MTH 552-A
Project Report

Topic: Bike Sharing Model

Submitted by:
Ayush Sekhari
Aayush Mudgal
Samadipa Saha
Sheallika Singh
Vibhuti Mahajan

Data source:

The dataset used is an hourly bike rental data spanning two years from the Capital Bikeshare program in Washington, D.C. provided by Hadi Fanaee Tork and hosted on UCI machine learning repository.
The training set comprises of the first 19 days of each month, while the task is to predict the count of rentals for the rest of the month.

Data Fields:

datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday - whether the day is considered a holiday
workingday - whether the day is a weekend or not
weather -
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 - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated
count - number of total rentals

registered	season	hrs	workingday	wk	holiday
Min. : 0.0	1:2686	12 : 456	0:3474	Friday :1529	0:10575
1st Qu.: 36.0	2:2733	13 : 456	1:7412	Monday :1551	1: 311
Median :118.0	3:2733	14 : 456	-	Saturday :1584	-
Mean :155.6	4:2734	15 : 456	-	Sunday :1579	-
3rd Qu.:222.0	-	16 : 456	-	Thursday :1553	-
Max. :886.0	-	17 : 456	-	Tuesday :1539	-
-	-	(Other):8150	-	Wednesday:1551	-

Figure Crude Data Statistics

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. These systems, enable interested people to rent a bike from one location and return it to a different place on an as-needed basis. There are about 500 bike-sharing programs around the world.
The data from such systems serve as a sensor network, and can be used for studying mobility in the city. The problem that we looked is from an ongoing Kaggle challenge titled, "Forecast use of a city bike-share system",

wherein we are supposed to forecast bike rental demand in the Capital Bikeshare program in Washington, DC, given some historical usage patterns with weather data.

Data Pre-Processing:

A first glance at the dataset suggests that the date-time field can not be efficiently used in the original form. The date-time field is split into respective hour, weekday, month and year fields, as they might be useful in further analysis. The provided data has regressors in the data, namely season, working day weekday and hour of the day which are categorical variables

Observing different pattern of bike renting by casual users and registered users :

Relations and dependence of the total bike rental counts initiated versus other factors is visualised first individually, to gain an insight into the provided data. Since the total bike rental counts is a sum of the casual bike rentals and registered bike rentals, initiated, we study the dependence of the various variables with respect to these individually. It is expected that the behavior of registered and casual bikers vary.

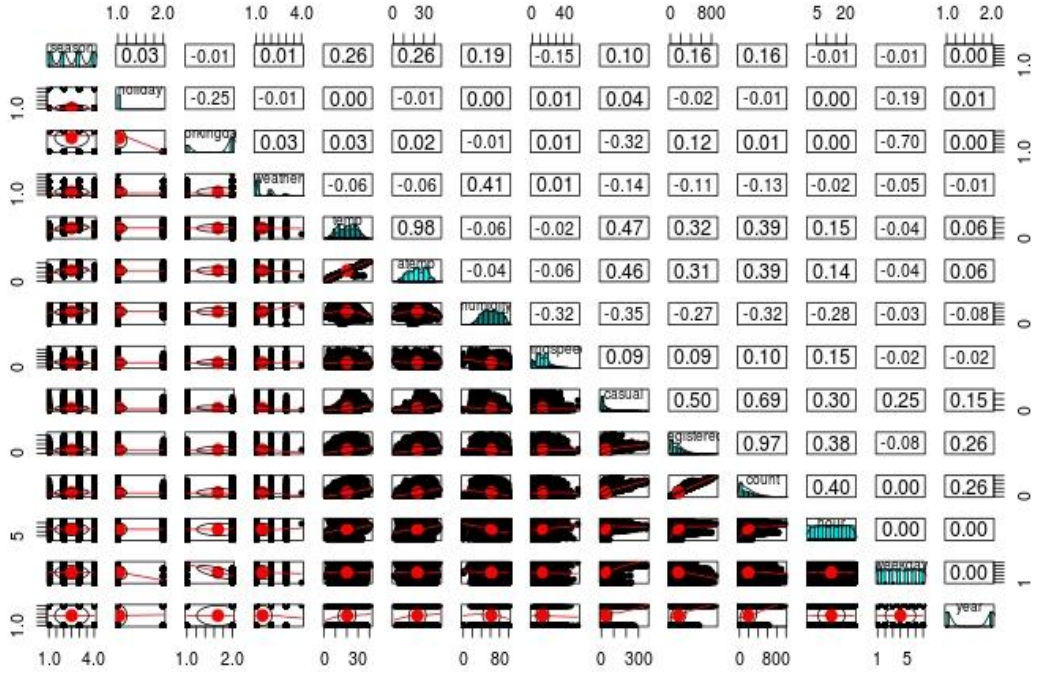
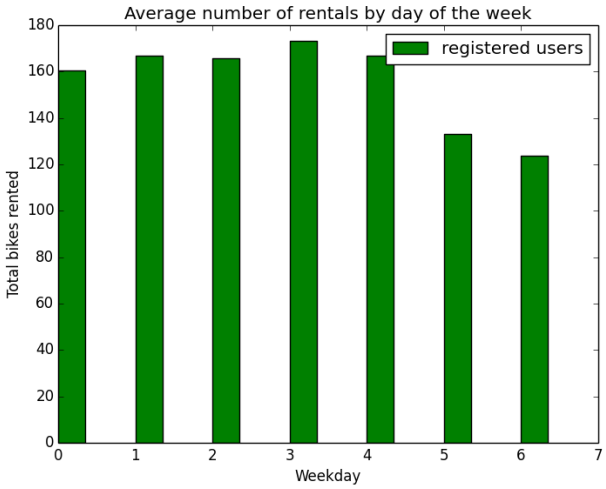


