PROJECT REPORT

PRACTICE MODULE FOR CERTIFICATE IN: PATTERN RECOGNITION SYSTEMS (PRS)

Home Credit Default Risk System

TEAM MEMBERS

SIVA KRISHNA THOTA

KOH HONG WEI

Contents

1.0 EXECUTIVE SUMMARY

An important fraction of the population finds it difficult to get their home loans approved due to insufficient or non-existent credit history. This prevents them to buy their dream homes and at times even forces them to rely on other sources of money which may be unreliable and have exorbitant interest rates. Conversely, it is a major challenge for banks and other finance lending agencies to decide for which candidates to approve housing loans. The credit history is not always a sufficient tool for decisions, since it is possible that those borrowers with a long credit history can still default on the loan and some people with a good chance of loan repayment may simply not have a sufficiently long credit history.

Thus, the project team sees the potential to develop a Home Credit Default Risk System (HCDRS) which will help to solve the above business problem. Specifically, it will help to identify if a new loan applicant shows a high / low risk for loan default. This will help to:
1) Reduce uncertainty from the home loan provider perspective
2) Reduce risk of lending to a client with high risk of default, and
3) Ensure home loan can be provided to segment such as the unbanked population or people with insufficient or non-existent credit history.

A number of recent researchers have applied machine learning to predict the loan default risk. This is important since a machine learning-based classification tool to predict the loan default risk which uses more features than just the traditional credit history can be of great help for both, potential borrowers, and the lending institutions. Hence, the project team aims to train a best-in-class classifier using a variety of machine learning techniques on the dataset to help determine the risk of loan default.

For testing the live system/demonstration, please refer to http://68.178.202.122:5000/login

2.0 PROBLEM STATEMENT / BUSINESS OPPORTUNITY

The problem statement can be summarized as a binary classification problem where the inputs are various features describing the financial and behavioural history of the loan applicants, in order to predict whether the loan will be repaid or defaulted.

2.1 ASSUMPTIONS

The project makes a key assumption on the data's accuracy. The source/raw data is assumed to be accurate as failing which, the model's accuracy will be affected.

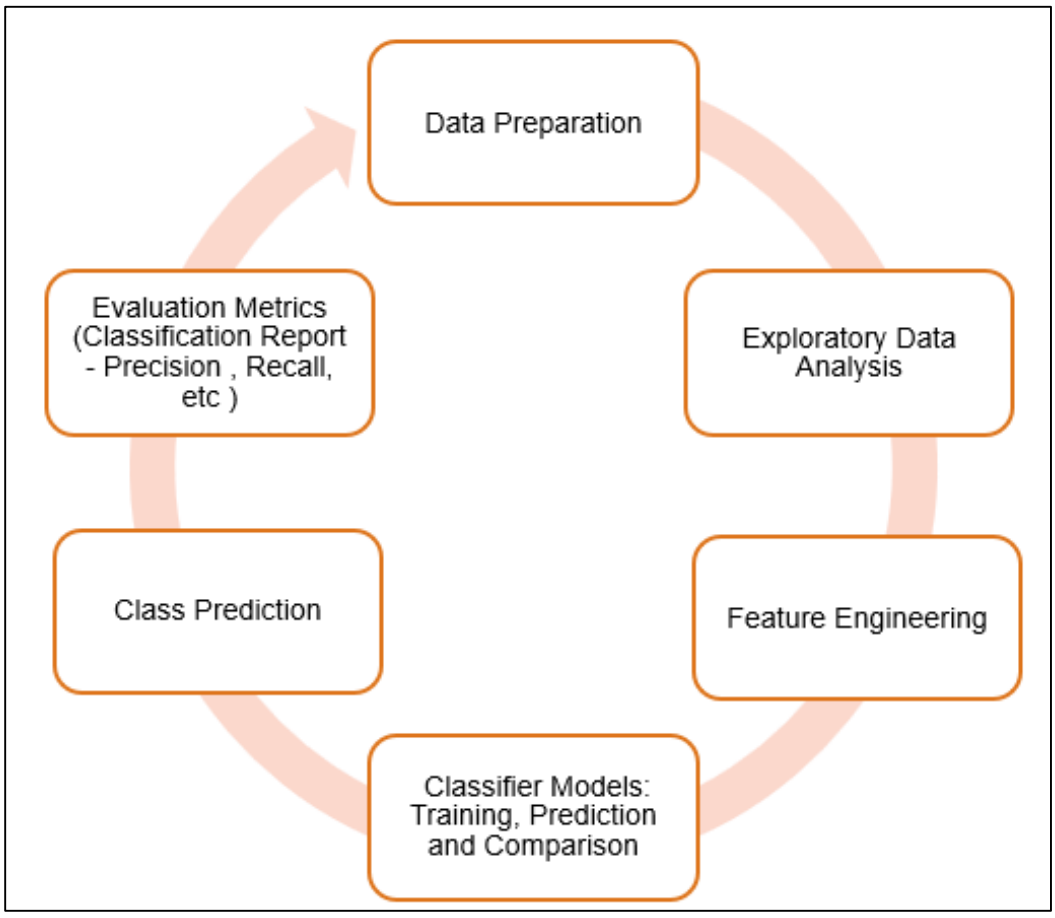
2.2 PROJECT OBJECTIVES

the team envisions developing a Home Credit Default Risk System (HCDRS) to:
1) Reduce uncertainty from the home loan provider perspective
2) Reduce risk of lending to a client with high risk of default and
3) Ensure home loan can be provided to segment such as the unbanked population or people with insufficient or non-existent credit history.

2.3 PROJECT SCOPE

The Project scope consists of 6 broad areas and is illustrated below.

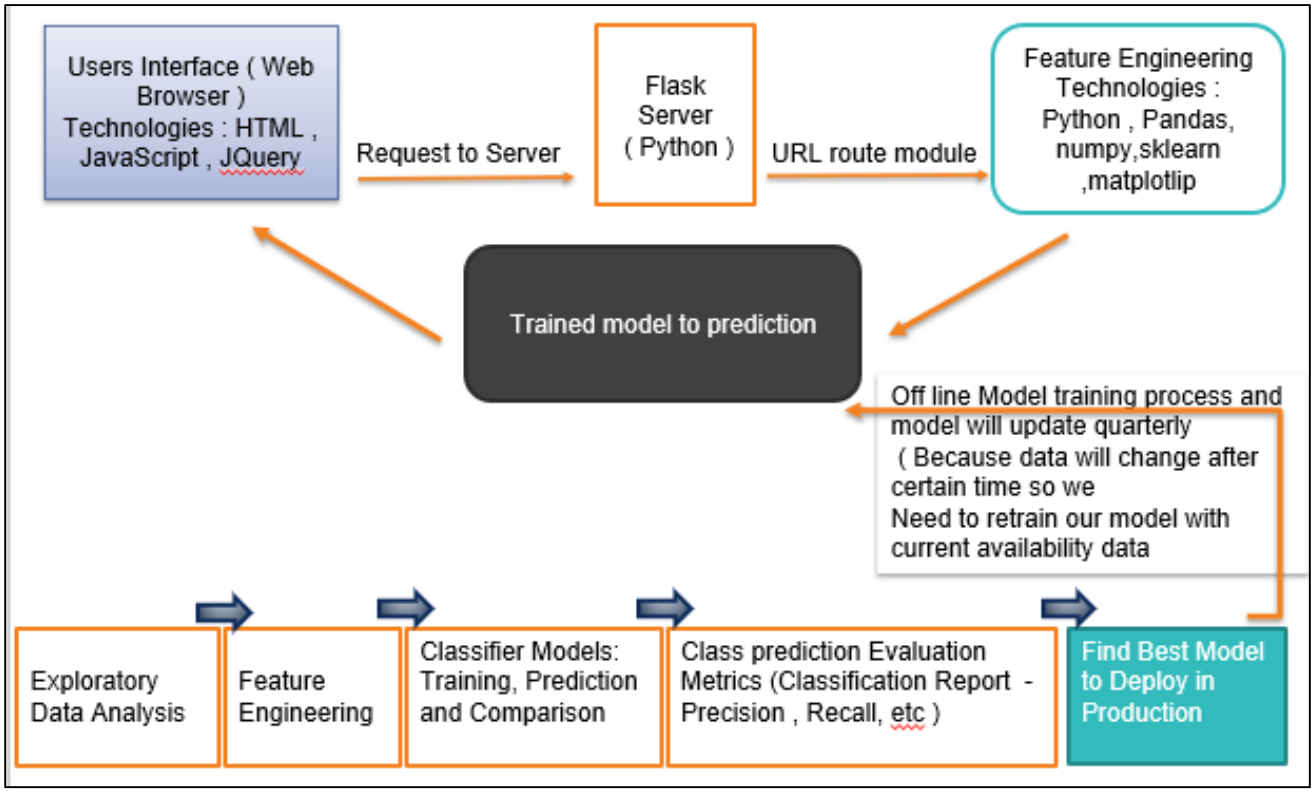
Figure 1: Project scope



2.4 SYSTEM ARCHITECHTURE / DESIGN

The System architecture / design is illustrated below.

Figure 2: System architecture / design



3.0 DATA SET & DATA CLEANSING

The Data is taken from Kaggle's competition which is available publicly.

There are ~300K records and 122 columns with different data types:

- 65 floating point numbers
- 41 integer numbers – some are numerical categories
- 16 Strings – Categories involved

Some samples of the data field are: Gender, Education, Income, Housing type etc.

Since the ML model can't inherently deal with text, the data must be converted to appropriate numbers. Any significant distortion/noise in the model must be removed as much as possible.

