# BFS Capstone Project

## Final -Submission

Group members:

1. Poornachandran R
2. Rajesh Bajagur
3. Rajesh Kusuma
4. Sagar Sheth

# Problem Statement

**Problem statement:**

- Analyse past data of CredX to determine the factors affecting credit risk
- Build predictive models to identify the right customers using predictive models
- Finding right customers will help CredX to reduce credit risk
- Application Score Board
  - On the basis of the scorecard, identify the cut-off score below which you would not grant credit cards to applicants.
  - For the rejected population, calculate the application scores and assess the results. Compare the scores of the rejected population with the approved candidates and comment on the observations

**Assumptions:**

- If there is no values in the performance tag, then those records are considered as rejected population
- If Performance tag has value i.e. 1/0, then those records are considered as approved population
- Variables with regards to number  of personal loan taken in last 6 and 12 months are important but it is assumed Total number of Trade variables includes personal loans taken in last 6 and 12 months hence only total number of trades is considered in the model
- If customer makes more number of enquires, then it is assumed that customer is looking to take new loans for additional requirements as opposed to assuming customer is going through financial issue
- If  number of months in current residence is less, then those customer are considered as fresher\the ones who recently employed, and they will not have financial discipline to pay back the credits

# Analysis Approach

- Using CRISP-DM framework for building models for predictive models

**Business Understanding**

- Two data sets are provided i.e. demographic and credit bureau
- Analysis should focus more on approved population to understand the driving factors for credit default

**Data Understanding**

- Performance tag has two values 1(default) & 0(no-default), and these values are considered as approved population
- If there are no values in the performance tag, then those records are considered as rejected population

**Data Preparation**

- Remove duplicate application ids in both data sets
- Merge both demographic and credit bureau data sets
- Create data sets for approved population i.e. records with performance tag with 0s & 1s
- Perform Data cleaning steps i.e. handing missing values, convert chars to factor
- Perform EDA analysis using charts, and also using WoE & IV values

**Model Building**

- Build models using logistic regression , decision tree and random forest techniques
- Planning to build two models
    - Demographic data model for approved population
    - Model using both demographic and credit bureau data approved population

**Model Evaluation**

- Use ROC\AOU along with accuracy , sensitivity & to evaluate & choose the best model
- Use KSS stats and lift & gain charts to evaluate the chosen model

