

The background of the slide is white and decorated with various geometric shapes in blue, pink, and brown. These include circles, semi-circles, and triangles, some of which are partially cut off by the edges of the frame. The shapes are arranged in a scattered, abstract pattern.

HOME CREDIT Default Risk

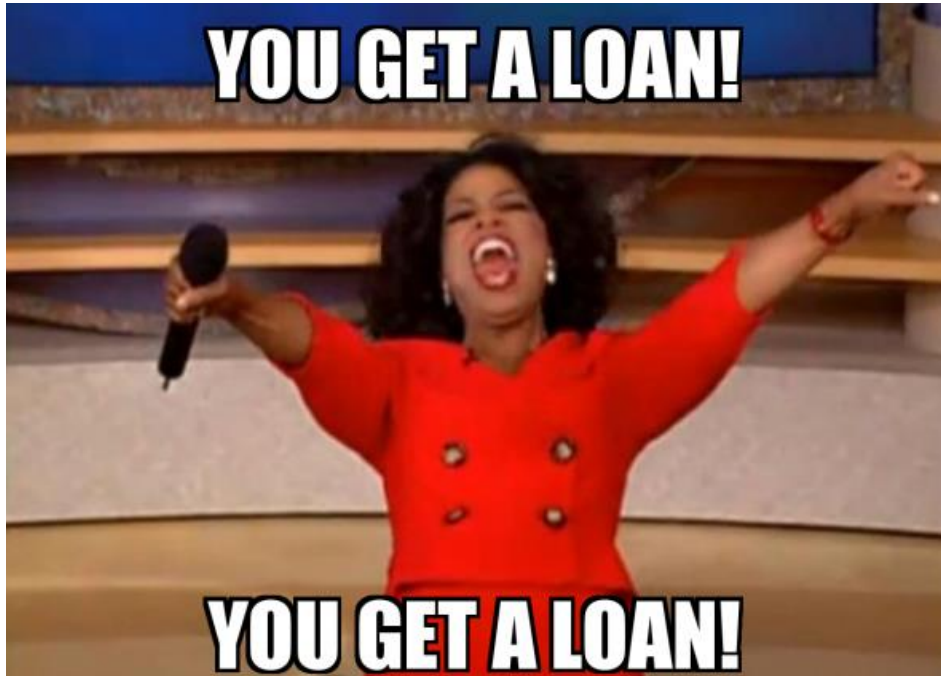
STEVEN L TRUONG

Friday, 05/14/2021

Motivation

**Can we get
accepted to the
loan that we want
when applying?**

Motivation



HOME CREDIT

HOME CREDIT



The Bank

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

HOME CREDIT



The Bank

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



Applicants

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.



Modeling

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



Language

Python



DataViz

Matplotlib, seaborn

Exploratory data analysis (EDA)



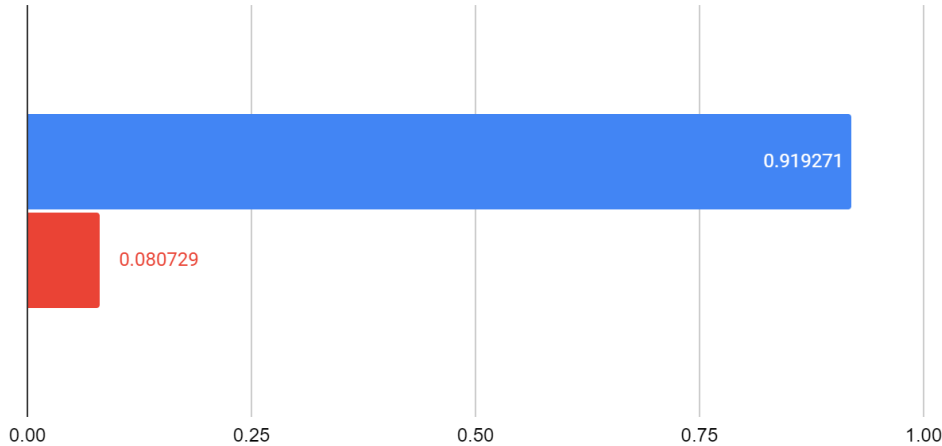


EDA and insights

EDA and insights

Defaulted vs. Non-defaulted

■ Non-defaulted ■ Defaulted



91.9 %

NON-defaulted

Accounts that essentially have good credit.

8.1 %

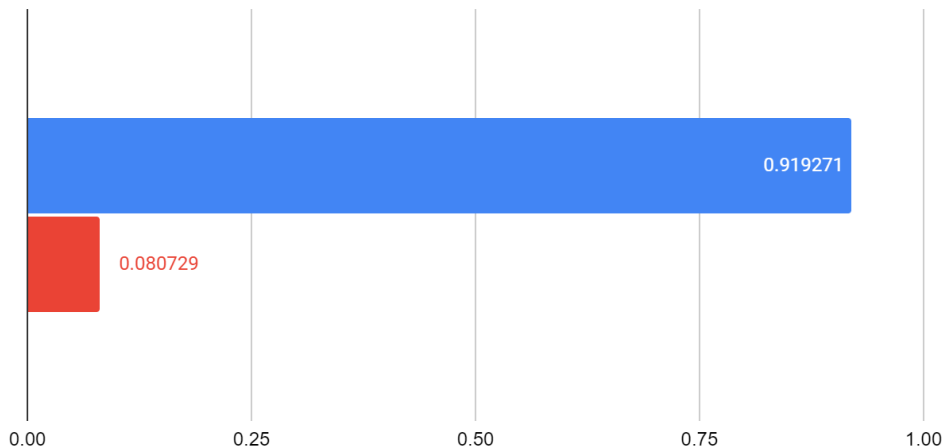
Defaulted

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

EDA and insights

Defaulted vs. Non-defaulted

■ Non-defaulted ■ Defaulted



91.9 %

NON-defaulted

Accounts that essentially have good credit.

