COMP3948 ASSIGMENT 2

Owen Arando

# Table of Contents

[Introduction 2](#_Toc89619053)

[Exploratory Data Analysis 2](#_Toc89619054)

[Numeric Data 2](#_Toc89619055)

[Numeric Data Dictionary 2](#_Toc89619056)

[Distribution of Significant Numerical Features 3](#_Toc89619057)

[Categorical Data 4](#_Toc89619058)

[Categorical Data Dictionary 4](#_Toc89619059)

[Distribution of Significant Categorical Features 5](#_Toc89619060)

[Distributions of All Features 6](#_Toc89619061)

[Correlations and Heatmaps 7](#_Toc89619062)

[Heatmap of All Correlations 7](#_Toc89619063)

[Heatmap of Significant Features 8](#_Toc89619064)

[The Target 9](#_Toc89619065)

[Distribution of Target Variable 9](#_Toc89619066)

[Significant Features VS Target 9](#_Toc89619067)

[Model Evaluation 11](#_Toc89619068)

[Model Preparation 11](#_Toc89619069)

[Models 11](#_Toc89619070)

[Conclusion 13](#_Toc89619071)

# Introduction

In this report we will attempt to predict whether a customer will commit fraud based on a given dataset. To do this we will perform an Exploratory Data Analysis on the dataset. Then we will create Models using Logistic Regression and SMOTE. Afterwards, we will evaluate our models using cross fold validation and examine their accuracy, precision, recall and f1 scores. Finally, we will choose our preferred model and state our conclusions.

# Exploratory Data Analysis

We will start our EDA by going over the numeric and categorical data in our dataset. For both types of data, we will display a data dictionary, describing key properties of each feature and then displaying the distributions of their significant features.

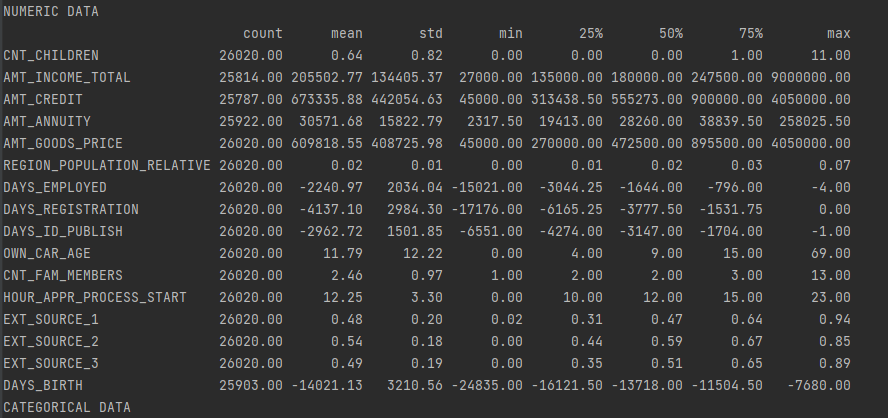
To select these significant features, we found all features with a chi-squared score >= 3.8. Feature selection was also tested with RFE and Forward Feature Selection, but Chi Square produced the best results.

After examining both numeric and categorical features, you will see a visualization of the distribution of all features. Afterwards, we will go over heatmaps and the correlations of all features as well as the correlations of significant features. And finally, we will analyze our target variable and its relationship with significant factors.

## Numeric Data

### Numeric Data Dictionary

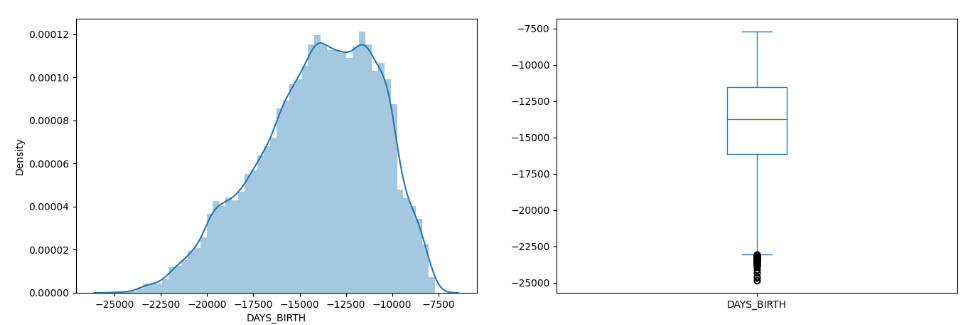
The data dictionary for numeric data can be seen below. In it we can see that AMT\_INCOLME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY and DAYS\_BIRTH have missing values, these will need to be imputed. Additionally, the features with the AMT prefix are also a much higher order of magnitude than the other features. For these features, scaling would help, but as we find out later, these features are insignificant and dropped.

