CREDX – Credit Loss Analysis

Finding the right customers

PGDDA Capstone Project
March 2018
IIIT-B

Vamsi Kiran Telikicherla
Suman Kumar Sinha
Talveen Singh Rakhra
Shounak Kundu

BUSINESS OBJECTIVE AND STRATEGY

Business Objective
- Formulate strategy to combat the increased loss of credit due to customer defaults in recent years.
- Possibly mitigate credit risk by acquiring the right set of customers.

Strategy
- Analyse, understand and clean data in-order to prepare a model capable of predicting the right set of customers

UNDERSTANDING CUSTOMER AND CREDIT DATA

Files
- Customer demographics
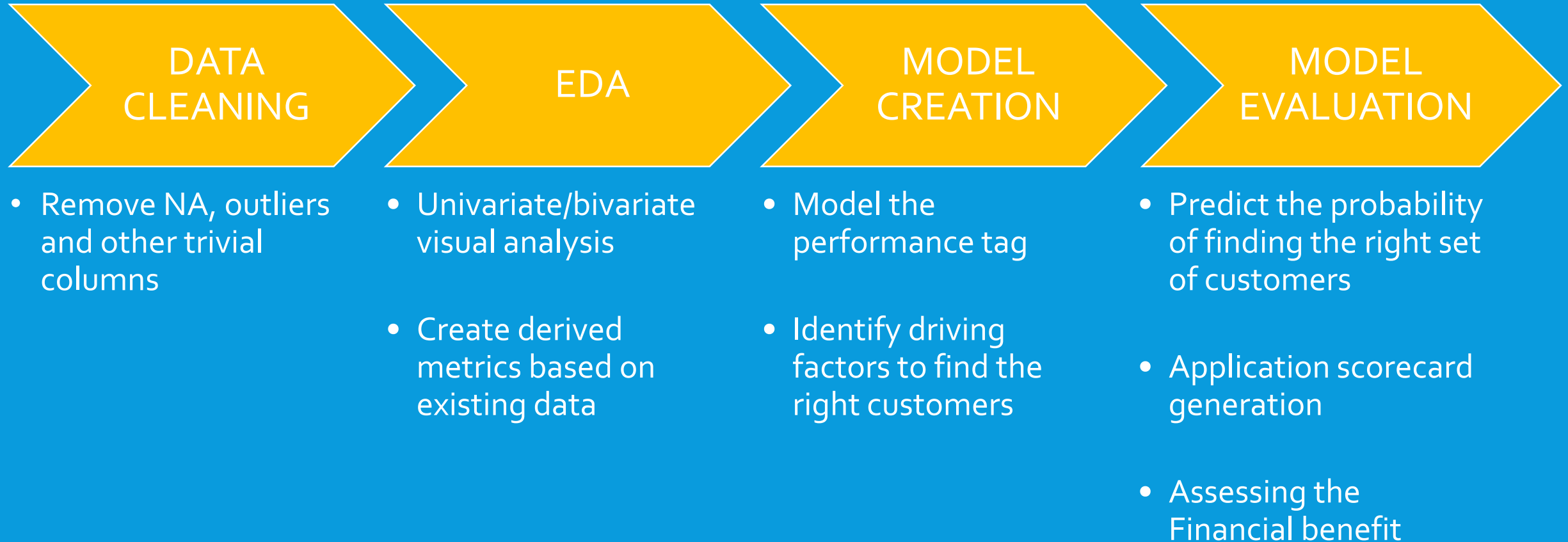- Customer credit bureau data

Nature of data
- 71295 records
- 71289 unique customers
- 29 total Attributes

Points of interest
- Overall Customer Age < 65 years
- Salaried employees >50%
- Male to Female Ratio ~ 3:1
- Performance Tag
 - Indicates whether the customer has defaulted or not

