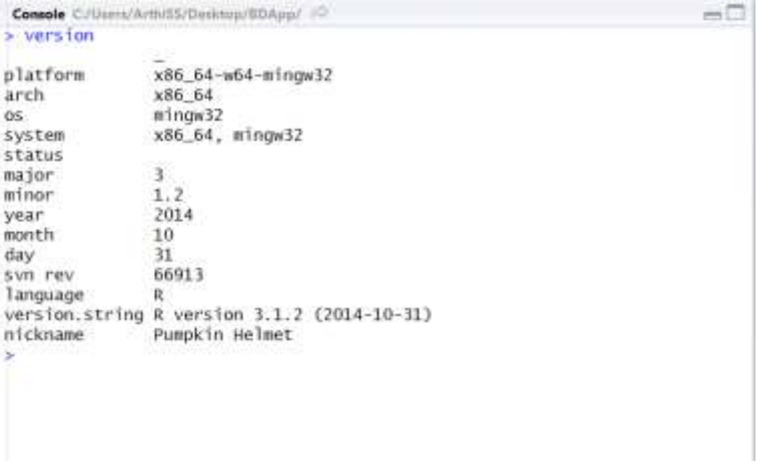
Prediction of whether a Customer would get a new credit card

Introduction:

In today's world, credit cards are the one of the most used payment methods. To customers, credit cards offer a way to conserve cash in hand and a period of time to buy goods without a psychological barrier of over-purchasing goods. To banks, the credit cards providers, credit cards offers profits provided the credit card holders pay their bills in timely fashion. According to an article published in the Forbes magazine, the credit card holders who pay their bills mostly on time but who don't pay in full are the most valuable customers. Therefore, credit card providers are always on a search for valuable customers to whom they can advertise their credit cards and whom they can convince to buy their cards. In order to make their search for these valuable customers easier, theses credit card providers collect a variety of information about their customers. This is a marketing strategy used by most credit card providers. The objective of this project is to predict whether a customer would get a new credit card based on the information collected about him/her.

Project deliverables: 1. R Code 2. Visualizations that would explain relation between the features and target variable 3. Evaluation of classifiers using accuracy and specificity as evaluation.

Software used and setup: For this project, R software version 3.1.2 released on 31st October 2014. In order to check the version of R present in the system, the "version" command can be used. The following snapshot shows the version command results for the machine used to run the code:



```
> version
               _
platform       x86_64-w64-mingw32
arch           x86_64
os             mingw32
system         x86_64, mingw32
status
major          3
minor          1.2
year           2014
month          10
day            31
svn rev        66913
language       R
version.string R version 3.1.2 (2014-10-31)
nickname       Pumpkin Helmet
>
```

Snapshot of results for version command

For this classification, a number of libraries have been used. The following table consists of the library used and the purpose of the library for this specific project. Installation of packages required before calling the library. Use install.package("library name")

Table 1: Libraries used and purpose of each library

Library	Purpose
corrplot	To plot the correlation between the features.
mlbench, caret, class, randomforest	To evaluate the importance of each feature given the predict credit card purchase.
MASS	Use stepAIC to generate top 10 models according to AIC criterion
Deducer	Plot ROC plot for logistic regression model
C50	Decision Trees
e1071	Svm and naïve bayes
ggplot2	Visualizations of feature interactions

Reproducibility of the results: The seed is set to '456' so that the same instances can be taken as train and test each time the program is run. This enables to have reproducible results.

Data: For this prediction problem, a dataset with 5000 records and 12 features would be used. The following information of 5000 customers would be used as features for this classification: (1) Age of the customer (2) Experience (3) Income level (4) ZIP code (5) Family size (6) CCAverage (average expenditure on credit cards per month (7) Education (8) Mortgage (9) Personal Loan (10) Securities.Account (11) CD Account (certificate of deposition) (12) Online (online facilities available or not).

Methods, Preprocessing and Evaluation metrics:

In this project, four methods have been used: (1) Logistic Regression, (2) Decision Trees (3) SVM (4) Naïve Bayes

(1) **Logistic Regression**: This model was chosen because it is known as good binary classification model. This method is similar to linear regression and is linearly biased. But the difference is that the log odds of the outcome are modeled as a linear combination of the attributes used. After a model was defined, the best model was selected based on AIC, where the least AIC value suggests the best model.

Using this trained model, the test data and evaluation data were tested and confusion matrices were built. Two error metrics, accuracy, specificity were used. The true positives are defined as true credit card non-purchasers.

Apart from error metrics, both diagnostic plots and ROC curve were generated in order to assess the model performance.

