Loan Application Assessment

A Machine Learning Business Proposal
by William Andrian

Introduction

Loans are a way for people to gather capital and finance their current need or goals. Banks and other financial institutions act as lenders and lend out money in a certain amount of time with interest.

To acquire a loan, an applicant's loan application must be assessed and approved by the lender. A usual key metric is the credit score, which is a credit track record of an individual.

Unfortunately, some individuals lack the credit histories to accumulate credit score, especially unbanked people. Which makes lenders reluctant to approve their loan application. This dataset here aims to represent those people to give a fair assessment and extend financial inclusion to them

MARKET ANALYSIS

Trust

A key factor in the financial industry

Rejection

Rejected loan applications could make individuals “scarred” and less likely to try again

Cowling, M., Liu, W. & Calabrese, R. Has previous loan rejection scarred firms from applying for loans during Covid-19?. Small Bus Econ 59, 1327–1350 (2022). https://doi.org/10.1007/s11187-021-00586-2

Defaulting

A defaulted individual could get deterred into loaning again in the future. Lenders could also accumulate losses and lose trust

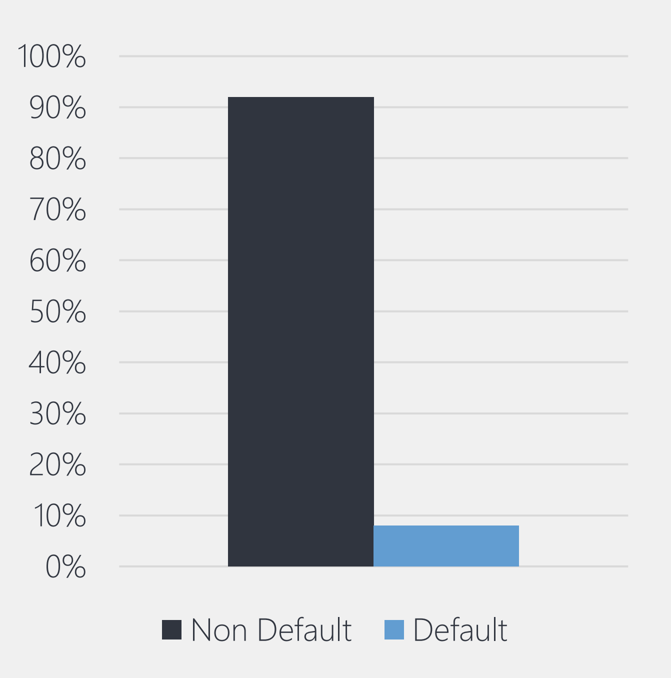
UNBANKED PERCENTAGE

Indonesia has 97.74 million or 48% of the country’s adult population being unbanked source: https://digitalfinance.worldbank.org/country/indonesia#

48%

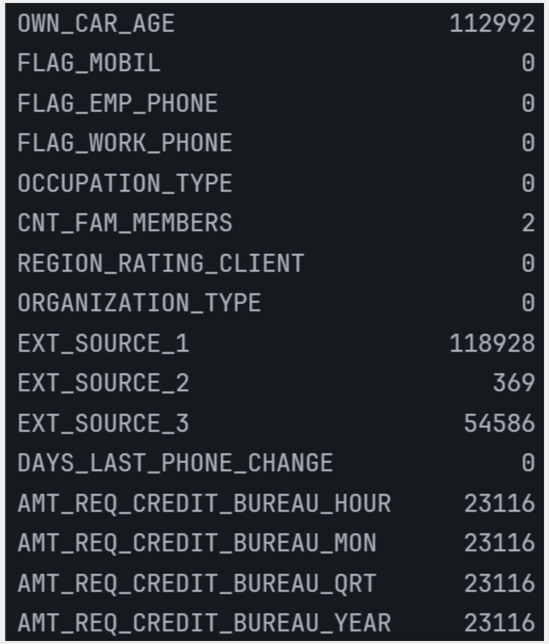
DATA ANALYSIS

Default Rate 8.07%



We have an extremely imbalanced dataset with the Default class only comprising 8.07% of the total applications

