Modelling Bike Rentals

Ricky Heinrich & Vimaljeet Singh

2023-03-24

Introduction

The dataset we have chosen to analyze is Hadi Fanaee-T's Bike Sharing Dataset, from the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, accessed in the UCI Machine Learning Repository. This dataset combines the Trip History Data for the years of 2011 and 2012 of 'Capital Bikeshare', which is metro Washington DC's bikeshare service, with weather data and the holiday schedule. We are hypothesizing that we are able to predict the count of bikes rented in a given hour given the predictor variables (weather, calendar date and time, etc). We are also hypothesizing that results will be better if we model casual and registered user bike rentals separately, rather than aggregated together in total bike users.

Description of Dataset

The data consists of an aggregated count of 'rides' by hour, over the span of the years 2011 and 2012. It contains 17379 rows and 17 columns.

Variable Name	Description	Type
instant	Record index	ordinal
dteday	Date	datetime
season	Season (winter, spring, summer, fall)	categorical
yr	Year (2011, 2012)	ordinal
mth	Month	categorical
hr	Hour	categorical
holiday	Whether day is a holiday or not	boolean
weekday	Day of the week	categorical
workingday	If day is neither weekend nor holiday	boolean
weathersit	Weather conditions	ordinal
temp	Temperature in Celsius	numerical
hum	Humidity	numerical
windspeed	Wind speed	numerical
casual	Count of bikes rented by casual users	numerical
registered	Count of bikes rented registered users	numerical
cnt	Count of total bikes rented	numerical

Plan overview

In this report, we will explore different models for predicting the rental bike count like - Linear Regression, Poisson Regression, and Random Forests. We chose these models due to their ability to handle the type of response variable we have and their relative simplicity in fiting, which allows for easy interpretation like in Linear model and Poisson model and more predictive power like in random forest. We will first fit these

models on the total rental count 'cnt', and then separately on the 'casual' and 'registered' rental counts. We will use variable selection methods to determine which predictors are most important. Our dataset will be split into a 60/40 training/testing set, and we will use mean squared error (MSE) and R^2 as measures of fit. Finally, we will present our conclusions and discuss the limitations of our models, as well as possible directions for future research.

We chose to remove 'atemp' variable because it had extremely high collinearity with 'temp', as well as the 'workingday' variable because it provides almost the same information as 'weekday'. We've also unscaled the temperature, humidity, and windspeed variables and stored those into 'rawtemp', 'rawhum', and 'rawwindspeed'. We are considering our full model to include the following predictors: yr, mth, hr, season, holiday, weekday, weathersit, rawtemp, rawhum, and rawwindspeed. The response variables are cnt, casual, and registered.

Linear Model (Parametric)

In the Bike Sharing dataset, the response variable 'cnt' represents the number of hourly users of a bike. Unlike qualitative or quantitative variables, this response takes on non-negative integer values of counts.

Total Count

Table 2 shows the statistics of the fitted linear model. The linear model has a R^2 of 0.6875 which means the model is able to explain 68.75% variation in the count data based on the given independent variables. The F-statistic is very high which means one or more of the coefficients is significant. Most of the variables that are in the model are significant. Overall, this seems to be a good fit for the model, but there seems to be an issue with the predicted values. 9.84% of the fitted values are negative which means that the linear model predicts a negative number of bikes rented during 9.84% of the hours in the data set (check 'Linear Model Fit' chart below). The negative expected values of bikers in certain situations raises doubts about the reliability of predictions made from the regression model. It also casts doubt on the accuracy of the coefficient estimates and confidence intervals of the model.

Moreover, it is plausible to assume that when the expected number of bikers is low, the variance associated with the number of users should also be small. For example, during a heavy December snow at 1 AM, we anticipate that only a few people will use a bike, and there will be less variation in the number of users during such conditions. In contrast, between 6am and 9am in summers, more number of riders are expected and hence the variance should be higher. Table 3 shows how these statistics vary.

