Bank Loan Default Risk

Purpose

Loan and Financial Intelligence

01

What is the burning platform?

Difficulty in approving those who need loans

02

Why does this matter?

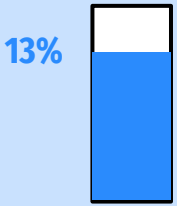
51.3% of Americans have taken a loan in their lifetime

03

What can this solve?

Loans to people in need and less risk for business

Background Research and Hypothesis



Total Balance of Personal Loans in the U.S.

$162M

$76 B

Fraudulent Loans

15% of Paycheck Protection Program Loans Est. Fraud

- Behavior features will affect target more than characteristic features

Data cleaning & Pre-processing

01

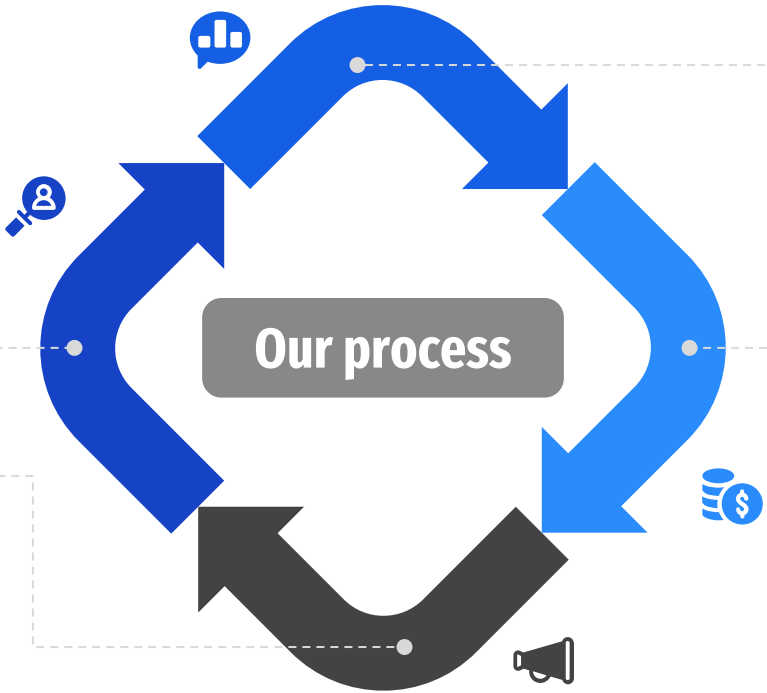
Missing Values

Filtered columns with more than 40% Null Value

02

Outliers

Capping outliers was performed



