Prediction of The Failure of Affordable Housing Projects

Byeongmo Kang

University of Minnesota Twin-Cities

Introduction

The purpose of this report is to provide affordable housing in San Francisco, California. Through this report, we can (1) understand the variables that most affect the failure of affordable housing projects, and (2) construct predictive models to help predict the probability of failure of future projects aimed at providing affordable housing for major San Francisco companies.

There were 29 variables in an initial dataset. However, I selected 6 predictor variables to predict the probability of failure of affordable housing projects. The response variable was Failing (1=Yes, 0=No) that does not provide affordable housing. I selected two continuous variables that were Market Rate and Family Unit and four categorical variables that were Status with 6 levels, Program Area with 8 levels Project Type with 2 levels and Housing Tenure with 4 levels.

In order to predict the failure of affordable housing projects, two datasets were used. First data excluded missing values and second data imputed missing values. Comparing various models in the two datasets, second datasets tended to have smaller the test error rate and larger AUC. Furthermore, unpruned tree and pruned tree model in dataset 2 was the best model to predict the failure of future projects.

Method

There were total 289 observations in the original data. However, there were 29 missing observations in Failing. Two datasets were used to manage missing values. In first dataset, I excluded the missing values in Failing. Since there were no missing values in continuous and categorical variables I chose, I did not imputed and excluded them. In second dataset, I created a new category *missing* when there were missing values in the categorical variable. Also, I used random regression imputation in order to deal missing values in the response variable. Then, I made model with variables I chose.

In the model, there were total six predictors that were four categorical variables and two continuous variables. To analyze the model, the most appropriate level of categorical variables was used.

Three classification methods were used to predict the failure of affordable housing projects; Binomial logistic regression, K-nearest neighbor(KNN), Classification tree. Using binomial logistic regression has the advantage of easy to interpret by using the regression coefficients in terms of Odds Ratios. Odds are an alternative to probabilities that represent probabilities. Odds are an alternative to probabilities that represent probabilities. (Handout 6 Generalized Linear Models, p.7) Pearson-x^2 was used to verify that the binomial logistic regression was performed well.

KNN classifier estimates by means of the K nearest points in the training set. (Handout 8 Classification, p.3). Also, KNN is a completely non-parametric approach and it does not require linearity of the decision boundary. (Handout 8 Classification, p.10) Therefore, I expect KNN to perform well in this model.

The reason for using tree classification is to create multiple regions corresponding to different values (or intervals) of the predictors. (Handout 9 Tree based methods, p.1) According to Sun(2008), “tree classifier is its ability to using different feature subsets and decision rules at different stages of classification.” Since there were categorical variables in model, best subset selection was used in the model. Therefore, I expect tree classification to perform well in this model. After making the tree, pruned tree was also used in order to avoid a risk of overfitting.

Finally, bagging was used in order to high variance in the tree.

To analyze the performance of the three classification methods, the error rate of each method was calculated through the validation set. The training set and validation set ratios are 70% and 30% respectively. ROC curves were used for accuracy prediction, with the smallest error rate and the largest AUC being selected as the best approaches.

Result

Dataset 1

I selected 6 predictors Market Rate, Family Unit, Status, Program Area, Project Type and Housing Tenure. There were no missing values in predictors. However, there were 29 missing values in Failing that was response variable. In dataset 1, I excluded the 29 missing values. By computing best subsets regressions, Market Rate, Family Unit, Status 2, Program Area.Inc, Project Type.Rehabilitation Housing Tenure.Ownership were selected. Status2 means that whether construction status of project at time of report is Predevelopment Feasibility or not. Program Area.Inc means that whether Program or OCII project area under which project is financed, underwritten or developed is Inclusionary or not. Project Type.Rehabilitation means that whether type of development project is rehabilitation or not. Housing Tenure.Ownership means that type of arrangement households can occupy the housing is ownership is or not.