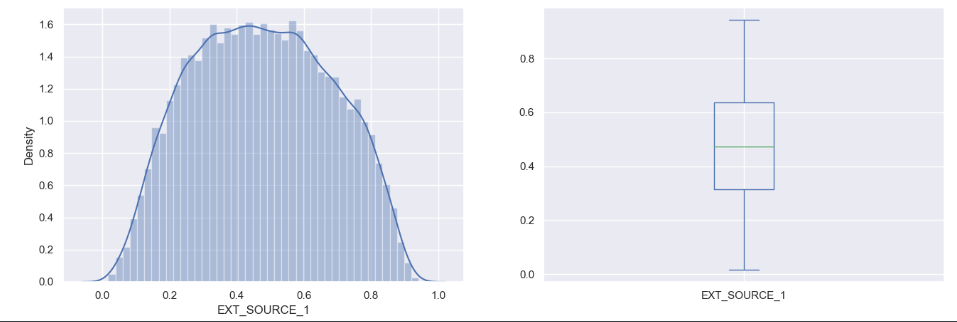


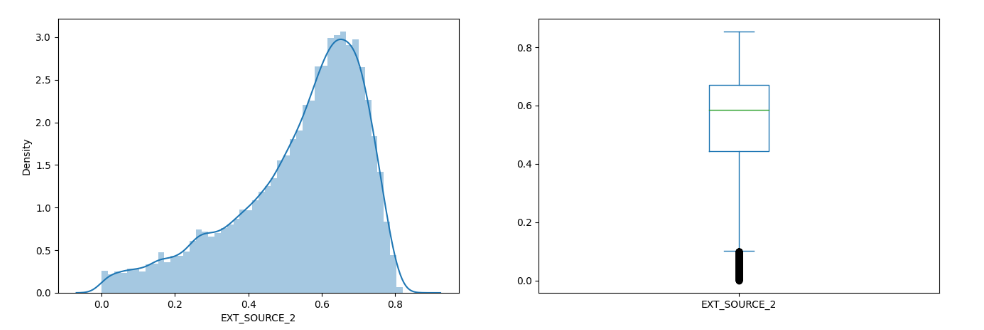
### Distribution of Significant Numerical Features

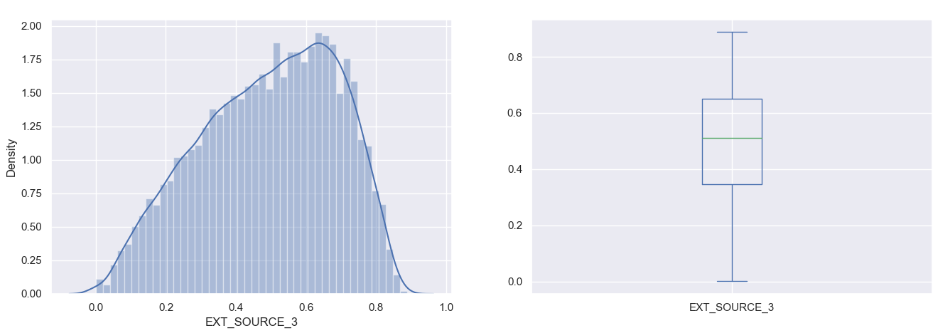
Below you can see the distributions of the most significant Numerical Features. The features with the EXT prefix are by far the most significant numeric features. DAYS\_BIRTH is part of a series of features with all negative values and was the only one to have a chi square score >= 3.8 in our feature selection process.

It should be noted that these values may already be scaled, possibly by a MinMaxScaler as they are all between 0 and 1. Additionally out of these four features only two have outliers and the number of outliers is quite mild compared to other datasets.





