What is the business problem in need of solving and what are the main questions at hand?

I plan to analyze Kaggle Dataset as shared by Home Credit Group, an International consumer financial provider. The problem that the client wants to solve by throwing open data challenge in Kaggle is as follows:

- Predict customer ability to repay loan as indicated by TARGET variable in the dataset where TARGET = 0 implies customer is able to repay loan while TARGET = 1 customer's difficulty in repaying loan
- Find correlation among independent feature set that includes exhaustive list of 122 independent features
- Which independent features are influencing the TARGET prediction
- Whether influence is in direct relation or inverse relation to TARGET prediction
- Impact of normalized score of External Source data
- Is there pattern in data distribution among correlated feature set that may help in prediction of TARGET

Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

Founded in 1997, Home Credit Group is an international consumer finance provider with operations in 10 countries. They focus on responsible lending

primarily to people with little or no credit history. 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.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. So there is running challenge in Kaggle which is the source for this project dataset. Through this project Client wants:

- What features among list of 122 in the dataset indicate customer’s ability to repay loan
- Is Customer’s age (DAYS_BIRTH) may be one criteria to assess prediction of ‘Repay Loan Credibility’
- If Loan in Cash or Revolving (NAME_CONTRACT_TYPE) impact the prediction
- External Sources (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) are normalized score from external data source if carry influence on prediction
- Correlation of the data features with the predictive ‘TARGET’ variable
- Is there any data distribution pattern among predictive features that may be useful to determine future data results.

What important fields and information does the data set have?

Important fields in our data include:

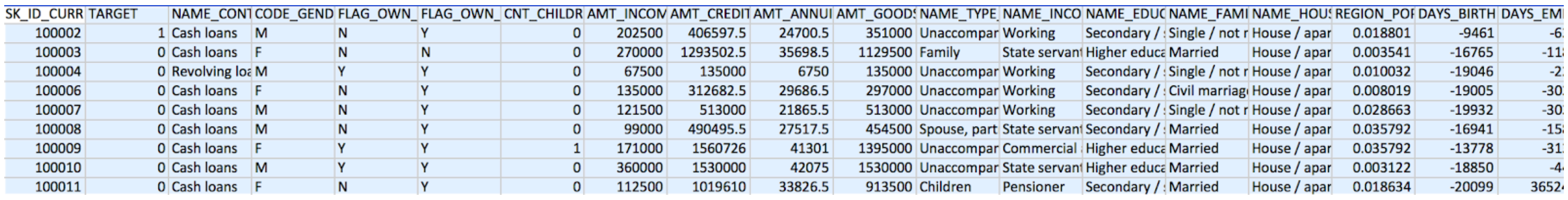
- Identification if loan is cash or revolving
- Gender of the client
- Normalized score from external data source 1
- Normalized score from external data source 2
- Normalized score from external data source 3
- Flag if the client owns a car
- Flag if client owns a house or flat
- Number of children 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, ...)

Exploration of the data shows differing Target prediction results based on impacts of wide feature datasets. Correlation analysis helped narrow focus on

meaningful feature set that is significantly affecting the TARGET prediction. Like customers in higher age group as 45-65 years are likeable to repay loan.

What kind of cleaning and wrangling did you need to do?

Predicting whether or not a client will repay a loan or have difficulty is a critical business need, and Home Credit is hosting this competition on Kaggle to see what sort of models the machine learning community can develop to help them in this task. application_train/application_test: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK_ID_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.
These data are readily and publicly available at
https://www.kaggle.com/c/home-credit-default-risk/data and appear as below:



SK_ID_CURR	TARGET	NAME_CONT	CODE_GEND	FLAG_OWN_	FLAG_OWN_	CNT_CHILDR	AMT_INCOM	AMT_CREDIT	AMT_ANNUI	AMT_GOODS	NAME_TYPE_	NAME_INCO	NAME_EDUC	NAME_FAMI	NAME_HOUS	REGION_POI	DAYS_BIRTH	DAYS_EMI
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.018801	-9461	-6
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher educa	Married	House / apar	0.003541	-16765	-11
100004	0	Revolving loa	M	Y	Y	0	67500	135000	6750	135000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.010032	-19046	-2
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompar	Working	Secondary /	Civil marriag	House / apar	0.008019	-19005	-30
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.028663	-19932	-30
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, part	State servant	Secondary /	Married	House / apar	0.035792	-16941	-15
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompar	Commercial	Higher educa	Married	House / apar	0.035792	-13778	-31
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompar	State servant	Higher educa	Married	House / apar	0.003122	-18850	-4
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary /	Married	House / apar	0.018634	-20099	3652

The datasets required extensive data wrangling for it involved not only fundamental steps of data preparation but also feature engineering and data imputation to be run during Machine Learning:

1. Extracting Data
2. Identifying Target Dataset among Multiple Data Sources
3. Identifying Missing Data
4. Identifying Data Types of the Feature Set into Non-Categorical and Categorical
5. Casting Data Types per Need
6. Feature Engineering (Date timestamp, One Hot Encoding)

Here is the detailed codebook showing above steps:

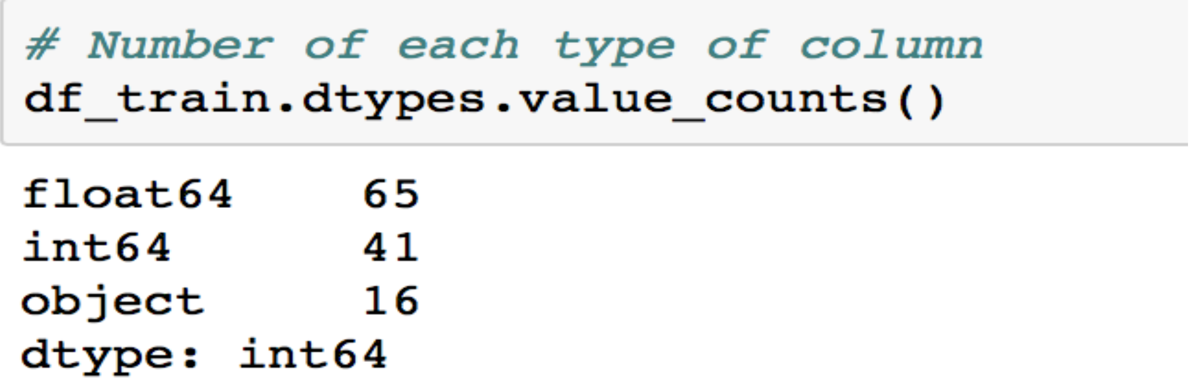
https://github.com/rashi-n/Machine-Learning-Projects/blob/master/Capstone%20Projects/Capstone%20II%20Project/Data%20Wrangling.ipynb

Any preliminary exploration you’ve performed and your initial findings.

