Acquisition Analytics for CredX

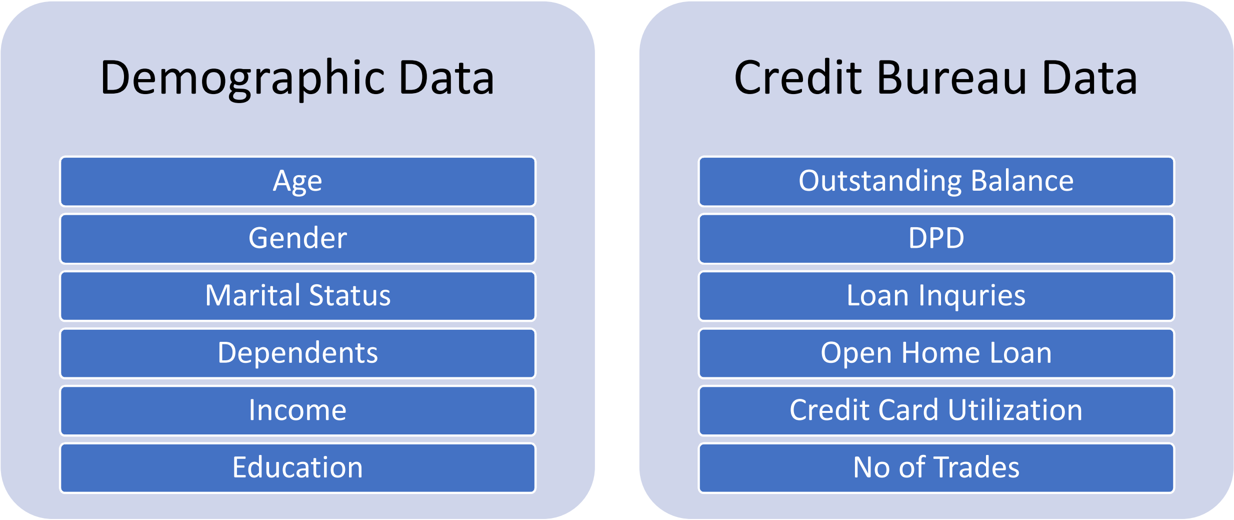
Final Submission

Ajoy Nambiar

Business Objective

- CredX wants to find out factors affecting credit loss. Current default rate is 4.2%
- Make strategy/ model to mitigate risk by building a scorecard to help ‘acquiring the right customers’ using bank’s records– demographic and from Credit Bureau
- Estimate financial benefits by automating the credit approval process