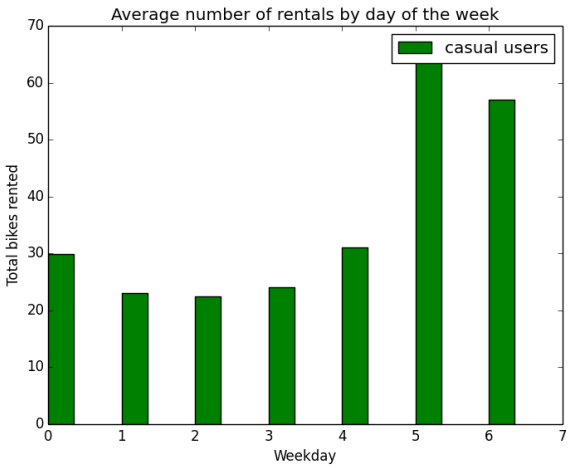
Figure: Scatter Plot
Bivariate Scatter Plots, below the diagonal and histograms on the diagonal

The scatter plot matrix, is drawn to capture any unusual behaviour or any potential correlations. It suggests that the weather, may have some correlation with humidity. And also that season, windspeed and temperature might be related to each other.

The above scatter plot matrix was used to try and observe any unusual behaviour, as well as potential correlations between variables in the dataset.

Days of the week:

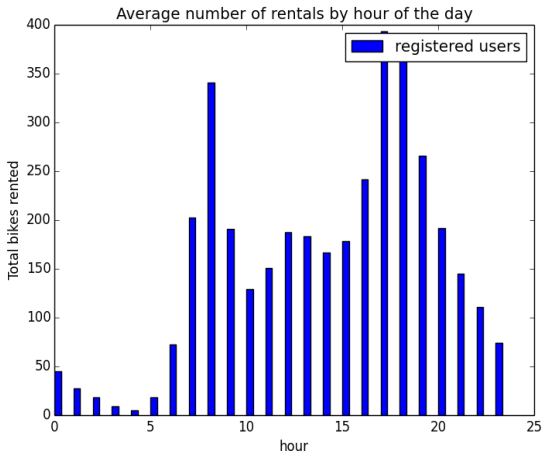


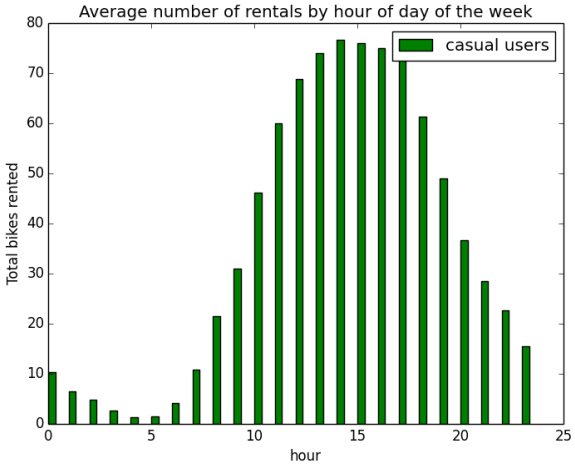


There is an observed drop of bike rented on weekends for registered users, but there is an increase of bike renting demand among casual users on weekends. And also casual users rent lesser bikes for weekdays. Such type of trend is logical and also supported by the dataset.

Hourly observation:

There are two peaks for bike rented by registered users at around 8am and 5-6pm and both are in working hours whereas for casual users is single peaked showcasing a “day-tripped” behaviour.

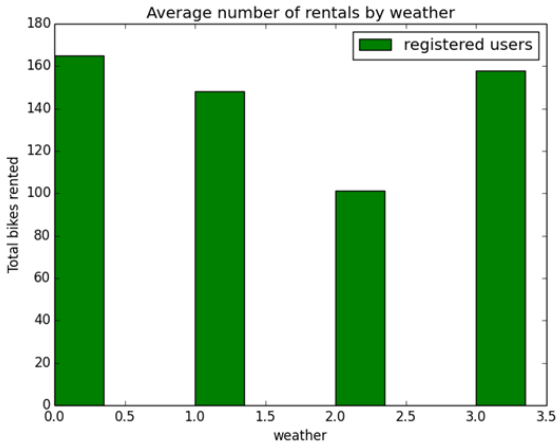


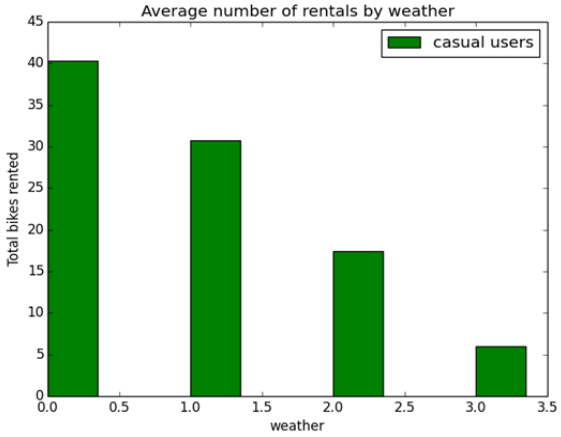


The registered users showed peaks for both a clear sky and worst of all weather condition. Casual users depicted a decreasing trend as the weather conditions deteriorated. But both casual and registered rentals increased with rise in temperature.

