Supervised Learning Capstone

Loan Risk Assessment

By Sebastian Rosado

1

Introduction

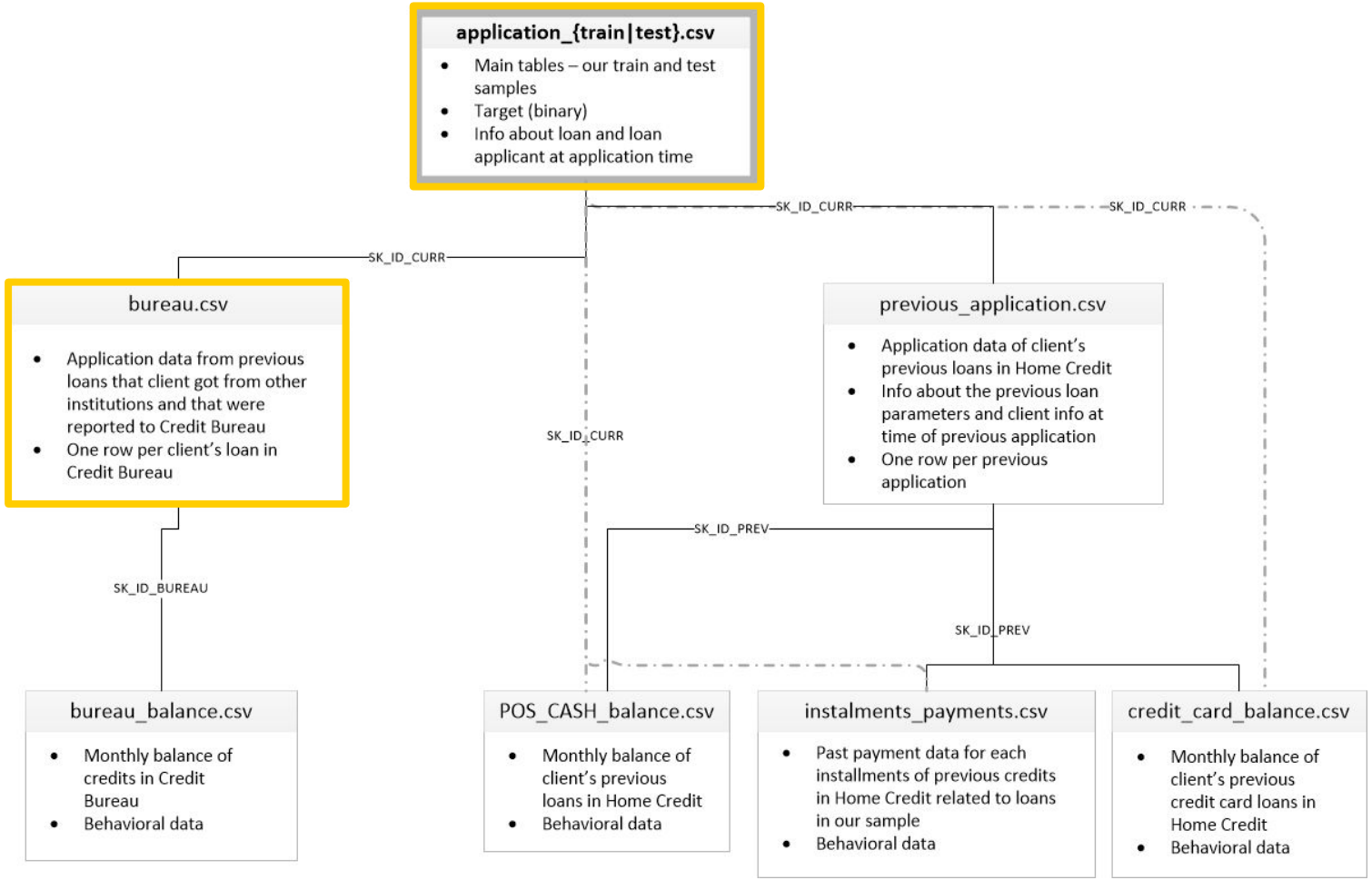
Motivation, Data Details

Motivation

- What variables are the most predictive of a client's difficulty to pay back their loans?
- Banking the unbanked
  - Machine learning + finance
- Real data
  - Direct application
  - Challenging
- Interesting, unconventional variables
  - Apartment sizes, social circle creditworthiness

