ANALYSES AND PREDICTION ON LOAN DEFAULTS

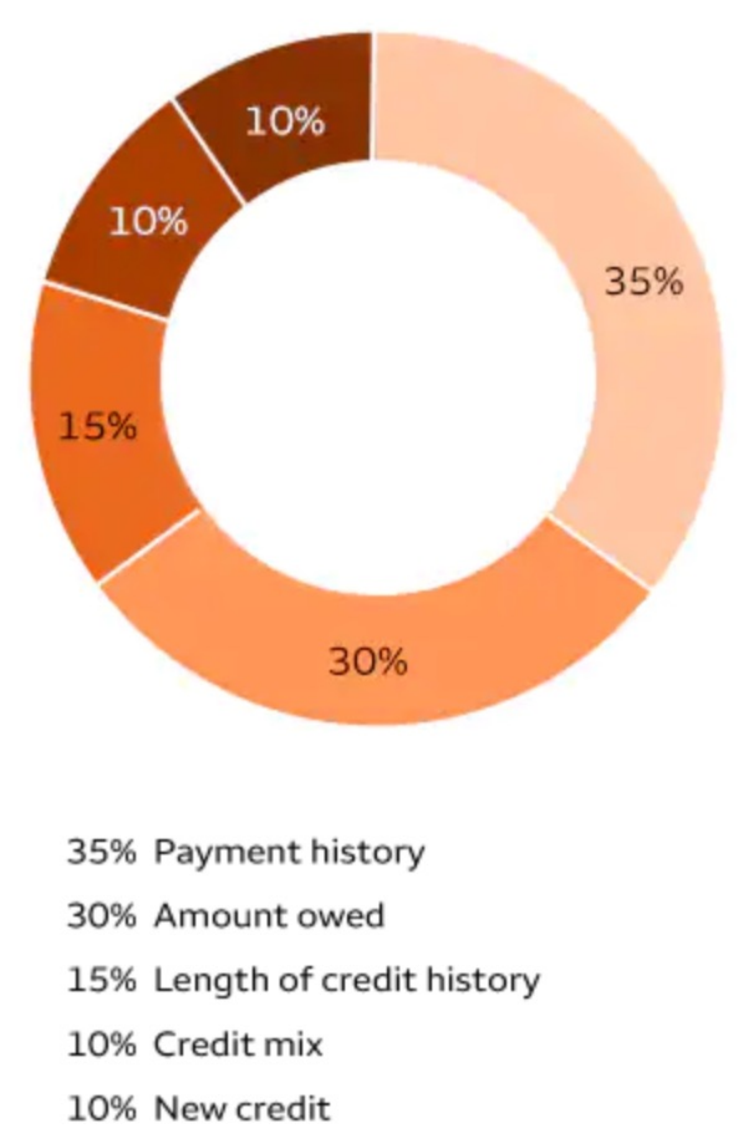
Beverly Yang

05MAR23

CONTENTS

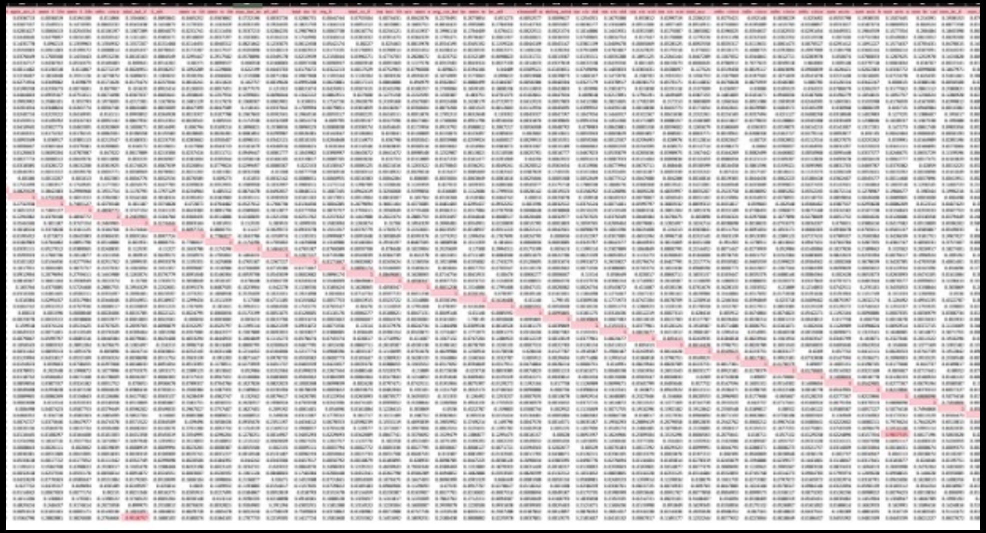
SUMMARY AND GOAL

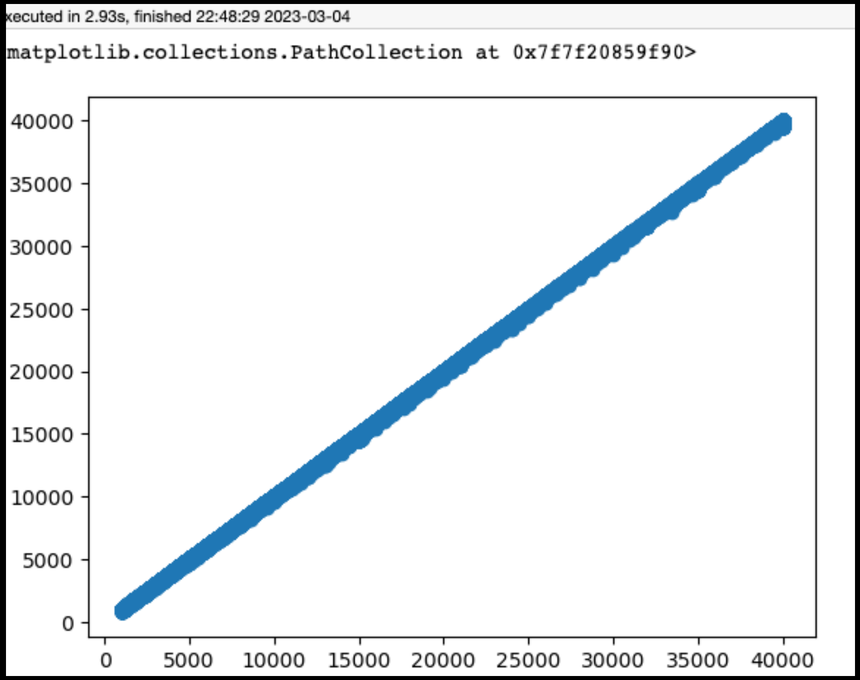
- Better assess the risk of potential borrowers
- Assumption:
  - Revolving credit features more important than installment features
  - Derogatory and delinquency remarks are considered
  - Select model based on highest ROCAUC score
- Most of categorical features removed
- Saga-lasso logistic regression model is the best at detecting loan default