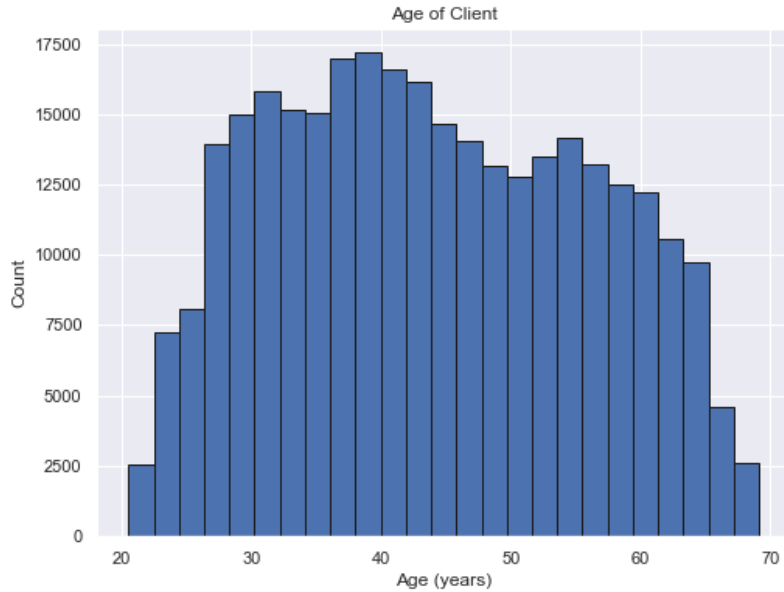
8.1 %

Defaulted

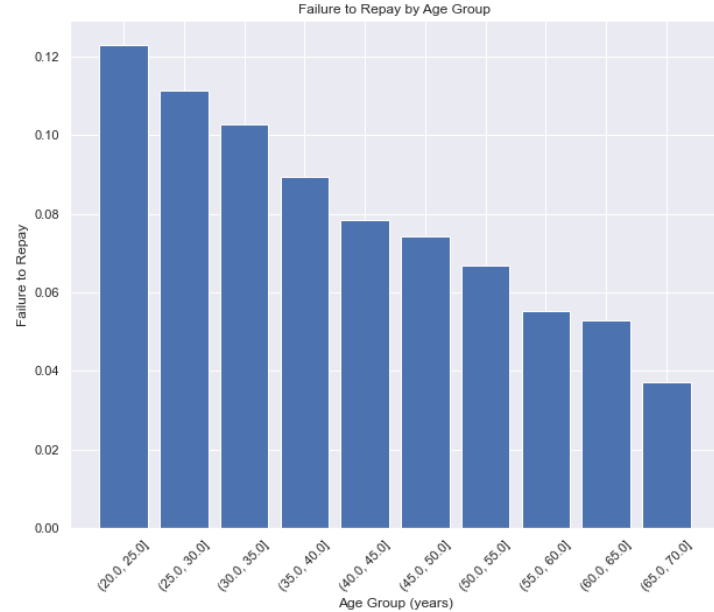
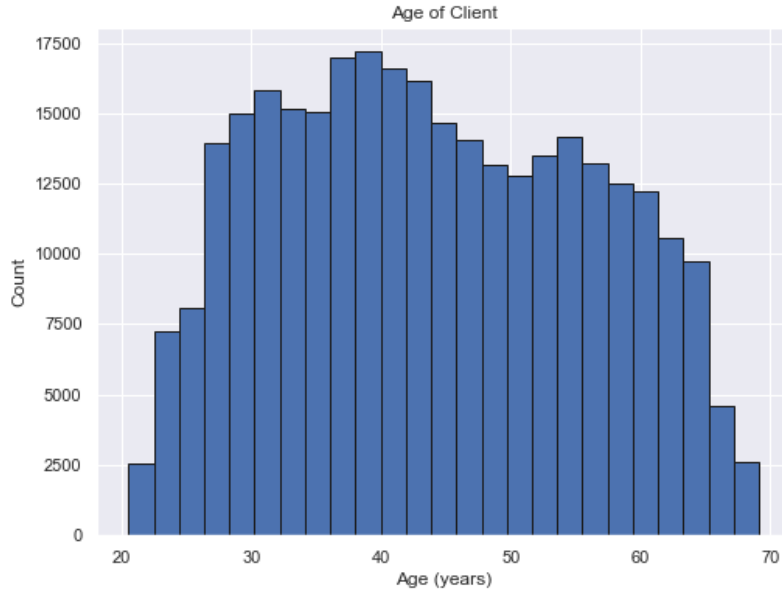
People that have poor credit or non-existent credit histories.