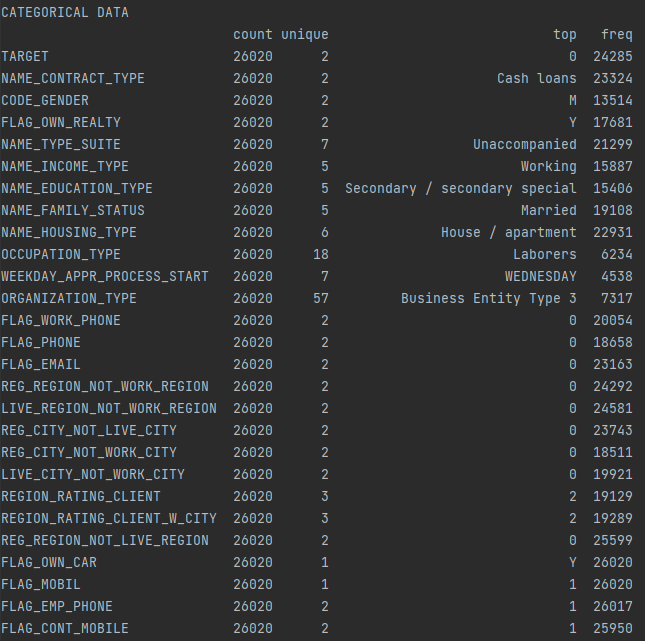




## Categorical Data

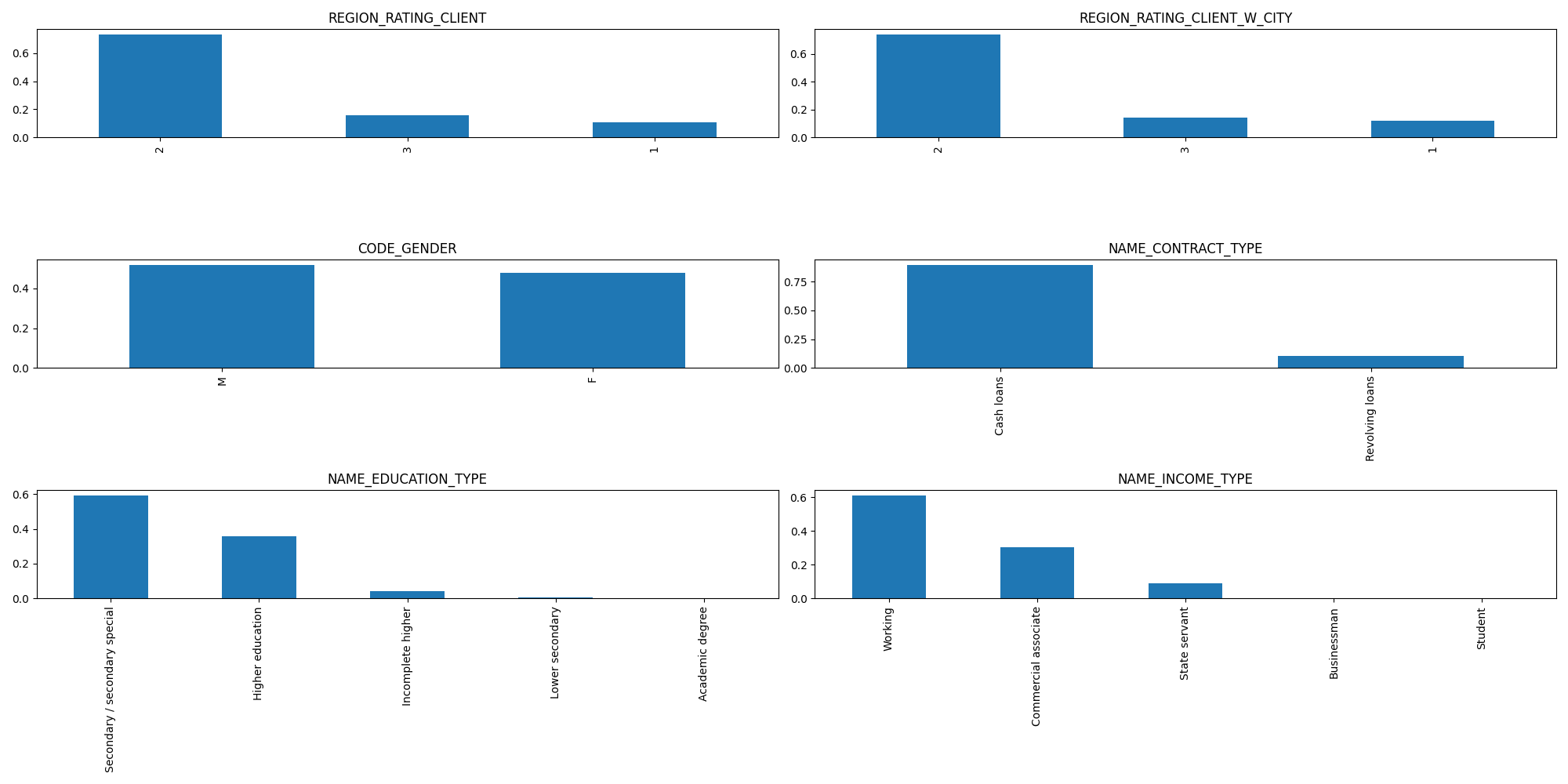
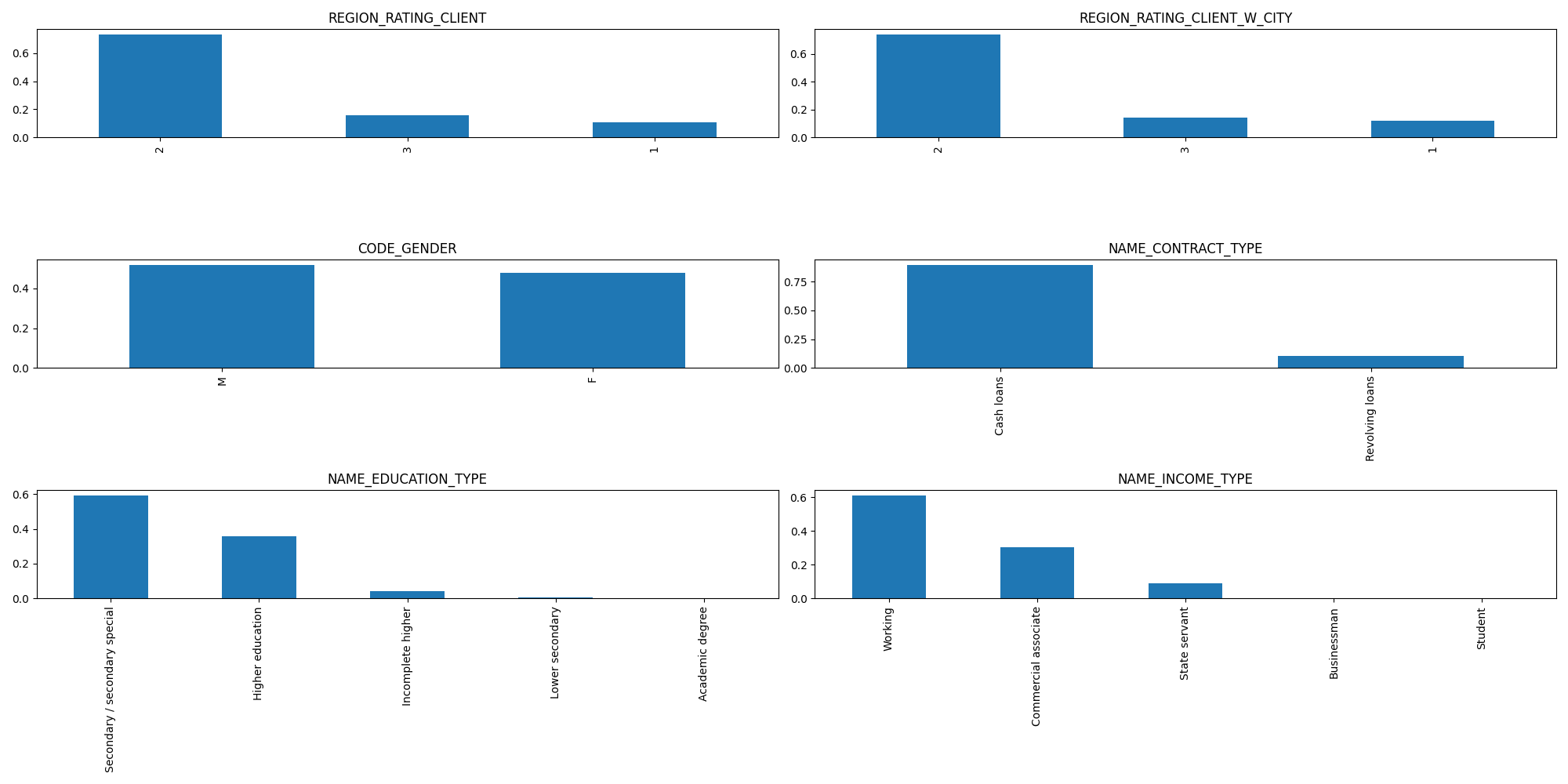
### Categorical Data Dictionary

