Introduction

This project analyzes information on the number of bikes shared daily from 2015 to 2016 in the city of London. The analysis aims to understand which variables explain the most in the number of bike shared for the bike sharing company and to construct a predictive model to forecast the probability of their business operation.

For statistical learning methods, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest were used.

This paper follows with three sections. In Method, the four statistical learning methods that were used in this project are explained. In Results, all the relevant outputs and plots of the analyses are presented. All the detailed interpretation is followed to look for meaning that are resulted through the analyses. The performance of the four implemented statistical models are compared by Test Error Rate, ROC, the area under the curve (AUC). The best predictive model is selected based on those standards. In Discussion, the results and limitations of the analyses are summarized, and the recommendation for the related future research is added.

Methods

The original dataset includes 500 observations with 9 variables. Among the variables, *N_bikes* is used as a measurement to decide the response variable of this project, *Profit (1= Profitable, 0=Not Profitable)* when at least 20,000 bikes were shared. The *date* variable is divided into *Year*, *Month*, and *Day* columns. The *holiday*, *weekend*, and season variables are qualitative, and the *temperature*, *feels_like*, *humidity*, and *wind_speed* are quantitative.

First, the four quantitative variables are being normalized. This process intends to avoid one predictor intrinsically influences the result more due to its larger value. The goal of normalization to change the values of numeric columns in the dataset to a common scale is achieved.

Logistic Regression is a statistical learning classification algorithm that predicts probability of a categorical response variable. This project uses *Profit*, which is binary as response. As the most important member of Generalized Linear Models, Logistic Regression predicts probabilities of *Profit* that ranges 0 to 1.

K-Nearest Neighbors (KNN) is another machine learning classification algorithm that looks at the neighboring data points to determine what this new data point will fall into. This nonparametric classification does not make any assumptions on the underlying data distribution so its flexibility gives an advantage to explore non-linear decision boundaries for this project.

Decision Tree is another non-parametric supervised learning method used for classification and regression. Decision tree builds classification or regression models in the form of a tree structure. The result is a tree with decision nodes and leaf nodes. Since this project includes both quantitative and qualitative predictors, the Decision Tree method can handle both properly. Specifically, Classification Tree algorithm is used for the response is binary variable.

Bootstrapping is used to improve the Decision Tree method further. Bootstrapping is a sampling technique in which we randomly sample with replacement from the data set. Bagging methods uses bootstrapping to create bagged trees by creating B number of decision trees that is trained on bootstrapped training sets. Random Forest also bootstrap trees but decorrelates the trees by introducing random subset of predictors. This avoids only few features are repeatedly selected which would have made the result strongly correlated.

Missingness of data is handled in two following ways. Discarding the missing values is adopted for the response variable and categorical variables in the first course of analyses. Then, the second analyses implement imputing method using iterative regression for the original set.

Results

i. Exploratory Data Analysis (EDA)

Before implementing statistical models on dataset, Exploratory Data Analysis performs initial investigations on data to discover certain patterns, to identify relationships between variables, and to check assumptions with the help of summary statistics.

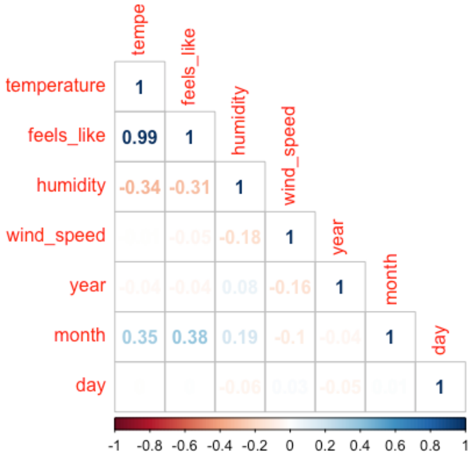

Figure 1. Correlation matrix of numerical variables

Figure 1 shows that *temperature* and *feels_like* variables are considerably correlated. The variables *temperature* and *feels_like* against *humidity* are subtly correlated but not significantly, as well as against *month*. Among the categorical variables, the variable season contains the pattern of what month variable is against *N_bikes*. Thus, the season variable replaces *year*, *month*, and *day* variables in the main analyses.

