Prediction for Bike Sharing Service Profitable for the Company

Introduction to Statistical Learning

Boram Shim

University of Minnesota

Introduction

The project is an analysis of prediction for bike sharing service profitable because the company which is in the city of London wants to know if its sharing service is profit or loss, depending on the number of bikes daily from 2015 to 2016. If the number of shared bikes is over 20000, it is profit. If not, it is loss. The goal of this analysis is to understand which variables contributed the most in making the bike sharing services profitable for the company which operates it.

This paper consists of Methods, Results and Discussion. In Methods, statistical methods, which are used logistic regression, K-Nearest Neighbors (KNN), classification trees and random forest, are described and explained reason why the four methods are selected. In Results, it is described how dataset is set up for this analysis. The result of the analysis is reported with all the relevant output and plots. Also, adequately interpretation is provided the performance of the four methods as comparing the in term of predictive accuracy with the area under the (ROC) curve (AUC). In Discussion, a summary of the analysis is provided and discussed, and the meaning of conclusion is explained. Last limitation for this analysis and further research is discussed.

Method

The datasets, which are contained information the number of bikes shared daily from 2015 to 2016 in the city of London, splits 80% of the data into the training set and the remaining 20% into the validation set.

The first main method to be implemented is logistic regression. It is an appropriate method of analysis if the value of response variable is binary. The response which is called “Profit” is binary as profit (= 1) or loss (= 0), so logistic regression is chosen for this project. Second main method to predict Profit response is KNN classifier which estimates posterior probability by means of the K nearest points in the training set. KNN is a completely non-parametric approach and thus it does not require linearity of the decision boundary and does not require assume a specific functional form. Unlike LAD and QDA, it does not follow Gaussian distribution, so the KNN algorithm works well with the numerical variables. Therefore, KNN is appropriate method for this analysis which has categorical variables and numerical variables in data. Third method is classification tree which is also non-parametric method and implement to predict Profit. Classification tree is similar to regression tree except that they are used when the response is qualitative. It is preferred when the relationship between response and variables are far from linear. Radom forest aims to further improve bagging by constructing decorrelated trees. Specifically, it is used bootstrapped training datasets but, in this case, at each split, instead of considering all the predictors, it is only considered M of them which we select randomly. M is usually $\sqrt{p}$. Random forest avoids situations where on very strong predictors may lead to bootstrapped trees which are very similar and thus correlated among each other. Therefore, this method is selected for this analysis. Boosting is also good method, but it will take a long time to select the best combination parameter and cost is expensive. Therefore, random forest is choosing for this analysis.

After using four methods, we will compare each method's ROC curves and AUC and test error rate to choose the best model to predict Profit with comparing between excluded missing data and imputed missing data.

Results

- Dataset Description

There are two types of data. In the first dataset, humidity variable in the dataset contains missing values impute them which the mean of that variables. Holiday variables in the dataset contains missing values exclude the observations with missingness from the analysis. N_bikes as response variable contains missing values exclude the observation with missingness from the analysis. In the second dataset, using the original dataset, if any of the categorical prediction in the dataset contains missing values, create a new category missing and assign such category to the missing observations. Whereas, impute all the quantitative variables in the dataset containing missing using iterative regression. Based on those two types of data, all method is repeated and get two version of result. In this result page, the two types of data results are compared and selected as best model. To make it simple, the first dataset is assigned as Data 1 and the other is Data2.

- Logistic Regression

Table 1. Logistic Regression Output with Data 1

	Estimate	Std. Error	P-Value
Intercept	7.9708	1.6107	**7.48e-07 *** **
holiday	-3.9245	1.1379	**0.000563 *** **
weekend	-4.7810	0.6516	**2.18e-13 *** **
season	-0.1101	0.2032	0.587894
temperature	9.5248	11.9517	0.425485
feels_like	0.2942	9.8998	0.976292
humidity	-9.3652	2.0733	**6.27e-06 *** **
wind_speed	-7.0077	1.9393	**0.000302 *** **

From the Table 1, logistic regression shows that p-value for holiday, weekend, humidity, and wind_speed is less than 0.05 and it means that those variables are significant statistically. On the other hand, season, temperature, feels_like has large p-value, those variables are not significant statistically. If for every 1 unit increase in humidity, the log odds of N_bikes decrease by 9,3652 with Data 1. Therefore, humidity has negative effect on the number of bikes, and it leads to loss. If for every 1 unit increase in temperature, the log odds of N_bikes increase by 9.5248 with Data 1. It has positive effect on the number of bikes, and lead to profit. Like this, temperature and feels_like has positive impact for getting profit.

