Home Credit Risk Analysis

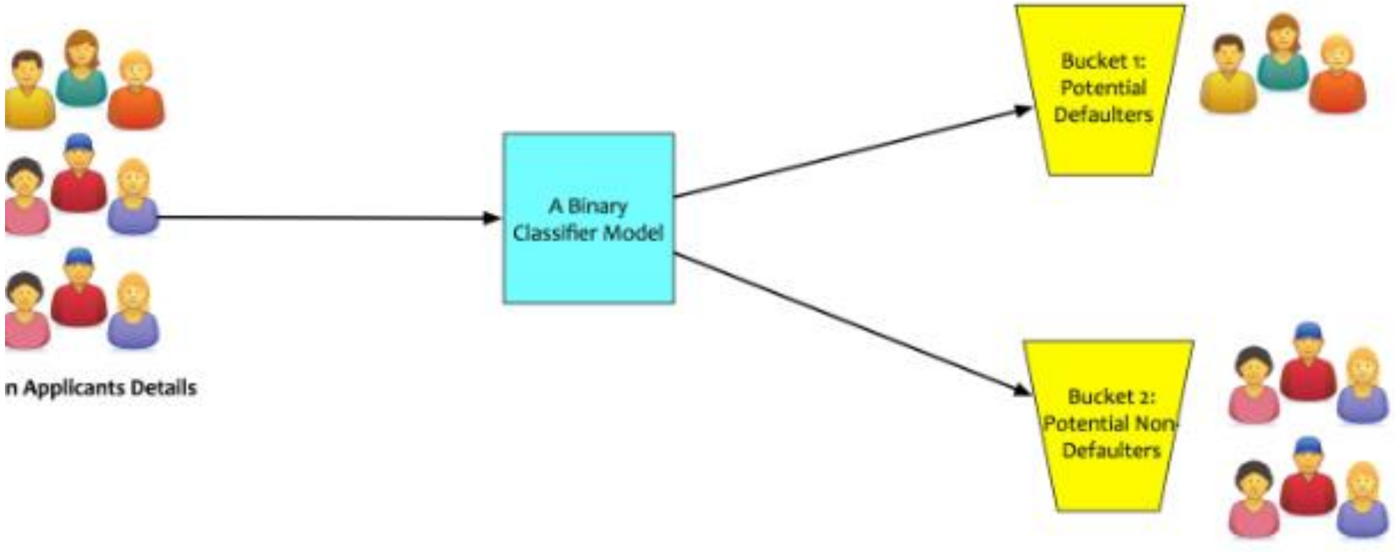
Ly Nguyen

Problems

- A number of increasing loan applicants
- Loss due to a borrower’s failure to make payments
- Loss of potential valued customers

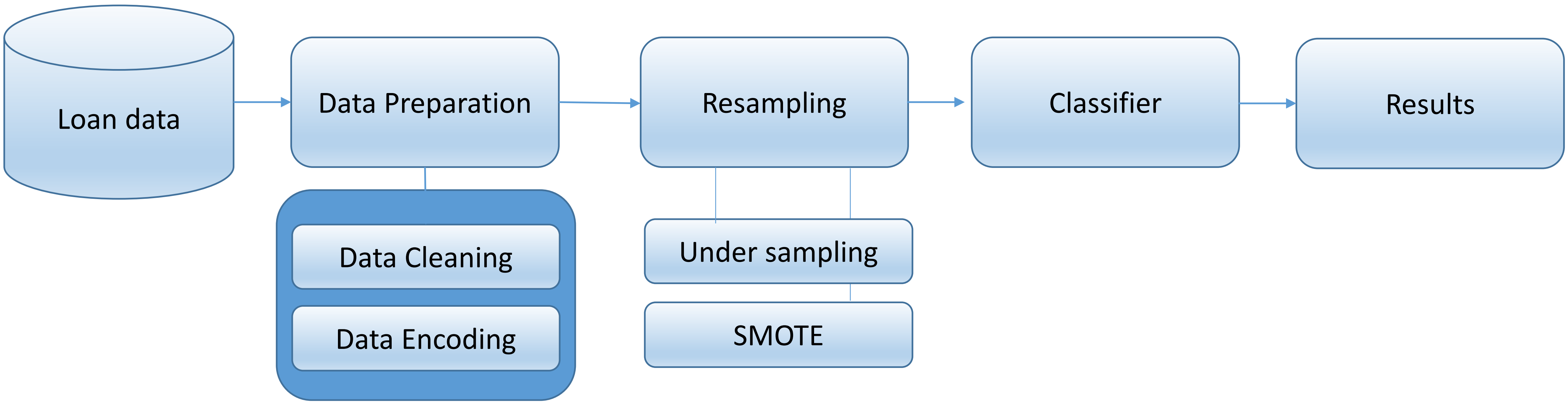
- How risky is the borrower?
- Should we lend him/her?