Heteroscedasticity refers to a violation of the assumption in the linear model:

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon$$

where the variance of the response variable (cnt) is not constant across the range of predictor variables. The most common form of heteroscedasticity in the response variable is that the variance of the response variable may change as the mean of the response variable changes. Because of this, the estimate for the variance of the slope and variance will be inaccurate. Heteroscedasticity can be detected by examining the scatter plot

Table 2: Linear Model Statistics

Statistic	Value
R_Squared	0.6875
MSE	9951.4900
F-Stat	447.5300

Table 3: Table of Variance vs Mean (for select times)

Months	Time	Riders_Mean	Riders_Variance
December, January, February	1am - 4am	12.10479	284.7236
April, May, June	6am - 9am	238.01648	32062.5637

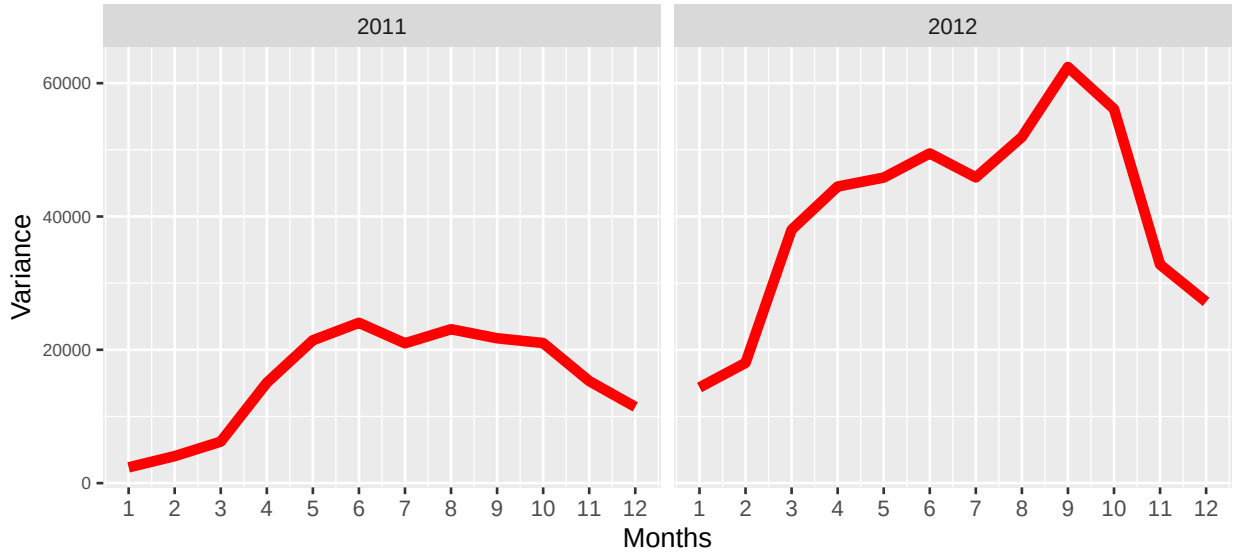

Figure 1: Variance of total count over 2011 and 2012

of the data before performing the regression. We will plot a graph of variance for the 'cnt' values for both the years to inspect this.

Figure 1 clearly shows that the variance varies throughout the year and the assumption of a linear relationship between the predictor variable and the response variable is severely violated due to unequal variance of the response variable. In fact, the variance in 2012 is visibly higher too. As a result, the assumption of homoscedasticity is not met, which raises concerns about the appropriateness of using a linear regression model to analyze the data.

Figure 2 shows the observed values vs predicted values using linear model. Our concern here is visualized when we see the red dots falling below zero on the x-axis.