- K - Nearest Neighbors (KNN)

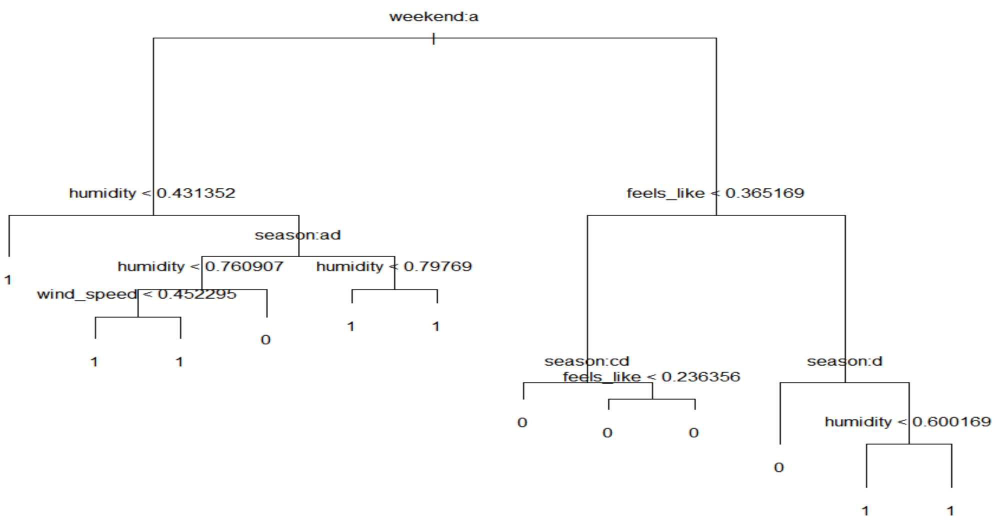
Table 2. The Result of Test Error Rate using KNN

	Test Error
K=3	0.1046512
K=5	0.08139535
K=15	0.09302326

KNN method id provided test error rate and ROC curves. From the Table 3, there are three KNN performance and test error rate with K= 3, 5, 15. It shows when K is increased, the test error rate is increased and decrease. It looks like "U" shape. For KNN method, we expect when K is increased, the model is less flexible. And, the training error rate increases, and the test error rate will have the usual "U" shape as K increase due to bias-variance trade off.

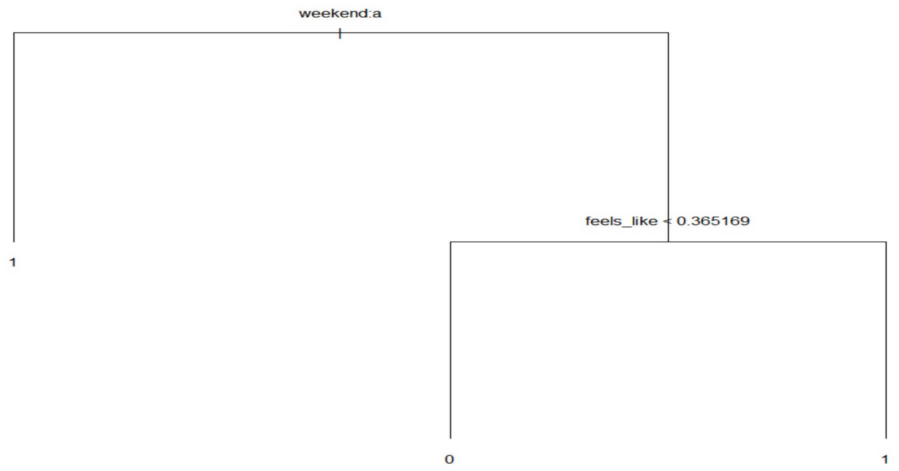
- Classification Tree

Plot 1. Classification Tree with Data 1



From Plot 1, we can see that all the variables were not used to construct the tree. It is because at each split the tree includes the variable which is importance. In this case, there are only 5 variables as weekend, humidity, season, wind_speed, feels_like. It means that only these variables lead to the larges reduction of error rate has large impact on response. The classification tree shows that weekend, humidity and wind_speed is significant like logistic regression while feels_liks and holiday shows different result.

