HOME CREDIT

Default Risk

STEVEN L TRUONG

Friday, 05/14/2021

Motivation

Motivation

The Bank

Determine if potential clients are capable of repayment to prevent losing money on bad credit clients.

Motivation

The Bank

Determine if potential clients are capable of repayment to prevent losing money on bad credit clients.

Clients

Ensure that people who are capable of repayment are not rejected and help people to achieve their dreams.

Data and tools

Data

Provided by Home Credit through Kaggle.

Language

Python

Modeling

Scikit-learn, xgboost, lightGBM, pandas, numpy.

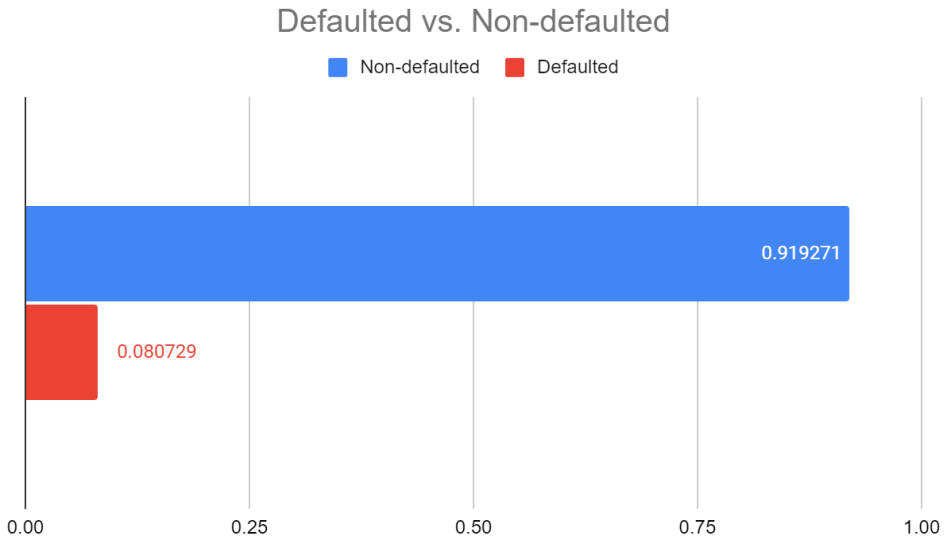
DataViz

Matplotlib, seaborn

Exploratory data analysis (EDA)

EDA and insights

EDA and insights



91.9 %

NON-defaulted

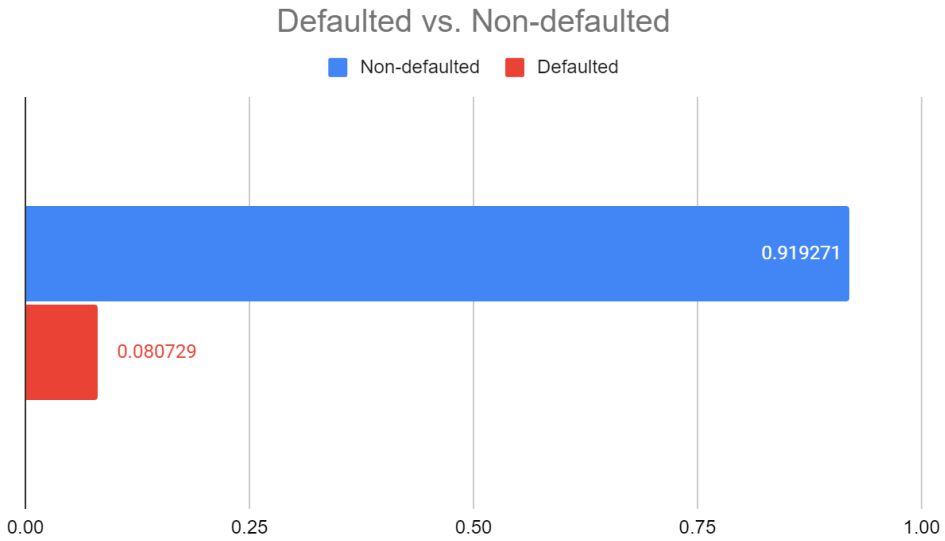
Accounts that essentially have good credit.

8.1 %

Defaulted

People that have poor credit or non-existent credit histories.

EDA and insights



91.9 %

NON-defaulted

Accounts that essentially have good credit.

