Credit Default Risk

Capstone 1

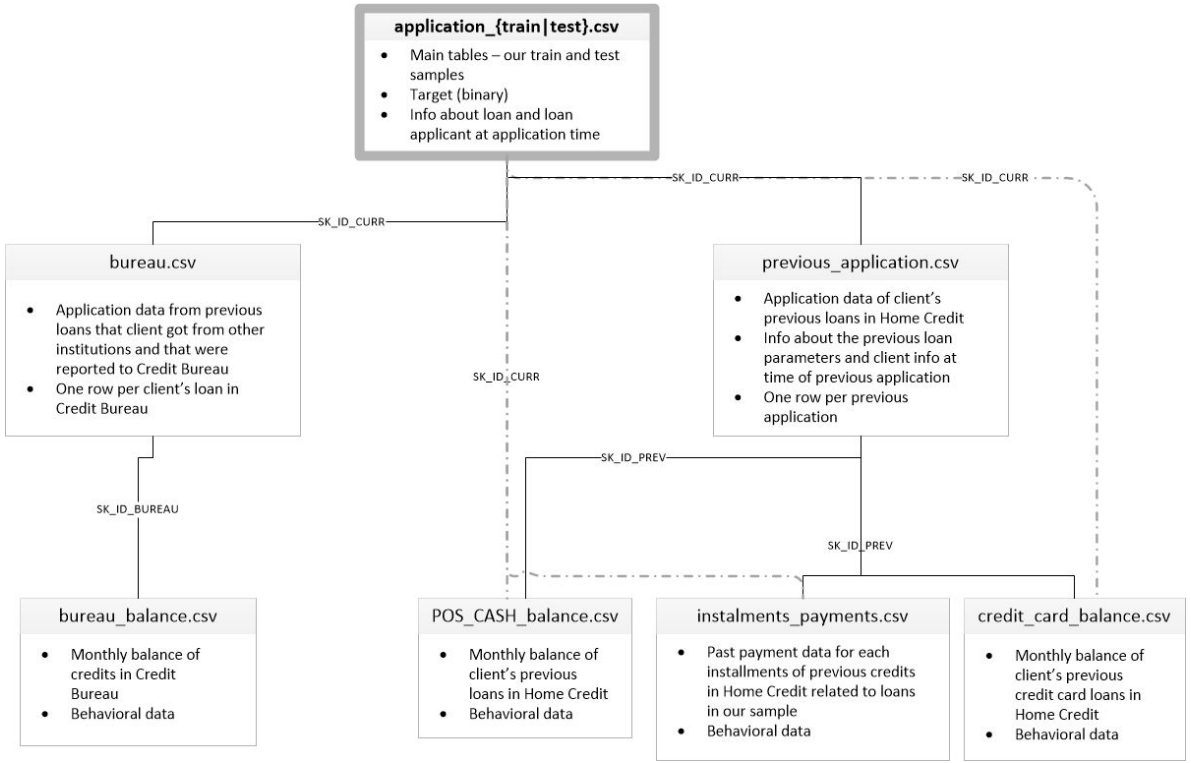
By: Eman Enain

PROBLEM

Applicants who don’t have credit history are often taken advantage off by paying very high interest rates. Are we able to identify whether an individual can pay or fail to pay their loan based on different factors such as: age, income level, education level and years of employment.

DATA OVERVIEW

- The data from Home Credit Default Risk Kaggle competition will be used for this capstone. The data contained application for 307,000 applicants and 122 features.



DATA CLEANING

Missing Data and Solution

- The dataframe has122 columns
  - 49 columns contained between 48% and 69% null values
  - The mean value was imputed for numerical values
  - The mode was imputed for categorical variables

OWN_CAR_AGE

Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct	62
Distinct (%)	0.1%
Missing	202927
Missing (%)	66.0%
Infinite	0
Infinite (%)	0.0%
Mean	12.06112067

APARTMENTS_AVG

Real number ($\mathbb{R}_{\geq 0}$)