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

EDA and insights

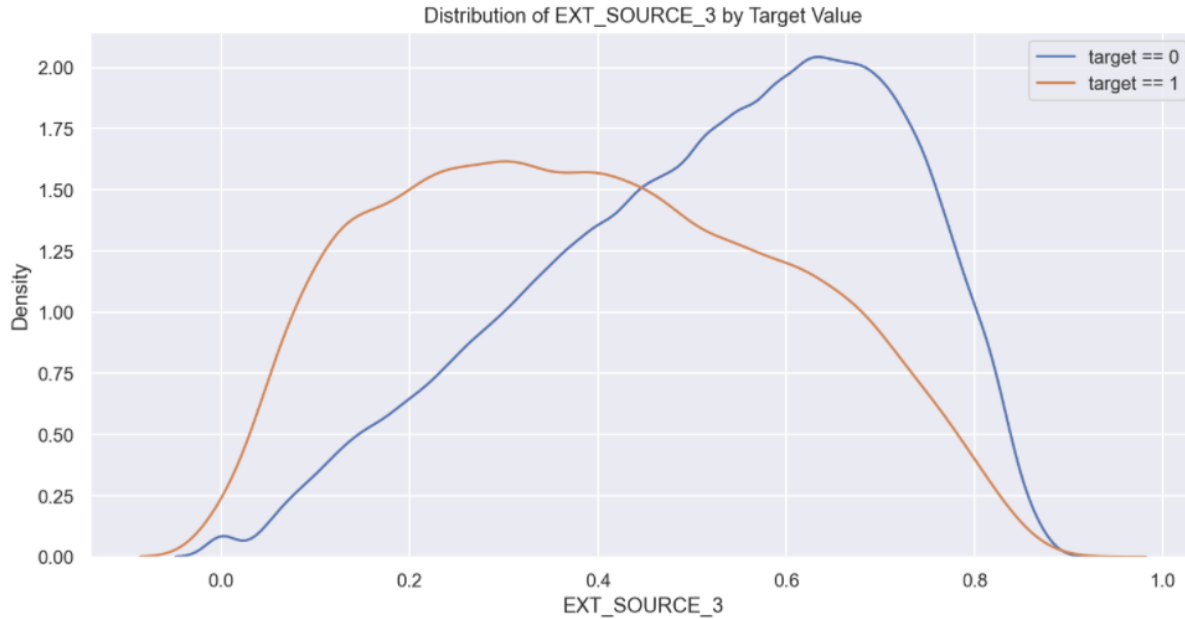


EDA and insights



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

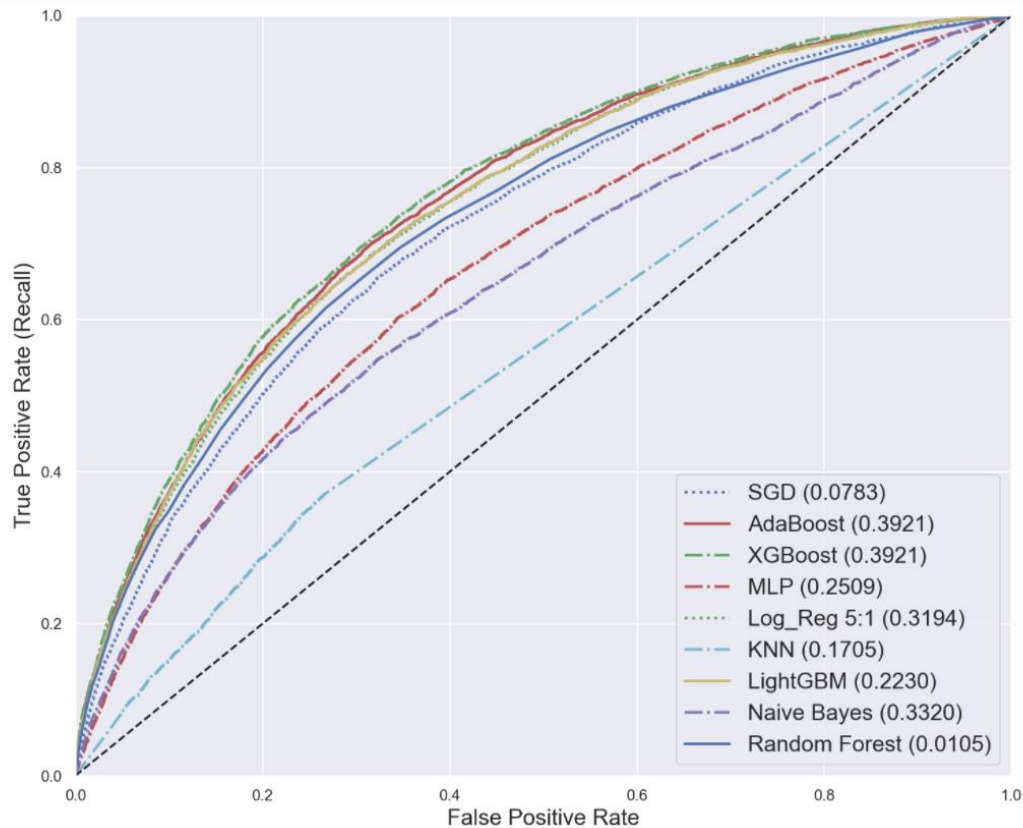
Use the power of data science and machine learning.