03

Feature Engineering

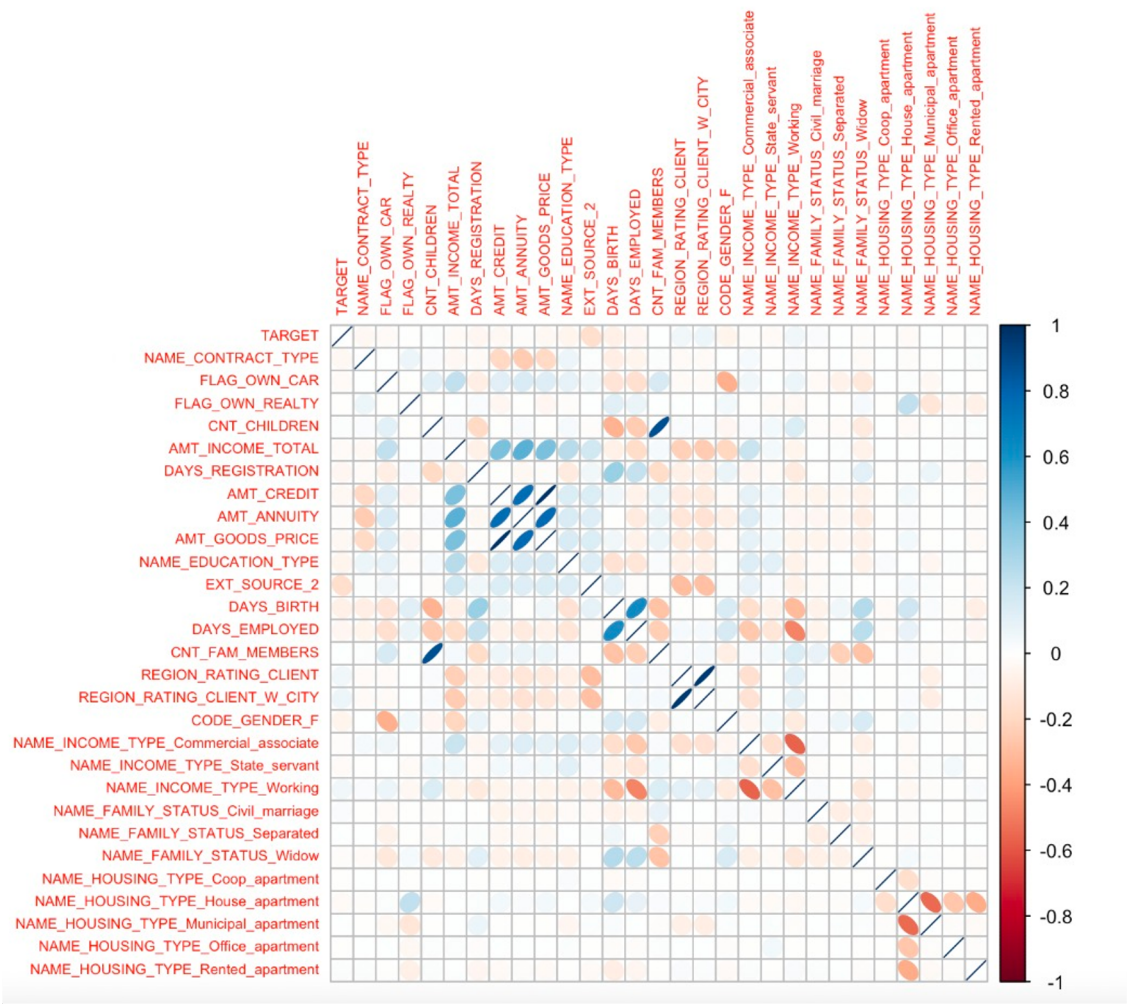
Filtered variables using correlation

04

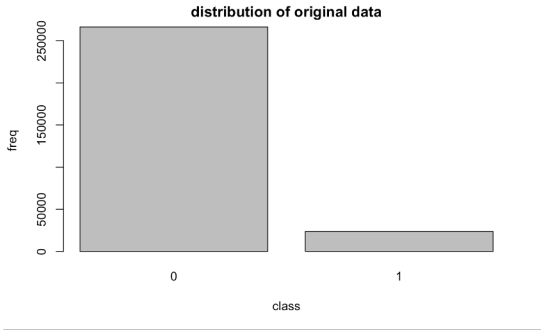
Create Dummy Variables

Create dummy and ordinal variables for categorical data

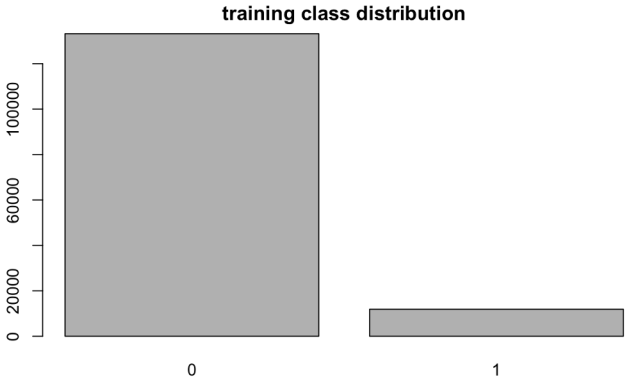
Attribute Correlation



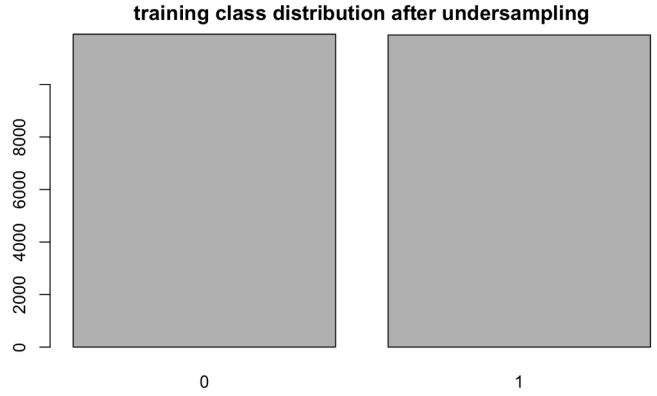
Unbalanced Dataset



- Class 0: over 250,000
- Class 1: about 27,000
- Split train and test 50/50



Undersampling



Model Benchmarks

Models	PPV (training/testing)	NPV (training/testing)	Overall accuracy (training/testing)
Logistic Regression	0.64/0.15	0.65/0.95	0.65/0.67
Classification Tree	0.65/0.14	0.65/0.95	0.65/0.66
Naive Bayes	0.61/0.12	0.63/0.95	0.59/0.59
KNN	NA/0.09	NA/0.93	NA/0.53
Random Forest	0.82/0.10	0.81/0.93	0.81/0.57