Dataset carries nearly 307K records with 121 features and one ‘TARGET’ variable to infer predictions on each transaction.

There were 67 columns that had missing values

Of the 122 features, 106 were non-categorical and 16 were categorical:



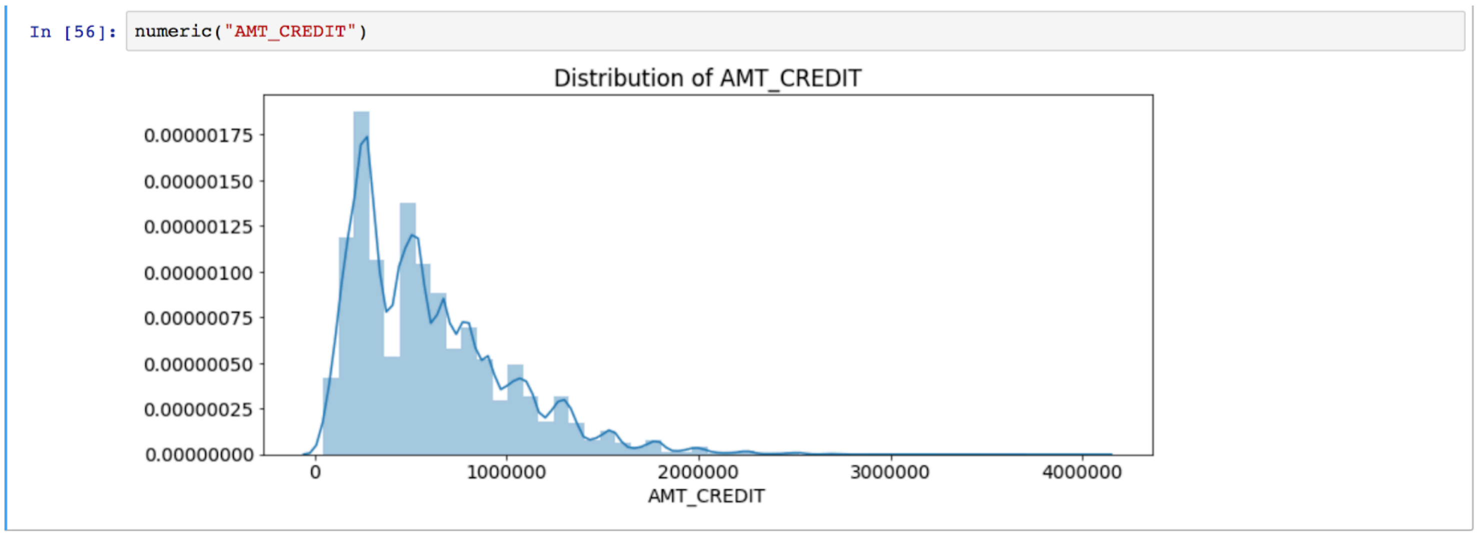
```
# Number of each type of column
df_train.dtypes.value_counts()
```

```
float64     65
int64       41
object      16
dtype: int64
```

1. Exploratory Data Analysis

Visualized distribution of numeric independent features

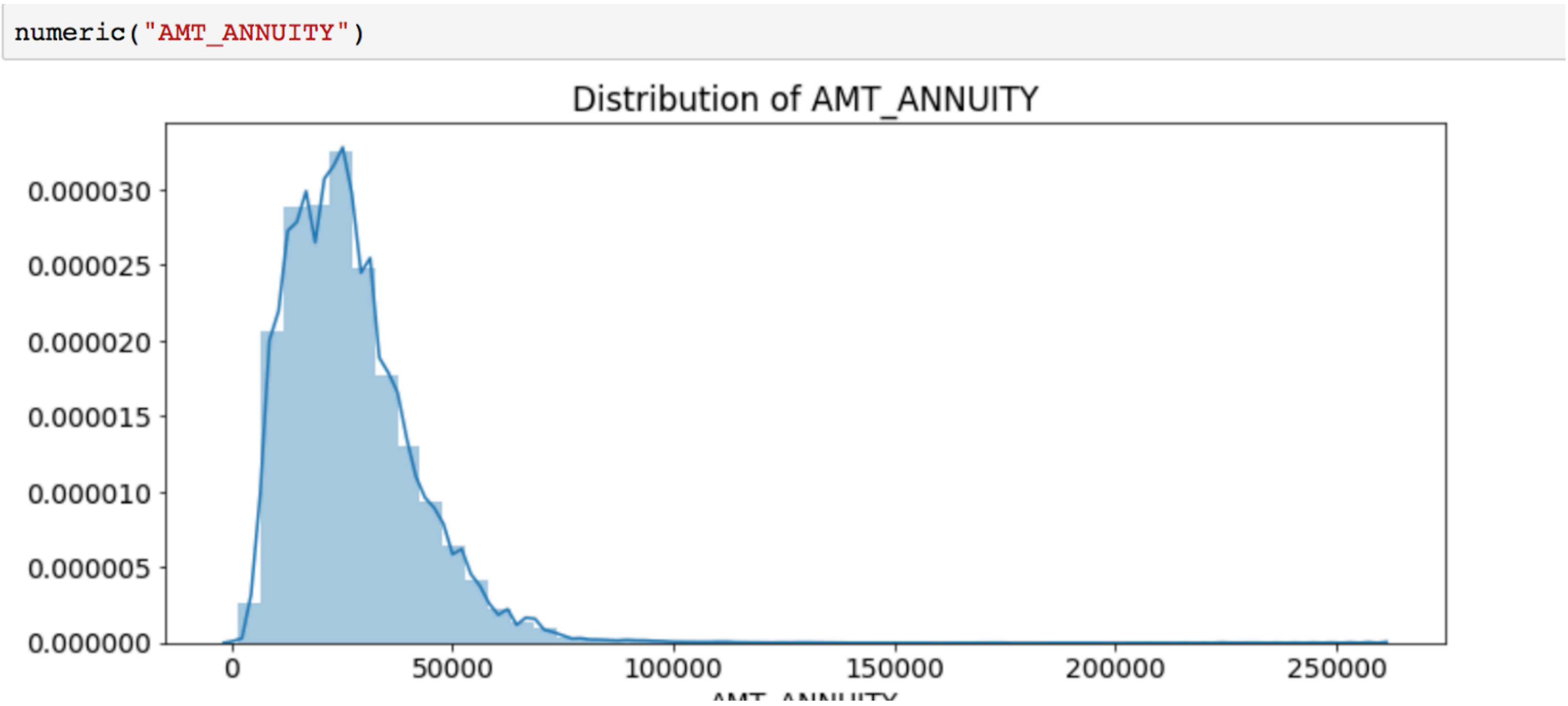
- AMT_CREDIT



```
In [56]: numeric("AMT_CREDIT")
```

