Home Credit Default Risk Prediction Description & EDA

To ensure that underserved people have a good loan experience, Home Credit uses various data, including telecommunications and transaction information, to predict the repayment ability of its customers.

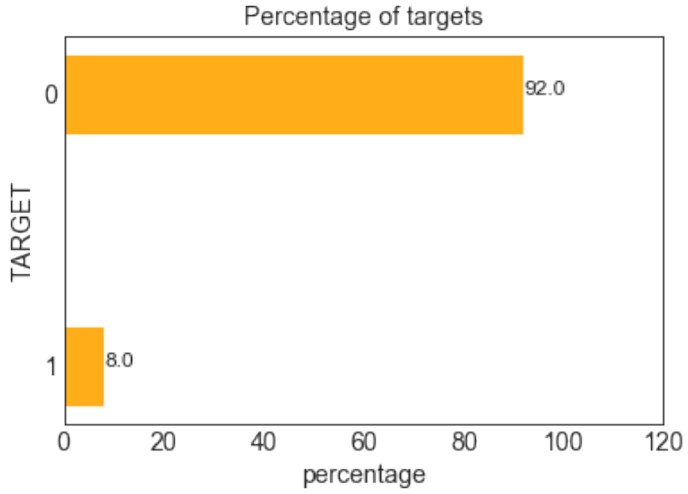
We have 218 variables including bureau balance data, installments payment data, credit card balance data, POS cash balance data and personal information as train dataset. Most of the variables are numeric and some of them, such as gender, whether their own realty and income type are strings and need to be coded.

To see the size of the data, we print their shape out and get:
Size of application_train data (307511, 122)
Size of POS_CASH_balance data (10001358, 8)
Size of bureau_balance data (27299925, 3)
Size of previous_application data (1670214, 37)
Size of installments_payments data (13605401, 8)
Size of credit_card_balance data (3840312, 23)
Size of bureau data (1716428, 17)

To make sure how many features we can put into use in our final dataset, we checked for the missing values. For the application train dataset, 68 variables have missing values. We filled those missing values with either the mean or directly deleted, which will be explicit in the concrete analysis below. We begin the EDA with the application dataset. There are 122 variables and 307511 observations in this dataset. As for the type of the122 variables, 67 of which are float, 43 are integer, 13 are object.

Target:

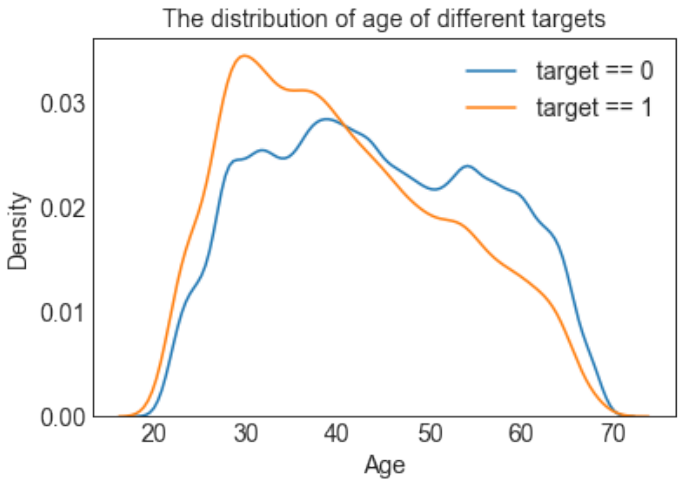
The target in our analysis is the "TARGET" column in the train dataset, which represents the result that the loan will be repaid (0) or not (1). 92% of people Target=0 repaied for their loan, the rest 8% are those who are not able to pay for their loan.



Age:

We wanted to see the distribution of the lender's age. Because in the dataset, it gives the days from birth rather than age, so that I divided the days by 365 to estimate their age and plot the figure below. The average age for our observations is 43.9. The minimum age is 20, and the maximum is 69.