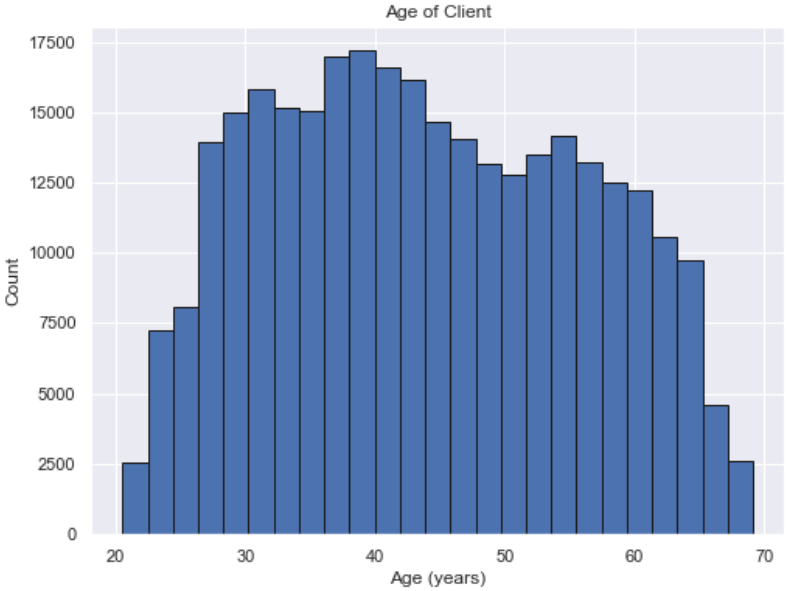
8.1 %

Defaulted

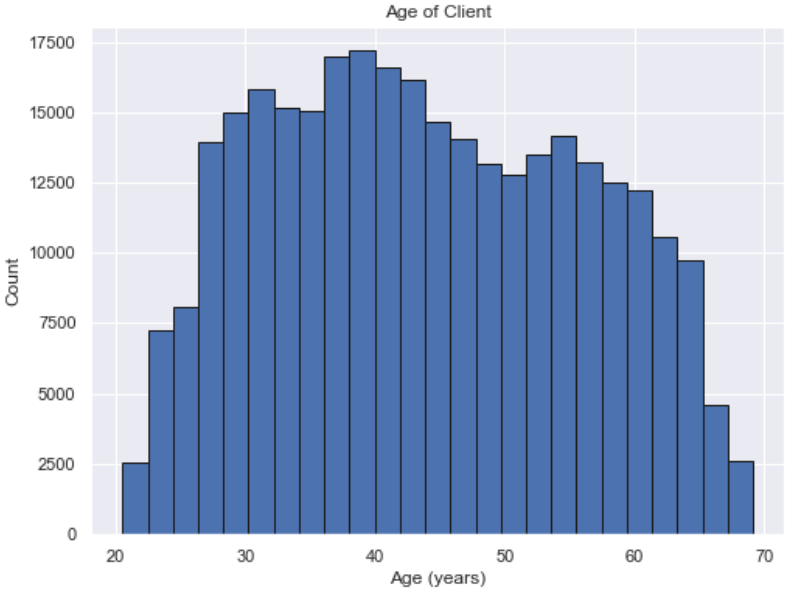
People that have poor credit or non-existent credit histories.

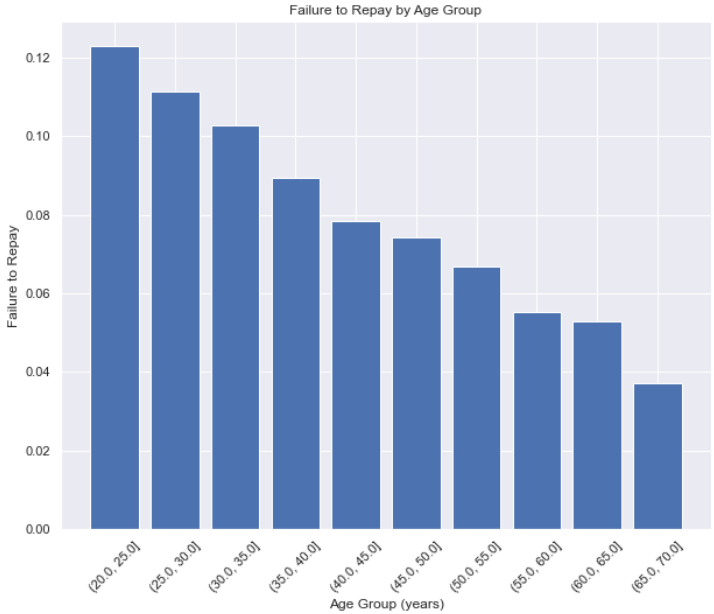
This is an **imbalance class** problem. The ratio is roughly 11:1

EDA and insights



EDA and insights





The younger the client, the more likely to get defaulted.