208 observations were randomly assigned to the training set, and 52 observations were randomly assigned to the validation set.

Table 1: Summary of binomial logistic regression from dataset 1

	Estimate	p-value
Intercept	$-7.873 * 10^{14}$	<2e-16
Market Rate	$2.504 * 10^{12}$	<2e-16
Family Unit	$1.612 * 10^{12}$	<2e-16
Status 2	$-7.300 * 10^{11}$	<2e-16
Program Area.Inc	$9.984 * 10^{14}$	<2e-16
Project Type.Rehabilitation	$3.365 * 10^{14}$	<2e-16
Housing Tenure.Ownership	$3.365 * 10^{14}$	<2e-16
Null deviance	239.50 on 181 degrees of freedom	
Residual deviance	504.61 on 175 degrees of freedom	

First of all, logistic regression was used in the training set. When I look at p-values, p-values from all predictors are smaller than 0.05. It means that all of predictors are statistically significant. Therefore, all of predictors affect the predicting failure of future projects. Especially, Program Area.Inc has the highest estimate($9.984 * 10^{14}$). It indicates that it is $2.7139326 * 10^{14}$ times more likely to fail the future project when program area is Inclusionary. The Pearson -x^2 was conducted in order to check whether this model has good fit or not after a biological logistic regression. However, p-value of binomial logistic regression model is small enough (close to 0). Hence, this model was rejected.

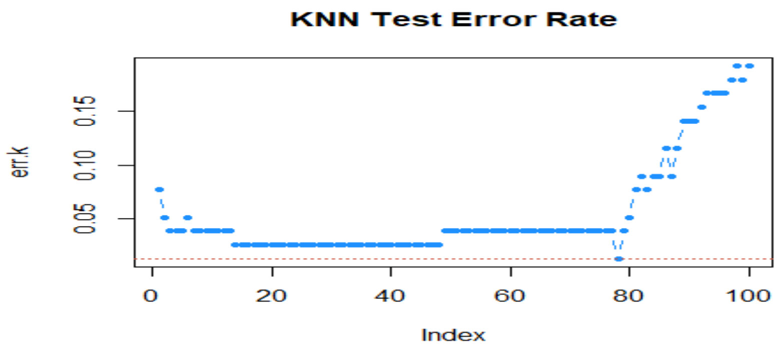


Figure 1:Graph for KNN test error rate from dataset 1

Second, I formed the KNN model. The number of k with the smallest test error rate (0.01282051) is selected. From the graph above, it indicates that k=78 has the lowest test error rate. Since KNN is the non-parametric method, k has automatically decided its flexibility. However, when k>5, the flexibility decreases. Since the model has k=78, the model would be improved.

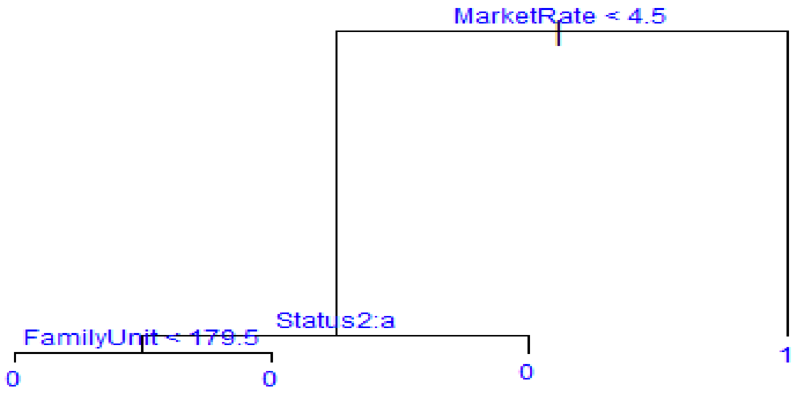


Figure 2: Classification unpruned tree from dataset 1

Third, I constructed the classification unpruned tree. From the classification unpruned tree, there are 0 and 1 at the bottom of the nodes. It indicates that the failing does not provide affordable housing (1=Yes, 0=No). If Market Rate is higher than 4.5, it indicates the failing. On the other hand, if Market Rate is lower than 4.5, it leads the choice to the left node. To sum up, start at the top of the tree and lead it to the left node if it meets the conditions, otherwise to the right node.

