MAFS 6010Z Project 1: Warm-up of Statistical Machine Learning

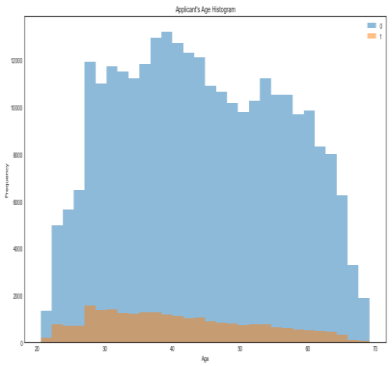
CHEN Yuying, LOU Ruoyu and SHANG Zhiheng {ychenhn, rlou, zshangaa}@connect.ust.hk
Department of Mathematics, HKUST

1. Introduction

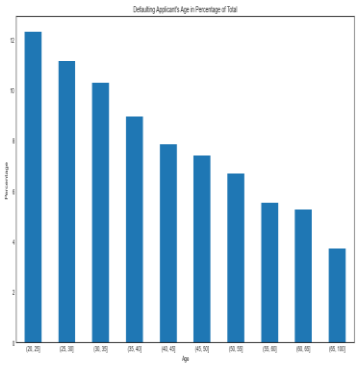
The objective of this project is to predict clients' repayment abilities through historical loan application data. Several features are visualized in correlation with the target variable and more emphasis will be placed on highly correlated ones in variable selection. After model comparing and selection, the **LGBM** is the final one for data fit and prediction.

2. Single Feature Correlation

Exploring single feature in correlation with the target variable is helpful to understand data relationship. Data modification and new feature extraction sometimes are also necessary for variable analysis. Taking feature [Days Birth] in dataset as example. For the figures comparing the applicant's age, there is no significant different between default and non-default. However, when we use percentage of defaulting in the age groups, we could see a correlation between the age and the default variable. **Younger** applicants are more likely to default rather than **older** ones. This feature will most likely be useful for the model.



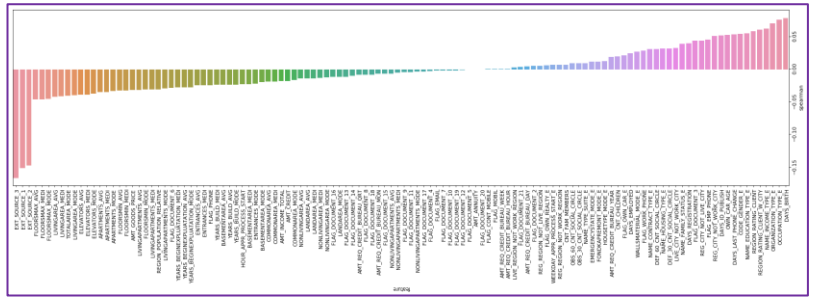
Applicant's Age Histogram



Defaulting Application's Age in Percentage of Total

3. Multivariable Correlation

Pearson correlation coefficient is applied in feature selection part. However, for this dataset, the method is not the effective one as it actually **worsen** the performance of the model. At last **all features** are included in the model fitting.



Correlation between Features and Response Variable

4. Missing Value and Multicollinearity

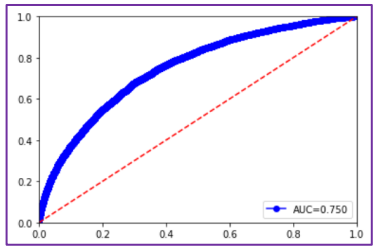
There are lots of missing values in dataset. Since missing values are also a valuable information, so they will be **left as null**.
However, considering the multicollinearity problem in multiple regression model may largely effect the variance, we use **VIF** to show and eliminate variables with high collinearity.

VIF Modification

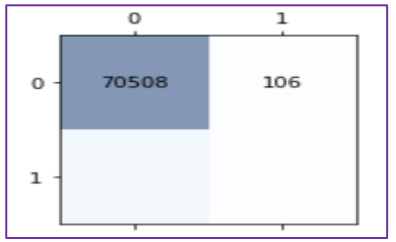
Feature	VIF
LIVINGAREA_AVG	467.7299613
LIVINGAREA_MEDI	428.0937897
YEARS_BUILD_AVG	366.7937811
YEARS_BUILD_MEDI	338.2328955
OBS_60_CNT_SOCIAL_CIRCLE	313.1557909
OBS_30_CNT_SOCIAL_CIRCLE	312.833265
APARTMENTS_MEDI	301.9062631
APARTMENTS_AVG	301.653541
ELEVATORS_MEDI	283.4109205
FLOORSMAX_MEDI	268.654587
ENTRANCES_MEDI	242.3252771
FLOORSMIN_MEDI	240.9464584
LIVINGAPARTMENTS_MEDI	239.4232216
ELEVATORS_AVG	234.9398056
FLOORSMAX_AVG	231.026847
LIVINGAPARTMENTS_AVG	219.3873828

5. Results

Several classifiers(LDA, LR, RF, GB, SGD, XGB and LGBM) are engaged to fit the data. **AUC** value is chosen as evaluation criteria and at last **LGBM** beats all the others. AUC curve and Confusion Matrix Plot are shown below. The Kaggle test is 0.748 (AUC value).



AUC Curve for LGBM



Confusion Matrix Plot for LGBM

6. Conclusion

In this project, the dataset contains lots of missing values and most features don't show strong correlation with target variable. It's not a surprise to see LGBM wins the game considering LGBM takes advantages of sparse optimization and parallel training.

7. References

Brownlee, Jason (March 31, 2020). "Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost"

8. Contribution

Data Processing and Analyzing
- CHEN Yuying 20744353
- SHANG Zhiheng 20738938

Models Testing and Comparing
- LOU Ruoyu 20743763