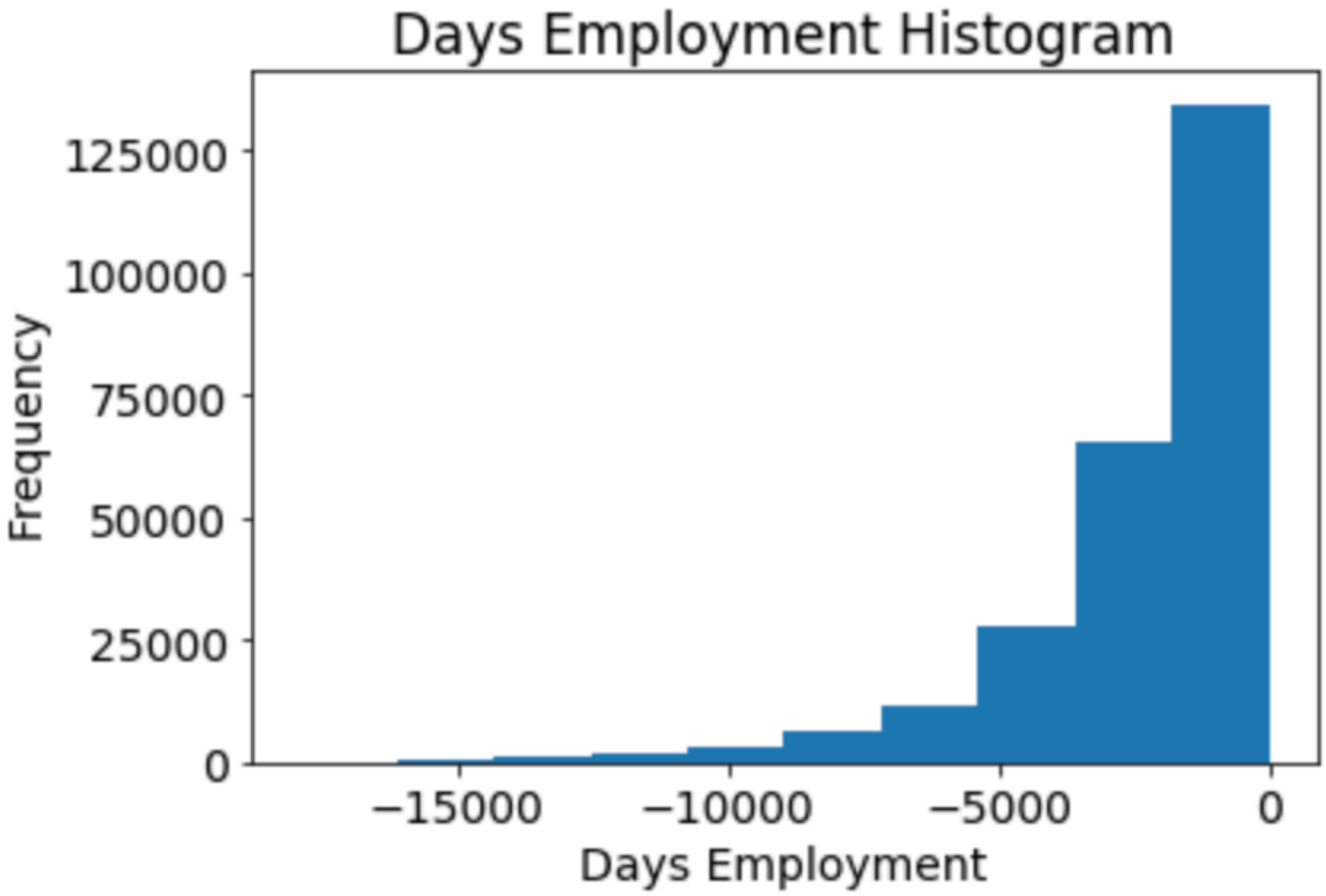
-

- AMT_ANNUITY

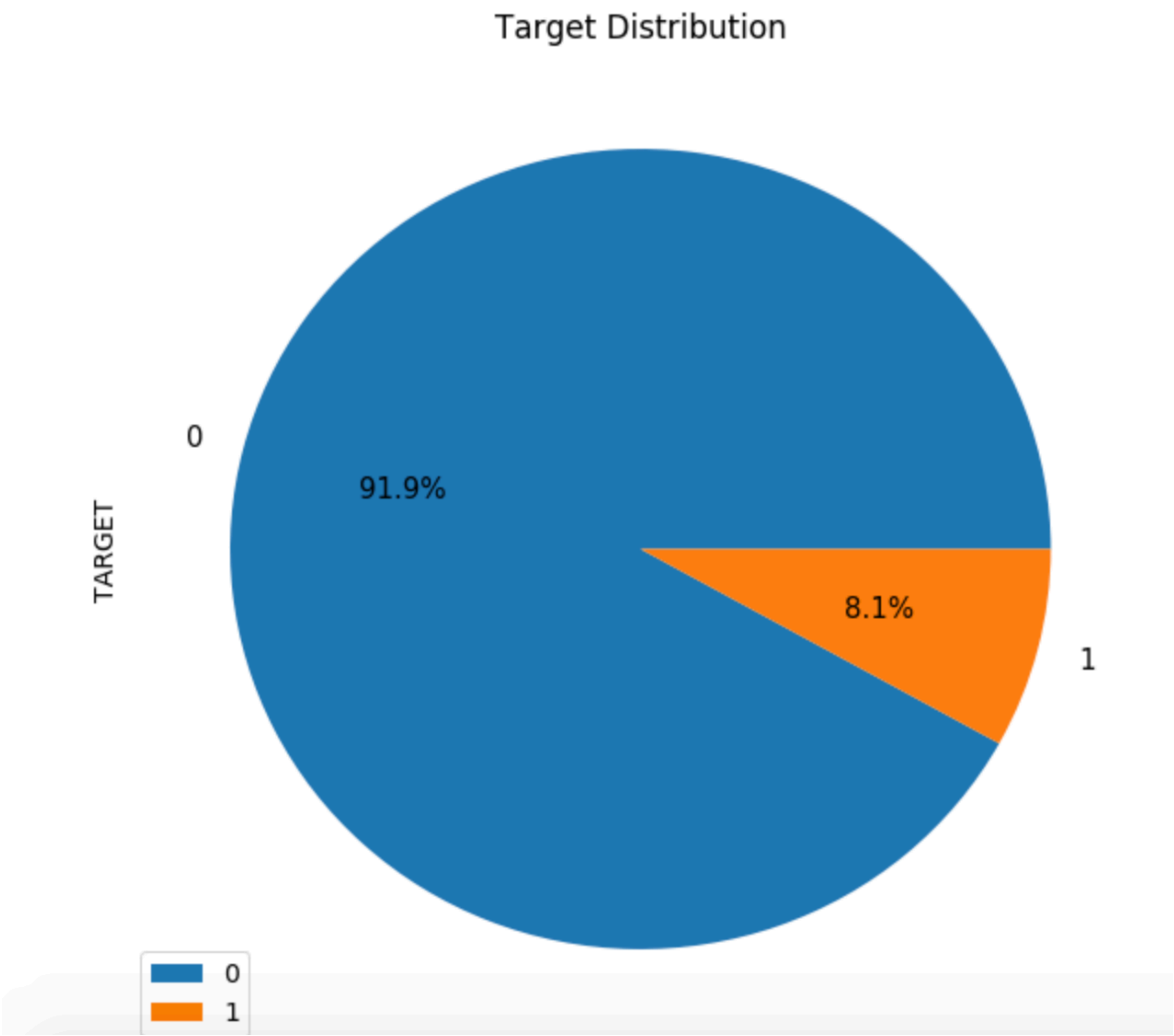


```
numeric("AMT_ANNUITY")
```

2. The ‘Days Employment’ distribution looks to be much more in line with what we would expect, and we also have created a new column to tell the model that these values were originally anomalous (because we will have to fill in the nans with some value, probably the median of the column). The other columns with DAYS in the dataframe look to be about what we expect with no obvious outliers. As an extremely important note, anything we do to the training data we also have to do to the testing data. Let's make sure to create the new column and fill in the existing column with np.nan in the testing data. The distribution looks to be much more in line with what we would expect, and we also have created a new column to tell the model