Objective



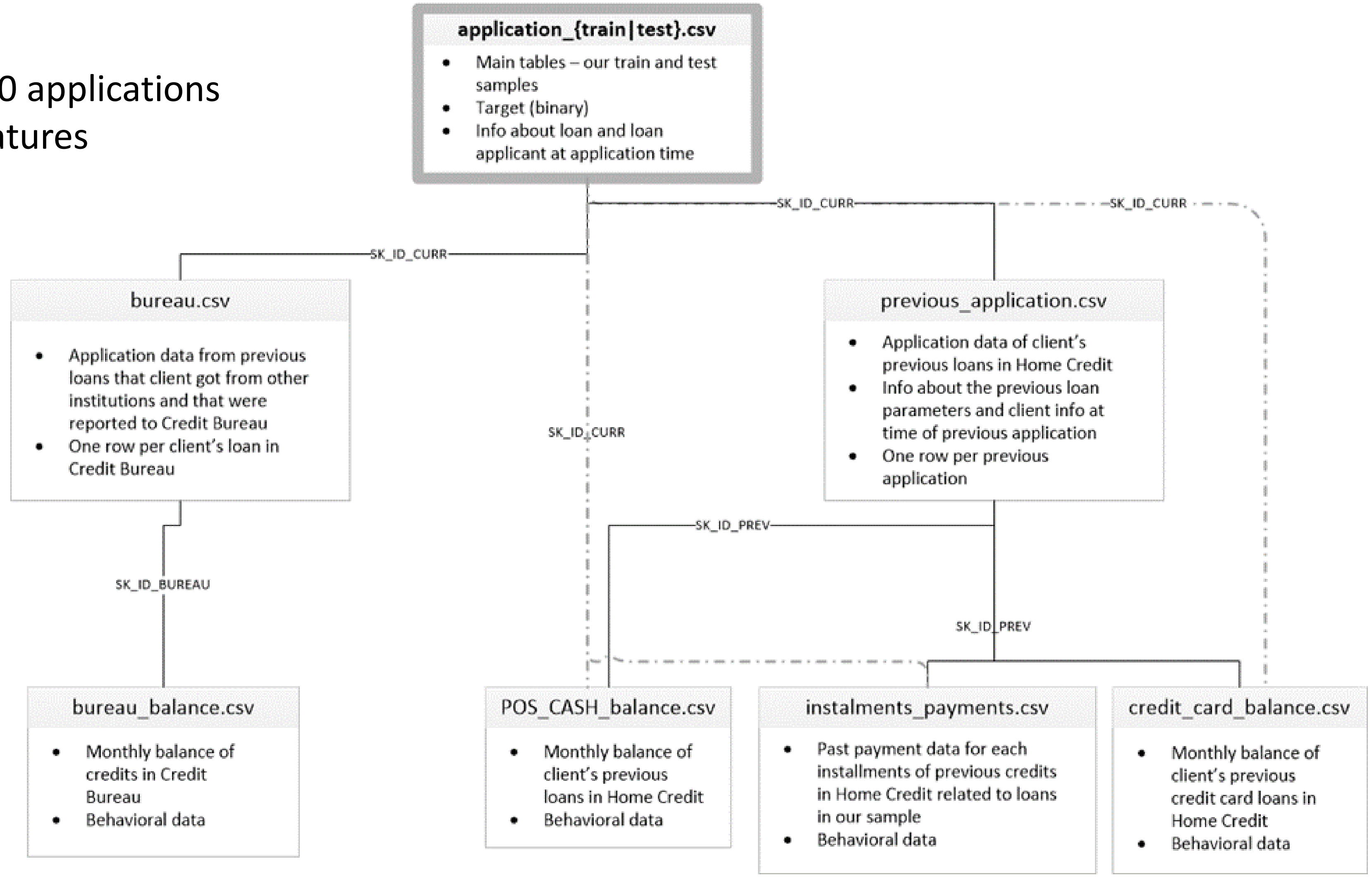
n Applicants Details
A Binary Classifier Model
Bucket 1: Potential Defaulters
Bucket 2: Potential Non-Defaulters