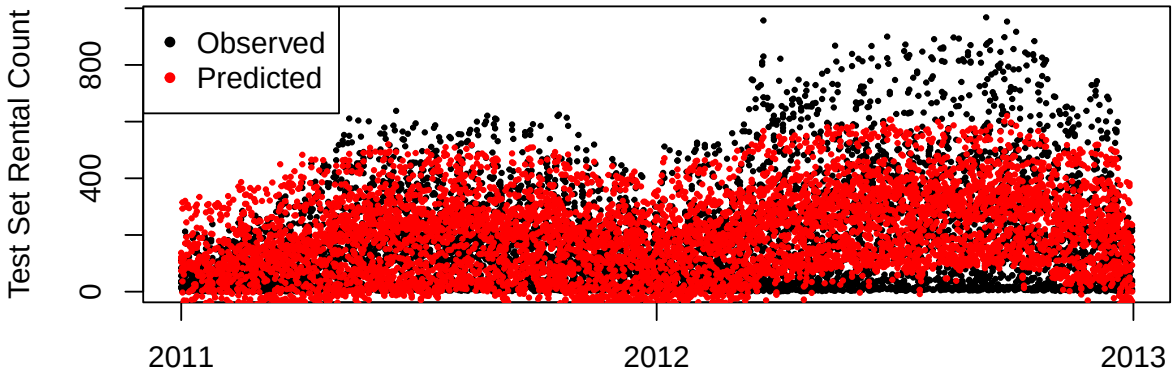

Figure 2: Linear model fit for total counts

In Figure 2, we see that some predicted values are below zero. We cannot have negative predicted values as the number of bikes rented in an hour can never be negative. This is another reason we should not prefer linear model for this data. Furthermore, the response in this dataset 'cnt' is in the form of integers, while a linear model assumes that the error term is continuous. This means that the response variable in a linear model must be continuous as well. Therefore, the integer nature of 'cnt' response implies that a linear regression model may not be entirely suitable for this dataset.

Transforming the response into `log` could help us eradicate some of the problems that we are facing with linear model. We could fit something like:

$$log(cnt) = \sum_{j=1}^{p} X_j \beta_j + \epsilon$$

Transforming the response variable in the Bike sharing data can be helpful in addressing two main issues associated with fitting a linear regression model: the occurrence of negative predictions and the presence of heteroscedasticity in the original data. By transforming the response, we can avoid negative predicted values and reduce heteroscedasticity, resulting in a more accurate and reliable model.

While transforming the response variable can address some issues in fitting a linear regression model, it is not entirely satisfactory. This is because the predictions and interpretations are made in terms of the logarithm of the response rather than the response itself, which can be challenging for interpretation. Moreover, this

Table 4: Fitted linear model statistics for Casual Count

Statistic	Value
R_Squared	0.5920
MSE	984.6553
F-Stat	295.2100

transformation cannot be applied to data set where the response can take on a value of zero. Therefore, although using a transformation of the response can be a reasonable approach for some count-valued data sets, it may not always be the optimal solution like in this case.

Casual Count

In Table 4, we see that the R^2 value for the model fitted on the response variable 'casual' indicates that this model accounts for 59.20% of the variation in the count of casual user rentals per hour, which is worse than the model for all users. Figure 3 shows that this model cannot predict the higher count values.

