Credit Card Approval Prediction

Rajeev Kanth Reddy
rmeda
rmeda@indiana.edu

Siddhartha Pagadala
sidpagad
sidpagad@umail.iu.edu

1. INTRODUCTION

In today's world, credit card approval is an important task in banking sector. This project aims at building an accurate and efficient credit card approval model, which is a crucial task in financial institutions.

2. Data

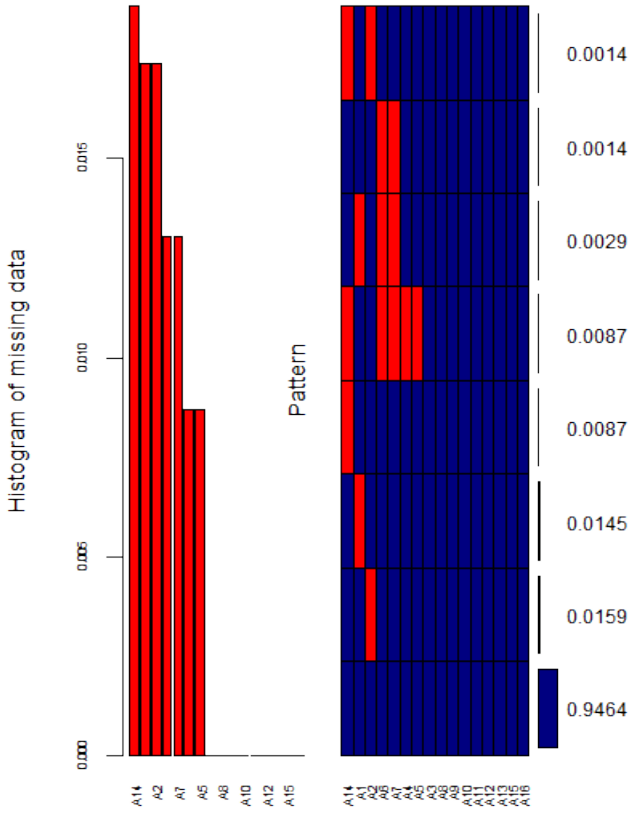
2.1 Description

Dataset for the project is obtained from http://archive.ics.uci.edu/ml/datasets/Credit+Approval. The dataset contains 15 predictor variables and a response variable. Being sensitive information, all attribute names and values have been changed to meaningless symbols to protect confidential data. The dataset contained mix of continuous, nominal with small number of values, nominal with large number of values which made the analysis challenging. The dataset contains 690 observations with 307(44.5%) being credit card approved and 383(55.5%) being credit card denied.

Attribute	Type	Values
A1	Nominal	a, b
A2	Continuous	
A3	Continuous	
A4	Nominal	l, t, u, v
A5	Nominal	g, p, gg
A6	Nominal	c, d, e, i, j, k, m, q, r, w, x, aa, cc, ff
A7	Nominal	h, j, n, o, v, z, bb, dd, ff
A8	Continuous	
A9	Nominal	f, t
A10	Nominal	f, t
A11	Continuous	
A12	Nominal	f, t
A13	Nominal	g, p, s
A14	Continuous	
A15	Continuous	
A16 (Response)	Nominal	+, -

2.2 Data Imputation

The dataset contains 37 (5%) missing values. The pattern of missing values in various predictor variables is shown in below figure:



The above pattern shows that around 95% of data are complete cases. The above plot shows proportion of missing values of each variable as well as combinations using color red.

The following techniques are performed to handle missing values:

- Deletion: Delete observation if it contains missing value in any one of the predictors.
- Iterative multiple imputation: Random regression imputation is executed on all missing predictors until their values converge.