Model Architecture



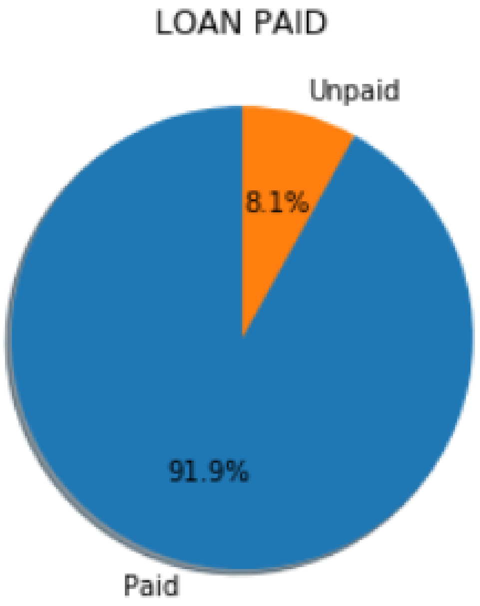
Dataset

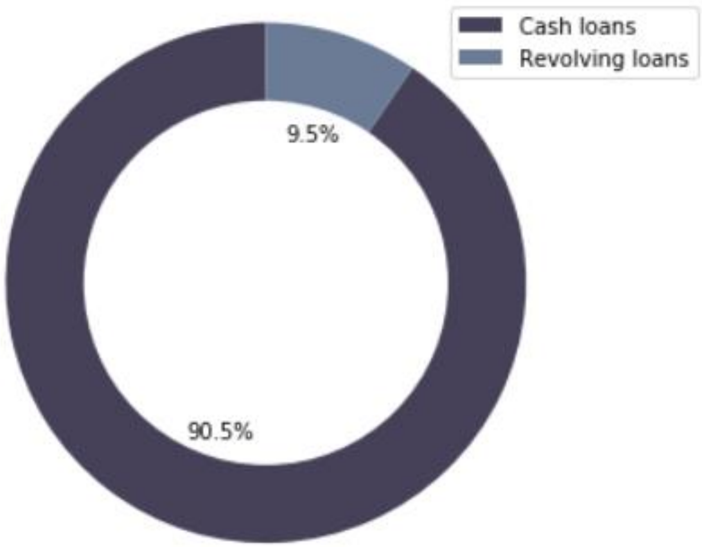
- 250,000 applications
- 200 features

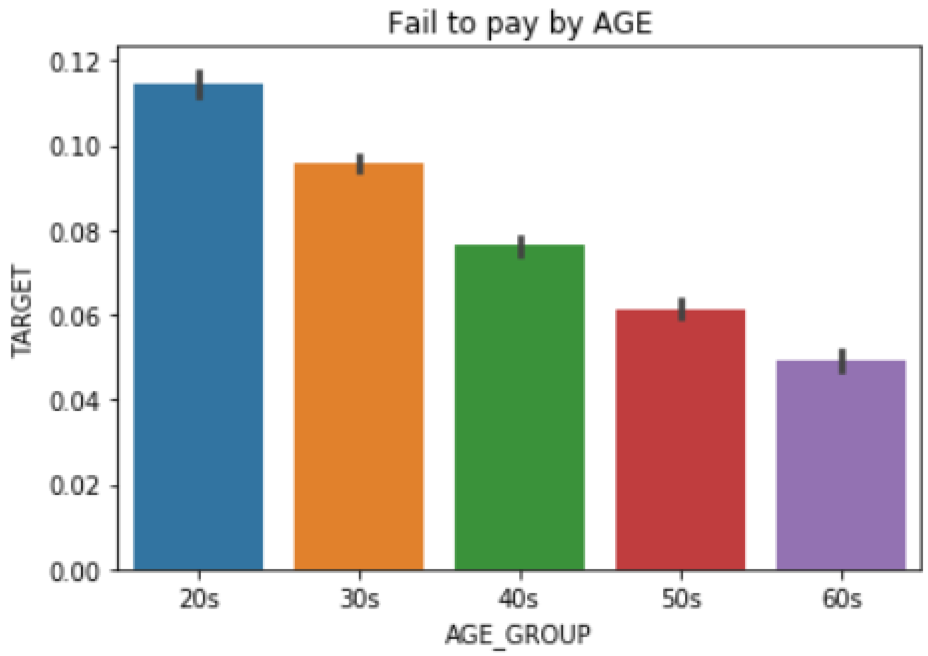


Exploratory Data Analysis

- **Question:** Is the loan was paid on time?

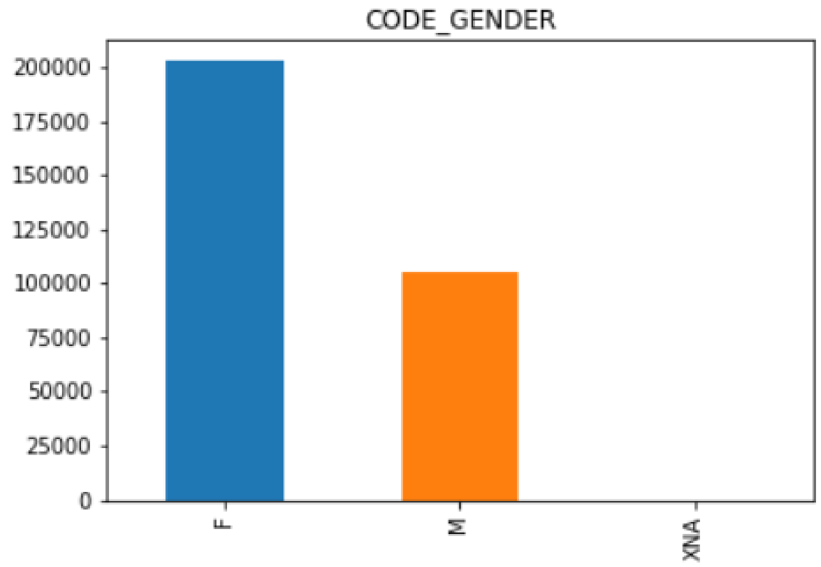


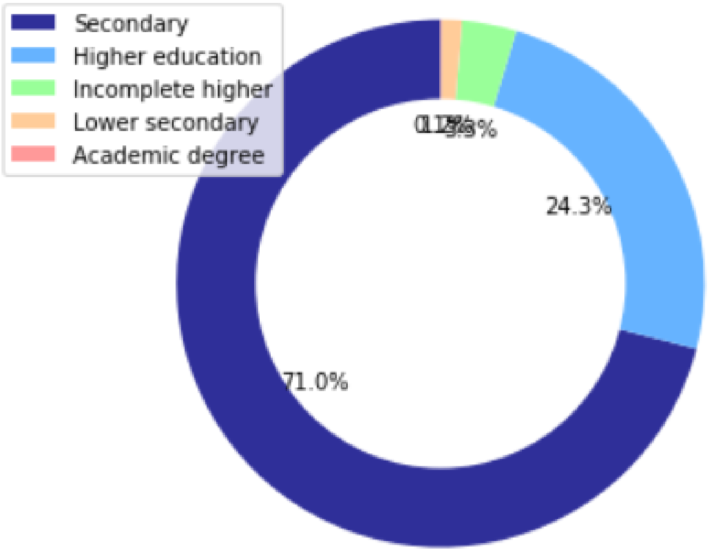


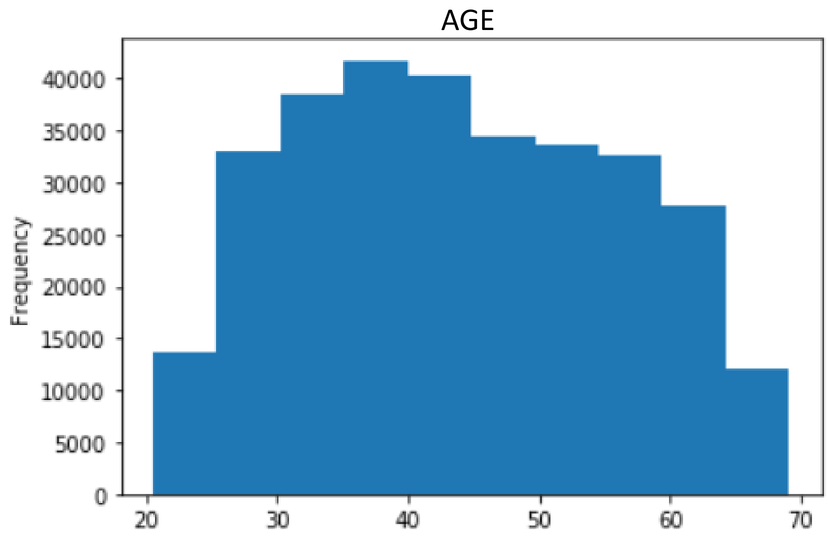


Exploratory Data Analysis

Question: Which customers mostly apply for loan?