Plot 2. Pruned Tree with Data 1

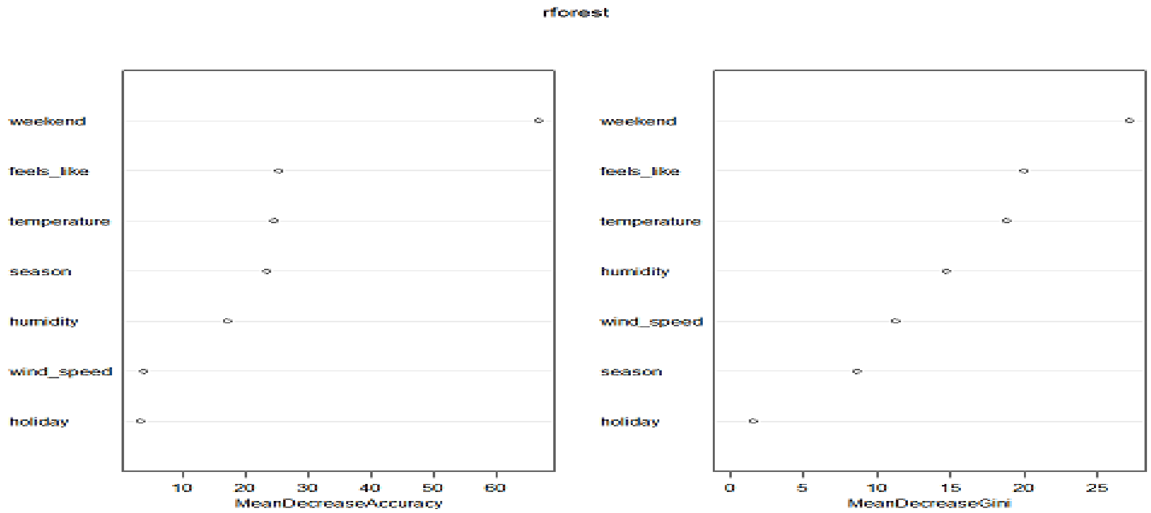


From Plot 2, there are only 2 variables as weekend and feels_like after pruning. In Data 1, the test error rate of tree is 0.127907 and the test error rate of prune is 0.1162791. The test error of pruned tree is lower than the test error rate of tree because the pruned tree involves less splits. It is more parsimonious, and it reduce the risk of overfitting.

Both tree and pruned tree shows that the most important variable is weekend and it is same as logistic regression result.

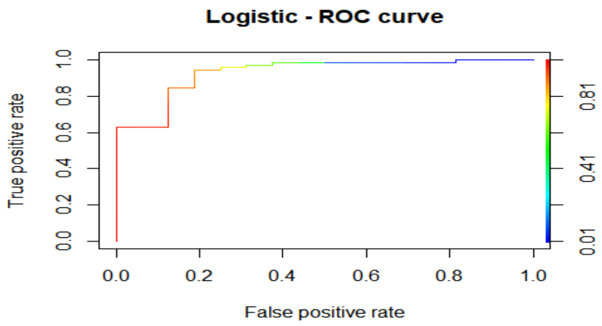
- Random Forest

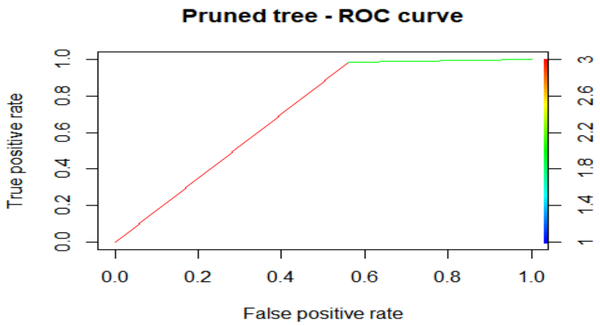
Plot 3. Importance Plot for Random Forest with Data 1



From the Plot 3, the graph shows that the most impactful variable is weekend. It shows same as logistic regression, tree, and pruned tree result. Next impactful variable is feel_like and temperature in order. Whereas holiday has almost no impact. When considering the reduction in the Gini index to evaluate the importance of the variables used when constructing the trees, the results are coherent with those of tree. There is one difference between logistic regression and random forest method. Temperature and feels_like p-values is large in logistic regression, but this method shows that the two variables are impactful.