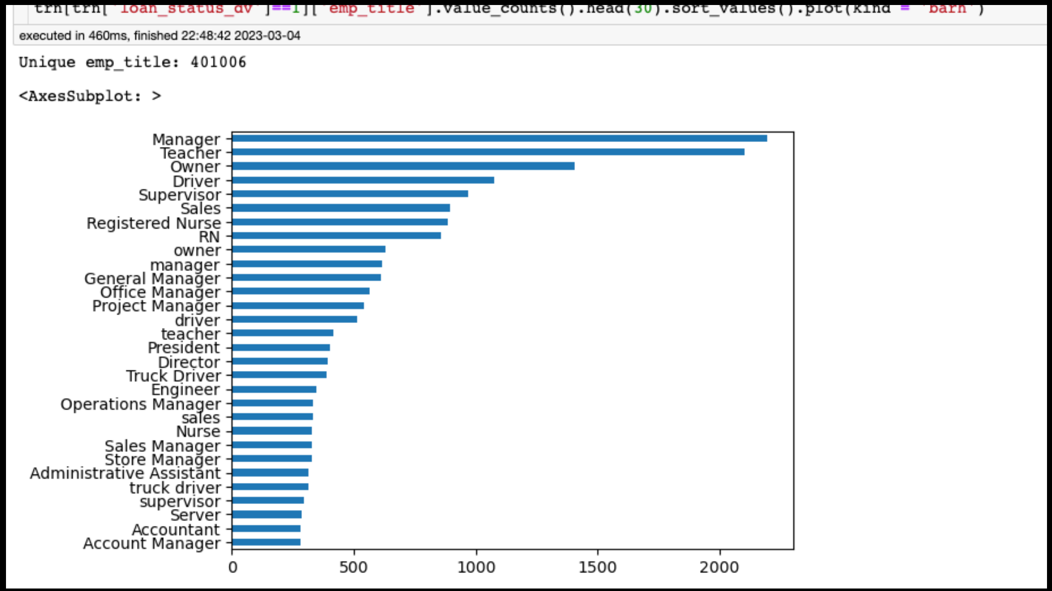


EXPLORATORY DATA ANALYSIS

- Data Cleaning and Preprocessing
    - Charged off event rate is: 7.31%
- Numeric: Pearson's correlation
- Categorical: Chi-square, and Cramer V score
- Most numerical data skew to the right



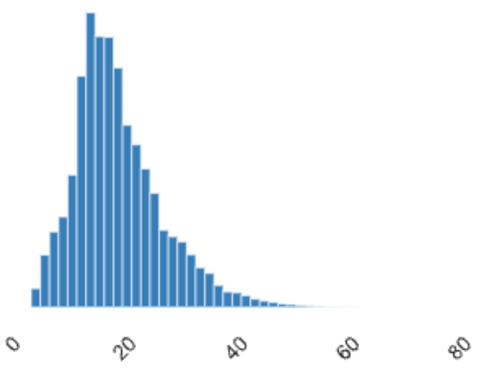




cr_dur_yr

Real number (ℝ)

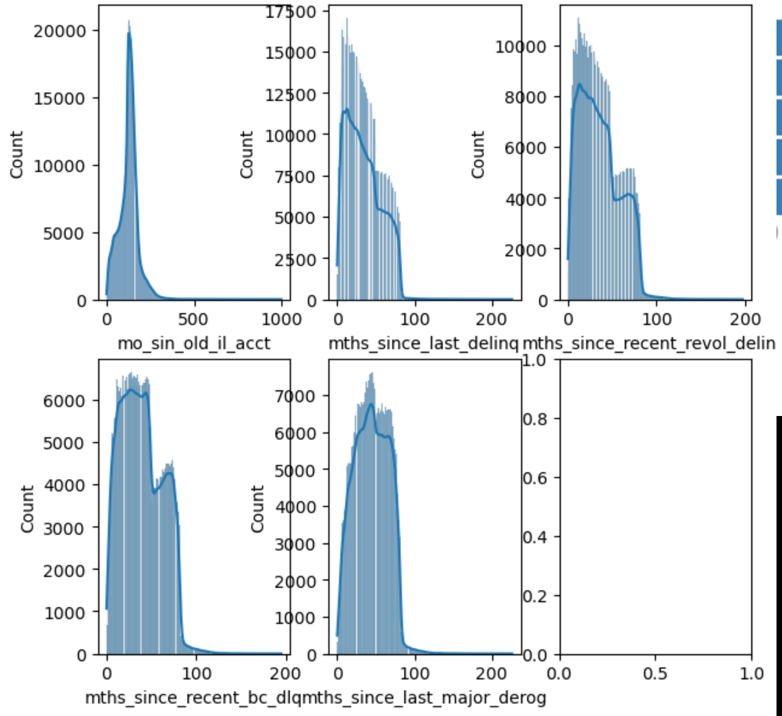
Distinct	2775	**Minimum**	3.079452055
Distinct (%)	0.2%	**Maximum**	84.89589041
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	18.68857259	**Memory size**	13.9 MiB

