Acquisition **A**nalytics for CredX

Submitted By: Anshul Roy

Business Objective

CredX wants to determine the factors affecting credit risk. Thereafter create strategies to mitigate the risk and assess the financial benefit of the risk model.

Demographic Data

- Age
- Gender
- Marital Status
- Dependents
- Income
- Education
- Profession
- Residence Type

+

Credit Bureau Data

- Outstanding Balance
- DPDs
- Loan Inquiries
- Open Home Loan
- Credit Card Utilization
- No. of Trades

Methodology	CRISP-DM Framework
Task	Predict Performance / Credit Default
Data set	Demographic & Credit Bureau
Features	As per data set. Data Dictionary available
Models	Supervised Classification Models like Naïve Bayes, Logistic Regression, Decision Tree & Random Forest

Problem Solving Methodology

Data Preparation

1. Merge Demographic and Credit Bureau Data on Applicant ID
2. Remove Duplicate Rows
3. Explore data by Univariate and Multivariate Analysis
4. Check NAs and NANs
5. Outlier Treatment

IV Analysis and WOE

6. Compute Information Value
7. Populate features with their WOE values
8. Observe variability using Principal Component Analysis
9. Combining IV Analysis With Variable Clustering
10. Create data frames with and without WOE Values

Model Building

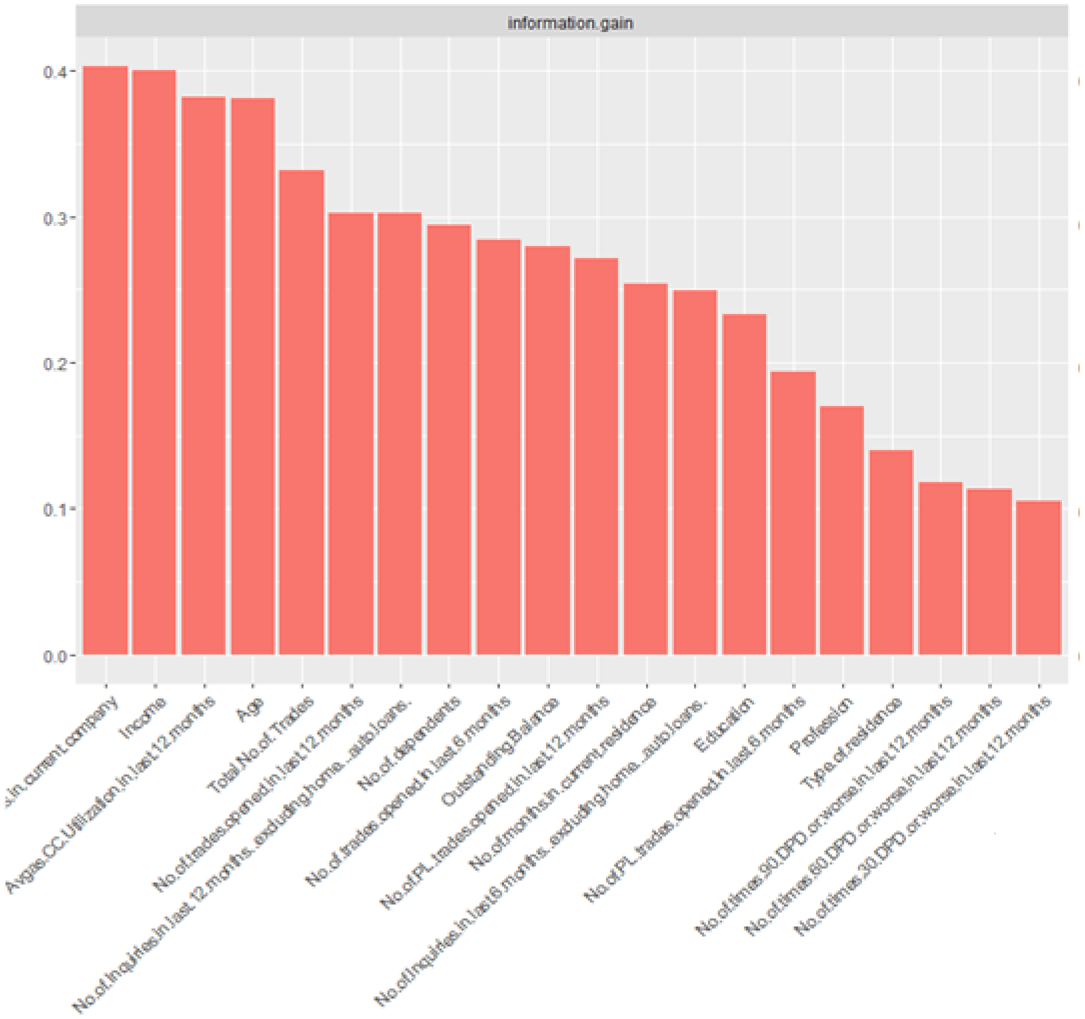
11. Split data into train and test
12. Use SMOTE for balancing data
13. Build models on Demographic Data and on All data separately
14. Obtain Performance Tag for 1425 rows having NAs in Performance Tag and merge with all data file
15. Rebuild Models using Logistic Regression, DT, RF, NB