that these values were originally anomalous (because we will have to fill in the nans with some value, probably the median of the column). The other columns with DAYS in the dataframe look to be about what we expect with no obvious outliers. As an extremely important note, anything we do to the training data we also have to do to the testing data. Let's make sure to create the new column and fill in the existing column with np.nan in the testing data.



3. Imbalance of Data

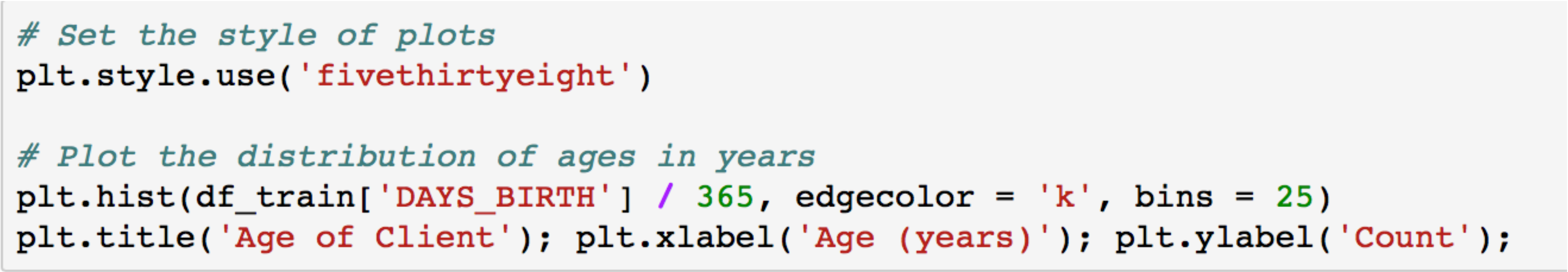


From this pie chart, we see this is an imbalanced class problem(http://www.chioka.in/class-imbalance-problem/). There are far more loans that were repaid on time than loans that were not repaid. Once we get into more sophisticated machine learning models, we can weight the classes by their representation in the data to reflect this imbalance.