ANALYSIS STRATEGY

DATA CLEANING

- Remove NA, outliers and other trivial columns

EDA

- Univariate/bivariate visual analysis

- Create derived metrics based on existing data

MODEL CREATION

- Model the performance tag

- Identify driving factors to find the right customers

MODEL EVALUATION

- Predict the probability of finding the right set of customers

- Application scorecard generation

- Assessing the Financial benefit

EDA : UNDERSTANDING KEY DATA

Application ID: 3 duplicate IDs

- 765011468
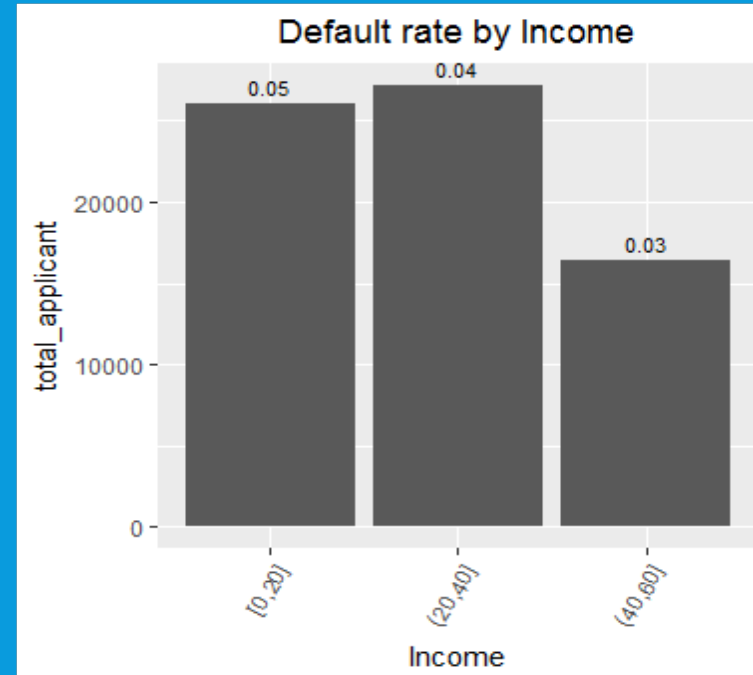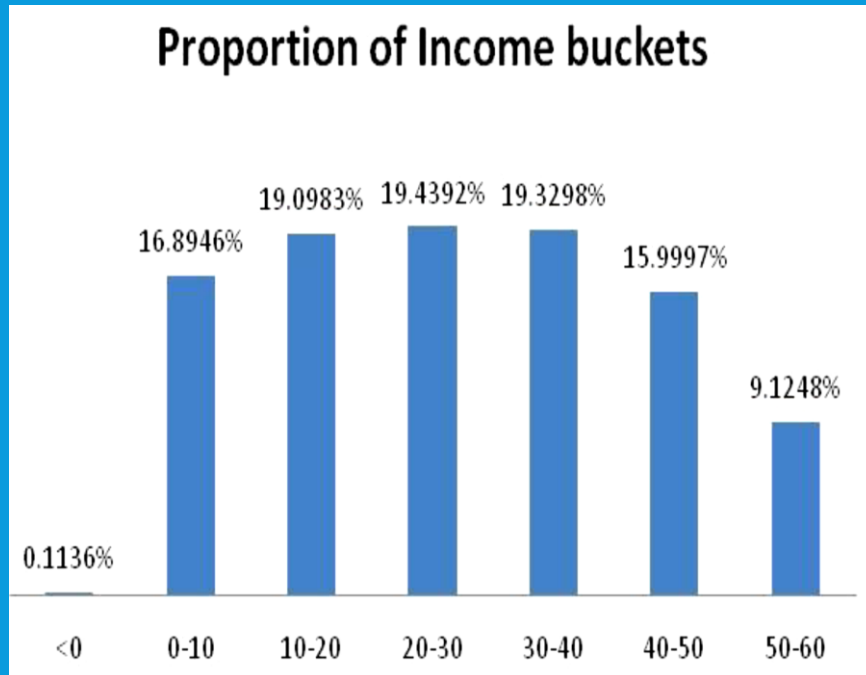- 653287861
- 671989187

Performance Tag:

- Non Defaulters ~67K applicants (94%)
- Defaulters ~3K applicants (4%)
- Rejected = 1425 (2%) applicants
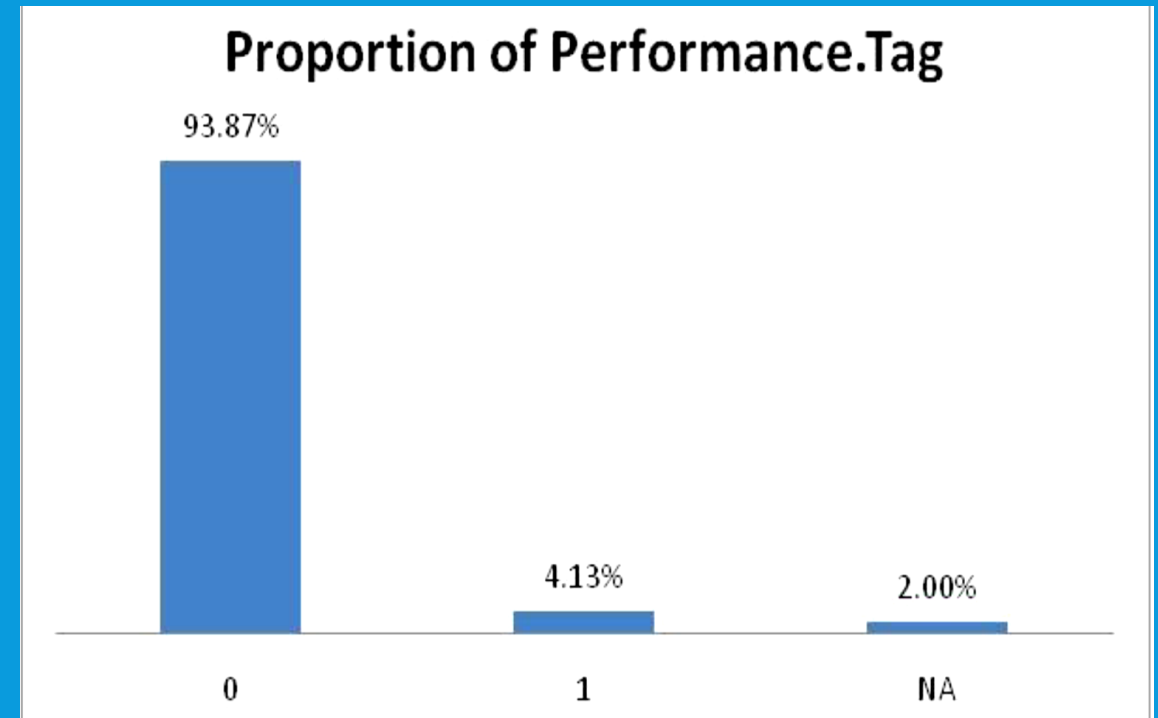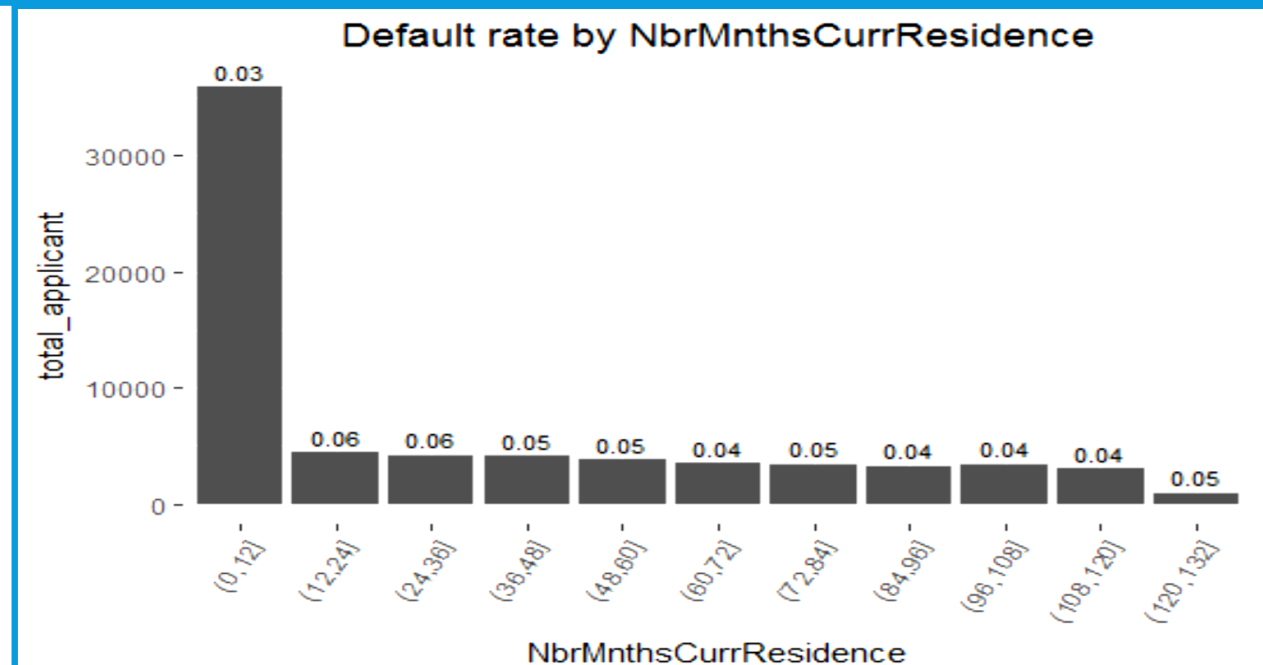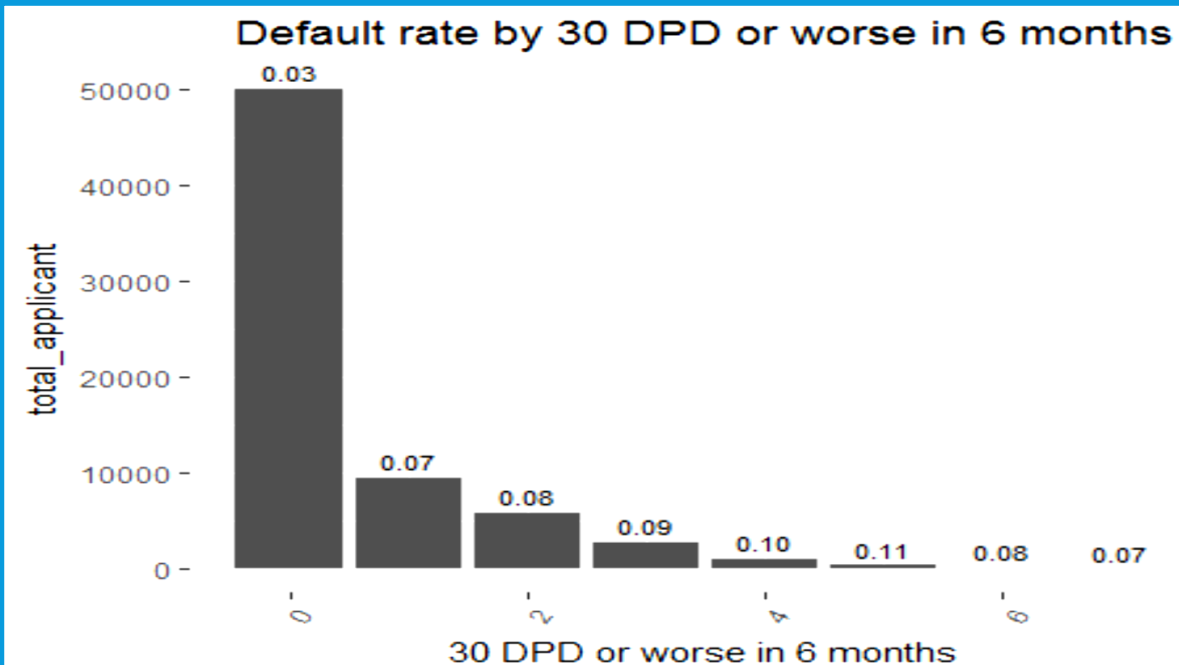
EDA : UNDERSTANDING KEY DATA