EDA and insights

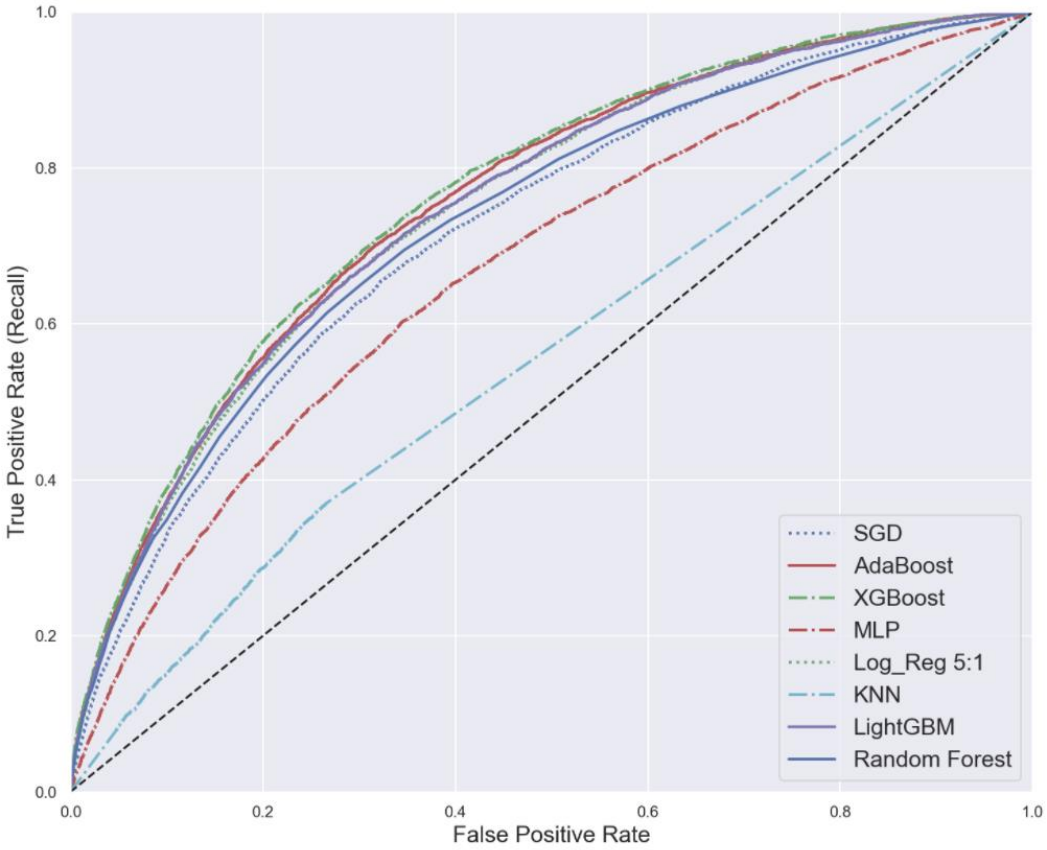
External source of income displays the **difference** between the values of the target. Hence there is some **relationship to the likelihood** of an applicant to repay a loan.

Let’s build some models

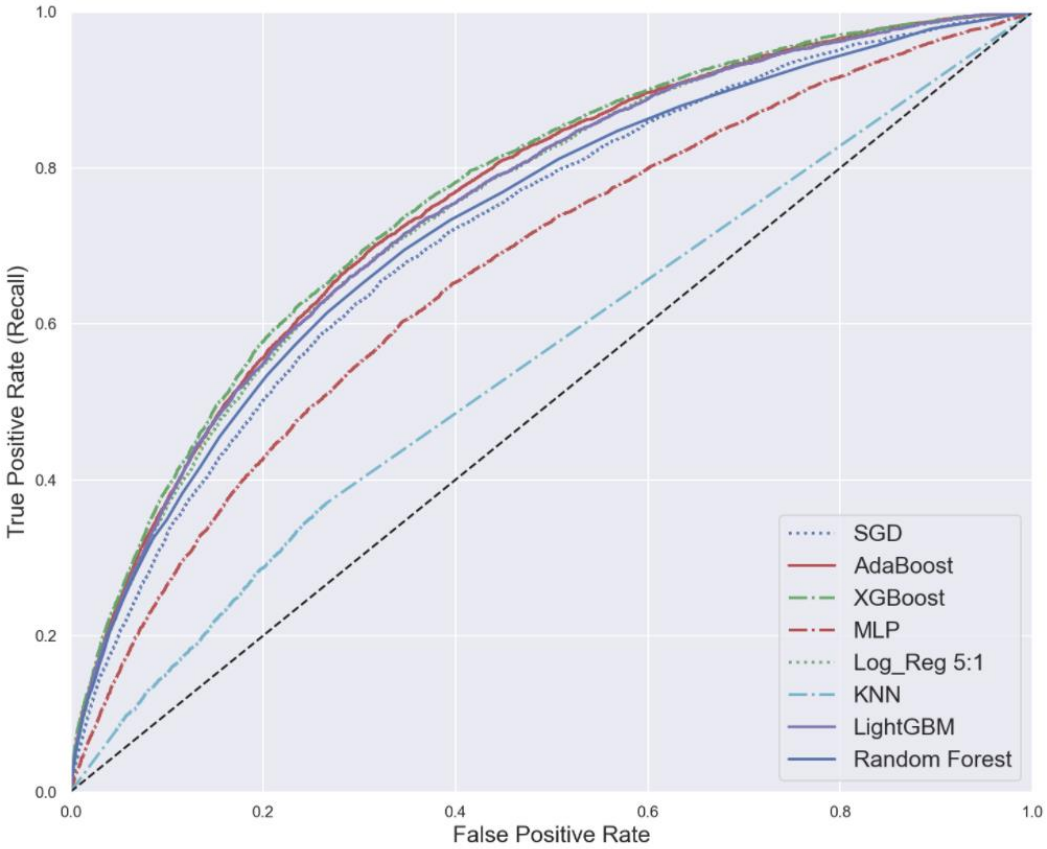
Use the power of data science and machine learning.