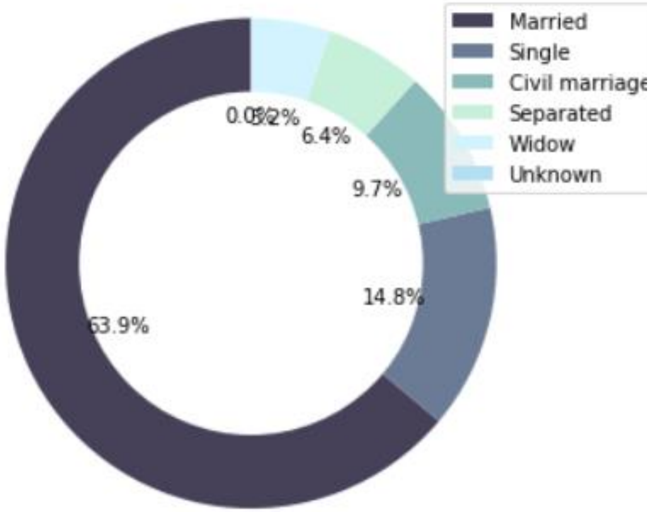


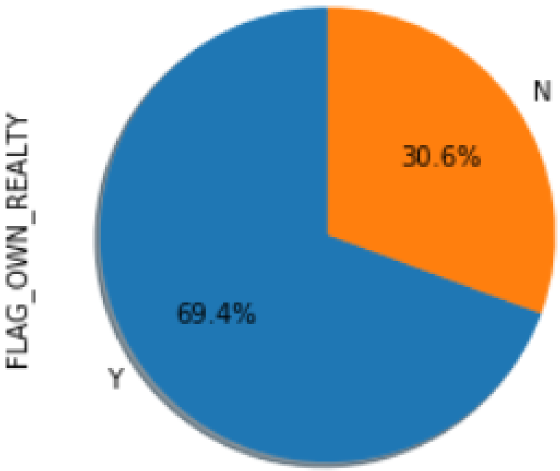


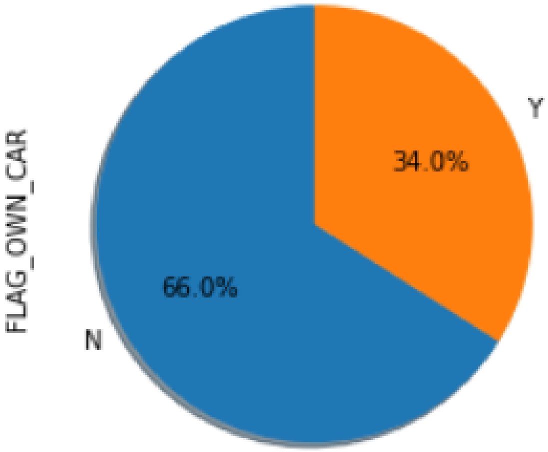


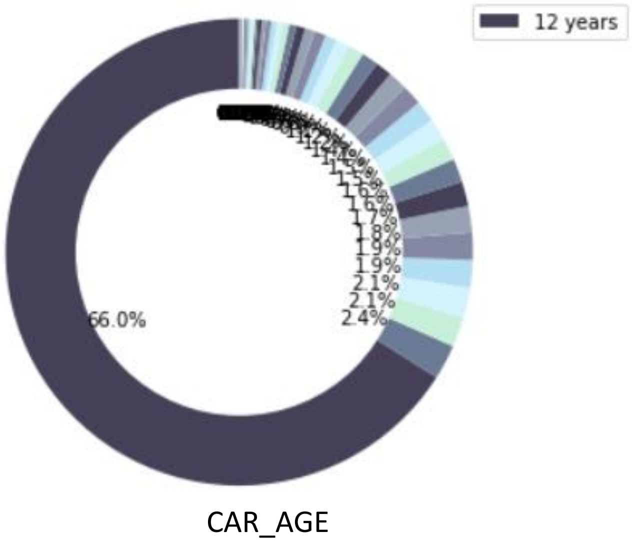
Exploratory Data Analysis

Question: Which customers mostly apply for loan?



