Predicting how capable each applicant is of repaying a loan at Home Credit

Final Project for Home Credit Indonesia Data Scientist Virtual Internship Program by Rakamin Academy

Created by: Tito Dwi Syahputra

Table of Content

Introduction

Business Background

Many people struggle to get loans due to insufficient or non-existent credit histories. Unfortunately, this population is often taken advantage of by **untrustworthy lenders** who are not able to repay. These clients will give negative impact to Home Credit's business performance.

Problem Statements

- Home Credit has to struggle to **find out trustworthy clients** who be able to repay their loan.
- Home Credit needs to find **best segment of clients by age and occupation** in order to focusing their marketing strategy on.
- There are **24.825** out of **307.511** clients with late payment, that is **8,1 %** of total clients.

Objective Statements

- To use historical loan application data to predict whether or not an applicant will be able to repay a loan.
- To find out Home Credit's **best segment of clients by age and occupation**.
- To find out what type of clients are not able to repay their loan.

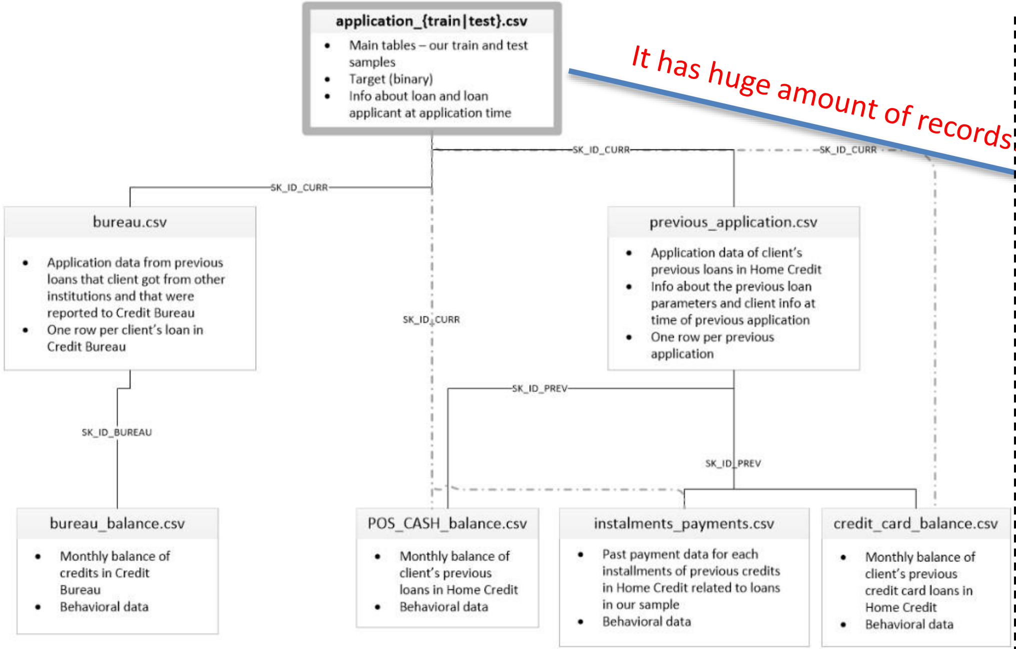
Methodology

- Machine learning algorithm using **Logistic Regression** and **LightGBM**.
- **Exploratory Data Analysis**.

Business Values

- We could help Home Credit to determine **early potential of untrustworthy client**.
- We could help Home Credit in deciding **efficient marketing strategy deployment** by age and occupation.

Dataset



It has huge amount of records!

```
training data shape: (307511, 122)
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREI
0	100002	1	Cash loans	M	N	Y	0	202500.0	40659
1	100003	0	Cash loans	F	N	N	0	270000.0	129350
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	13500
3	100006	0	Cash loans	F	N	Y	0	135000.0	31268
4	100007	0	Cash loans	M	N	Y	0	121500.0	51300

```
testing data shape: (48744, 121)
```

	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_
0	100001	Cash loans	F	N	Y	0	135000.0	568800.0	
1	100005	Cash loans	M	N	Y	0	99000.0	222768.0	
2	100013	Cash loans	M	Y	Y	0	202500.0	663264.0	
3	100028	Cash loans	F	N	Y	2	315000.0	1575000.0	
4	100038	Cash loans	M	Y	N	1	180000.0	625500.0	