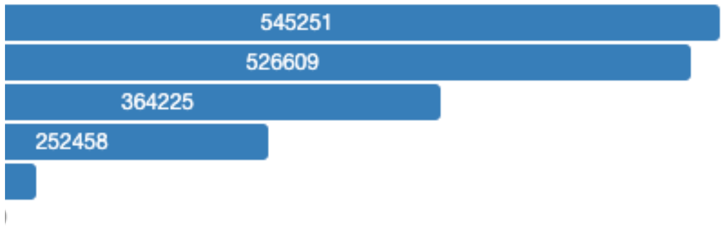


gr

Categorical

Distinct	
Distinct (%)	
Missing	
Missing (%)	
Memory size	





More details

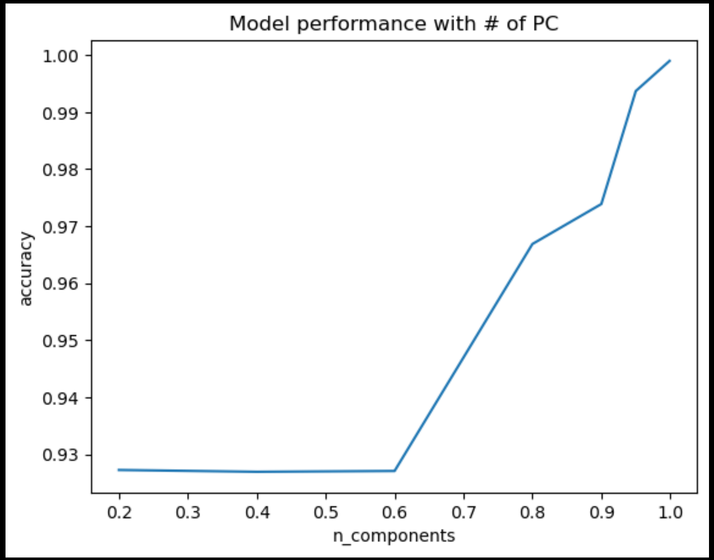
FEATURE SELECTION AND ENGINEERING

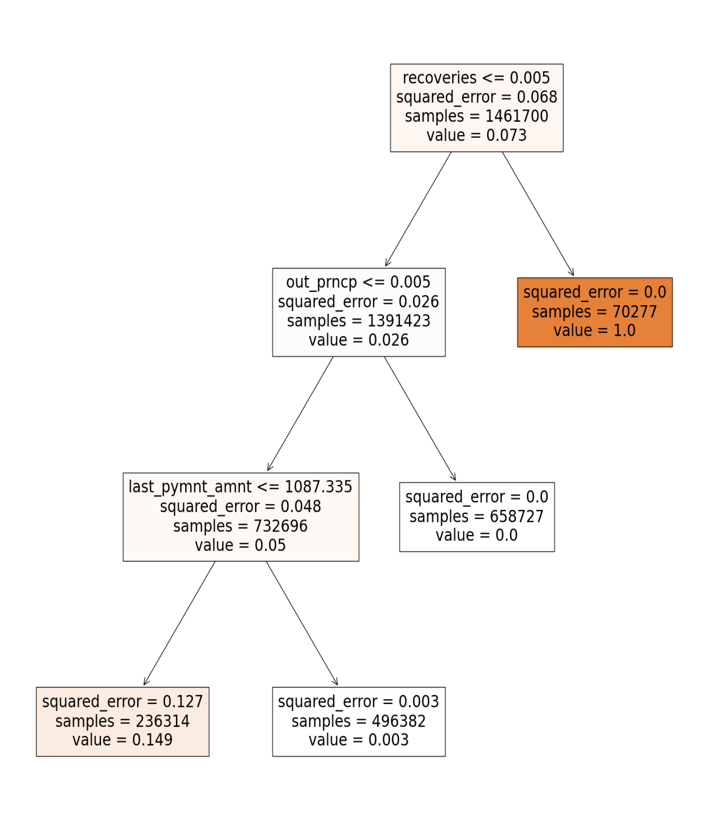
total_rec_int	6.33E-03	num_accts_ever_120_pd	-1.01E-06
loan_amnt	5.67E-03	mort_acc	-6.02E-07
recoveries	3.80E-03	num_tl_op_past_12m	-4.03E-07
total_rec_late_fee	2.67E-05	pub_rec_bankruptcies	-3.22E-07
revol_bal	1.95E-05	pub_rec	-2.99E-07