(2) **Decision Tree:** Since the problem at hand is a classic classification problem where certain conditions leads to either of the binary outcome, decision trees was an excellent choice for modelling. As expected by this method, rules were used to classify the data. The method used was C5.0 since pruning is included and over fit is prevented. This method uses information gain index to classify the data. The same error metrics as used on logistic regression were used for this method as well. A disadvantage of this method was that a plot of the tree classify the data isn't provided as given in CART.

(3) Naïve Bayes: Naïve is another linear classifier which is usually a good selection for binary classification. The difference between Naïve Bayes classifier and logistic regression classifier is that the weights for features are set according to the features correlation with the target variable's classes. In logistic regression, all weights are the same.

(4) SVM: SVM is classifier which uses only the points (known as support vectors) near the margin that divides the two classes. SVM can be a linear as well as non-linear classifier. Examples of non-linear SVM are Radial basis function and Gaussian classifiers. For the purpose of the classification problem, linear SVM would be used.

Preprocessing steps:

There were few problems with the data such as missing values, incorrect range of values for the attributes etc. Apart from solving these issues, the data was checked for class imbalance. As suspected, there was class imbalance which made it evident that accuracy would not be a good evaluation metric. The metric of interest for this classification problem would be specificity (the number of customers who are predicted to purchase credit card and who are actually credit-card purchasers). In order to select the relevant features, Pearson's correlation between the features was calculated. The plot below was generated using corrplot package. As shown in figure 1, dark blue color denotes highly correlated features (>0.75). Experience and Age were found to be highly correlated. Using the caret package, the feature importance for all features given to predict credit card purchase was generated. Since age was more importance than experience, experience was removed from the data set.

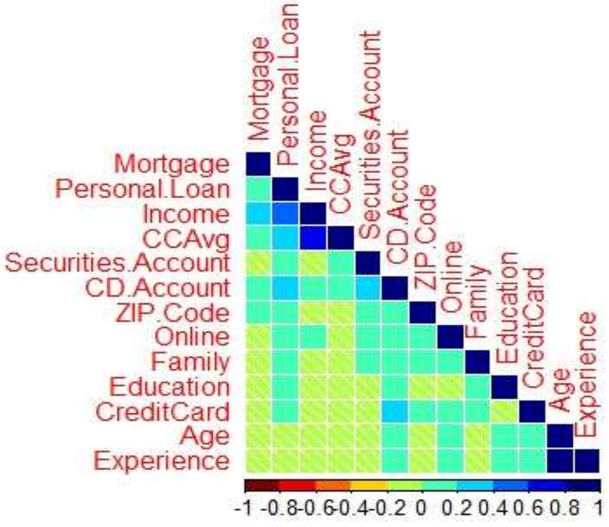


Figure 1: Correlation plot for features correlations

Finally, the data was split into 70% train and 30% test. Classifiers logistic regression and decision trees were run on a subset of features that were selected using AIC. The AIC criterion was used to generate top 10 subsets of features in terms of low information loss and model complexity. AIC criterion is usually used for model evaluation. Out of the top 10 models, the least AIC signifies the best model to use. The least AIC value obtained was “3876.07”. The features used for these two classifiers were Mortgage, Personal.Loan, Securities.Account, CD.Account and Online.

Results:

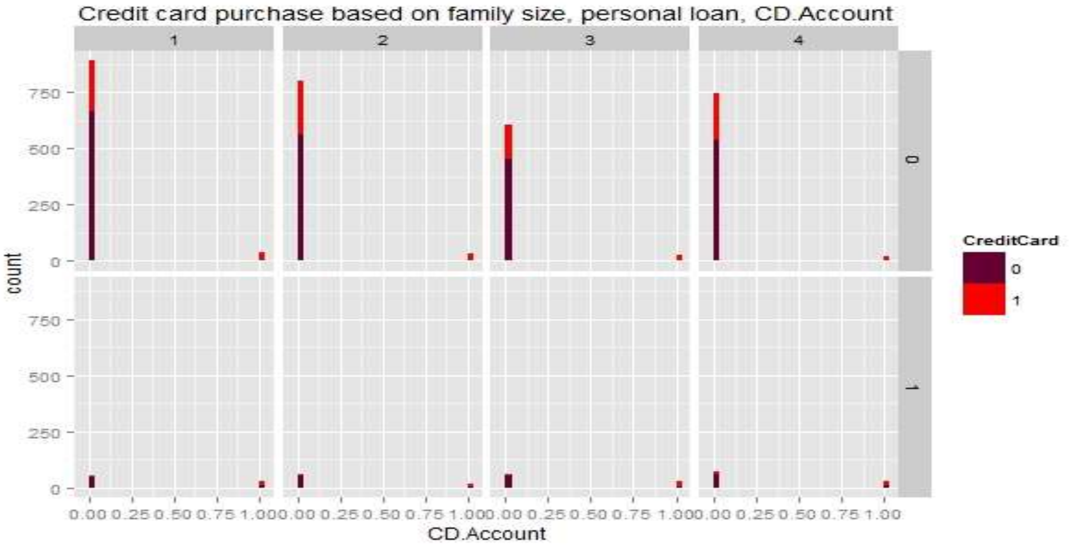


Figure 2: Credit Card purchase based on family size, personal loan and CD Account