Data used for analysis



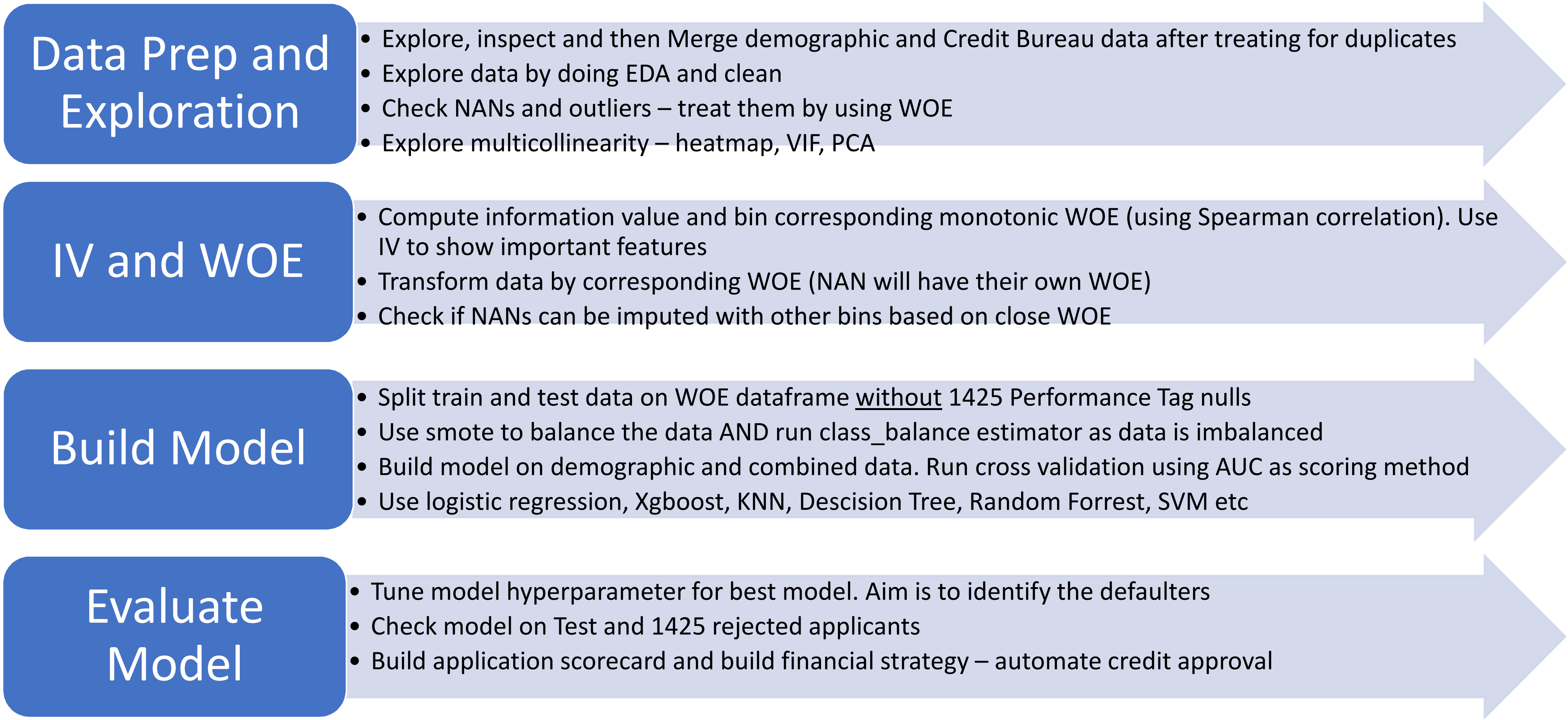
Demographic Data

- Age
- Gender
- Marital Status
- Dependents
- Income
- Education

Credit Bureau Data

- Outstanding Balance
- DPD
- Loan Inquries
- Open Home Loan
- Credit Card Utilization
- No of Trades

Methodology



Data Prep and Exploration

- Explore, inspect and then Merge demographic and Credit Bureau data after treating for duplicates
- Explore data by doing EDA and clean
- Check NANs and outliers – treat them by using WOE
- Explore multicollinearity – heatmap, VIF, PCA

IV and WOE

- Compute information value and bin corresponding monotonic WOE (using Spearman correlation). Use IV to show important features
- Transform data by corresponding WOE (NAN will have their own WOE)
- Check if NANs can be imputed with other bins based on close WOE

Build Model

- Split train and test data on WOE dataframe without 1425 Performance Tag nulls
- Use smote to balance the data AND run class_balance estimator as data is imbalanced
- Build model on demographic and combined data. Run cross validation using AUC as scoring method
- Use logistic regression, Xgboost, KNN, Descision Tree, Random Forrest, SVM etc