Models comparison (ROC AUC)

Models comparison (ROC AUC)



Models comparison (ROC AUC)

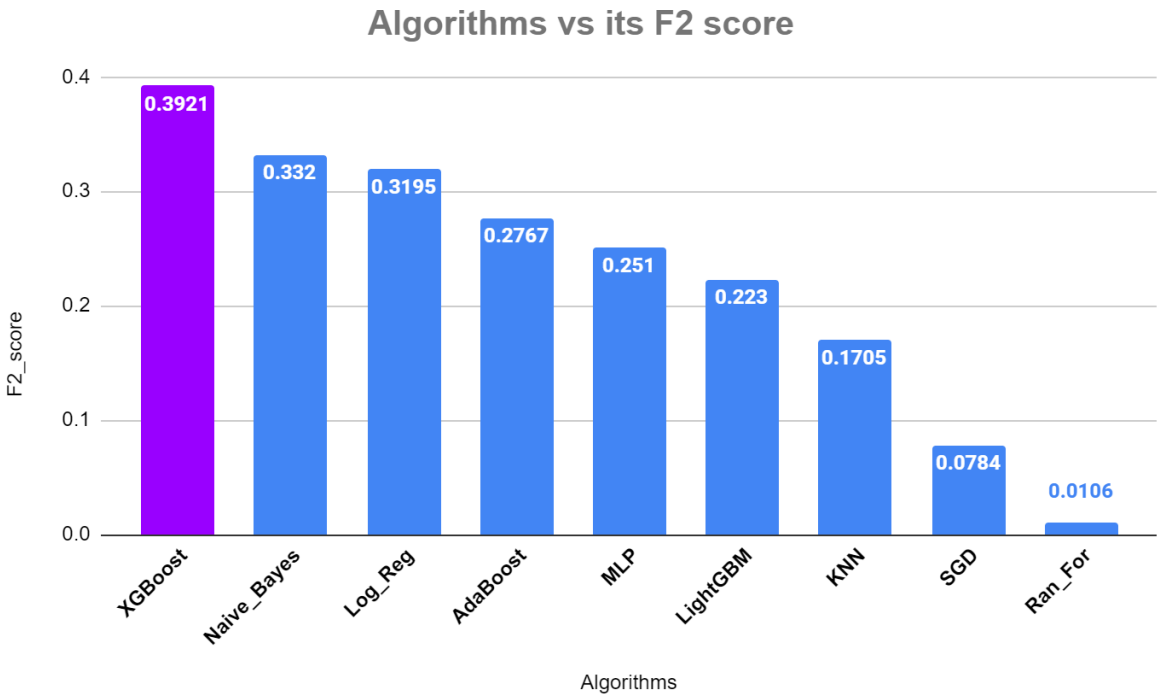


XGBoost wins.

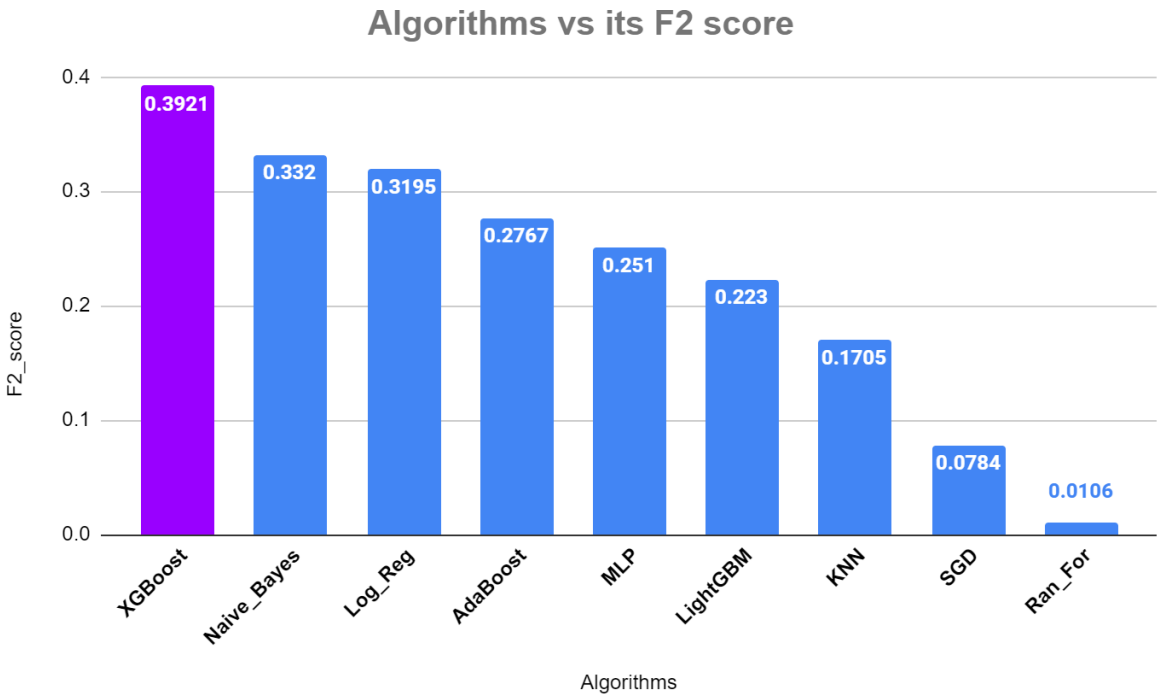
(GBT on steroid)

Models comparison (F2 Score)

Models comparison (F2 Score)



Models comparison (F2 Score)



- Metrics chosen: F_beta with **beta** = 2
- Again, our good buddy **XGBoost** wins.