Due to the huge data set and inherent noises, there are many cleaning steps involved and listed as follows:

- Converting string categorical columns into numerical – Label encoding
- Converting string categorical columns into numerical and adding new columns to indicate the presence of categorical variables – One hot encoding
- Replacing illogical outliers with empty values (NAN values)
- Imputing empty cells with the median of the values. In some cases, imputation is approached with a certain grouping
- Dealing with a few anomalies (see Tables 1-3)
- Changing invalid entries into valid entries

Table 1: Training Data Set Top 15 Columns with Missing values

	Missing Count	Missing Count Ratio	Missing Count %
COMMONAREA_MEDI	214865	0.698723	69.9
COMMONAREA_AVG	214865	0.698723	69.9
COMMONAREA_MODE	214865	0.698723	69.9
NONLIVINGAPARTMENTS_MODE	213514	0.694330	69.4
NONLIVINGAPARTMENTS_AVG	213514	0.694330	69.4
NONLIVINGAPARTMENTS_MEDI	213514	0.694330	69.4
FONDKAPREMONT_MODE	210295	0.683862	68.4
LIVINGAPARTMENTS_MODE	210199	0.683550	68.4
LIVINGAPARTMENTS_AVG	210199	0.683550	68.4
LIVINGAPARTMENTS_MEDI	210199	0.683550	68.4
FLOORSMIN_AVG	208642	0.678486	67.8
FLOORSMIN_MODE	208642	0.678486	67.8
FLOORSMIN_MEDI	208642	0.678486	67.8
YEARS_BUILD_MEDI	204488	0.664978	66.5
YEARS_BUILD_MODE	204488	0.664978	66.5

Table 2: Test Data Set Top 15 Columns with Missing values

	Missing Count	Missing Count Ratio	Missing Count %
COMMONAREA_AVG	33495	0.687161	68.7
COMMONAREA_MODE	33495	0.687161	68.7
COMMONAREA_MEDI	33495	0.687161	68.7
NONLIVINGAPARTMENTS_AVG	33347	0.684125	68.4
NONLIVINGAPARTMENTS_MODE	33347	0.684125	68.4
NONLIVINGAPARTMENTS_MEDI	33347	0.684125	68.4
FONDKAPREMONT_MODE	32797	0.672842	67.3
LIVINGAPARTMENTS_AVG	32780	0.672493	67.2
LIVINGAPARTMENTS_MODE	32780	0.672493	67.2
LIVINGAPARTMENTS_MEDI	32780	0.672493	67.2
FLOORSMIN_MEDI	32466	0.666051	66.6
FLOORSMIN_AVG	32466	0.666051	66.6
FLOORSMIN_MODE	32466	0.666051	66.6

Table 3: Different kinds of classes in Every categorical column

	Categorical Column	Count
0	NAME_CONTRACT_TYPE	2
1	CODE_GENDER	3
2	FLAG_OWN_CAR	2
3	FLAG_OWN_REALTY	2
4	NAME_TYPE_SUITE	7
5	NAME_INCOME_TYPE	8
6	NAME_EDUCATION_TYPE	5
7	NAME_FAMILY_STATUS	6
8	NAME_HOUSING_TYPE	6
9	OCCUPATION_TYPE	18
10	WEEKDAY_APPR_PROCESS_START	7
11	ORGANIZATION_TYPE	58
12	FONDKAPREMONT_MODE	4
13	HOUSETYPE_MODE	3
14	WALLSMATERIAL_MODE	7
15	EMERGENCYSTATE_MODE	2

Table 4: Clients Repayment Abilities By Suite Type

	NAME_TYPE_SUITE	TARGET	SK_ID_CURR
0	Children	No Payment Difficulties	3026
1	Children	Payment Difficulties	241
2	Family	No Payment Difficulties	37140
3	Family	Payment Difficulties	3009
4	Group of people	No Payment Difficulties	248
5	Group of people	Payment Difficulties	23
6	Other_A	No Payment Difficulties	790
7	Other_A	Payment Difficulties	76
8	Other_B	No Payment Difficulties	1596
9	Other_B	Payment Difficulties	174
10	Spouse, partner	No Payment Difficulties	10475
11	Spouse, partner	Payment Difficulties	895
12	Unaccompanied	No Payment Difficulties	228189
13	Unaccompanied	Payment Difficulties	20337

Table 5: Clients Repayment Abilities by Realty Ownership Status

	FLAG_OWN_REALTY	TARGET	SK_ID_CURR
0	No	No Payment Difficulties	86357
1	No	Payment Difficulties	7842
2	Yes	No Payment Difficulties	196329
3	Yes	Payment Difficulties	16983

3.1 EXPLORATORY DATA ANALYSIS (EDA)

After cleaning the data, the Project team did an exploration of the data (i.e. EDA). The results suggest that:

- most people paid off their loan, i.e. returned the borrowed money (Figure 3)
- Labourers is the most frequent occupation type while the least frequent is IT Staff (Figure 4)

Figure 3: EDA

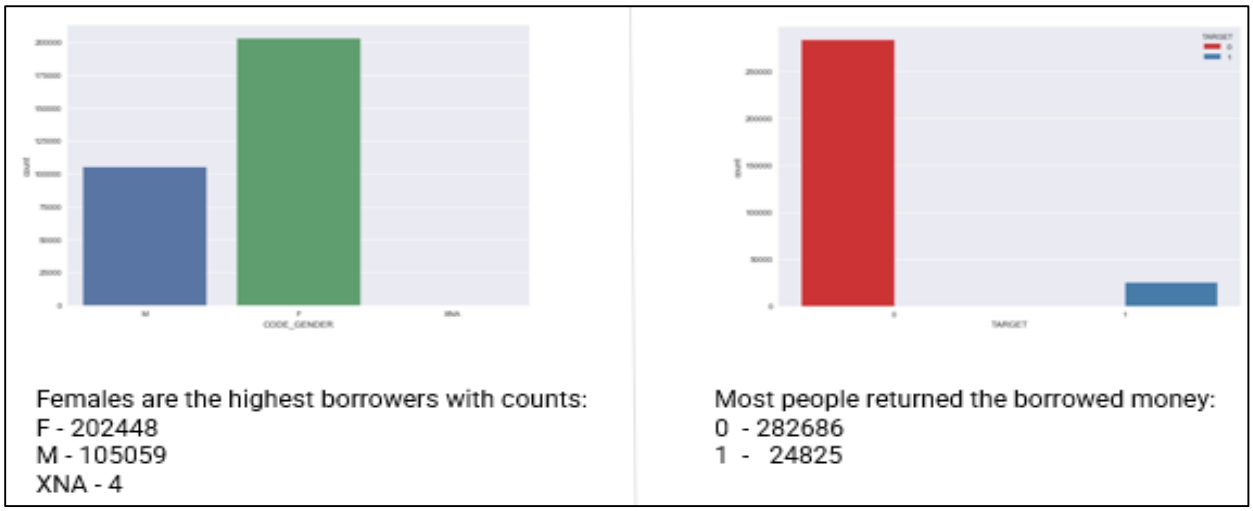
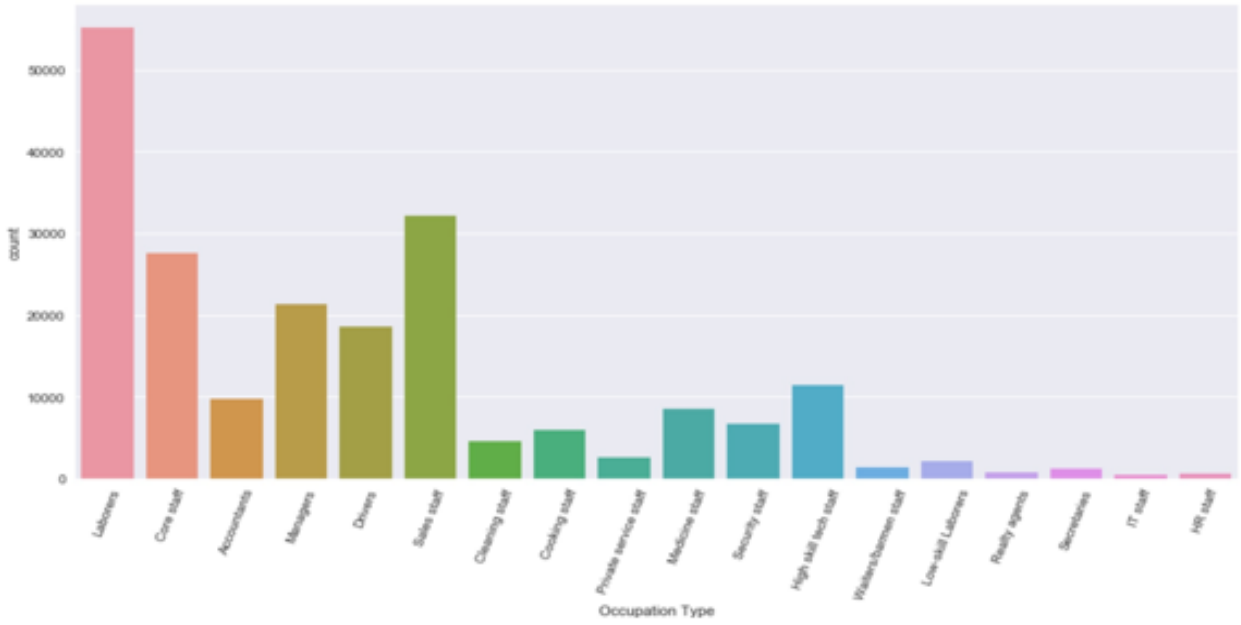
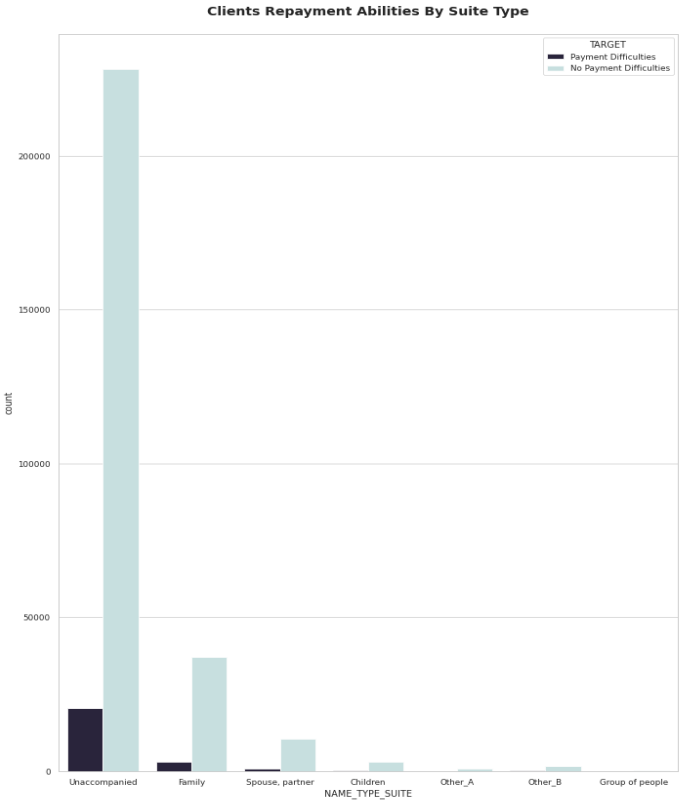
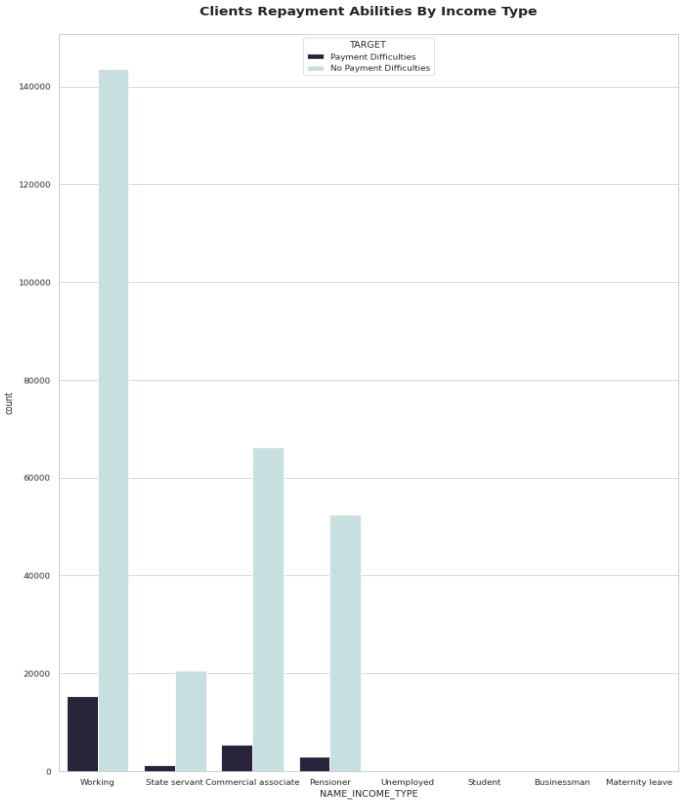
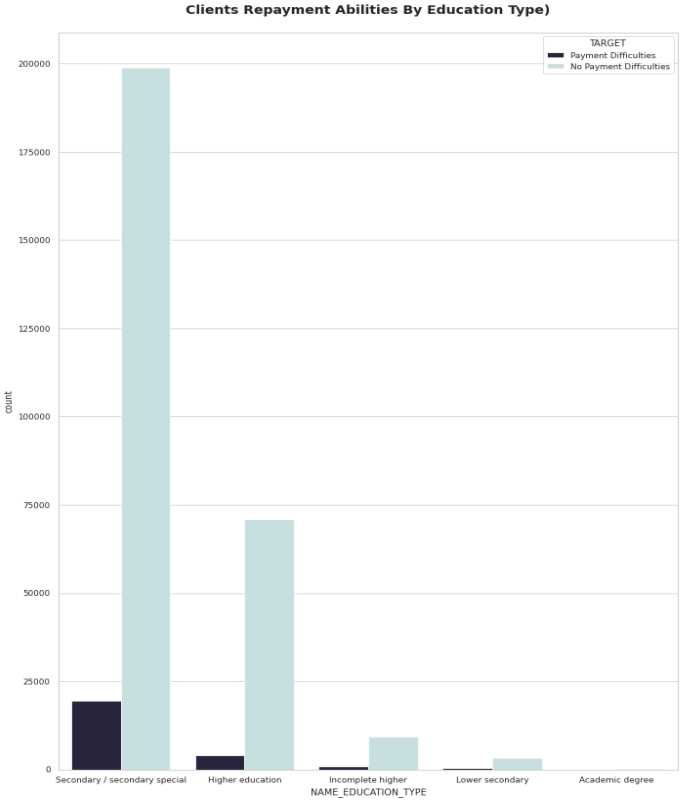


