Home Credit Default Risk

Rashi Nigam 08-14-2018

Overview

- Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.
- Founded in 1997, Home Credit Group is client for this project and an international consumer finance provider with operations in 10 countries

Why the Need

- Help better the low income group to get loans who struggle due to insufficient or non-existent credit histories

- Provide a positive and safe borrowing experience to broader community

- Whether influence is in direct relation or inverse relation to TARGET prediction

- Help clients realize their dreams and ambitions in a financially responsible way

- Encourage economic development through supporting domestic consumption, thereby improving living standards

Project Objective

The client Home Credit Group needed a research analysis on their dataset to optimize their service & operations that

- Predict how capable each applicant is of repaying a loan
- Features that influence Home Credit Default Risk most

About the Data

- Home Credit Group posted its data readily & publicly on Kaggle https://www.kaggle.com/c/home-credit-default-risk/data
- Dataset carries nearly 307K records with 121 features and one 'TARGET' variable to infer predictions on each transaction

Data Wrangling Steps

Home Credit Group dataset required extensive data wrangling in terms of

- Extracting Data
- Identifying Target Dataset among Multiple Data Sources
- Identifying Missing Data
- Identifying Feature Set into Non-Categorical and Categorical
- Casting Data Types per Need
- Feature Engineering(Date timestamp, One Hot Encoding)
- Baseline Machine Learning Modeling

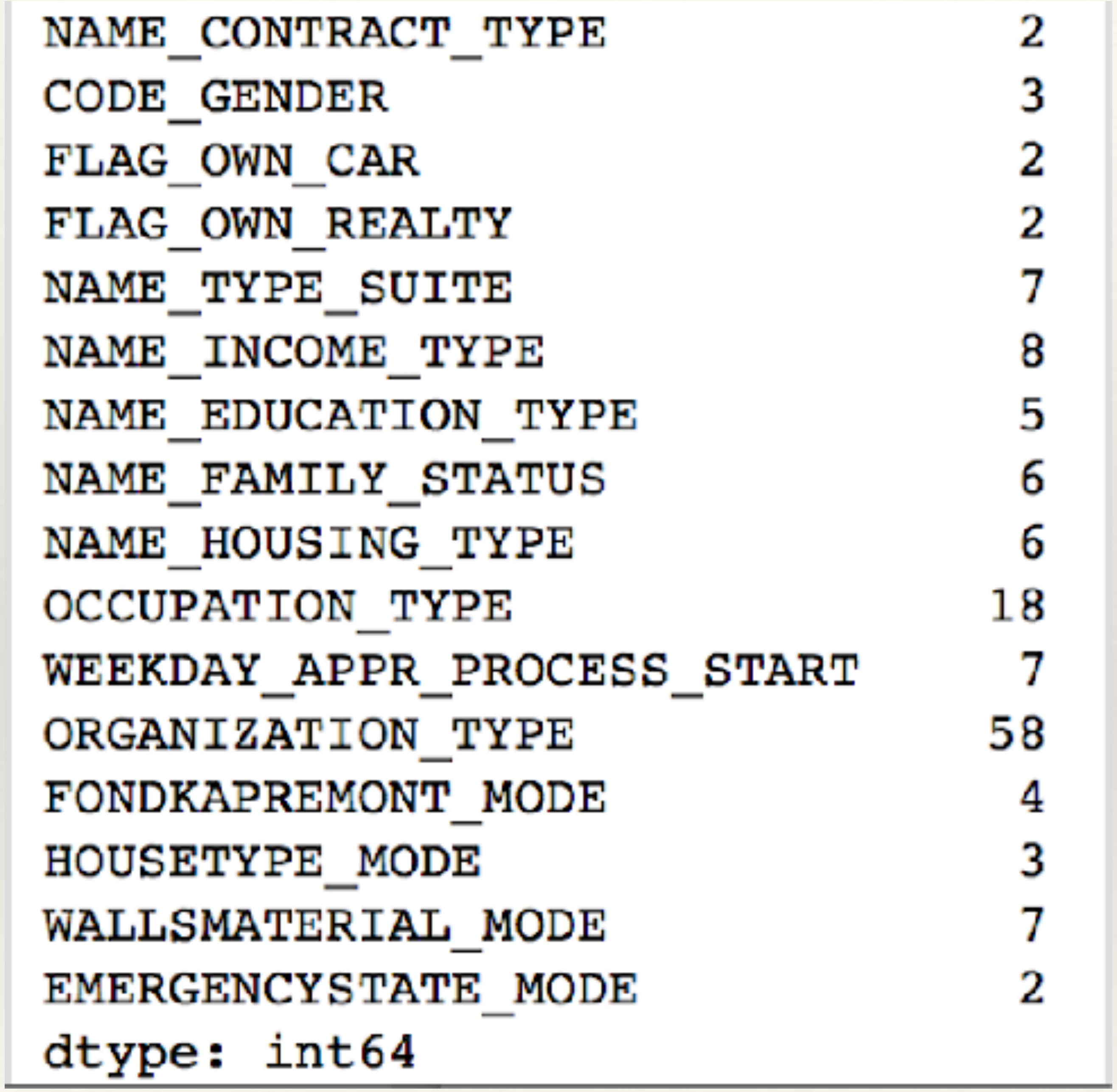
Data Wrangling Steps

With 70 % data missing in some of the 67 columns (total 122 features) it is important to baseline machine learning model (here Logistic Regression):