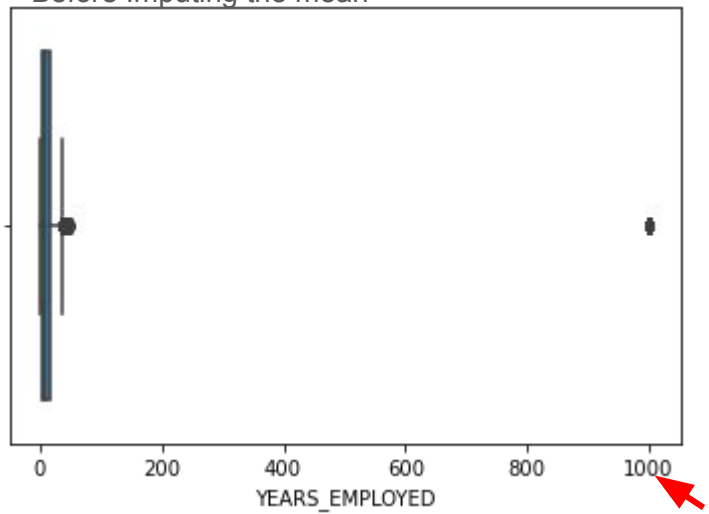
MISSING

Distinct	2339
Distinct (%)	1.5%
Missing	156060
Missing (%)	50.8%
Infinite	0
Infinite (%)	0.0%
Mean	0.1174420226

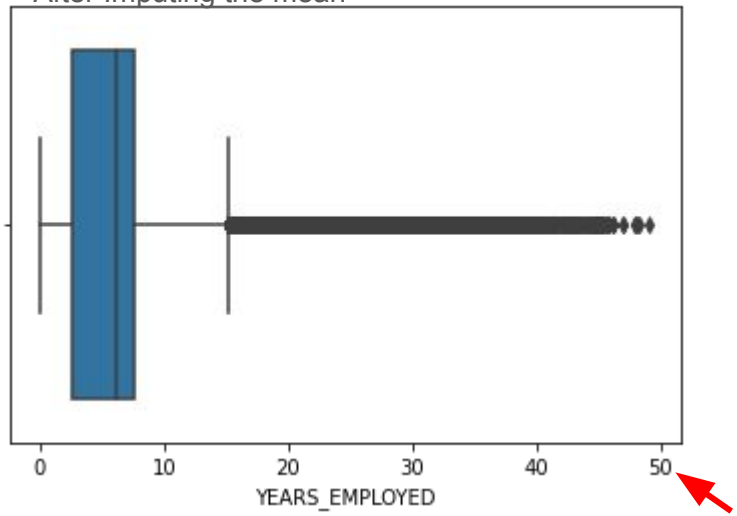
OUTLIERS: Data Processing Error

- Years_employed contained 55374 observations where the years employed is 1000.7 years.
- The median value for years employed (without the 1000.7 yrs) is 6.1 years - this value was replaced with 1000.7 for all 55374 observations.

Before Imputing the mean



After Imputing the mean



Data Subsets: Do more features mean better predictions?

- According to a domain expert many of the features provided are not usually required by financial institutions.
- Created 2 subsets of the data:
  1. The first: with the original number of features (122)
  2. A subset of the data was created with less features (53)