The data dictionary for categorical data can be seen below. In it we can see that several of our categorical values have only one value or are extremely close to only having one value. These categories: FLAG\_OWN\_CAR, FLAG\_MOBIL, FLAG\_EMP\_PHONE, and FLAG\_CONT\_MOBILE can be removed.



### Distribution of Significant Categorical Features

The distribution of significant categorical features can be seen below. Most of these features have only   
a few categories and are concentrated on an even smaller subset of categories. This makes these categories prime candidates for dummy variables. All these categories were identified as significant by Chi Squared.



## Distributions of All Features

Below we can see histograms of all numeric data and ordinal categorical data—any data that is represented by a number without the need to be dummied.

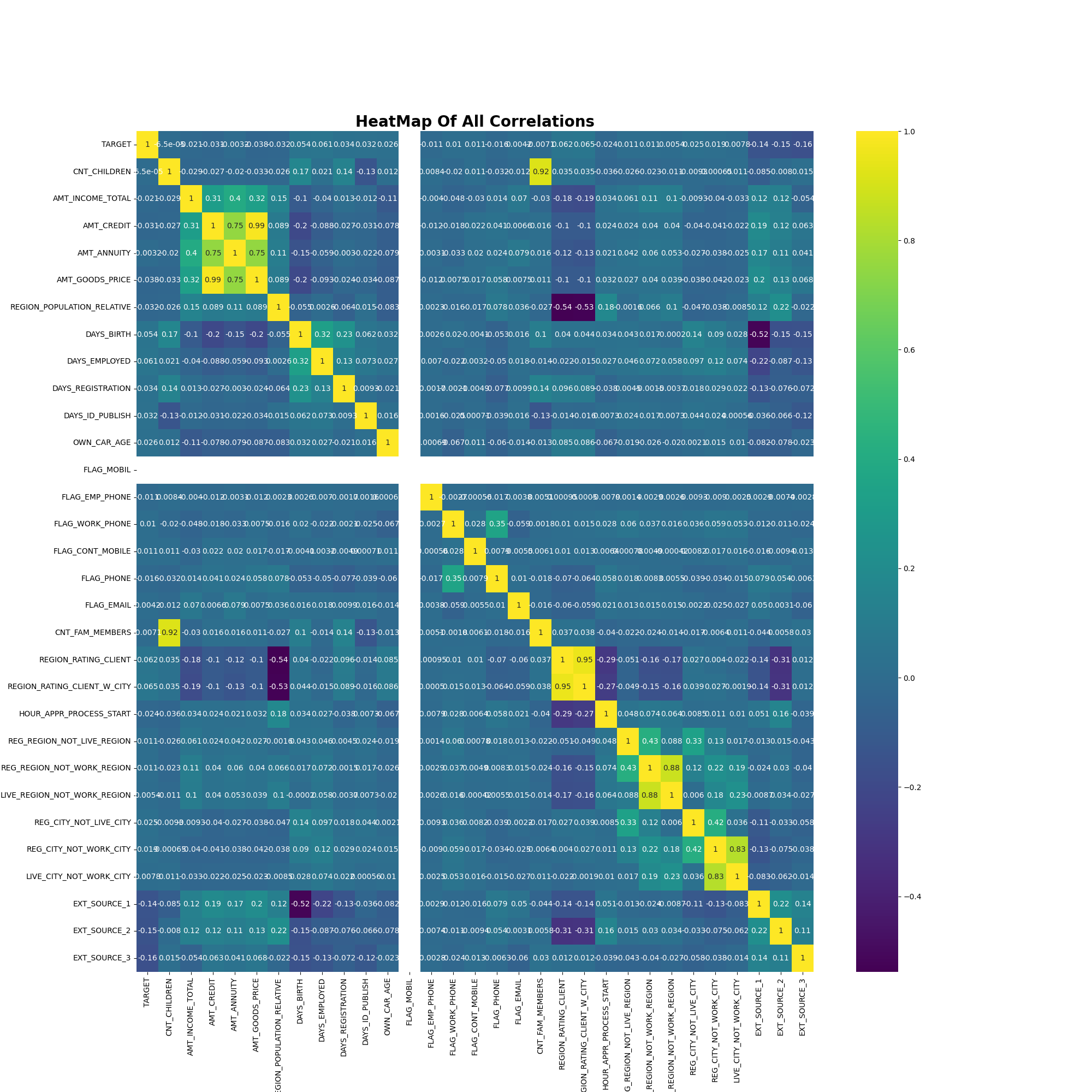
Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