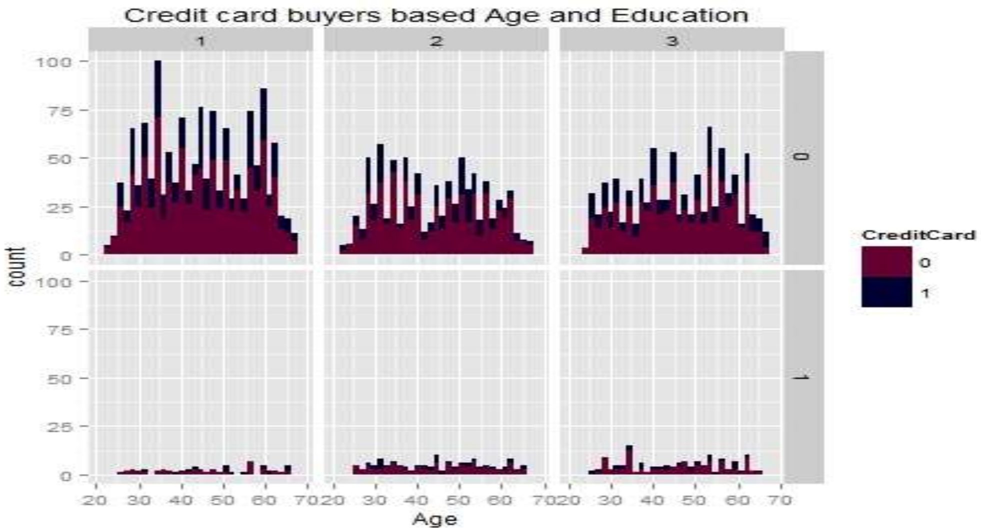


Figure 3: Credit Card purchase based on Age and Education

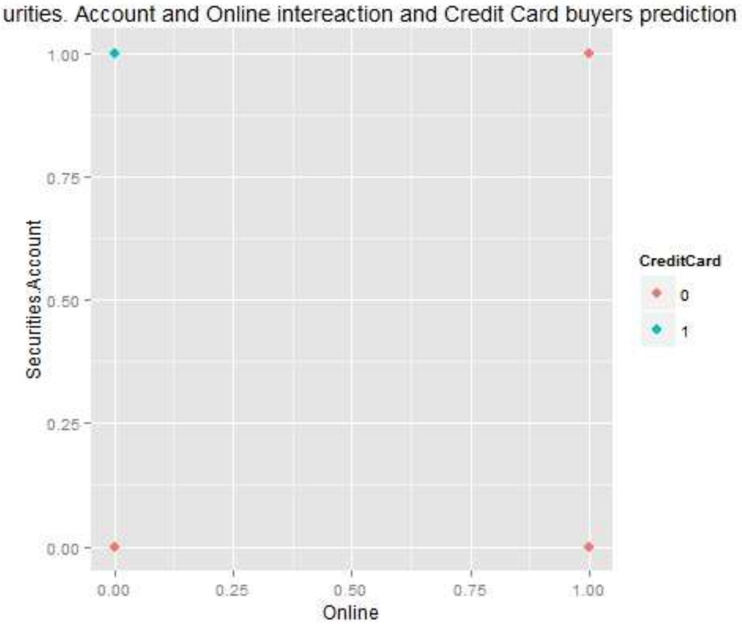


Figure 4: Securities.Account and Online facilities interaction for credit card purchase

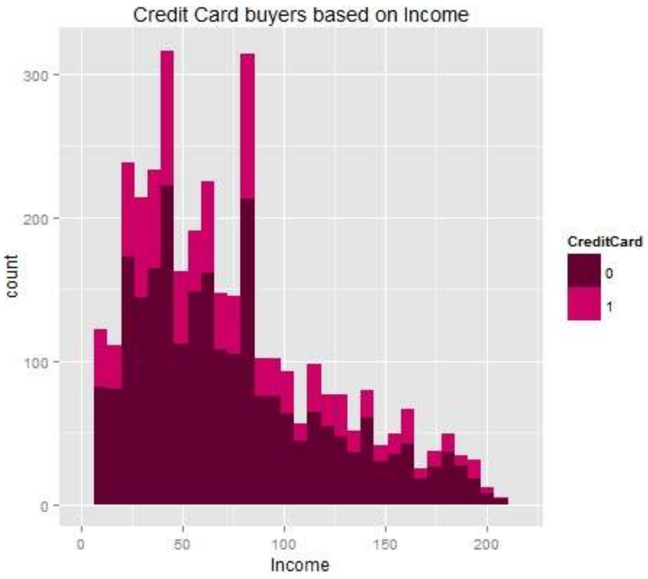


Figure 5: Credit Card purchase based on Income

Table 2: Classifiers performance evaluation based on accuracy and specificity

Classifier	Specificity	Accuracy
Logistic Regression	14.47%	73.46%
Decision Trees	14.47%	73.46%
SVM	88.40%	74.06%
Naïve Bayes	76.19%	73.46%