	Predicted	
Actual	0	1
0	88984	44218
1	4380	7503

	Predicted	
Actual	0	1
0	88316	44886
1	4709	7174

	Predicted	
Actual	0	1
0	78344	54858
1	4158	7725

	Predicted	
Actual	0	1
0	70785	62417
1	5403	6480

	Predicted	
Actual	0	1
0	76496	56706
1	5425	6458

Logistic Regression

Significant Variables	Coefficient
NAME_CONTRACT_TYPE	0.6425
FLAG_OWN_CAR	0.7733
DAYS_REGISTRATION	0.9999
AMT_GOODS_PRICE	0.9999
NAME_EDUCATION_TYPE	0.8036
EXT_SOURCE_2	0.1138
DAYS_BIRTH	0.9999
DAYS_EMPLOYED	0.9999
REGION_RATING_CLIENT	1.1714
Pre_approved_num	0.9170
Pre_canceled_num	1.0428
Pre_refused_num	1.1243

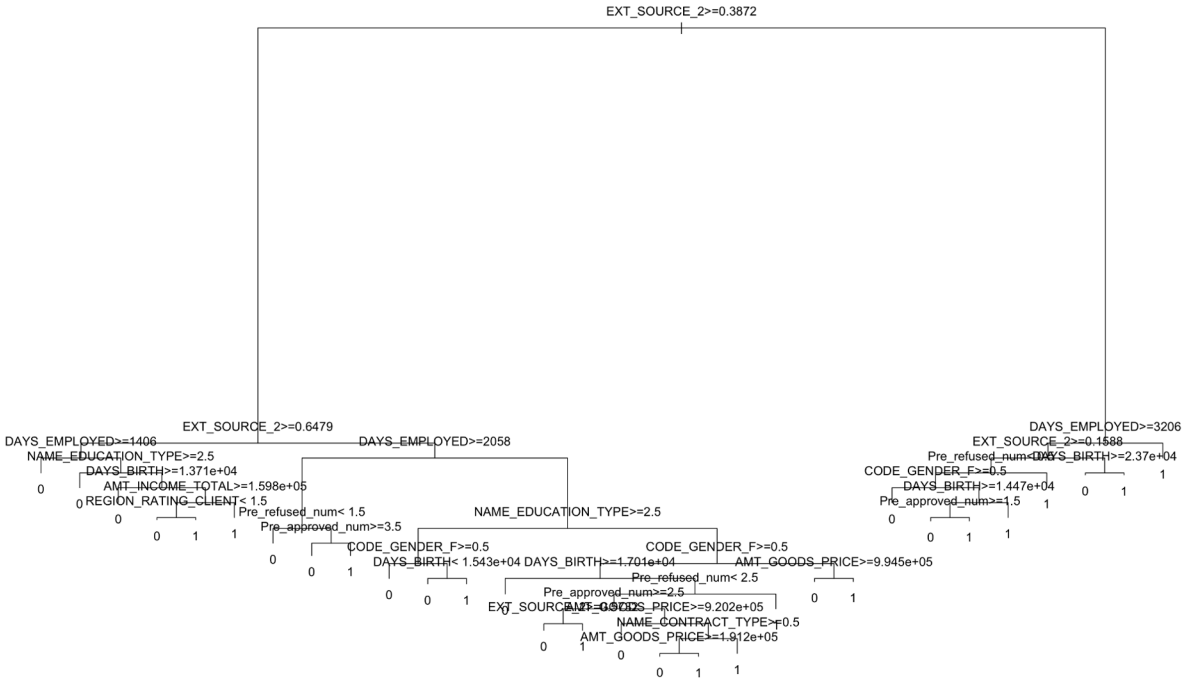
Accuracy

Overall: 0.665
Class 0: 0.9531
Class 1: 0.1451

	Predicted	
Actual	0	1
0	88984	44218
1	4380	7503

Previous application information may have significant influence.
Region_rating_client may reflect how clients' geographic related to money flow.
Ext_source_2 is extremely relevant (credit score)
Name_contract_type: revolving loans decreases the odds of defaulting
Owning a car decreases the odds of late payments too

Classification Tree



Accuracy

Overall: 0.6582
Class 0: 0.9494
Class 1: 0.1378

Actual	Predicted 0	1
0	88316	44886
1	4709	7174

Key Findings

Actionable Insights

Name_Status_Contract
Status of previous application

Days_Employed
days before the application applicant started working

Own_Car
Did the applicant own a car (1/0)?

01 02 03 04 05 06

Ext_Source_2
Credit score from an external source

Contract_Type
What kind of loan the application is for

Education_Type
Applicant's highest level of education

Business Recommendations

Actionable Insights

Data collection and governance to benefit data understanding

Different use cases dependent on preferred performance of target class

Expand applicant profiles utilizing external sources Credit Score

Conclusion

Purpose	• Helping banks select qualified applicants that won't default the loan and avoid losing money
Takeaways	• Different group of people may have different prediction models, more classification based on clients features may improve model performance
Lesson Learned	• Unbalanced dataset needs to be resampled. • Using sampling methods on an imbalance dataset could generate bad results on a class accuracy, since the imbalance might be reflected in the real world but not on our trained model (depending on the proportion you choose)