HOME CREDIT SCORECARD MODEL

Using Logistic Regression & Decision Tree

CASE UNDERSTANDING

Home Credit is currently using various statistical methods and Machine Learning to make credit score predictions. Now, we ask you to unlock the full potential of our data. By doing so, we can ensure that customers who are able to repay are not turned down when applying for a loan, and that loans can be provided with a principal, maturity, and repayment calendar that will motivate customers to succeed. Evaluation will be carried out by checking how deep your understanding of the analysis you are working on is.

GOALS

Predict the credit scores of customers who make loans using machine learning models and provide insights and recommendations so that customers who are able to pay off loans are not rejected when applying for loans.

OBJECTIVE

Predict customer credit scores and minimize the potential for default on loans submitted by customers.

METRICS

Use Recall Metrics and F1-Score to capture more customers with potential credit problems.

FLOW ANALYSIS

EDA

data visualization and discovering insights about the business.

Modelling

Using Logistic Regression & Decision Tree for train and test data.

Data Cleaning

Drop unnecessary features, Handling missing value, Drop duplication & outliers

Data Preparation

Normalize data, Encoding data, Future selection, Solve imbalance data using SMOTE.

DATA VIEW AFTER CLEANING

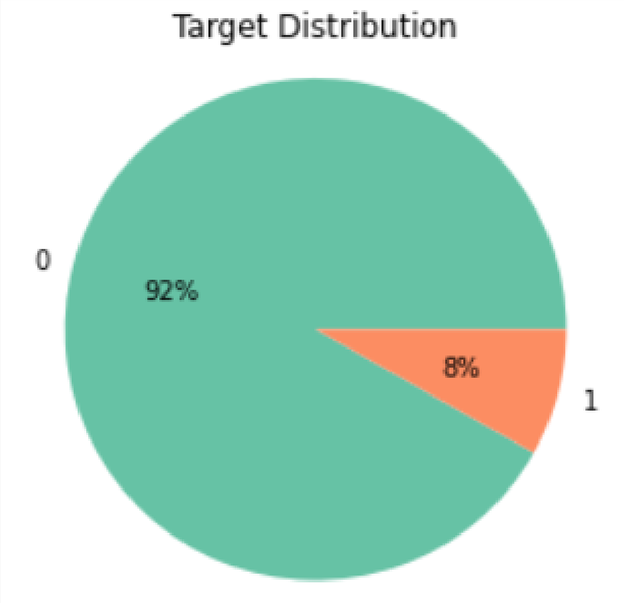
Removing some features that are not used or have no effect on the model in general. That leaves 46 columns and 293462 rows.

	TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	...	OCCUPATION_TYPE_Low skill Laborer
0	1	0.567100	0.094783	0.103342	0.088348	0.888663	0.045086	0.852140	0.705433	2	...	
1	0	0.783550	0.327260	0.152575	0.309859	0.476287	0.043648	0.951929	0.959566	1	...	
2	0	0.134199	0.023591	0.022985	0.026889	0.347505	0.046161	0.827335	0.648326	2	...	
3	0	0.350649	0.070165	0.125662	0.072983	0.349819	0.038817	0.601451	0.661387	2	...	
4	0	0.307359	0.122673	0.090651	0.134443	0.297482	0.038820	0.825268	0.519522	2	...	
...	...	...	...	...	...	...	...	...	...	...	...	
307506	0	0.422799	0.054967	0.116134	0.052497	0.896229	0.046133	0.657263	0.724607	1	...	
307507	0	0.148629	0.058859	0.046494	0.052497	0.249887	1.000000	0.822147	0.431708	2	...	
307508	0	0.408369	0.165835	0.126972	0.154930	0.577857	0.026076	0.726937	0.284424	3	...	
307509	1	0.466089	0.085218	0.083218	0.079385	0.747516	0.034258	0.896158	0.870641	2	...	
307510	0	0.422799	0.165137	0.212647	0.180538	0.471150	0.043455	0.792153	0.943032	1	...	

293462 rows × 46 columns

*Previously there were 122 columns and 307511 rows.