**Model Deployment**

- Build a score card to decide whether or not to approve the credit card

# Driving Variables

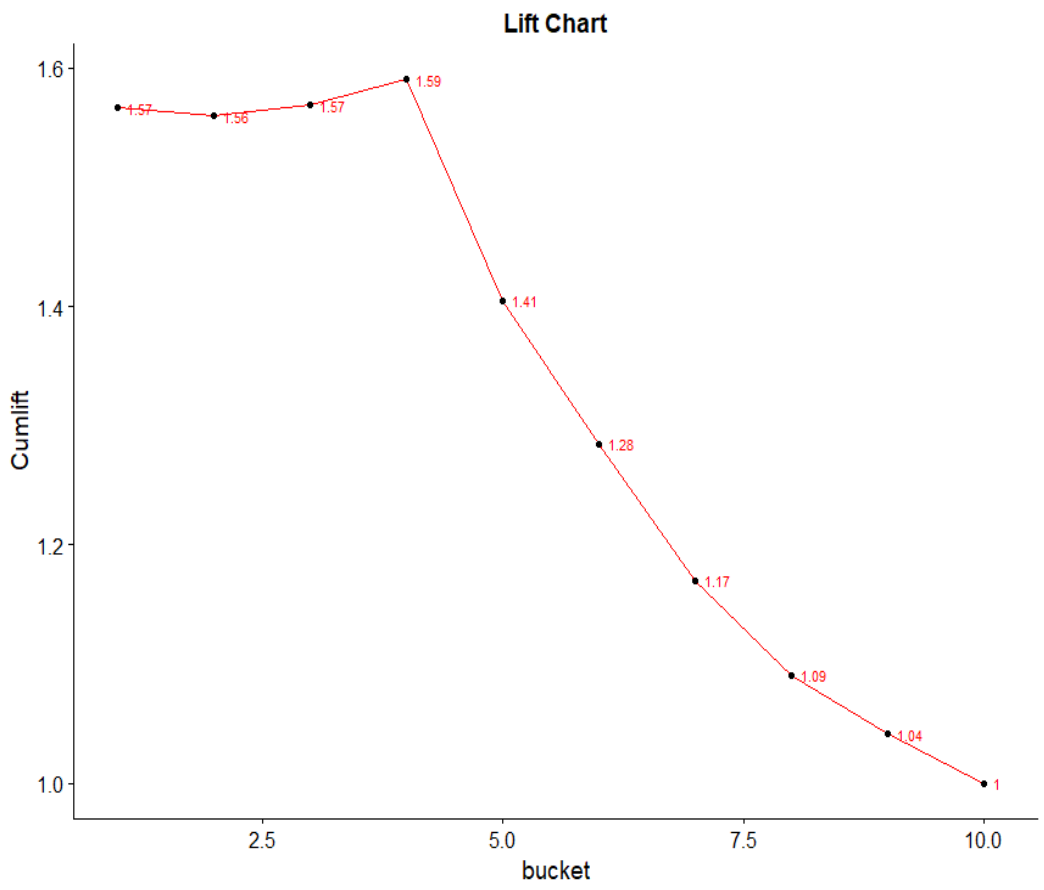
| Variable Name | Explanation |
|---|---|
| **No.of.times.30.DPD.or.worse.in.last.6.months** | 30 days past due is one of the influencing factor i.e. higher the number most likely customer would be defaulted as it is an indication of whether or not customer is facing financial issues |
| **Avgas.CC.Utilization.in.last.12.months** | If credit card utilization is high in last 12 months, then customers are more likely to default. |
| **Total.No.of.Trades** | High number of trades means customer is living closer to the edge of defaulting. |
| **Presence.of.open.home.loan** | If customer has a home loan, then perform due diligence in approving credit card\monitoring the credit performance. As customer will more likely to get defaulted if there more commitments |
| **No.of.months.in.current.company** | If number of months in current company of customers is less, then customer is more likely to get defaulted |
| **No.of.months.in.current.residence** | If number of months in current residence is less, then customer is more likely to get defaulted |
| **Presence.of.open.auto.loan** | If customer has an Auto loan, then perform due diligence in approving credit card\monitoring the credit performance. As customer will more likely to get defaulted if there more commitments |