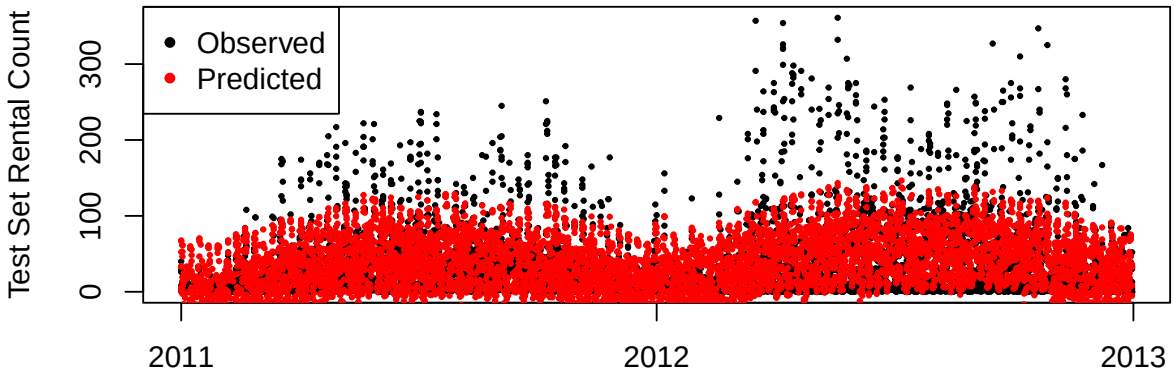

Figure 3: Linear model fit for casual count

Registered Count

In Table 5, we see that the R^2 value for the model fitted on the response variable 'registered' indicates that this model accounts for 68.48% of the variation in the count of casual user rentals per hour, which is only slightly worse than the model for all users (68.75%). Figure 4 is similar to Figure 3, where it shows that this model cannot predict the higher count values.

It seems for that in the case of the linear model, the success in predicting casual user and registered user ridership separately is ambiguous, as the separate models don't explain more of the variation in their response variables.

Table 5: Fitted linear model statistics for Registered Count

Statistic	Value
R_Squared	0.6848
MSE	7245.4982
F-Stat	442.0000

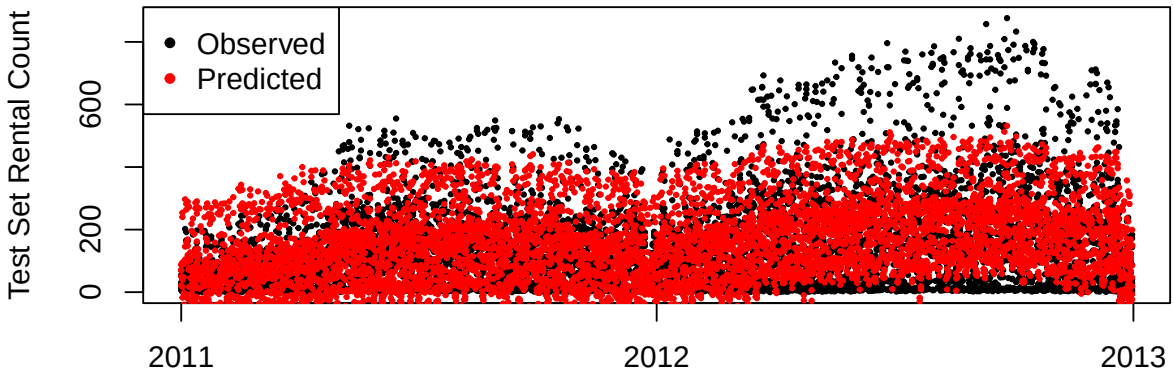

Figure 4: Linear model fit for registered counts

Poisson Model (Parametric)

The Poisson distribution is commonly employed to model counts due to several reasons, such as the fact that counts, like the Poisson distribution, are restricted to non-negative integer values. This makes it a suitable and natural choice for modeling count data. Poisson model can be fit as:

$$log(\lambda(X_1, ..., X_p)) = \beta_0 + \beta_1(X_1) + \text{ůůů} + \beta_p(X_p)$$

Here if the coefficient β_1 is equal to 0.5, then we can interpret it as follows: for a one-unit increase in X_1, the expected mean count will increase by a factor of $e^{0.5} = 1.65$, holding all other variables constant.

Total Count

All the variables in the model are significant. The residual deviance is way lower then the null deviance which shows that the independent variables help in explaining the response variable. Also, from the Figure 5, we can see that all the predicted values are above the x-axis which is one of the problems we were trying to overcome after using linear model. We can also see that the predicted values have a greater range than in the linear model, and may match up with the larger count values better.