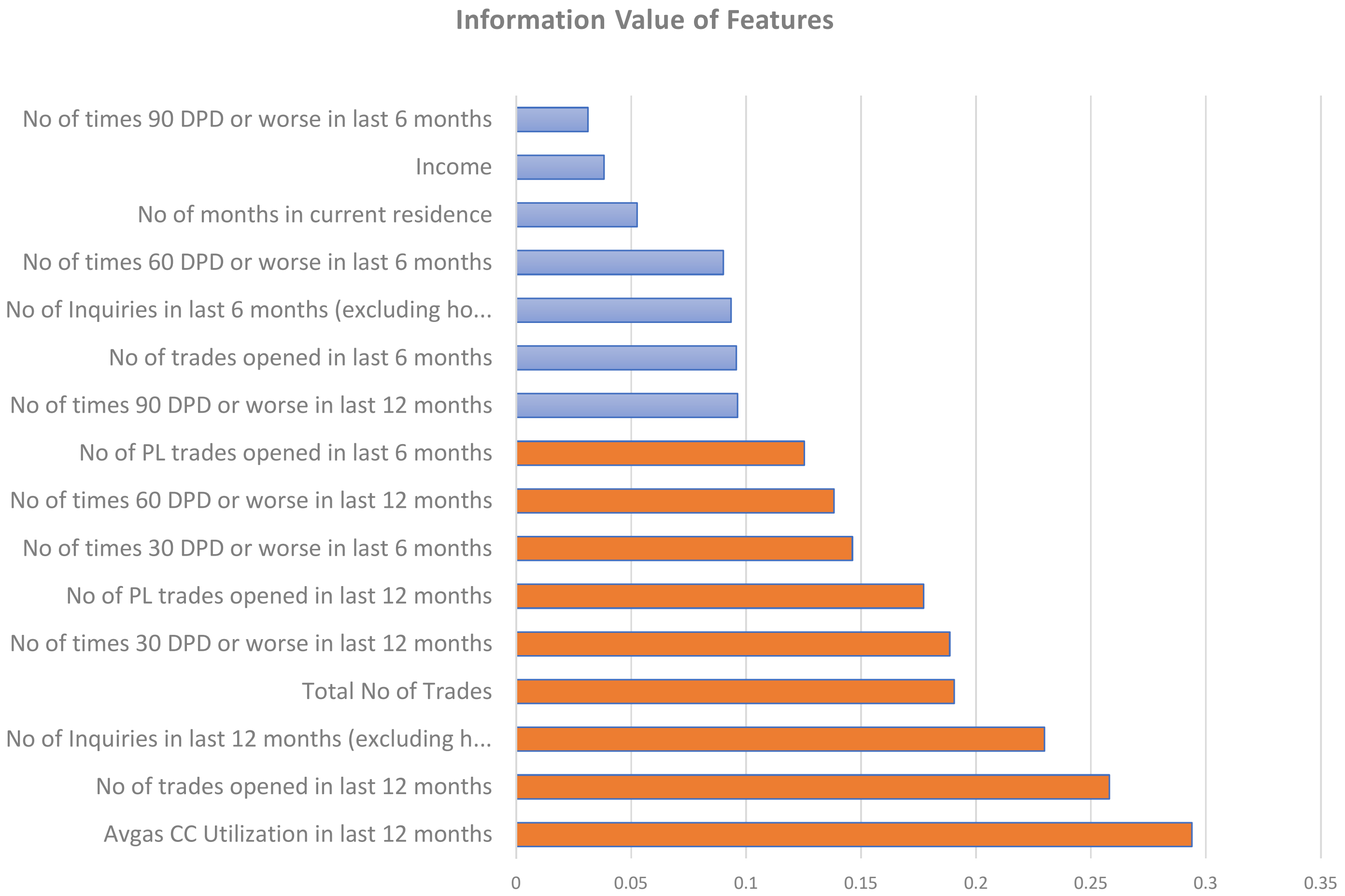
Evaluate Model

- Tune model hyperparameter for best model. Aim is to identify the defaulters
- Check model on Test and 1425 rejected applicants
- Build application scorecard and build financial strategy – automate credit approval