Description of the Data

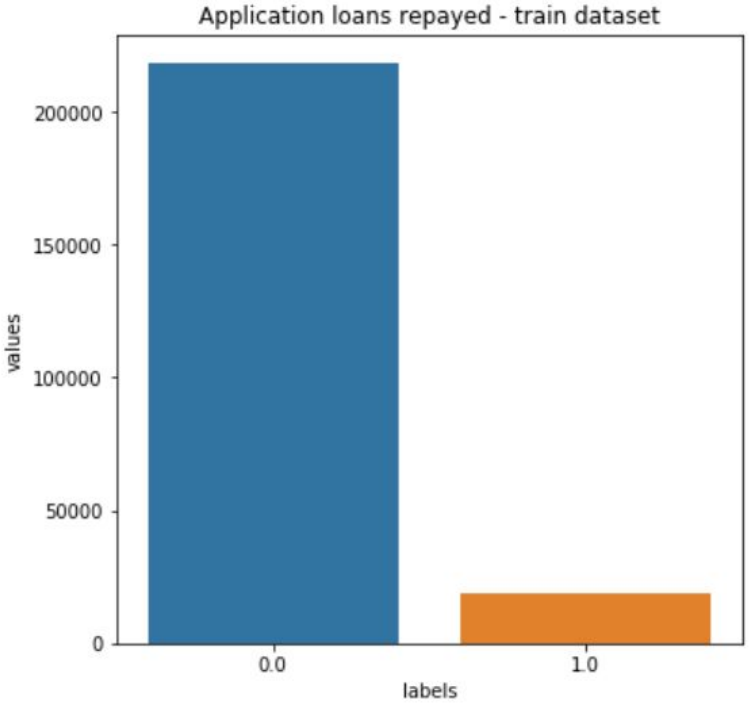
- Home Credit → non-bank financial institution
  - Applicants with little / no credit history
- Two merged DataFrames (one-to-one)
  1. Clients with active Home Credit loans
  2. All previous loans for those clients
- 236,630 rows, 194 columns
- Target: Payment Difficulties (1/0)


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_CURR
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