4. **Effect of Age on Repayment**

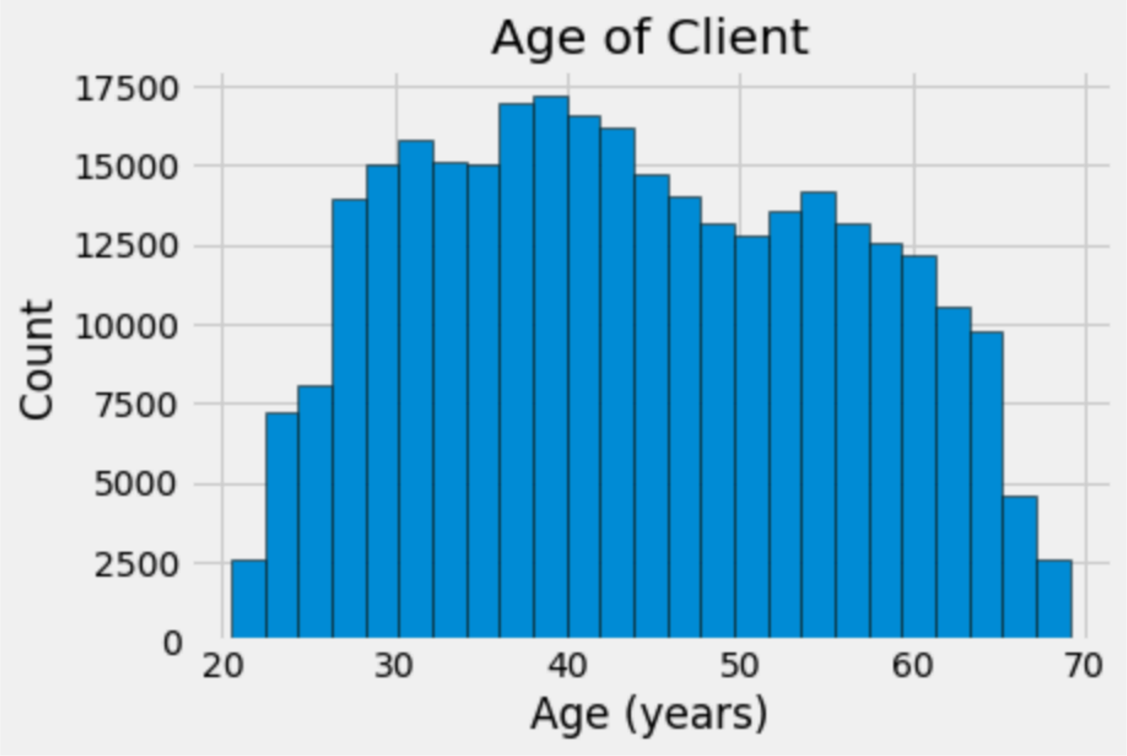
By itself, the distribution of age does not tell us much other than that there are no outliers as all the ages are reasonable. To visualize the effect of the age on the target, we will next make a kernel density estimation plot (KDE) colored by the value of the target. A kernel density estimate plot shows the distribution of a

single variable and can be thought of as a smoothed histogram (it is created by computing a kernel, usually a Gaussian, at each data point and then averaging all the individual kernels to develop a single smooth curve). We will use the seaborn kdeplot for this graph.



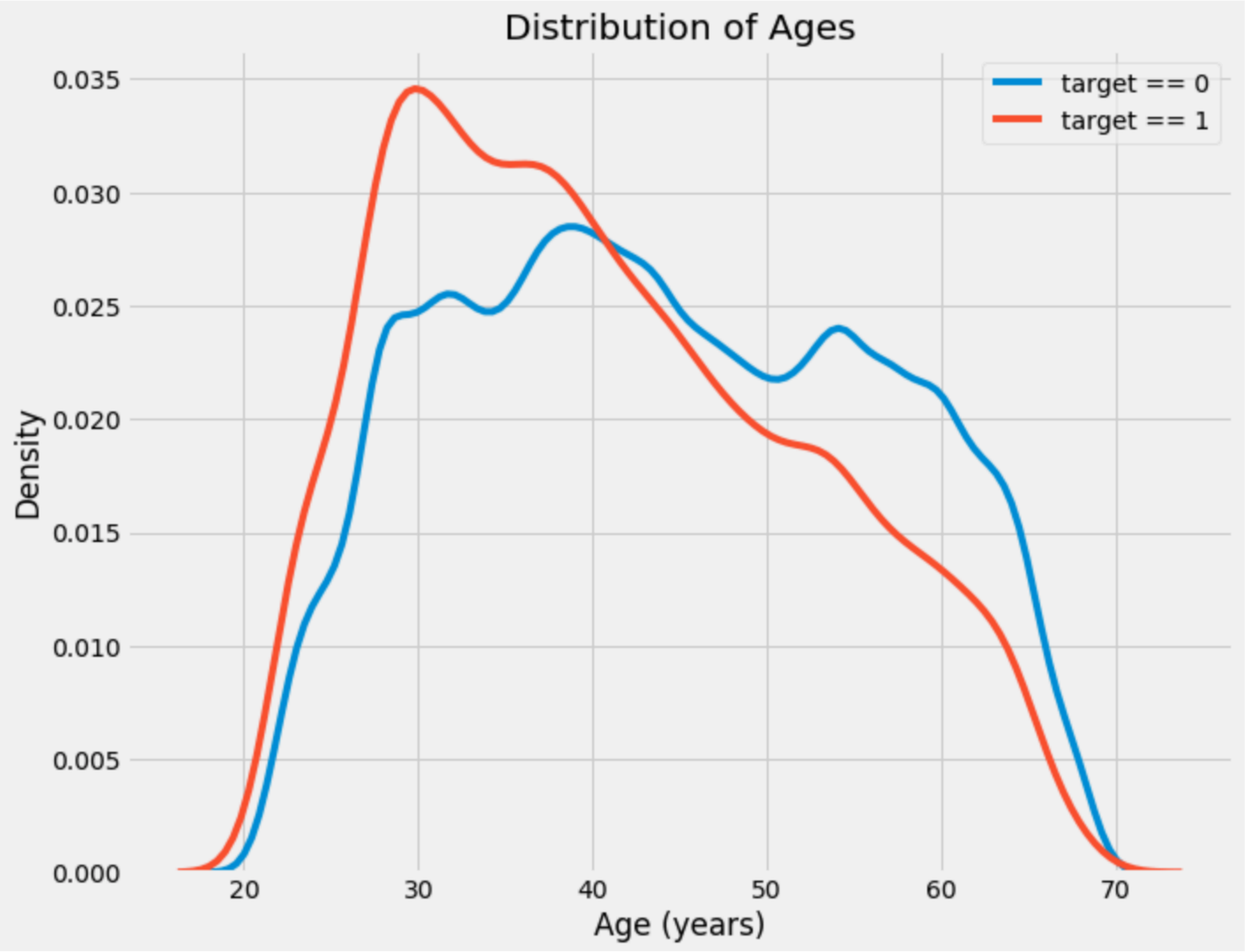
```
# Set the style of plots
plt.style.use('fivethirtyeight')

# Plot the distribution of ages in years
plt.hist(df_train['DAYS_BIRTH'] / 365, edgecolor = 'k', bins = 25)
plt.title('Age of Client'); plt.xlabel('Age (years)'); plt.ylabel('Count');
```