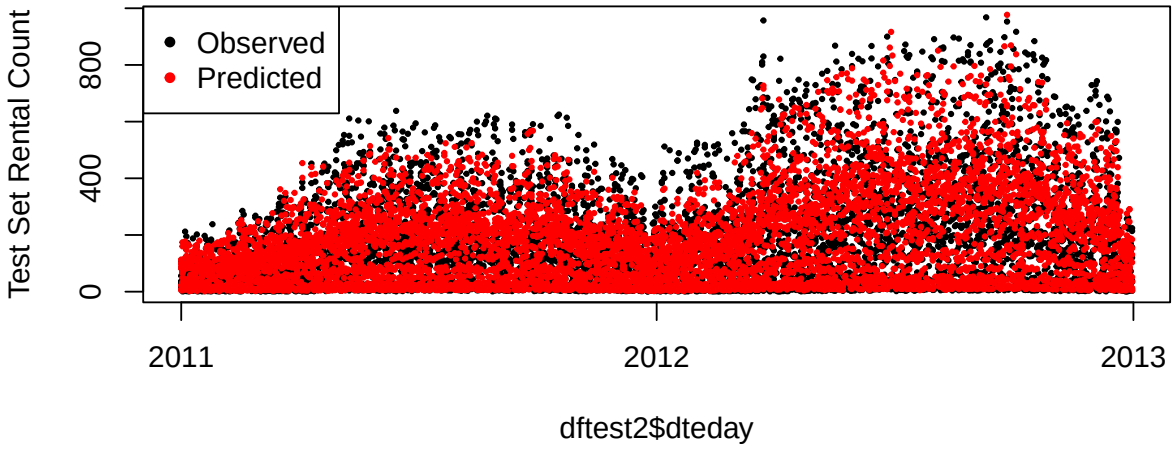

Figure 5: Actual vs predicted values of test set with GLM-Poisson model

When using a Poisson regression to model bike usage, we make an implicit assumption that the mean bike usage in an hour is equal to the variance of bike usage during that hour. In contrast, a linear regression model assumes that the variance of bike usage always takes on a constant value. Therefore, the Poisson regression model is better suited to handle the mean-variance relationship observed in the Bike sharing data compared to the linear regression model.

In fact from Table 6 we can see that the variance in 'cnt' appears to be much higher than the mean, a situation referred to as "overdispersion" which can seemingly be handled by quasi-poisson model. We checked the results from a quasi-poisson model as well and the result is exactly the same as that of a Poisson model and hence it was not included.

Table 6: Table: Mean vs Variance for Poisson

Year	Month	Mean	Var
2011	1	55.51	2363.97
2011	2	74.29	4048.27
2011	3	87.73	6221.97
2011	4	131.95	15056.42
2011	5	182.56	21431.03
2011	6	199.32	24050.94
2011	7	189.97	20977.03
2011	8	186.99	23089.84
2011	9	177.71	21731.45
2011	10	166.23	21014.77
2011	11	142.10	15313.78
2011	12	117.84	11436.88
2012	1	130.56	14351.25
2012	2	149.04	18032.86
2012	3	221.90	38013.55
2012	4	242.65	44495.40
2012	5	263.26	45834.14
2012	6	281.71	49466.60
2012	7	273.67	45863.84
2012	8	288.31	51927.01
2012	9	303.57	62430.32
2012	10	280.85	56122.56
2012	11	212.62	32788.45
2012	12	166.73	27190.42

Casual Count

We see in Figure 6 the improvement in the range of predicted values compared to the linear model.