```
app_train['TARGET'].value_counts()
```

```
0    282686
1     24825
Name: TARGET, dtype: int64
```

```
missing_values_table(app_train).head()
```

```
You are selecting dataframe which has "122" columns
There are "67" columns that have missing values.
```

	Missing Values	% of Total Values	Total Rows
COMMONAREA_MEDI	214865	69.872	307511
COMMONAREA_AVG	214865	69.872	307511
COMMONAREA_MODE	214865	69.872	307511
NONLIVINGAPARTMENTS_MEDI	213514	69.433	307511
NONLIVINGAPARTMENTS_MODE	213514	69.433	307511

```
missing_values_table(app_test).head()
```

```
You are selecting dataframe which has "121" columns
There are "64" columns that have missing values.
```

	Missing Values	% of Total Values	Total Rows
COMMONAREA_MODE	33495	68.716	48744
COMMONAREA_MEDI	33495	68.716	48744
COMMONAREA_AVG	33495	68.716	48744
NONLIVINGAPARTMENTS_MEDI	33347	68.413	48744
NONLIVINGAPARTMENTS_AVG	33347	68.413	48744

- There are **7 different sources** of data!
- The **main data** for training and testing are **application train and test**.
- Other than that, there are **additional data** which are obtained from previous loans in Home Credit (**previous_application**) and other Institutions (**bureau**)
- Each entity has unique ID in which define their relationships.

- The class distribution (**TARGET**) is **imbalanced**
- There are **some missing values** in the data. When doing modeling,
  - **Impute** them with median for **Logistic Regression** model.
  - **Keep** missing values for **LightGBM** model.
- No duplicate records.
- Data types seem valid at all.
- Some invalid entry values are handled.

EDA

Feature Selection with ANOVA Test and Chi-Squred Test

Check data distribution in general

Check data distribution whether clients be able to repay their loan or not

Deep dive

"ANOVA Test"	TARGET p_value	result
EXT_SOURCE_3	0.000000e+00	reject H0 (significant)
EXT_SOURCE_2	0.000000e+00	reject H0 (significant)