- Comparing AUC and Test Error Rate





Plot 4. ROC Curves for using all methods with Data 1

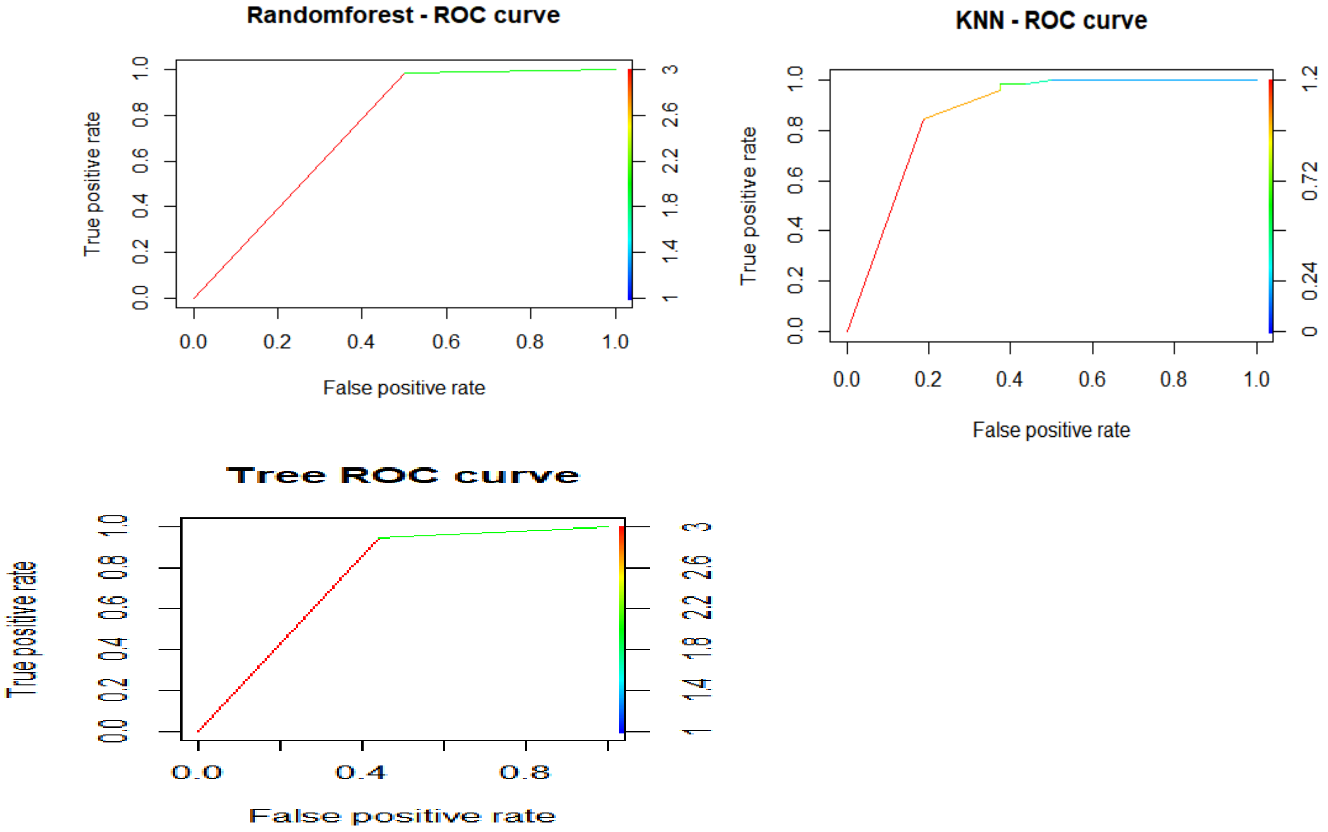


Table 3. AUC and Test Error Result with Data 1

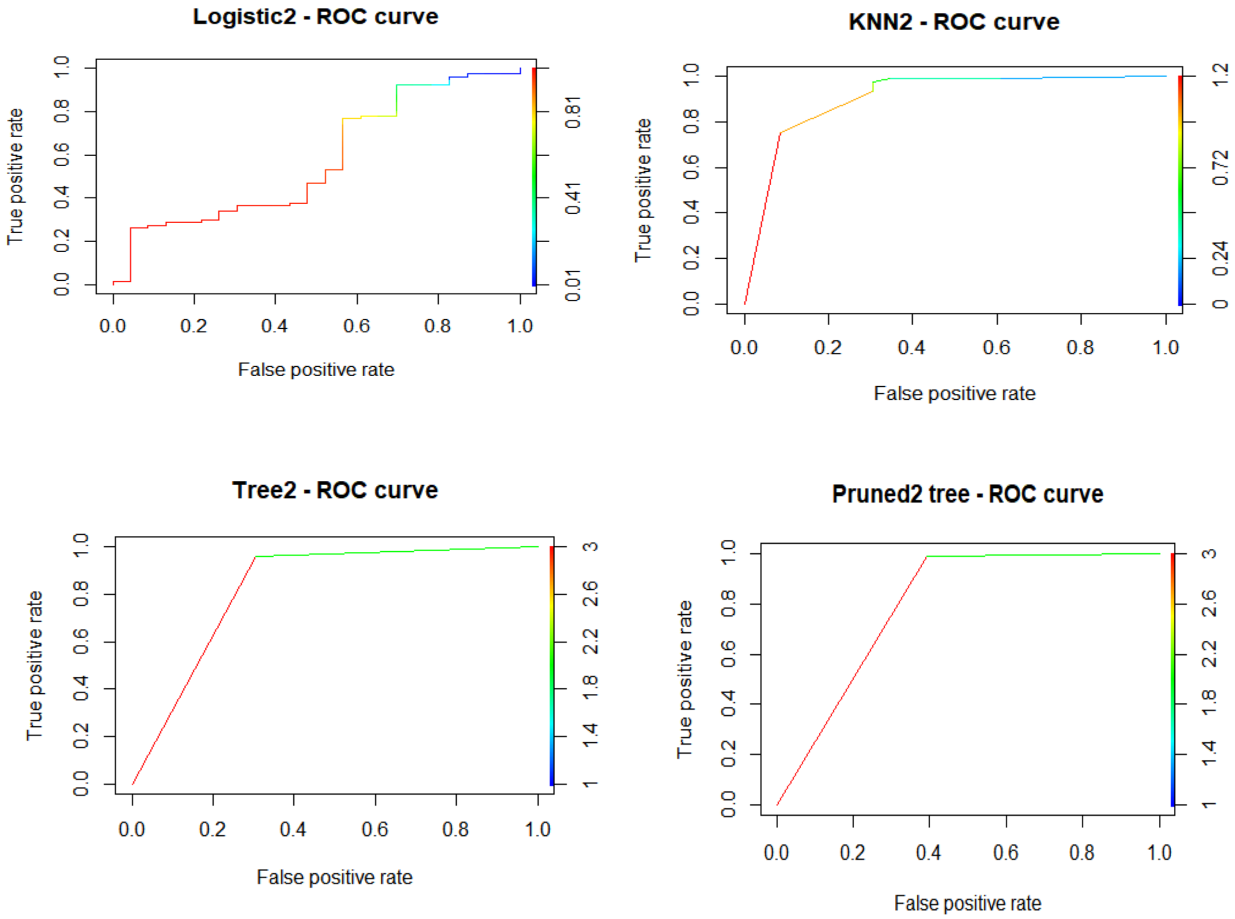
	Logistic Regression	KNN K=5	Classification Tree	Pruned Tree	Random Forest
AUC	0.9295	0.8714	0.7527	0.7116	0.7429
Test Error	0.08139	0.08139	0.1279	0.1163	0.093

From Plot 4, there are ROC curve with using all method. It will do well when the ROC curve is close to 1 and AUC is large. In other words, the largest the AUC, the better the performance as this implies high sensitivity and high specificity. From Table 3, logistic regression and KNN provides the lowest test error rate as 0.08139. The largest AUC is logistic regression as 0.9295. It means that logistic regression performs better than other methods in Data 1.

Tree has the largest test error rate and small AUC. The results indicate that there is a possibility that variables which are important to each tree splits, may have been missed. So, tree performs poorly in Data1.

Contrary to expectations, random forest AUC is lower than classification tree AUC slightly. However, we can guess random forest method is better than classification tree because random forest has low variance than classification tree. It means if thresholds are different, random forest outperform compare to other methods.