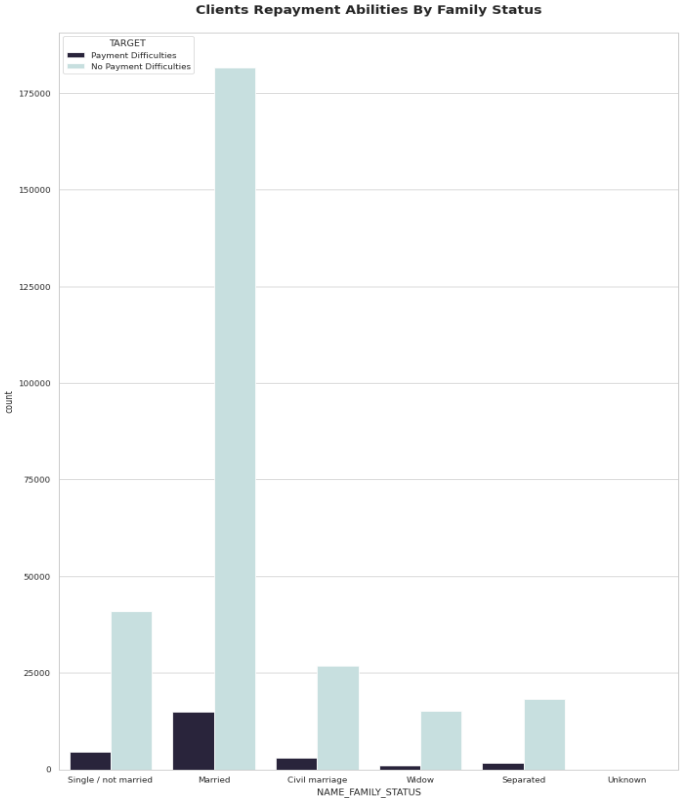
Figure 4: Occupation Type histogram

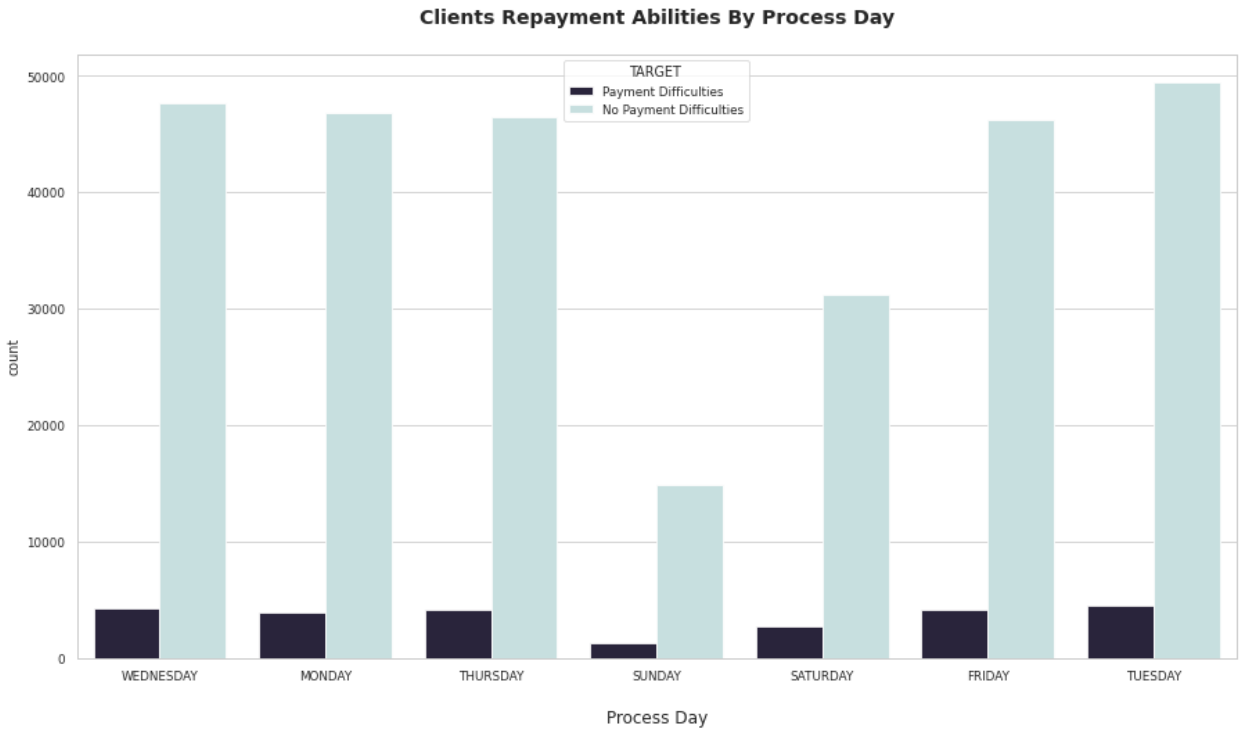



Clients Repayment Abilities By Suite Type
TARGET
Payment Difficulties
No Payment Difficulties
count
NAME_TYPE_SUITE
Unaccompanied
Family
Spouse, partner
Children
Other_A
Other_B
Group of people

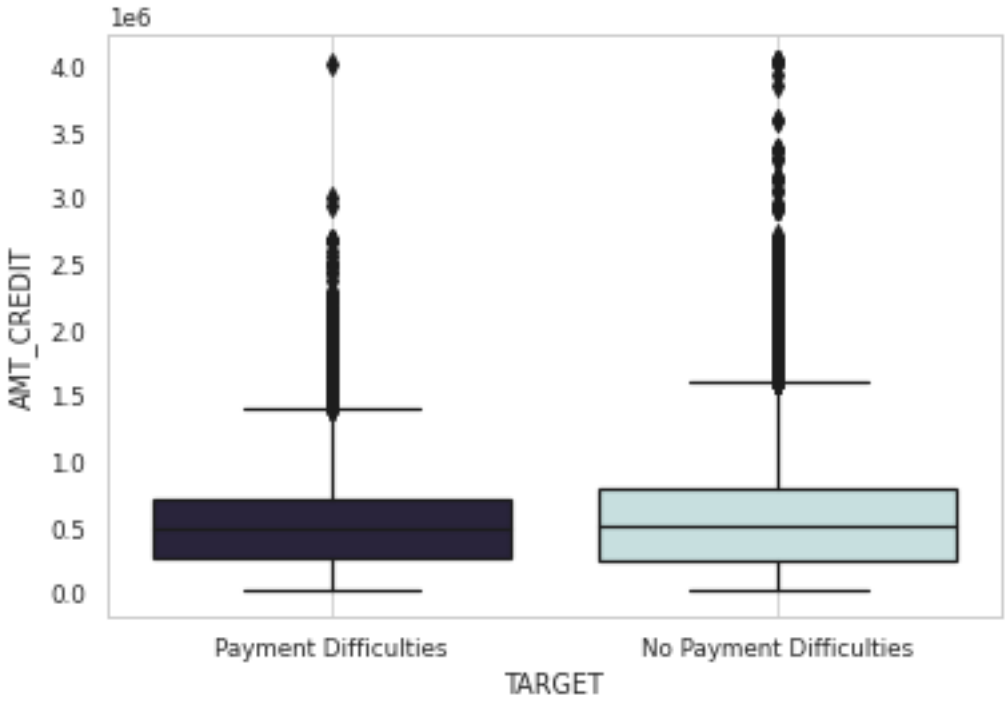

Clients Repayment Abilities By Income Type
TARGET
Payment Difficulties
No Payment Difficulties
count
NAME_INCOME_TYPE
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave


Clients Repayment Abilities By Education Type)
TARGET
Payment Difficulties
No Payment Difficulties
count
NAME_EDUCATION_TYPE
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

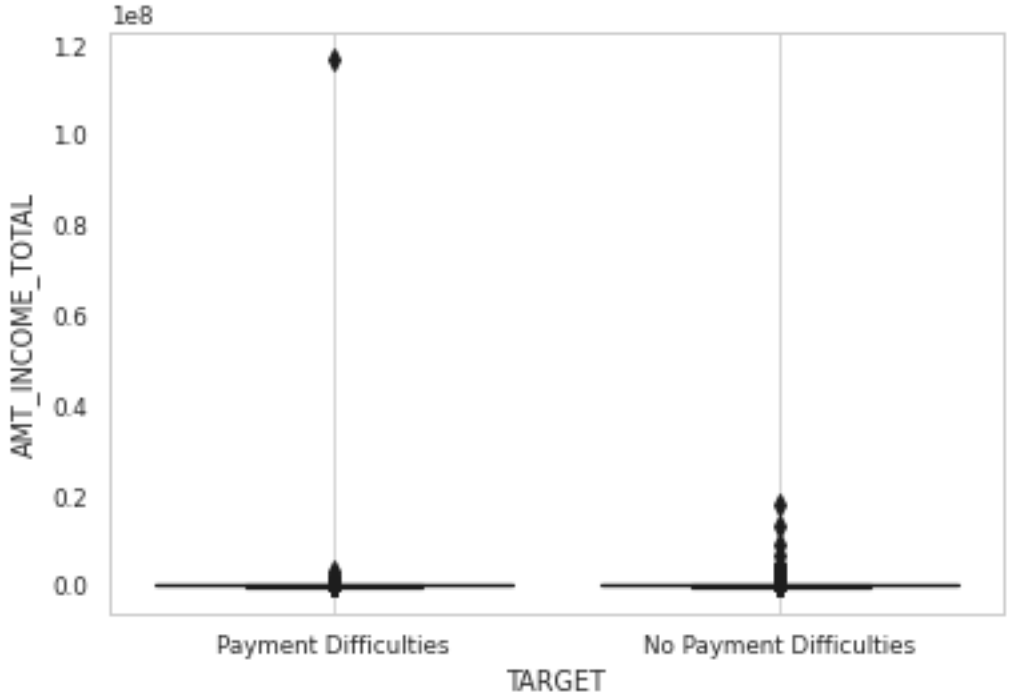

Clients Repayment Abilities By Family Status
TARGET
Payment Difficulties
No Payment Difficulties
count
NAME_FAMILY_STATUS
Single / not married
Married
Civil marriage
Widow
Separated
Unknown