EXT_SOURCE_1	0.000000e+00	reject H0 (significant)
DAYS_BIRTH	0.000000e+00	reject H0 (significant)
DAYS_EMPLOYED	8.444512e-301	reject H0 (significant)
...	...	...
AMT_REQ_CREDIT_BUREAU_WEEK	6.845546e-01	fail to reject H0
FLAG_MOBIL	7.669698e-01	fail to reject H0
FLAG_CONT_MOBILE	8.373783e-01	fail to reject H0
FLAG_DOCUMENT_5	8.609936e-01	fail to reject H0
FLAG_DOCUMENT_20	9.049243e-01	fail to reject H0

105 rows × 2 columns

Sort the features by their p-values

"Chi-Squared Test"	TARGET p_value	result
TARGET	0.000000e+00	reject H0 (significant)
OCCUPATION_TYPE	3.784500e-288	reject H0 (significant)
NAME_INCOME_TYPE	1.928146e-266	reject H0 (significant)
ORGANIZATION_TYPE	6.582184e-257	reject H0 (significant)
NAME_EDUCATION_TYPE	2.447681e-219	reject H0 (significant)
CODE_GENDER	4.183493e-202	reject H0 (significant)
NAME_FAMILY_STATUS	7.744842e-107	reject H0 (significant)
NAME_HOUSING_TYPE	1.099089e-88	reject H0 (significant)
NAME_CONTRACT_TYPE	1.023515e-65	reject H0 (significant)
FLAG_OWN_CAR	9.330994e-34	reject H0 (significant)
WALLSMATERIAL_MODE	1.453180e-27	reject H0 (significant)
HOUSETYPE_MODE	9.992328e-07	reject H0 (significant)
EMERGENCYSTATE_MODE	1.138680e-06	reject H0 (significant)
NAME_TYPE_SUITE	1.132931e-05	reject H0 (significant)
FLAG_OWN_REALTY	6.681470e-04	reject H0 (significant)
FONDKAPREMONT_MODE	7.732982e-04	reject H0 (significant)
WEEKDAY_APPR_PROCESS_START	1.744737e-02	reject H0 (significant)

- **Select all significant features** (reject H0 null hypothesis) to be used in **modeling section**.
- **Select the top 5 features** from ANOVA Test and Chi-Squared Test in order to **limit our exploration**.

EDA

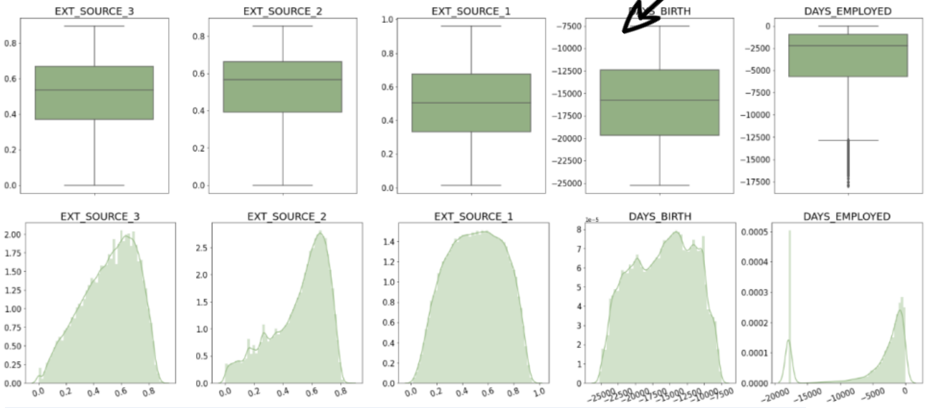
Top 5 numerical features from ANOVA Test

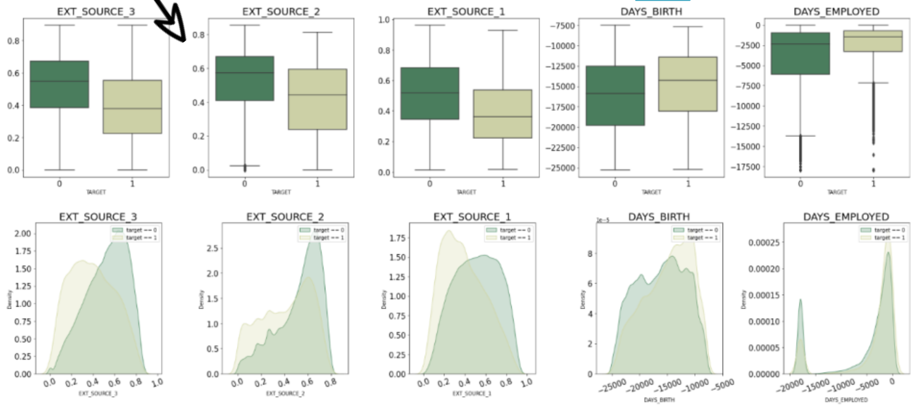
Feature Selection with ANOVA Test and Chi-Squred Test

Check data distribution in general

Check data distribution whether clients be able to repay their loan or not

Deep dive