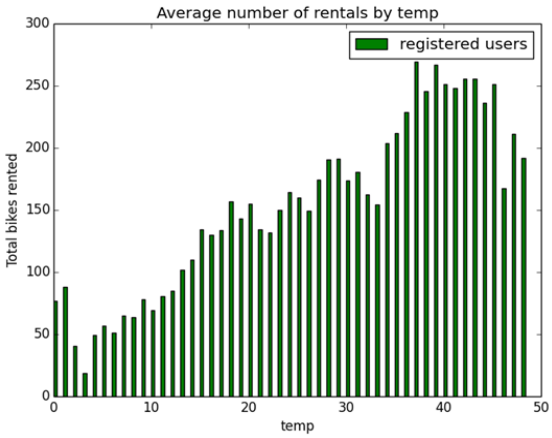
Weather Dependence:

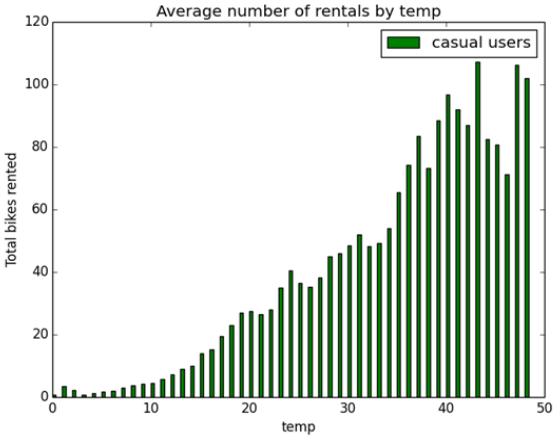
Registered rentals were more in case of both best and worst weather defined, but casual rentals decreased as weather deteriorated.

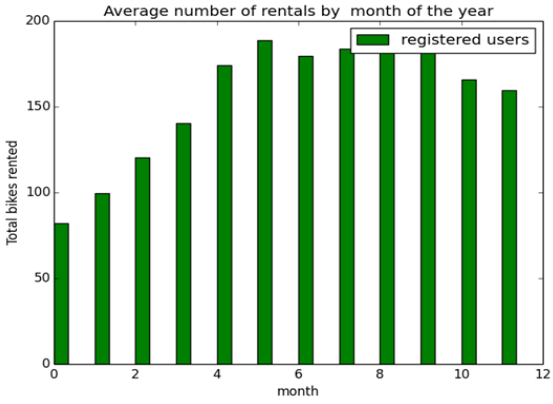


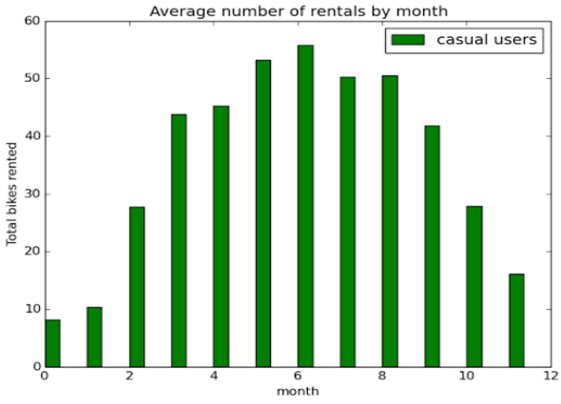


Rentals increased with temperature for both registered and casual.





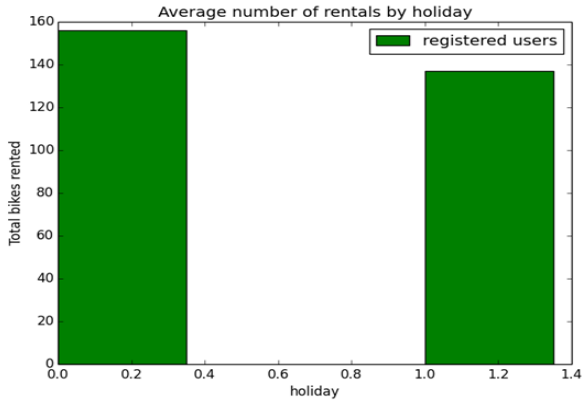


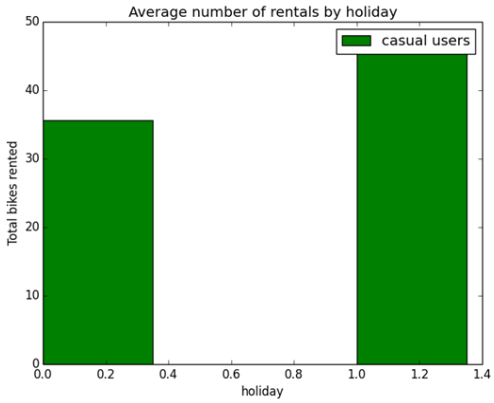


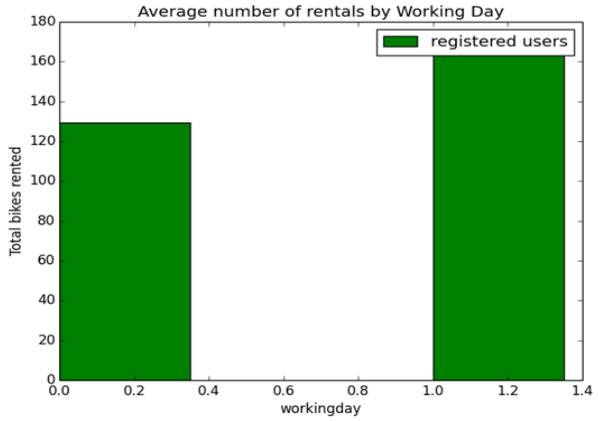
Hot months, thus definitely showed a high bike renting percentage from both casual and registered users

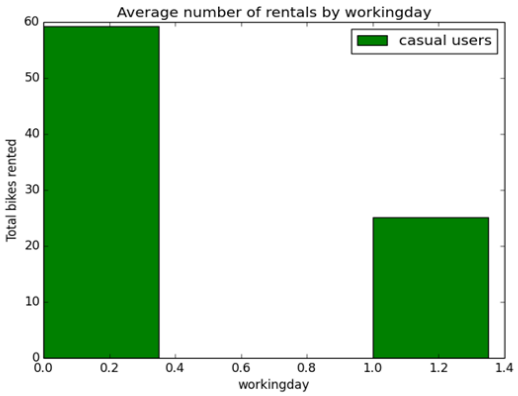
Working day/ Holiday Patterns:

Casual bike rentals were more in case of a holiday as compared to registered ones, but the working day rentals showed an opposite picture









Model Fitting

Different type of machine learning algorithms (as deemed adequate and list below) were tried. It is expected that there is high correlation between set of features and mild correlation among others. For each of the model tried, the technique of principal component analysis prior to the model training.