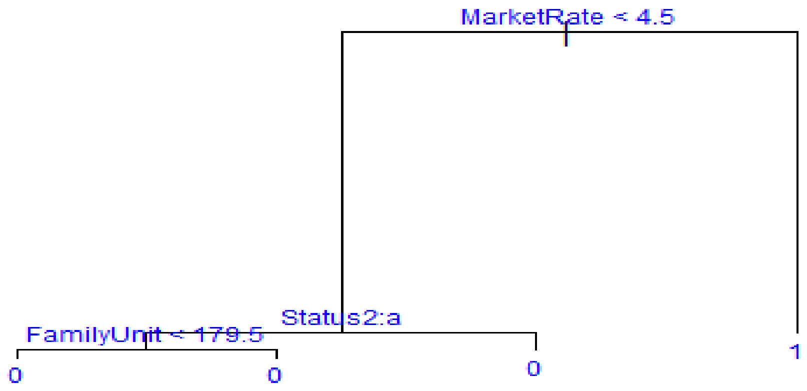


Figure 3: Classification pruned tree from dataset 1

Fourth, I constructed classification pruned three. When creating a tree, it sometimes has too many splits. A lot of splits are so flexible that data can be overfitting. The classification tree uses the Gini index. It is used to measure the purity of nodes and prune them. In this case, unpruned tree is already simple, I am able to see that pruned tree is no different from unpruned tree.

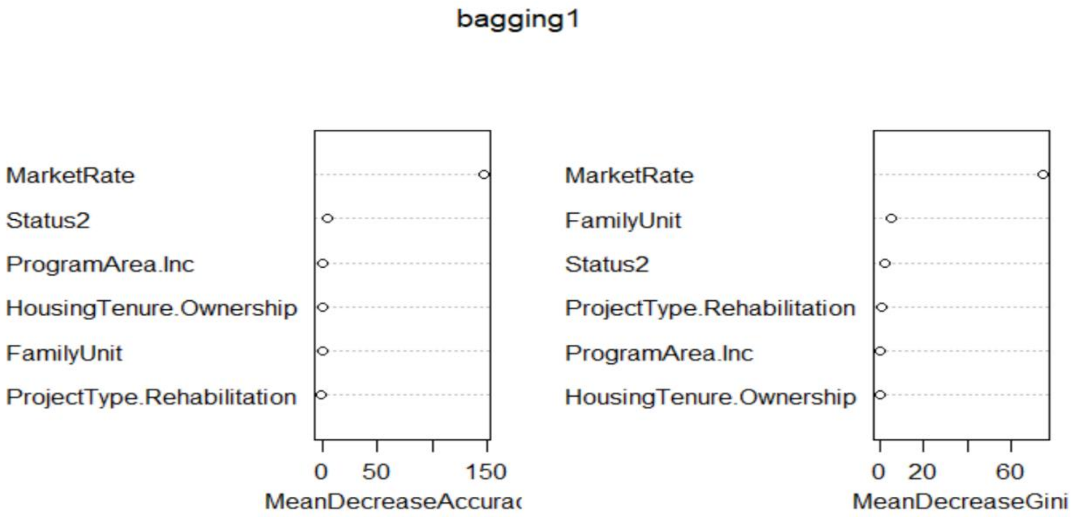


Figure 4 : Variable importance of bagging from dataset 1

Fifth, I performed bagging. Baggig is a tree based approach which allows to reduce the variance of decision trees by means of the Bootstrap. (Handout 9 Tree based methods, p.7) The advantage of using Bagging is that it avoids the risk of overfitting. Also, bagging provides which predictors are important or not. In the left plot, the importance of predictor variables was determined by the mean decrease accuracy. In the right plot, on the other hand, the importance of predictor variables was determined by the mean decrease Gini index. The plot on the left shows that it is important in order of market rate, Status2, Program Area.Inc, Housing Tenure.Ownership, Family Unit and Project Type.Rehabilitation. However, the plot on the right shows that it is important in order of market rate, Family Unit, Status2, Project Type.Rehabilitation, Program Area.Inc and Housing Tenure.Ownership.