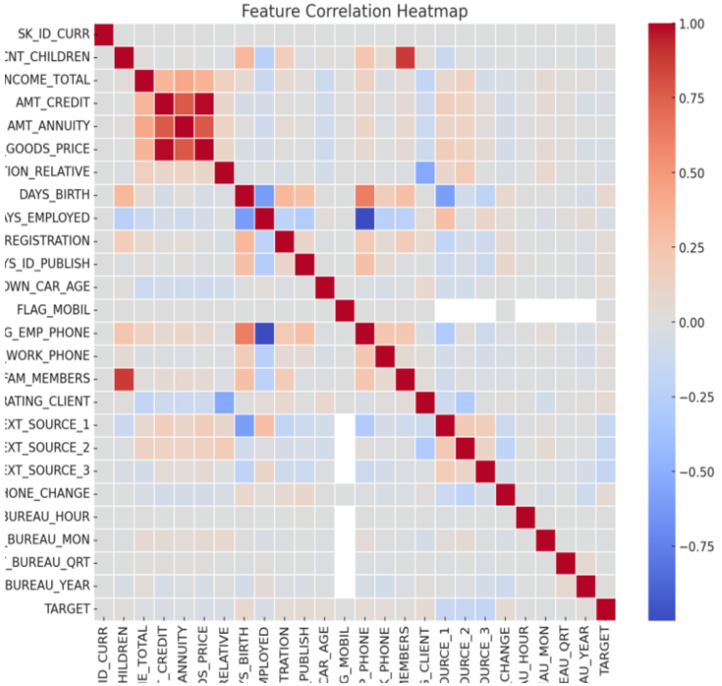
Missing Entries



```
OWN_CAR_AGE                    112992
FLAG_MOBIL                          0
FLAG_EMP_PHONE                      0
FLAG_WORK_PHONE                     0
OCCUPATION_TYPE                     0
CNT_FAM_MEMBERS                     2
REGION_RATING_CLIENT                0
ORGANIZATION_TYPE                   0
EXT_SOURCE_1                   118928
EXT_SOURCE_2                      369
EXT_SOURCE_3                    54586
DAYS_LAST_PHONE_CHANGE              0
AMT_REQ_CREDIT_BUREAU_HOUR      23116
AMT_REQ_CREDIT_BUREAU_MON       23116
AMT_REQ_CREDIT_BUREAU_QRT       23116
AMT_REQ_CREDIT_BUREAU_YEAR      23116
```

Lots of missing values are present in the dataset. Various techniques will need to be implemented

Low Correlation with Target



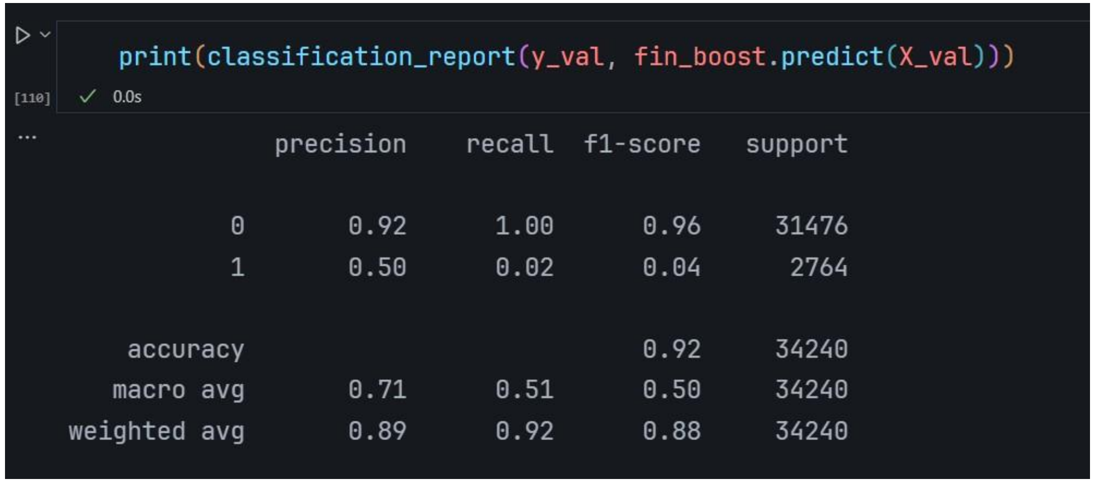
No obvious correlation can be drawn to the target variable. Complex relationships need to be discovered

PROBLEM STATEMENT

This is a severely imbalanced problem where both classes are equally important to us, because denying an application would lose the company customers and approving a likely-to-default application would also lose customers and external trust. So choosing a metric to evaluate our model would be important.