- Replace missing values with NaN strategy
- Machine Learning Imputation method

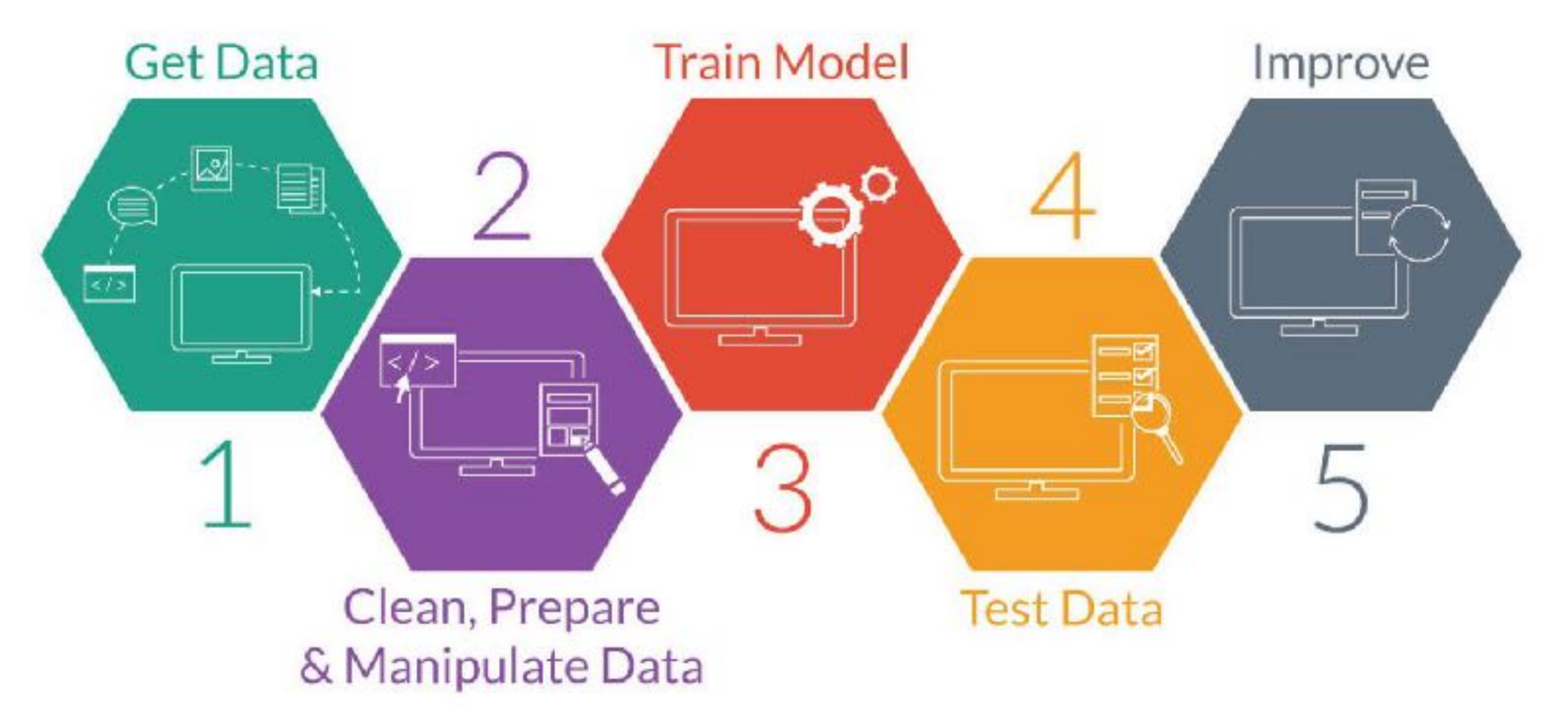
Data Wrangling Steps

As part of initial feature engineering, there were 16 categorical variables to be encoded:



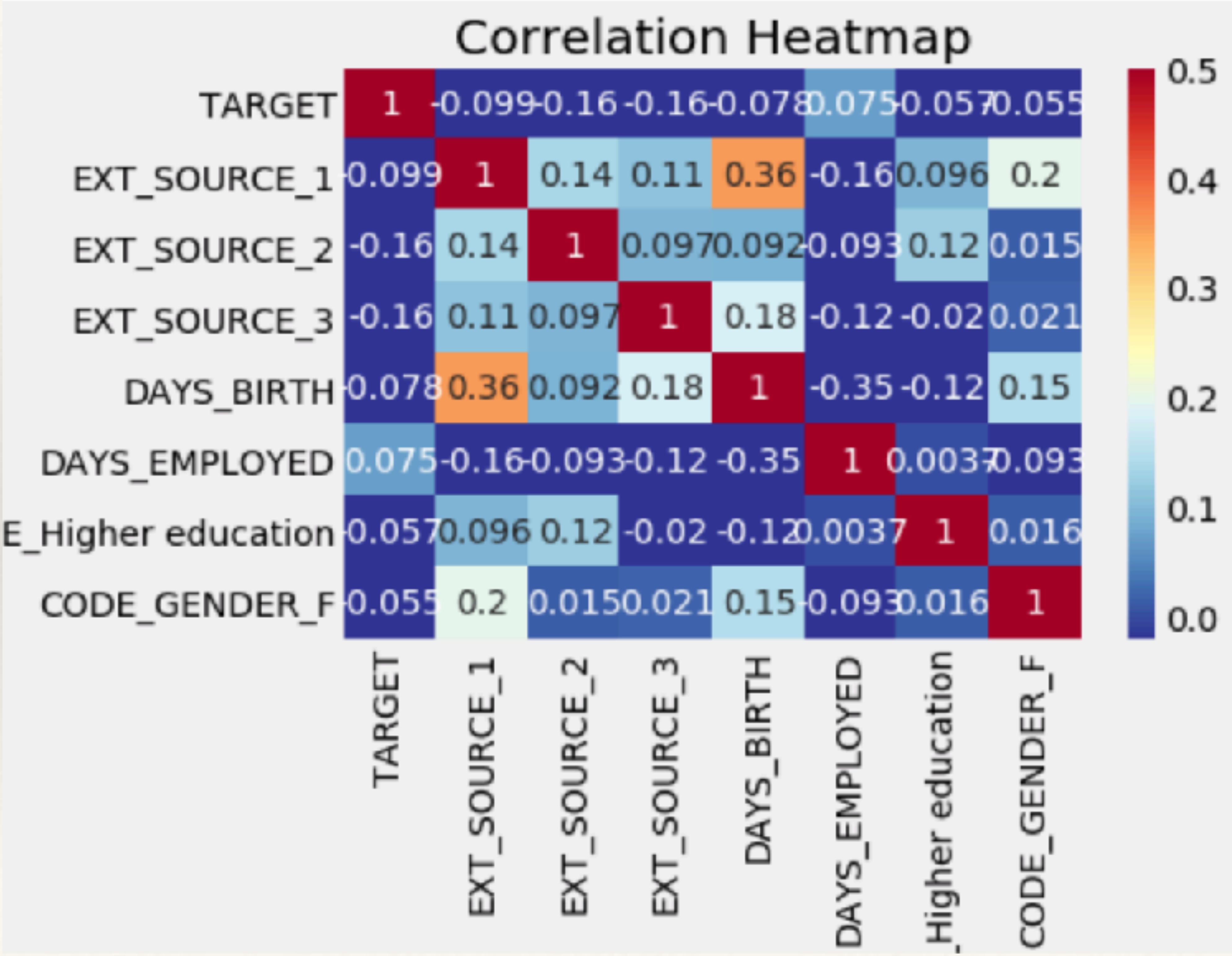
```
NAME_CONTRACT_TYPE             2
CODE_GENDER                    3
FLAG_OWN_CAR                   2
FLAG_OWN_REALTY                2
NAME_TYPE_SUITE                7
NAME_INCOME_TYPE               8
NAME_EDUCATION_TYPE            5
NAME_FAMILY_STATUS             6
NAME_HOUSING_TYPE              6
OCCUPATION_TYPE               18
WEEKDAY_APPR_PROCESS_START     7
ORGANIZATION_TYPE             58
FONDKAPREMONT_MODE             4
HOUSETYPE_MODE                 3
WALLSMATERIAL_MODE             7
EMERGENCYSTATE_MODE            2
dtype: int64
```