- In overall, **the higher** \`EXT_SOURCE_1\`, \`EXT_SOURCE_2\`, and \`EXT_SOURCE_3\`, then **the more likely** the clients be able to repay their loans. According to the documentation, these features represent a "normalized score from external data source".
- External sources may be a cumulative sort of credit score rating made using numerous sources of data.
- For the \`DAYS_BIRTH\`, **the older clients tend to repay their loans**.
- For the \`DAYS_EMPLOYED\`**, the longest employed clients tend to repay their loans**.

Mean aggregation group by \`TARGET\`

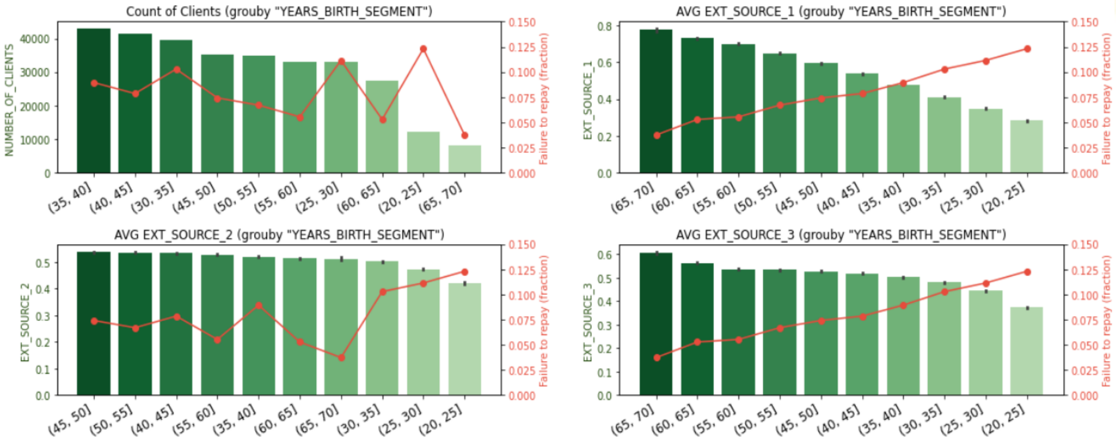
TARGET	EXT_SOURCE_3	EXT_SOURCE_2	EXT_SOURCE_1	DAYS_BIRTH	DAYS_EMPLOYED
0	0.520969	0.523479	0.511461	-16138.176397	-5305.571401
1	0.390717	0.410935	0.386968	-14884.828077	-3753.701188

Click to navigate to repository of this project

Tito Dwi Syahputra

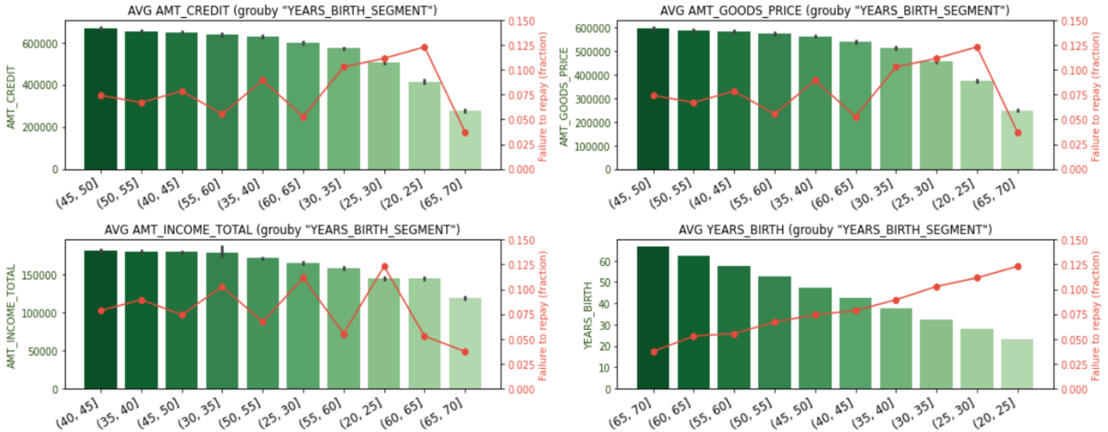
EDA (deep dive)

Segment of Client by Age



- ➢ Home Credit base clients have age between 30-45 and the smallest number of clients are in age group 20-25.
- ➢ Younger age groups are more likely to have failure repayment, especially **age group between 20-25, 25-30, and 30-35 where the failure repayment fraction is above 10%** and **below 5% for the oldest age group**.

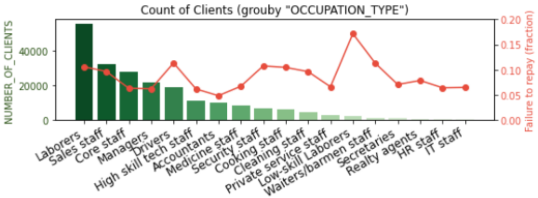
- ➢ Older age group of clients tend to loan more money compared to younger clients.
- ➢ Older age group of clients tend to have more \`EXT_SOURCE_1\`, \`EXT_SOURCE_2\`, and \`EXT_SOURCE_3\`.
- ➢ It seems like age has a bit relationship with income.

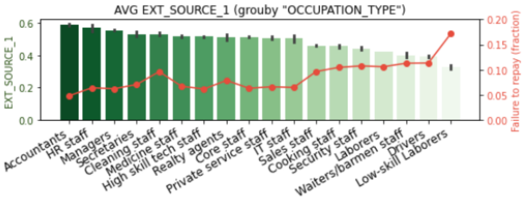


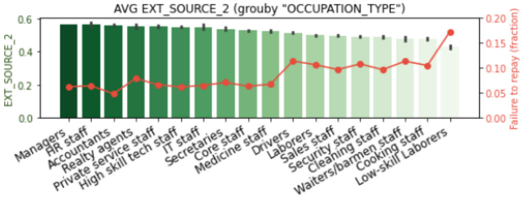
EDA (deep dive)

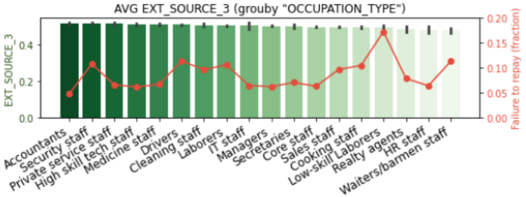
Segment of Client by Occupation

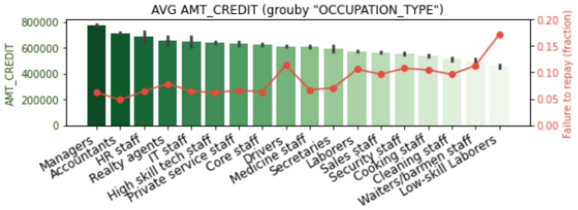
- **Most of the Home Credit's clients are employed as "Laborers"** then followed by "Sales staff", "Core staff", and "Managers". Only a few from "IT staff", "HR staff", "Realty agents", and "Secretaries".