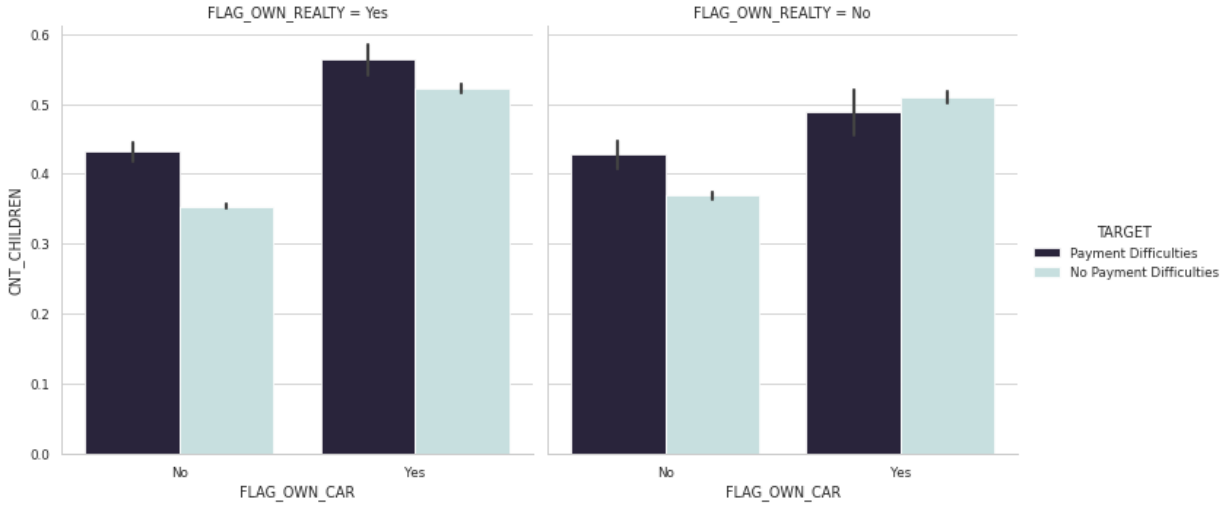


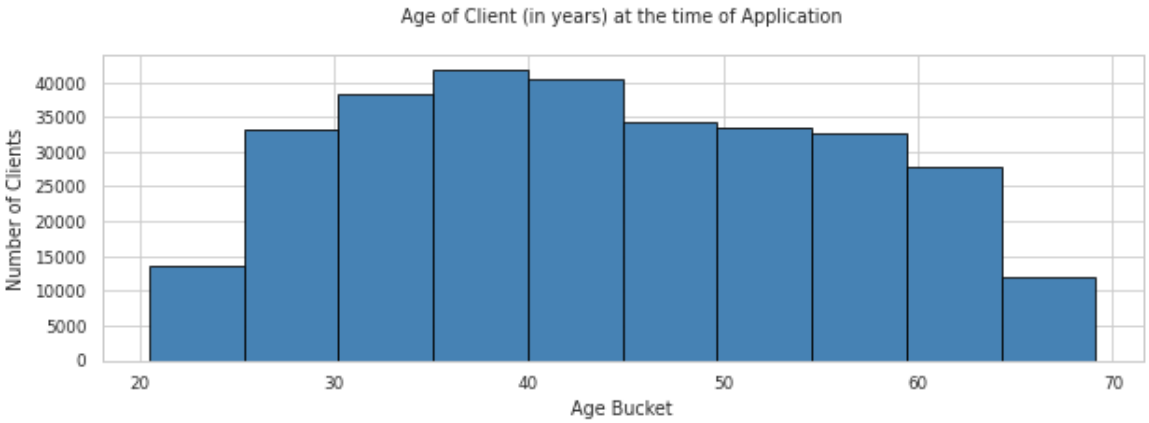
Amount Credit of the Loan vs Target

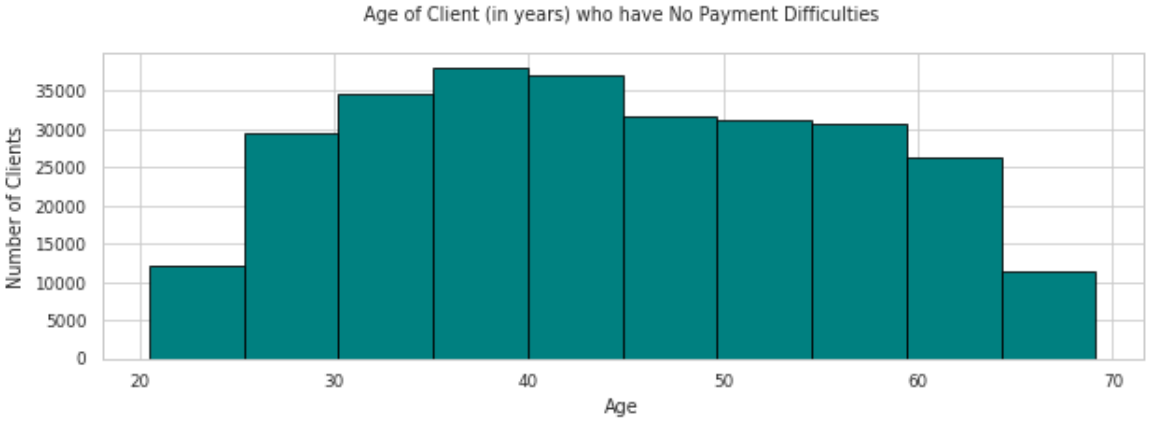


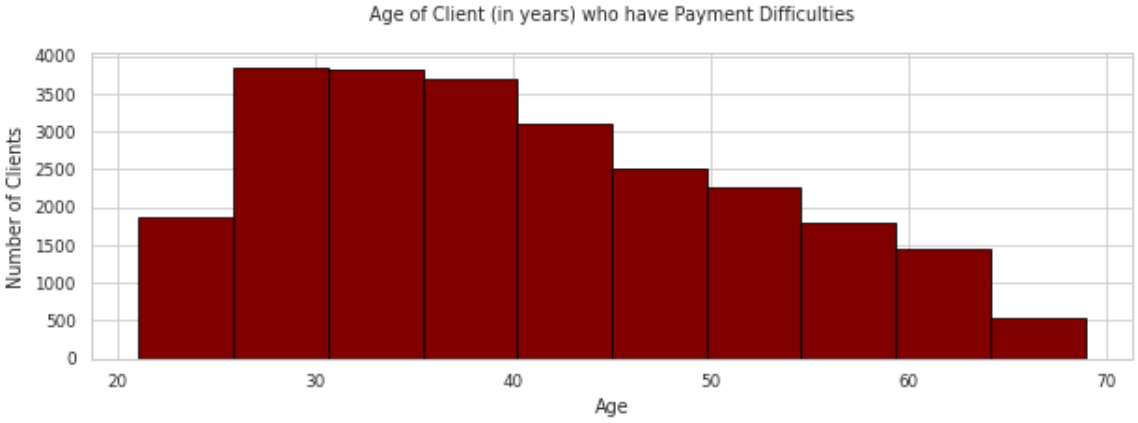
Amount Income vs Target

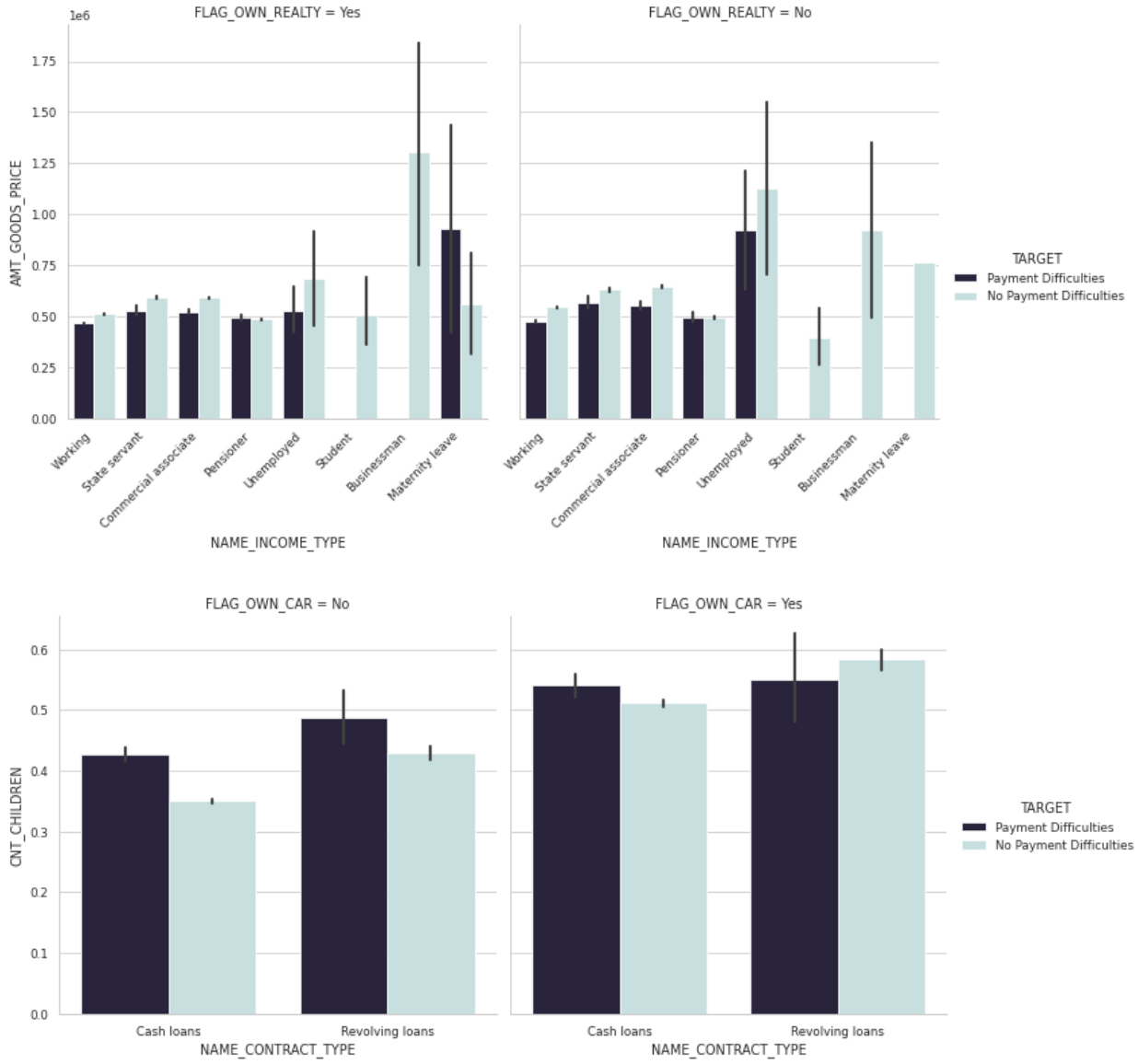


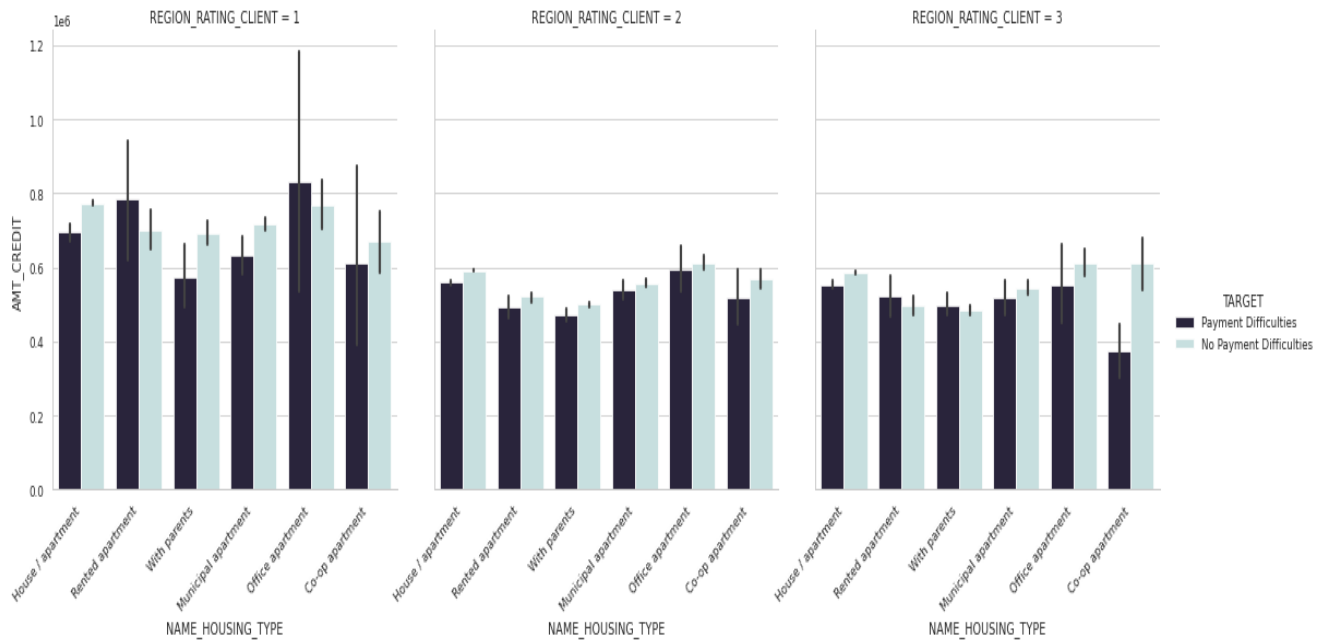


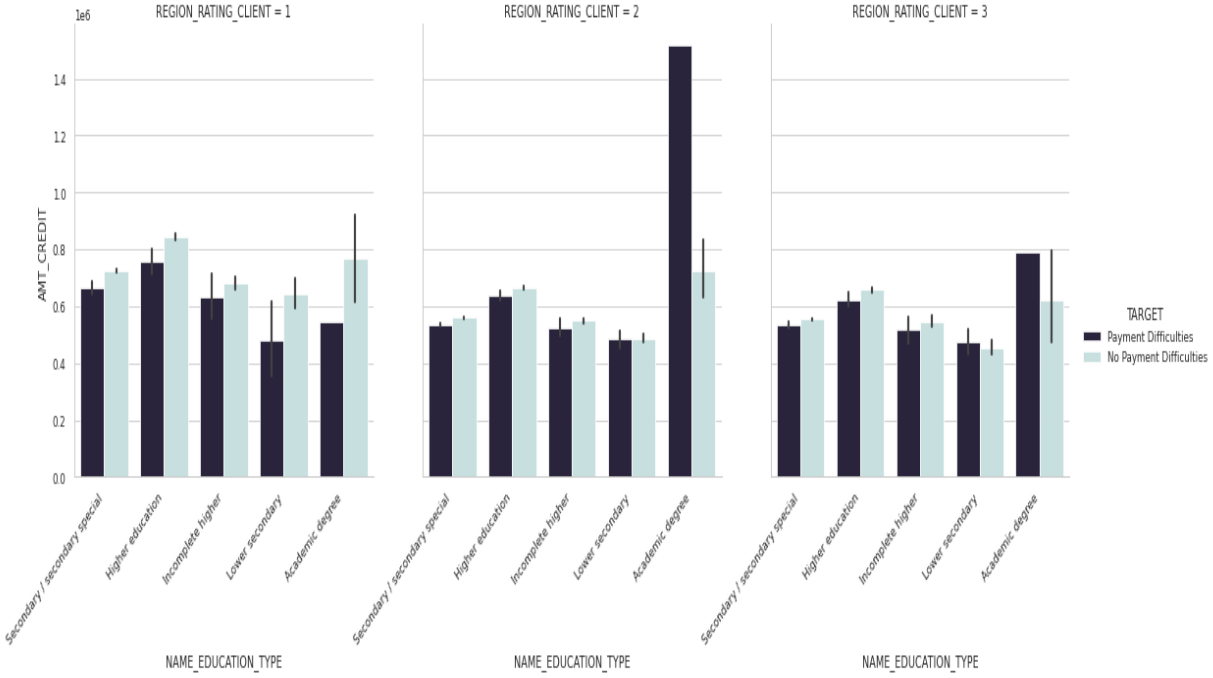

Age of Client (in years) at the time of Application