EXPLORATORY DATA ANALYSIS

Imbalanced Data

Over ~92% of applicants successfully pay the loan back, while 8.1% fail to repay

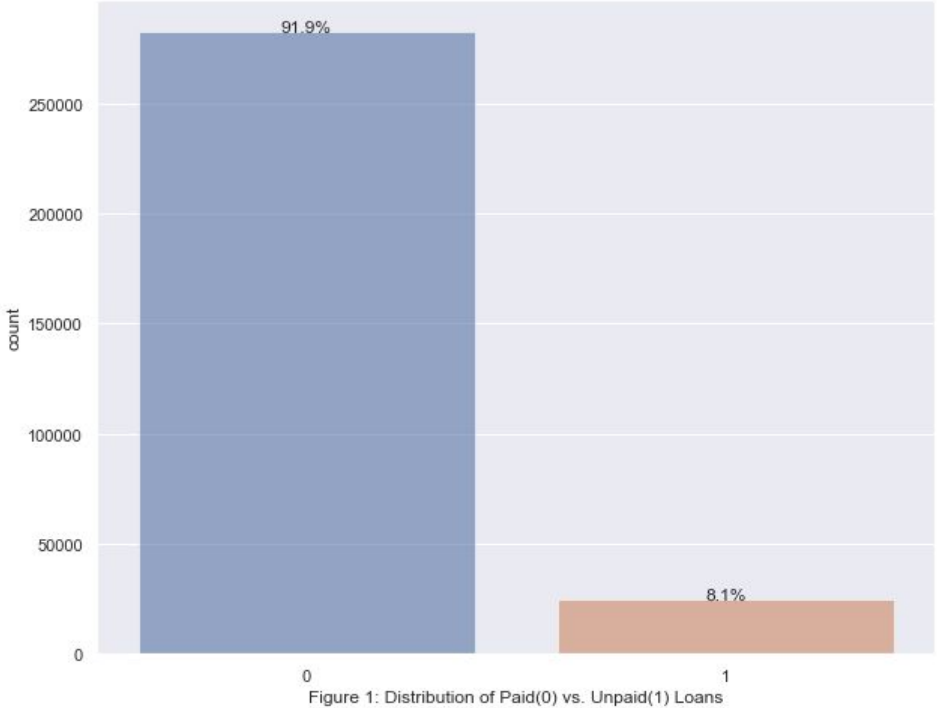


Figure 1: Distribution of Paid(0) vs. Unpaid(1) Loans

- Females are double the male applicants
- Males have a higher rate of defaulting on loans than females.