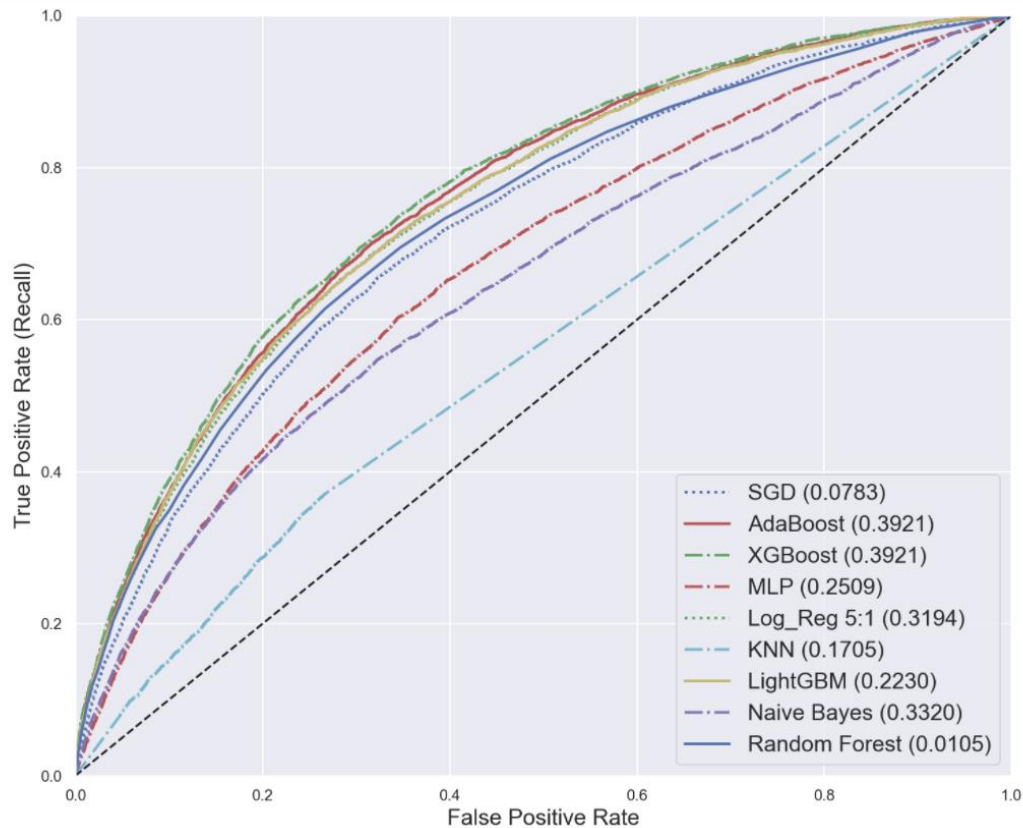
Models comparison (ROC AUC)



Models comparison (ROC AUC)

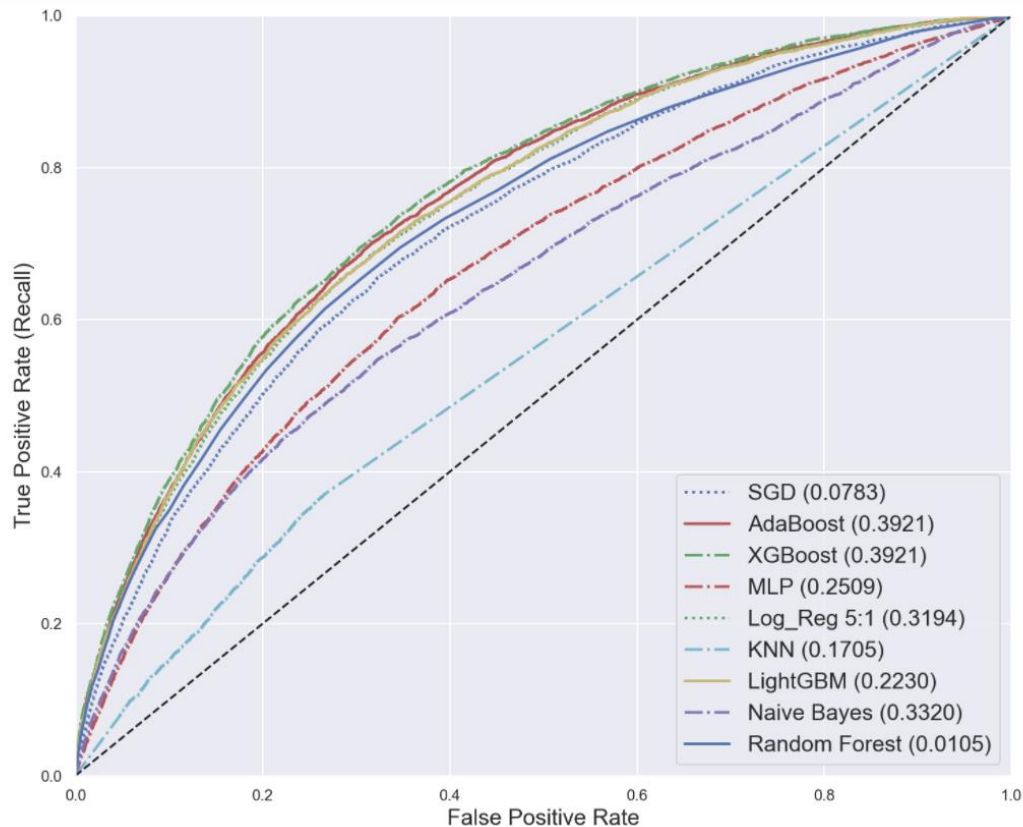


Models comparison (ROC AUC)



XGBoost wins.