2

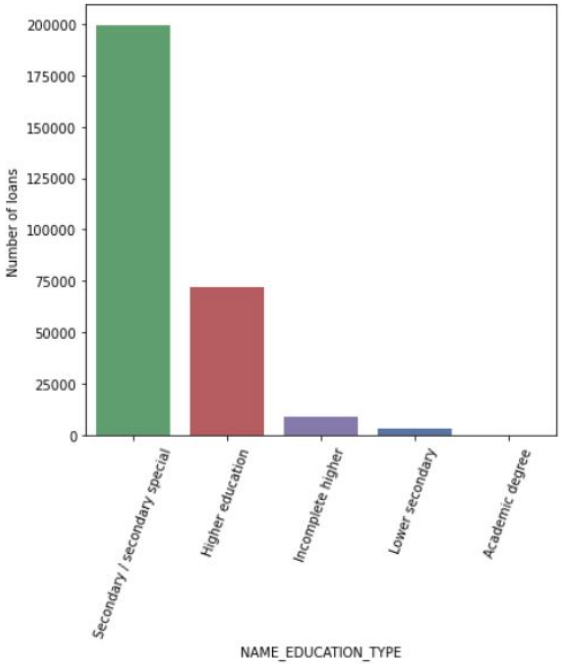
Data Exploration & Cleaning

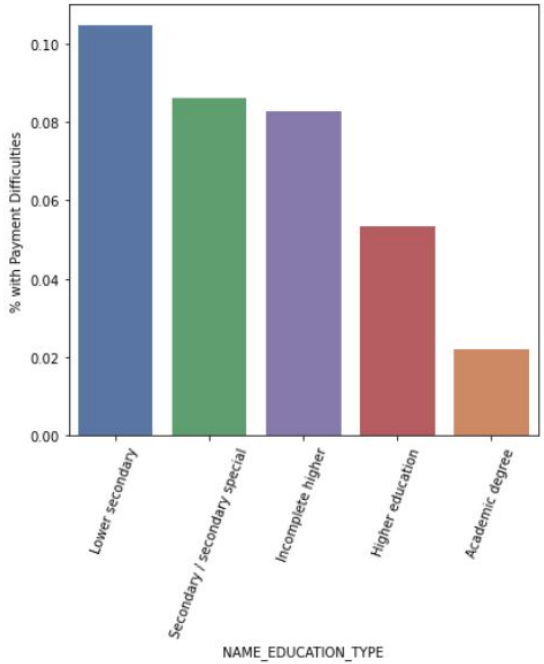
Target Variable



Educational Achievement

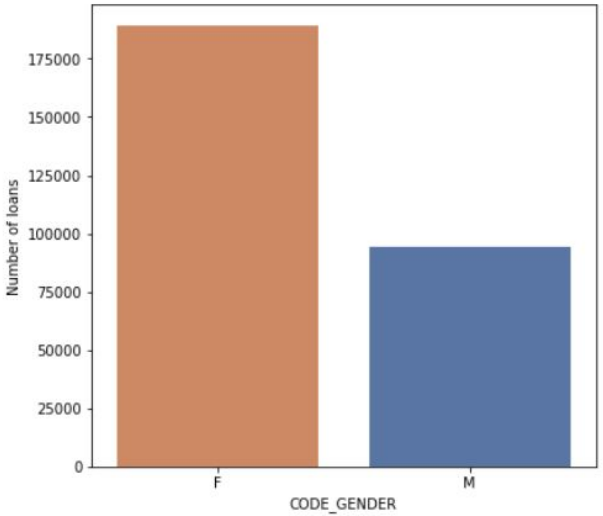
```
Description of NAME_EDUCATION_TYPE: 13    Level of highest education the client achieved
Name: Description, dtype: object
```

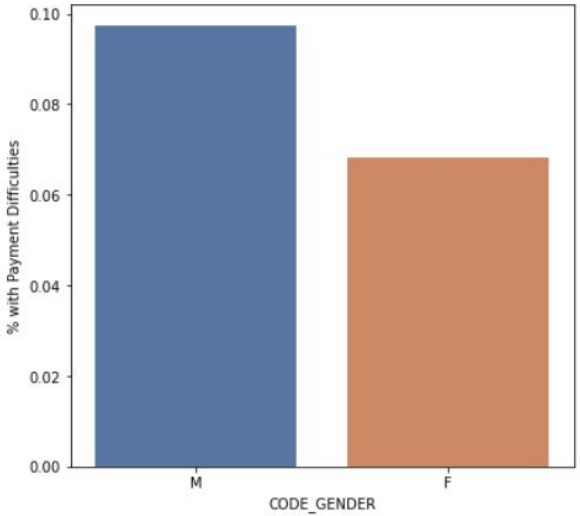




Gender

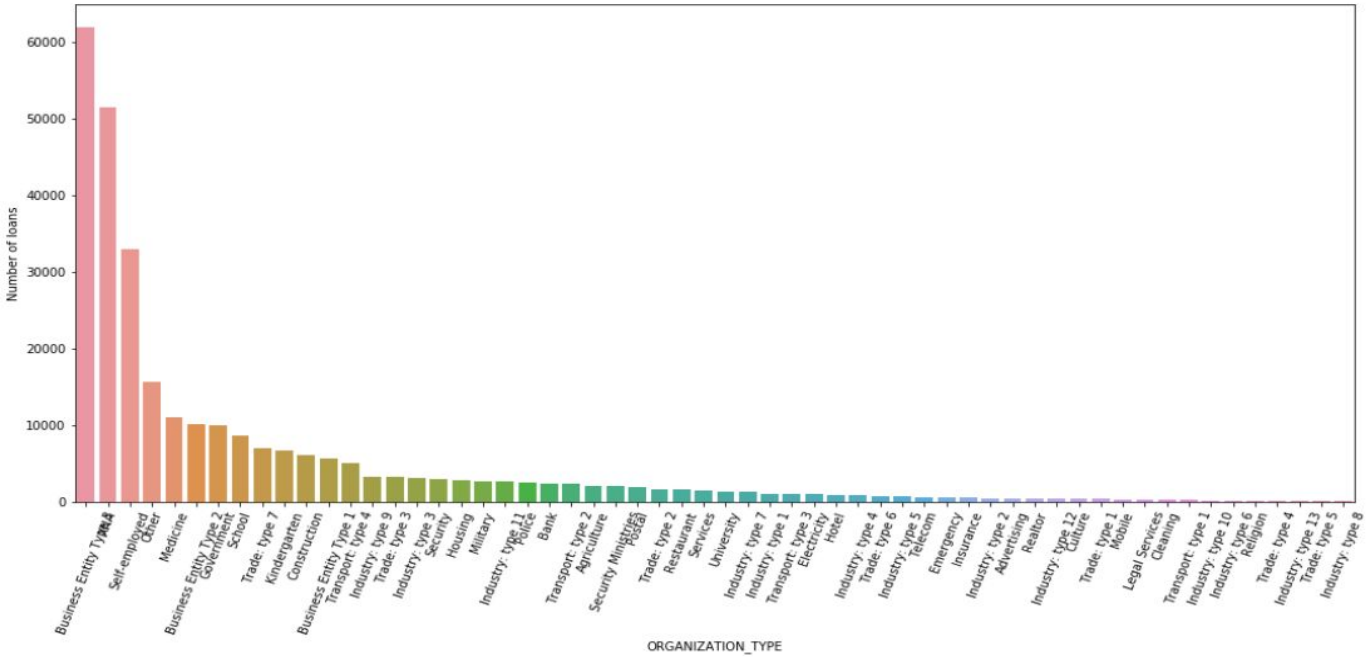
```
Description of CODE_GENDER: 3    Gender of the client
Name: Description, dtype: object
```