Model optimizaton

Now we have our chosen model,
let's tune and improve **XGBoost.**

Optimized XGBoost Model

Optimized XGBoost Model

Train Set

F2 Score

0.463

ROC AUC

0.804

Optimized XGBoost Model

Train Set

F2 Score

0.463

ROC AUC

0.804

Test Set

F2 Score

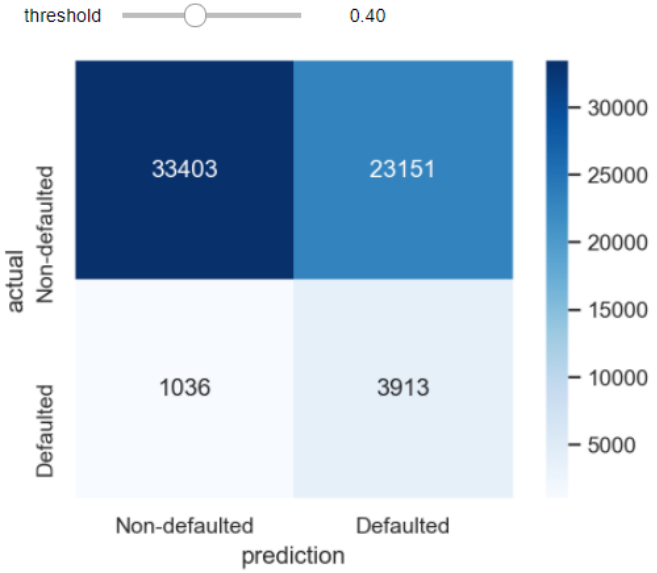
0.432

ROC AUC

0.768

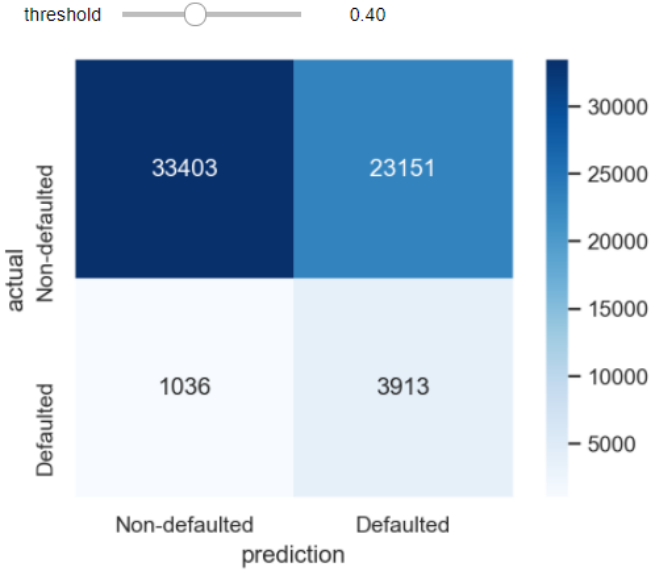
Results

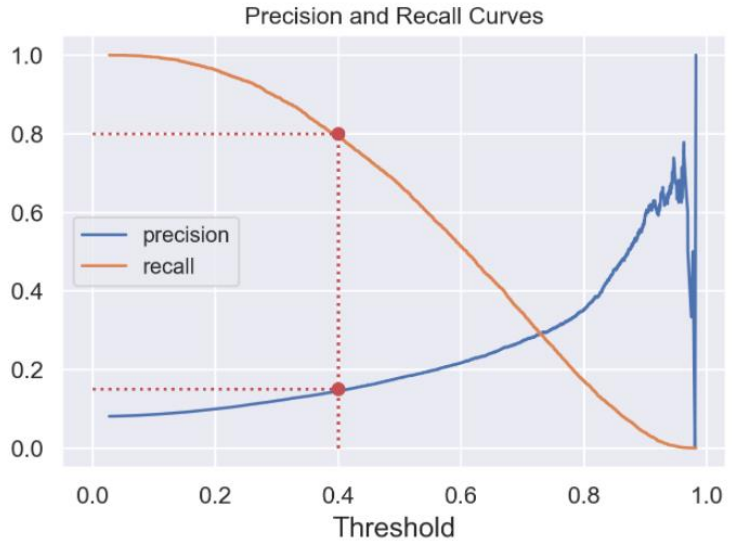
Results



We lean a little bit towards recall (not too strict on precision either)