- **"Low-skill Laborers" have the highest fraction of failure to repay** then followed by "Drivers", "Waiters/barmen staff", "Laborers", "Security staff", "Cooking staff", and "Sales staff".
- **"Accountants" have the lowest fraction of failure to repay** then followed by "High skill tech staff", "Managers", "Core staff", "HR staff", and "IT staff".

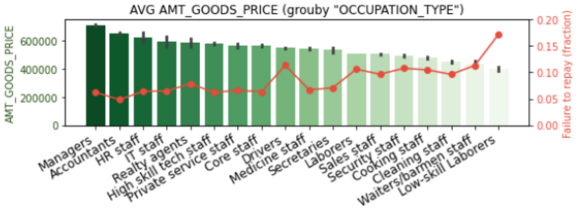


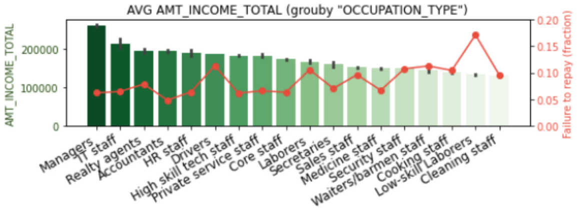


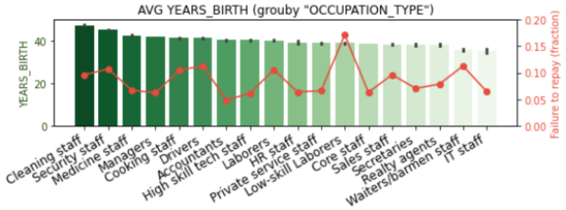












- "Managers" have the highest average income and "Cleaning staff" have the lowest.
- In sum, high level paid job more likely to loan more amount credit and goods compared to low level job.

Modeling and Evaluation

Feature Engineering

1

Modeling and Evaluation

2

Hyperparameter Tuning

3

Evaluate Final Model

	Accuracy train	ROC AUC 5-CV train	ROC AUC 5-CV validate	ROC AUC train	ROC AUC test
LogReg_baseline	0.686	0.746 ± 0.001	0.743 ± 0.003	0.746	0.730
LogReg_poly	0.684	0.746 ± 0.001	0.744 ± 0.003	0.746	0.730
LogReg_domain	0.687	0.751 ± 0.001	0.748 ± 0.003	0.751	0.736
LogReg_bureau	0.688	0.75 ± 0.001	0.747 ± 0.002	0.750	0.736
LogReg_previous	0.701	0.765 ± 0.001	0.761 ± 0.003	0.766	0.752

	Accuracy train	ROC AUC 5-CV train	ROC AUC 5-CV validate	ROC AUC train	ROC AUC test
LGBM_baseline	0.712	0.797 ± 0.0	0.757 ± 0.004	0.790	0.740
LGBM_poly	0.708	0.797 ± 0.001	0.757 ± 0.003	0.790	0.741
LGBM_domain	0.717	0.803 ± 0.001	0.764 ± 0.003	0.798	0.756
LGBM_bureau	0.720	0.807 ± 0.001	0.763 ± 0.002	0.800	0.750
LGBM_previous	0.728	0.822 ± 0.001	0.775 ± 0.004	0.814	0.763
LGBM_total	0.736	0.829 ± 0.0	0.781 ± 0.003	0.821	0.777
LGBM_missing50	0.735	0.825 ± 0.0	0.778 ± 0.003	0.818	0.772
LGBM_missing0	0.682	0.761 ± 0.001	0.725 ± 0.003	0.755	0.703
LGBM_importance	0.734	0.825 ± 0.0	0.777 ± 0.002	0.817	0.772

	Accuracy train	ROC AUC 5-CV train	ROC AUC 5-CV validate	ROC AUC train	ROC AUC test