Data Preparation and Exploration – Quality Summary

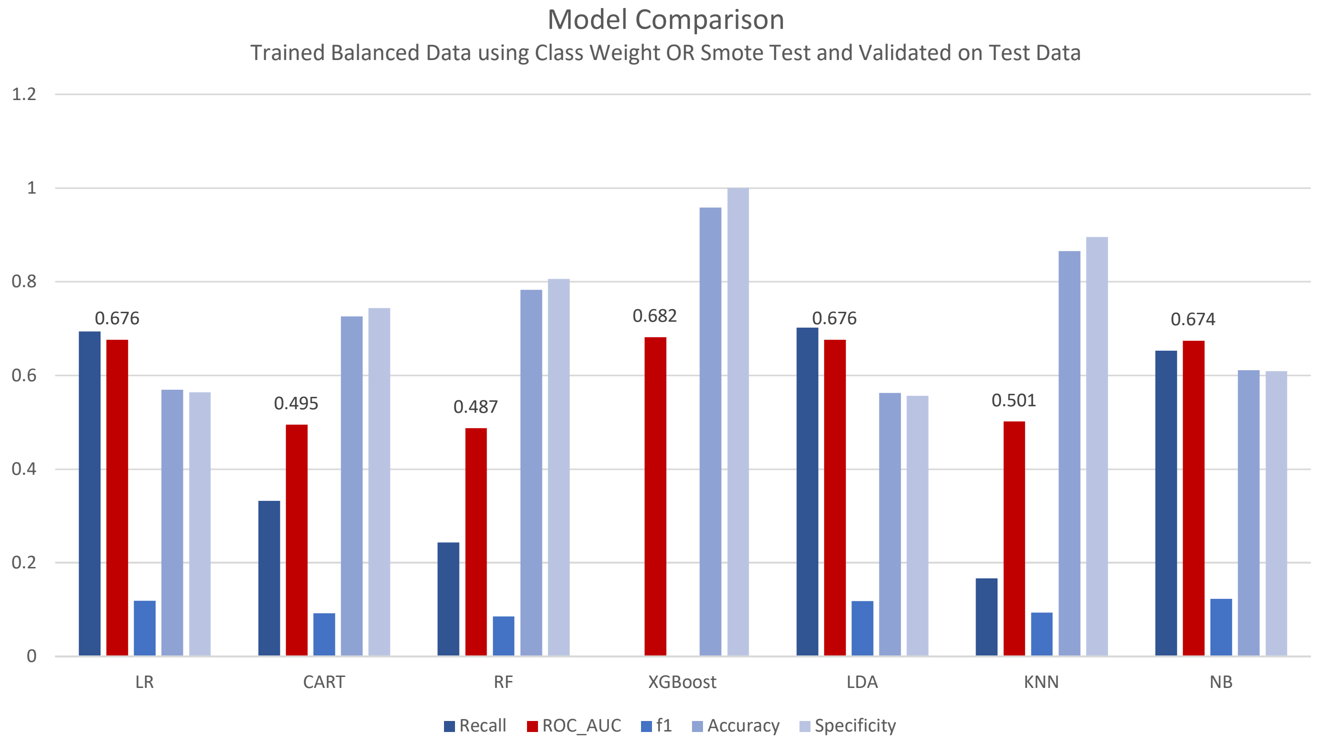
- 3 Duplicate ID show up of 3 different applicant. Delete duplicate with aim to retain applicants who defaulted
- 1425 performance Tag are null. These most likely are applicants who were rejected credit
- Remove Performance tag in one sheet (Credit bureau) before merging
- Missing data handled by imputing appropriately (shown above) but features with large number of null like ‘Avgas CC Utilization in last 12 months’ will be treated and imputed by using their WOE
- WOE will also help in treating outliers
- Gender will not be used in model (may amount to discrimination). Only used to explore data

Information Value – Identify Important Features



- Shown here moderate and weak predictive features
- The most important features are:
  - Avgas CC Utilization in last 12 months
  - No of trades opened in last 12 months
  - No of Inquiries in last 12 months (excluding home & auto loans)
  - No of times 30 DPD or worse in last 12 months
  - No of PL trades opened in last 6 months
  - No of trades opened in last 6 months