Income:

- 81 incorrect values (Negative)
- Notable inverse trend between income and default rate

EDA : UNDERSTANDING KEY DATA

**No of times 30 DPD or worse in last 6 months:**
- No "NA" or blank values
- Trend: Customers who have crossed 30DPD once or more in last 6 months tend to default more

Number of months in current residence:
- No blanks or incorrect values
- Trend: Customers who stay one or more year at the same residence tend to default more
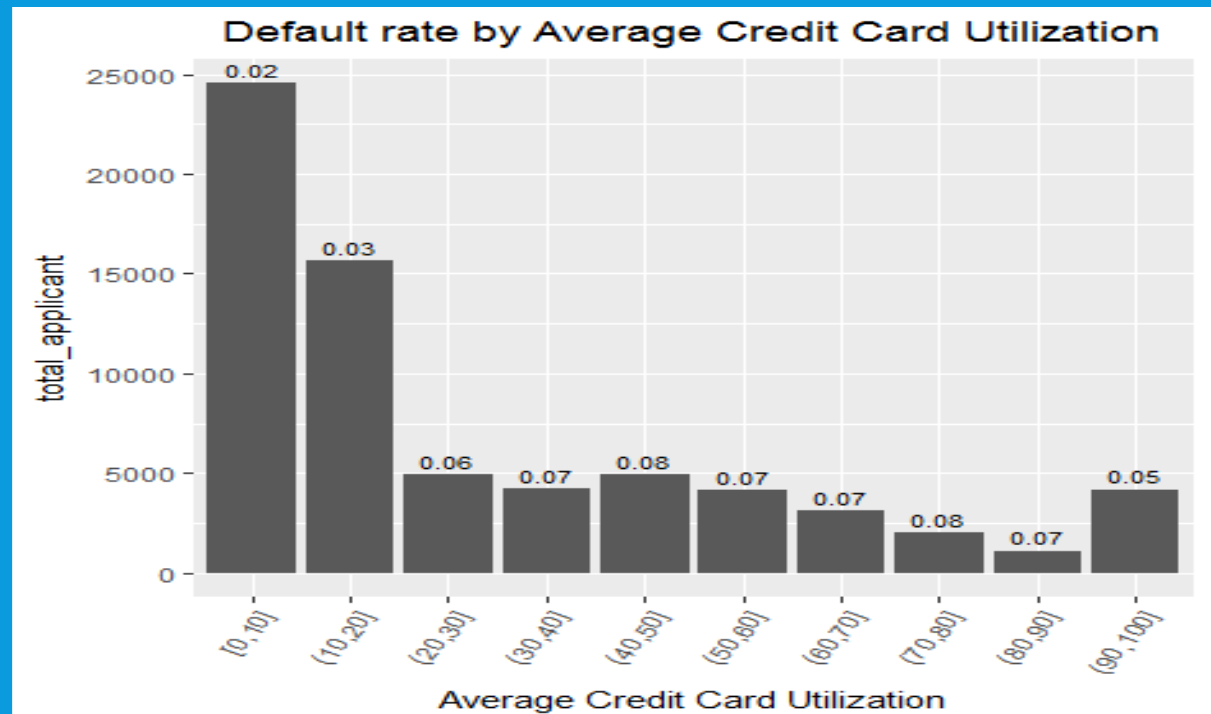
EDA : UNDERSTANDING KEY DATA

