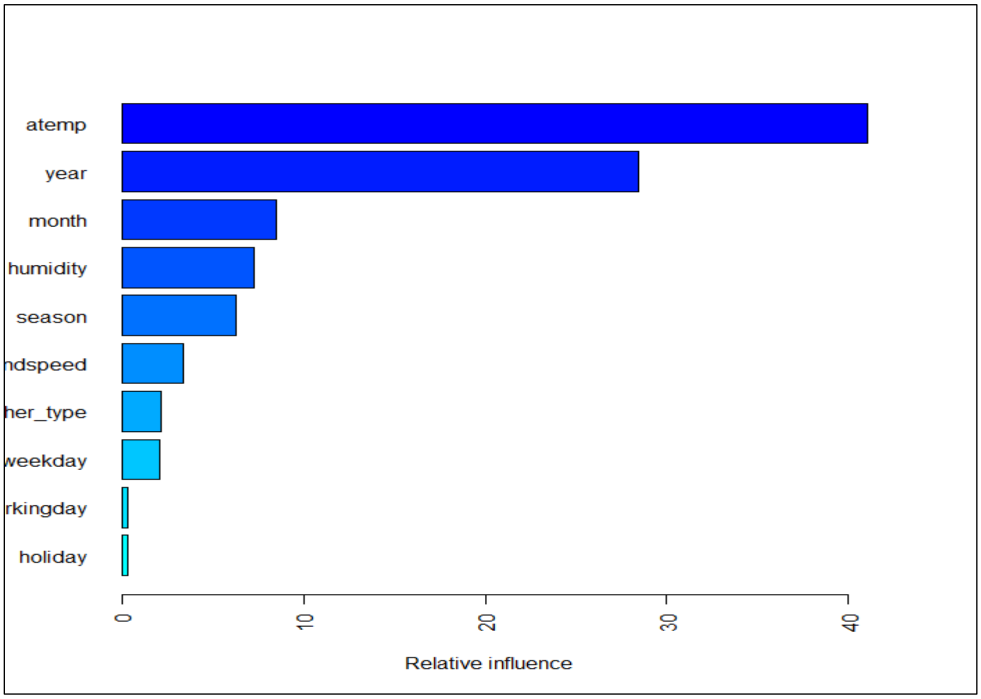


Table 4.3: gbm relative influence values

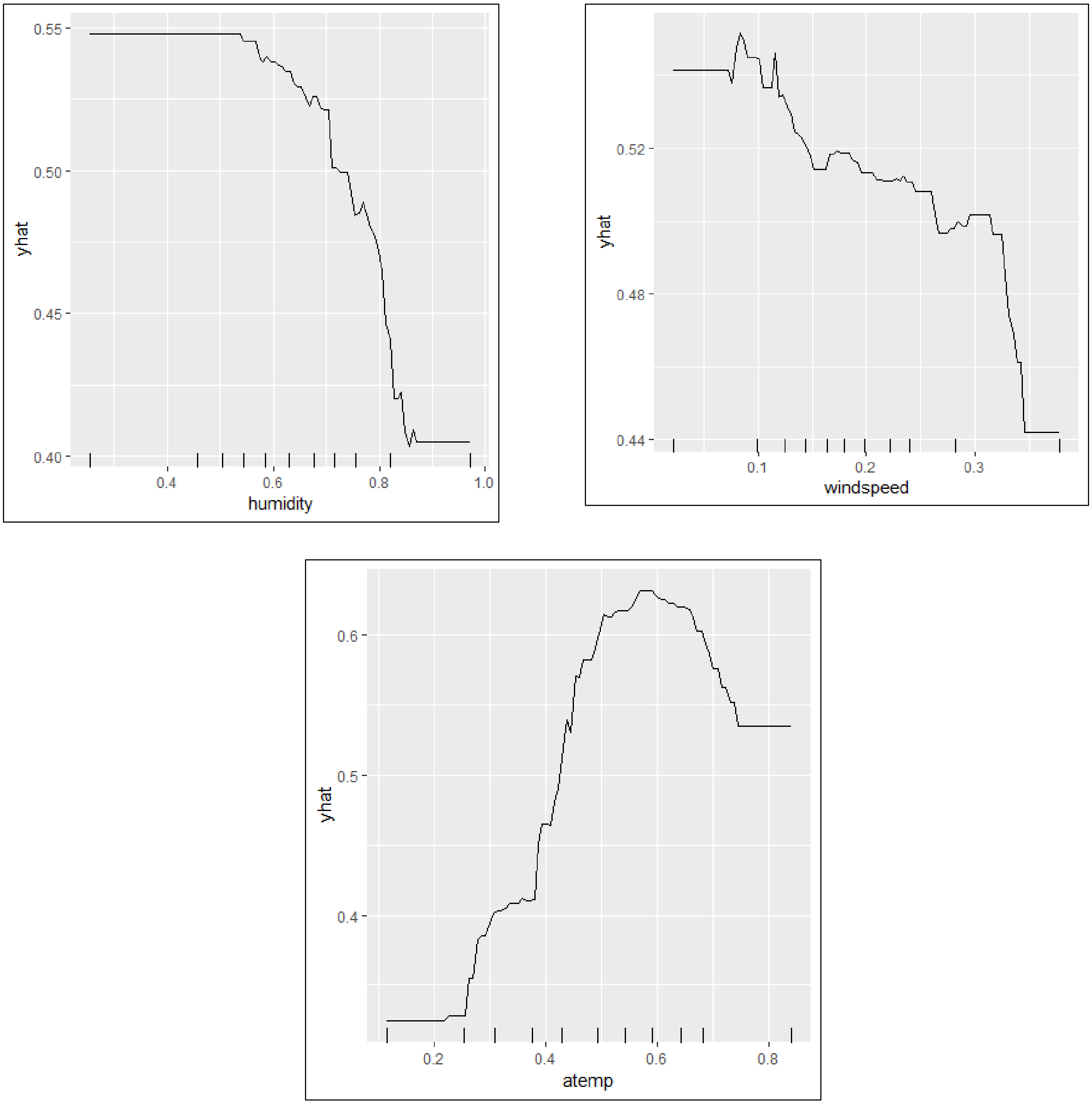
var	rel.inf
atemp	41.0175929
year	28.4525106
month	8.5374088
humidity	7.2989401
season	6.3020311
windspeed	3.4086015
weather_type	2.1482351
weekday	2.0976222
workingday	0.3692905
holiday	0.3677673

4.1.2. Partial dependence Plot

After the most relevant variables have been identified, the next step is to attempt to understand how the response variable changes based on these variables. For this we can use partial dependence plots (PDPs) and individual conditional expectation (ICE) curves.

PDPs plot the change in the average predicted value as specified feature(s) vary over their marginal distribution. For example, consider the atemp (Feeling temperature). The PDP plot below displays the average change in predicted total_count as we vary atemp while holding all other variables constant. This is done by holding all variables constant for each observation in our training data set but then apply the unique values of atemp for each observation. We then average the total_count across all the observations. This PDP illustrates how the predicted total_count increases as atemp increases.