Various ML Model – Validate on Test Data



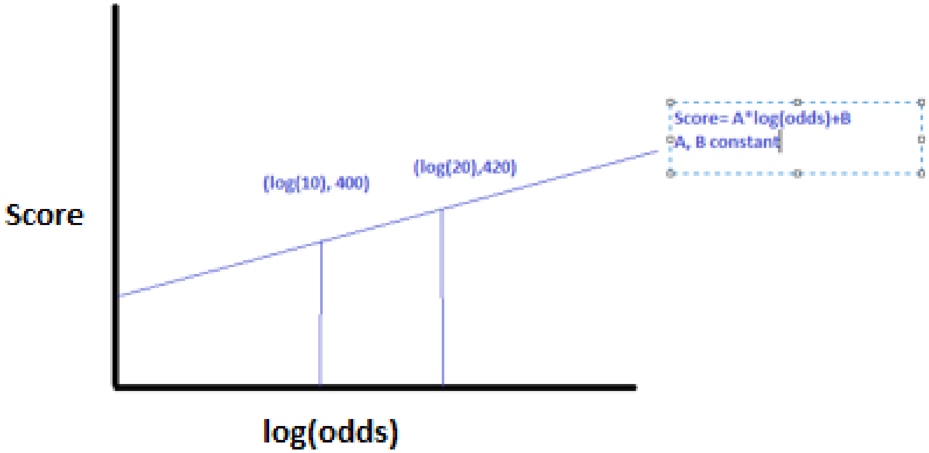
- Models Trained on class_balance estimator AND SMOTE balanced train data and tested on 'unbalance' train data
- Primary metric is ROC_AUC
  - Close to 1 means ideal model
  - Close to 0.5 is naïve model
- Best ROC_AUC and Recall is of **Logistic regression model & XGBoost** – will tune this further

Best Model

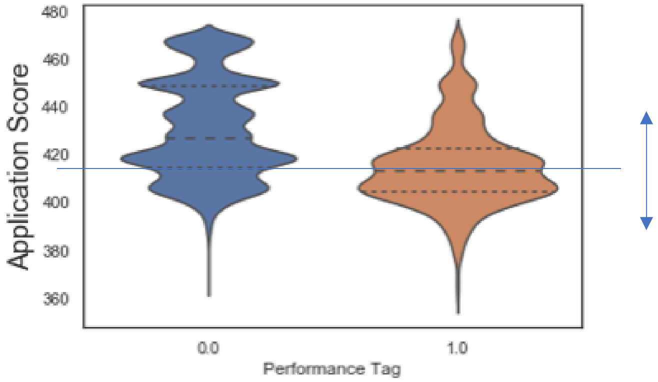
- **Simple Model**: Used RFE and VIF to prune the features from 29 to 6 with highest VIF without drop in model AUC using Logistic Regression. AUC (on test)=0.6747
- **Best Model**: Delivered by PCA and XGBoost using balanced_estimator. AUC (on test)=0.6748. Gain-Lift chart is the best among all tuned models - **Captures *85% of all defaulters in 6 deciles***
- At optimum threshold model has
  - Accuracy= Sensitivity=Specificity= 65%
- Test model on rejected applicants
  - Total Rejected size = 1425
  - Predicted defaulters 1414 candidates

Decile	total	bad	cum-bad	bad_gain	cumlift
1	6987	780	780	26.46	2.65
2	6987	513	1293	43.86	2.19
3	6986	422	1715	58.18	1.94
4	6987	319	2034	69	1.72
5	6986	278	2312	78.43	1.57
6	6987	203	2515	85.31	1.42
7	6987	177	2692	91.32	1.3
8	6986	111	2803	95.08	1.19
9	6987	91	2894	98.17	1.09
10	6987	54	2948	100	1