Attribute	# of missing values (out of 690)
A14	13
A1	12
A2	12
A7	9
A6	9
A5	6
A4	6

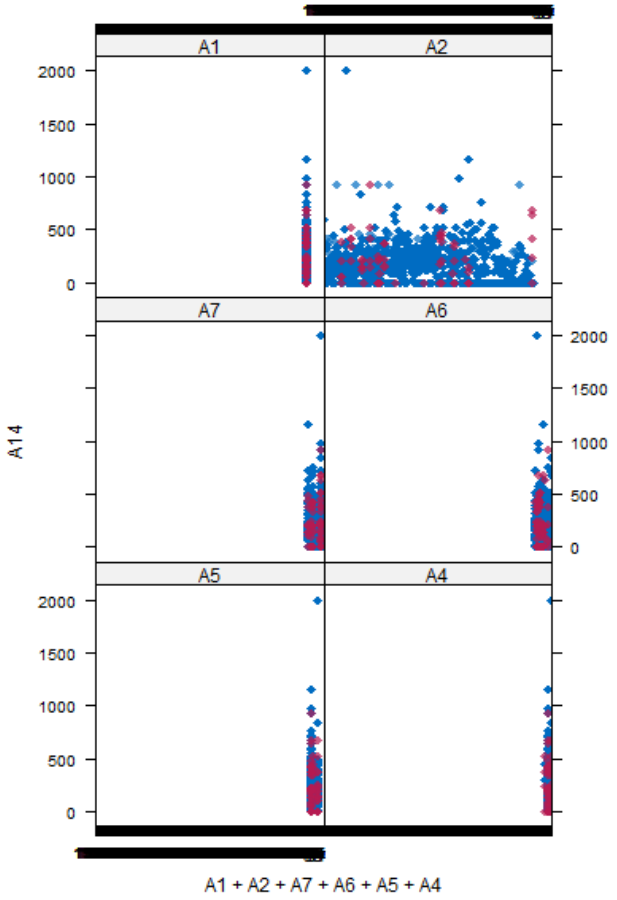


Figure 2: Distribution of original and imputed data

The shape of the magenta points (imputed) matches the shape of the blue ones (observed). The matching shape tells us that the imputed values are indeed “plausible values”.

2.3 Data Visualization:

Figure 3 shows the scatter plot of numerical predictor variables. From the scatter plot no pattern can be observed among the pairs of variables, hence no correlation among them. It suggests that all the predictor variables are significant in prediction of response variable.

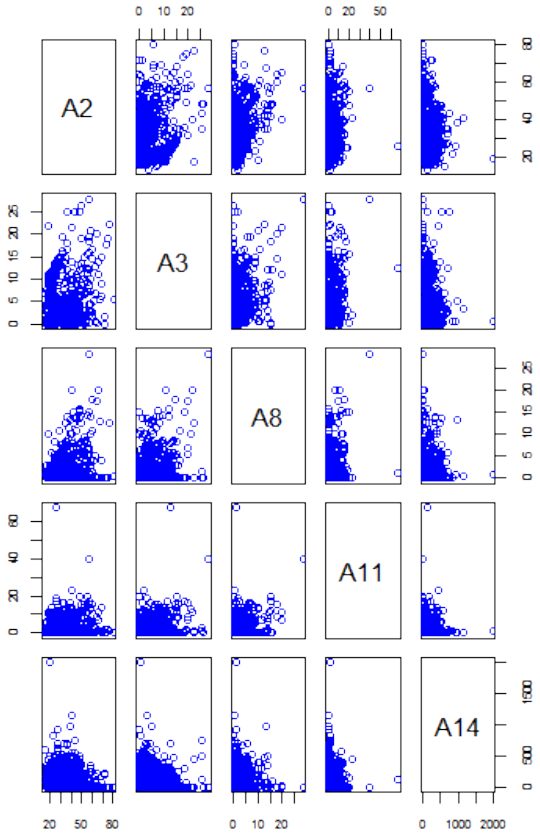
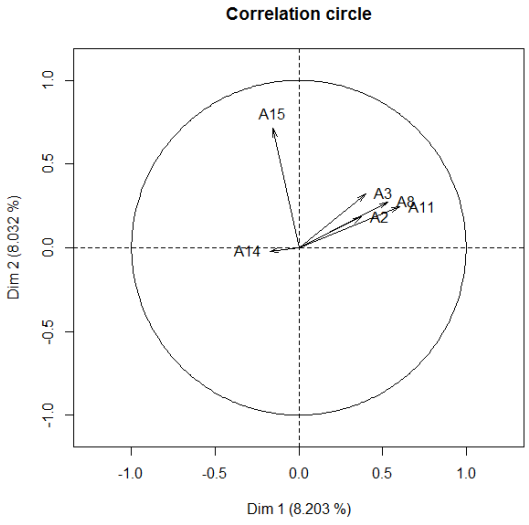


Figure 3: Scatter Plot

2.4 Principal Component Analysis



The above plot shows there are no dominant directions among any variables which shows no significant correlation among any variables.