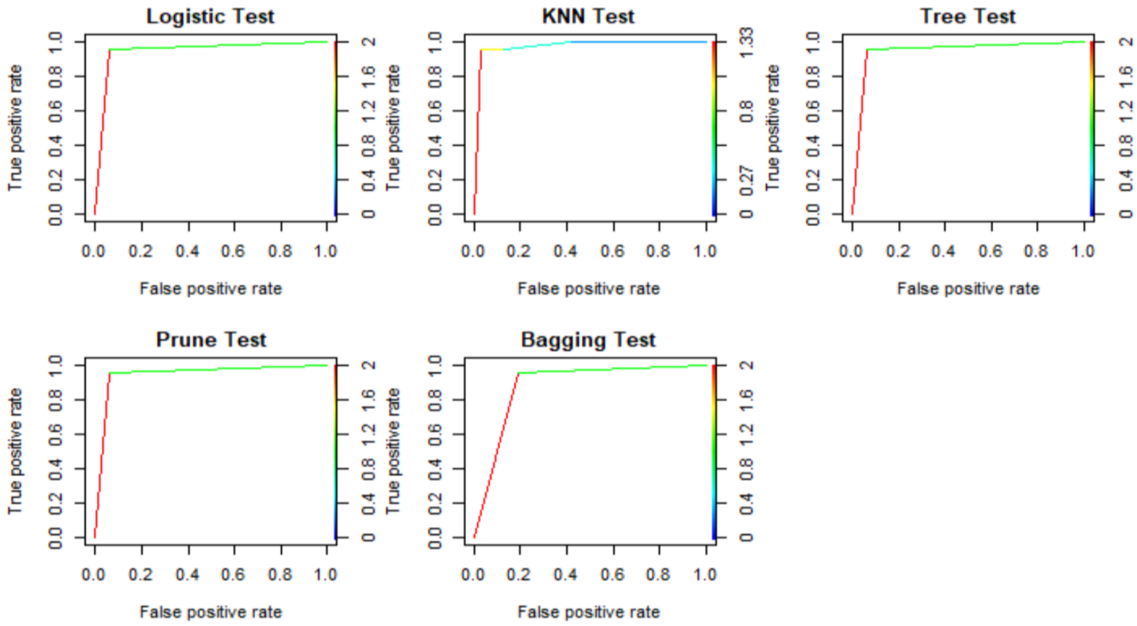


Figure 5: Roc curve of all methods from dataset 1

Finally, I formed ROC curves and recognized AUC to compare the models I have formed. Typically, the rules we want to select are based on domain knowledge, and if not available, we can use to the ROC curve. The Roc curve shows how sensitivity and specificity vary for all possible thresholds and the overall performance of the classifier is summarized by the *Area Under the* (ROC) *curve* (AUC). (Handout 8 Classification, p.10) When ROC curve is close to 1 and AUC is large enough, it is likely to perform well.

Table 2: Result of AUC and test error rate from dataset 1

Method	AUC	Test Error Rate
Binomial Logistic Regression	0.9464653	0.05128205
KNN	0.9728895	0.03846154
Unpruned Tree	0.9464653	0.05128205
Pruned Tree	0.9464653	0.05128205
Bagging	0.8819492	0.1025641

From the analysis result, KNN had the lowest test error rate and the highest AUC. However, since k was 78 in KNN, the model would be improved. Binomial logistic regression, unpruned tree, pruned tree had the second smallest test error rate and the second highest AUC. Since Pearson-x^2 indicated that p-value of binomial logistic regression model was small enough, however, binomial logistic regression model was rejected. Therefore, unpruned tree and pruned tree was the best model in dataset 1.