Metrics Evaluation

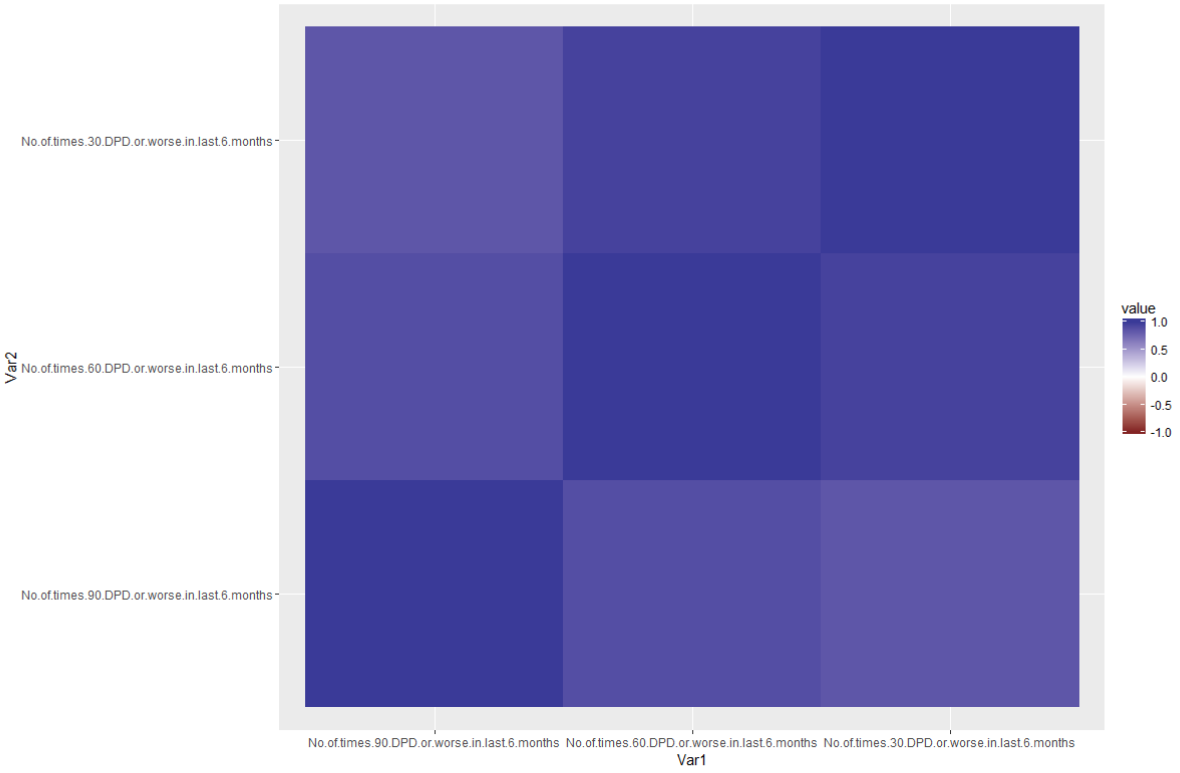
16. Check Model on Test Data
17. Evaluate Model Metrics
18. Build Application Score Card from Logistic Regression Model
19. Build Financial Strategies using the optimum model