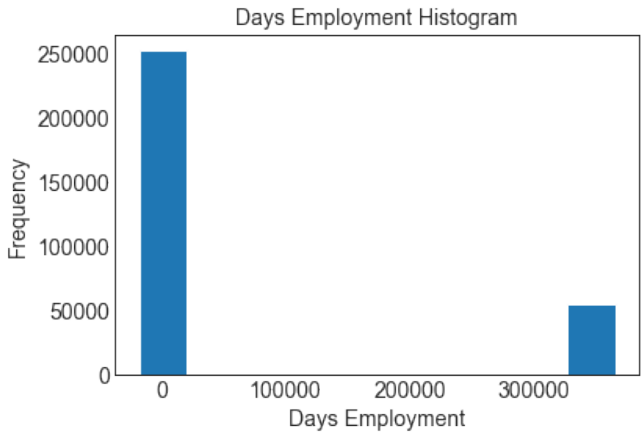
We wanted to know whether the age of lenders' contribute to their final result. From the plot below, we found that young people more tend not to pay back than older people, which makes sense. Thus, age is an important factor that we should included in our model.



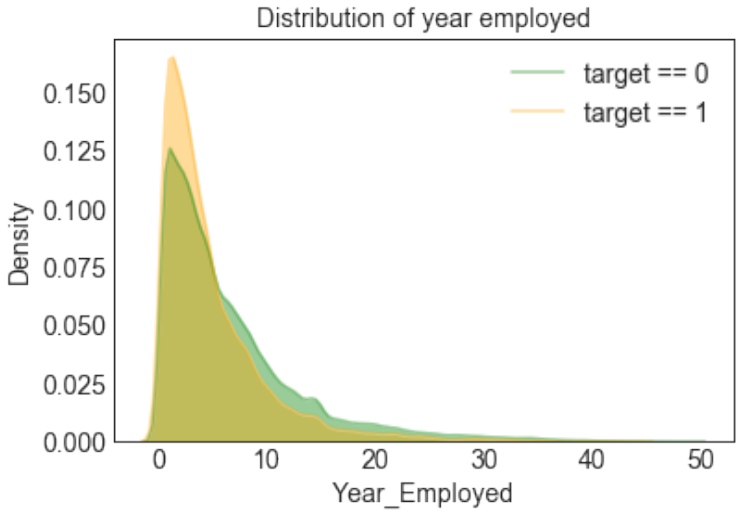
Employment:

We then went deep into the days employed of our observations, and we found that

there are many outliners who have worked 365243 days are like errors because that about 1000 years.



So I plot only those with employment days less than 365243 for the two target classes and divided the days by 365. Some difference between class 0 and class 1 could be found in this figure, that the distribution for people who tend to default is more skewed than those who are employed for a longer period of time.



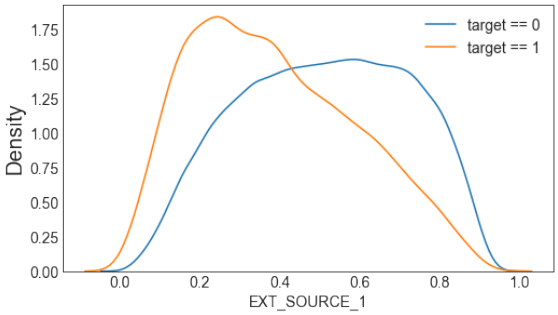
Exit Sources:

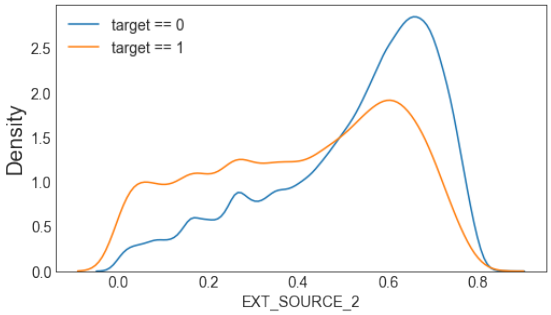
The correlation map shows that the three exit sources have the highest relation with our target.