tuned_LGBM_1	0.745	0.843 ± 0.001	0.784 ± 0.003	0.833	0.779

Feature engineering results:

- Baseline features (model**_baseline**)
- Baseline + Polynomial features (model**_poly**)
- Baseline + Domain knowledge features (model**_domain**)
- Baseline + Merge and aggregate with additional data sources (**_bureau** and **_previous**)
- Baseline + Domain + Merge and aggregate with additional data sources (model**_total)**
- Baseline + Domain + Merge and aggregate with additional data sources, then remove missing values >50% (model**_missing50**)
- Baseline + Domain + Merge and aggregate with additional data sources, then remove missing values >0% (model**_missing0**)
- Baseline + Domain + Merge and aggregate with additional data sources then remove zero importance features from LGBM feature importance (model**_importance)**

The implemented models are:

- **LogReg**: Logistic Regression
- **LGBM**: LightGBM

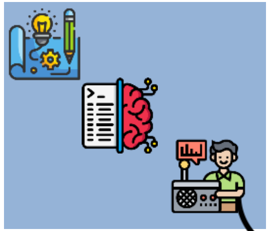
Optimize hyperparameters of the best model, **LGBM_total**, using random search cross validation.

Results:

- The best model is **LGBM_total**. This model was trained using features from Baseline + Domain + Merge and aggregate with additional data sources with **793 total features**.
- After hyperparameter tuning using RandomSearchCV (i.e. **tuned_LGBM_1**), we have improved overall model performance metrics!
- The **ROC AUC test** are obtained after submitting submission file into Home Credit Kaggle Competition.

Tito Dwi Syahputra

Previous workflows



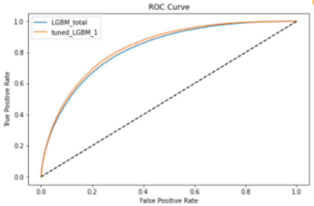
Evaluate Final Model

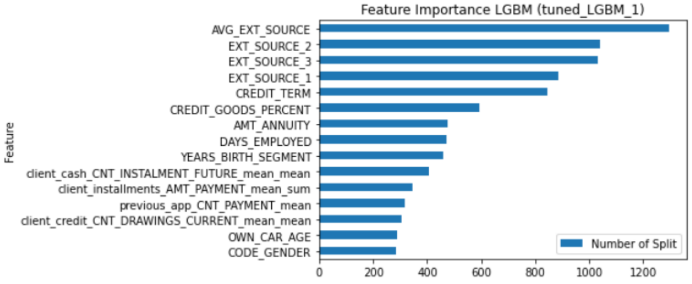
4

Modeling and Evaluation

	Accuracy train	ROC AUC 5-CV train	ROC AUC 5-CV validate	ROC AUC train	ROC AUC test
LGBM_total	0.736	0.829 ± 0.0	0.781 ± 0.003	0.821	0.777

	Accuracy train	ROC AUC 5-CV train	ROC AUC 5-CV validate	ROC AUC train	ROC AUC test
tuned_LGBM_1	0.745	0.843 ± 0.001	0.784 ± 0.003	0.833	0.779





LGBM_total model