Figure 4.3: partial dependency plot for humidity, windspeed and atemp

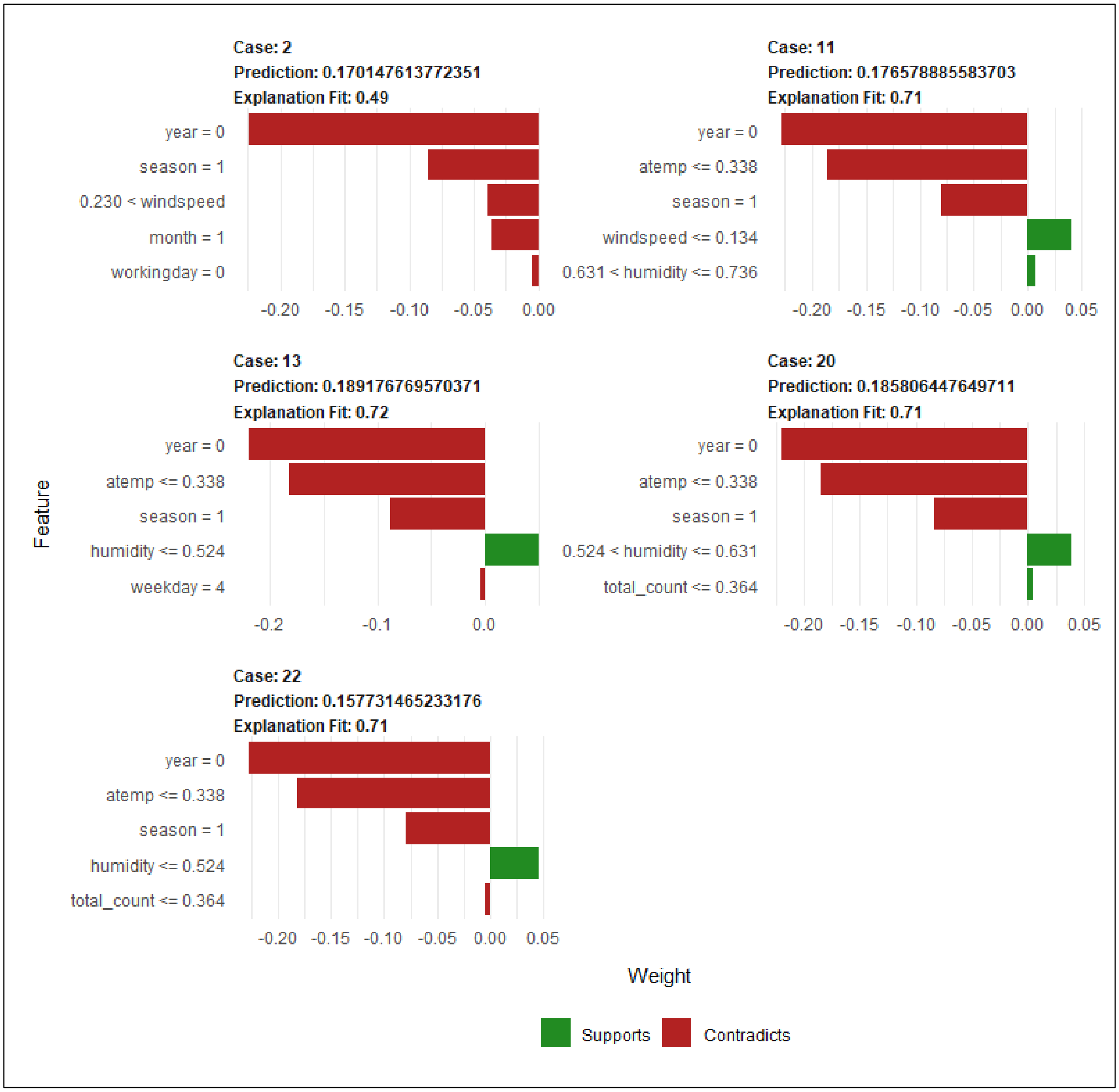


4.2. Support and Contradict features

LIME is a newer procedure for understanding why a prediction resulted in a given value for a single observation**.**