In order to meet those requirements, I've selected PR-AUC and F1 score because they are more sensitive to the minority class. This prevents getting a high score only because the model predicts the majority class. [1]

Example of modelling not sensitive to the minority class. An accuracy of 0.92 seems high but when we remember the majority class was 92%, we can achieve that score by guessing everything as the non-default.



```
print(classification_report(y_val, fin_boost.predict(X_val)))
```

```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     31476
           1       0.50      0.02      0.04      2764

    accuracy                           0.92     34240
   macro avg       0.71      0.51      0.50     34240
weighted avg       0.89      0.92      0.88     34240
```

[1] Saito, Takaya, and Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets." PloS one vol. 10,3 e0118432. 4 Mar. 2015, doi:10.1371/journal.pone.0118432

WHY MACHINE LEARNING?

As found with the Data Analysis, no obvious relations to the target variable can be determined. That means the underlying relationships are complex and not easy to figure out. That makes the perfect case for machine learning, which are great for discovering deep and hidden relationships between variables that are not obvious to the human eye.

Also, the speed and volume a machine learning model is capable to train and test on makes formulating and validating a solution extremely fast compared to slower human based analysis

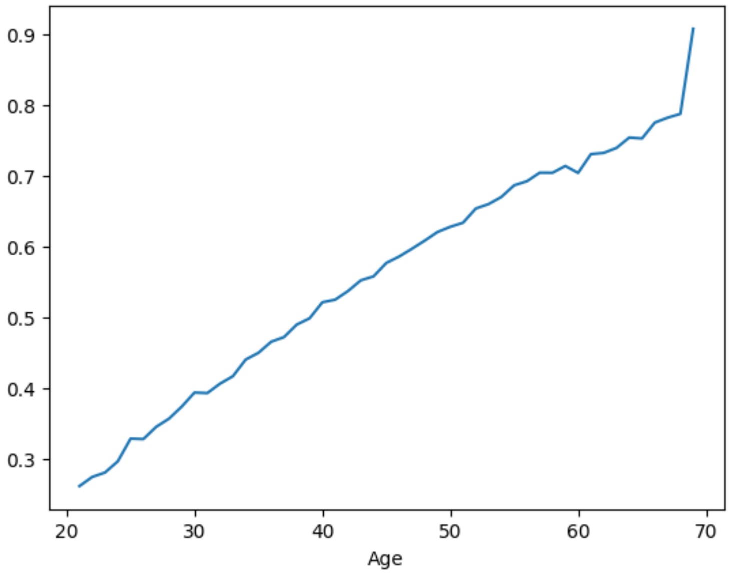
FEATURE ENGINEERING

※EXT_SOURCE

Imputing EXT_SOURCE_1

EXT_SOURCE_1 is strongly correlated with age so I've imputed them based on average values on age