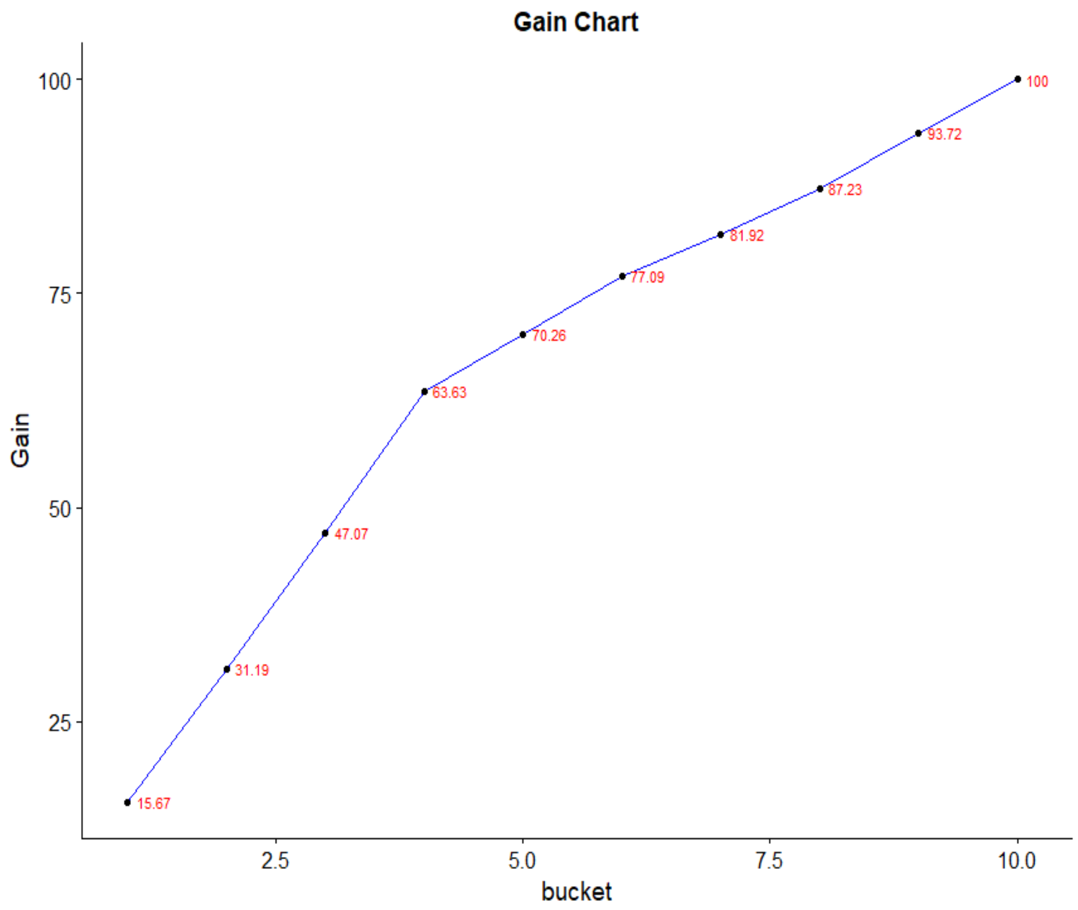
# Model Evaluation\Selection

| Model Name | Accuracy | Sensitivity | Specificity | KSS | AUC |
|---|---|---|---|---|---|
| Logistic | 0.605673 | 0.644582 | 0.603962 | 0.248544 | 0.6215 |
| Random Forest | 0.6138979 | 0.621118 | 0.613580 | 0.234698 | 0.6173 |
| Decision Tree | 0.8636 | 0.212560 | 0.892222 | 0.104790 | 0.5523 |

- More weightage is given for sensitivity while choosing the model as model is used to find credit risk of the customer
- Sensitivity & AUC of logistic models greater than other models , hence logistic regression has been chosen for the deployment

# Logistic Regression : Lift and Gain Chart




Logistic regression  model catches 1.59 times more defaulters than a random model would have caught

Gain for a final model is 70% by the 5th decile,this means is that if we sort all defaulters according to probability, then among the top 50% employees of this sorted list, we would find 70% of all customers that were likely to faulted

# Score Card Approach

- Score card used as early indicator of finance strain
- Used probability of non – default i.e. 1-P and offset value calculated based on good to bad odds of 10 to 1 at a score of 400 doubling every 20 points
- Score Card Formula = Offset + (1-P) * 500, and Score card will be ranging from 350\lower to 800\higher
- Least risky customer will have higher score
- Cut-off score is 625.6451 , below which credit cards will not be granted to applicants
- **Score Card Evaluation\Model performance** : Tested score card for rejected population, and it is observed 96% rejected population have less than cut-off score which means score card can be used for real population

# Financial benefit assessment

- Final model catches 1.59 times more defaulters than random model, this means company can put procedures to increase the chances of predicting defaulters thus reducing the credit risk of the bank. Also concentrate on top 50% of total population would reduce the recovery cost
- **<u>Assumption(while approving the credit card)</u>** : Correct decision of the bank would result in 30% of the profit at the end of a specific period say 2–4 years

| | Good Customer | Bad Customer |
|---|---|---|
| Good Customer(observed) | +0.3 | 0 |
| Bad Customer(observed) | -1.0 | 0 |

- Results from logistic regression model

| | Good (predicted_positive) | Bad (predicted _negative) | Row Total |
|---|---|---|---|
| Good(observed) | 19906 (97% True_Positve) | 13053 | 32959 |
| Bad(observed) | 515 (3% False_Positive) | 934 | 1449 |
| Colum Total | 20405 | 14003 | 34408 |

# Financial benefit assessment

- If the company gives everybody the credit card which would result in the following profit per customer: (32959*0.3- 1449*1.00)/34408= 0.24 unit profit.
- Assuming $5000 credit is approved for ALL customers i.e.34408, then total profit for the company : 0.24 * 5000 * 34408 = $ 41,289,600 \$ 41m. Assuming none of the customer is defaulted
- Profit using the model = True Positive*5000*0.3 – False Positive*5000 = 19906*5000*0.3 – 515*5000 = 27,284,000 \$27m
- Result = 27,284,000 - 41,289,600 = -$ 14,005,600 \ -$14m
- **Model Implications :**
  - 14m is the potential loss of revenue due to rejection of good customers
  - Loss is controlled when compared to other models i.e. final model captures 1.59 more defaulters than other models

# Recommendations

- Use final logistic model as it is easy to understand and implement, however continuous improvement of the model is important
- Improve procedures around capturing missing values
- Reducing miss-classification should be the objective of continuous improvement should be
- Regular checks on score card is very important to avoid customer default\ taking pro-active measures
- Demographic model is also important to perform screening of credit card searly in the process, thus reducing processing cost