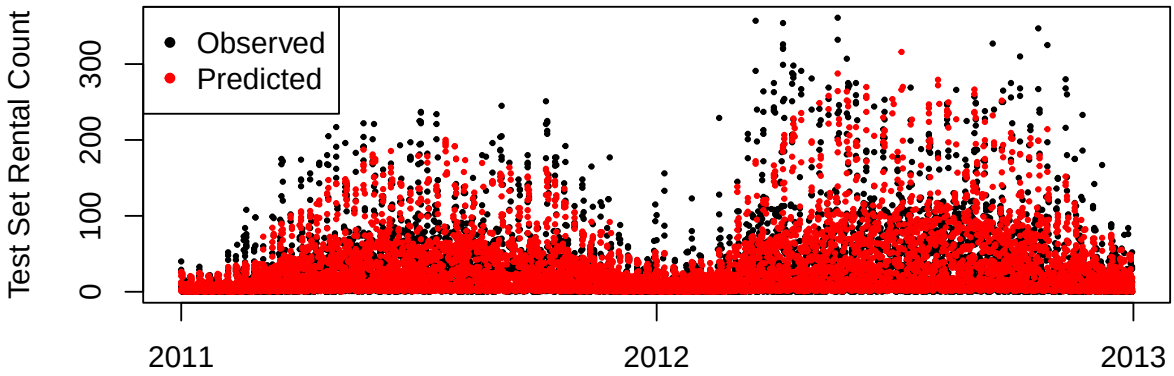

Figure 6: Linear model fit for casual count

Registered Count

Figure 7, however, shows that the predicted values are overshooting the observed values, already indicating a poor fit.

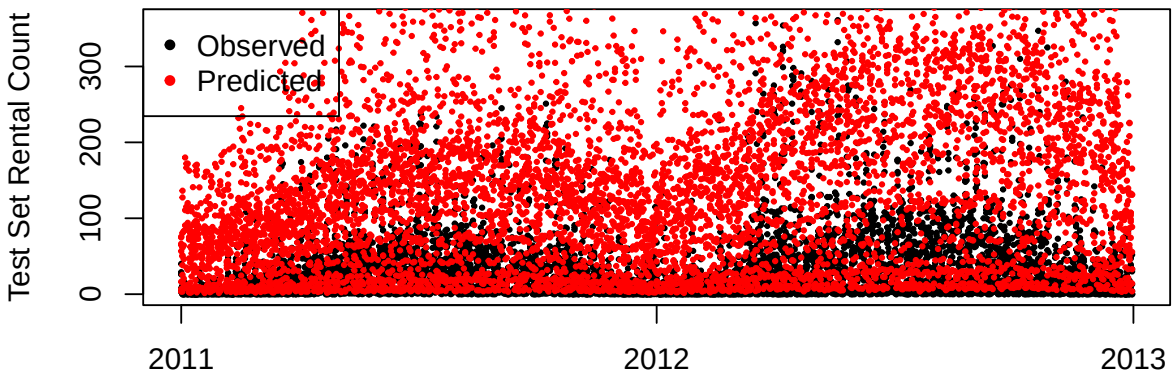

Figure 7: Linear model fit for casual count

Random Forest Model (Non-Parametric)

Because we are interested in prediction, we are ready to lose some interpretability in our model in exchange for better predictive power. Using random forest over a simple decision tree decreases the variance as well as the bias, usually resulting in closer predictions.

Total Count

As seen in Figure 8, the percentage increase in MSE shows that when either "hr" or "yr" are permuted, then the MSE increases by over 150%: the values of these variables are really important in prediction. "Weekday" and "temperature" variables are also important, as MSE can increase by 75% or 100% as well when they get jumbled up. All variables except for the "windspeed" seem to sit at %IncMSE values of over 25%, suggesting that they all carry some importance in the model.

The IncNodePurity shows that the "hr" variable is by far the most important variable in regards to decreasing the node impurity.

Casual Count

Modeling for just the Casual users, we see in Figure 9 below that the most important variable in terms of the mean decrease of accuracy in predictions on the out of bag samples when the variable is permuted is now weekday, at over 200%, followed by hr just below 150%, and then a year just under 100%. Still, there's a slew of variables that change over 50% in MSE when they are permutated.

In terms of IncNodePurity, the hr variable still ranks the highest, but the value is not as high as in the total cnt, and the weekday and rawtemp values follow closer behind than other models.