Models comparison (ROC AUC)



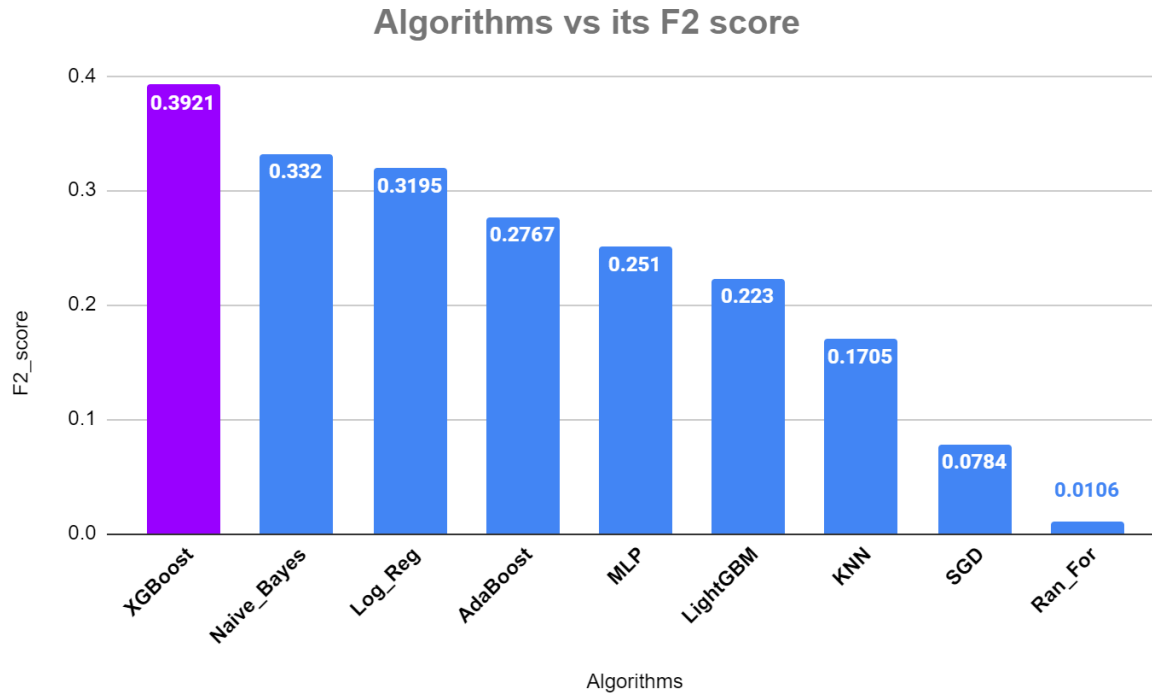
XGBoost wins.
(GBT on steroid)



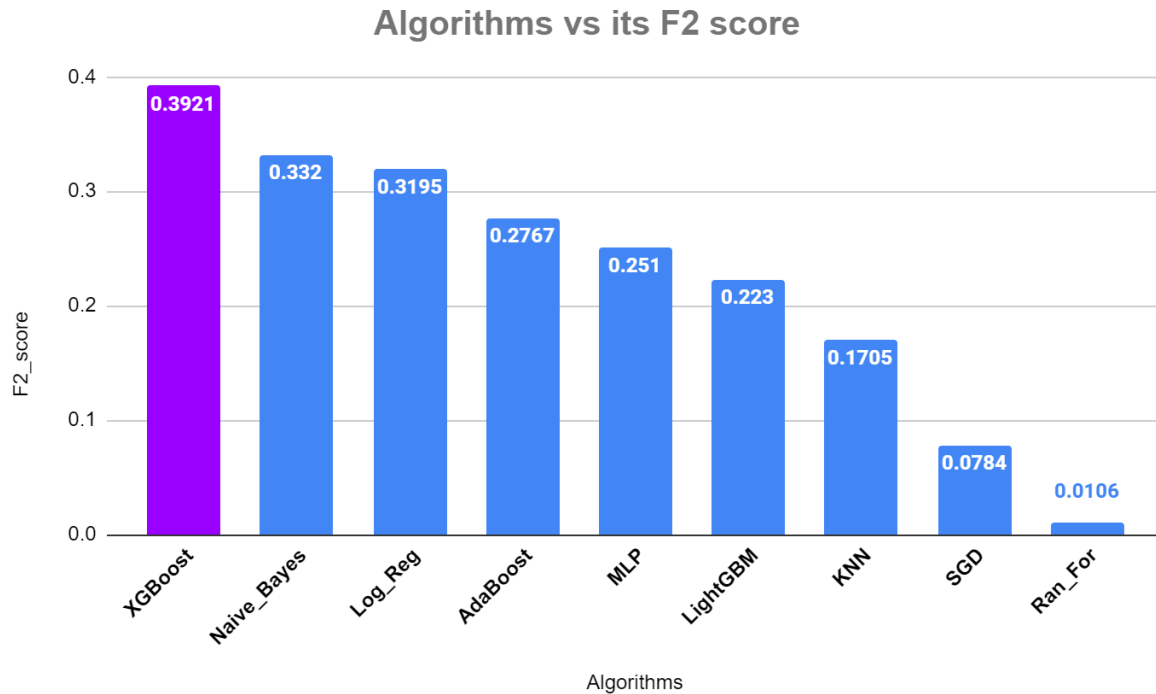
Models comparison (F2 Score)