- Impute numeric values: dti, inq
- Create credit history feature
- 44 features are used for training

Feature importance:

- Baseline logistic regression model
- PC performance
- Decision Tree Regressor





MACHINE LEARNING MODELING

Simple logistic regression modeling

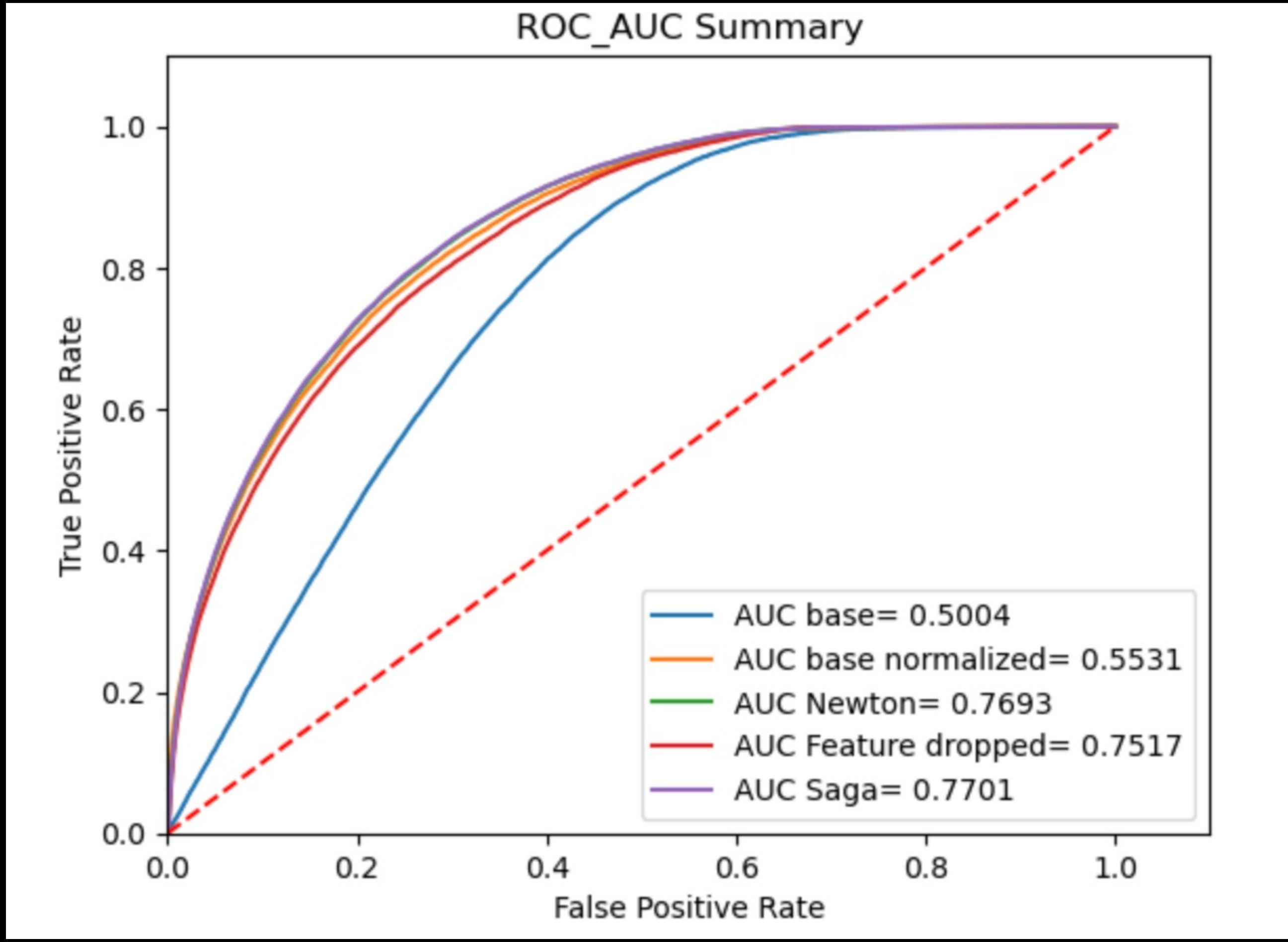
1. Helper functions for one-hot encoding, scoring, plotting ROCAUC
2. Prepare features and target
3. Split training dataset into training and validation set. Use all test data for prediction
4. Normalization
5. Fit various logistic models
6. Predict use trained model