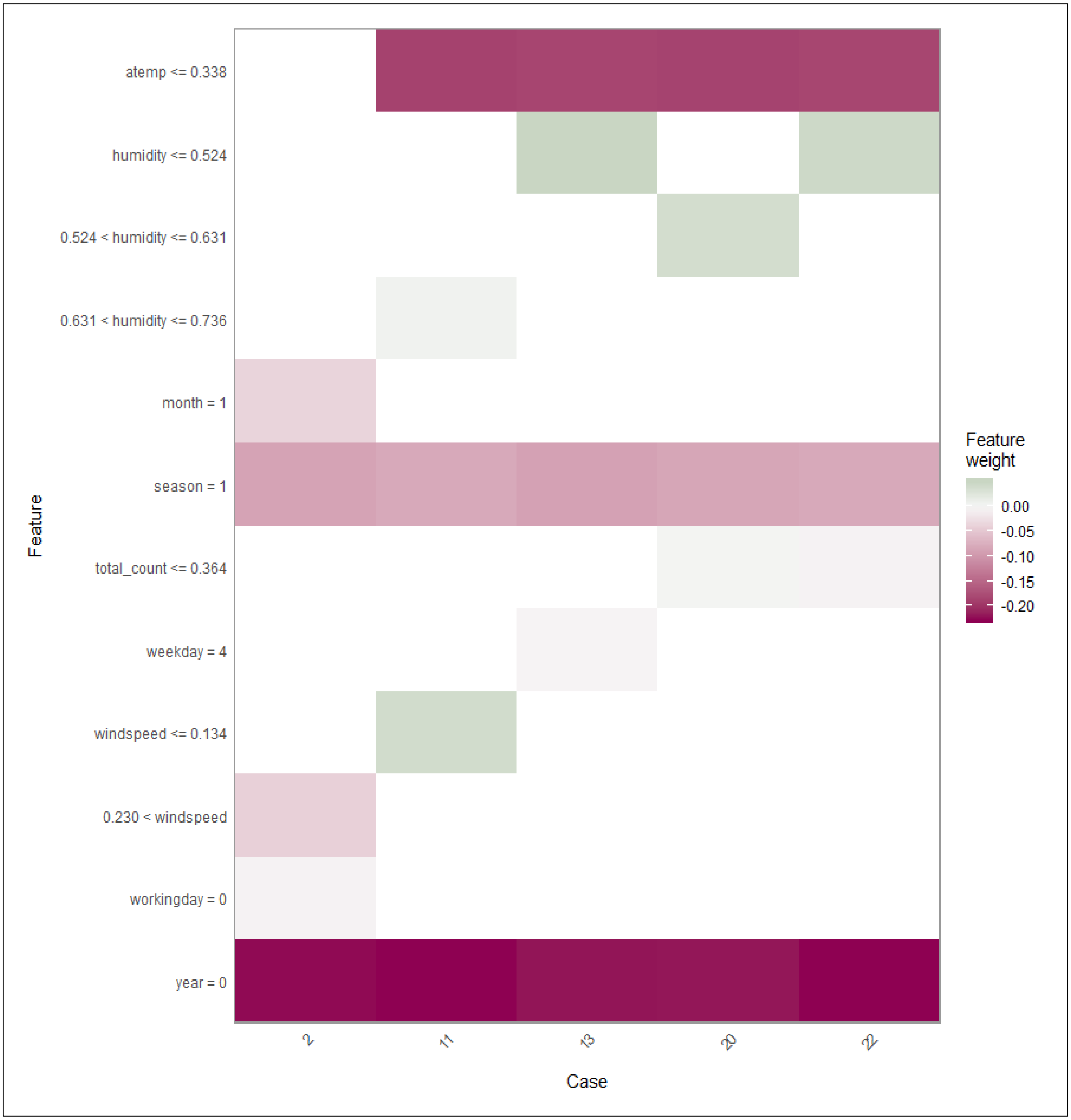
We can now apply to our two observations. The results show the predicted value (Case 1: 0.17014, Case 2: 0.17657), local model fit (both are relatively poor), and the most influential variables driving the predicted value for each observation.

Figure 4.4: Predictors support and contracts plot for first 5 observations



The other plot we can create is a heatmap showing how the different variables selected across all the observations influence each case. This plot becomes useful if you are trying to find common features that influence all observations or if you are performing this analysis across many observations which makes plot_features difficult to discern.

Figure 4.5 Heat map of different variables across first 5 observations



Chapter 5

5. Conclusion

5.1. Model Evaluation

We have made our gbm model with 10 fold cross validation. We are going to evaluate our model based on Rsquare on 10 fold cross validation fitted value and fitted model values with the observed value of the target variable , *Rsquare tells us how well the model is able to define the variance of the target variable based on the predictors*. We have got the RSquare value of **0.94** which is pretty good. We have got the RMSE value of **0.076** after tuning the model.

5.1.1. RSquare on CV fold and Fitted model

Below is the Rsquare value calculated on *gbm.fit$cv.fitted* and *gbm.fit$fit* with *train$total_count*

Cross Validation Fitted RSquare = **0.8771373**

Fitted Model RSquare = **0.9465938**

5.1.2. Root Mean Square Error

RMSE shows the mean square error. We have got the error of 0.07 which looks good.

```
➢ caret::RMSE(pred, test$total_count)
```

RMSE = **0.0766395**

5.1.3. Mean Absolute Error

MAE is one of the error measures used to calculate the predictive performance of the model.

```
➢ mean_absolute_error = caret::MAE(pred, test$total_count)
```

MAE = **0.05384554**

6. References

Blog article by Ando Sabaas on feature selection