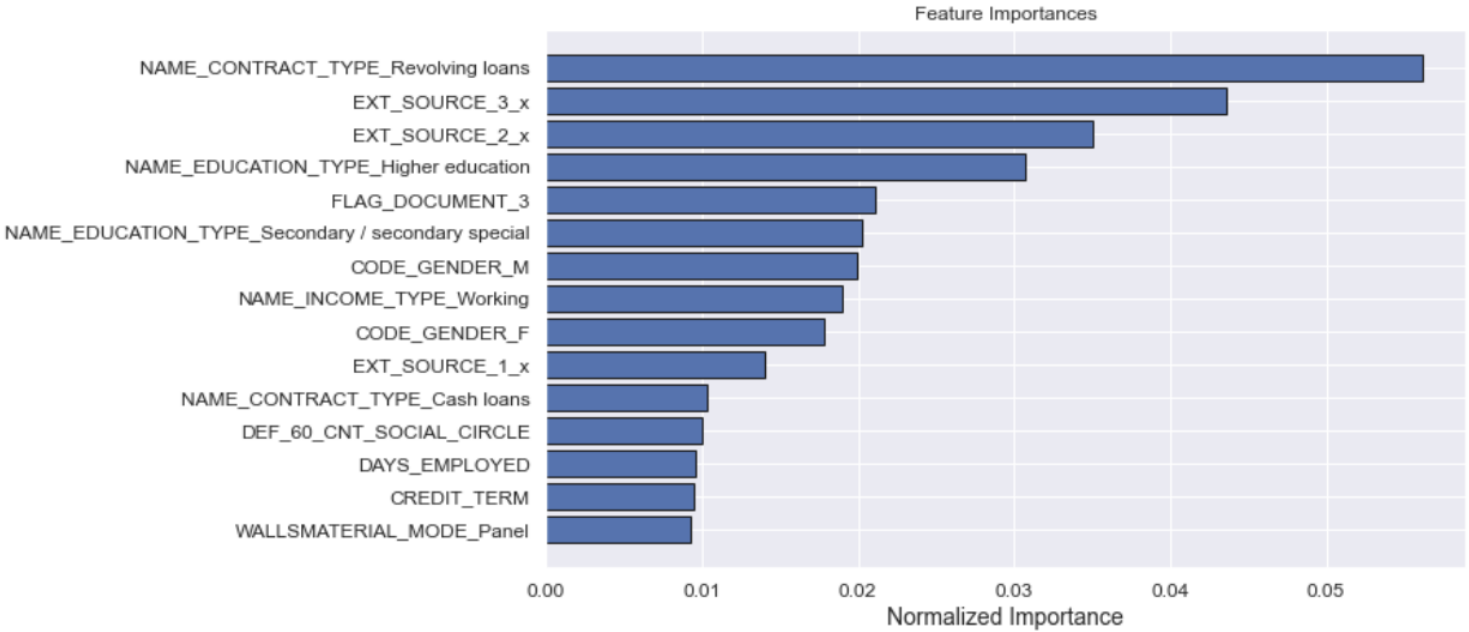
Results





We lean a little bit towards recall (not too strict on precision either)

Features importance



Mainly external **income sources, education,** and **type of loans** determine the decision.