Model Workflows



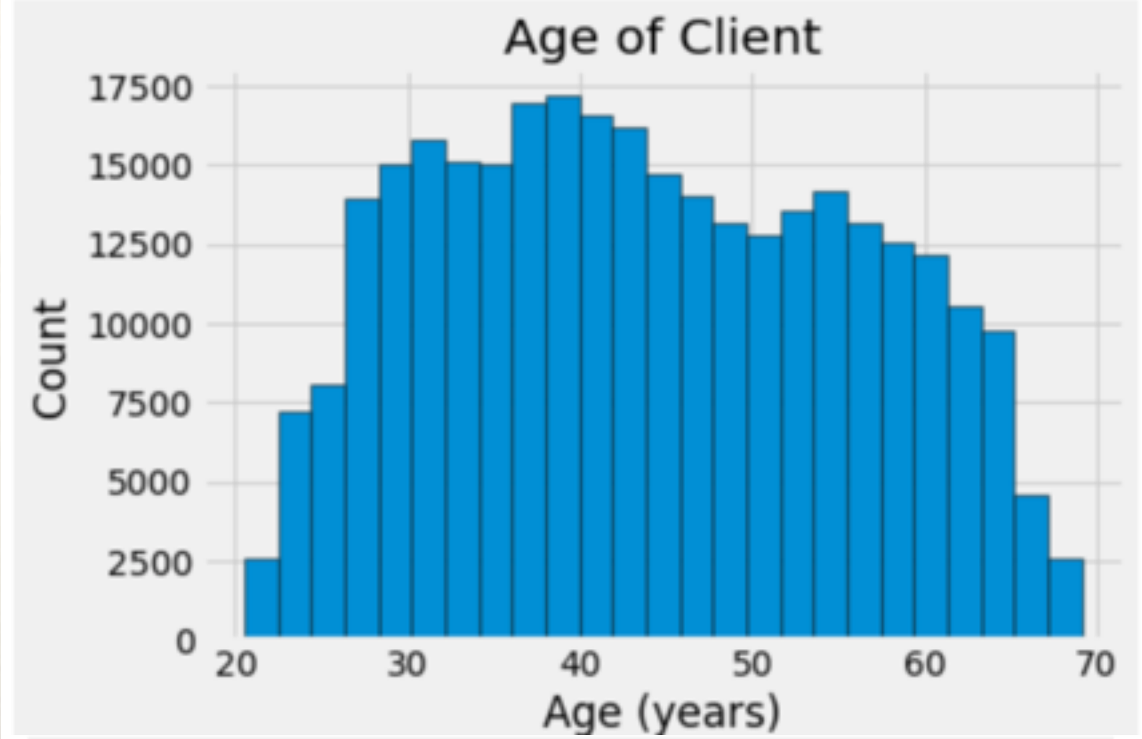
Exploratory Data Analysis

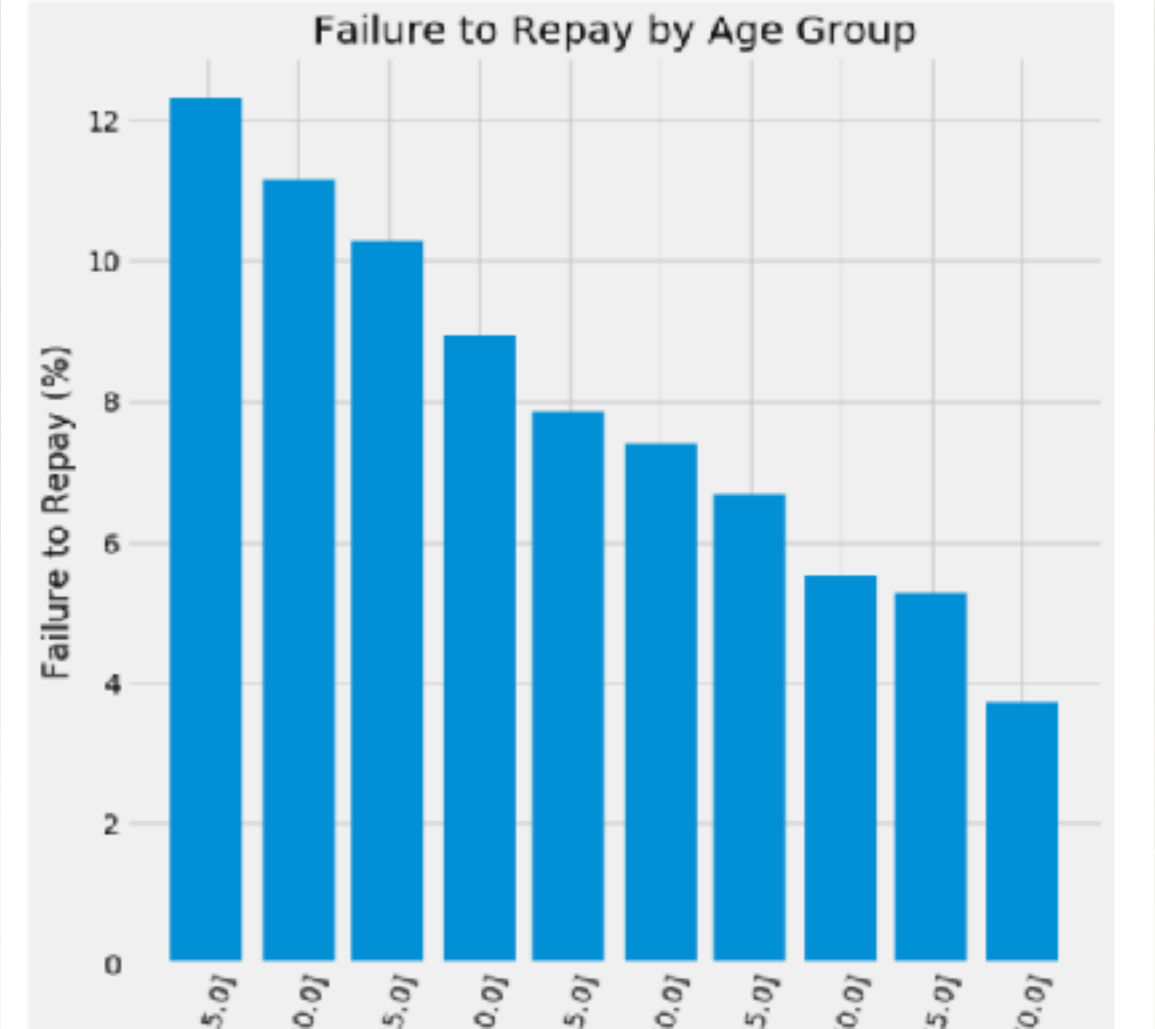
What impacts the customer repaying a loan ability