EXPLORATORY DATA ANALYSIS



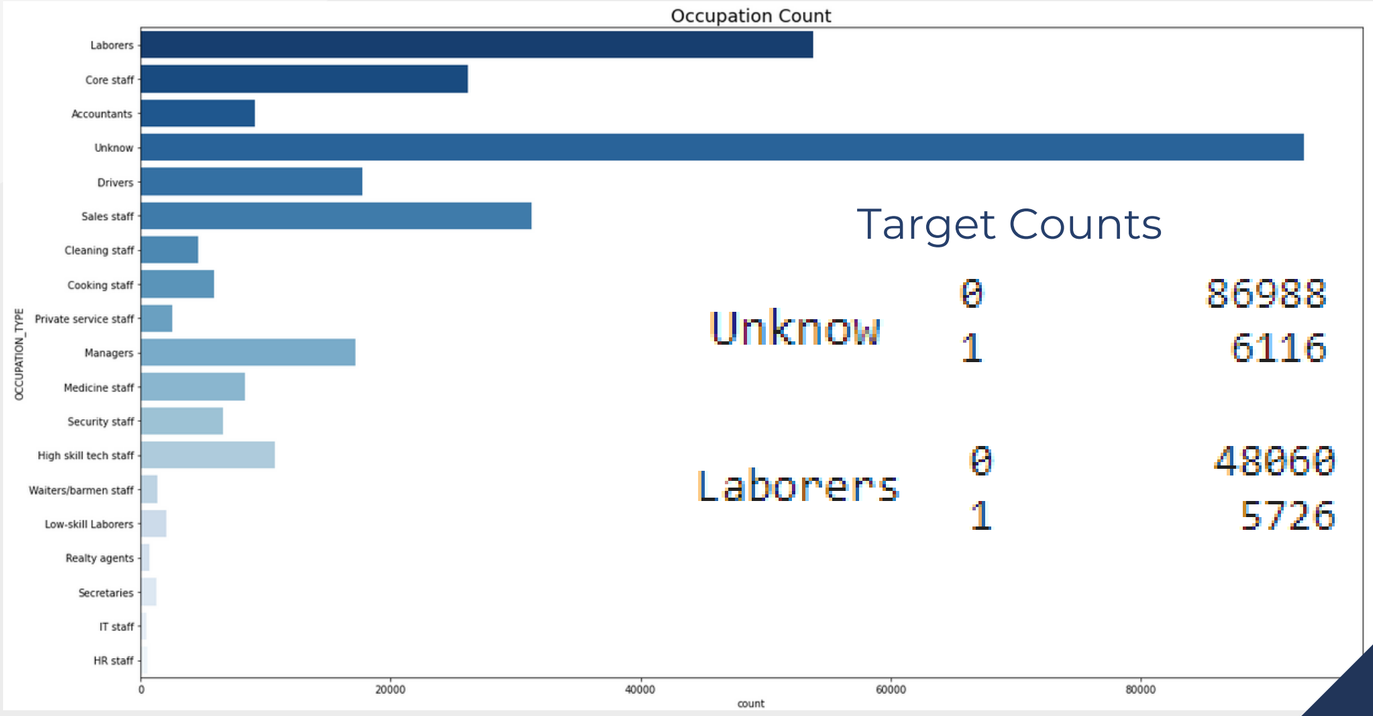
0	269454
1	24008

The target distribution based on the data provided looks unbalanced between the values 1 and 0 so it needs to be more careful when doing modeling.