Dataset 2

The predictor variables were used as same as dataset 1 by computing best subsets regressions. There were 29 missing values in the response variable (Failing). In dataset 2, I created a new category *missing* when there were missing values in the categorical variable. Then, missing values in the response variable were imputed by using random regression imputation.

202 observations were randomly assigned to the training set, and 87 observations were randomly assigned to the validation set. The process of forming models was the same as dataset 1.

Table 3: Summary of binomial logistic regression from dataset 2

	Estimate	p-value
	-2.6203753	0.000319
Market Rate	0.0235473	0.119047
Family Unit	$1.612 * 10^{12}$	0.862105
Status 2	2.2808864	0.000355
Program Area.Inc	4.2684607	$2.87 * 10^{-5}$
Project Type.Rehabilitation	0.1412050	0.839342
Housing Tenure.ownership	2.3169413	0.020384
Null deviance	264.30 on 201 degrees of freedom	
Residual deviance	93.08 on 195 degrees of freedom	

From the table above, p-values of Market Rate, Family Unit and Project Type.Rehabilitation were higher than 0.05. It means that the variables are not statistically significant to predict the failing. P-values of Status 2, Program Area.Inc and Housing Tenure.ownership1 were smaller than 0.05. Therefore, these variables are statistically significant to predict the failing. Especially, ProgramArea.Inc has the highest estimates (4.2684607). It indicates that it is 11.6 times more likely to fail the future project when program area is Inclusionary. Since p-value of binomial logistic regression model was 0.005 from Pearson -x^2, this model was rejected.

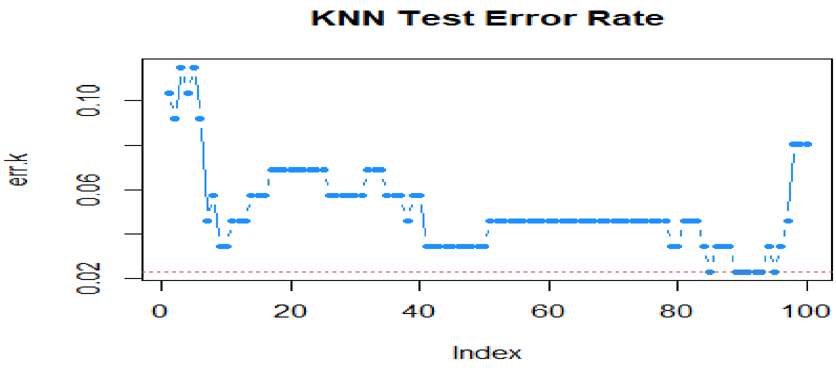


Figure 6: Graph for KNN test error rate from dataset 2

The number of k with the smallest test error rate (0.02298851) was selected. From the graph above, it indicates that k=85, 89, 90, 91, 92, 93 and 95 have the lowest test error rate. Since the model have k=85, 89, 90, 91, 92, 93 and 95, the model would be improved.