Models comparison (F2 Score)



Models comparison (F2 Score)



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

Model optimization

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

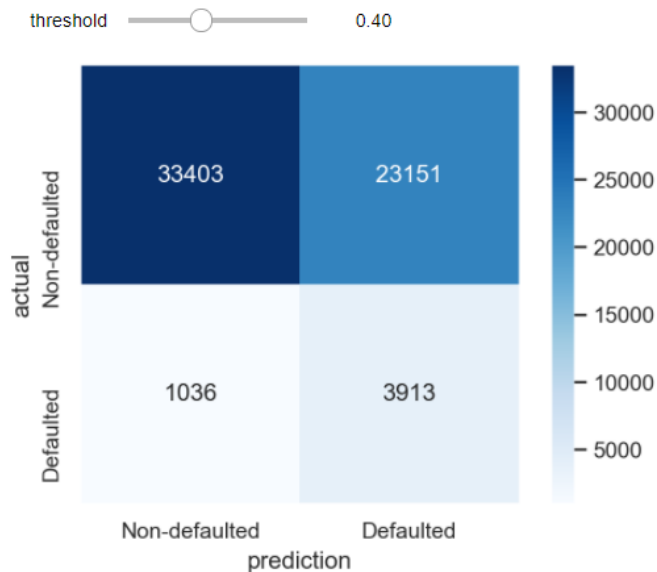
ROC AUC

0.768



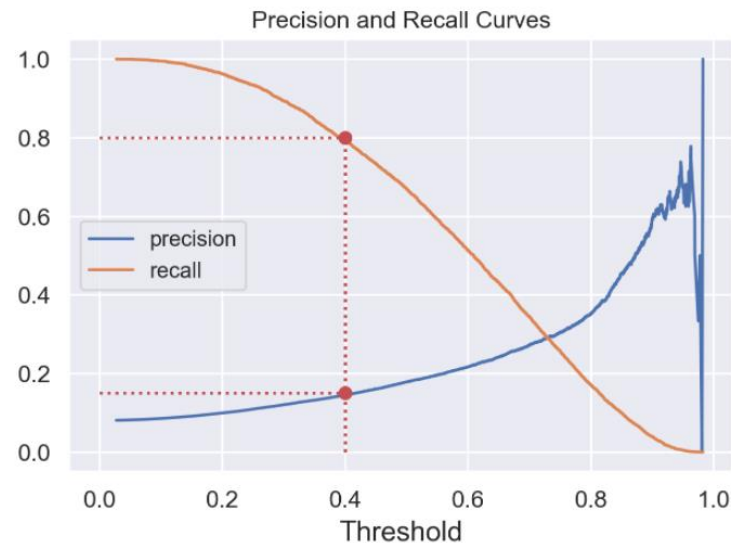
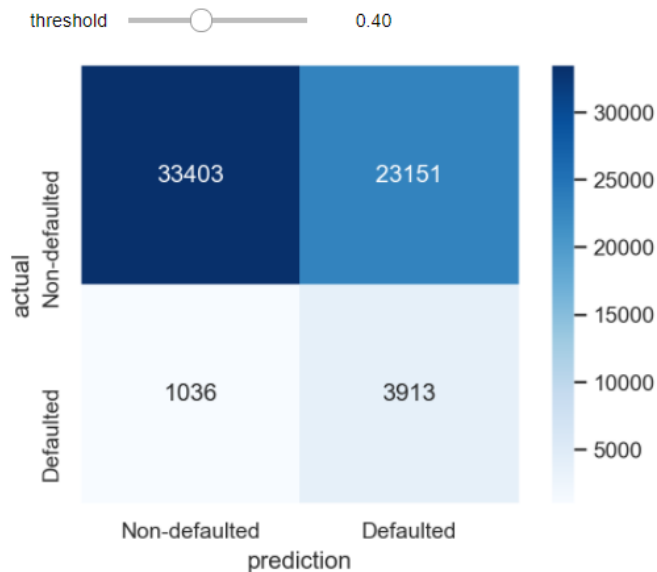
Results