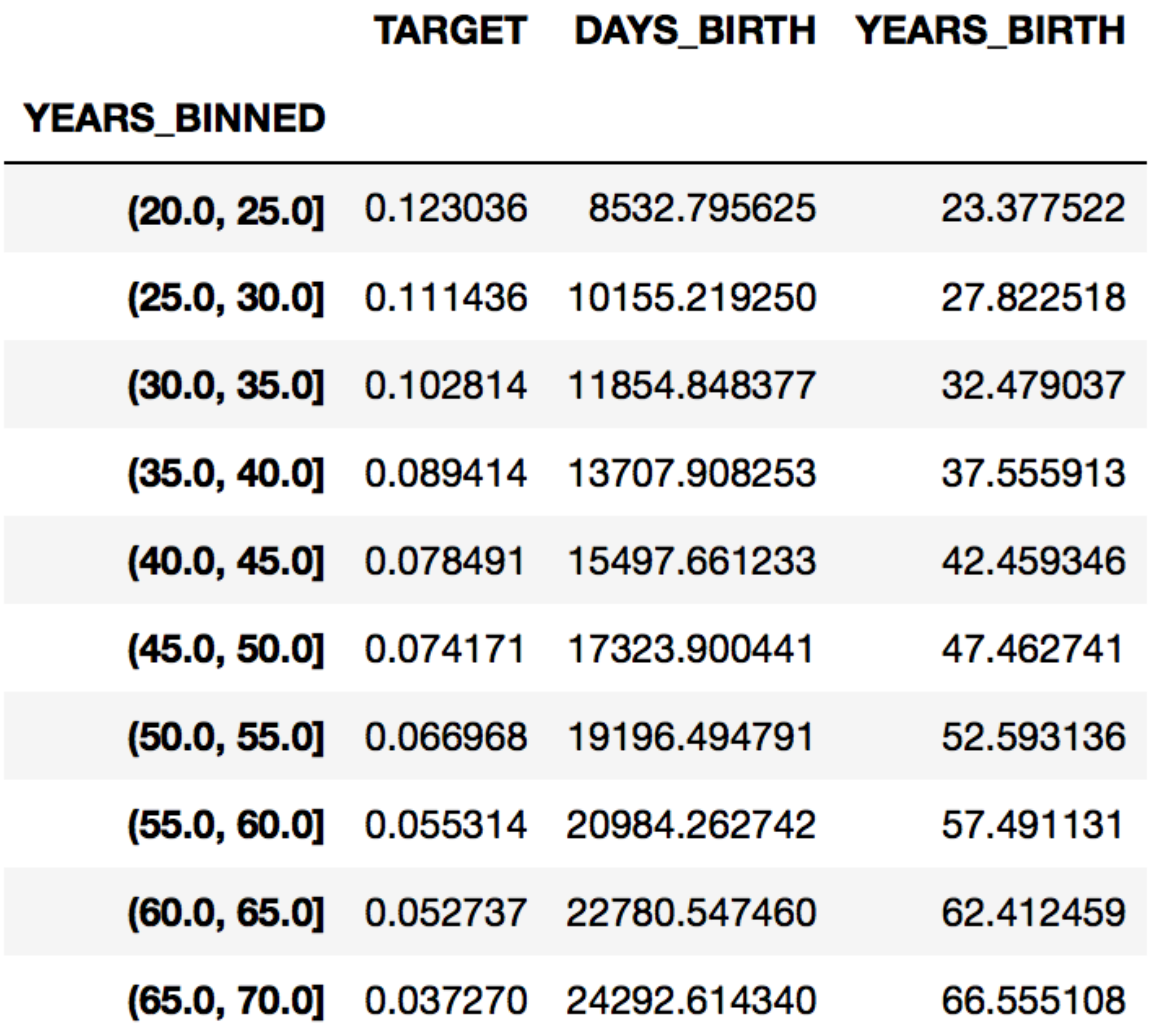


5. The target == 1 curve skews towards the younger end of the range. Although this is not a significant correlation (-0.07 correlation coefficient), this variable is likely going to be useful in a machine learning model because it does affect the target. Let's look at this relationship in another way: average failure to repay loans by age bracket.

To make this graph, first we cut the age category into bins of 5 years each. Then, for each bin, we calculate the average value of the target, which tells us the ratio of loans that were not repaid in each age category.