1. Linear Regression
2. Logarithmic Linear Regression
3. Ridge Regression
4. Random Forest
5. Logarithmic Random Forest
6. Gradient Boosting method
7. Exponential Gradient Boosting

The task is to predict the total number of bike rentals, which is an aggregation of bike rentals by casual users and bike rentals by registered users. Each model fitting is tried using both the approaches, i.e. predicting the total count and also by predicting the casual users count along with registered users count.

Model 1: Multiple Linear Regression Model

Categorical variables, namely hour, weekday, month, season and weather are replaced by dummy variables (one-hot encoded features).
A multiple linear regression model is trained on the data to predict the total users, casual and registered users count.

The **Summary** of the fits are as follows:

linear model = count ~ season.f + holiday + workingday + weather.f + temp + atemp + humidity + windspeed + hour.f + month.f + weekday.f

linear model= casual ~ season.f + holiday + workingday + weather.f + temp + atemp + humidity + windspeed + hour.f + month.f + weekday.f

linear model= registered ~ season.f + holiday + workingday + weather.f + temp + atemp + humidity + windspeed + hour.f + month.f + weekday.f

Though the standard error for count model is higher than standard error sum of both registered model and causal model error but model on count has highest adjusted **R-squared =0.6363** which shows that this model will be the better for prediction rather than separately for both casual users and registered users .

Plots:
residual v/s fitted response plot for total count as response variable

Normal Q-Q plot for linear model having count as response:

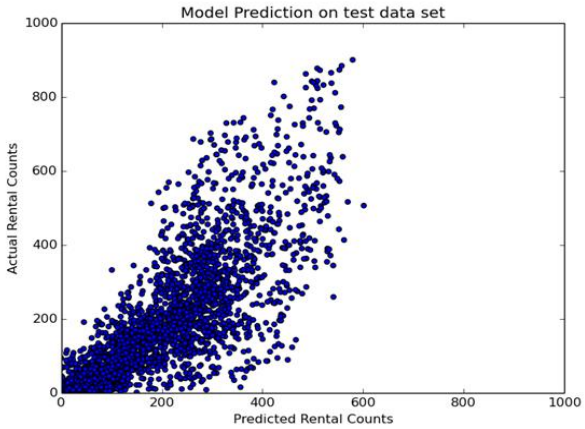
Opening funnel type of residual v/s fitted response where count is response variable shows that there is heteroscedasticity in the model.
Response variable is the increasing function of regressors variables.

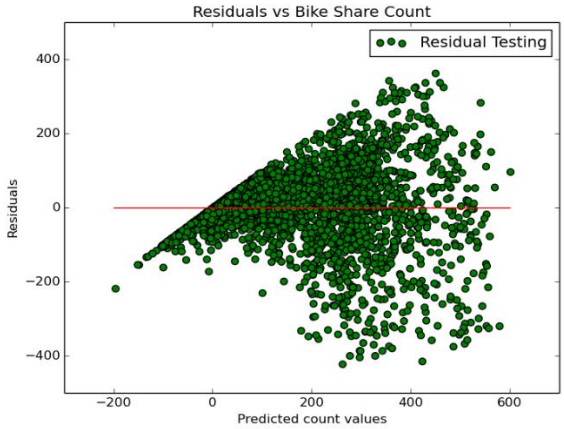
Plot of actual observation for count variable v/s predicted values for count variable from MLR model:

Summary Statistics:
mean square error (MSE)whole model: 11931.07
MSE(casual):1068.64
MSE(registered)=RMSLE(reg)=8543.23
Root mean square Lograthmic error(RMSLE):9533.93

Training accuracy(count): 64.02%

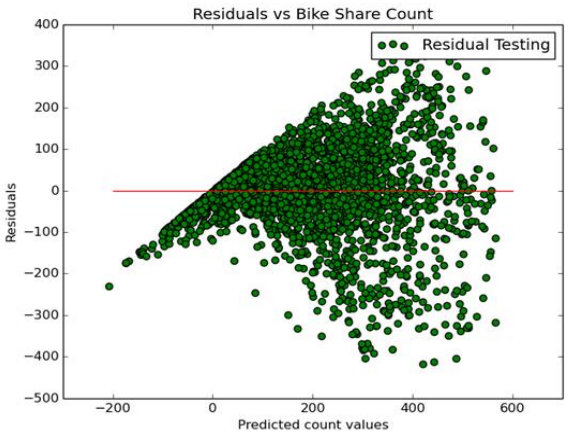
Test Accuracy: 62.89%
Training accuracy(casual): 58.78%
Test Accuracy: 56.58%
Training accuracy(reg): 63.81%
Test Accuracy: 60.53%

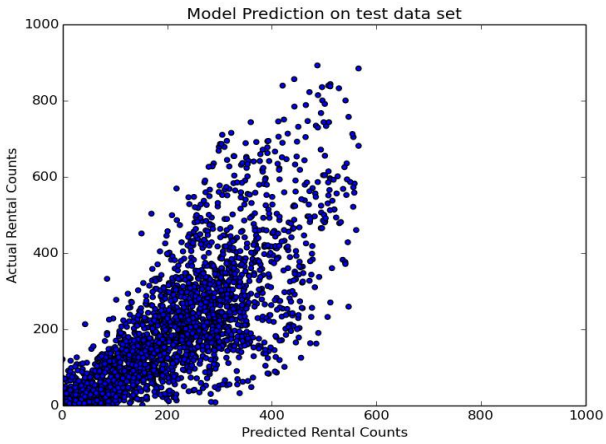




Linear Regression with PCA (to see if model fits better):

Summary Statistics:
MSE(count)=12261.31
MSE(casual)=1025.301
MSE(reg)=8312.73
RMSLE=9362





Training accuracy(count)=64.35%
Test accuracy(count)=61.88%
Training accuracy(casual)=58.04%
Test accuracy(casual)=58.88%
Training accuracy(reg)=63.35%
Test accuracy(reg)=62.17%
A higher error with all the components made no sense to further reduce the components and continue with the PCA reduced model.