$$Profit \sim temperature + feels\_like + humidity + wind\_speed + holiday + weekend + season$$

Equation 1. Basic equation that will be built upon the statistical learning analyses

ii. Data Split

Validation set approach is used to split the dataset into train and test set. 80 percentage of data is randomly contained to train data set, and remaining 20 percentage of data is test set. Following statistical learning models utilize this validation set approach to tune the model with the train sets and validate it with test sets to obtain the test error rate.

iii. Logistic Regression

The regression result of Equation 1 is given in Table 1 below. This trial is resulted from when we discard the missing values in response and categorical variables.

	(Intercept)	temperature	feels_like	humidity	wind_speed	holiday	weekend	season
Estimate	**7.4539**	**3.7374**	**1.5105**	**-6.7678**	**-5.6196**	**-3.0938**	**-3.5898**	**-0.44**
P-value	**6.92e-08** **(1.3821)**	**0.6928** **(9.4590)**	**0.8511** **(8.0461)**	**1.57E-05** **(1.5671)**	**6.02E-05** **(1.4007)**	**0.0162** **(1.2868)**	**1.42E-12** **(0.5069)**	**0.0242** **(0.1952)**
Residual deviance	**172.51**							
AIC	**188.51**							

Table 1. Regression Results for Equation 1. Standard Errors in the Parentheses.

From the result, it is indicatable that *humidity*, *wind_speed*, *holiday*, *weekend*, and *season* variables are statistically significant, whereas temperature and feels_like variables are not. Since it is already known the *temperature* and *feels_like* variable are correlated, the equation excluding *feels_like* (Equation 2) is tested as well. In result, *temperature* becomes a significant variable.

	(Intercept)	temperature	humidity	wind_speed	holiday	weekend	season
Estimate	**7.3855**	**5.4982**	**-6.7015**	**-5.7264**	**-3.0947**	**-3.5988**	**-0.4396**
P-value	**2.67E-08** **(1.3821)**	**2.16E-05** **(9.4590)**	**1.08E-05** **(1.5671)**	**7.83E-06** **(1.4007)**	**0.016** **(1.2868)**	**1.02E-12** **(0.5069)**	**0.0243** **(0.1952)**
Residual deviance	**172.54**						
AIC	**186.54**						

Table 2. Regression Results for Equation 2.

For both equations, the residual deviance and AIC are very similar but the Equation 2 has one less variable and slightly lower AIC, which tells the better fit of the model. From next analyses, results with the Equation 2 are presented.

Signs of the estimates shows that the *temperature* only positively impacts the probability of *Profit* of the bike sharing services as opposed to other variables. Interpretation of the estimates include that for 1 unit of degree increase in *temperature*, the log odds of *Profit* increases by 5.4982. Likewise, the negatively influential variables against profit can be understood by same methodology.

iv. K-Nearest Neighbors

K-Nearest Neighbors (KNN) uses the train data to learn the model and the test data to see how well the model performs on unseen data. It is significant to choose the best K, a parameter that refers to the number of nearest neighbors to include in most the voting process, that makes the model optimal.

K was chosen through cross-validation. It takes the small portion from the training dataset and calls it a validation dataset, and then uses the same to evaluate different possible values of K. This approach predicts the label for every instance in the validation set using K ranging from 1 to 10 (the Figure 2 takes the range of 1 to 50 to explain the bias-variance tradeoff in “U” shape) and then looks at what value of K gives us the best performance on the validation set. Once it finds the optimal value, it is used as the final setting of KNN to minimize the validation error.