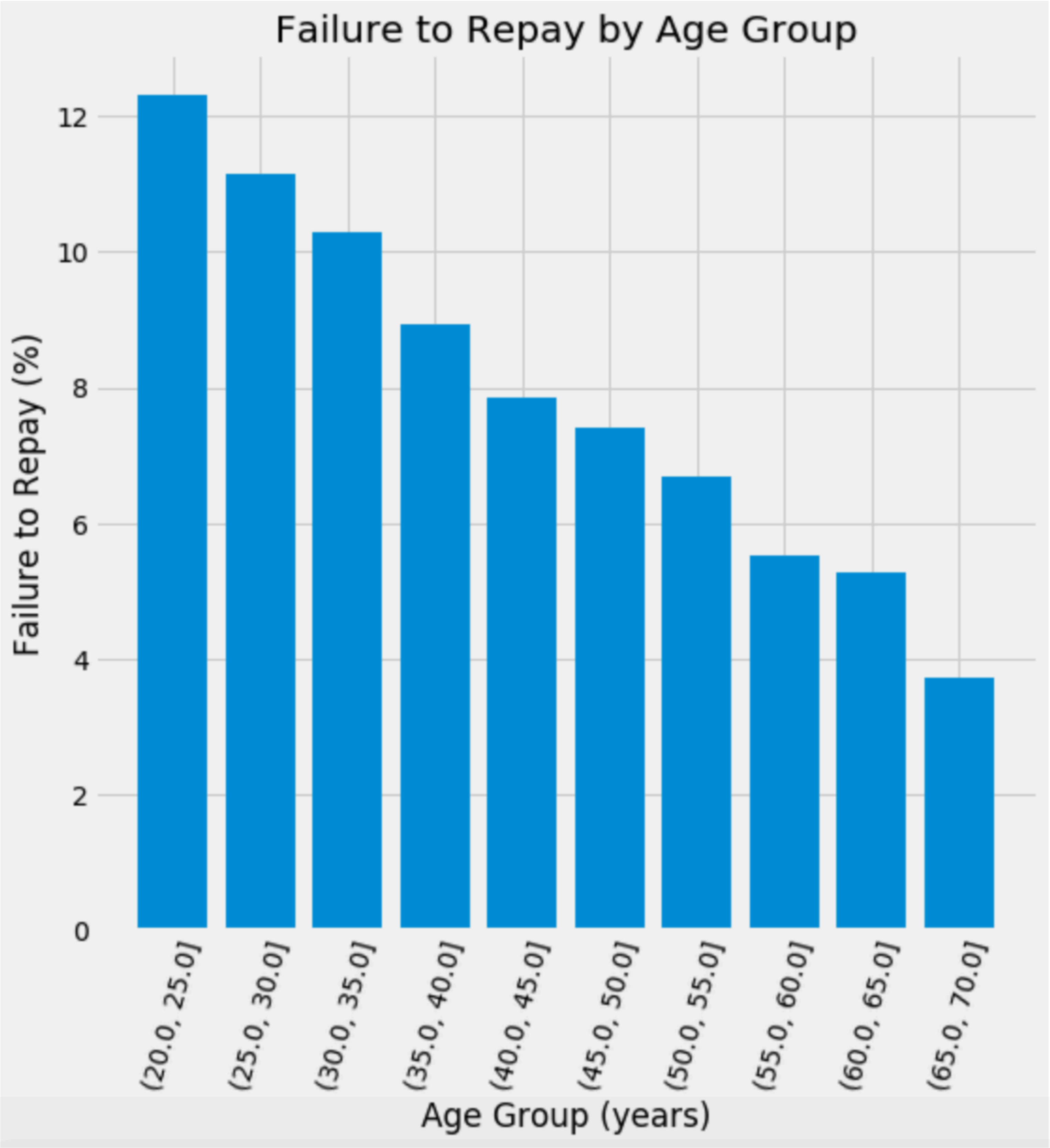
6. Age Group by the Bin & Mean Probability of the TARGET



	TARGET	DAYS_BIRTH	YEARS_BIRTH
YEARS_BINNED			
(20.0, 25.0]	0.123036	8532.795625	23.377522
(25.0, 30.0]	0.111436	10155.219250	27.822518
(30.0, 35.0]	0.102814	11854.848377	32.479037
(35.0, 40.0]	0.089414	13707.908253	37.555913
(40.0, 45.0]	0.078491	15497.661233	42.459346
(45.0, 50.0]	0.074171	17323.900441	47.462741
(50.0, 55.0]	0.066968	19196.494791	52.593136
(55.0, 60.0]	0.055314	20984.262742	57.491131
(60.0, 65.0]	0.052737	22780.547460	62.412459
(65.0, 70.0]	0.037270	24292.614340	66.555108

7. Failure to Repay by Age Group

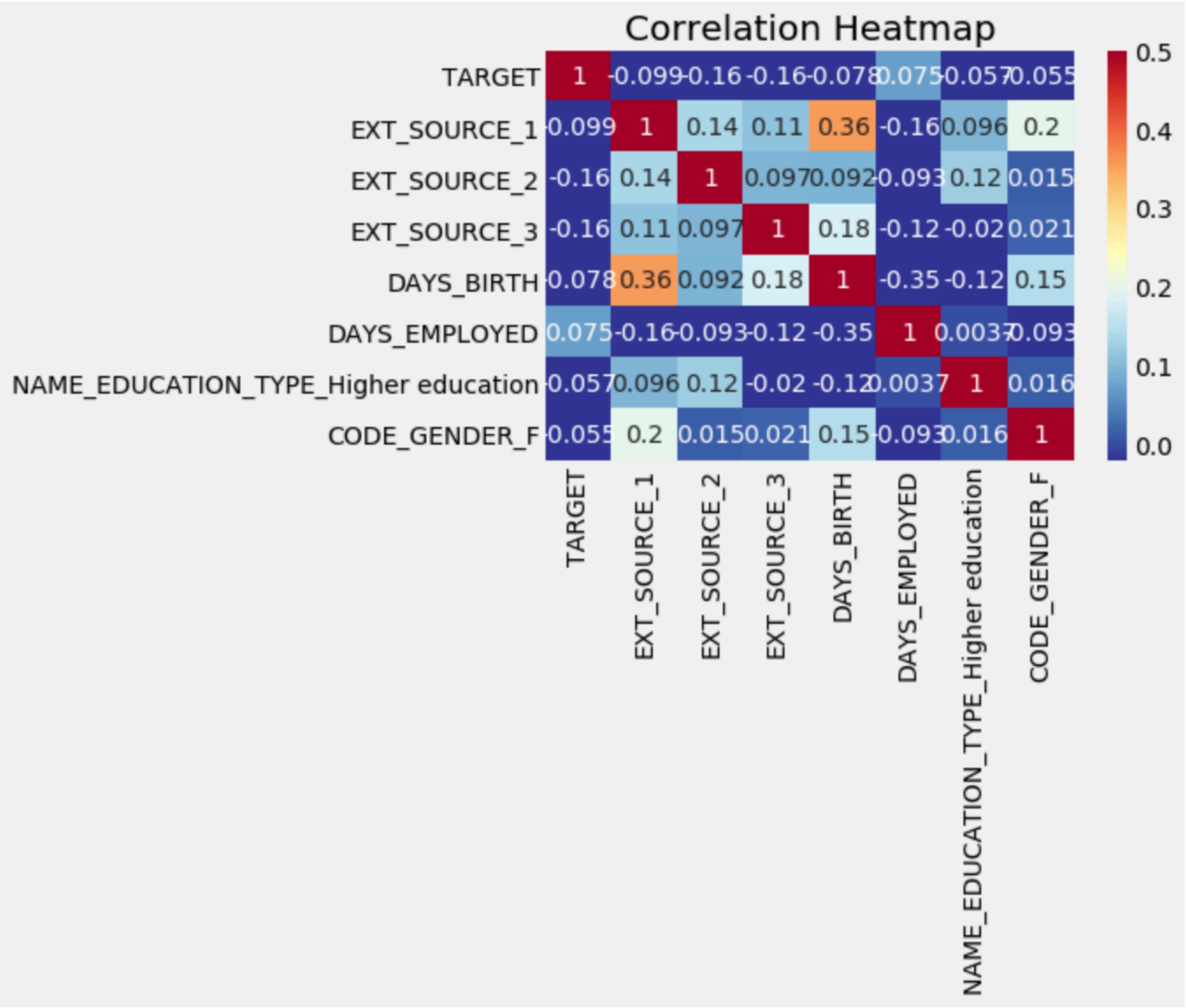
There is a clear trend: younger applicants are more likely to not repay the loan! The rate of failure to repay is above 10% for the youngest three age groups and beolow 5% for the oldest age group.
This is information that could be directly used by the bank: because younger clients are less likely to repay the loan, maybe they should be provided with more guidance or financial planning tips. This does not mean the bank should discriminate against younger clients, but it would be smart to take precautionary measures to help younger clients pay on time.