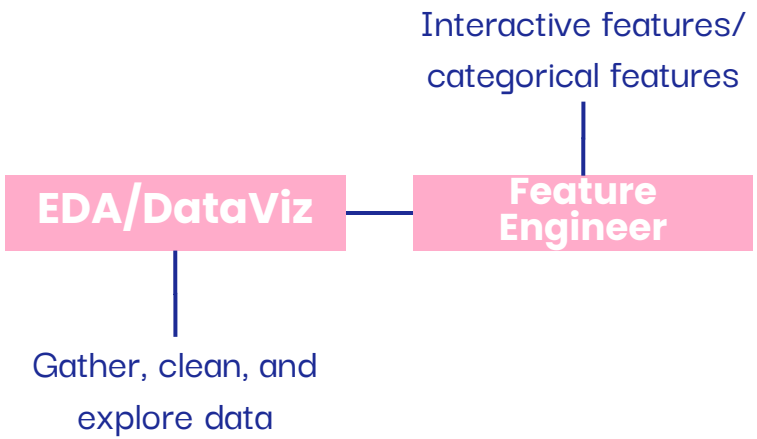
Recap

Recap

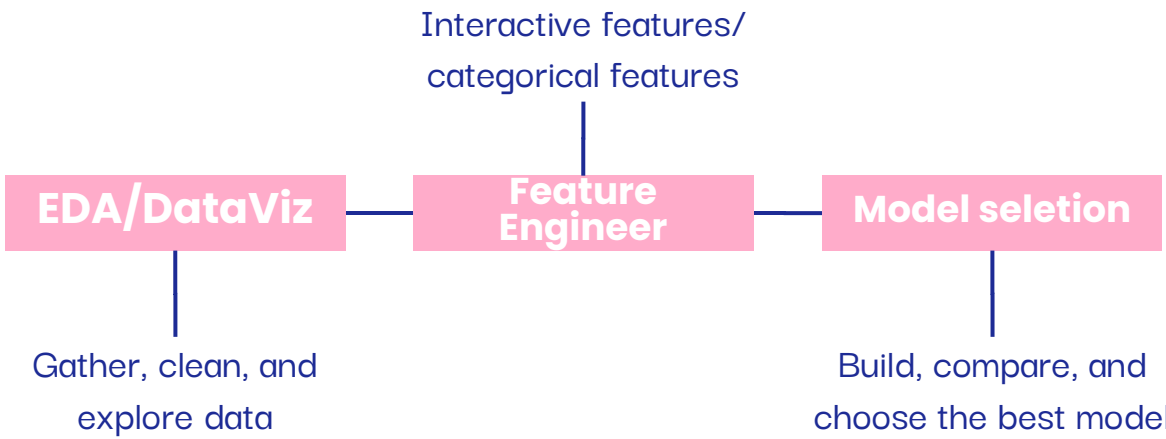
EDA/DataViz

Gather, clean, and explore data

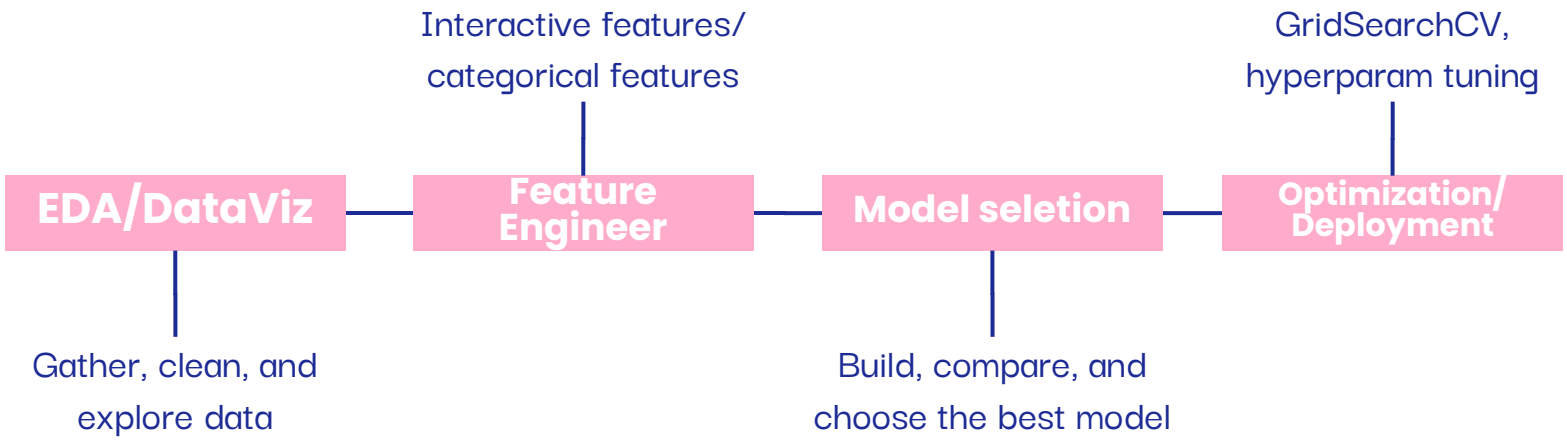
Recap


Interactive features/
categorical features
EDA/DataViz
Feature Engineer
Gather, clean, and
explore data

Recap


Interactive features/
categorical features
EDA/DataViz
Feature Engineer
Model seletion
Gather, clean, and explore data
Build, compare, and choose the best model

Recap


EDA/DataViz
Gather, clean, and explore data
Feature Engineer
Interactive features/ categorical features
Model seletion
Build, compare, and choose the best model
Optimization/ Deployment
GridSearchCV, hyperparam tuning