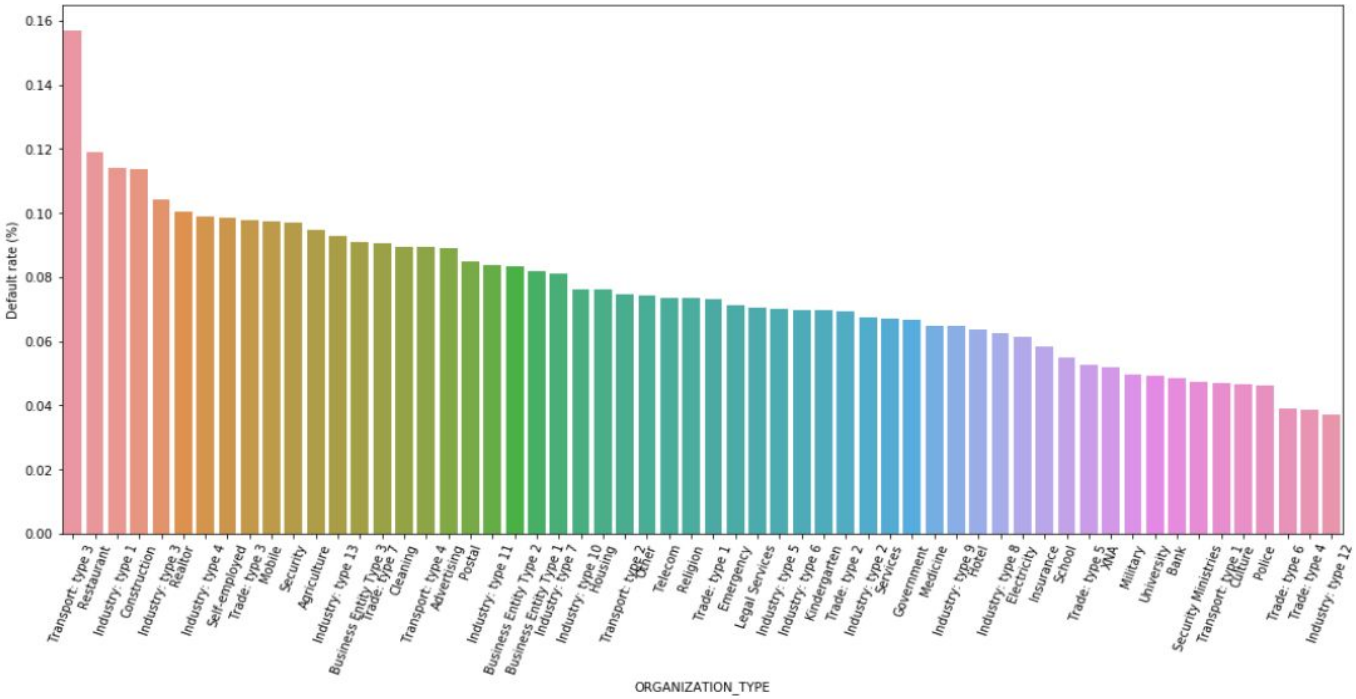


Organization Type

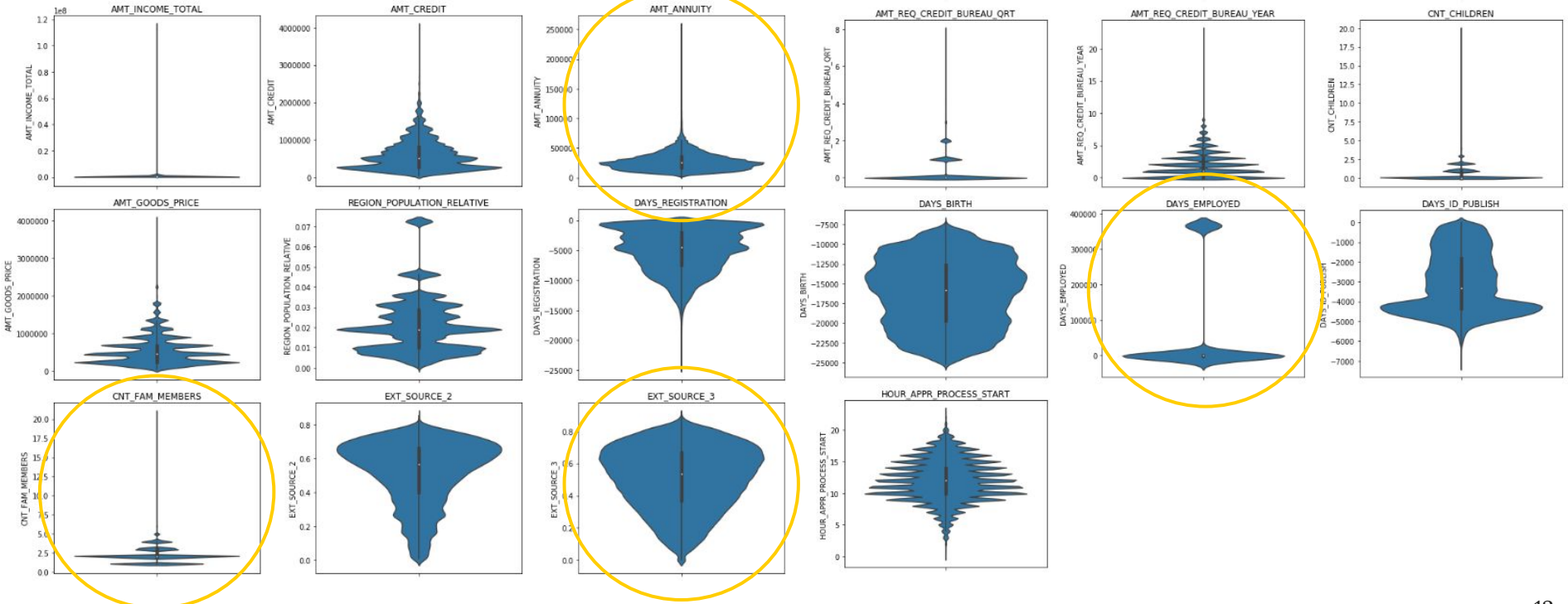
```
Description of ORGANIZATION_TYPE: 40    Type of organization where client works
Name: Description, dtype: object
```



Organization Type



Continuous & Discrete



3 Feature Engineering

Created Features

- Third Degree Polynomial Transformations:
  - 'EXT_SOURCE_2', 'EXT_SOURCE_3', 'DAYS_BIRTH', 'bureau_DAYS_CREDIT_mean'
- Financial Variables
  - Annuity payment as % of income (kept)
  - Credit payments as % of income
  - Credit overdue as % of total credit
- PCA
  - Tried four, kept one (['AMT_INCOME_TOTAL', 'AMT_ANNUITY', 'AMT_CREDIT', 'AMT_GOODS_PRICE'])

4

Modeling & Tuning Process

Modeling Process

	SMOTE-Balancing	Imbalanced Original
Random Forest	Y	Y
Logistic Regression	Y	Y
XGBoost	Y	Y
Dimen.- Reduced XGBoost	Y	N
Dimen.- Reduced RF	Y	N

5 Results

Metric Comparison - Recall

	SMOTE-Balancing	Imbalanced Original
Random Forest	87.75%	0.0%
Logistic Regression	54.41%	0.01%
XGBoost	92.86%	0.17%
Dimen.- Reduced RF	84.62%	N
Dimen.- Reduced XGBoost	90.2%	N

6 Conclusions

Key Takeaways and Lessons Learnt

Key Takeaways

- Most Important Variables:
  - Completed Higher Ed. (26.37%)
  - Owns phone (7.95%)
  - Applied on Sunday (5.88%)
  - Active external loans (5.64%)
  - Owns a car (5.51%)
- Best Model (judging efficiency vs. score)
  - Dimen. Reduced XGBoost

Lessons Learned

- SMOTE increases recall, though sometimes at the expense of accuracy
- Averaging feature importances reduces blind spots (don't trust a single algorithm)
  - This can be used for efficient dimensionality reduction
- Visualization functions have a high ROI
- Good data science is a marathon, not a race

Room for Improvement

- Precision-recall curves
- More feature engineering
- Add other models (e.g. SVM & KNN)

7

Thank You

Data Source

- https://www.kaggle.com/c/home-credit-default-risk