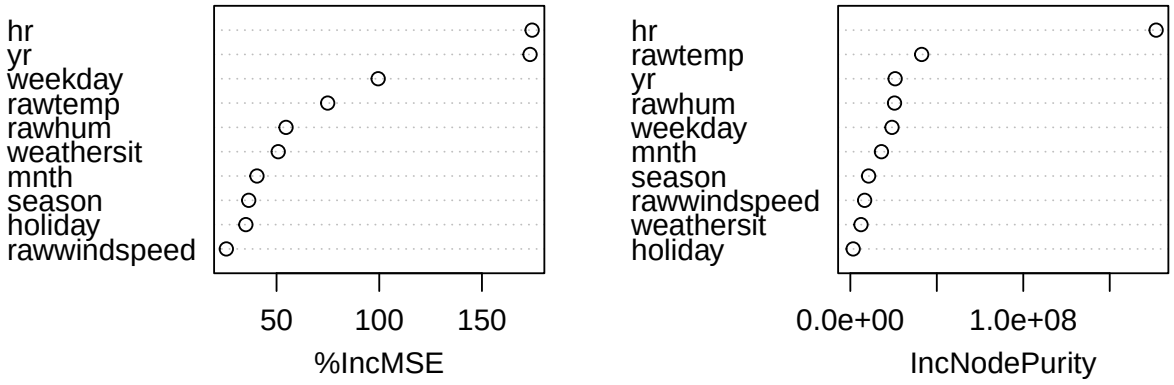

Figure 8: Variable importance for total count

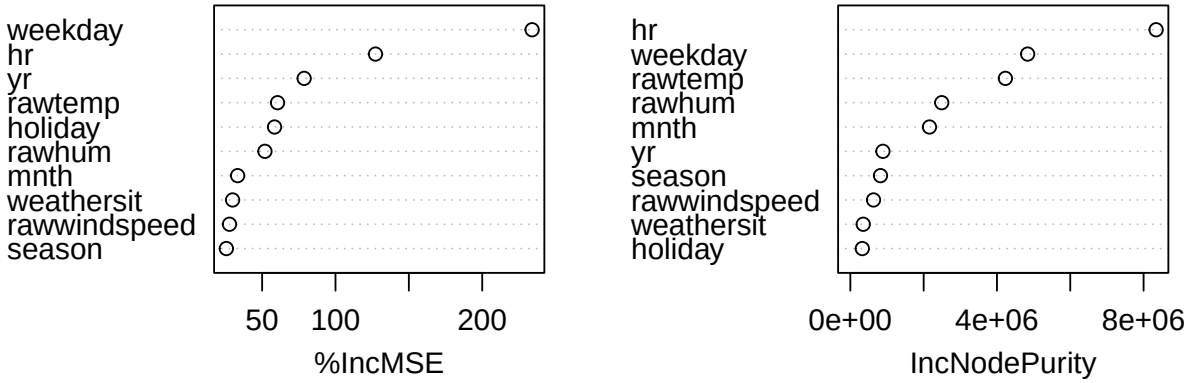

Figure 9: Variable importance for casual count

Registered Count

Modeling for just the registered users follows more closely to the total count, as we can see in Figure 10, which makes sense given that registered users take up a greater proportion of total rides than casual users. We see that hr, year, and weekday are again top variables when it comes to %IncMSE, and that hr is vastly more important when it comes to IncNodePurity.