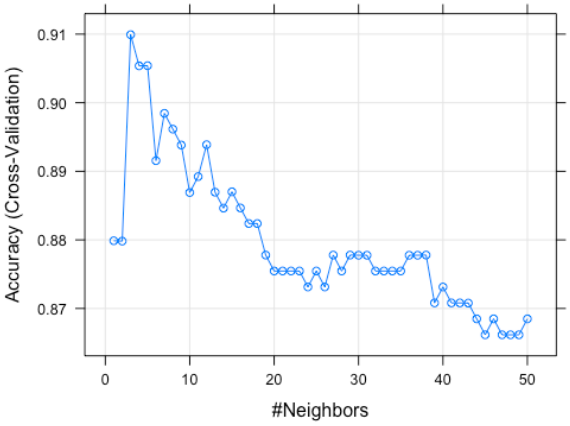

Figure 2. Plot of Accuracy of the Model against K

Figure 2 clearly shows that the accuracy (1-Error Rate) is highest when K = 3. Noticeably, the accuracy continuously decrease as K increase, and this phenomenon shows an example of why overfitting happens when the model become more flexible, with larger K in this method.

v. Decision Tree (Classification Tree)

Decision Tree uses the train set to build a tree model which can use to predict class or value of target variables by learning decision rules inferred from training data. This project predicts the binary response so the classification tree that predicts belonging class is used.

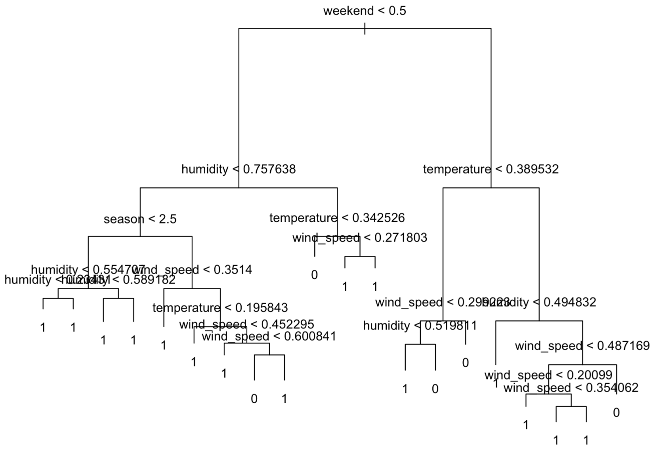

Figure 3. Result of the Classification Tree Model

Figure 3 depicts that the 6 variables are used as leaf to build the classification tree. Especially, it is notable that the weekend variable functions as a root of the tree, which means the most influential predictor of the model against Profit. With the 6 variables used, the lowest error rate is reached when the stopping rule is achieved.

vi. Pruning Tree

Pruning is a technique that reduces the size of decision trees by removing features of the tree that provide little explanation to class instances. The smaller tree with fewer splits avoids overfitting and achieves lower variance.

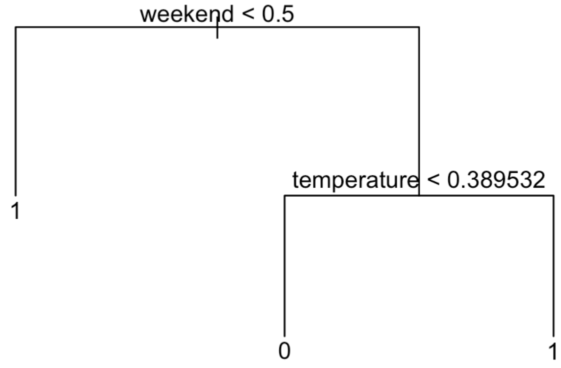

Figure 4. Result of the Pruned Tree Model

As same as classification tree, the *weekend* variable still plays the most important role in this method as a root leaf. The *temperature* comes as next branch and it can be regarded as the second most influential variable to be selected through the pruned tree method.

vii. Random Forest

The Random Forest algorithm builds multiple decision trees through bootstrapping train sets and merges them together to get a more accurate and stable prediction. Despite trading easy interpretation and clear visualization of decision tree, the Random Forest method provides variable importance plot to indicates which variables are the most important predictors.