Model 2: Logarithmic Linear Regression Model :

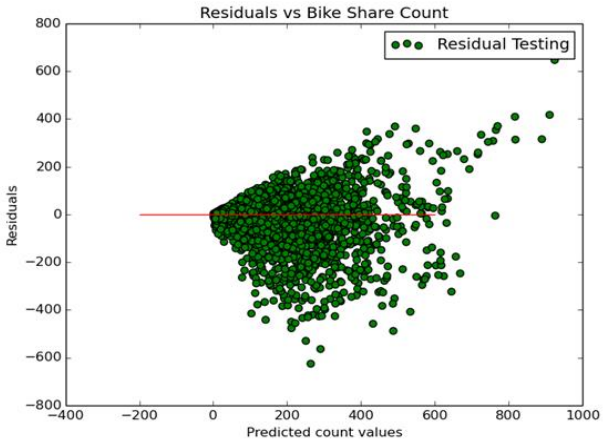
From the above analysis, we realise that modelling Y as a linear function of the regression coefficients may not be sensible and a more realistic model might be to use the natural log of the response variable, ln(Y), as a linear function of the coefficients instead. This will ensure that our response variable can only be defined in the interval [0, +inf).

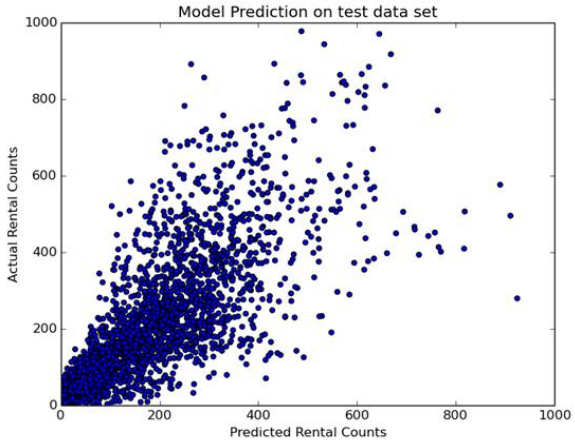
So, now ***linear model*** =log_count ~ season.f + holiday + workingday + weather.f + temp + atemp + humidity + windspeed + hour.f + month.f + weekday.f
Q-Q plot:
Adjusted R-square(0.7978) is high for this transformed model, Standard error is quite less for this model.
Residuals v/s fit plot shows that residuals are random and finally the predicted v/s actual plot of ln(count) variable has points along line with slope 1.

All these observations show that transformed model is better for prediction of count variable than MLR model and thus will have higher accuracy of prediction.

Summary Statistics:

RMSLE(count)=0.6211
RMSLE(casual)=2.055
RMSLE(reg)=0.685
RMSLE(whole model)=0.502



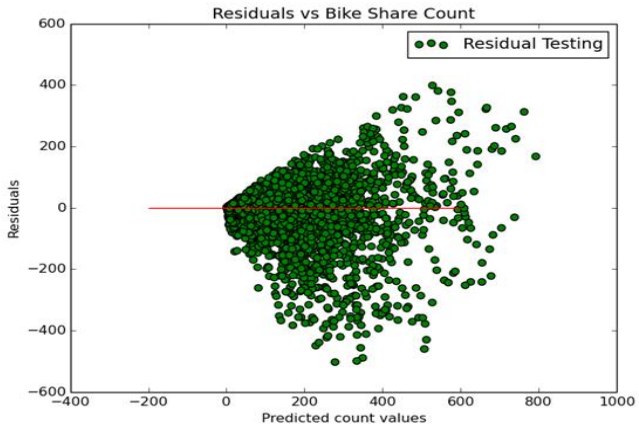


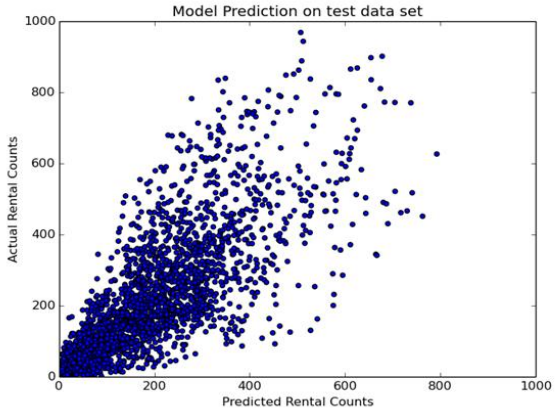
Training accuracy(count)=80.51%
Test accuracy(count)=80.86%
Training accuracy(casual)=82.26%
Test accuracy(casual)=83.06%
Training accuracy(reg)=79.56%
Test accuracy(reg)=77.58%

Customarily we apply PCA in hope of better model but the linear model fit does not show any improvements:

Summary Statistics:

RMSLE(count)=0.6407
RMSLE(casual)=2.073
RMSLE(reg)=0.670
RMSLE(whole model)=0.5005





Training accuracy(count)=80.51%
Test accuracy(count)=80.86%
Training accuracy(casual)=82.26%
Test accuracy(casual)=83.06%
Training accuracy(reg)=79.56%
Test accuracy(reg)=77.58%

Model 3: Ridge Regression Model

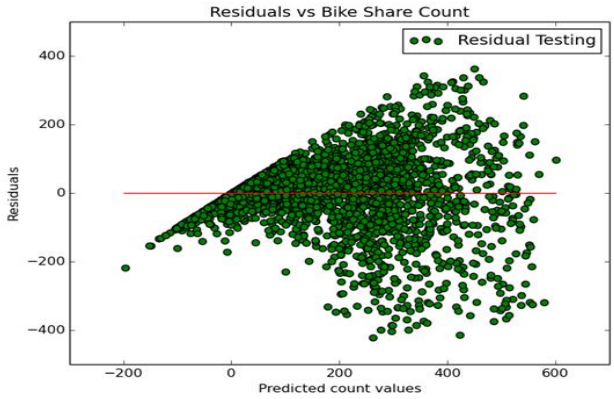
To account for the issue of over-fitting in the model and instability of the least square estimates, ridge regression model was incorporated that although gave biased estimates of the regression parameters but the variation was considerably reduced.