Results



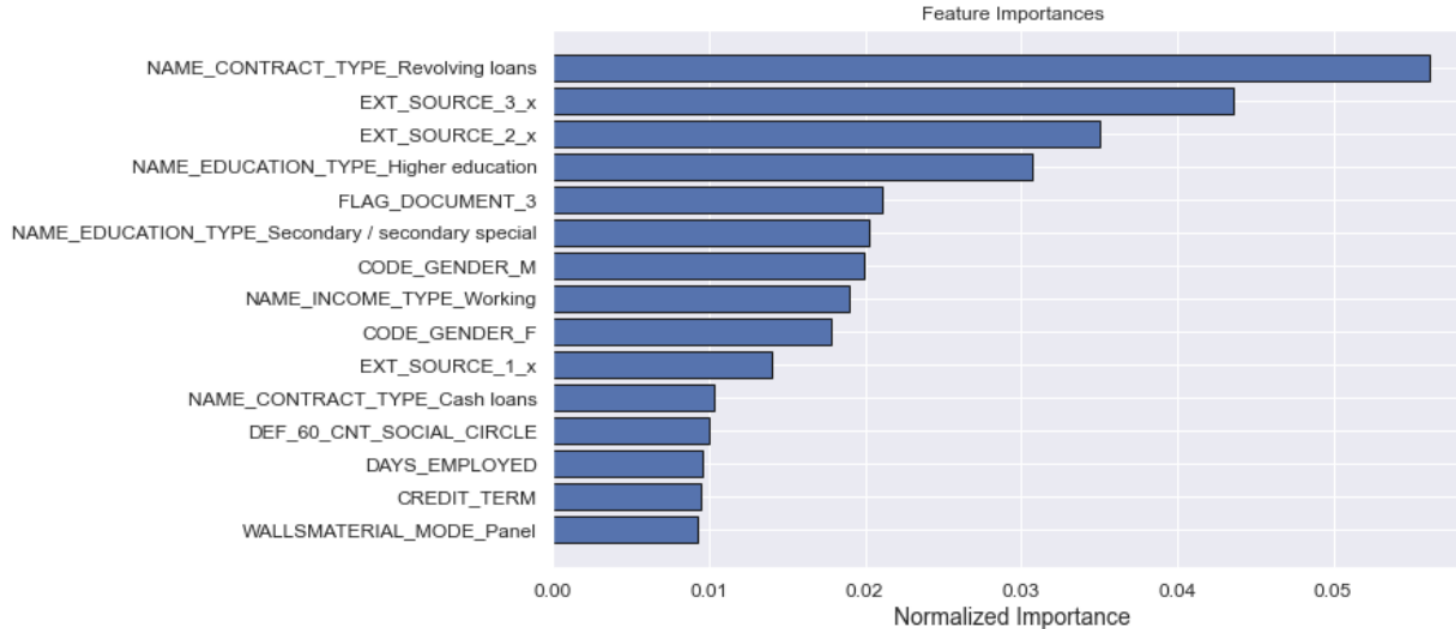
We lean a little bit towards recall (not too strict on precision either)

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.