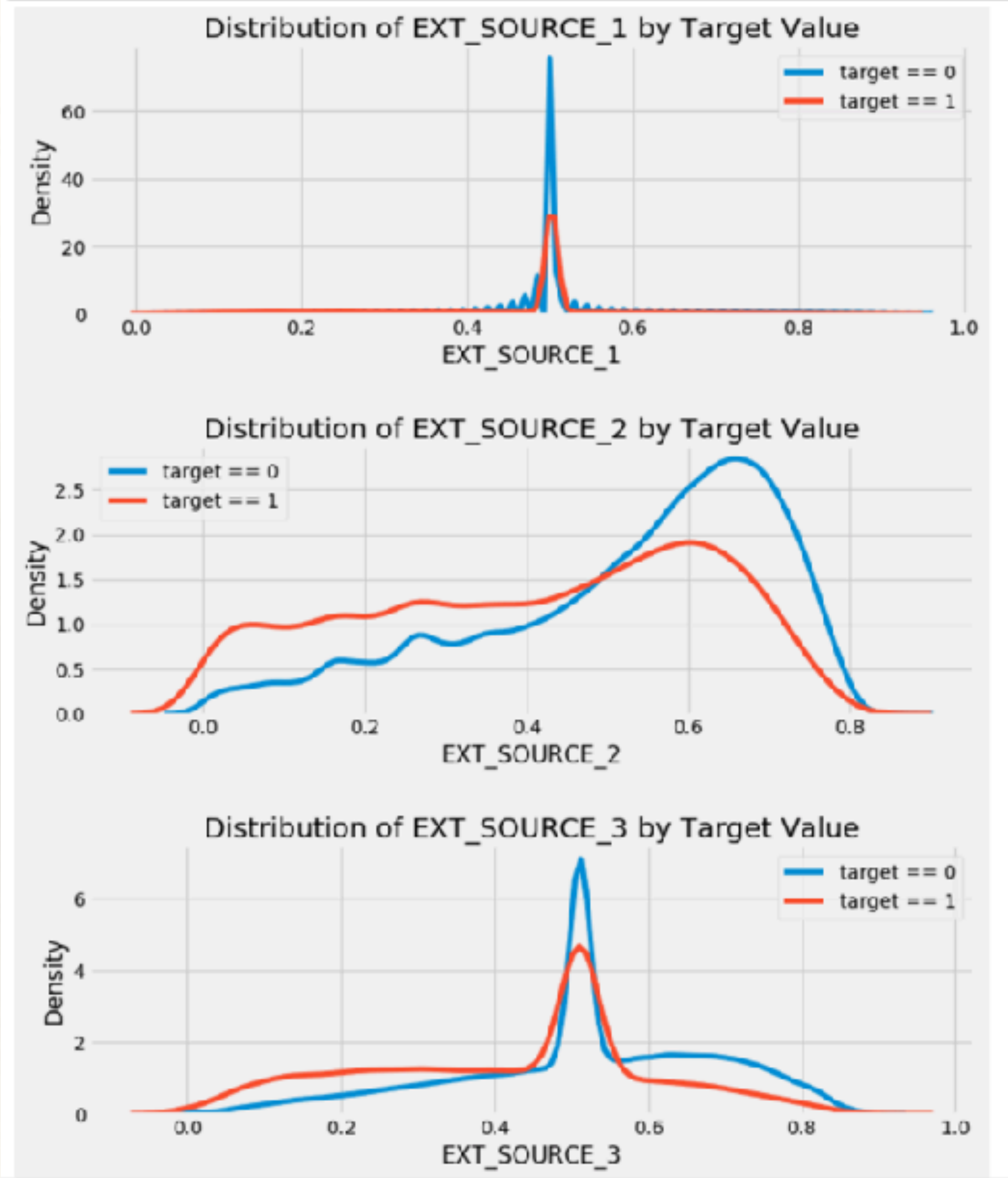


Exploratory Data Analysis

How and When impacts customer repaying a loan ability

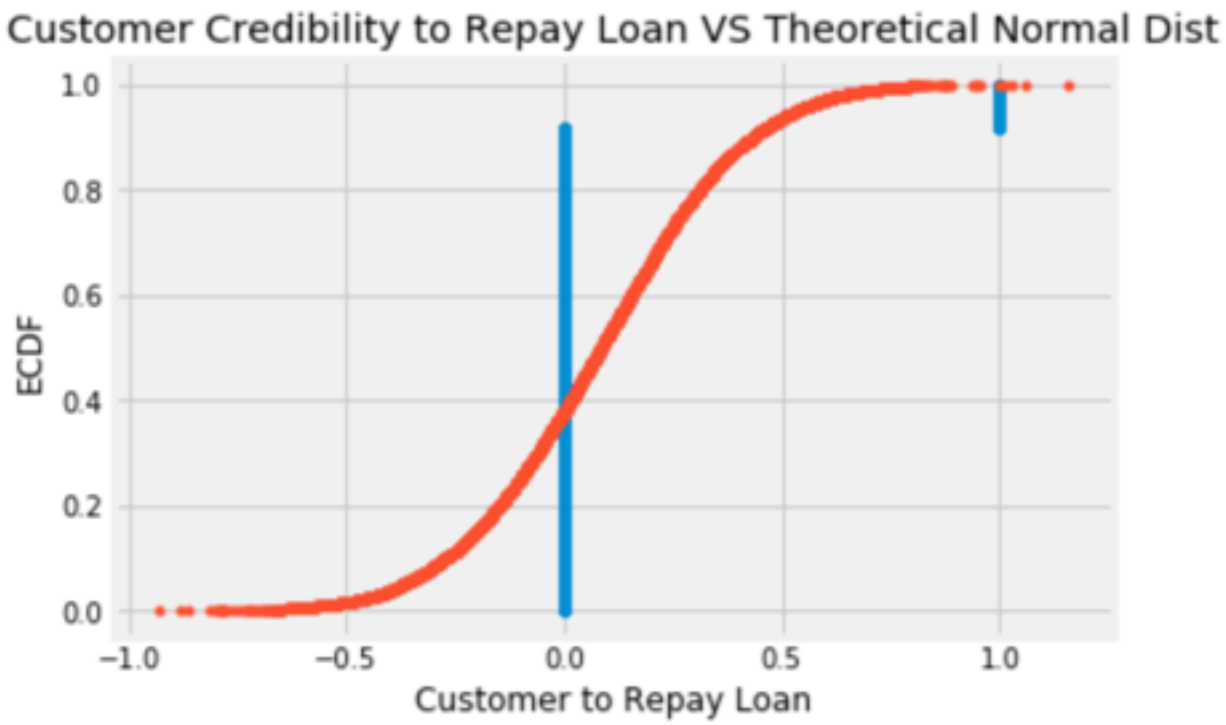


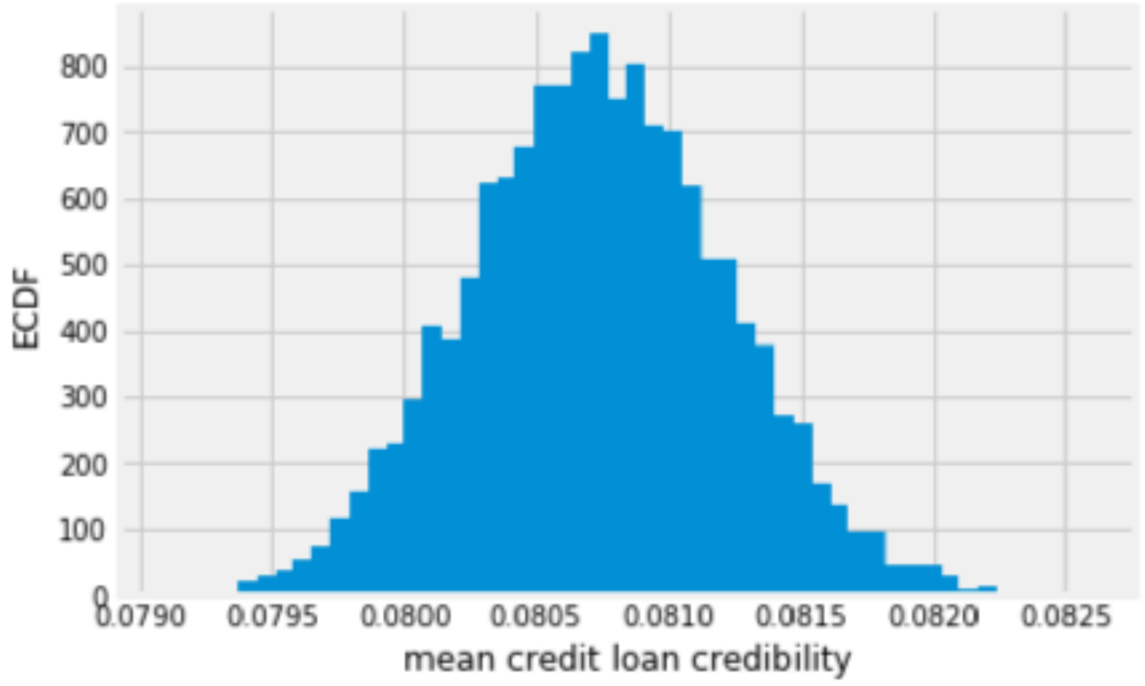




EDA/Inferential Statistics