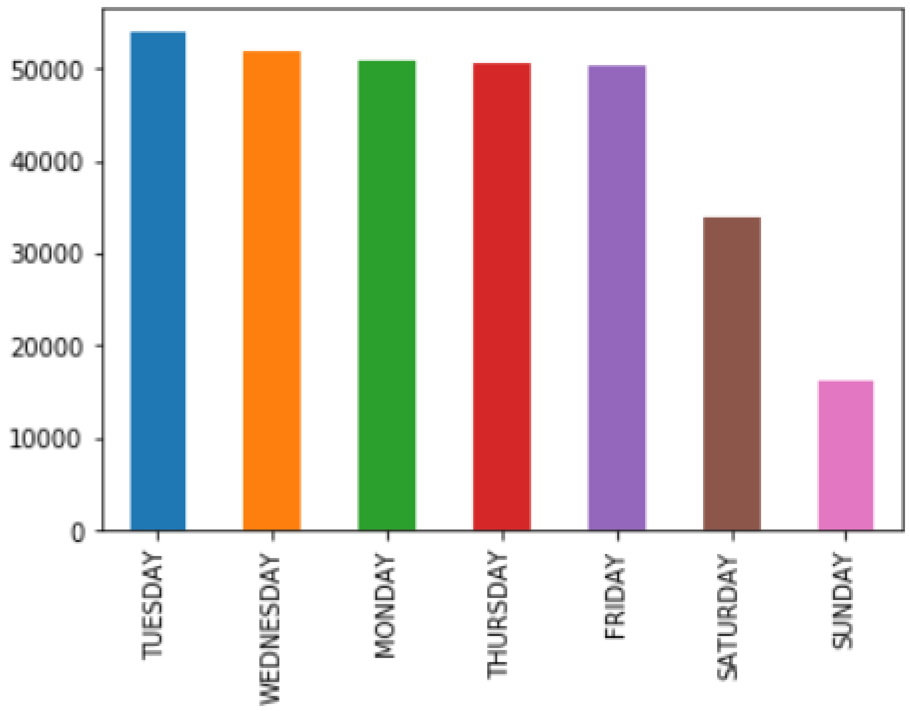


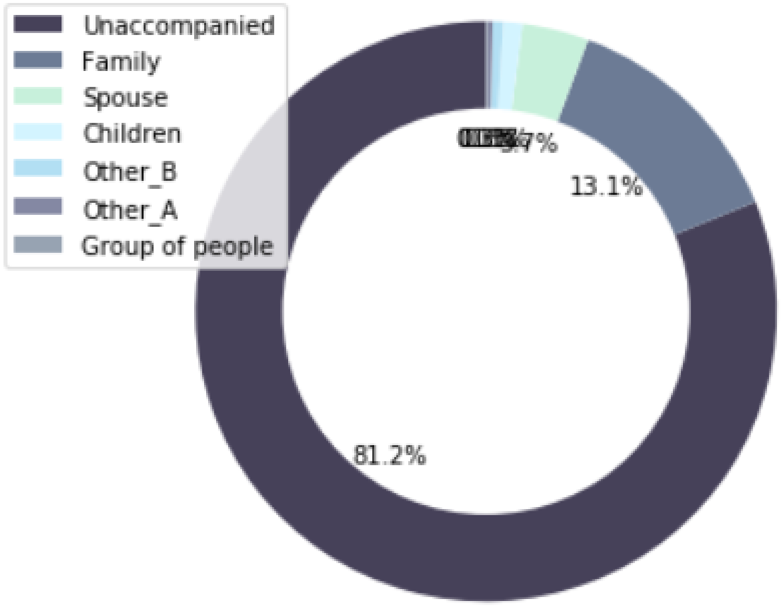




Exploratory Data Analysis

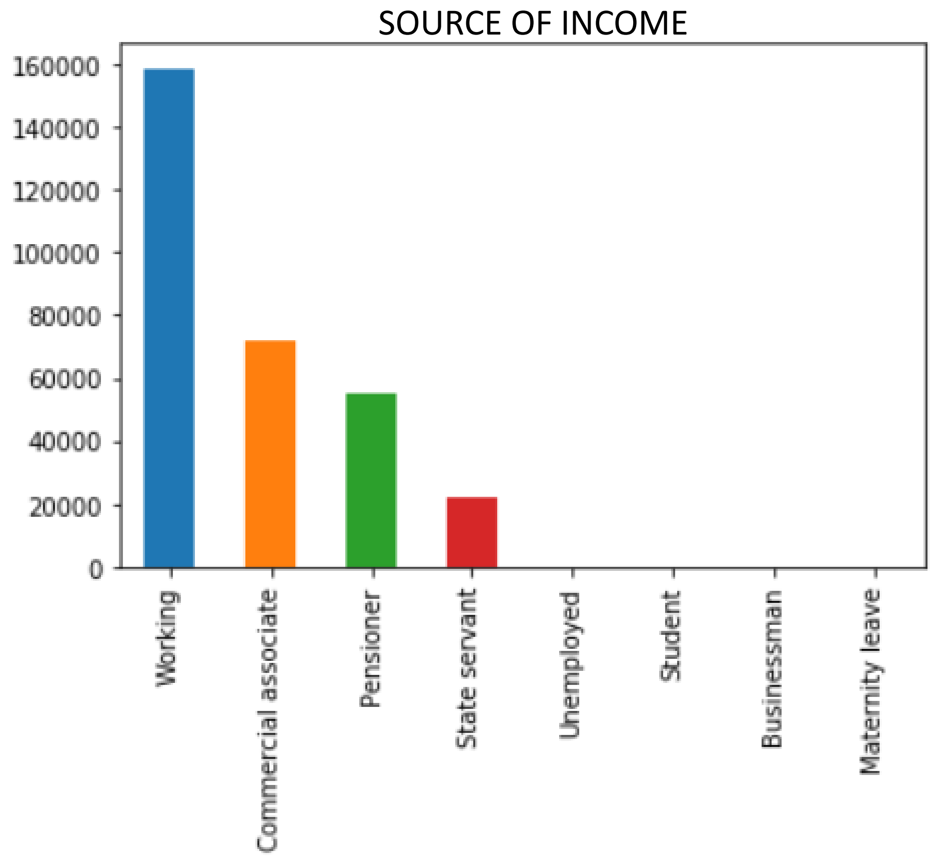
Question: Which day customers mostly apply for loan?