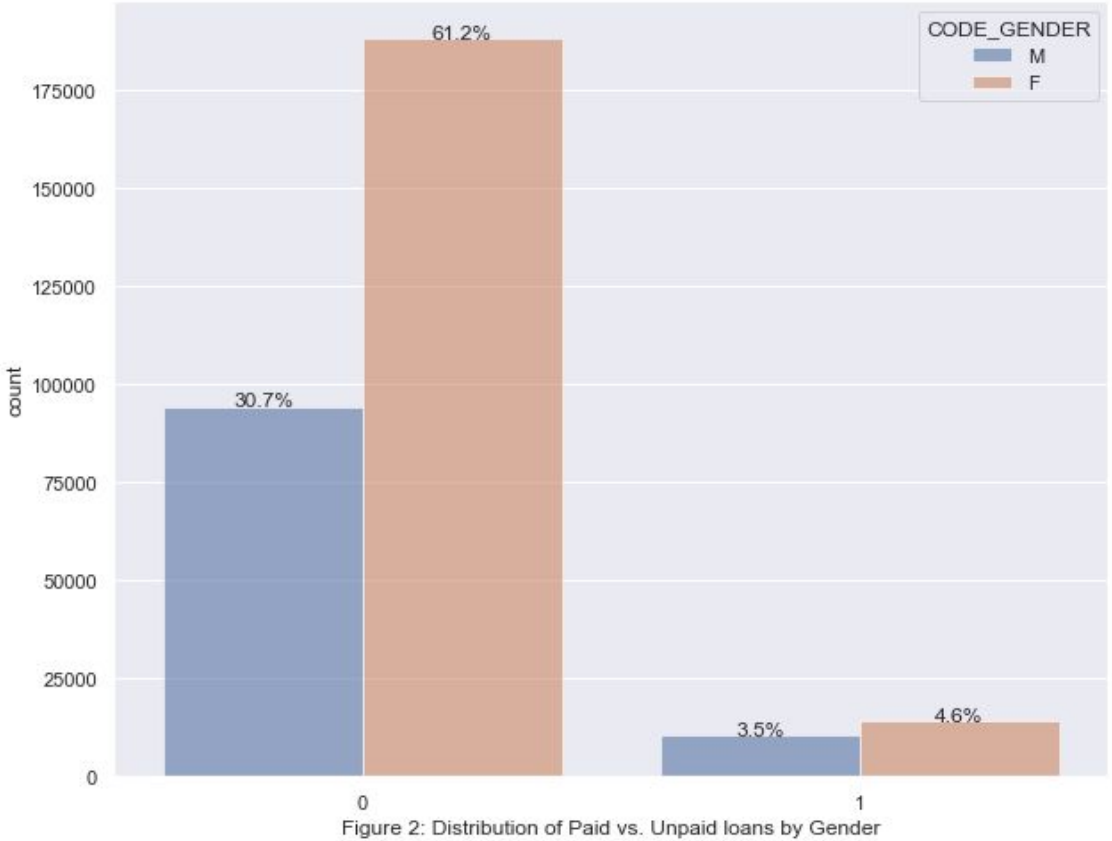
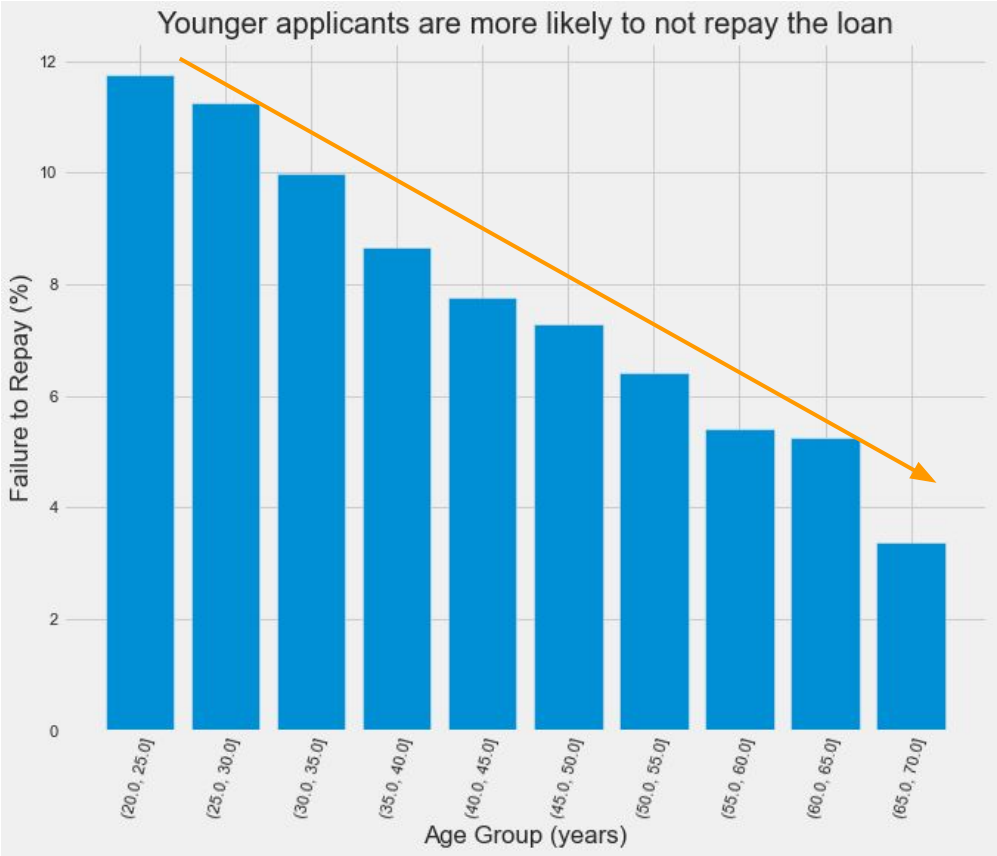
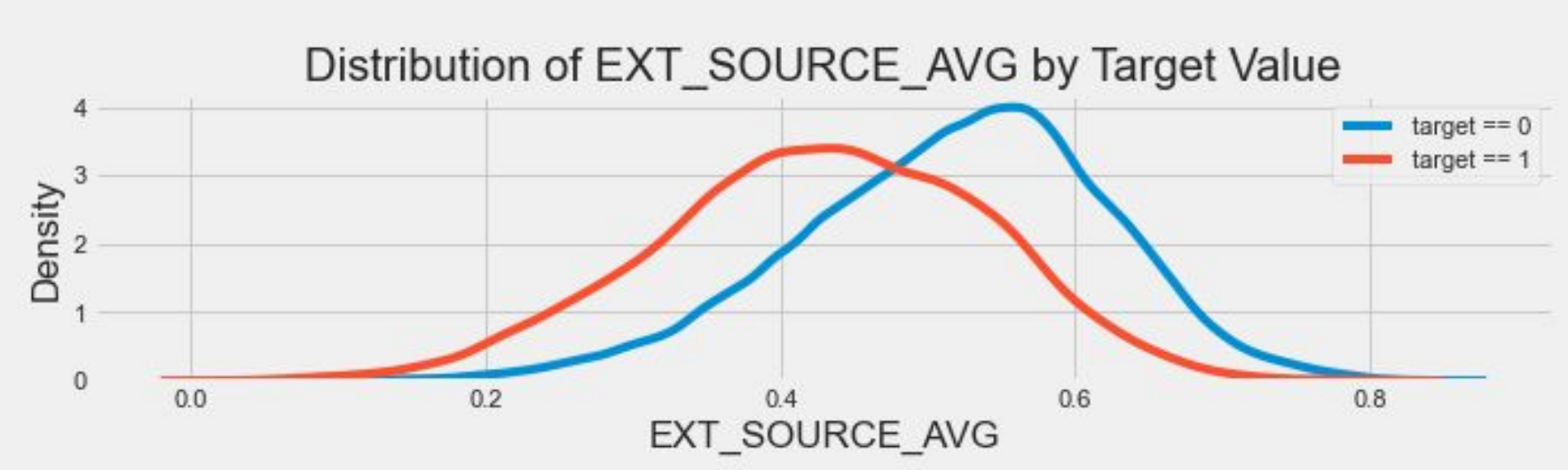