Customer Repaying Loan Ability Distribution

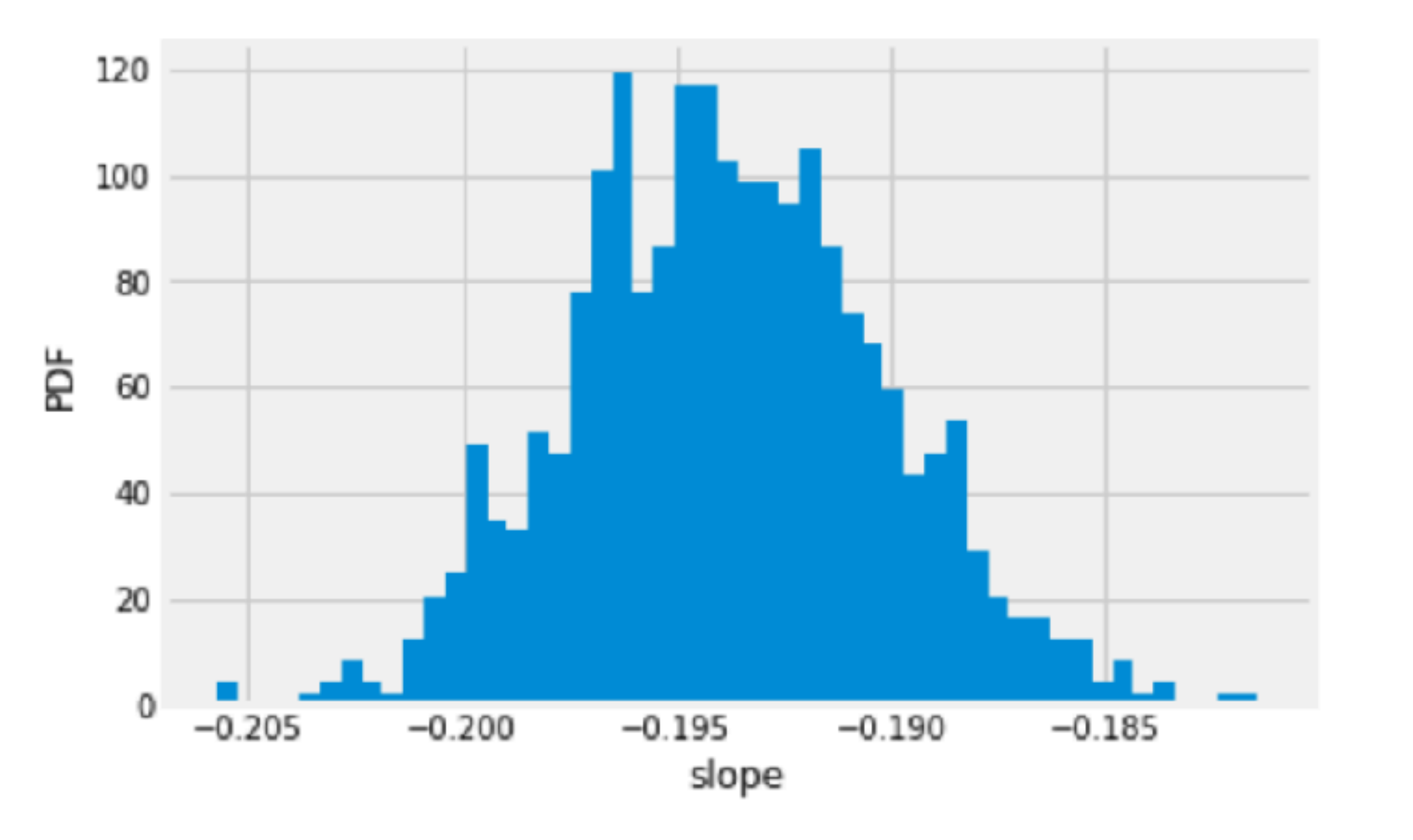




This is bootstrap estimate of the probability distribution function of the mean of 'Credit Loan Default Risk' at Home Credit Group. It assumes 95% Confidence Interval.