Application Scorecard



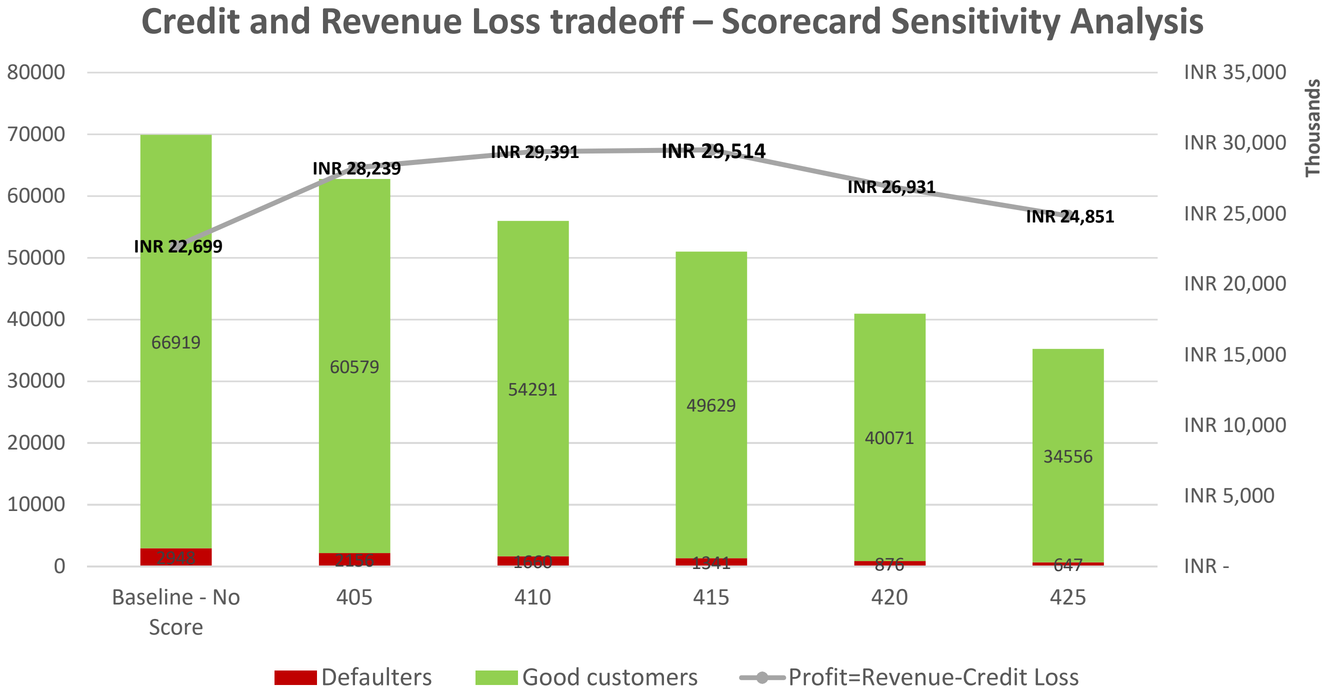
- Solving the linear equation- Base score is 400 at odds 10:1 and pdo=20
  - Score= A*log(odds)+B
  - A=20/(log(20)-log(10))=28.8539
  - B=400-A*log(10)=333.5614



Score Threshold set to optimize business objective – **maximize operational profit** in P&L

Financial Impact – Credit Loss vs Revenue Loss* tradeoff

*Revenue Loss occurs due to good customer identified as 'bad' by model



Credit and Revenue Loss tradeoff – Scorecard Sensitivity Analysis

- Profit= Revenue-Credit Loss
- **Auto approval score threshold** sensitivity analysis performed
- Assuming for each defaulter the operation credit loss is 15K INR & for each good customer revenue is 1K INR
- The baseline profit (NO model) is 22.7 m INR and best threshold of score 415 would yield profit of 29.5 m INR i.e. **approx. 30% increase in profits**
- At score 415 threshold the default rate would 2.63% compared to baseline default rate of 4.2%
- ***Model assumption:*** All applicants are independent and there is no correlation between default behavior of two individuals