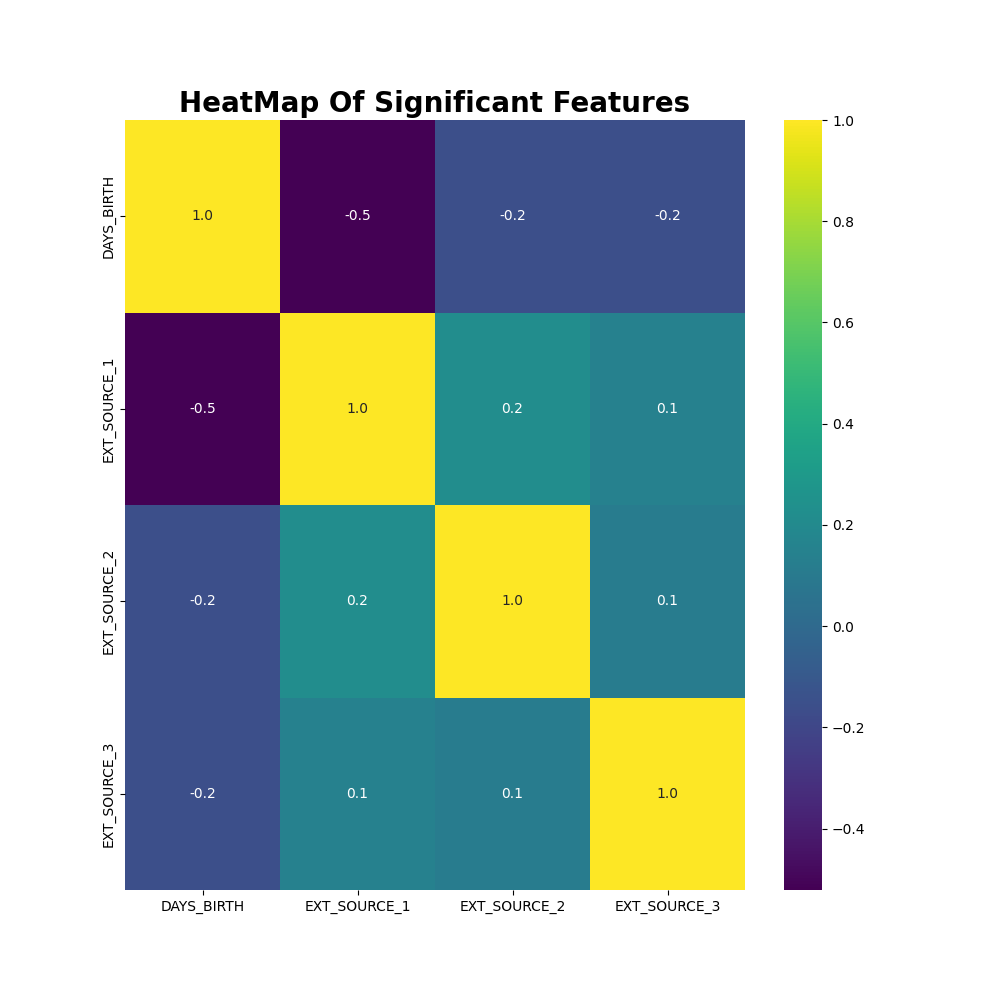
## Correlations and Heatmaps

### Heatmap of All Correlations



### Heatmap of Significant Features

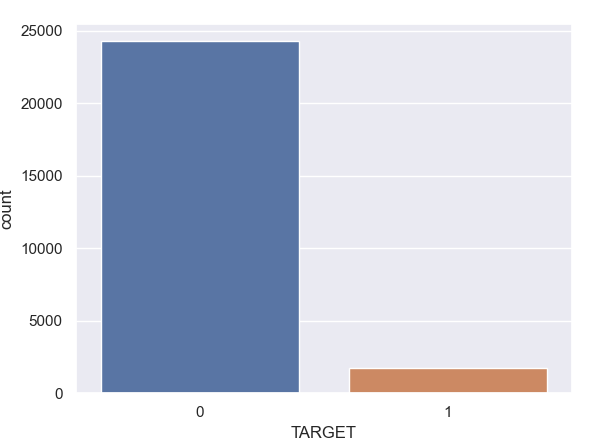
Below you can see a heatmap of the significant numerical model features identified by Chi Square. Most features are independent which is good. However, there is a notable negative correlation between EXT\_SOURCE\_1 and DAYS\_BIRTH.