```
======== model evaluation metrics "LGBM_total" ========
confusion matrix and classification report "app_train":
[[207873  74813]
 [  6228  18597]]
              precision    recall  f1-score   support

      repaid       0.97      0.74      0.84    282686
  not repaid       0.20      0.75      0.31     24825

    accuracy                           0.74    307511
   macro avg       0.59      0.74      0.58    307511
weighted avg       0.91      0.74      0.79    307511
```

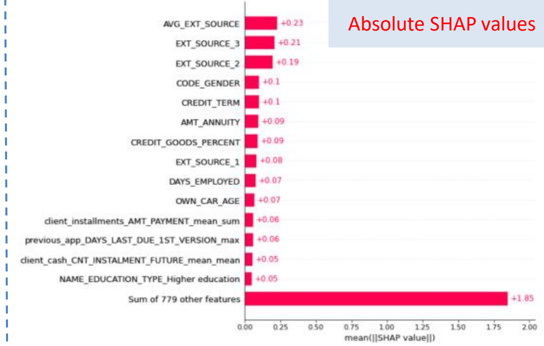
tuned_LGBM_1 model

```
======== model evaluation metrics "tuned_LGBM_1" ========
confusion matrix and classification report "app_train":
[[210263  72423]
 [  5876  18949]]
              precision    recall  f1-score   support

      repaid       0.97      0.74      0.84    282686
  not repaid       0.21      0.76      0.33     24825

    accuracy                           0.75    307511
   macro avg       0.59      0.75      0.58    307511
weighted avg       0.91      0.75      0.80    307511
```

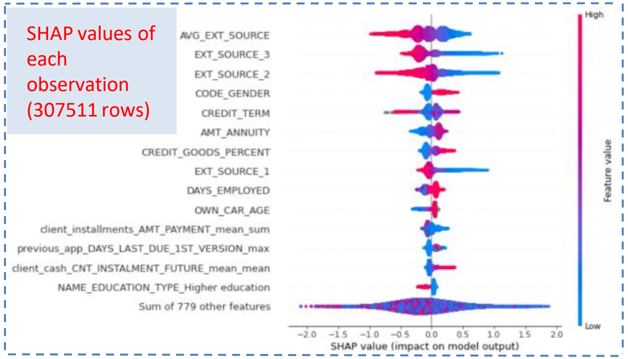
Feature importance:
- The **top 4 feature importances** are "normalized score from external data source".
- Domain knowledge feature engineering i.e. `CREDIT_TERM`, `CREDIT_GOODS_PERCENT`, `YEARS_BIRTH_SEGMENT` are within top 15 of feature importance.
- Some of feature engineering from other data sources seems to be within feature importance.

SHAP values:
- `AVG_EXT_SOURCE`, `EXT_SOURCE_3`, `EXT_SOURCE_2` are the top 3 features which have high impact on LGBM model output. These three features negatively affect the model outcome meaning that as these feature values increase, then the probability or odds of failure repayment decreases.
- The LGBM model output interpret that male gender (encode as 1 meaning high feature value) will increase the probability or odds of failure repayment.

Absolute SHAP values



SHAP values of each observation (307511 rows)



Conclusions and Recomendations

Conclusions:

- We have successfully built a machine learning model using Logistic Regression and LightGBM algorithm to predict whether or not an applicant will be able to repay a loan. The best model is **LightGBM** with **77,7%** ROC AUC score and after hyperparameter tuning the ROC AUC score increases to **77,9%**.
- The best segment of clients by age are clients with age above 50 (failure to repay rate < **7,5%**) and the best segment of clients by occupation are "Managers" and "Accountants"(failure to repay rate < **7,5%** and also high amount of credit and income).
- Based on SHAP values, most of failure repayment of loan comes from clients with **low "normalized score from external data source"**. This normalized score might be interpreted as credit score, so low credit score means **untrustworthy clients**.

Recommendations on client's age:

- Marketing strategy has to be more **focus on older age group** since in overall they have lower failure repayment rate and higher external source score (more trustworthy). Essentially, these groups may have better and stable financial condition.
- As the **younger clients** are less likely to repay the loan, maybe **Home Credit should be provided with more guidance or financial planning tips** to younger clients. This doesn't mean that Home Credit should discriminate against younger clients, but it would be smart to take precautionary measures to help younger clients pay on time.

Recommendations on clients' occupation:

- The higher the level paid job that clients have, then the more likely that clients repay the loan. In particular, clients with **"IT staff"** and **"HR staff"** as occupation are only a few in Home Credit. In addition, those clients more likely to repay their loan and apply good amount of credit. Therefore, **Home Credit need to focus their marketing strategy to increase the number of clients with high level paid job such as "IT staff" and "HR staff" since these are potential profitable clients**.
- Most of Home Credit clients' occupation are **"Laborers"** (the number of these clients are above 40.000 which is the highest over all other types of occupations). Hence, **Home Credit should put more attention and special treatment to these clients** such as giving them lower monthly payment rates, creating a special loan program or loyalty program for those clients who always repay their loan on time.

looking forward your feedback, as long as it's constructive and honest!

THANK YOU!

Please, connect with me and check my other projects!

https://id.linkedin.com/in/tito-dwi-syahputra

https://github.com/titods/Projects