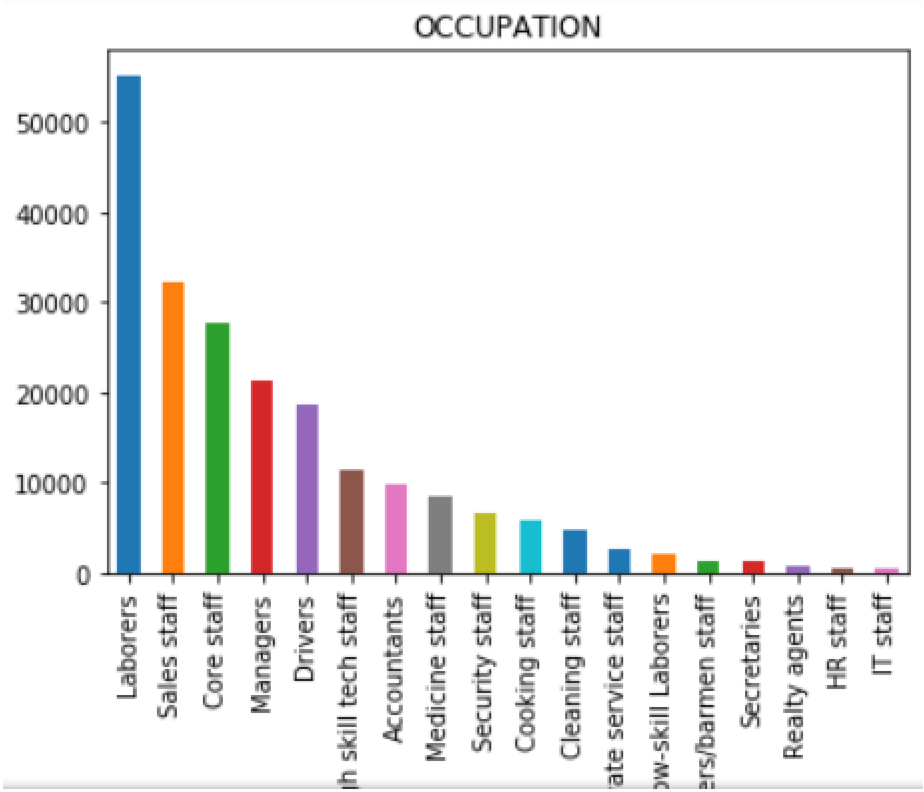




Exploratory Data Analysis

Question: **Which occupation mostly applied for the loan?**

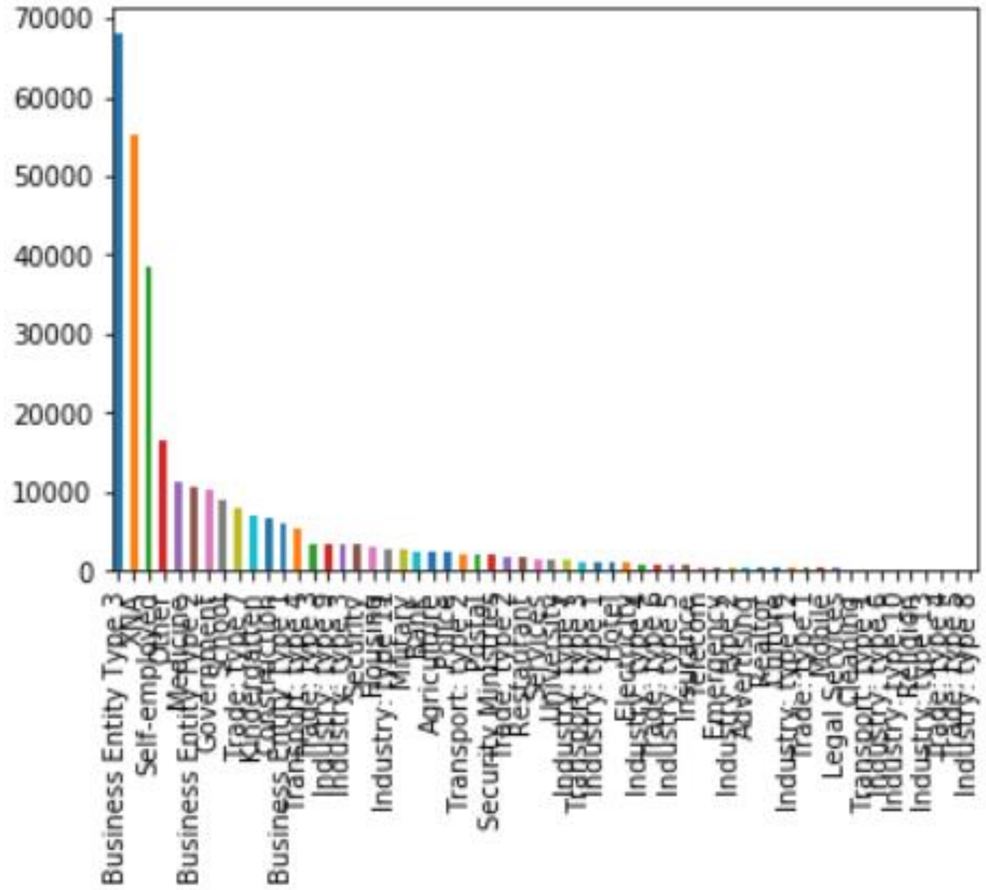




Exploratory Data Analysis

Question: **Which organization mostly applied for the loan?**

ORGANIZATION

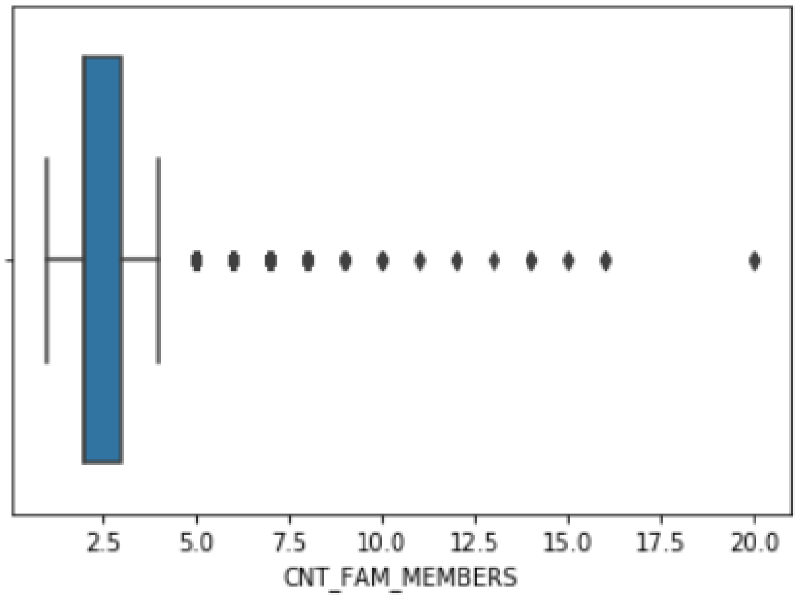


Missing values

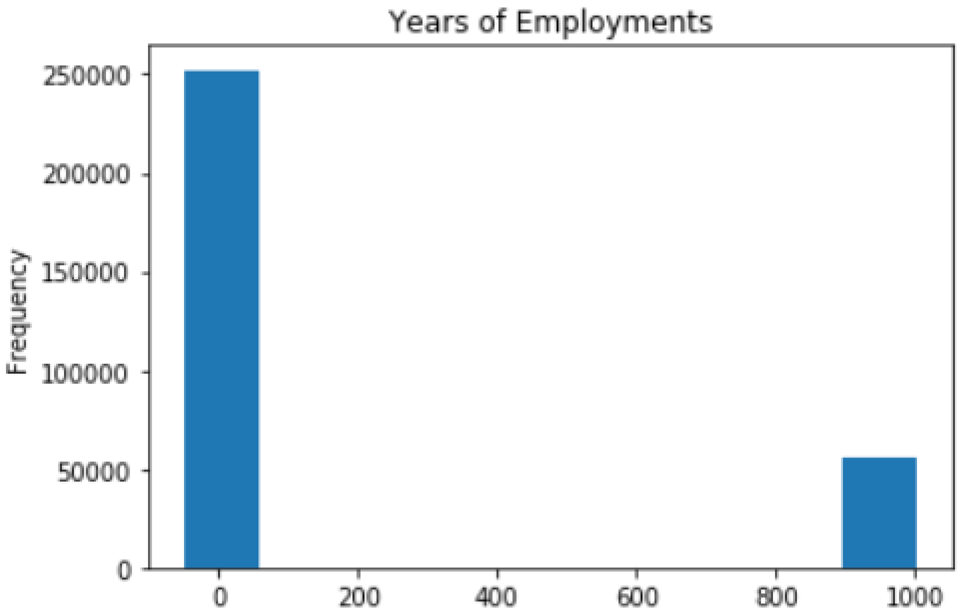
- 67 features have missing values

	Missing values	% of missing values
COMMONAREA_AVG	214865	69.87
COMMONAREA_MODE	214865	69.87
COMMONAREA_MEDI	214865	69.87
NONLIVINGAPARTMENTS_MEDI	213514	69.43
NONLIVINGAPARTMENTS_MODE	213514	69.43
NONLIVINGAPARTMENTS_AVG	213514	69.43
FONDKAPREMONT_MODE	210295	68.39
LIVINGAPARTMENTS_AVG	210199	68.35
LIVINGAPARTMENTS_MEDI	210199	68.35
LIVINGAPARTMENTS_MODE	210199	68.35
FLOORSMIN_AVG	208642	67.85
FLOORSMIN_MEDI	208642	67.85
FLOORSMIN_MODE	208642	67.85
YEARS_BUILD_MODE	204488	66.50
YEARS_BUILD_AVG	204488	66.50