## The Target

### Distribution of Target Variable

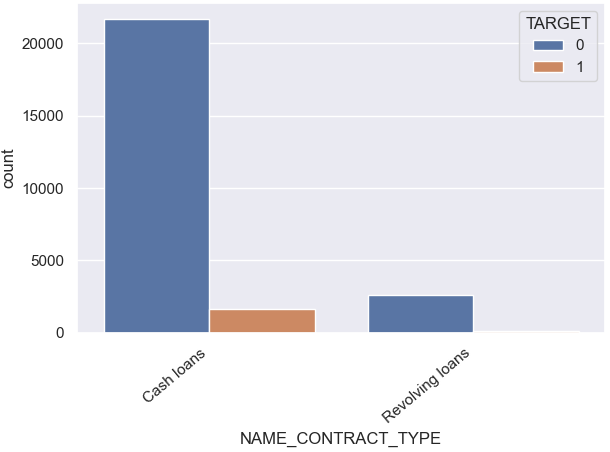
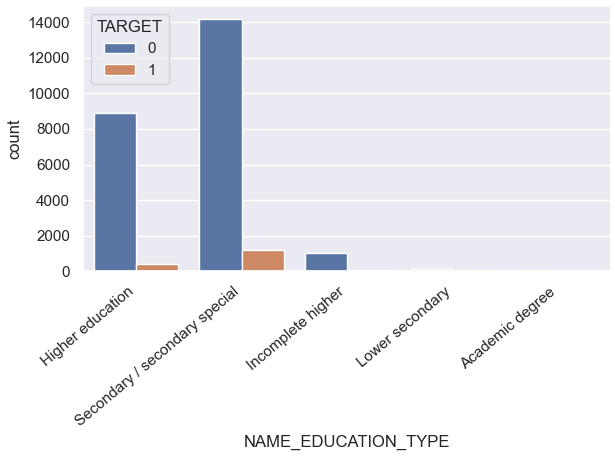
Below you can see the distribution of the target variable. In this case, our data is heavily unbalanced. Most of our data is for samples in which no fraud has occurred. This resembles reality but will make for poor predictions. We will need to use SMOTE to resample our dataset.

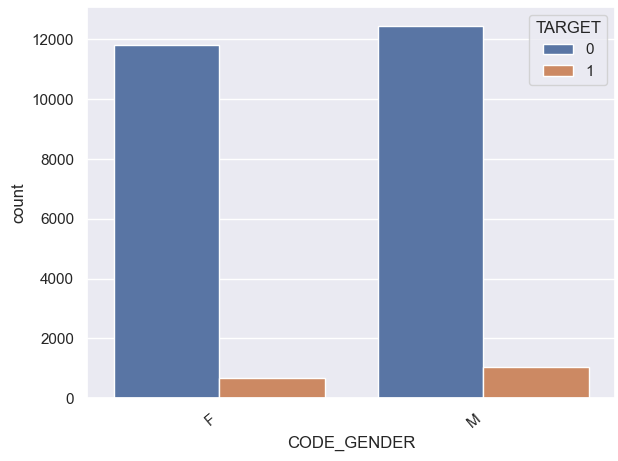


### Significant Features VS Target

Shown below are significant categorical features and their relationships to target variables. When looking for significance, one should look at the ratio of the target value for each category, not the overall amount of data in a category. For example, in CODE\_GENDDER even though the amount of data is almost equivalent for men and women it is clear that proportionally, men are more likely to commit fraud than women.



# Model Evaluation

In this section we will explain our modelling process. We will first outline the data preparation behind the models, then have a look at the models themselves and finally, we will select a model.

## Model Preparation

All models below share the following preparations. First categorical features with only one value were removed. Then categorical features were dummied. For these, we had K-1 dummies made for K values in a category to be efficient with the number of features produced. After creating dummies, we had 155 features. After all our features could be processed as a numerical value, we used KNN imputation to fill in the missing values we had for some numerical features. Then, we reduced these 155 features down to 41 using Chi Square as explained previously in the EDA. From this point we made several models based on binning and further reducing our selected features to keep our models as simple as possible. All models were evaluated using Cross Fold Validation.