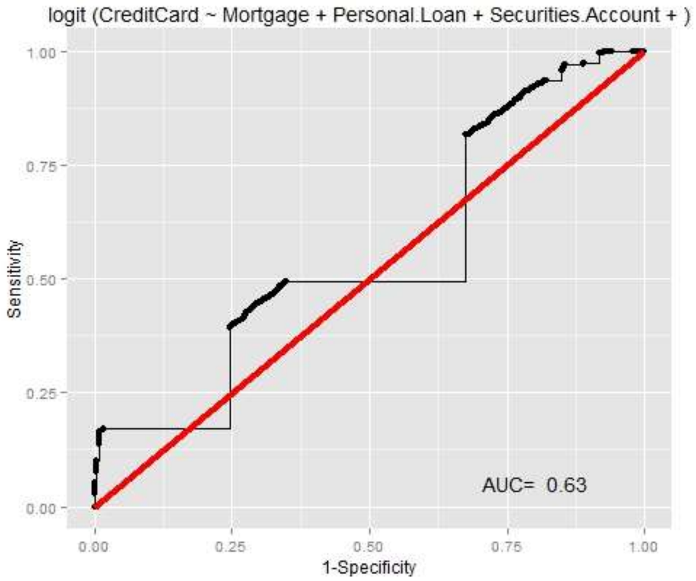


Figure 6: AUC under ROC for Logistic Regression

Discussion:

Before discussing the classifier performance, it is important to appreciate the interaction amongst the features. As shown in figure 2, there is more number of customers who do not have a personal loan. Given that there is class imbalance, it is also evident the more number of people who wouldn't purchase a credit card than who would purchase. But from the figure, it can be inferred that if a customer is willing is purchase a credit card, his/her family size is more likely to be 1 than 4. Similarly, a customer with a certificate of deposit (CD.Account = 1) and no personal loan will be likely to purchase a credit card. As shown in figure 3, there is more number of customers in age group of 30-40 and of education level 1 who will be likely to buy credit card. From figure 4, it is evident that a customer with securities account and with no online facilities will purchase a credit card. In all other combinations of securities account and online facilities, the customer will not purchase the credit card. From the last figure, it is evident that the number of customers with income greater than 80-90 range is less. The figure shows that there is more number of customers willing to purchase credit card than unwilling to purchase credit card irrespective of income. Nevertheless, using the caret package, income was found to be an important variable in predicting credit card purchase.

In of terms of classifier performance results, SVM performed the best. It was able to predict 88.40% of customers who are actually willing to purchase credit card as willing to purchase credit card. The remaining 11.6 % were falsely predicted as unwilling to purchase credit card. Logistic Regression and Decision trees did not do well probably because only a subset of features was used to run. The caret package evaluated Income and Zip code as two most important features. But these two classifiers used the subset of features suggested by AIC. This could be a possible reason for the low specificity performance of these classifiers. However, the accuracy performance of these classifiers is on par with the accuracies of using naïve bayes and SVM. This indicates that the logistic regression and decision tree must have overfit such that the true negative instances were overlooked and the classifiers were biased towards positive class. This inference is also supported by the ROC curve for logistic regression. The curve at few points passes through the diagonal of the plot which suggests that at these instances, the classifier predicted the instances inaccurately. The flat vertical lines which are part of the curve suggest that there were more number of false positives (1- specificity) than true positives (sensitivity). For Naïve Bayes, Laplace smoothing was used to decrease the overfit of the model. But the parameter did not significantly help even when set at 10. In future, the classifiers can be run on multiple models suggested by AIC criterion. While in the given list of four classifiers, SVM performs the best, boosting and bagging techniques can probably increase the specificity. Since bagging also suffers due to class imbalance, oversampling the data can help reduce the effects of imbalance. Another alternative to oversampling would be try the prediction problem as an unsupervised problem and perform k-means clustering for it. Since unsupervised learning classifications cannot be evaluated by metrics such as accuracy, recall etc., comparing the class labels with the prediction of the k-means clustering can lead heuristic evaluation of specificity for the classifier.

Conclusion:

Therefore, the best classifier is SVM and the most importance variables are zip code and income. For decision trees, the CD.Account feature was used completely. But zip code and income were not part of dataset used for decision tree classifier and hence a comparison cannot be made.