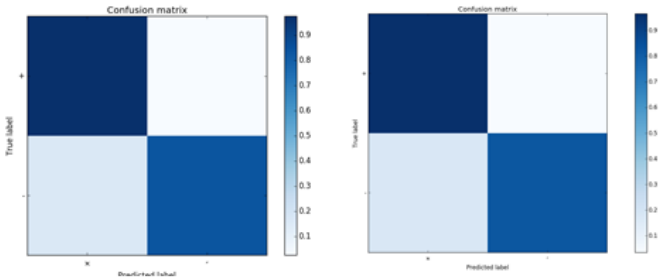
3. Modelling and Results

The following models are developed on the data, with and without missing values, and its performance is evaluated on both test and training data. Training and test data are being split in ratio of 80:20.

- Logistic Regression: Since response variable contains two classes, Approved (+) or Denied (-), probabilities are calculated using logistic regression and appropriate classes are assigned, '+' if probability >= 0.5 else '-'.
- Support Vector Machine: Support Vector Machine is applied on the dataset with various kernel functions such as linear, radial and sigmoid.

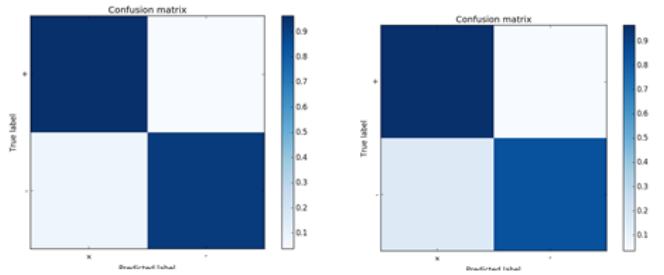
Model	Data with deleted missing values		Data with imputed missing values	
	Test error	Train error	Test error	Train error
Logistic Regression	0.1837	0.1006	0.0869	0.1195
SVM(linear)	0.1531	0.1247	0.0797	0.1413
SVM(radial)	0.1531	0.1247	0.0797	0.1504
SVM(sigmoid)	0.1479	0.1291	0.0797	0.1504

3.1 Confusion matrix:



logistic with missing values deleted

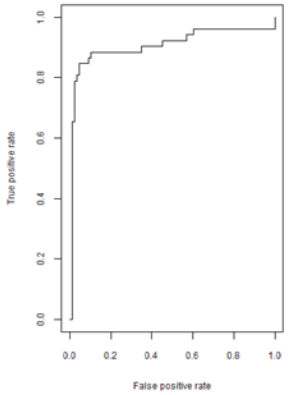
logistic missing values imputed



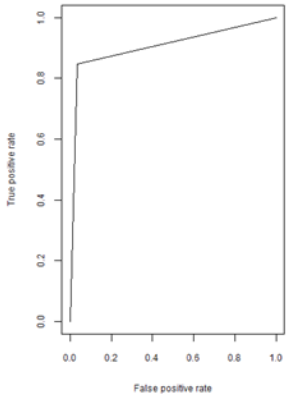
svm with missing values deleted

svm missing values imputed

3.2 ROC (Receiver Operating Characteristic) performance metric:



ROC curve for Logistic



ROC for SVM

The AUC calculated from the above ROC is around 0.96 which is close to 1 for both Logistic Regression and Support Vector Machine models, which suggest that the predictive ability of the models is good.

4. Conclusion:

The analysis of performance of various implemented models show that all the models performed better after imputing the missing data than data with deleted missing values. Of all the models Support Vector Machine showed better results with accuracy of 92 %.

5. REFERENCES:

[1] Quinlan. UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Credit+Approval

[2] Richard J. Bolton and David J. Hand, "Unsupervised Profiling Methods for Fraud Detection", Technical Report (Department of Mathematics, Imperial College, London), 2002.