Chart->



Imputing EXT_SOURCE_2 & EXT_SOURCE_3

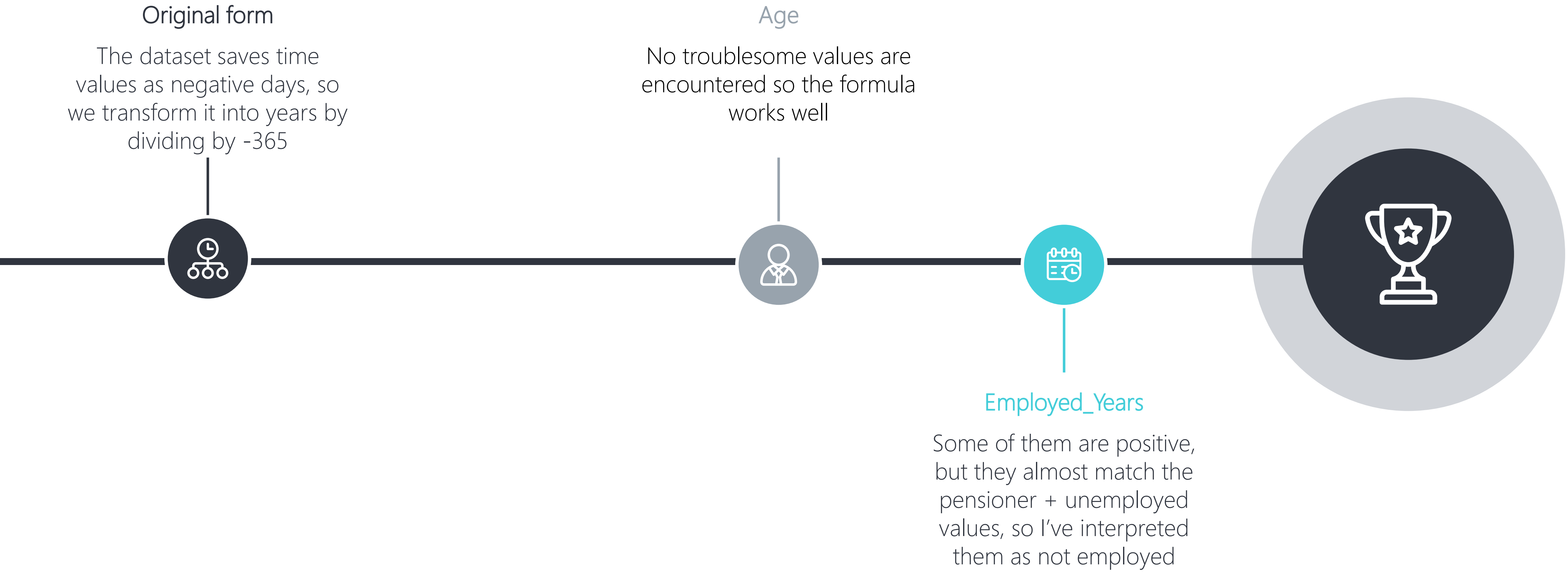
Since they are not strongly correlated to anything, I imputed them based on their mean

RATIOS

Since they're weakly correlated, ratio features could extract more information

FEATURE ENGINEERING

※To Years



Original form

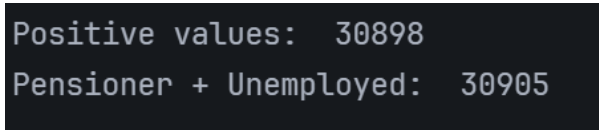
The dataset saves time values as negative days, so we transform it into years by dividing by -365

Age

No troublesome values are encountered so the formula works well

Employed_Years

Some of them are positive, but they almost match the pensioner + unemployed values, so I've interpreted them as not employed



```
Positive values:  30898
Pensioner + Unemployed:  30905
```

FEATURE ENGINEERING

※Neighboring_Mean

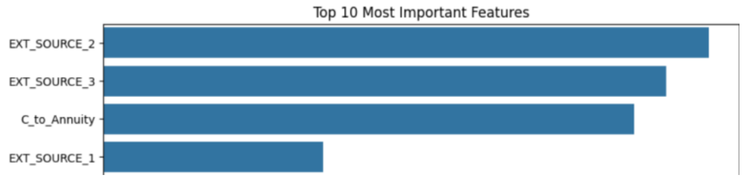
Original Idea

Get the target mean of the nearest k-rows of a row using nearest neighbors. Considered columns are normalized.