YEARS_BUILD_MEDI	204488	66.50
OWN_CAR_AGE	202929	65.99
LANDAREA_AVG	182590	59.38
LANDAREA_MODE	182590	59.38
LANDAREA_MEDI	182590	59.38
BASEMENTAREA_MEDI	179943	58.52
BASEMENTAREA_MODE	179943	58.52
BASEMENTAREA_AVG	179943	58.52
EXT_SOURCE_1	173378	56.38
NONLIVINGAREA_MEDI	169682	55.18
NONLIVINGAREA_MODE	169682	55.18
NONLIVINGAREA_AVG	169682	55.18
ELEVATORS_MODE	163891	53.30
ELEVATORS_MEDI	163891	53.30
ELEVATORS_AVG	163891	53.30

Outlier



2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS



Years of Employments
Frequency
250000
200000
150000
100000
50000
0
0
200
400
600
800
1000

Dataset Characteristics

- Imbalance of classes – 1:9 ratio
- Poor recall in baseline model

Baseline Modeling

	precision	recall	f1-score	support
0	0.91	1.00	0.95	46041
1	0.20	0.00	0.00	4387
accuracy			0.91	50428
macro avg	0.56	0.50	0.48	50428
weighted avg	0.85	0.91	0.87	50428

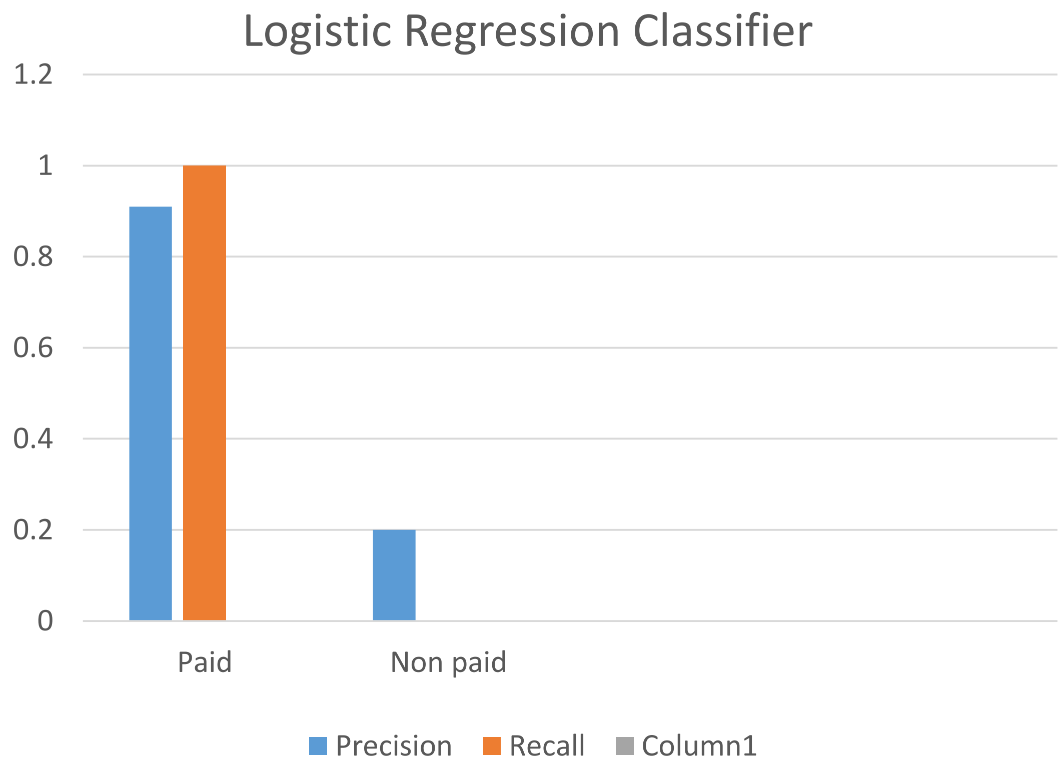
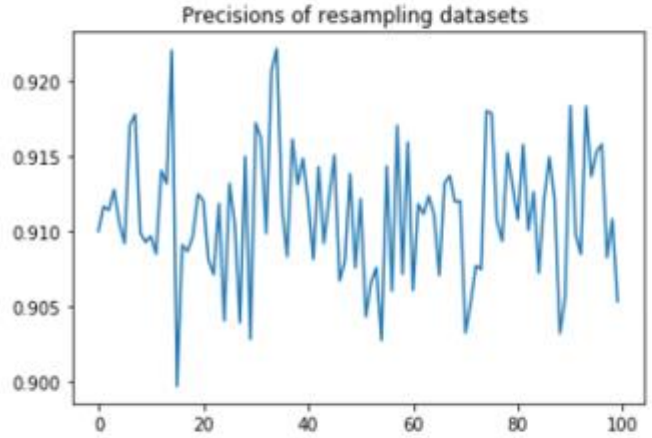