Future Work

Data

Incorporate multiple datasets

Algorithm

Do better on XGBoost and LightGBM

Deployment

Build interactive app and deploy to streamlit/AWS

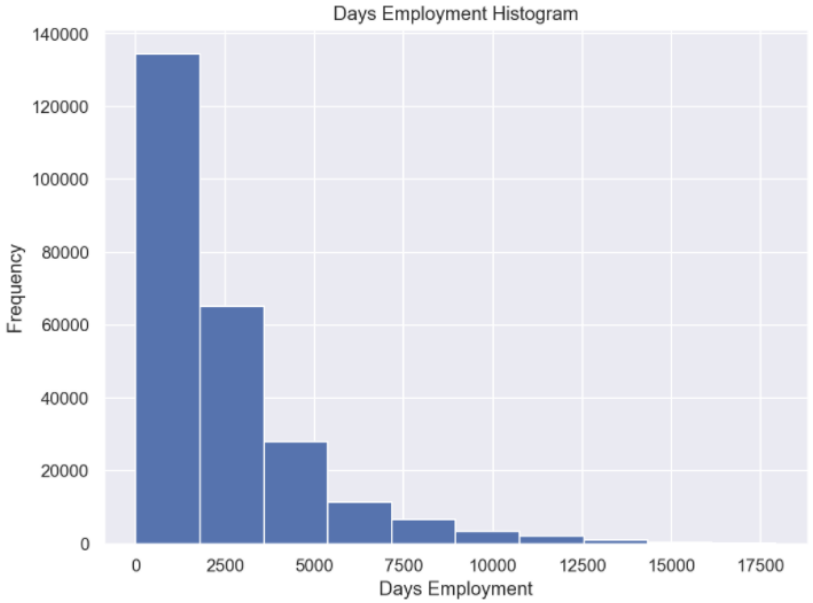
Thank you

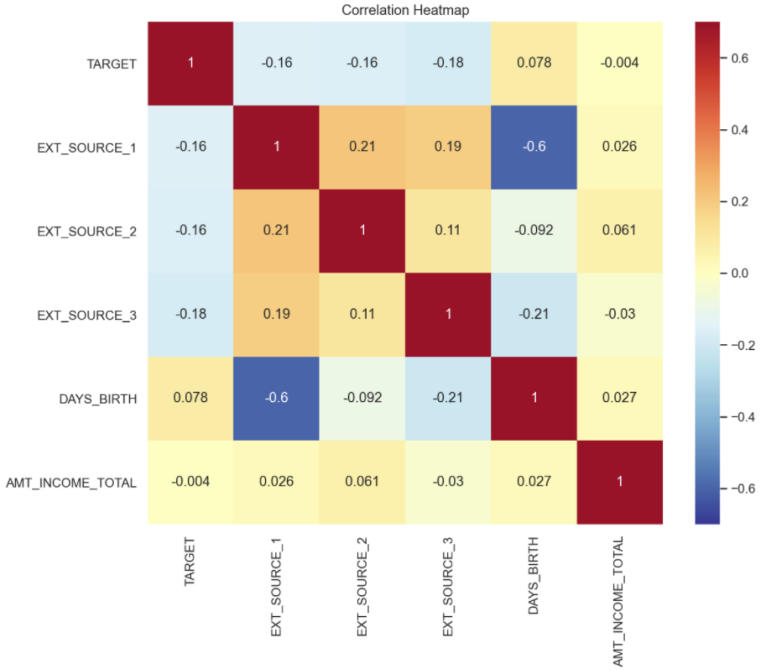
Questions?

Please reach out to me at https://www.linkedin.com/in/luongtruong77/

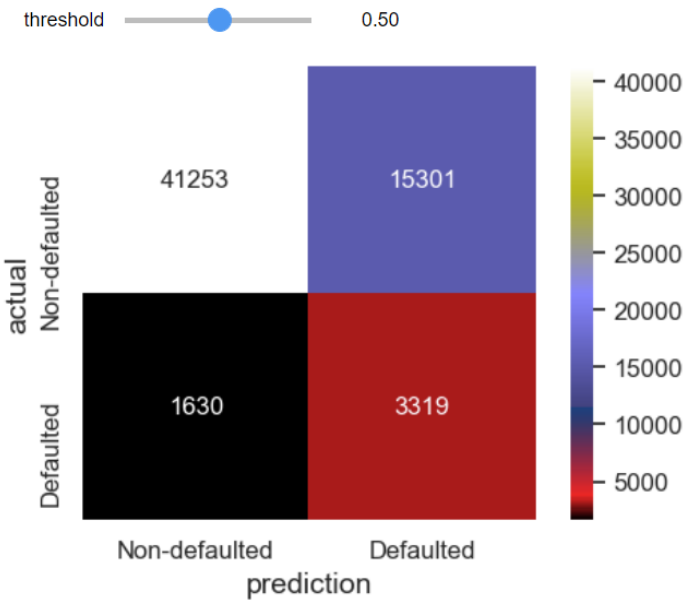
Steven L Truong

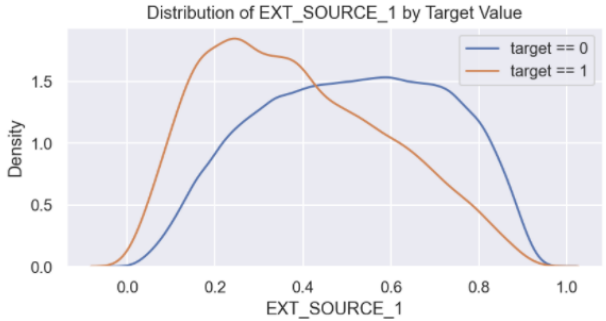
Appendix

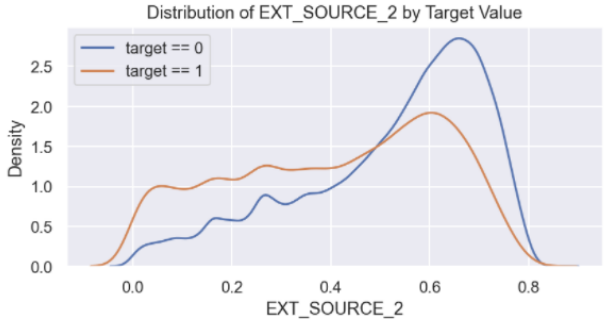

Days Employment Histogram
Frequency
140000
120000
100000
80000
60000
40000
20000
0
0
2500
5000
7500
10000
12500
15000
17500
Days Employment


Correlation Heatmap
TARGET
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
DAYS_BIRTH
AMT_INCOME_TOTAL
1 -0.16 -0.16 -0.18 0.078 -0.004
-0.16 1 0.21 0.19 -0.6 0.026
-0.16 0.21 1 0.11 -0.092 0.061
-0.18 0.19 0.11 1 -0.21 -0.03
0.078 -0.6 -0.092 -0.21 1 0.027
-0.004 0.026 0.061 -0.03 0.027 1
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Appendix


threshold
0.50
41253
15301
1630
3319
actual
Non-defaulted
Defaulted
Non-defaulted
Defaulted
prediction
40000
35000
30000
25000
20000
15000
10000
5000


Distribution of EXT_SOURCE_1 by Target Value
target == 0
target == 1
Density
EXT_SOURCE_1


Distribution of EXT_SOURCE_2 by Target Value
target == 0
target == 1
Density
EXT_SOURCE_2