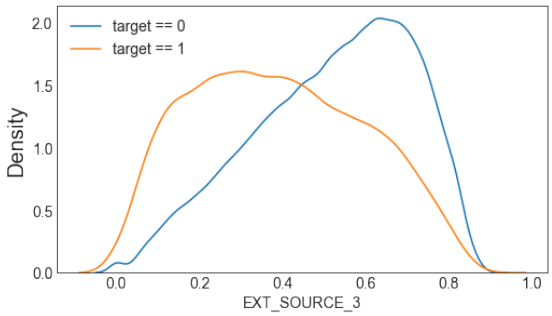
```
correlations = train_data.corr()['TARGET'].sort_values()
correlations.head(5)
```

```
EXT_SOURCE_3     -0.178919
EXT_SOURCE_2     -0.160472
EXT_SOURCE_1     -0.155317
FLOORSMAX_AVG    -0.044003
FLOORSMAX_MEDI   -0.043768
Name: TARGET, dtype: float64
```

So then, we looked into the three sources by plotting figures, which show significant patterns between different exit sources and the results.

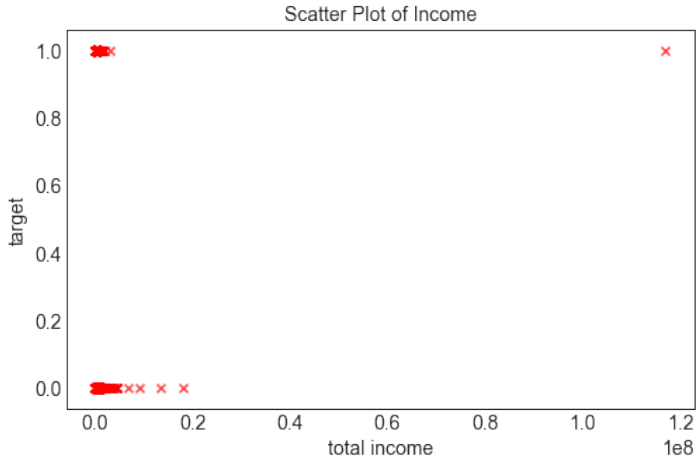






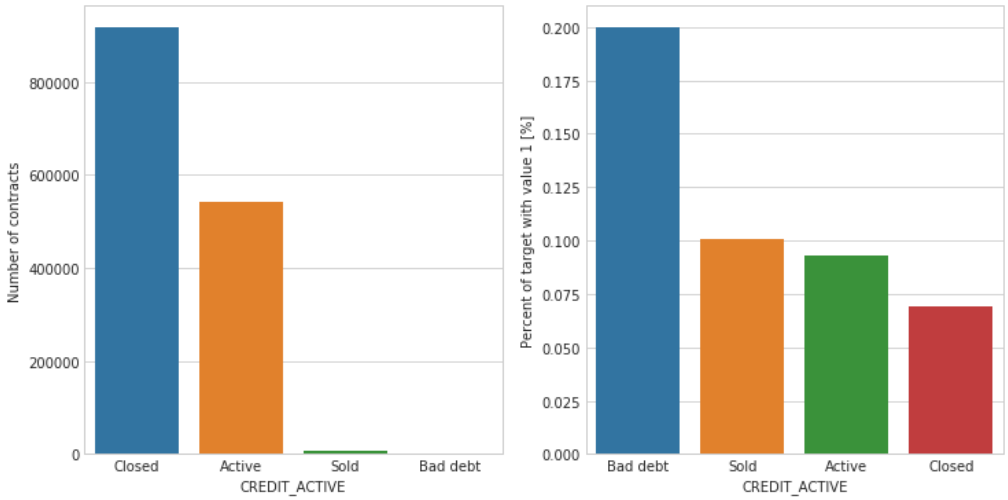
Annual Income:

The mean annual income for the population is 1.687979e+05, but there is only one record that has over 1.0e+08 annual income and shows the default result. Also, we can know from the scatter plot that people who earned more money annually tend to not default in the end.