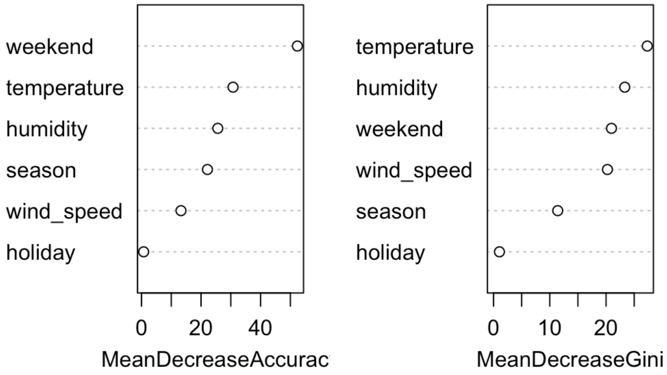

Figure 5. Variable Importance Plot of the Random Forest Model

The Mean Decrease Accuracy plot shows that the *weekend* variable is the most important variable among predictors, just like Decision Tree methods found out. The Mean Decrease Gini describes the *temperature* variable has the highest purity among variables, followed by *humidity*, *weekend*, and *wind_speed*. If Equation 1 were used where *feels_like* was not excluded, the purity of *temperature* would have been not that high. Likewise, the variable *holiday* has lowest MDG as it often comes along with the *weekend* variable.

viii. Model Selection

K-Fold Validation approach is adopted to compare the four classification models and select the best fitting model. The K-fold validation approach randomly assigns 20% data into training set and 80% data into test set, and it repeats same procedure for 10 times (k=10). Once we get the test ER and AUC and obtain the average of them to compare the models.

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes.

	Logistic Regression	KNN	Classification Tree	Pruned Tree	Random Forest
Test Error Rate	0.146324	0.128392	0.1204545	0.1273256	0.1041226
AUC	0.9483235	0.436233	0.841811	0.7751465	0.9249169

Table 3. Test Error Rate and AUC for the dataset that discards missing data

Table 3 explains that the Logistic Regression performs poorly on this dataset as its test error rate is largest among models despite largest AUC. This linear classification method does not work with the dataset well which describes the lack of linearity among variables that are enough to predict the Profit in the equation. Unlike Logistic Regression, non-linear classifier including KNN, Classification Tree, Pruned Tree, and Random Forest performed better to this dataset. Especially, it is noticeable that the test error rate of the Random Forest is the lowest which indicates that the model reduces the variance of the model as they reduced the complexity, in other words, the number of features that are used to predict the response and make the tree uncorrelated.

In addition, the AUC of the Random Forest is the larger than other non-linear classification. As we set up the threshold in the classifier to 0.5, the Random Forest model outperforms the Classification Tree and the Pruned tree as well. It can be challenged, however, as the threshold is adjusted to check how it can alter the results of the non-linear classification methods.

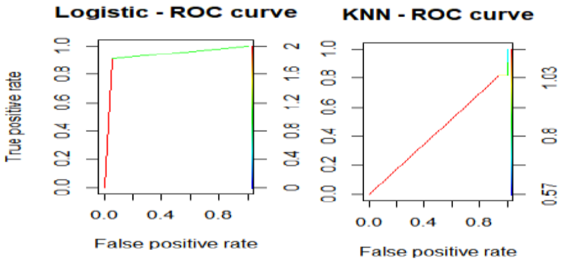

Figure 6. ROC curves for the Logistic Regression and KNN classifications

Figure 6 portraits that the ROC curves for the Logistic Regression fits close to 1 which is already observed through high AUC. However, it is indicatable that the KNN's ROC curve diagonally half which goes hand in hand with the lowest AUC, 0.43623.