Figure 2: Distribution of Paid vs. Unpaid loans by Gender

Younger Applicants are more likely to default on a loan that older applicants:



As the external credit score increases, failure to repay a loan significantly decreases.



Modeling Algorithms

Two different datasets were used on Gradient Boost and Random Forest.

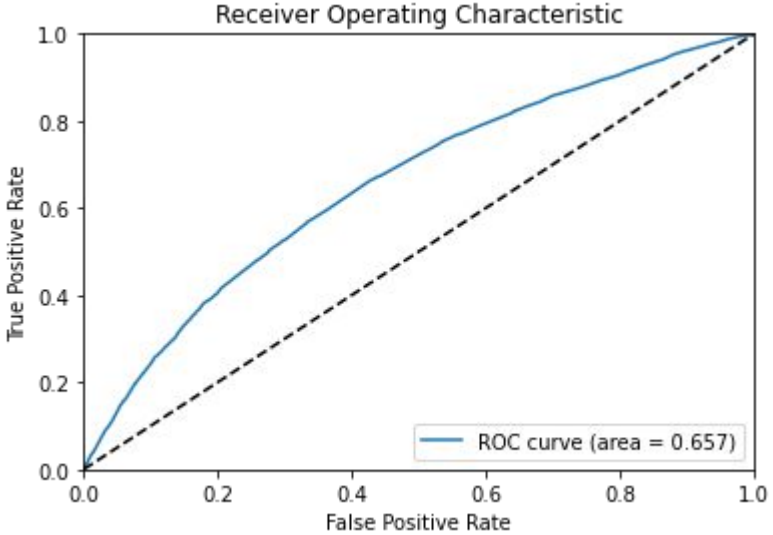
- The first dataset had all the original columns (122)
- The second dataset had a great reduction in the number of columns (53)

	All Features Accuracy	Reduced Features Accuracy
Gradient Boost	61.6%	59.4%
Random Forest	67.9%	68.2%

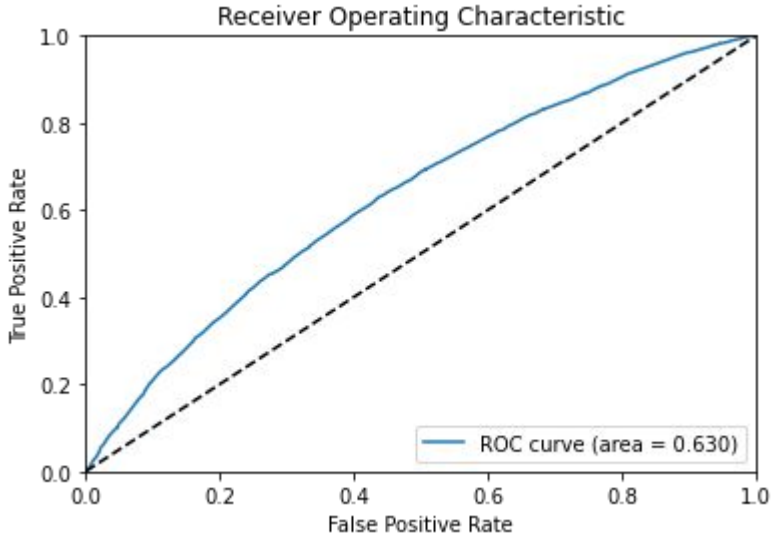
- The BEST performing model is the Random Forest model with reduced columns.