No of PL trades opened in last 12 months:
- Applicants with 1-7 PL trades opened in last 12 months tend to default more
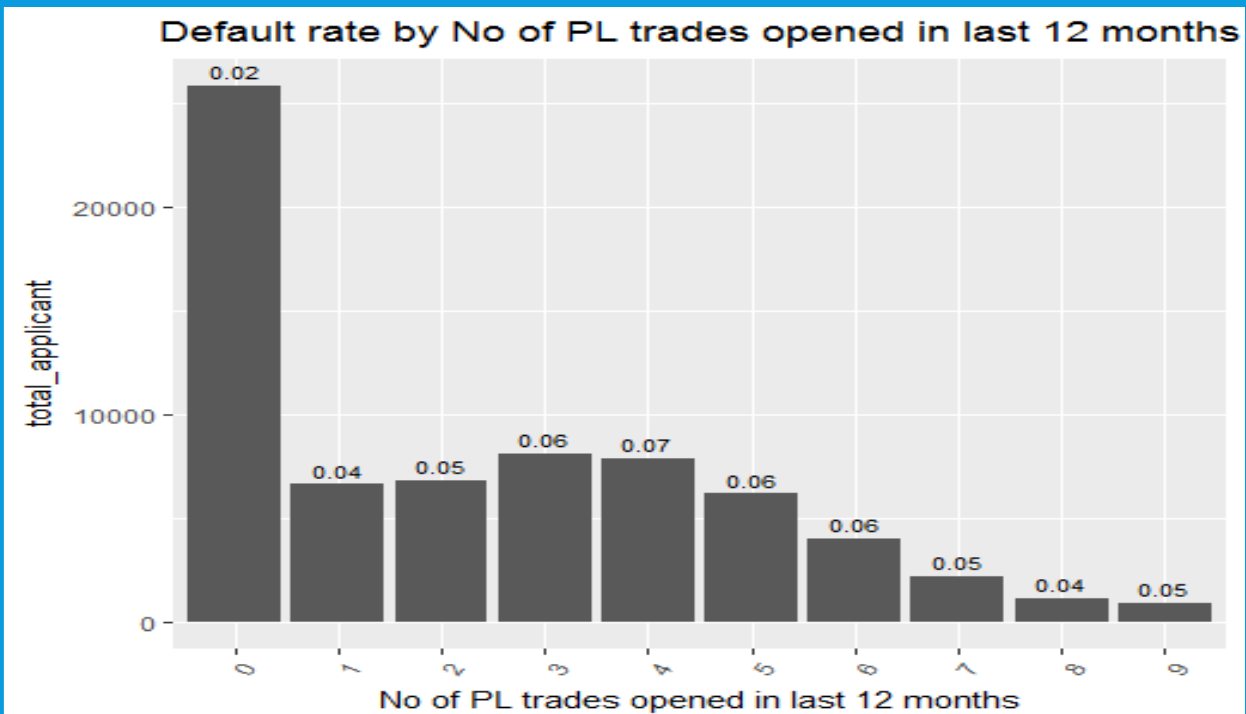
Average CC Utilization in last 12 months:
- 1058 first time CC user
- Trend: Defaulters tends to use more Credit card

Default rate by No of PL trades opened in last 12 months

Default rate by Average Credit Card Utilization

EDA : UNDERSTANDING KEY DATA

Number of months in current company:
- No "NA" or blank values
- Trend: Customers who stay less than 2 years in current company tend to default more.