EDA/Inferential Statistics

Who impacts Credit Loan Default Risk Distribution

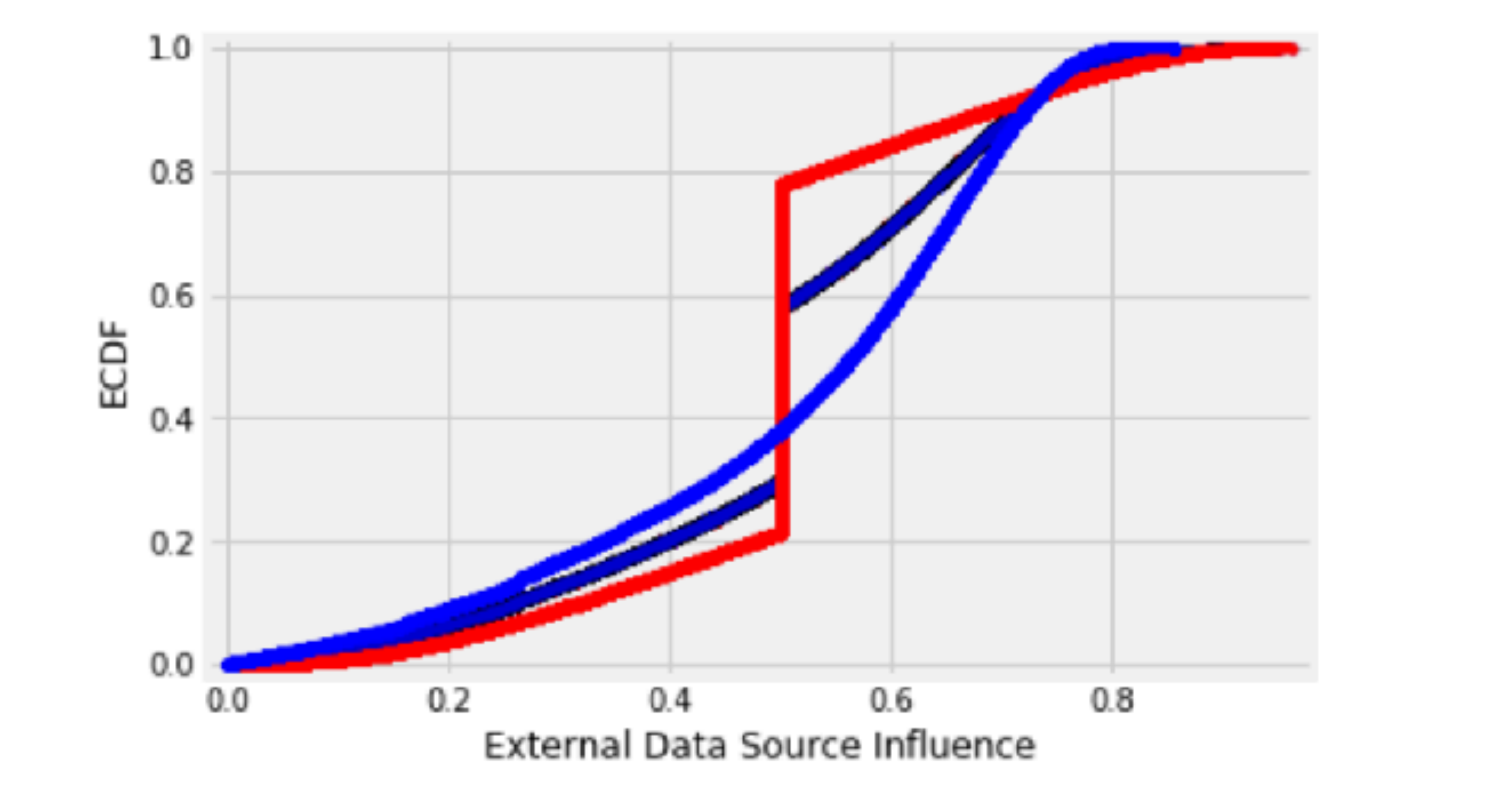


Extending Confidence Interval Concept to Pairs Bootstrap between TARGET (Credit Loan Default risk) and External_Source_1 data

Exploratory Data Analysis

How much is impact to the Home Credit Default Risk

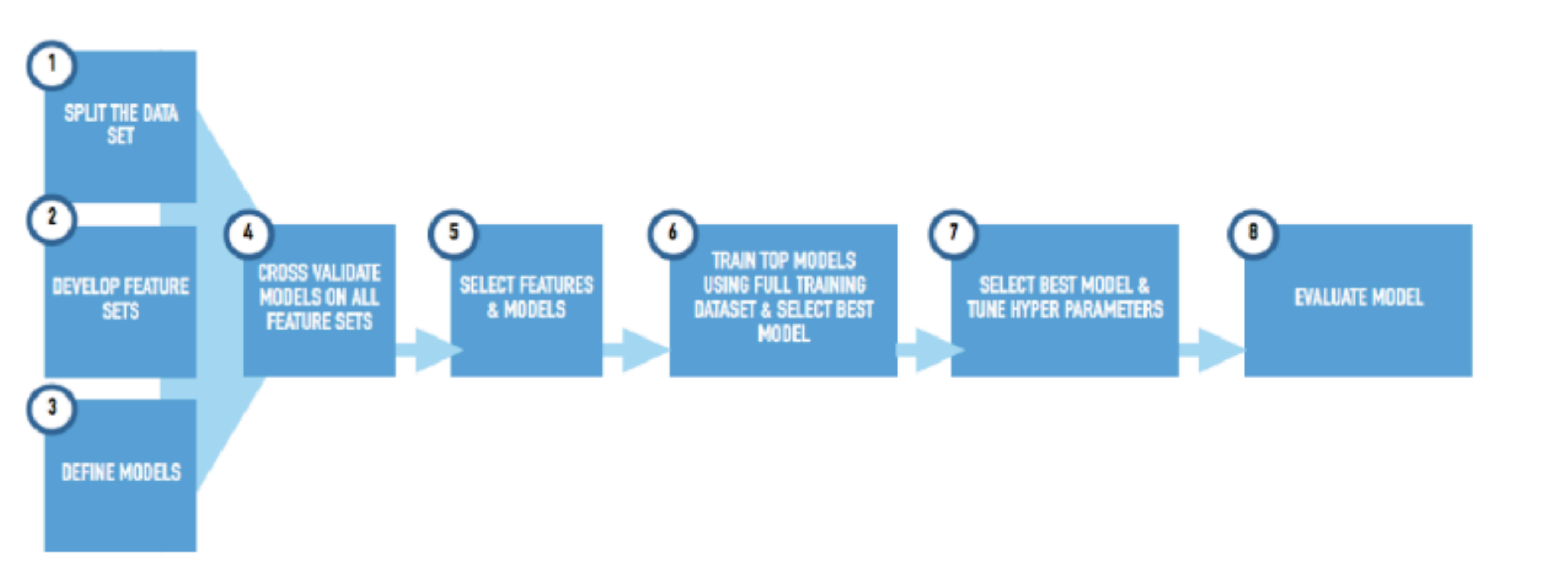
Null Hypothesis- There is no significant difference between EXT_SOURCE_1 and EXT_SOURCE_2 mean on 'Ability to Repay Loan'