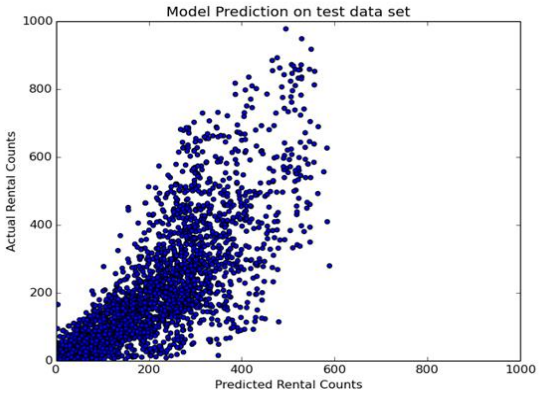
Part A: Ridge Regression was applied to the whole model.

Summary Statistics:

Mean square error for count: 12960.97 (upto 2 decimal places)
Mean square error for registered users: 1045.19
Mean square error for casual users:8215.84

Root Mean squares lograthmic error (RMSLE) (for full model): 9231.91





Training accuracy whole model: 64.0%
Test accuracy: 62.85%
Training accuracy casual: 58.31%
Test Accuracy: 58.02%
Training accuracy registered: 62.77%
Test accuracy registered: 63.89%

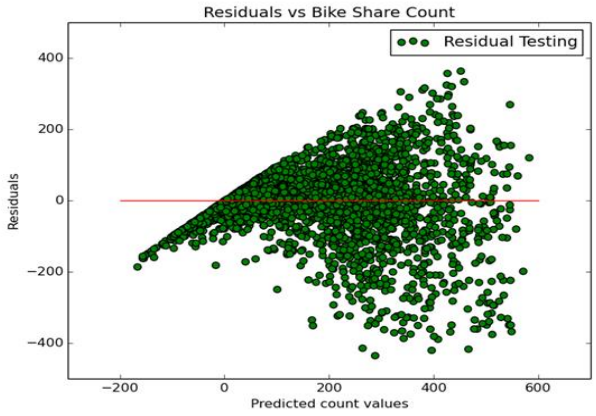
Part B: PCA was applied to the same model to test for a better fit.

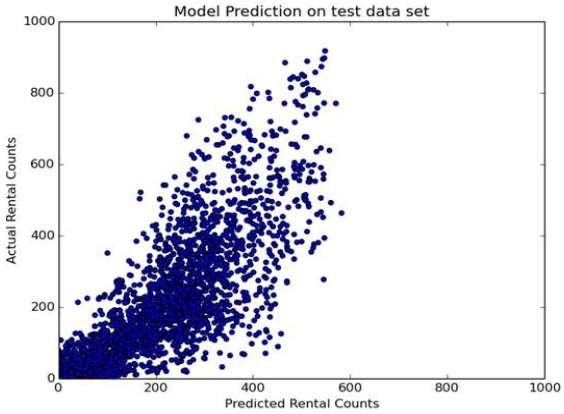
The RMSE value was high even considering all the components.

Summary Statistics:

Mean Squared error(whole model): 11308.91
MSE(casual): 939.76
MSE(reg): 8095.20
RMSLE(whole model): 9069.51

Training accuracy(count): 62.91%
Test accuracy(count): 66.12%
Training accuracy(casual): 58.44%
Test accuracy(count): 57.43%
Training accuracy(reg): 62.62%
Test accuracy(reg): 64.19%

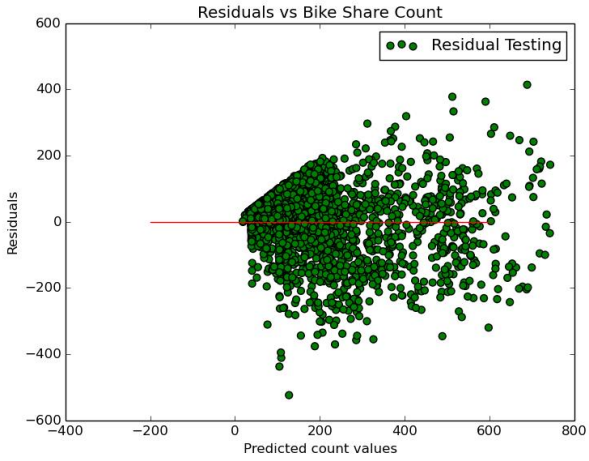


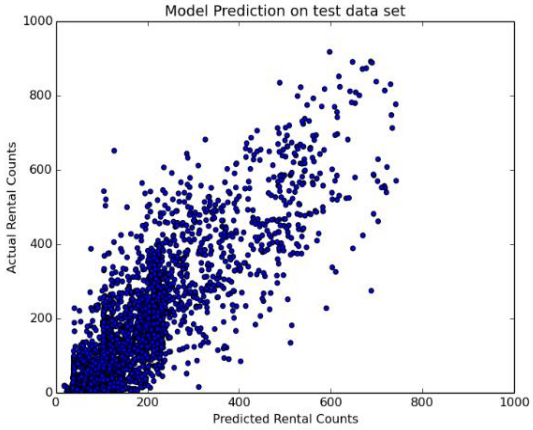


No further component analysis was done as PCA gave similar results for linear model