solver	'liblinear'	'lbfgs'	'newton-cg'	'sag'	'saga'
Multinomial + L2 penalty	no	yes	yes	yes	yes
OVR + L2 penalty	yes	yes	yes	yes	yes
Multinomial + L1 penalty	no	no	no	no	yes
OVR + L1 penalty	yes	no	no	no	yes

Source

MODEL PERFORMANCE USING TRAINING SET

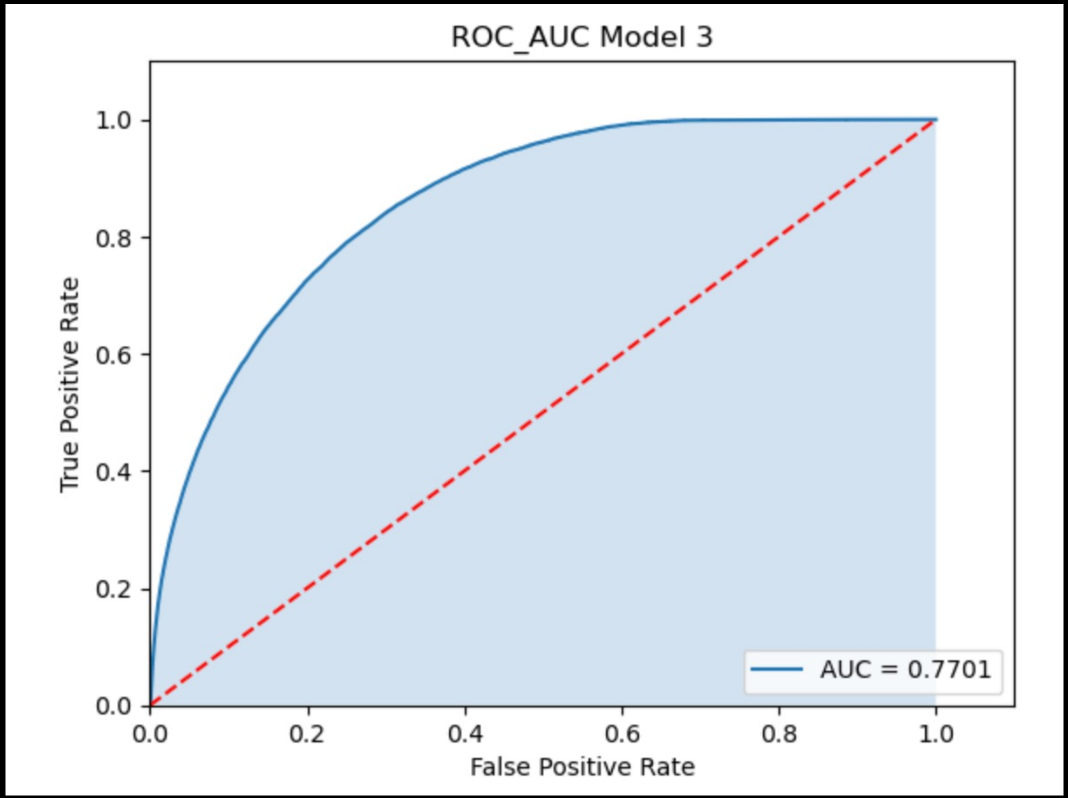
- Most accurate and precise:
  - base normalized
    - 93% accuracy
    - 70% Precision
- Best recall:
  - newton-Cholesky: 83%