Default rate by NbrMnthsCurrCompany
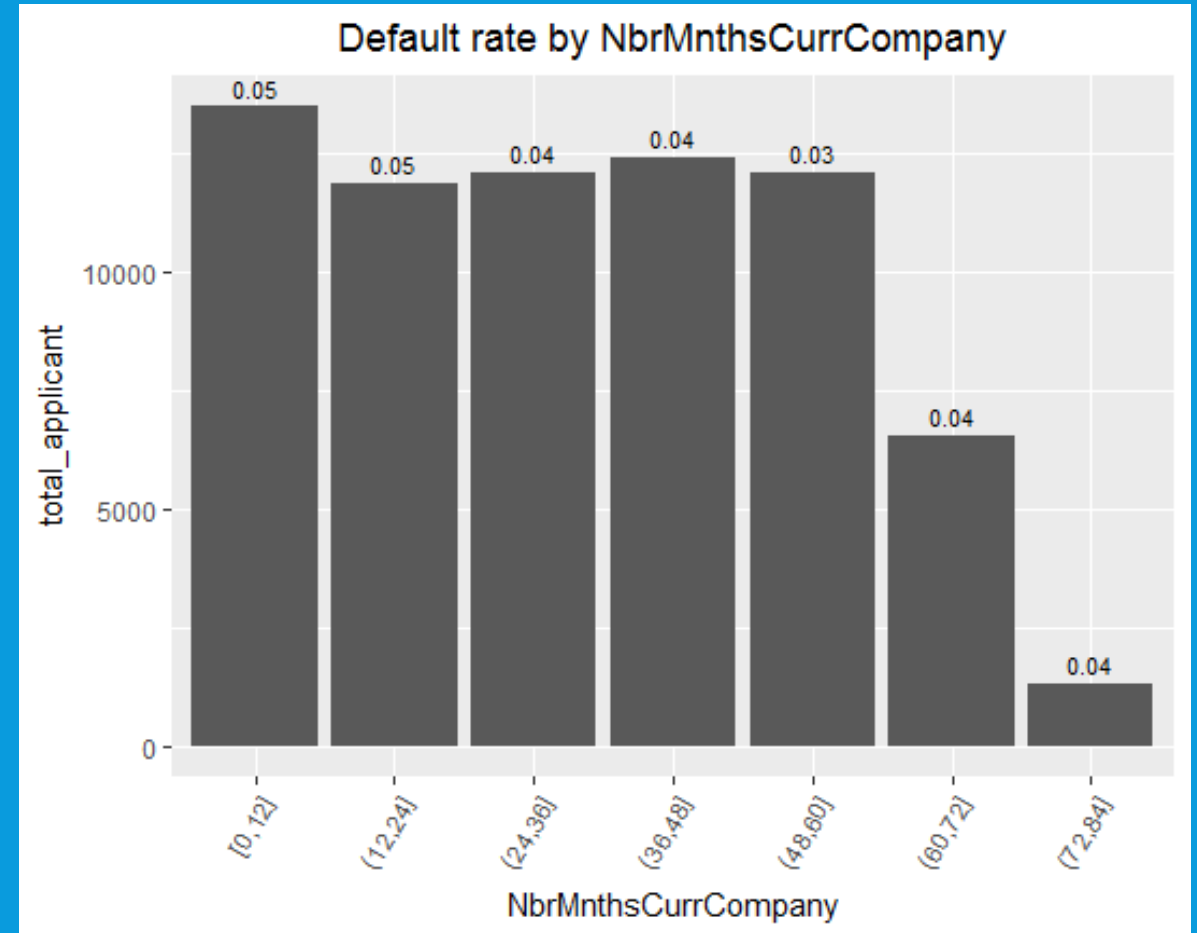
DATA PREPARATION

IV AND WOE:
- IV helps in identifying important predictors and reduce dimensions
- WOE denotes the effect of individual variable on the target
- Found 18 variables as significant (Refer table)

SMOTE:
- Data is balanced using SMOTE to avoid making a biased model

$$IV = \sum (DistributionGood_i - DistributionBad_i) \times WOE_i$$

$$Weight\,of\,Evidence = ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

| Columns	IV
binning.AvgCCUtilLast12Mnths	0.315294
NbrOfPLTradesOpenLast12Mnths	0.315274
NbrOf30DPDLast6Mnths	0.24877
NbrOfPLTradesOpenLast6Mnths	0.232858
NbrOf30DPDLast12Mnths	0.221618
NbrOf90DPDLast12Mnths	0.218707
NbrOf60DPDLast6Mnths	0.213864
binning.TotalNbrOfTrades	0.20797
binning.OutstandingBal	0.198815
NbrOfTradesOpenLast6Mnths	0.197141
NbrOf60DPDLast12Mnths	0.191041
binning.NbrOfTradesOpenLast12Mnths	0.190322
binning.NbrOfInq_ExclHomeAutoLoan_Last12Mnths	0.175896
NbrOf90DPDLast6Mnths	0.163895
binning.NbrOfInq_ExclHomeAutoLoan_Last6Mnths	0.074198
binning.income	0.039888
binning.NbrMnthsCurrResidence	0.030396
binning.NbrMnthsCurrCompany	0.019481

MODEL BUILDING

- Applied Logistic Regression, Decision Tree and Random Forest to iteratively build model by removing insignificant variables and variables exhibiting high levels of multicollinearity

- Model was evaluated for accuracy, sensitivity and specificity values

- As the business objective is to find the right customer, the final variable was selected in such way as to attain higher accuracy and higher sensitivity

MODEL EVALUATION

- In Logistic Regression, we eliminated insignificant variable one by one
- In Decision tree, we tuned the hyper parameters to find the best model
- In Random Forest, we varied the hyper parameters and created a model
- Got the best results using **Logistic Regression**