Income Average

Unknow	0	142753.46
	1	142150.27
Laborers	0	157577.80
	1	154787.74

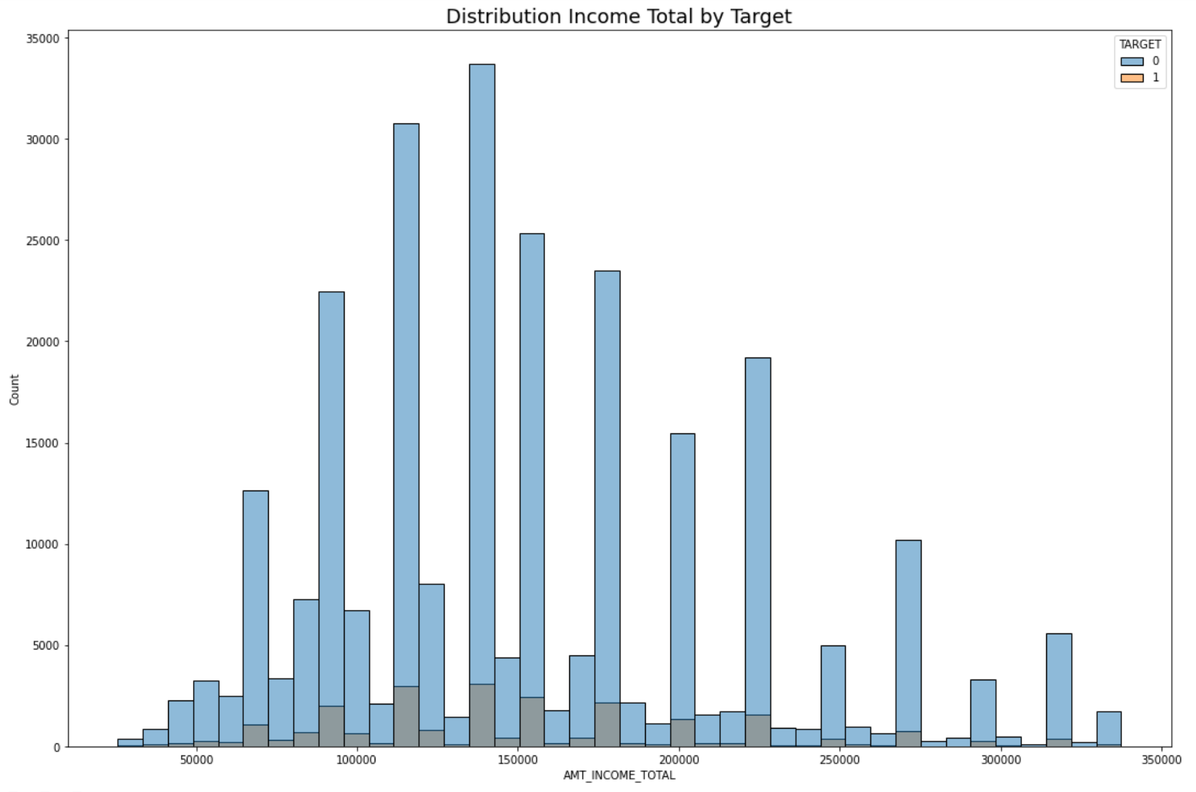
The type of work is unknown and Laborers have a high number of credit applications but the possibility of bad credit is also high, so it is necessary to be careful in handling credit requests from these 2 categories.