Number of Clients
0
5000
10000
15000
20000
25000
30000
35000
40000
20
30
40
50
60
70
Age Bucket


Age of Client (in years) who have No Payment Difficulties
Number of Clients
0
5000
10000
15000
20000
25000
30000
35000
20
30
40
50
60
70
Age

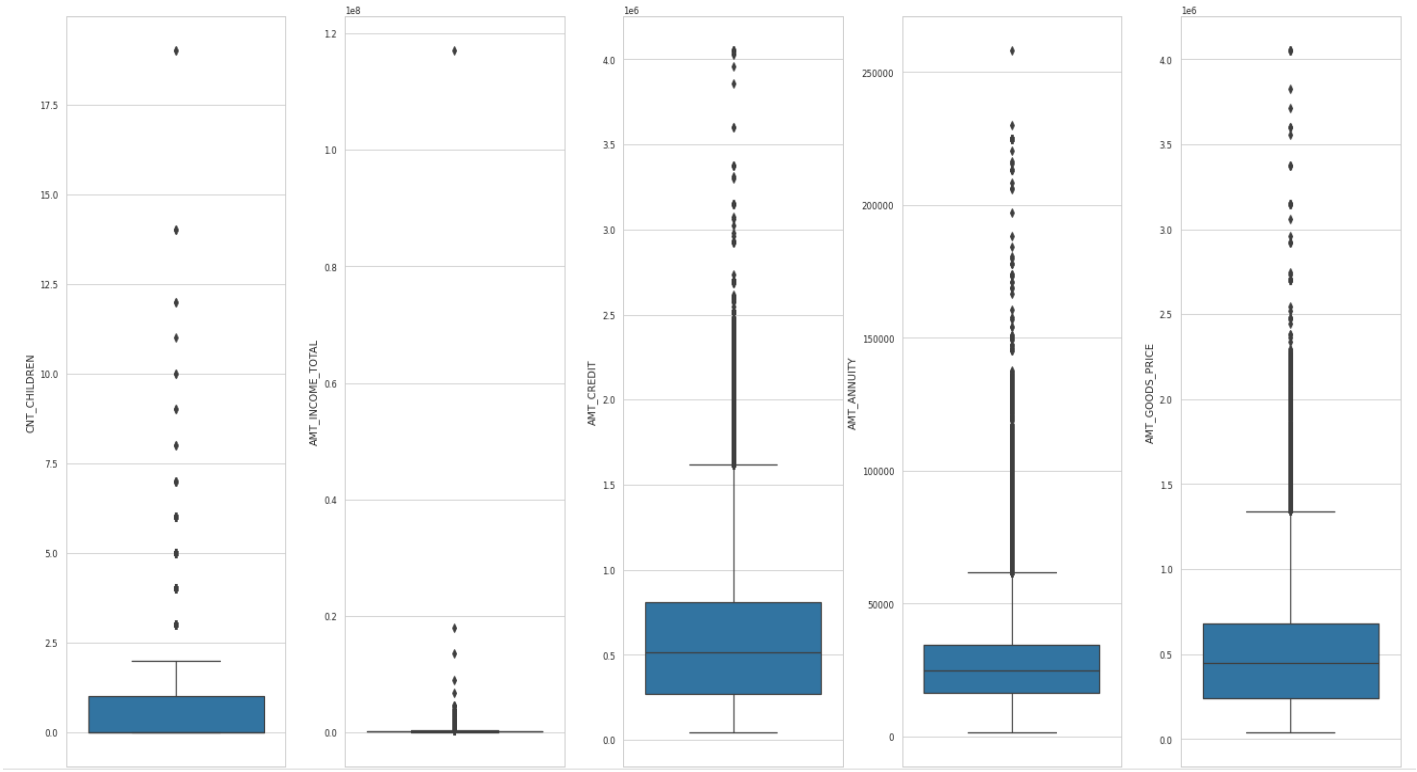

Age of Client (in years) who have Payment Difficulties
Number of Clients
0
500
1000
1500
2000
2500
3000
3500
4000
20
30
40
50
60
70
Age