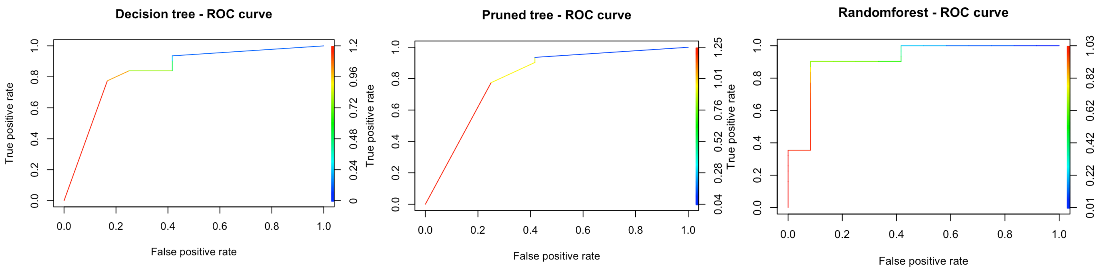

Figure 7. ROC curves of the Decision Tree, Pruned Tree, and Random Forest models

Figure 7 shows that the Random Forest has the largest AUC that is measured through the ROC curve that fits close to 1 of the plot. However, it must be noticed that when we adjust the threshold, the result can be different that other classification methods could outperform the Random Forest method, which is the best performing method for this dataset with 0.5 threshold.

ix. Analyses of imputed datasets

	Logistic Regression	KNN	Classification Tree	Pruned Tree	Random Forest
Test Error Rate	0.131232	0.121252	0.1157505	0.1251586	0.1087738
AUC	0.647743	0.565754	0.849327	0.7448383	0.9196451

Table 4. Test Error Rate and AUC for the dataset that imputes missing data

Using the original dataset, categorical predictors in the dataset contains missing values creates a new category missing and assign such category to the missing observations. Whereas, all the quantitative variables in the dataset containing missing values are imputed using iterative regression.

The results that are shown in Table 4 are not distinguishably different from the results that is obtained with the dataset that discards the missing observations. Unlike the previous result in Table 3, the Logistic Regression performed poorer indicating that the imputed data made the datasets more centered distribution, which might have led to faint the linearity between variable and response. Likewise, the lowest Test Error Rate is shown through the Random Forest method, 0.1087738. For the dataset that was proceeded with the iterative regression imputation, the non-linear classification outperforms the linear classifier, and we can know that from the start that the variables are not assumed that they are jointly normally distributed.

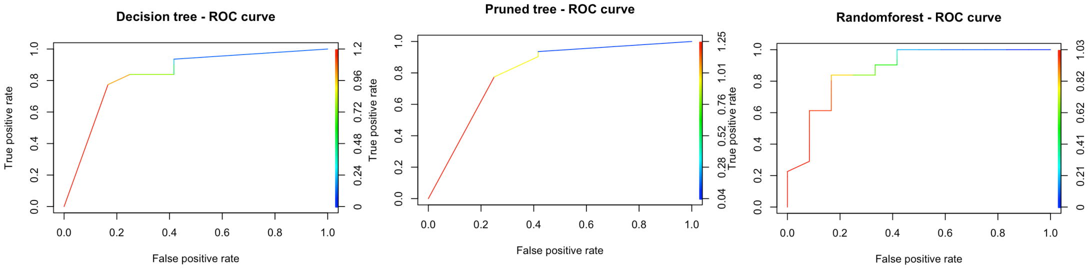

Figure 8. ROC curves of the Decision Tree, Pruned Tree, and Random Forest models

Also, when the iterative regression imputed the data, it would have introduced biasness to the data that may not be relatable to the missing values. Because of this biasness of the dataset, we observe the lower AUC between same methods, described in Figure 8.

Discussion

The goal of this project to identify which variables explain the most in the number of bike shared for the bike sharing company and to construct a predictive model to forecast the probability of their business operation is achieved through analyses.

The results showed that the non-linear classification method that specifically introduced the randomness of selecting features of the dataset, which is Random Forest method, predicts best to the profitability of the bike sharing operation. The linear classification might have performed better when more quantitative variables are collected that are especially able to form Gaussian distribution. The lowest test ER was provided through the Random Forests, and the model indicates that the importance of variables is in order of *weekend, temperature, humidity,* and so on.

The limitations could improve this research better. One suggestion is to collect the sample size. Because we have small number of datasets as well as a good amount of missingess to observation, the interpretation of these statistical learning methods can be not reliable. Another suggestion can be adding more variables that can predict the response better. The *temperature* variable was already strongly correlated to *feels_like* variables, and future studies can avoid these problems.