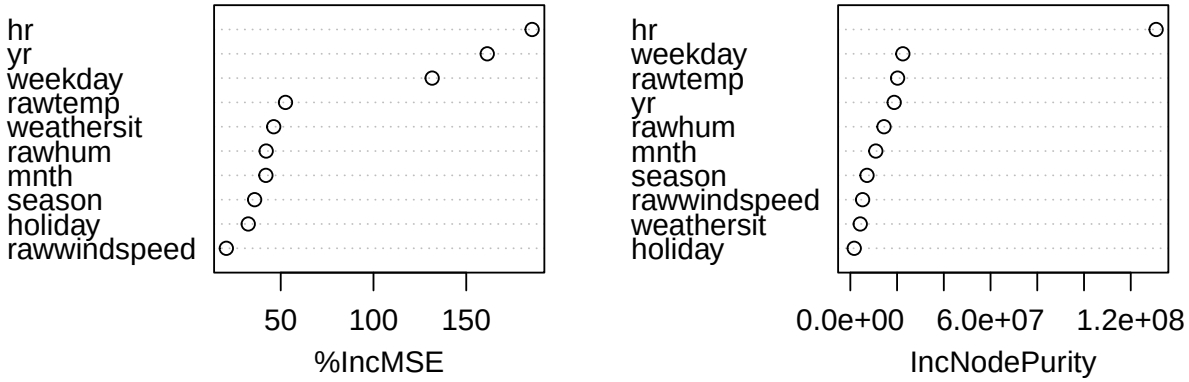

Figure 10: Variable importance for registered count

Perfomance of all models on test set

To compare the model performances on the test set, we have calculated the MSE for each model for all the three responses i.e. 'cnt', 'casual', 'registered' and the results are tabulated in Table 7. We observe that MSE of total counts in linear model is the highest followed by MSE of poisson model and then random forest. We should note that the MSE of random forest model is lowest for all three models, regardless of what response variable we are regressing it on. This tells us that random forest model has the best fit as far as the predictions are concerned using the fitted random forest model.

Table 7: MSEs for models

Model	MSE.Total.Count	MSE.Casual.Count	MSE.Registered.Count
Linear Model	9951.490	984.6553	7245.498
Poisson Model	8635.064	498.8537	4890.431
Random Forest	3361.434	247.0523	1983.619

Limitations

- The data set that was aggregated on the UCI Machine Learning Repository excluded all the data where the count of bikes rented was zero. This is a limitation of the dataset from a data collection standpoint. The dataset which has zeros for 'cnt' variable would not fit well with log transformation of the response and hence it was excluded in case kaggle updates the dataset to include zeros for the response variable.
- We see that 'yr' is an important variable all around: it makes sense, as Capital Bikeshare began operations in 2010, and so usage growth would grow a lot in the first few years. To predict count of bike rentals per hour in future years, we would have to take into account the growth in the business and popularity. Therefore, a limitation in our data is that is only covers two years and so it is hard to forecast usage in future years.
- Random Forests models provide excellent prediction capability but they can be difficult to interpret, especially if there are a large number of trees. Even though the MSE is lowest for random forest, the model is hard to interpret in the context of how increase in a unit of a specific independent variable will affect the response. Sometimes, overfitting could also be a problem with this model. Also, in some cases Random forest models tend to perform better when there are categorical variables in the data. This can be a problem if the data is mostly continuous.
- The Linear Model (LM) is a good fit but the negative predictions for the count of rented bikes make us question the reliability of the model. This model would be better for interpretability as compared to random forest. Furthermore, the data violates the normality assumptions of the model which limits its capacity to provide meaningful results.

Conclusion

We observed that the Poisson model is a better fit than linear model both in terms of having lower MSE and having predicted values that are positive. Even though Poisson model is less easily interpretable than linear model, minimal effort is required to interpret how the coefficients of Poisson model relate to/explain the response variables as was shown before. Furthermore, the Random Forests provide much less interpretability, but makes up for it via greater accuracy in predictions, as can be seen from its relative MSE.

If we want to understand how the different variables contribute to explaining the response variables, the Poisson model should be picked. If we'd rather get good predictions, random forest model should be picked. We are expecting a non-parametric model with splines to fit even better, but were not able to run such a model with our limited computer power.

We observe in Table 7 that the MSE for 'casual' variable is significantly lower than those of 'registered' and 'cnt'. In fact, in all models, the MSE of the total count models is greater than the sum of MSEs for the casual and registered mode. This means that the separated models are better able to capture and accurately predict the number of bike renters, as was hypothesized.

In the linear models and poisson models, all predictors came out as significant, and so variable importance was hard to discern. The random forest model gave us the insight that the predictors most important in predicting the counts of riders, whether casual or registered, were the hour of the day, the year, and the weekday. This suggests that time affects the ridership more than the weather conditions.