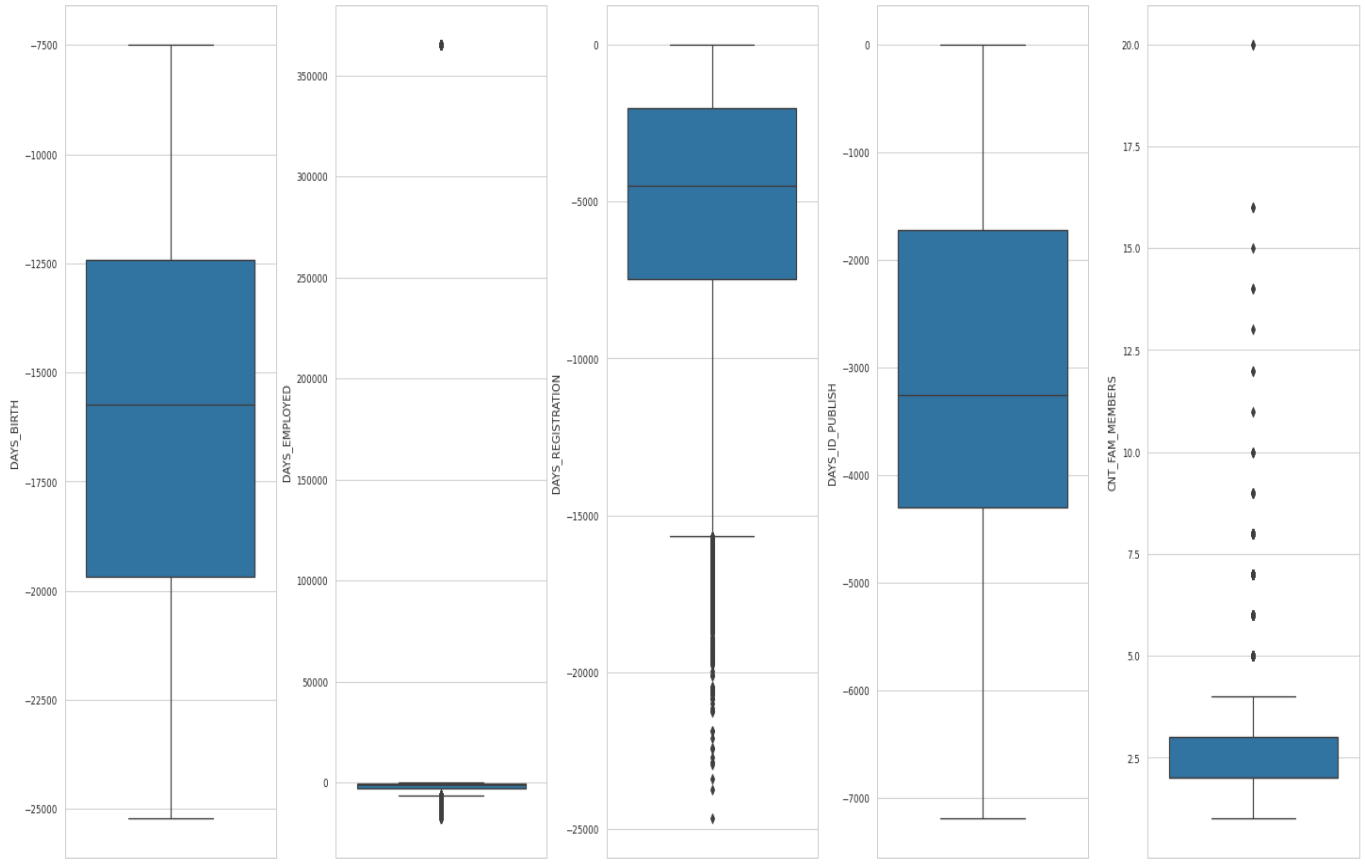

FLAG_OWN_REALTY = Yes
FLAG_OWN_REALTY = No
1e6
AMT_GOODS_PRICE
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE
TARGET
Payment Difficulties
No Payment Difficulties
FLAG_OWN_CAR = No
FLAG_OWN_CAR = Yes
CNT_CHILDREN
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
TARGET
Payment Difficulties
No Payment Difficulties

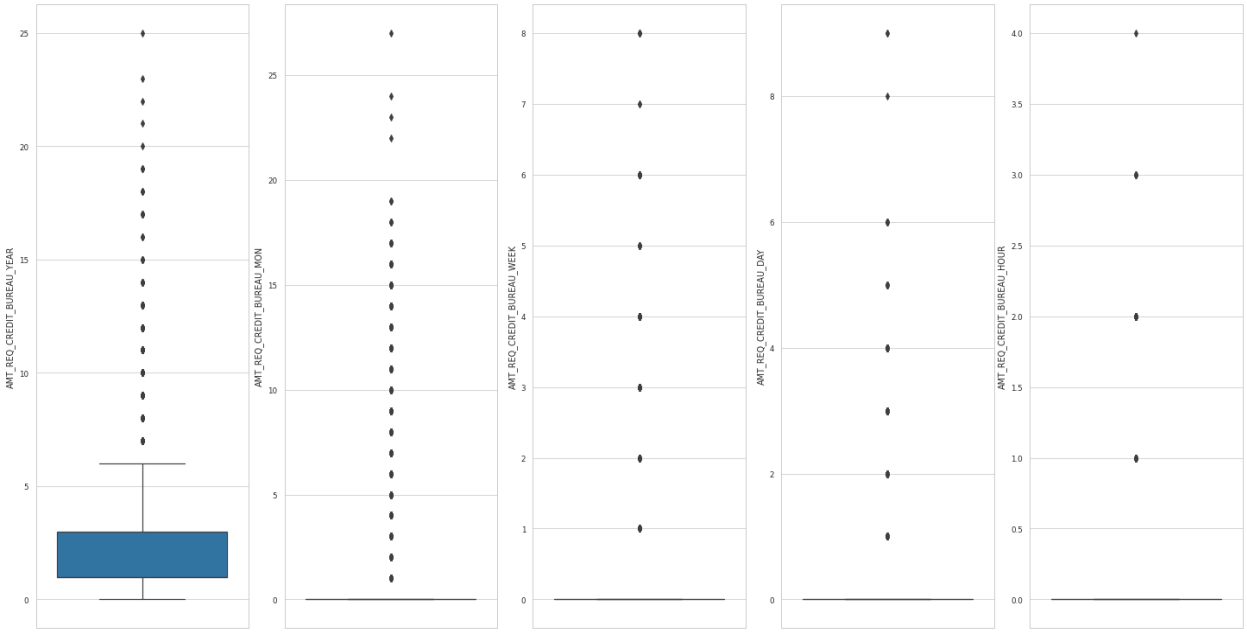

REGION_RATING_CLIENT = 1
REGION_RATING_CLIENT = 2
REGION_RATING_CLIENT = 3
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE
TARGET
Payment Difficulties
No Payment Difficulties

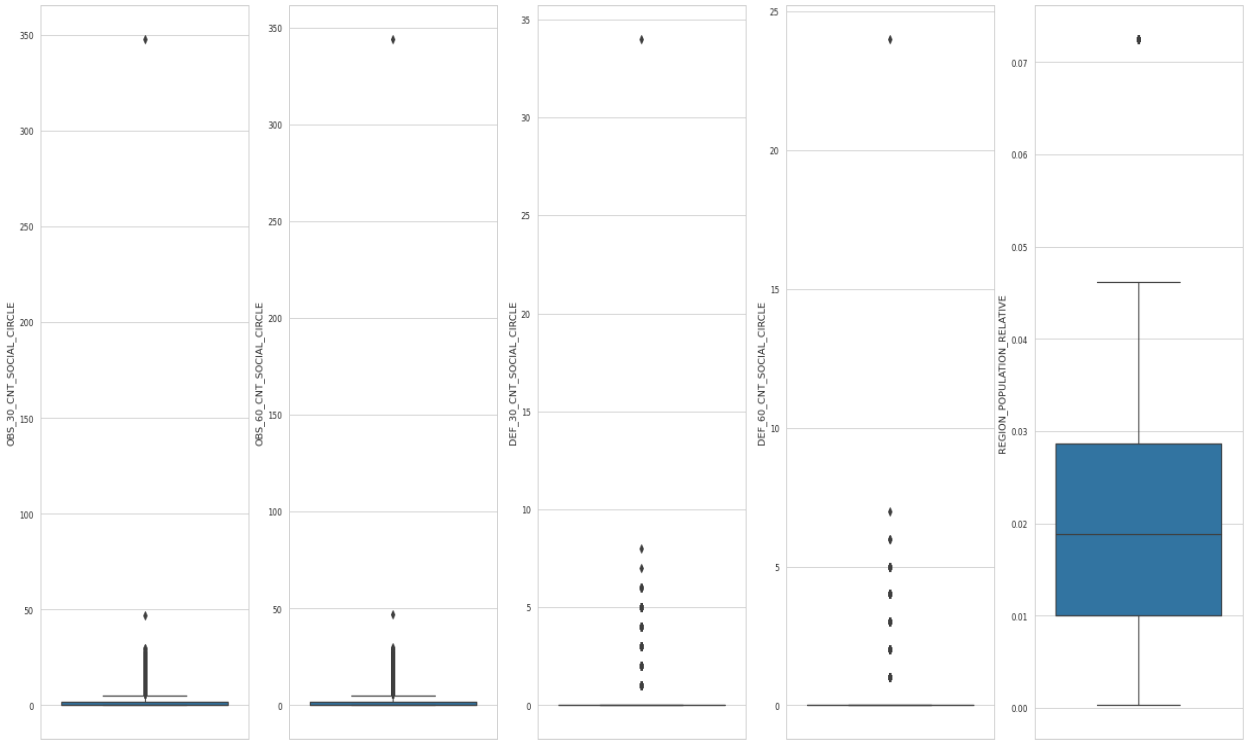

REGION_RATING_CLIENT = 1
REGION_RATING_CLIENT = 2
REGION_RATING_CLIENT = 3
1e6
AMT_CREDIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE
TARGET
Payment Difficulties
No Payment Difficulties

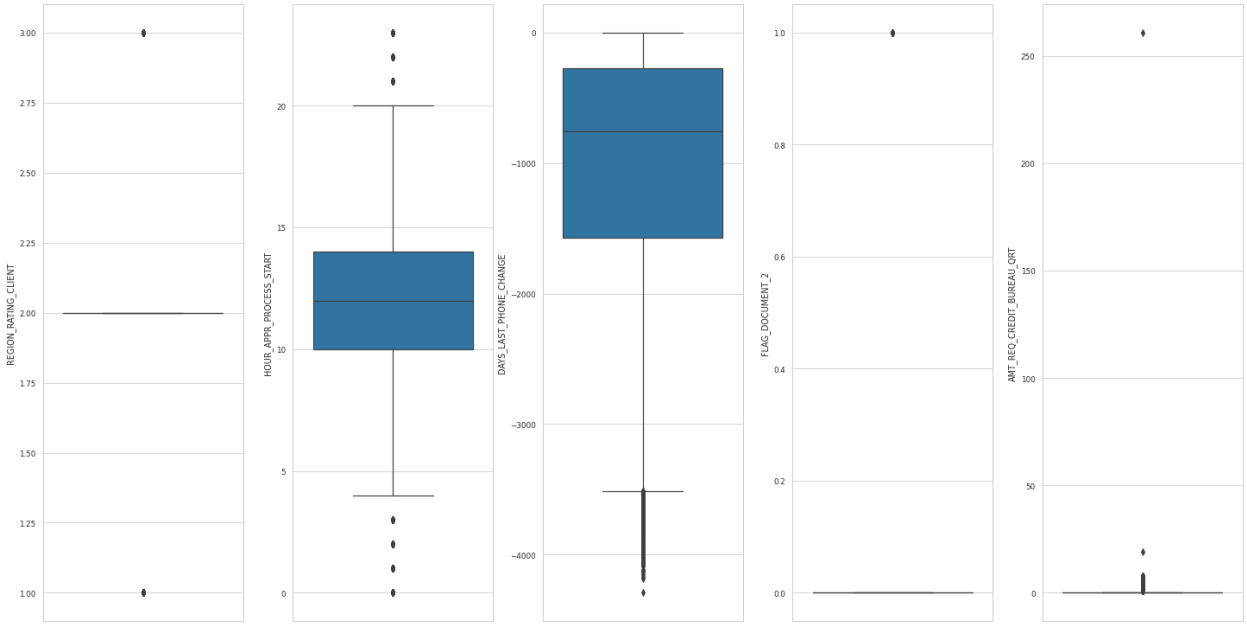
Detecting Outliers:



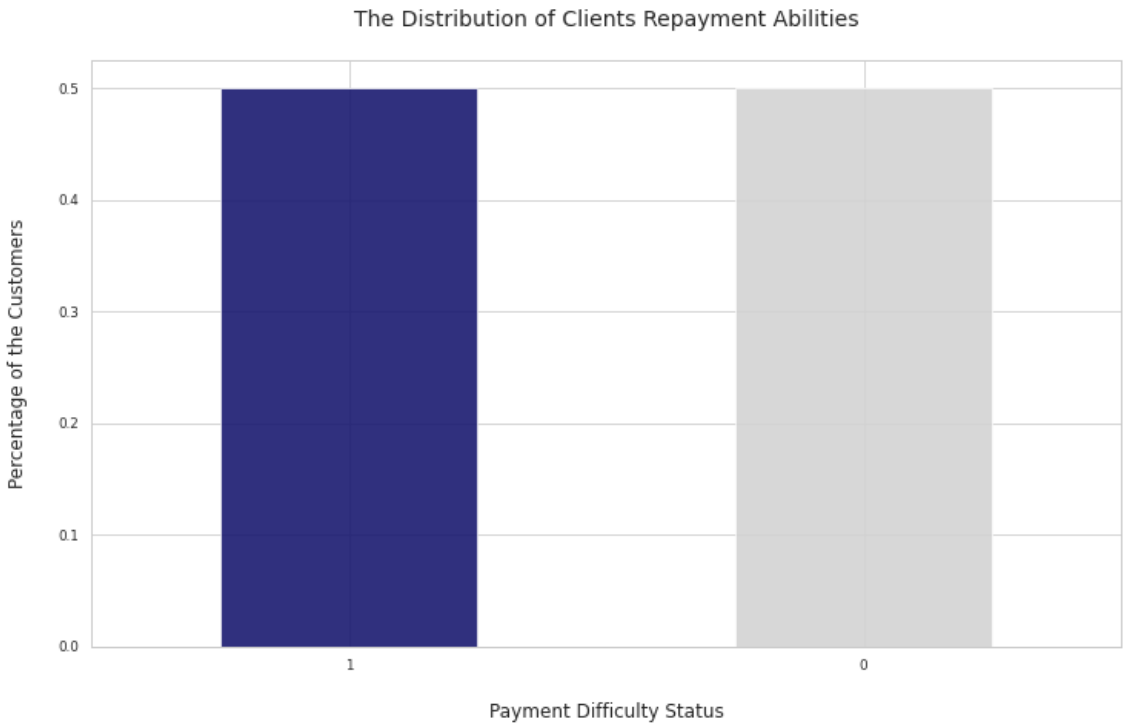



AMT_REQ_CREDIT_BUREAU_YEAR
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_HOUR


OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
REGION_POPULATION_RELATIVE



Handling Data Imbalance (Up sampling Techniques)



3.2 MACHINE LEARNING (ML) MODELS

The team uses predictive modelling technique whereby the outcome of the model is expected to identify the potential that someone will default on a loan (see Figure 5).

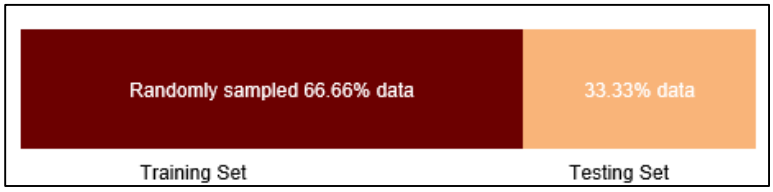
Figure 5: ML Model

The Expected Target Outcome will yield either 0 or 1; 0 refers to Not a defaulter, 1 refers to potential defaulter.

The key performance metric to evaluate the usefulness of the ML model is accuracy. The Training and Testing datasets were subjected to the same feature engineering to evaluate the model.

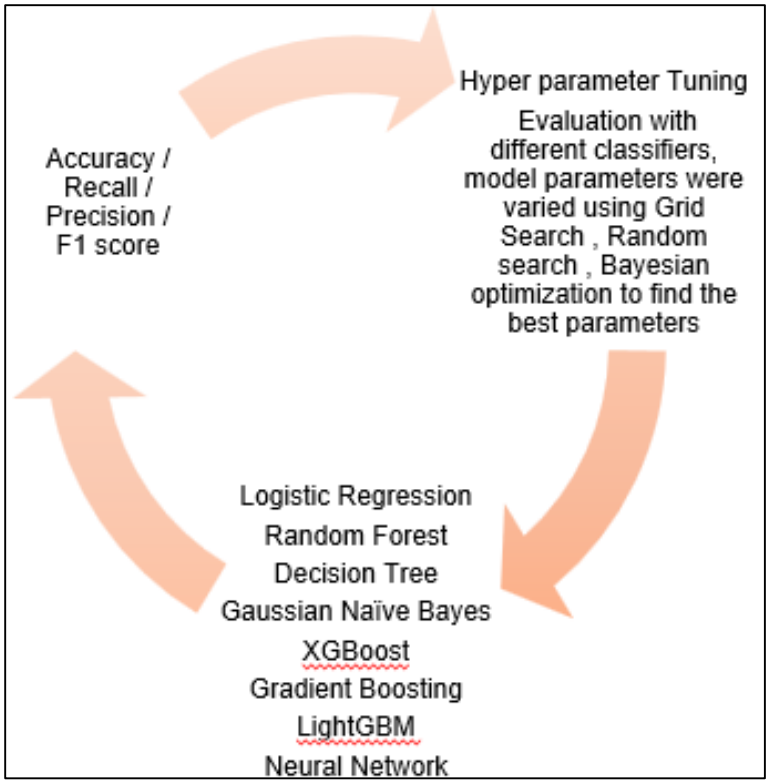
Out of the main training data set, a certain percentage is kept untrained to test the model's performance. Training set and validation set are split in following percentages: 66.66%: 33.33%. On the testing set, the target labels are hidden, until the performance is evaluated (see Figure 6).

Figure 6: Training / Testing data set



The ML Models used in this Project are: Logistic regression, Random forest, XGBoost, LightGBM, Naïve bayes and ensemble (see Figure 7)...

Figure 7: ML Techniques


Hyper parameter Tuning
Evaluation with different classifiers, model parameters were varied using Grid Search , Random search , Bayesian optimization to find the best parameters
Logistic Regression
Random Forest
Decision Tree
Gaussian Naïve Bayes
XGBoost
Gradient Boosting
LightGBM
Neural Network
Accuracy / Recall / Precision / F1 score

3.3 MACHINE LEARNING PREDICTION & RESULTS

Table 4 shows the accuracy of various ML models. The prediction accuracy of the training and testing data in Random Forest model has a value that is not much different and hence demonstrating that the model is the best among all models; there is no underfitting or overfitting. The confusion matrix for various ML models is shown in Figures 8 – 13.

Table 4: ML results

	Models	Training Accuracy Score	Testing Accuracy Score	ROC Score
0	Random Forest	1.000000	0.996500	0.996500
1	Decision Tree	1.000000	0.882600	0.882600
2	K-Nearest Neighbor	0.915600	0.880700	0.880600
3	Neural Network	0.700100	0.694800	0.694800
4	Logistic Regression	0.671600	0.672900	0.672900
5	Gaussian Naive Bayes	0.602400	0.603900	0.604000

Figure 8*: Confusion matrix (Logistic Regression)*

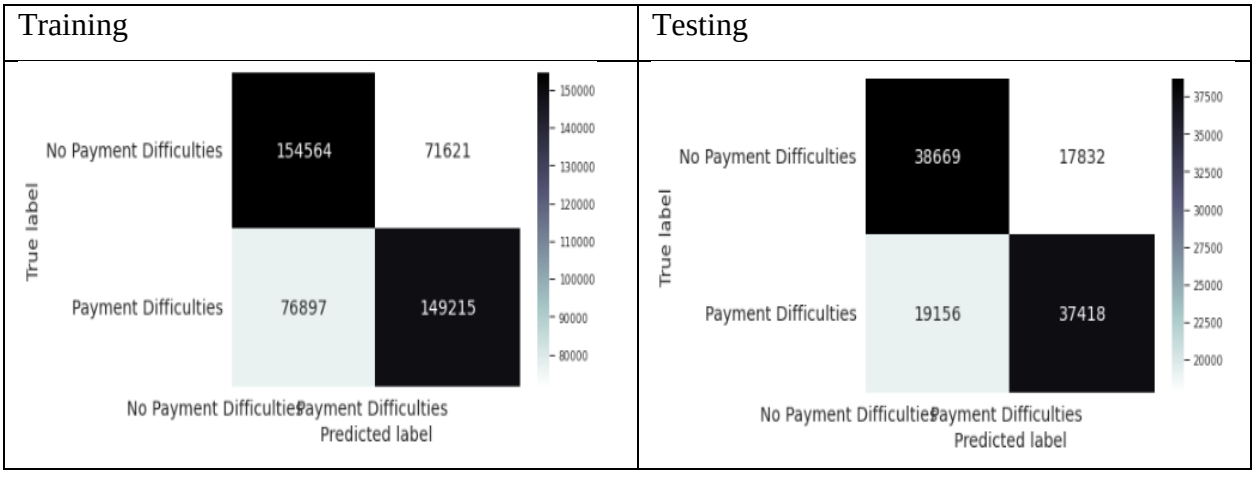


Figure 9*: Confusion matrix (Gaussian Naïve Bayes)*

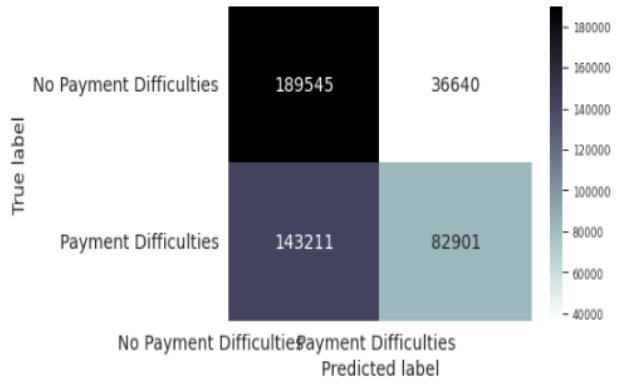
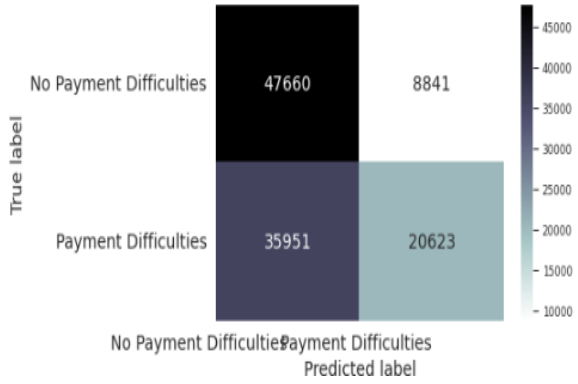
Training	Testing
	