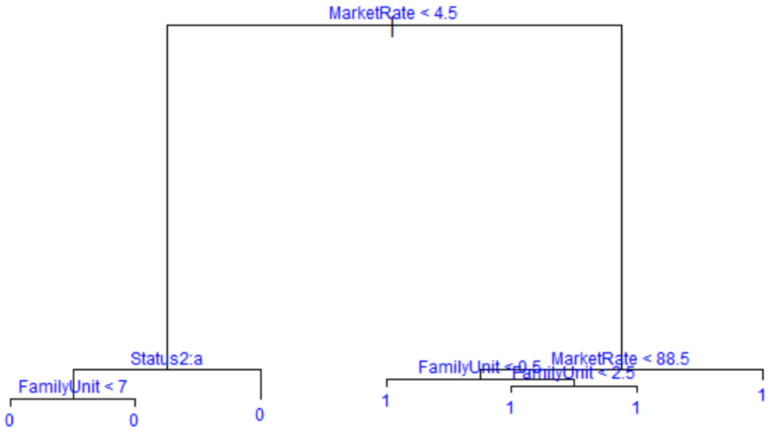
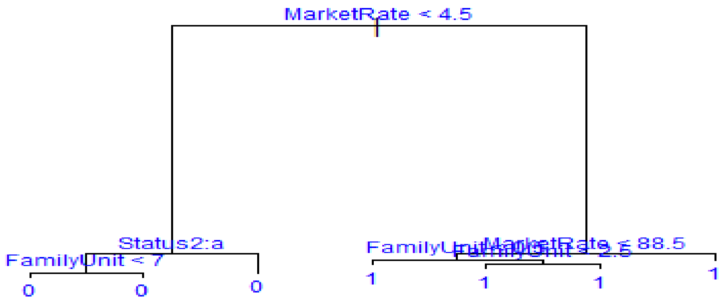


Figure 7: Classification unpruned tree from dataset 2

Compared to unpruned tree from data set 1, unpruned tree from dataset 2 had more splits. However, if the Market Rate is higher than 4.5, it indicates the failing which is same result as data set 1. If the market rate is less than 4.5, it has the same result as data set 1



.

Figure 8: Classification pruned tree from dataset 2

Classification pruned tree from dataset 2 was same as unpruned tree from dataset 2.

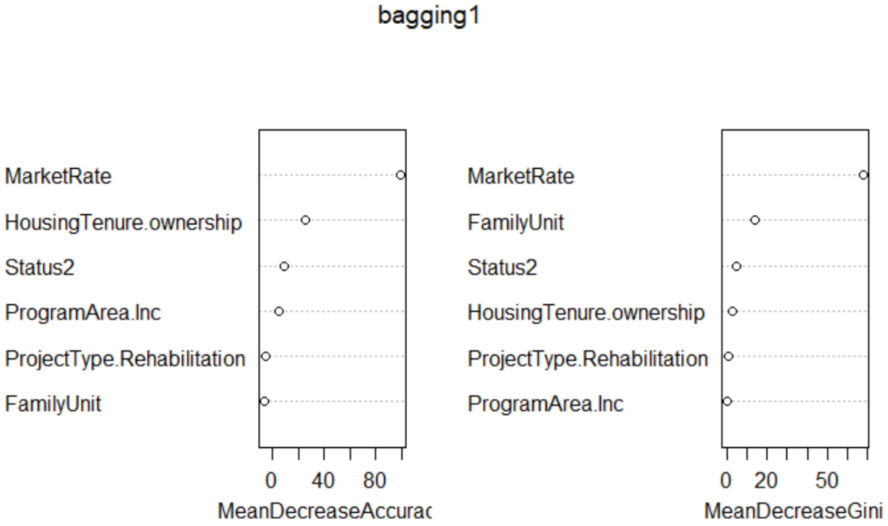


Figure 9: Variable importance of bagging from dataset 2

The plot on the left shows that it is important in order of Market Rate, Housing Tenure.Ownership, Status2, Program Area.Inc, Project Type.Rehabilitation and Family Unit. However, the plot on the right shows that it is important in order of market rate, Family Unit, Status2, Housing Tenure.Ownership, Project Type.Rehabilitation and Program Area.Inc.