EXT_SOURCE_1 & EXT_SOURCE_2 Means are not identically distributed and do not influence data in similar way. So Null Hypothesis is rejected.

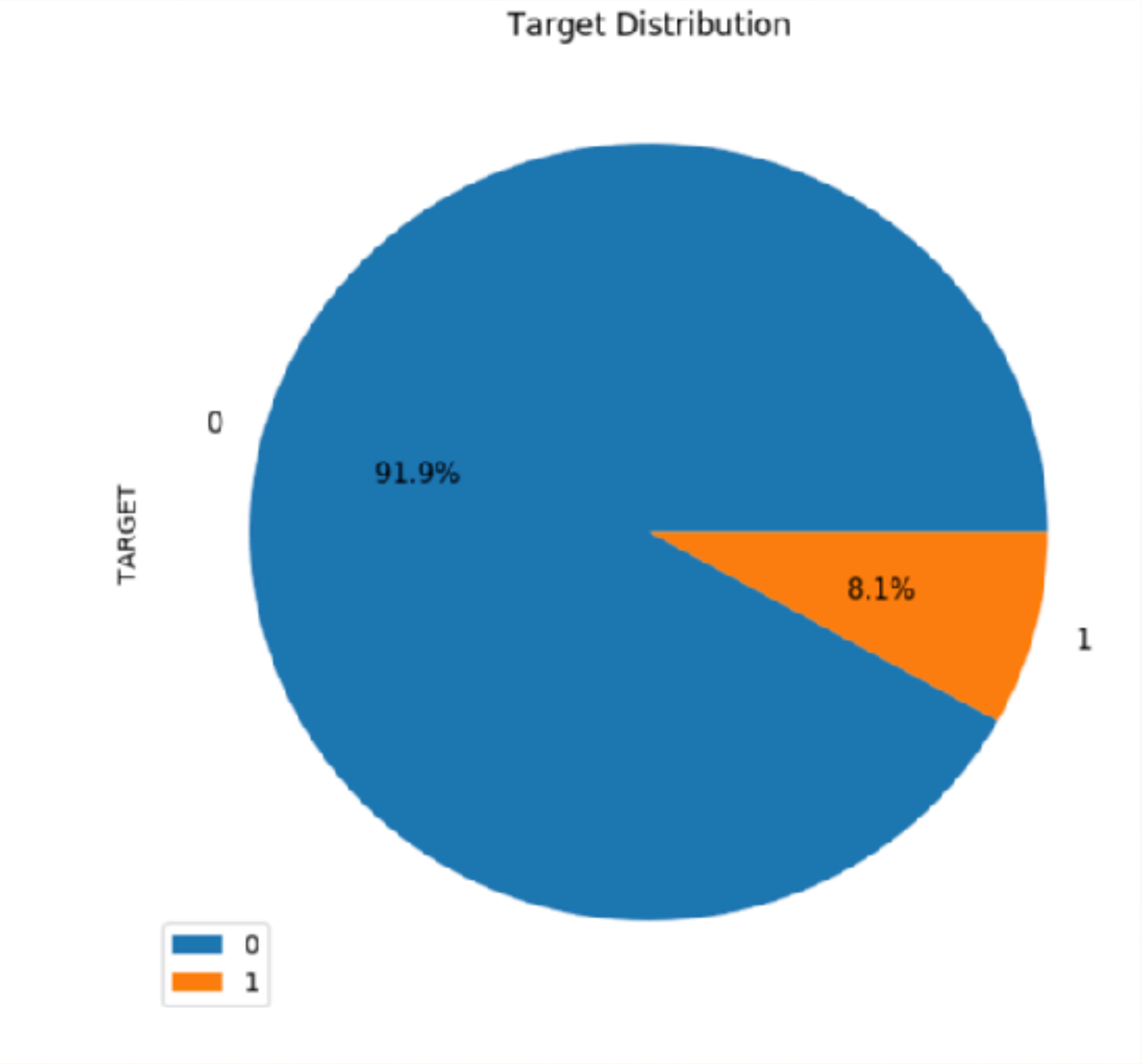
Supervised Classification

Steps in Machine Learning Modeling


1
SPLIT THE DATA SET
2
DEVELOP FEATURE SETS
3
DEFINE MODELS
4
CROSS VALIDATE MODELS ON ALL FEATURE SETS
5
SELECT FEATURES & MODELS
6
TRAIN TOP MODELS USING FULL TRAINING DATASET & SELECT BEST MODEL
7
SELECT BEST MODEL & TUNE HYPER PARAMETERS
8
EVALUATE MODEL