Cut Off Value = **0.54**
Accuracy = **62.51%**
Sensitivity = **62.04%**
Specificity = **62.53%**
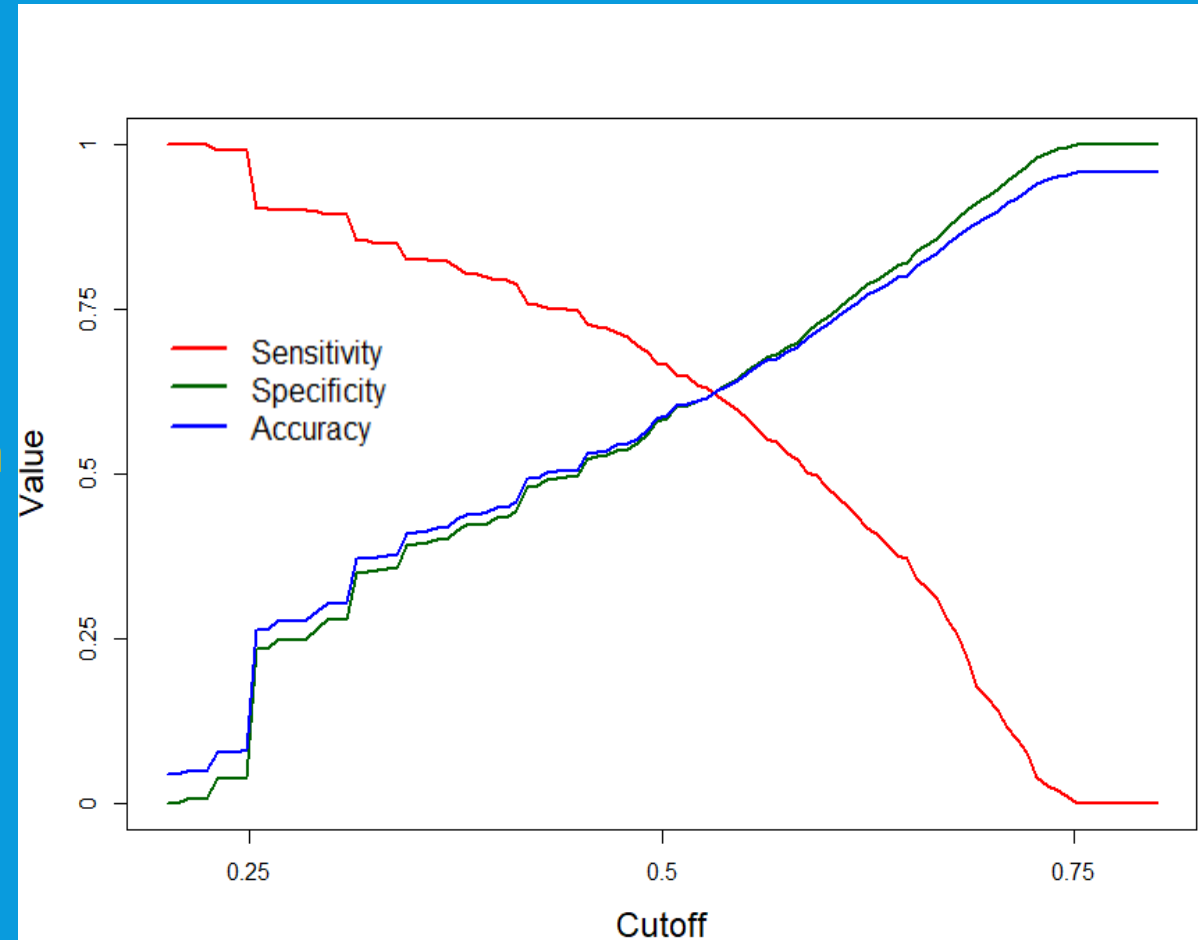
MODEL EVALUATION AND INTERPRETATION

Final Model produced 4 factors which are significant and most decisive
1. Applicants who have crossed **30DPD once or more in last 6 months**
2. Number of **PL trades open in last 12 months** by Applicants
3. **Tenure** of Applicants **residing in current residence**
4. **Average credit card utilization** of Applicants

APPLICATION SCORECARD: TEST DATA

- Cut Off derived = **0.54**
- Based on above cut-off, CredX should approve application with score **>330**
- The graph shows the distribution of application in test data where the applications should be approved/rejected
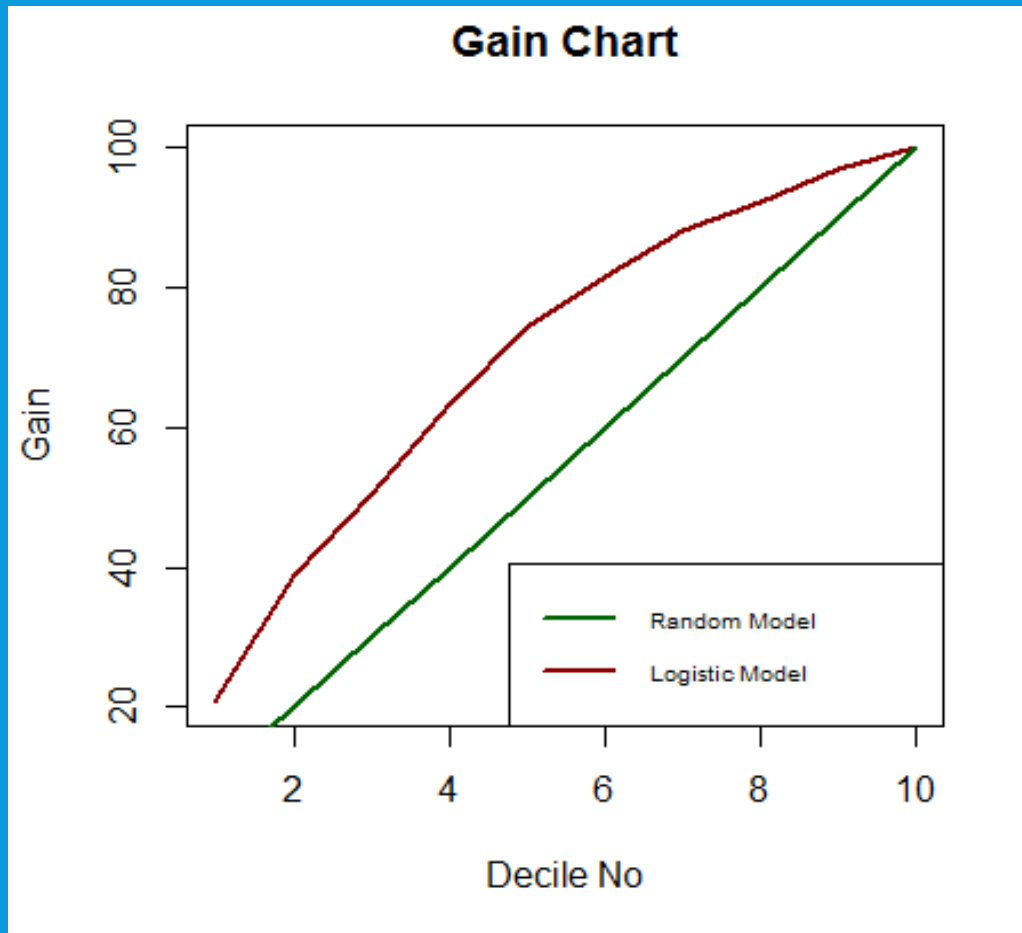