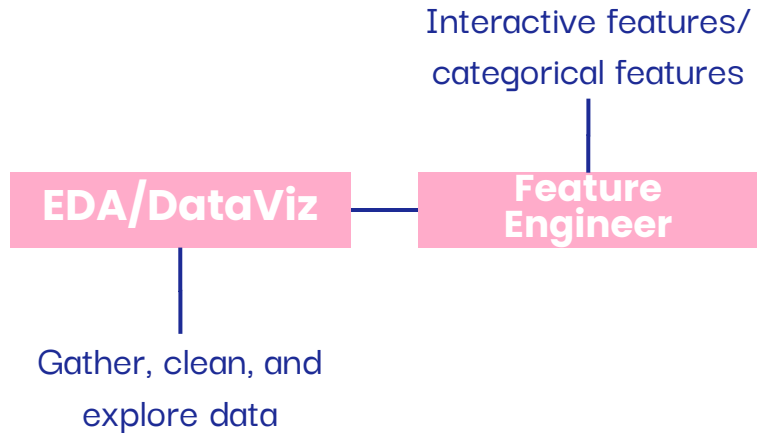
Recap

Recap

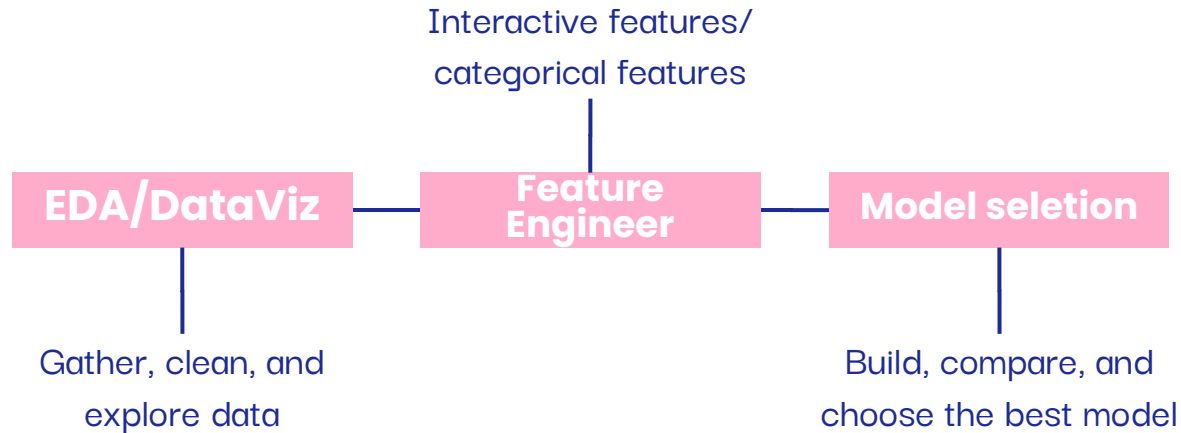
EDA/DataViz

Gather, clean, and
explore data

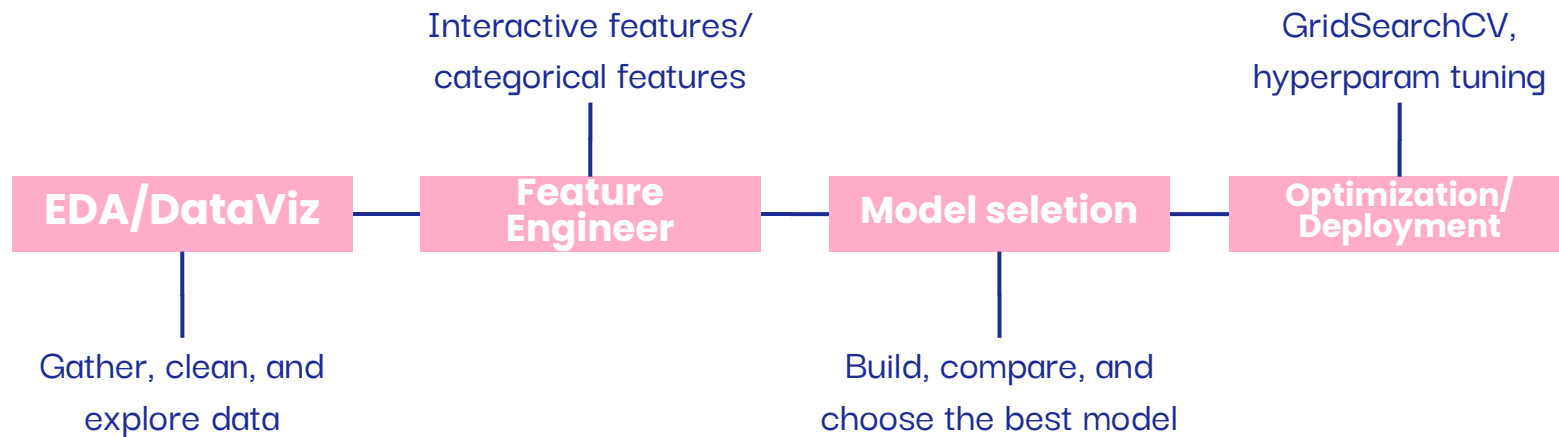
Recap



Recap



Recap



Future Work



Data

Incorporate multiple
datasets



Algorithm

Do better on XGBoost
and LightGBM



Deployment

Build interactive app
and deploy to
streamlit/AWS

How to confuse machine learning





Thank you

Questions?

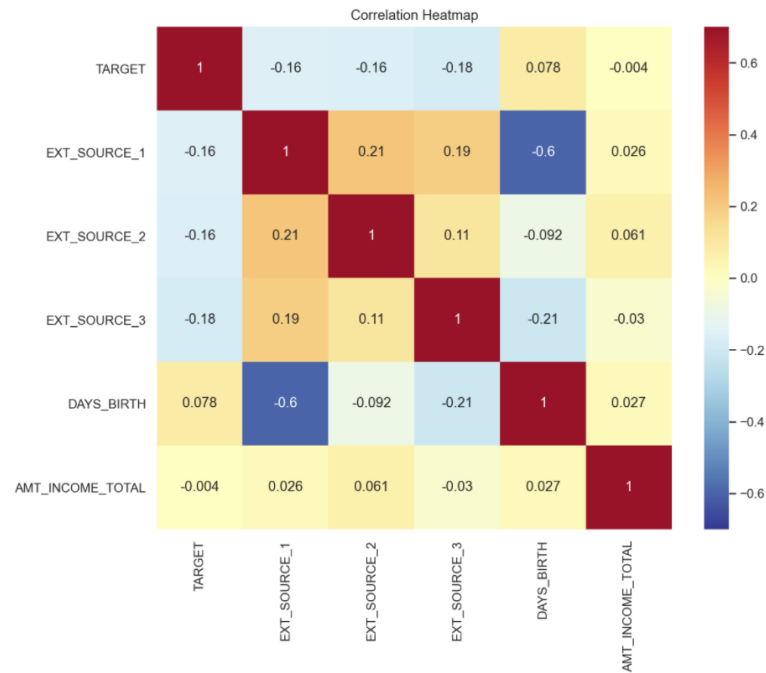
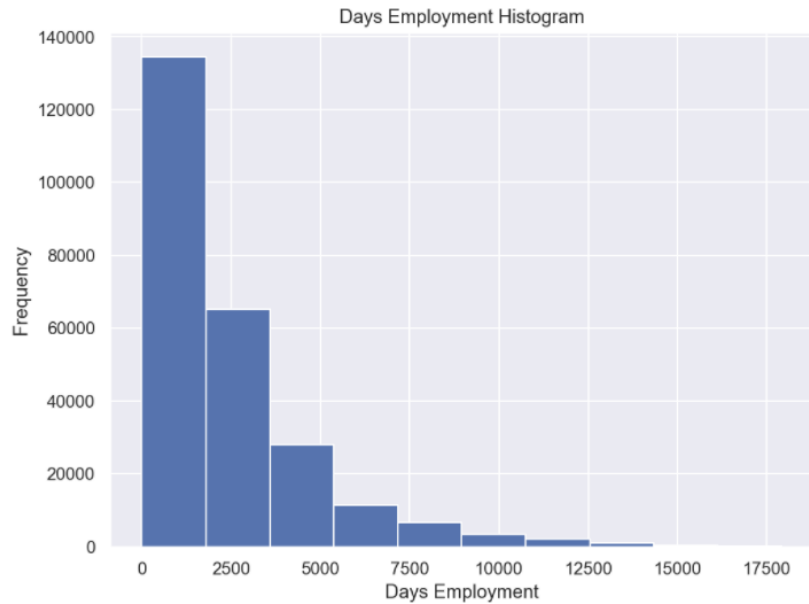
Please reach out to me at



<https://www.linkedin.com/in/luongtruong77/>

Steven L Truong

Appendix



Appendix