8. **Exterior Sources**

Exterior Sources The 3 variables with the strongest negative correlations with the target are EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3. According to the documentation, these features represent a "normalized score from external data source". I'm not sure what this exactly means, but it may be a cumulative sort of credit rating made using numerous sources of data.

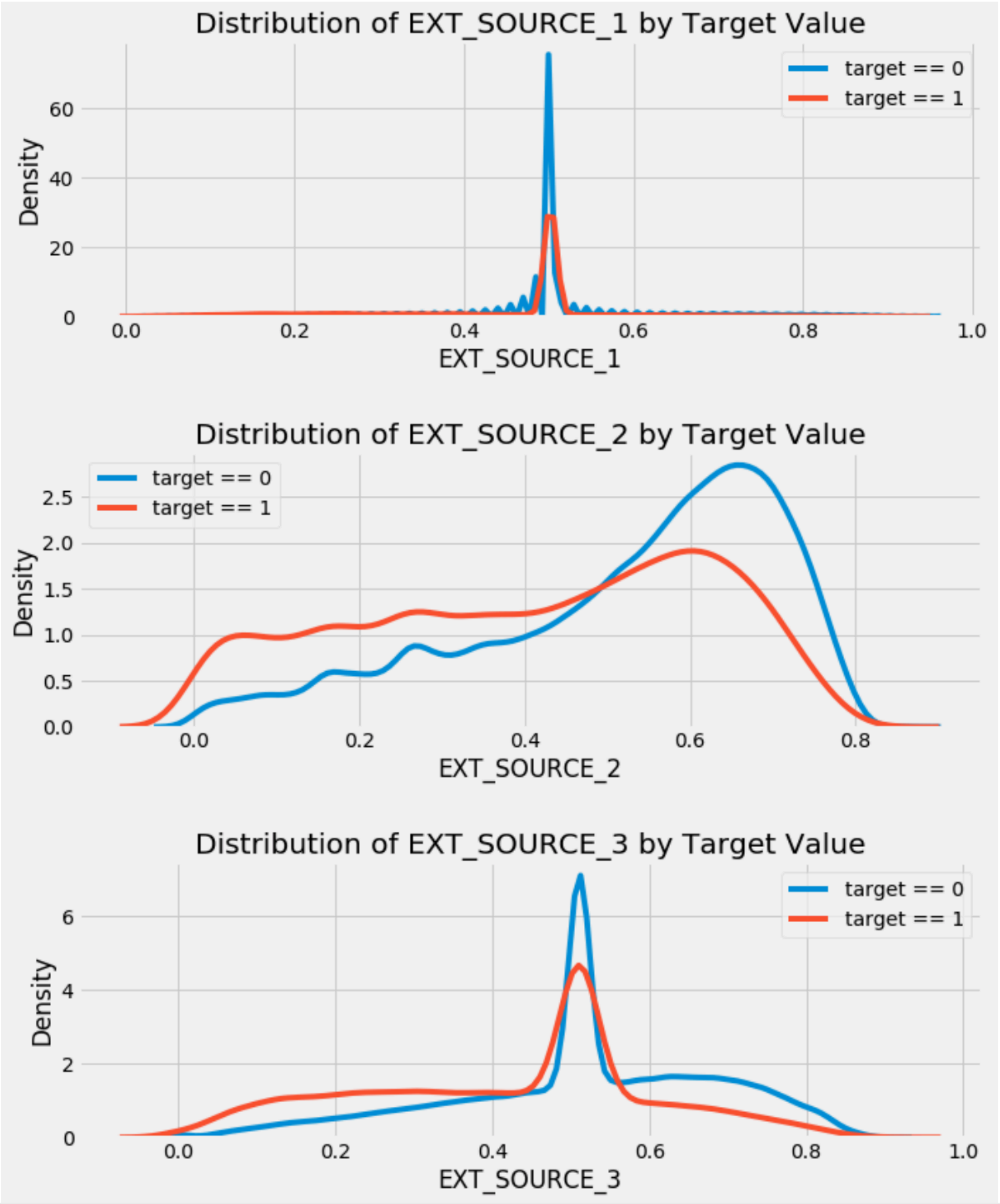
Let's take a look at these variables.
First, we can show the correlations of the EXT_SOURCE features with the target and with each other.



Based on statistical analysis exploring the strength of relationships between TARGET and independent variables such as age, gender, employment duration, external data source, we uncovered the following insights:

- Age distribution indicates by increasing age TARGET prediction to repay loan increases
- External Data Sources influences TARGET inversely
- Gender correlates with Target prediction

- Employment Duration



Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

The next step in our analysis will be to apply inferential statistics: Continuous Distribution, Normal distribution test, Bootstrap sample mean, Confidence Interval to assess Bootstrap replicates, Paired Bootstrap test, Null Hypothesis, Alternate Hypothesis, regression modeling to a majority proportion of the historical data, T-Test, p-value. We will be seeking to determine a best line of fit over the data based on selective application of the variables described earlier. Having chosen the optimal arrangement of our variables, we will test the predictive strength of this model on the remaining portion of our data. This will serve as a secondary check and ensure a minimal amount of model predictions are false positives or negatives. Once this testing phase has validated our model, we can confidently plan to apply the model to future TARGET predictions for the next LOAN transactions in application_train Kaggle datasets.

References:

read Chloe's article.