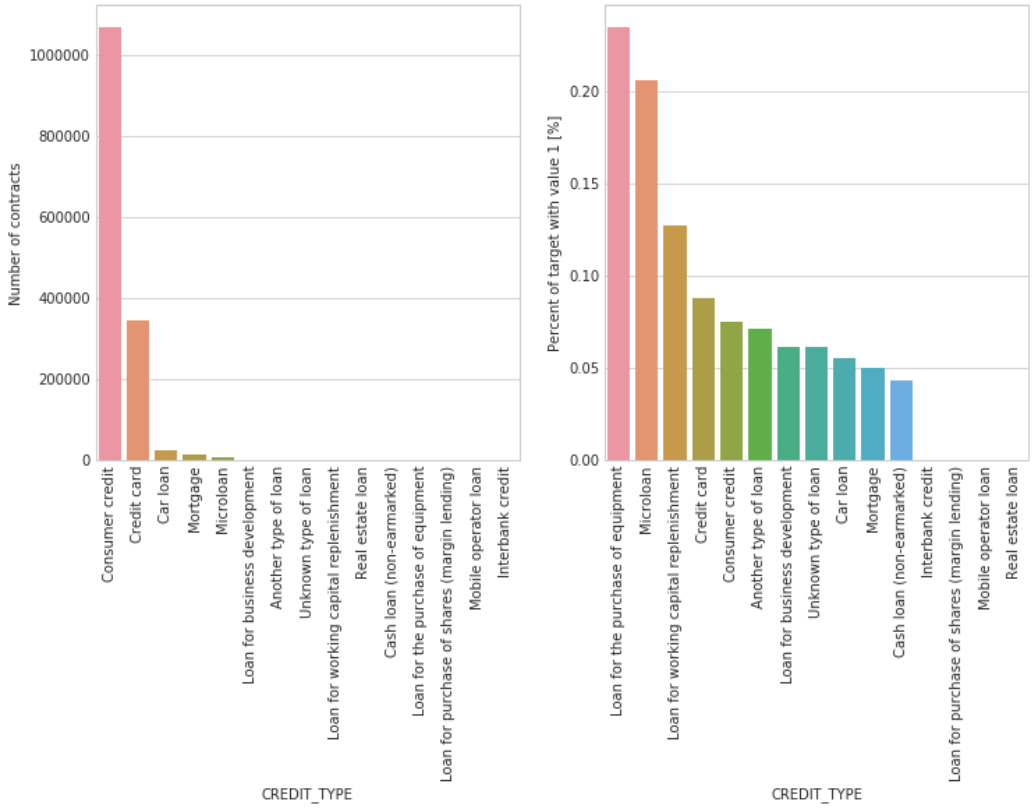
Then we looked at the Bureau dataset, and we found that this dataset contains some variables that seem valuable. So we merged the Bureau dataset with train dataset and did the visualizations.

Credit_Status:



We can see that most of the credit status is closed. Sold and bad debt only take up a little portion. Then let us see the percentage of those people who do not repay. It makes sense that bad debt status has the most percentage of people who do not repay and closed, with the most number of contracts, has the lowest rate. We concluded that the credit active status is an important contributor to the final result.