Plot 5. ROC Curves for using all methods with Data 2



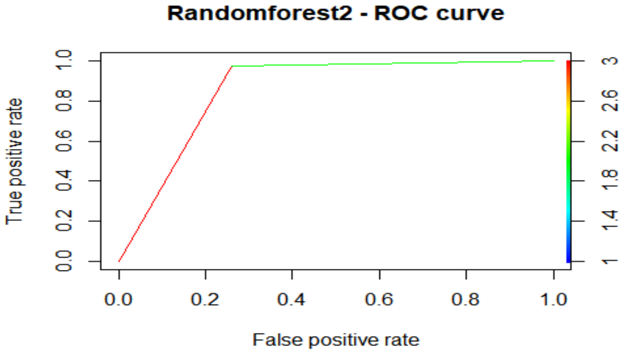


Table 4. AUC and Test Error Rate Result with Data 2

	Logistic Regression	KNN K=5	Classification Tree	Pruned Tree	Random Forest
AUC	0.5737	0.9051	0.8283	0.7979	0.8566
Test Error	0.09	0.12	0.1	0.1	0.08

From Table 4, it shows different result from the result of Data 1. Logistic regression and random forest are provided the low test error rate as 0.09 and 0.08. However, logistic AUC has the lowest. The largest AUC is KNN as 0.9051. Next high AUC is random forest as 0.8566. Based on overall result with Data 2, KNN performs well in this analysis with Data 2. Tree is belonging to large test error rate compare to other methods in Data 2. The results indicate that there is a possibility that variables which are important to each tree splits, may have been missed. So, tree performs poorly in Data 2 also.

Table 5. Comparing AUC and Test Error Rate Result with Data 1 and Data 2

Data 1 | Data2	Logistic Regression		KNN K=5		Classification Tree		Pruned Tree		Random Forest	
AUC	0.9295	0.5737	0.8714	0.9051	0.7527	0.8283	0.7116	0.7979	0.7429	0.8566
Test Error	0.08139	0.09	0.08139	0.12	0.1279	0.1	0.1163	0.1	0.093	0.08

From Table 5, in Data 2, AUC results are larger than results in Data 1, except for logistic regression. Logistic regression and KNN in Data 2 test error rate are larger than Data 1, others are lower than Data 1. What is interesting about this analysis is that when K= 15 the test error rate is 0.06 which is the smallest test error rate compare to other methods. When K=5, KNN has largest test error rate in Data 2. K has been designated as 3, 5, and 15, as in Data 1, but when K=5, 15, 20 in Data 2, test error rate looks like “U” shape and get AUC result is 0.9288 which is pretty large AUC. KNN is a non-parametric method so when we pick K, it has automatically decided its flexibility. In this case, K= 5 may have made the model too flexible, so it leads low bias with low training error rate and high variance with large test error rate. So, we improve it picking K > 5 because when K increases the flexibility is decreased. Therefore, we need to consider this case and pick K=5, 15, and 20 for better performance.

In case of Data 1, we discard holiday and N_bikes missing variables and impute humidity with the mean of that variables. So, the sample size is reduced with all the problems which come with it and we are throwing away a lot of the information available. In case of Data 2, we impute all missing values of the quantitative variables in the data set using iterative regression. This approach allows a natural generalization of the univariate approach. It does not assume any joint distribution for the missing values, whereas in some cases it may sensible to do so.

Discussion

This project analysis to predict for bike sharing service profitable because the company which is in the city of London wants to know if its sharing service is profit or loss, depending on the number of bikes daily from 2015 to 2016 with given effected variables.

As a result, KNN with imputing missing values and K= 15 is best model to predict this project based on the result of low test error rate and close to 1 for AUC. After discarding missing values, sample size will reduce, so imputing missing values are better. In addition, we do not know whether a given model is linear or not. Therefore, it is best to use KNN, QDA, and classification tree that are not linear. However, QDA is excluded from the method used because variables have to follow Gaussian. And the tree method has risk to miss important variables, so we can see that KNN is provided the best result in the end.

One interesting point is the logistic regression. Unlike other methods, the AUC variation is greatest despite the similar test error rate for the logistic regression model. Because the logistic regression is a linear model, and it is one of the parametric methods which are assumed a functional form for a model which is characterized by some unknown parameters, it is possible to predict that it result in different outcomes depending on the data.

The limitation of this analysis is that the number of observations is not enough to predict Profit or not very well. After discarding missingness, observations are only 429 and it is insufficient. Test error rate is not reliable since the randomization with the observations leads to high variance when using the validation set approaches. If there are more observations and variables, the analysis will outperform and clear to predict for bike sharing service profitable for the company.