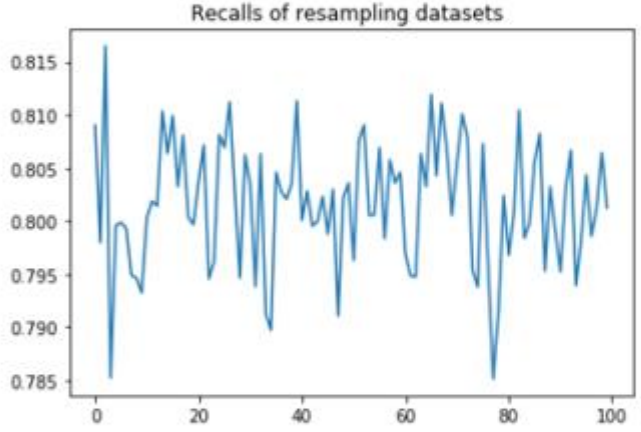
Table 1: Logistic Regression Classifier

Resampling

- To improve the gap between Precision and Recall

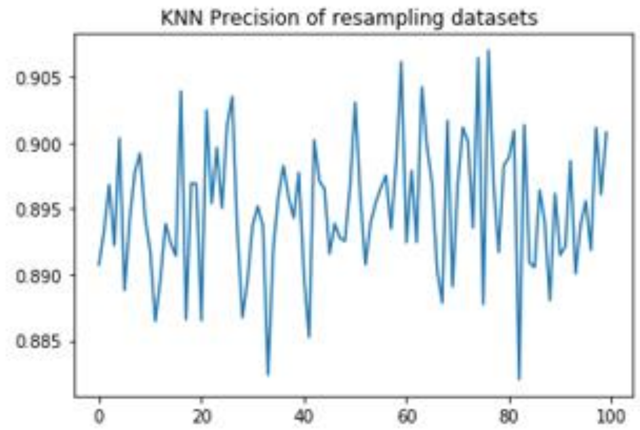
Under sampling



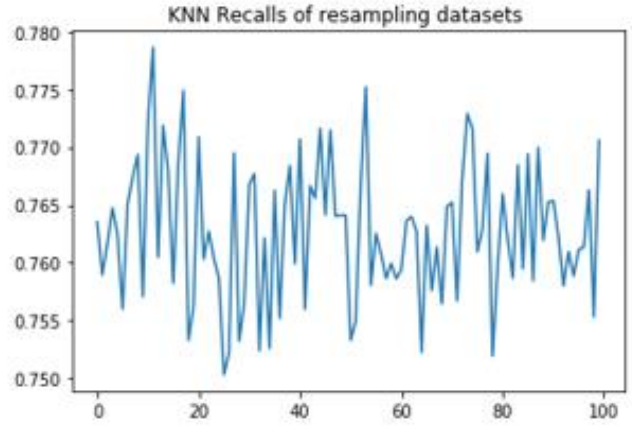


Logistic Regression

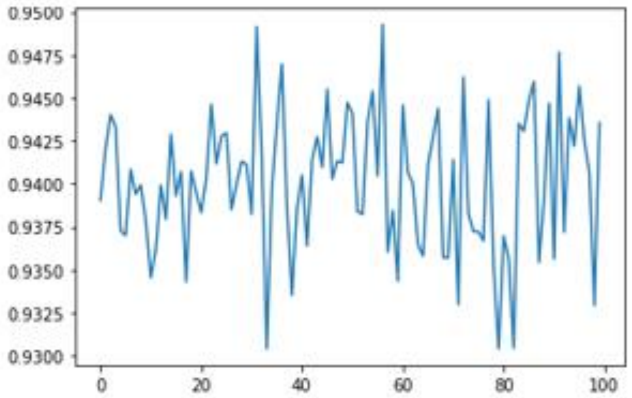
Average Precisions 0.895015570274



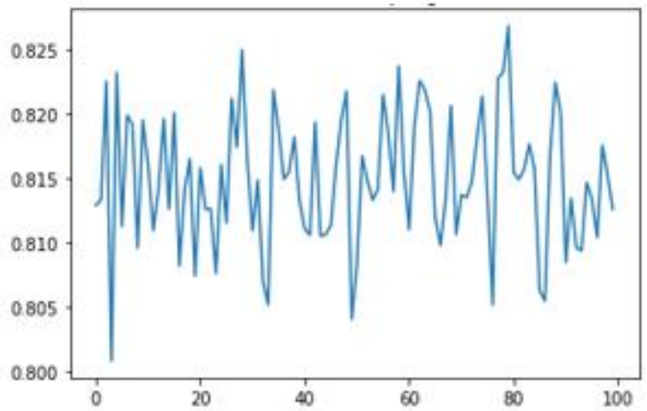
Average Recalls 0.762710377329



kNN



RF's Precision



RF's Recalls

Random Forest

Over sampling (SMOTE)

	precision	recall	f1-score	support
0	0.94	0.60	0.73	46098
1	0.12	0.61	0.21	4329
accuracy			0.60	50427
macro avg	0.53	0.60	0.47	50427
weighted avg	0.87	0.60	0.69	50427

Results

	Precision	Recall
Logistic Regression - Under sampling	0.91	0.80
Logistic Regression - Over sampling	0.87	0.6
kNN – Under sampling	0.895	0.763
Random Forest – Under sampling	0.895	0.815

Table 2. Results of classification models

Regardless of techniques used, there is a tradeoff between Precision and Recall.

Conclusions

- Explored and analyzed risk with a variety of models
- Further studies can be performed to analyze to improve the models e.g. cross validation, feature importance, dimension reduction, etc.

Recommendations

- The models are good for predicting the loan risk analysis with +60% of improvement
- The remaining applicants' data should be further analyzed by financial experts
- The model should be used in combination with human analysis for decision making. They cannot replace the fully decision making, but aids in decision making workflow.
- For better precision, more relevant features will be needed.