Shulei Yang (sy8uu)
STAT 5330 – Data Mining Mini Project

1. Exploratory Data Analysis

a. Exploratory on response variable

Proportion of each class of observations:

	Percentage of 'yes'	Percentage of 'no'
Training	11.27%	88.73%
Testing	10.95%	89.05%

b. Exploratory on response of predictors in training data set.

i. First column (a column name 'X'), looks like customer index. Therefore, I won't use it for my model fitting.

ii. Continuous Variables and Categorical Variables.

There are 4 continuous variables and 16 categorical variables in our data set (exclude variable 'x').

2. Modeling and Data Analysis

a. LDA & QDA

LDA and QDA should not be used for this data set. Both LAD and QDA assumes a multivariate normal distribution for the predictors they use. However, there are too many categorical variables in our data set. Therefore, if we try to run 'lda' or 'qda' functions in R, we will receive error notice.

b. Logistic Regression and Linear SVM

	Accuracy	Sensitivity	Specificity
Logistic Regression	91.16%	31.26%	98.53%
Linear SVM	88.93%	42.35%	94.66%

c. Non-linear SVM

	Accuracy	Sensitivity	Specificity
Non-linear SVM	92.4%	45.23%	98.2%

Comparing results from linear SVM and nonlinear SVM, nonlinear SVM preforms better since it has larger accuracy, sensitivity and specificity.

d. Comparison

From the result we can see that Non-linear SVM produces the best model, since its Accuracy and Sensitivity are all the highest. LDA and QDA are no appropriate

here. Like I have discussed above, both LDA and QDA assume a multivariate normal distribution for the predictors. Therefore, they are not appropriate for a data set with a lot of categorical variables.

e. Continuous predictors

	Accuracy	Sensitivity	Specificity
LDA	91.16%	46.34%	96.67%
QDA	88.13%	62.31%	91.3%
Logistic Regression	91.19%	30.82%	98.61%
Linear SVM	91.04%	32.15%	98.28%

For LDA and QDA, when we build the model with all the variables, we will receive error notice in r. As I discussed above, this is because LDA and QDA are not appropriate for data set contains too much categorical variables. Therefore, when we us only continuous predictors, we successfully built the LDA and QDA models.

For both logistic regression and linear SVM, when we built the model with only continuous predictors, we find accuracy and specificity increase for both models, but sensitivity decreases for both models.

f. Random Prediction

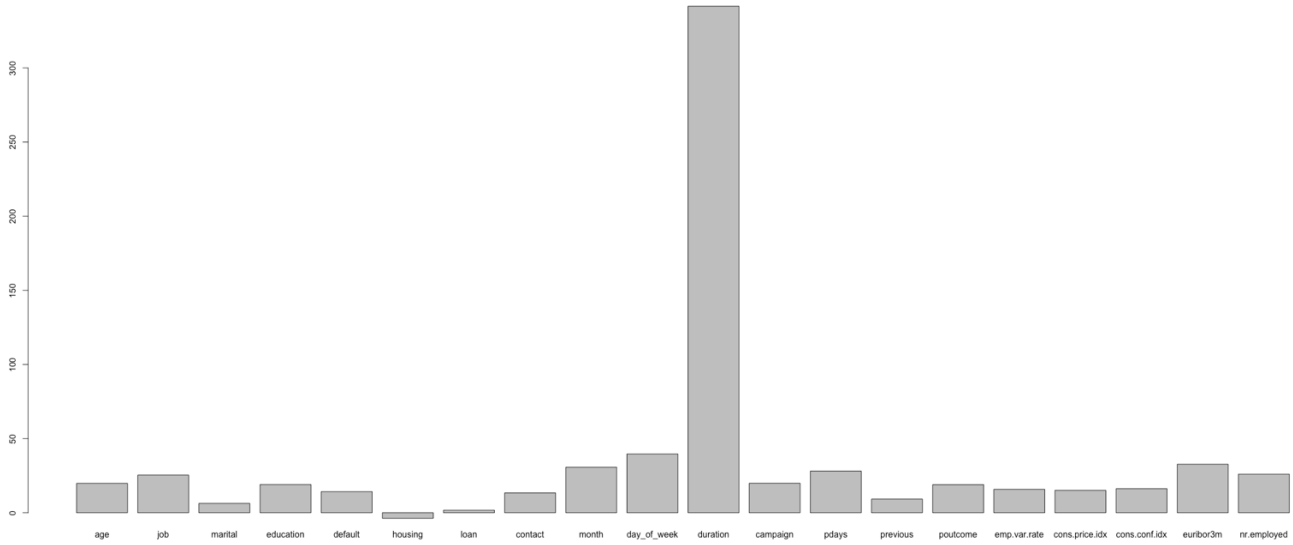
	Accuracy	Sensitivity	Specificity
Random Prediction	50.11%	48.56%	50.3%

g. Random forest

	Accuracy	Sensitivity	Specificity
Random Forest	97.91%	84.04%	99.62%

h. Barplot of MeanDecreaseGini

From the barplot we can see there are some variables like loan, housing and marital which only reduce a little impurity. Therefore I think we can build a reduced model with fewer number of predictors.



i. Boosting Model with Exponential Loss

	Accuracy	Sensitivity	Specificity
Boosting	87.42%	50.11%	92.01%