Application Score Card
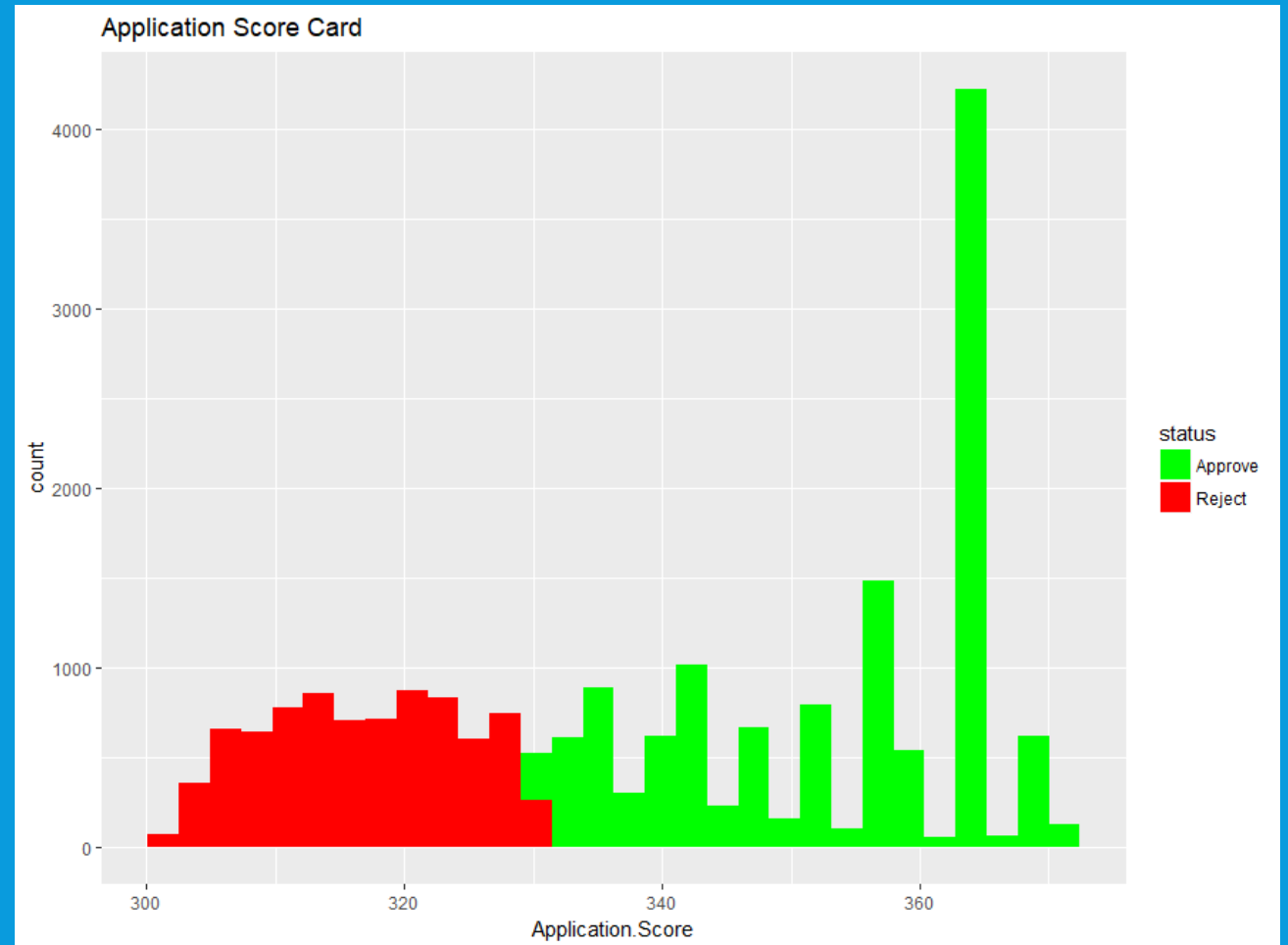
APPLICATION SCORECARD: REJECTED DATA

- We can see that majority of rejected application have score **less than 330**
- The proposed model correctly predicts **~98%** rejected applicants
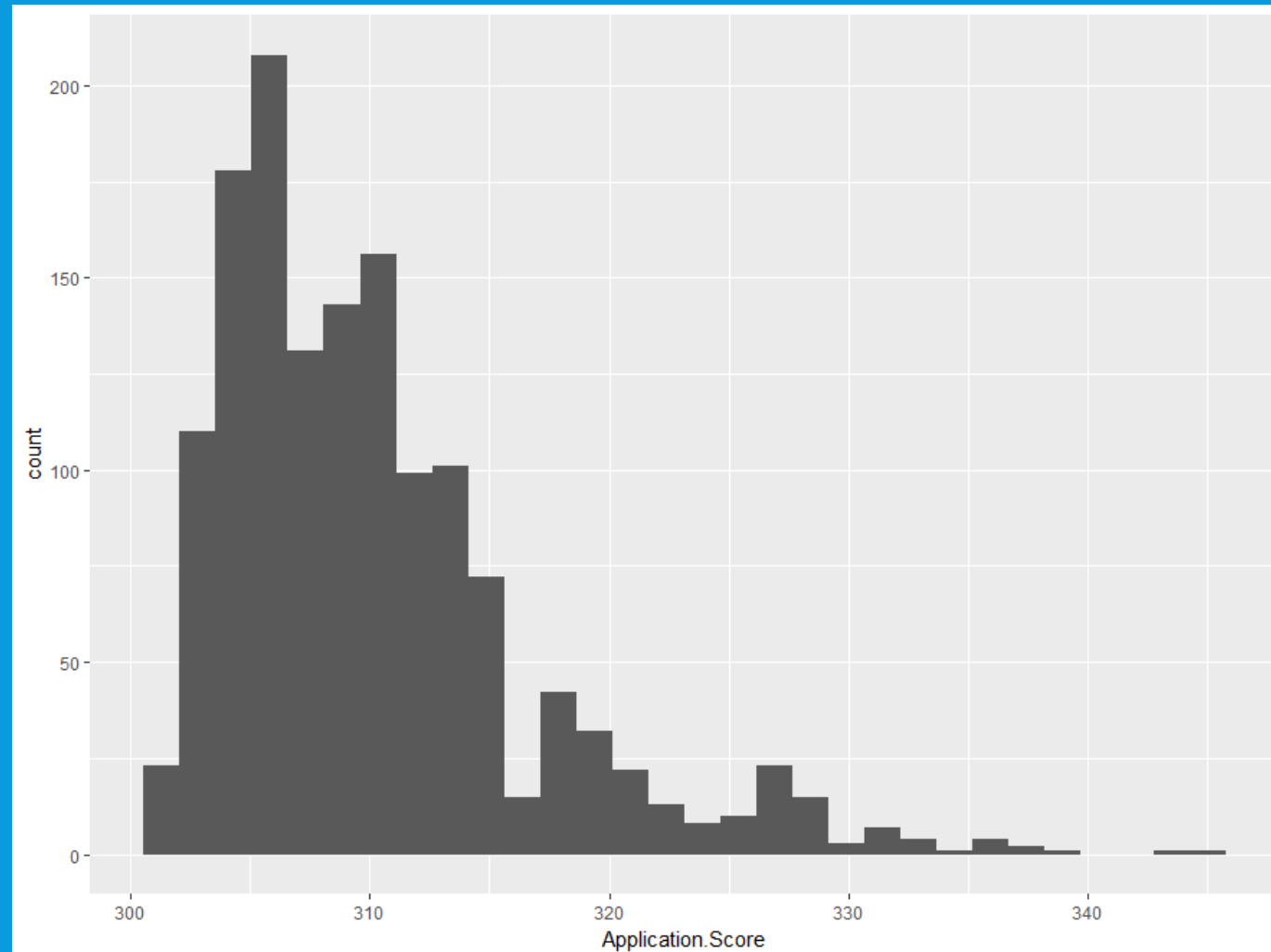
FINANCIAL BENEFIT