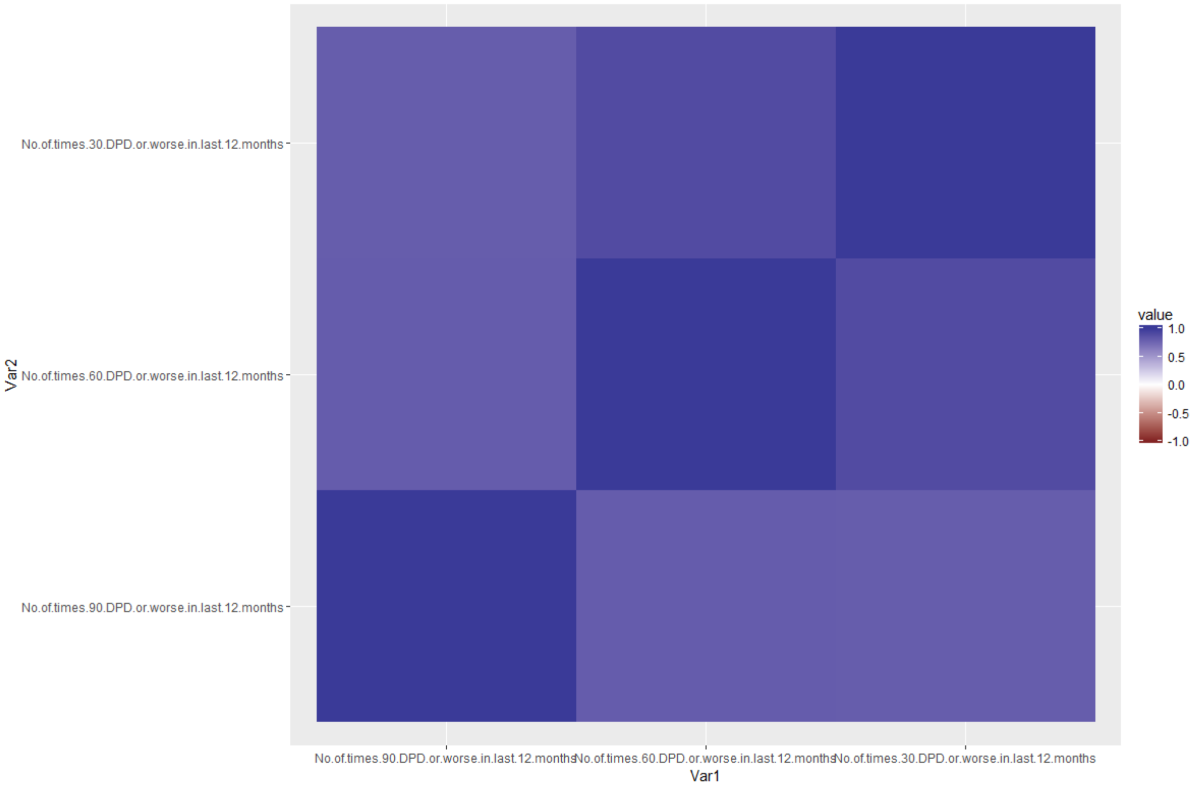
Top Variables with most Information Value

Variable	IV
Avgas.CC.Utilization.in.last.12.months	0.31709560595
No.of.PL.trades.opened.in.last.12.months	0.29589547357
No.of.Inquiries.in.last.12.months..excluding.home...aut...	0.29498356705
No.of.trades.opened.in.last.12.months	0.29203196749
Outstanding.Balance	0.24629513857
No.of.times.30.DPD.or.worse.in.last.6.months	0.24156273923
Total.No.of.Trades	0.23117467072
No.of.PL.trades.opened.in.last.6.months	0.21970498073
No.of.times.90.DPD.or.worse.in.last.12.months	0.21387483771
No.of.times.60.DPD.or.worse.in.last.6.months	0.20583387648
No.of.Inquiries.in.last.6.months..excluding.home...auto...	0.20518701285