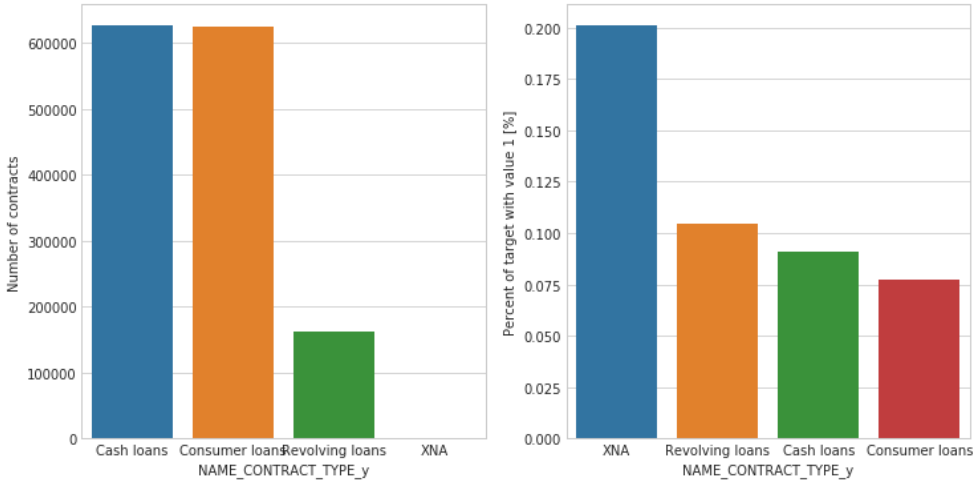
Credit_Type:



We can also check the credit type registered in the Bureau. It is obvious that loans for the purchase of equipment and microloan have relatively high rates of those not repaid. For consumer credit, which takes up the most part of the contracts, has about only 8% default rate.

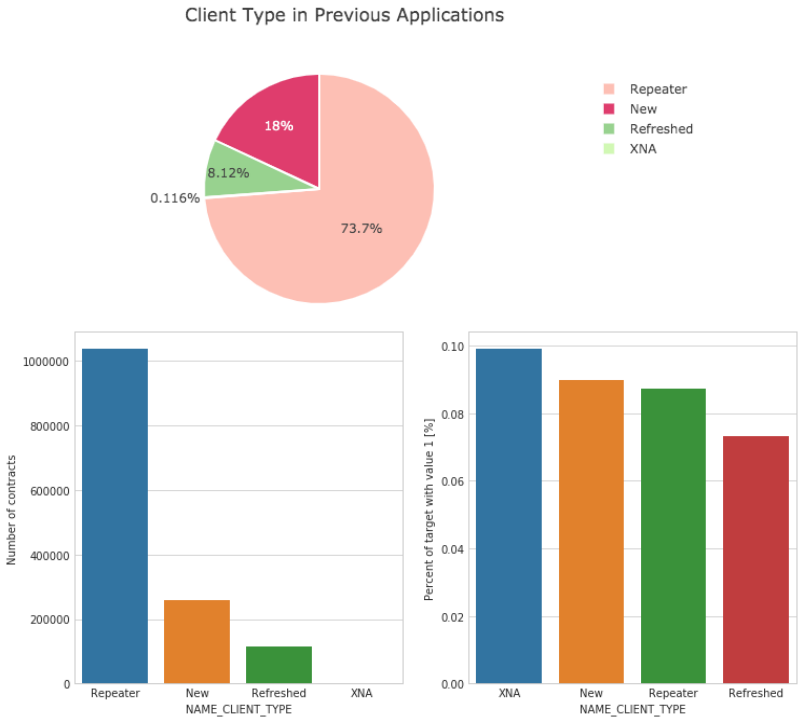
Also, we looked into the application dataset, and analyzed some variables that might be important, such as Contract_Type and Client_Type.

Contract_Type:



When comparing the contract type in previous application dataset, we can see that cash loans and consumer loans are pretty high in total numbers, and have the lowest default rate. XNA, which has the highest rate, are very close to 0 in total number so that it will not affect the overall outcome largely.

Client_Type:



By analyzing the client type in the previous application, we can see that repeaters are more than 70%. The percentage of New clients and refreshed clients are 18% and 8% separately. The percentage of different types of clients in default population (target=1) shows very different in the left bar graph. So we believe that the client type is also an important factor to our model.