Original idea from `https://www.kaggle.com/competitions/home-credit-default-risk/discussion/64821`

Columns Considered

The considered columns are EXT_SOURCE 1 to 3, and Credit to Annuity Ratio since they were found to be important features


Top 10 Most Important Features

Validation

To validate this method, the preprocessing is "trained" on the train split to ensure no data leakages happen during training and validation

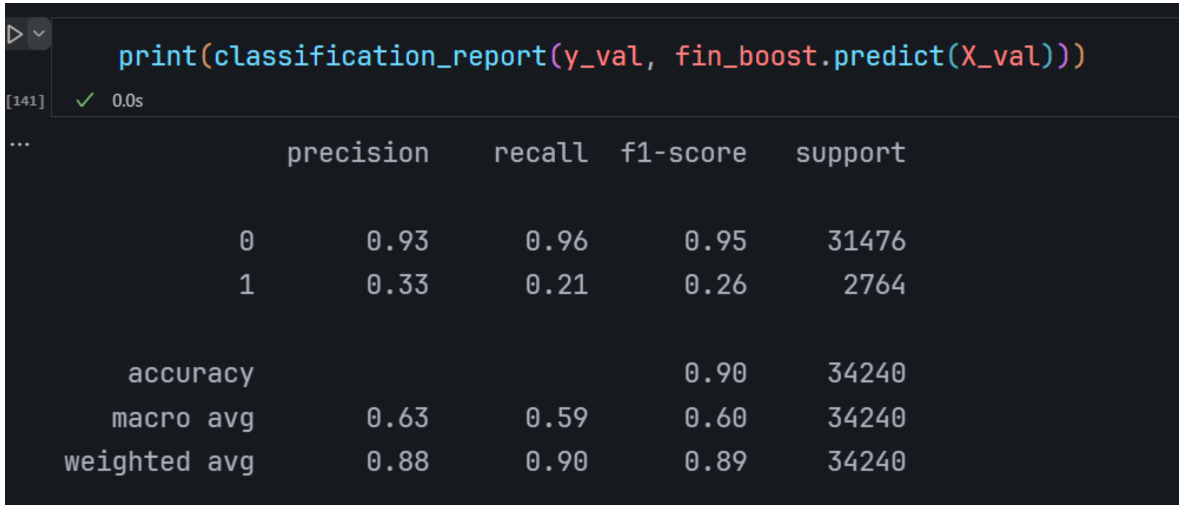
MODELLING

The model first tried out was Catboost and LightGBM since they are both renowned for their performance and speed especially for LightGBM. Both are gradient boosting models which can also reduce overfitting

CatBoost

But after trials, Catboost was the higher performing model and the LightGBM model was dropped

The results were quite well, with me trading off accuracy for a higher F1 macro to more accurately predict the minority class



```
print(classification_report(y_val, fin_boost.predict(X_val)))
```