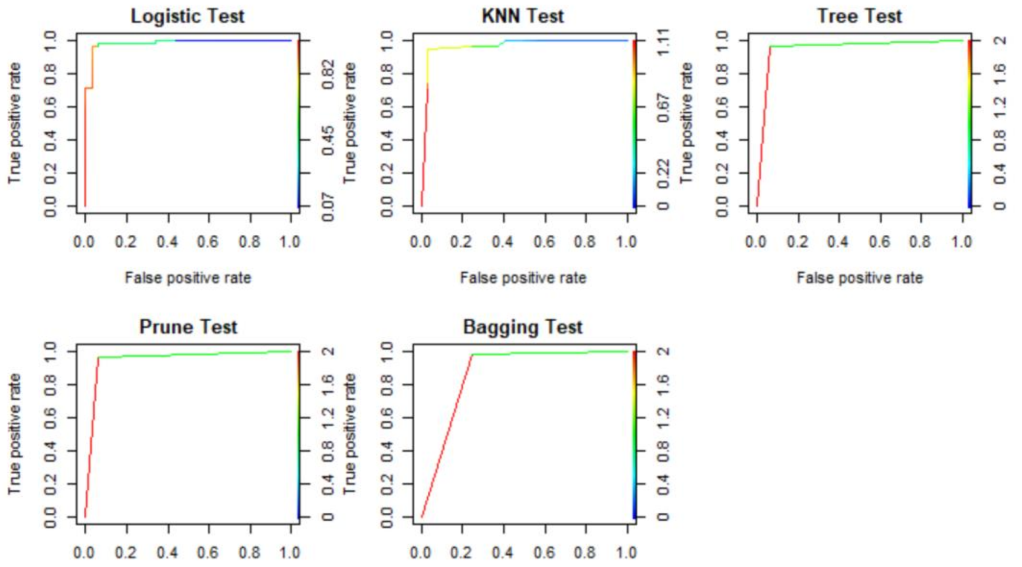


Figure 10: Roc curve of all methods

Table 4: Result of AUC and test error rate from dataset 2

Method	AUC	Test Error Rate
Binomial Logistic Regression	0.9846591	0.03448276
KNN	0.9647727	0.1149425
Unpruned Tree	0.9505682	0.04597701
Pruned Tree	0.9505682	0.04597701
Bagging	0.8659091	0.1034483

From the analysis result, binomial logistic regression had the smallest test error rate and had the highest AUC. Since p-value of binomial logistic regression model was 0.005 from Pearson-x^2, however, this model was rejected. KNN had the second largest AUC, but it had too large Ks. Unpruned tree and pruned tree had the second smallest test error rate and the third highest AUC. Therefore, unpruned tree and pruned tree was the best model in dataset 2.

When comparing models from datasets 1 and 2, the model from dataset 2 had a smaller test error rate and a larger AUC. Since dataset 2 had more 29 observations than dataset 1, dataset 2 is likely to have lower variance and higher power of test than dataset 1. (Handout 11 Missing data imputation, p.3) If too many meaningless variables are excluded, the sample size becomes excessively small. Small sample sizes cause high variance. Imputing means artificially change the distribution of variable with missings. It leads to reduce the correlation between the variable with missings and the other variables. Therefore, imputing reduces the correlation between variables and missing values and prevents the departure of variables. As a result, the model from dataset 2 had a smaller test error rate and a larger AUC.

Discussion

The most appropriate model was selected through two datasets. The binomial logistic regression model had the lowest test error rate and the highest AUC on both datasets. Through the Pearson-x^2, however, the binomial logistic regression model was rejected. As a result, the unpruned tree and pruned tree models were selected on both datasets.

The most important predictor variable in unpruned and pruned tree models was Market Rate. If the Market Rate is greater than 4.5, the project is likely to fail. On the other hand, if the Market Rate is less than 4.5, the project is unlikely to fail.

Minimizing missing values in the data will be necessary to create a more accurate, more predictable models in further study. Imputing can cover missing values, but that is not the unique value of the variable. In other words, the actual values of the variables cannot be represented. Minimizing missing values in the data will build better predictions, better models.

References

Sun, D. (2016). Computer vision technology for food quality evaluation (2nd ed., Food science and technology international series). *Amsterdam ; Boston: Elsevier/ Academic Press*.