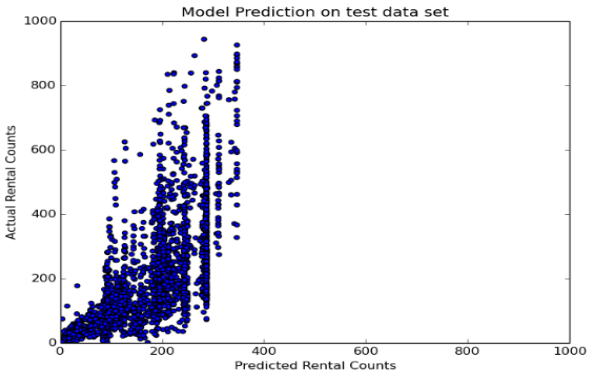
Model 4: Random Forests:

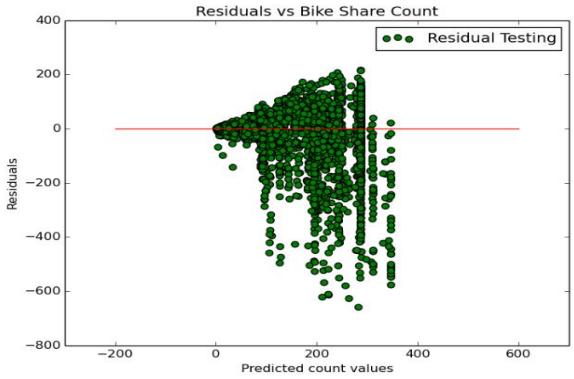
Random forests are an ensemble learning technique for classification and regression. It uses a number of decision trees at training time and outputs the class, which either can be chosen to be the mode of classes (classification problem) or mean prediction (regression of individual trees)





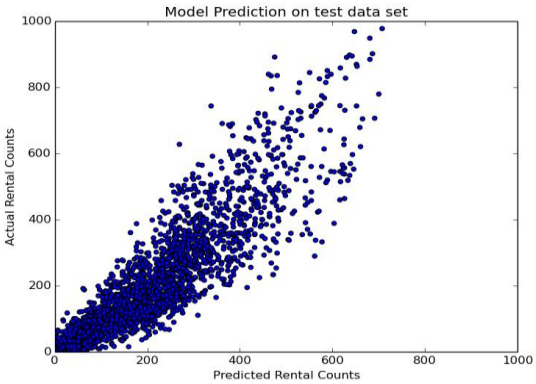
Plot: Residual and prediction plot for Random Forest Model



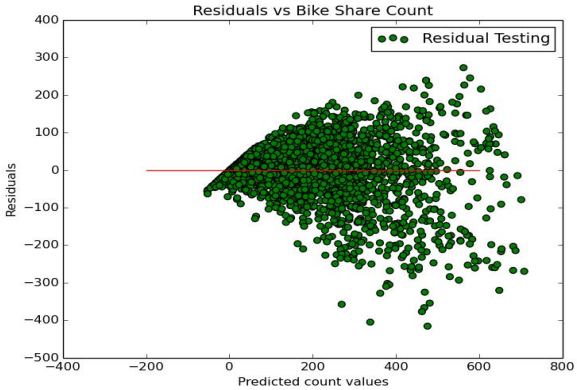


Plot: Residual and prediction plots for logarithmic random forest

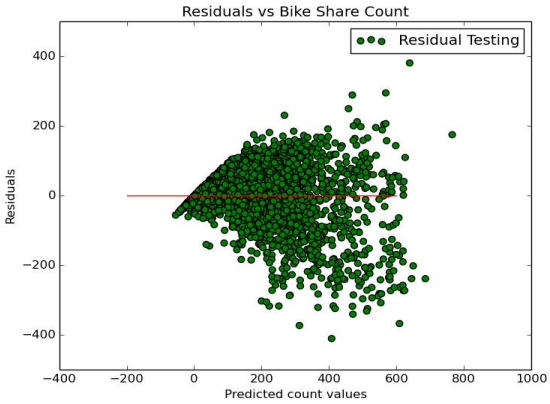
Model 5: Gradient Boosting:



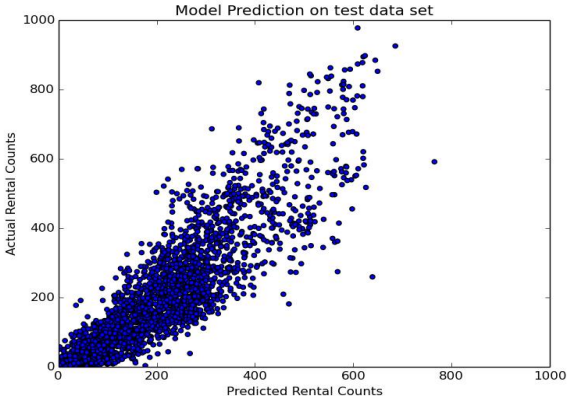
Prediction plot gradient boosting + PCA



Residual Plot of gradient boosting + PCA

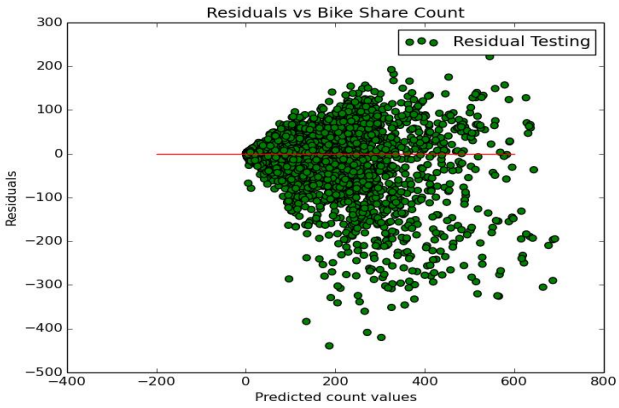


Residual Plot for gradient boosting

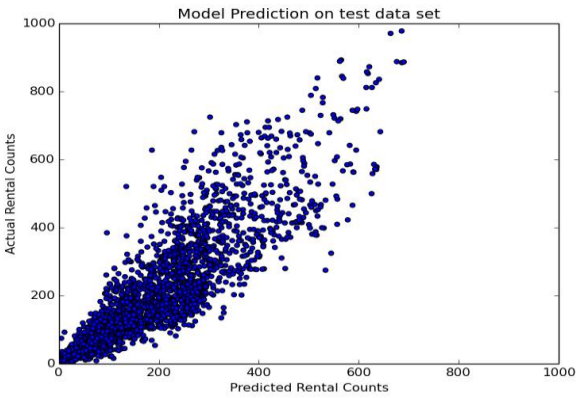


Prediction Plot for gradient boosting

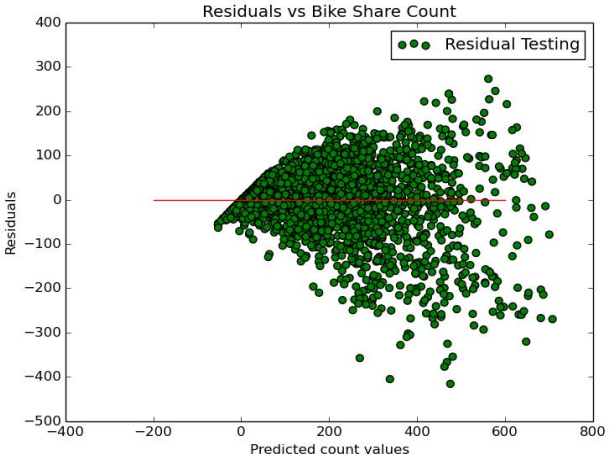
Exponential Gradient Boosting



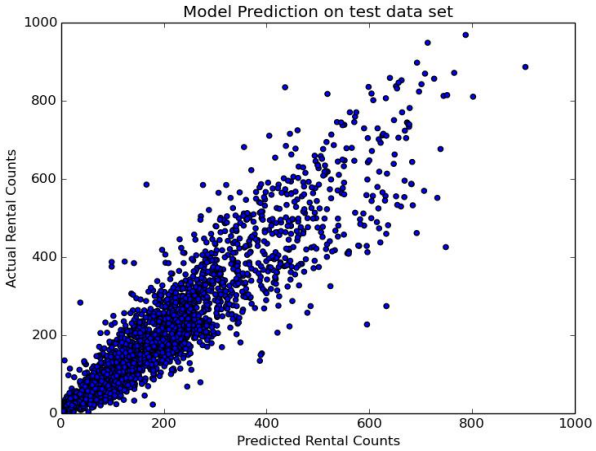
Plot for residuals



Prediction Count Plot



Residuals :Exponential Gradient Boosting+ PCA



Prediction: Exponential Gradient Boosting +PCA

Conclusions

Model	Error : Total rentals	Error : casual rentals	Error: registered rentals	Error on combined results (combined)
Linear Regression	11931.075	1068.645	8543.238	9533.932
Linear Regression + PCA	12261.3183175	1025.30143657	8312.7347094	9361.99912
Logarithmic Linear Regression	0.621125	2.055288	0.68553	0.50271
Logarithmic Linear Regression + PCA	0.64078	2.0735	0.6702	0.50052
Ridge Regression	12960.9690416	1049.19105427	8215.83914121	9231.91065263
Ridge Regression +PCA	11308.911407	939.762690265	8095.20493292	9069.51635324
Random Forest	9281.41135645	1419.57104252	14218.8442677	15550.7342488
Logarithmic Random Forest	0.573611186141	2.08582654967	0.633126757344	0.837809727457
Gradient Boosting	7044.75413279	549.016689845	5164.1756597	5809.9210437
Gradient Boosting + PCA	6048.59202504	423.155997018	4414.17601006	4863.42048033
Exponential GBM	0.418082132506	2.07649862424	0.453996824344	0.3697752905
Exponential Gradient Boosting + PCA	0.390918826588	2.05931175299	0.354472028901	0.379172566378

Error: Square error for non-logarithmic models
Error: Root mean squared logarithmic error for logarithmic models

Observation:It is clearly observed that the error is lesser (in all cases) if we split the total count into casual rentals and registered rentals. Thus our initial intuition of predicting for casual and registered users separately is a valid one. It is observed that there is a difference in rental patterns among registered and casual users.

Non-linear models in general give a better prediction.

Model	Training Accuracy (on total rentals)	Test Accuracy (on total rentals)	Training Accuracy (on casual rentals)	Test Accuracy (on casual rentals)	Training Accuracy (on Registered Rentals)	Test Accuracy (on registered rentals)
Linear Regression	64.02%	62.89%	58.78%	56.58%	63.81%	60.53%
Linear Regression + PCA	64.35%	61.88%	58.04%	58.88%	63.35%	62.17%
Logarithmic Linear Regression	80.51%	80.86%	82.26%	83.06%	79.56%	77.58%
Logarithmic Linear Regression + PCA	80.76%	80.16%	82.50%	82.39%	78.96%	79.25%
Ridge Regression	64.00%	62.85%	58.31%	58.02%	62.77%	63.89%
Ridge Regression +PCA	62.91%	66.12%	58.44%	57.43%	62.62%	64.19%
Random Forest	77.80%	72.07%	47.51%	45.66%	36.79%	36.88%
Logarithmic Random Forest	84.45%	83.73%	49.30%	49.32%	45.13%	45.13%
Gradient Boosting	80.94%	79.04%	83.11%	78.39%	81.63%	78.95%
Boosting+PCA	86.64%	81.29%	89.31%	83.86%	88.12%	80.57%
Exponential GBM	92.82%	91.09%	87.79%	86.22%	90.86%	89.45%

Exponential Gradient Boosting + PCA	98.72%	92.68%	90.76%	87.38%	91.34%	88.50%

Among all the tested models Exponential Gradient Boosted Method performs the best. In almost all the cases the models where the response variable was predicted to be logarithmic with input vectors. In few cases models after PCA didn't result in better results.

References:

Kaggle forums: https://www.kaggle.com/c/bike-sharing-demand/forums
Leo Breiman, Random Forests, Machine Learning, Volume 45 Issue 1, October 1 2001, Pages 5-32