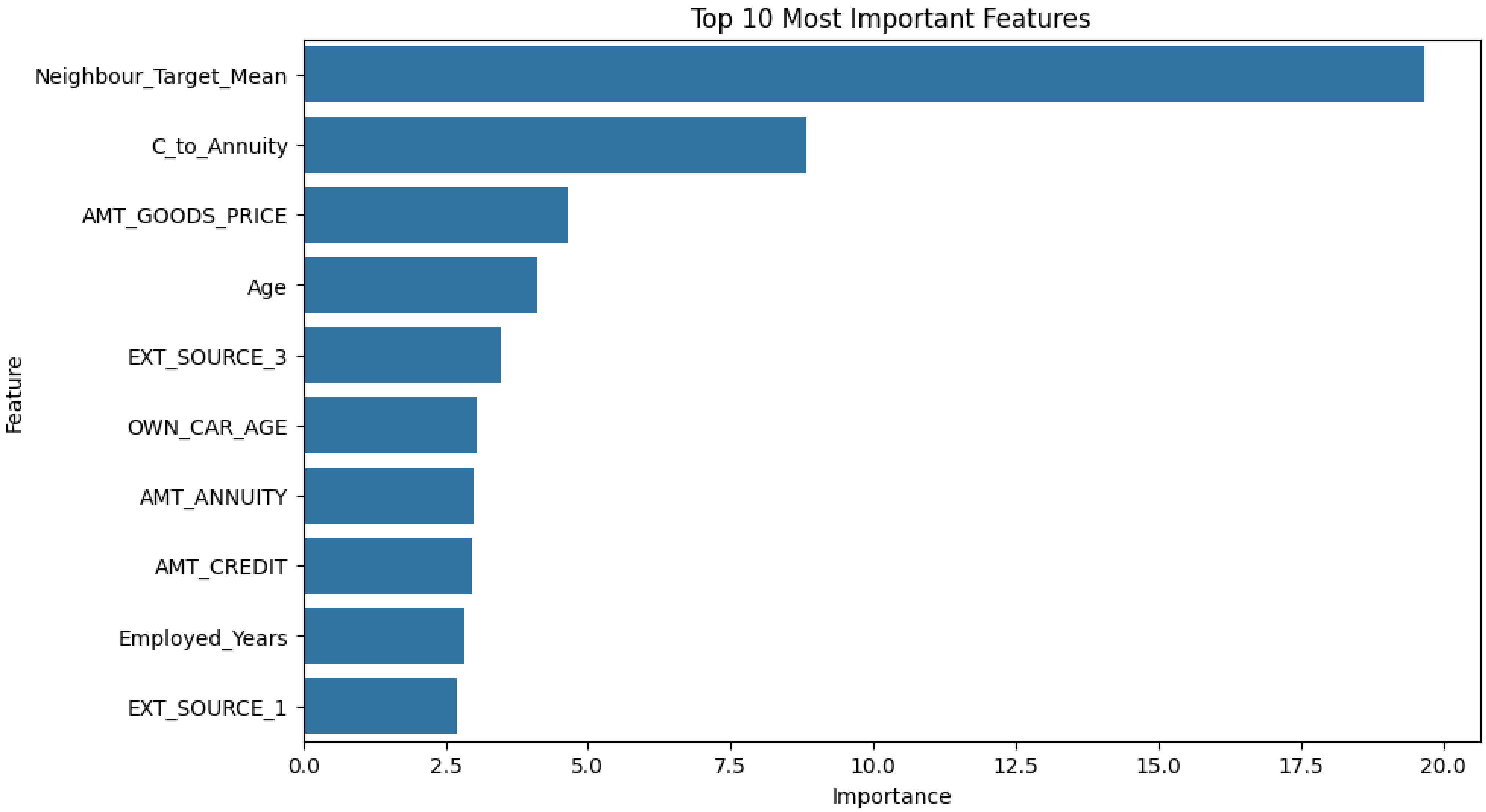
[141] ✓ 0.0s

```
              precision    recall  f1-score   support

           0       0.93      0.96      0.95     31476
           1       0.33      0.21      0.26      2764

    accuracy                           0.90     34240
   macro avg       0.63      0.59      0.60     34240
weighted avg       0.88      0.90      0.89     34240
```

FEATURE IMPORTANCE



FEATURE IMPORTANCE

The feature engineering proved to work with the Neighbour_Target_Mean feature discussed before being the most important feature, followed by C_to_Annuity, another ratio feature engineered, AMT_GOODS_PRICE was untouched, Age was derived from the formula, and EXT_SOURCE_3 was imputed. With Neighbour_Target_Mean being calculated by EXT_SOURCEs and normalized C_to_Annuity, I think that if EXT_SOURCEs contain less missing data which I had to impute, the model would have an improvement.

BUSINESS PROPOSAL

- The model is very performant in predicting, taking less than a second to predict 35,000 rows
- Preprocessing 136,000 rows of data for training takes 5 minutes on a single laptop but a parallelized distributed system could speed up the process significantly.
- The model could be deployed to an app or website to ensure constant availability anywhere and anytime
- Adding these facts to the prediction performance of the model means that using such model could replace the need of human assessment especially for early screenings, and speed up the process since it doesn’t have to wait for a human to assess. This means less waiting time for an application and the applicant can be reached out faster to discuss the results.

CONCLUSION

The model trained would likely help a lender company to assess default risks of applicants. This model is the proposed solution to a company since it will give additional assessments besides from human assessors. A model will be faster and readily available than a human so integrating it as an early checker could be beneficial for the company. Also, a suggestion to improve this model is gathering more information especially EXT_SOURCE because they are very important to this model and reducing the number of missing data of those features would improve this model.

THANK YOU