## Models

The models can be seen and are described below.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | A | B | C |
| Features | ORGANIZATION\_TYPE$Trade: type 3  ORGANIZATION\_TYPE$Legal Services  DAYS\_BIRTH  NAME\_FAMILY\_STATUS$Married  ORGANIZATION\_TYPE$Construction  ORGANIZATION\_TYPE$School  FLAG\_PHONE$1  ORGANIZATION\_TYPE$Self-employed  OCCUPATION\_TYPE$Low-skill Laborers  NAME\_HOUSING\_TYPE$Rented apartment  NAME\_FAMILY\_STATUS$Civil marriage  NAME\_INCOME\_TYPE$Commercial associate  REG\_CITY\_NOT\_WORK\_CITY$1  NAME\_FAMILY\_STATUS$Single / not married  ORGANIZATION\_TYPE$Bank  ORGANIZATION\_TYPE$Government  OCCUPATION\_TYPE$Managers  ORGANIZATION\_TYPE$Business Entity Type 3  OCCUPATION\_TYPE$Sales staff  OCCUPATION\_TYPE$Security staff  NAME\_INCOME\_TYPE$Working  ORGANIZATION\_TYPE$Transport: type 3  OCCUPATION\_TYPE$High skill tech staff  OCCUPATION\_TYPE$Core staff  OCCUPATION\_TYPE$Laborers  REG\_CITY\_NOT\_LIVE\_CITY$1  NAME\_INCOME\_TYPE$State servant  OCCUPATION\_TYPE$Accountants  OCCUPATION\_TYPE$Drivers  NAME\_HOUSING\_TYPE$With parents  CODE\_GENDER$M  NAME\_CONTRACT\_TYPE$Revolving loans  NAME\_EDUCATION\_TYPE$Secondary / secondary special  EXT\_SOURCE\_2  REGION\_RATING\_CLIENT$1  REGION\_RATING\_CLIENT\_W\_CITY$1  EXT\_SOURCE\_1  REGION\_RATING\_CLIENT$3  EXT\_SOURCE\_3  NAME\_EDUCATION\_TYPE$Higher education  REGION\_RATING\_CLIENT\_W\_CITY$3 | ORGANIZATION\_TYPE$Trade: type 3  ORGANIZATION\_TYPE$Legal Services  NAME\_FAMILY\_STATUS$Married  ORGANIZATION\_TYPE$Construction  ORGANIZATION\_TYPE$School  FLAG\_PHONE$1  ORGANIZATION\_TYPE$Self-employed  OCCUPATION\_TYPE$Low-skill Laborers  NAME\_HOUSING\_TYPE$Rented apartment  NAME\_FAMILY\_STATUS$Civil marriage  NAME\_INCOME\_TYPE$Commercial associate  REG\_CITY\_NOT\_WORK\_CITY$1  NAME\_FAMILY\_STATUS$Single / not married  ORGANIZATION\_TYPE$Bank  ORGANIZATION\_TYPE$Government  OCCUPATION\_TYPE$Managers  ORGANIZATION\_TYPE$Business Entity Type 3  OCCUPATION\_TYPE$Sales staff  OCCUPATION\_TYPE$Security staff  NAME\_INCOME\_TYPE$Working  ORGANIZATION\_TYPE$Transport: type 3  OCCUPATION\_TYPE$High skill tech staff  OCCUPATION\_TYPE$Core staff  OCCUPATION\_TYPE$Laborers  REG\_CITY\_NOT\_LIVE\_CITY$1  NAME\_INCOME\_TYPE$State servant  OCCUPATION\_TYPE$Accountants  OCCUPATION\_TYPE$Drivers  NAME\_HOUSING\_TYPE$With parents  CODE\_GENDER$M  NAME\_CONTRACT\_TYPE$Revolving loans  NAME\_EDUCATION\_TYPE$Secondary / secondary special  EXT\_SOURCE\_2  REGION\_RATING\_CLIENT$1  REGION\_RATING\_CLIENT\_W\_CITY$1  EXT\_SOURCE\_1  REGION\_RATING\_CLIENT$3  EXT\_SOURCE\_3  NAME\_EDUCATION\_TYPE$Higher educationa  REGION\_RATING\_CLIENT\_W\_CITY$3  DAYS\_BIRTH$(-24835.0, -23691.333]  DAYS\_BIRTH$(-23691.333, -22547.667]  DAYS\_BIRTH$(-22547.667, -21404.0]  DAYS\_BIRTH$(-21404.0, -20260.333]  DAYS\_BIRTH$(-20260.333, -19116.667]  DAYS\_BIRTH$(-19116.667, -17973.0]  DAYS\_BIRTH$(-17973.0, -16829.333]  DAYS\_BIRTH$(-16829.333, -15685.667]  DAYS\_BIRTH$(-15685.667, -14542.0]  DAYS\_BIRTH$(-14542.0, -13398.333]  DAYS\_BIRTH$(-13398.333, -12254.667]  DAYS\_BIRTH$(-12254.667, -11111.0]  DAYS\_BIRTH$(-11111.0, -9967.333]  DAYS\_BIRTH$(-9967.333, -8823.667] | REGION\_RATING\_CLIENT\_W\_CITY\_3  NAME\_EDUCATION\_TYPE\_Higher education  EXT\_SOURCE\_3  REGION\_RATING\_CLIENT\_3  EXT\_SOURCE\_1  REGION\_RATING\_CLIENT\_W\_CITY\_1  REGION\_RATING\_CLIENT\_1  EXT\_SOURCE\_2  NAME\_EDUCATION\_TYPE\_Secondary / secondary special  NAME\_CONTRACT\_TYPE\_Revolving loans  CODE\_GENDER\_M  NAME\_HOUSING\_TYPE\_With parents  OCCUPATION\_TYPE\_Drivers  OCCUPATION\_TYPE\_Accountants  NAME\_INCOME\_TYPE\_State servant  REG\_CITY\_NOT\_LIVE\_CITY\_1  OCCUPATION\_TYPE\_Laborers  OCCUPATION\_TYPE\_Core staff  OCCUPATION\_TYPE\_High skill tech staff  ORGANIZATION\_TYPE\_Transport: type 3  NAME\_INCOME\_TYPE\_Working  OCCUPATION\_TYPE\_Security staff  OCCUPATION\_TYPE\_Sales staff  ORGANIZATION\_TYPE\_Business Entity Type 3  OCCUPATION\_TYPE\_Managers  ORGANIZATION\_TYPE\_Government  DAYS\_BIRTH\_BINS\_(-24835.0, -23119.5]  DAYS\_BIRTH\_BINS\_(-23119.5, -21404.0]  DAYS\_BIRTH\_BINS\_(-21404.0, -19688.5]  DAYS\_BIRTH\_BINS\_(-19688.5, -17973.0]  DAYS\_BIRTH\_BINS\_(-17973.0, -16257.5]  DAYS\_BIRTH\_BINS\_(-16257.5, -14542.0]  DAYS\_BIRTH\_BINS\_(-14542.0, -12826.5]  DAYS\_BIRTH\_BINS\_(-12826.5, -11111.0]  DAYS\_BIRTH\_BINS\_(-11111.0, -9395.5] |
| Number of Features | 41 | 54 | 35 |
| Dummied | Yes | Yes | Yes |
| Binned | No | Yes | Yes |
| Avg. Accuracy | 0.6957340507302076 | 0.769292851652575 | 0.7121445042275173 |
| Std. Accuracy | 0.010376704529035385 | 0.006962554574831821 | 0.010654741016585319 |
| Avg. Precision | 0.13718994156085168 | 0.15205089780562867 | 0.13948581405893565 |
| Std. Precision | 0.00764451505849101 | 0.012531239479750078 | 0.012723189267146464 |
| Avg. Recall | 0.6741565753184371 | 0.5378085903115423 | 0.6400277198754254 |
| Std. Recall | 0.025129241055697065 | 0.043053653137572104 | 0.03773513038186724 |
| Avg .F1 | 0.22787514734390038 | 0.23686853939917568 | 0.22895521485730588 |
| Std. F1 | 0.011030689696610245 | 0.018103733409393963 | 0.01927567686203369 |
| Confusion Matrix |  |  |  |