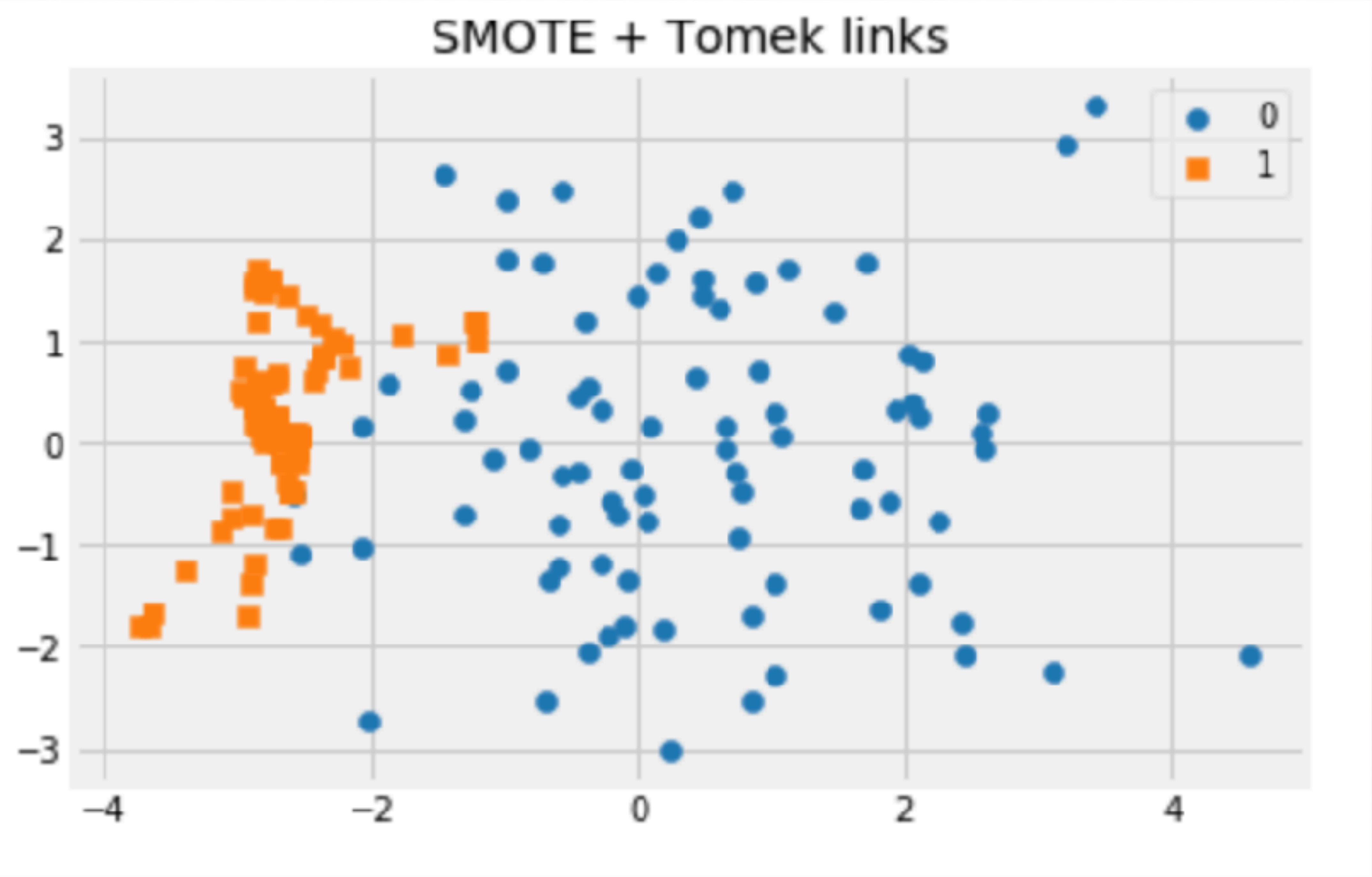
Supervised Classification

Class Imbalance Issue in Baseline Model Data set



Supervised Classification

Resampling, PCA, Tomek Links, Cluster Centroids, SMOTE resolves Class Imbalance



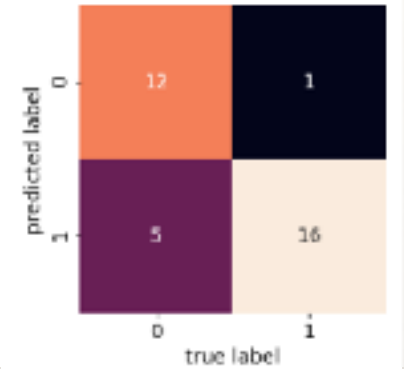
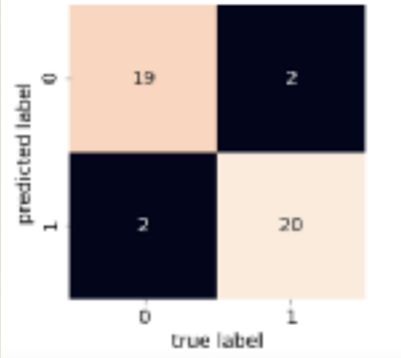
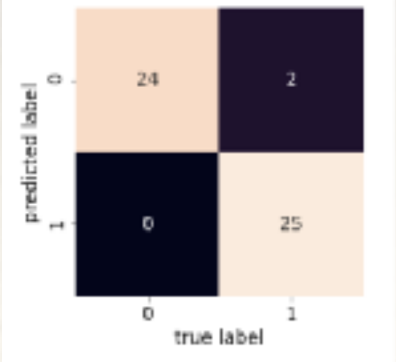
Supervised Classification

Deploy Machine Learning Model over Resampled Dataset

Classifier	Parameters
Logistic Regression	C=1C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False
Random Forest	n_estimators=100, random_state=0
Decision Tree	class_weight=None, max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_split=1e-07, min_samples_leaf=1,min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best'

Supervised Classification

Model Evaluation

Classifier	Accuracy	Precision	Recall	F1-score	Confusion Matrix
Logistic Regression	0.82	0.84	0.82	0.82	
Random Forest	0.91	0.91	0.91	0.91	
Decision Tree	0.96	0.96	0.96	0.96	

Recommendations

With this predictive model, Client benefits in better prediction of Home Credit Default Risk:

- Age distribution indicates by increasing age TARGET probability to repay loan increases
- External Data Sources influences TARGET inversely
- Gender correlates with Target prediction
- Employment Duration indicates by increasing Employment Duration, TARGET probability to repay loan increases
- Identification if loan is cash or revolving
- Normalized score from external data source 1/data source 2/data source 3
- Flag if the client owns a car
- Flag if client owns a house or flat
- Number of children the client has
- Income of the client
- Credit amount of the loan
- Loan annuity
- For consumer loans it is the price of the goods for which the loan is given
- Who was accompanying client when he was applying for the loan
- Clients income type (businessman, working, maternity leave,…)
- Level of highest education the client achieved
- Family status of the client
- What is the housing situation of the client (renting, living with parents, ...)

Future Work

There is lot of potential to enhance the model by:

- Collection of more features in the dataset like Geographic Region and Credit History dataset inclusion to help client identify if Risk is none or some to repay loan by customer

- Model improvement using other Classification models like k-Nearest Neighbor, Support Vector

References

- Blog post: Simple guide to confusion matrix terminology by me
- Videos: Intuitive sensitivity and specificity (9 minutes) and The tradeoff between sensitivity and specificity (13 minutes) by Rahul Patwari
- Notebook: How to calculate "expected value" from a confusion matrix by treating it as a cost-benefit matrix (by Ed Podojil)
- Graphic: How classification threshold affects different evaluation metrics (from a blog post about Amazon Machine Learning)
- scikit-learn documentation: Model evaluation
- Guide: Comparing model evaluation procedures and metrics by me
- Video: Counterfactual evaluation of machine learning models (45 minutes) about how Stripe evaluates its fraud detection model, including slides

Thank You