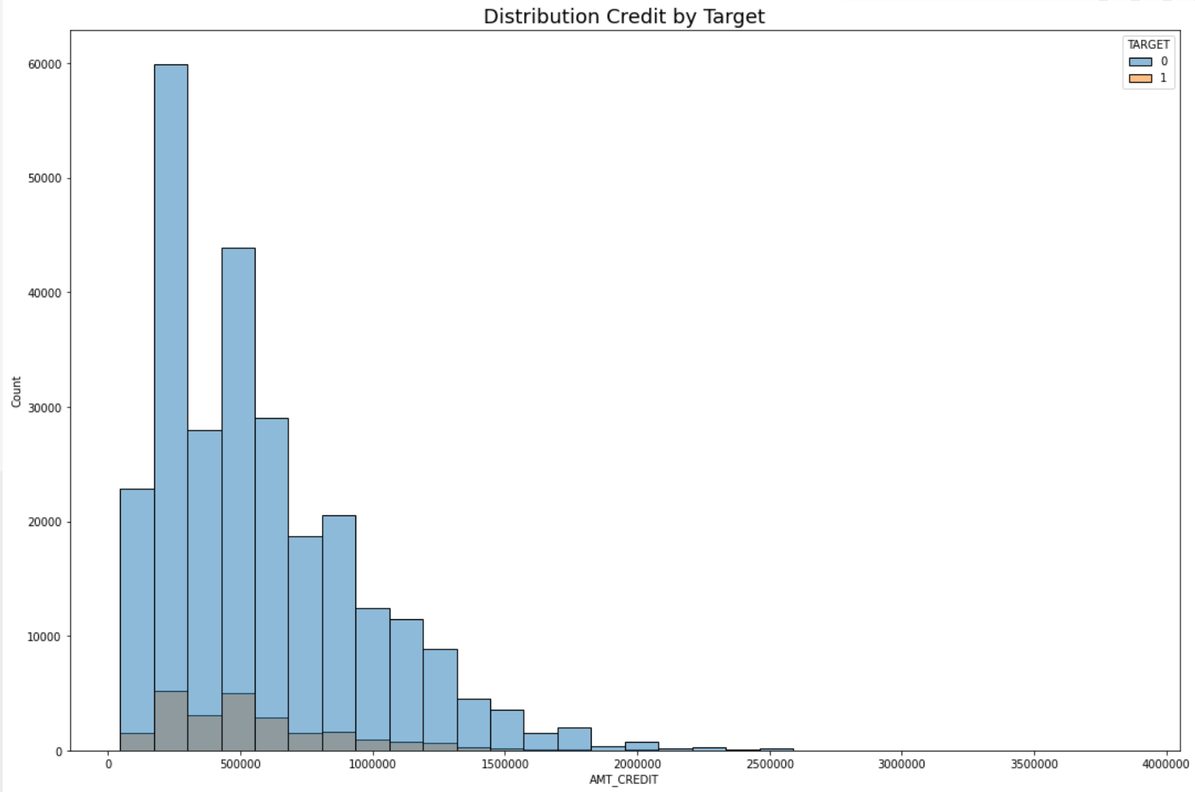


Target Counts

Unknow	0	86988
	1	6116
Laborers	0	48060
	1	5726

EXPLORATORY DATA ANALYSIS





The higher the customer's total income, the smaller the percentage of loans and the higher the number of loans submitted, the smaller the percentage of bad loans, so a campaign is needed to attract customers with high income.

MODELLING

Preprocessing

- Normalize data using MinMaxScaller()
- Encoding feature ['NAME_CONTRACT_TYPE', 'NAME_INCOME_TYPE','OCCUPATION_TYPE']
- Feature selection using PPScore & Feature Importance Random Forest

selected features

'DAYS_EMPLOYED', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'AMT_ANNUITY', 'DAYS_BIRTH', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE', 'AMT_INCOME_TOTAL'

Model	Train	
	Recall	F1-Score
Logistic Regression	0.55	0.54
Decision Tree	0.90	0.90

RUN MODEL IN TEST DATA

Logistic Regression

AMT_CREDIT	AMT_GOODS_PRICE	AMT_ANNUITY	DAYS_BIRTH	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE	AMT_INCOME_TOTAL	Credit_Score
568800.0	450000.0	20560.5	-19241	-5170.0	-812	-1740.0	135000.0	0
222768.0	180000.0	17370.0	-18064	-9118.0	-1623	0.0	99000.0	0
663264.0	630000.0	69777.0	-20038	-2175.0	-3503	-856.0	202500.0	0
1575000.0	1575000.0	49018.5	-13976	-2000.0	-4208	-1805.0	315000.0	0
625500.0	625500.0	32067.0	-13040	-4000.0	-4262	-821.0	180000.0	0
...	...	...	...	...	...	...	...	...

0	46363
1	2381

Decision Tree

AMT_CREDIT	AMT_GOODS_PRICE	AMT_ANNUITY	DAYS_BIRTH	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE	AMT_INCOME_TOTAL	Credit_Score
568800.0	450000.0	20560.5	-19241	-5170.0	-812	-1740.0	135000.0	0
222768.0	180000.0	17370.0	-18064	-9118.0	-1623	0.0	99000.0	0
663264.0	630000.0	69777.0	-20038	-2175.0	-3503	-856.0	202500.0	0
1575000.0	1575000.0	49018.5	-13976	-2000.0	-4208	-1805.0	315000.0	0
625500.0	625500.0	32067.0	-13040	-4000.0	-4262	-821.0	180000.0	0
...	...	...	...	...	...	...	...	...

0	40367
1	8377

BUSINESS RECOMENDATION

The Logistic Regression Model can help to predict potential customers with more successful paying rates and the Decision Tree Model can help to predict potential customers with high default rates. The use of the model can be adjusted to the case you want to look for or choose a model according to the results of the meeting with bos.

Furthermore, the recommendation that I can give is to conduct a campaign to attract potential customers with high salaries because according to the analysis, customers with high salaries will have a smaller percentage of defaults. But you need to be careful with some customers with certain jobs because according to the analysis they can default on credit.

THANK YOU

Linkedin
https://www.linkedin.com/in/dewa-adji/

Email
dewaadji12@gmail.com

Github Repository
https://github.com/dewaadji/HCI-Credit-Credit-ScoreCard-Model