Model A is our most basic model and is a good control for our other two models. Model A is composed of the significant features analyzed from our initial set of 155 original and dummied features. Model A performs the worst among the three models when purely looking at F1 scores, and it has the second highest number of features.

Binning was used for the next two models. Binning for the most part produced only mild improvements, if any. Binning was done for both our most significant and least significant numerical columns and produced mixed results. Due to the volatility of binning, we decided to bin significant, but not the most significant features. We did this to mitigate the risk that if binning ultimately provided unfavorable results for our model on a larger dataset, at least it would not affect our most significant features. We were still able to find mild improvements with binning the DAYS\_BIRTH feature using 10 bins. However, it should be noted that these bins increased the number of features by a decent margin.

The first model with bins, Model B is the best performing model, with an F1 score above 0.23. 0.23 is a threshold that many of the hundreds of models tested could not pass. That being said, it has the highest number of variables and the difference in f1 score is not too far off from Model C. Additionally while these F1 scores were calculated with cross fold validation, our models may perform differently on new data. For this reason, we chose Model C as our best model. it has the second best F1 score, 0.2289 and it has the least amount of features by far. Model C has a lower accuracy and precision than Model B, but it has a higher recall which should be an important statistic when trying to detect fraud.

# Conclusion

Now we will end the report an analysis of the strengths of Model C, how it can help to improve fraud detection and insights gained for future improvements.

Model C may not have as high of an F1 score as Model B, but it has a recall score that is ~20% higher than the recall score of Model B and only slightly lower than that of Model A. For cases like fraud where it is crucial to identify all positive cases, recall matters. Now, this does come with a caveat, The precision score of Model C (as well as our other models) is very low. This is because all the models are overpredicting fraud. This can also be seen in the confusion matrixes of the models, where there is a significant number of false positives. However, in cases like fraud where it is crucial to identify when it occurs this may be an acceptable trade off. Because when investigating fraud, it may be better to ask for forgiveness than for permission.

Model C could be used as an initial warning for fraud detection. A warning that fraud may be occurring, and further investigation could be required. Because of its decent recall score, it may do a decent job of not missing any cases of fraud, but it should not be used as a sure-fire indicator that fraud is occurring due to its poor precision.

In addition to this while developing our models, it was clear that SMOTE had a positive impact and it improved all metrics for all models by a large margin. And SMOTE is an excellent technique for solving problems with unbalanced datasets like detecting fraud.

As for likely improvements, our models had a large number of features, so factor analysis and PCA could reduce the number of features in our model, reduce collinearity where present and make our models easier to interpret.