ROC CURVE - Gradient Boost

ALL FEATURES



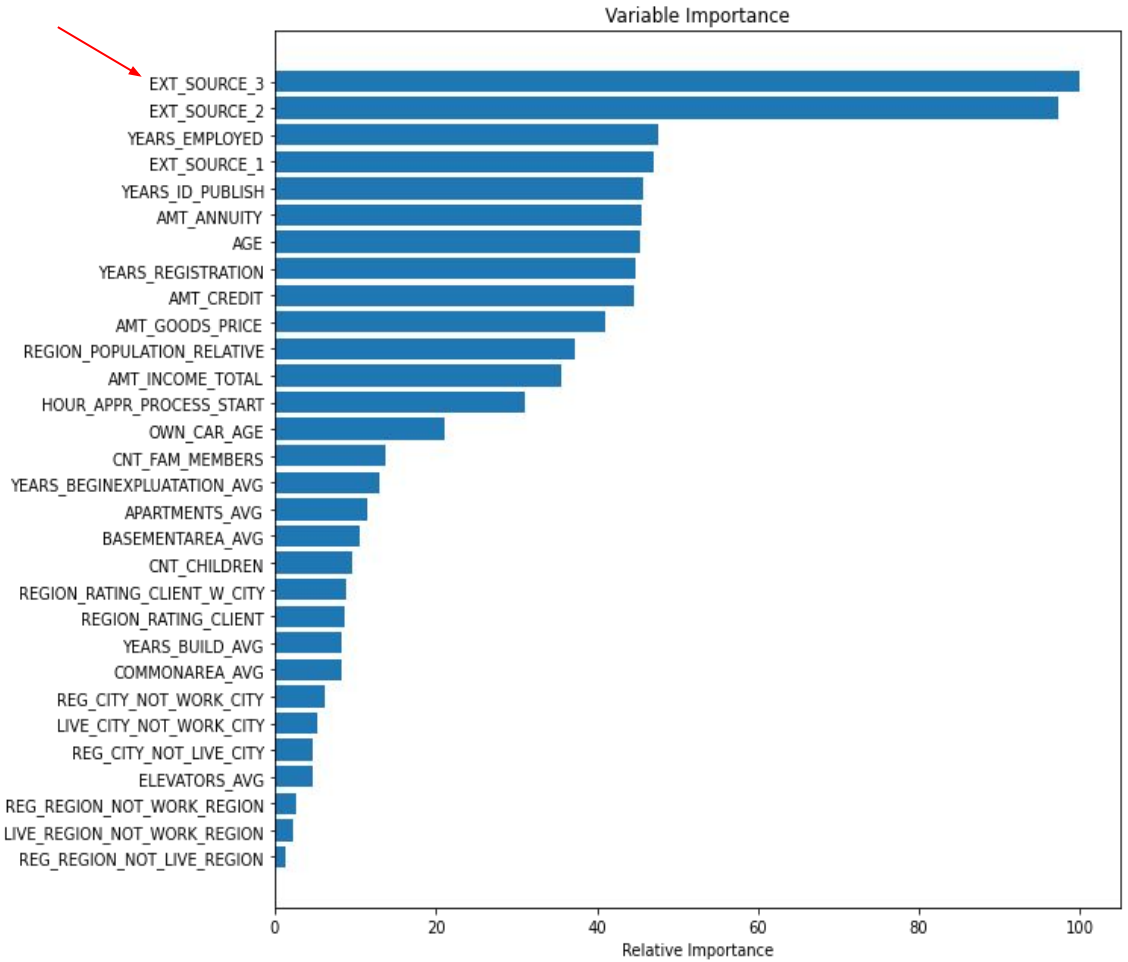
REDUCED



CONCLUSION BASED ON MODELS

- It’s difficult to predict whether an applicant will default on a loan without using existing credit score - EXT_SOURCE variables had the strongest negative correlation.

- Those variables also had close to 100% for random forest feature importance



FUTURE IMPROVEMENTS

- Create a user interface where users are able to enter their information (age, income, total owing credit, number of active loans, years of employment) and get a prediction of the likelihood of defaulting on a loan.
- Filter the datasets and remove EXT_SOURCE and check performance.
- Use the dataset with added engineered features and check whether the accuracy increases or decreases.