Gain Chart

- Using the proposed model, we can avoid credit risk of **81%** of defaulters by targeting **60%** of the applicants

bucket	total	totalresp	Cumresp	Gain	Cumlift
1	2081	181	181	20.56818	2.056818
2	2080	161	342	38.86364	1.943182
3	2081	103	445	50.56818	1.685606
4	2080	114	559	63.52273	1.588068
5	2080	95	654	74.31818	1.486364
6	2081	63	717	81.47727	1.357955
7	2080	58	775	88.06818	1.258117
8	2081	36	811	92.15909	1.151989
9	2080	42	853	96.93182	1.07702
10	2080	27	880	100	1

FINANCIAL BENEFIT
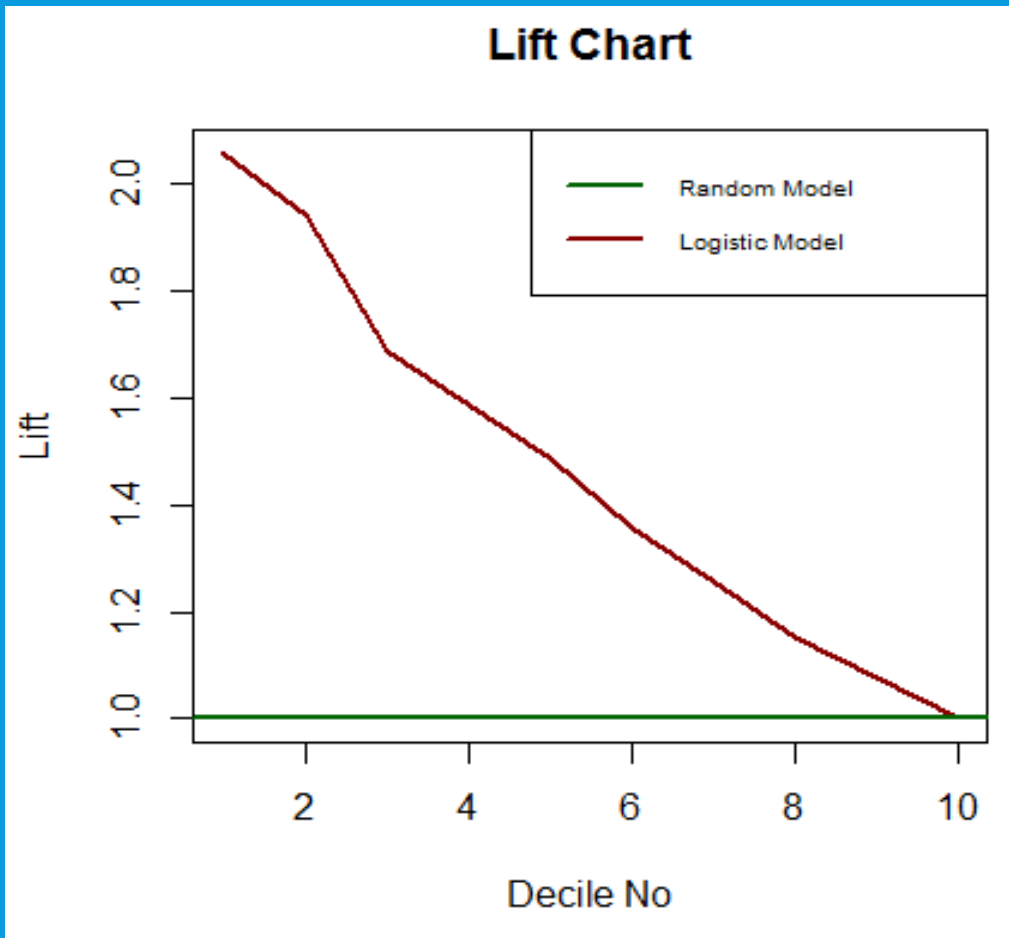
Lift Chart

- The lift chart shows we can predict defaulters **~1.4** times compared to any random model when we are targeting 60% of applicants

RECOMMENDATIONS

- On an average our model is able to predict with **63%** accuracy whether the applicant will default or not

- Management should focus on following characteristics of the applicant
  - **Crossed 30DPD in last 6 months**
  - **Number of PL Trades opened in last 12 months**
  - **Tenure of Applicants residing in current residence**
  - **Average credit card utilization**

- CredX should approve applicants having score **> 330**