MODEL SELECTION

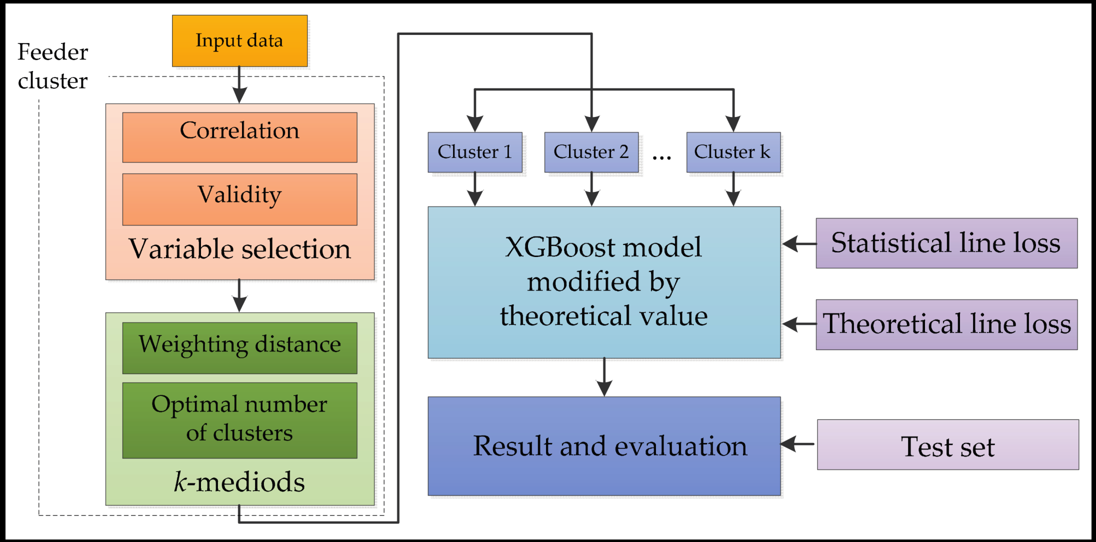
- Saga for detection
- 42.9% default predicted in test
- Low Precision & High Recall

```
        Prediction
Actual        0       1
     0 240782   98072
     1   4529   22042
None
Accuracy: 0.7192282958199356
Precision:    0.18350899978353896
Recall:       0.829551014263671
Specificity: 0.7105774168225844
F1 SCORE for Default:  0.30053516037768
```



IMPROVEMENTS

- Under-sampling
- EDA:
  - Text mining and refining emp_title columns
  - Remove all rows with NaNs
  - Investigate geographic implications
- Feature Engineering
  - Assign class weight
  - Advanced imputations
- Narrow down more features (GridSearch)
- XGBoost Tree
- Business understanding



Source

THANK YOU!