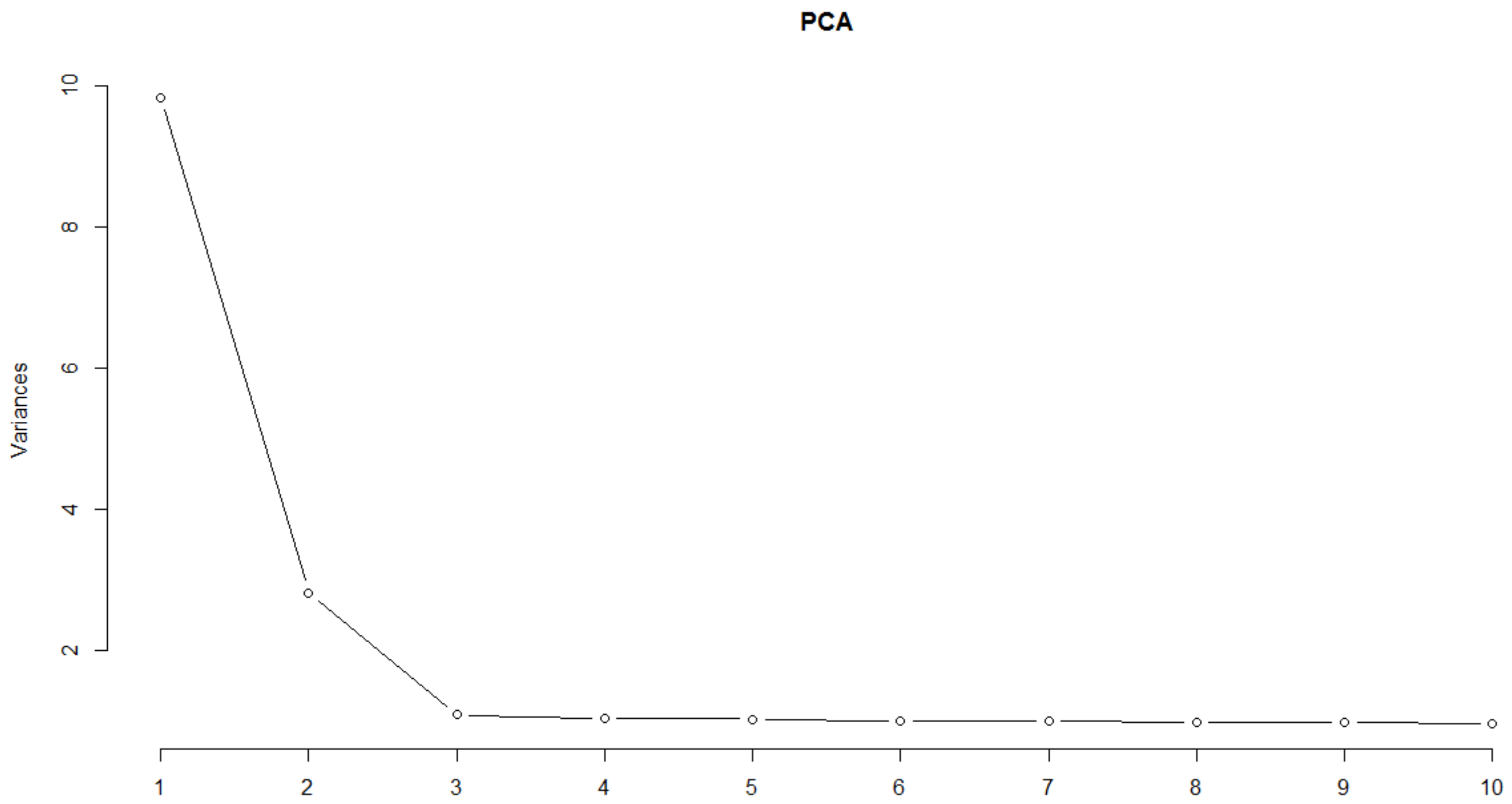


Multicorrelation among Features





Principal Component Analysis



The Data set has highly multicollinear features. Over 95% variability in the data can be explained from top 8 variables

Naïve Bayes model insights and results

- Assumption: Features are independent given the class. Ensure that the correct ratio of class label is maintained both in train & test data set.
- Data is balanced using SMOTE. Use 10-fold Cross-Validation technique.
- The sensitivity and AUC of the model improved but the overall accuracy of the model decreased (as compared with KNN).

Model Metrics	Values (Numeric)
Overall Accuracy	0.7151
Sensitivity	0.97040
Specificity	0.07217

Decision Tree model insights and results

- Use 3-fold Cross-Validation technique.
- MinSplit = 37, MinBucket = 30 and CP = 0.001 obtained as best parameters.
- The sensitivity and overall accuracy of the model increased (as compared with NB)

Model Metrics	Values (Numeric)
Overall Accuracy	0.919
Sensitivity	0.95910
Specificity	0.05889

Random Forest model insights and results

- Use 3-fold Cross-Validation technique with 50 iterations.
- Data is balanced using SMOTE. Use 10-fold Cross-Validation technique.
- The sensitivity and AUC of the model improved but the overall accuracy of the model decreased (as compared with KNN).

Model Metrics	Values (Numeric)
Overall Accuracy	0.956
Sensitivity	0.958423
Specificity	0.086207

Logistic Regression model insights and results

- C-statistic for both train and test data were found close to 0.6, which shows the model has good proportion of concordant pairs.
- KS-statistic for both train and test data lies at the first decile => model can distinguish between the binary classes.

Model Metrics	Values (Numeric)
Overall Accuracy	0.9581564
Sensitivity	0.001144165
Specificity	0.9998008

Application Score Card and Financial Advantage

- ❑ We have PDO = 20, Base Score=400 & odds = 10
- ❑ Score = Offset + { Factor* log(Odds) }

  where Offset = 400 - (28.8539*log(10)) = 333.5614

  and Factor = = 20/log(2) = 28.8539

  and log(odds) = log(odds(good)) = log(probability(0)/probability(1))
- ❑ Threshold Score is 260, below which we will not suggest to acquire the customer.
- ❑ Our model provides good discriminatory power over pre-identifying risky customers.
- ❑ With the Acquisition Model, we have set the base application score. This will help business to avoid acquiring customers who have high probability (over 91%) of defaulting.
- ❑ We have successfully identified top 8 features among the 28 given features. Data collection strategies for these features should have quality check and control.
- ❑ Our model developed have 90% more accuracy (95% Overall) than a model developed at random with the available features.