Figure 10*: Confusion matrix (Decision Tree Classifier)*

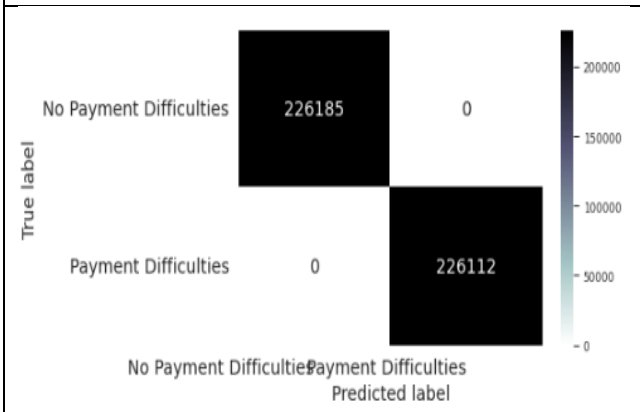
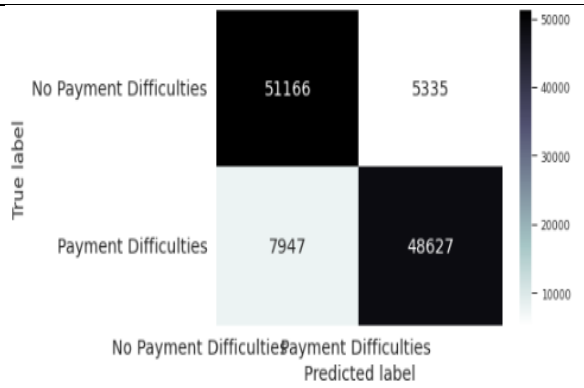
Training	Testing
	

Figure 11*: Confusion matrix (Random Forest)*

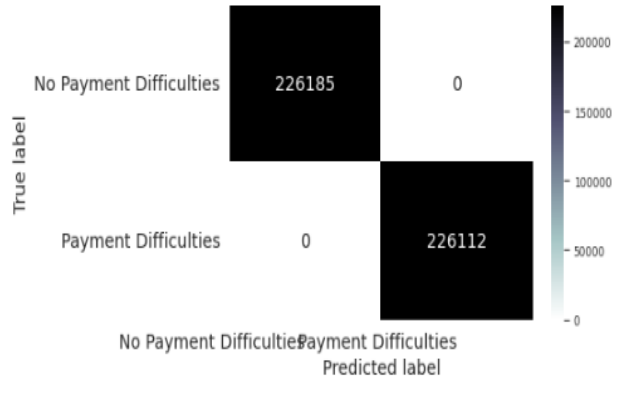
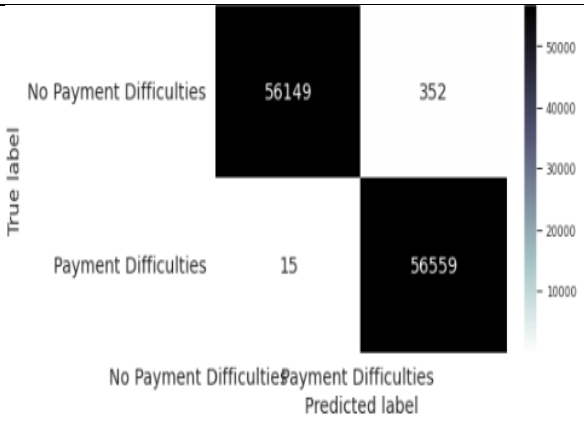
Training	Testing
	

Figure 12*: Confusion matrix (KNN)*

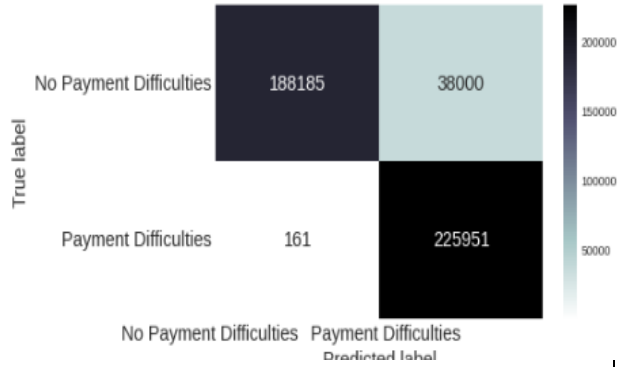
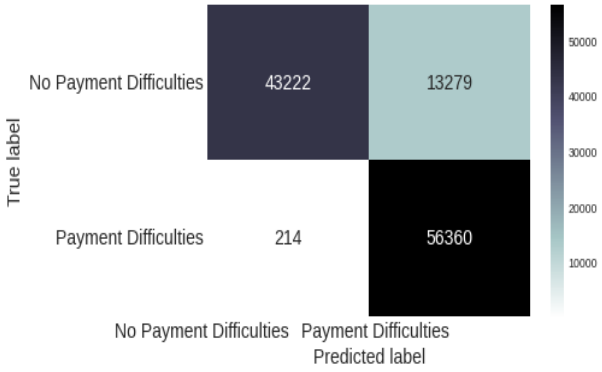
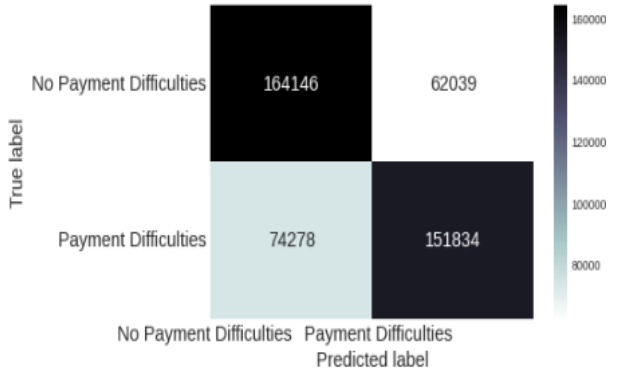
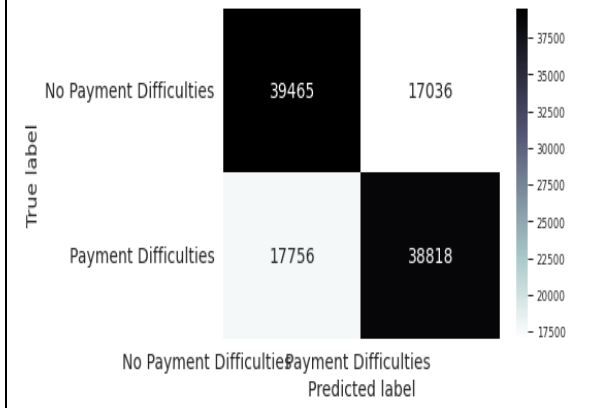
Training	Testing
	

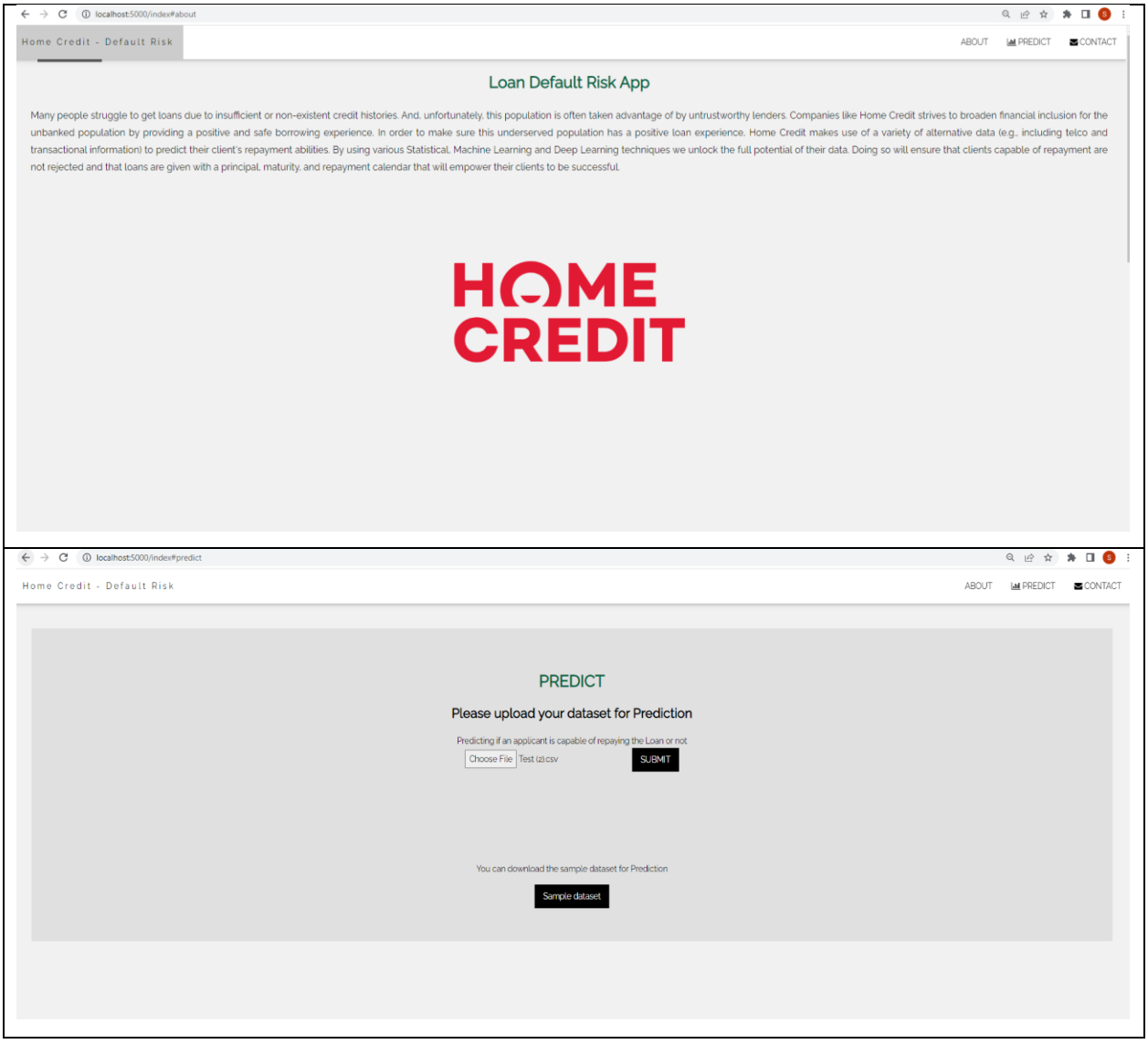
Figure 13*: Confusion matrix (Neural Network)*

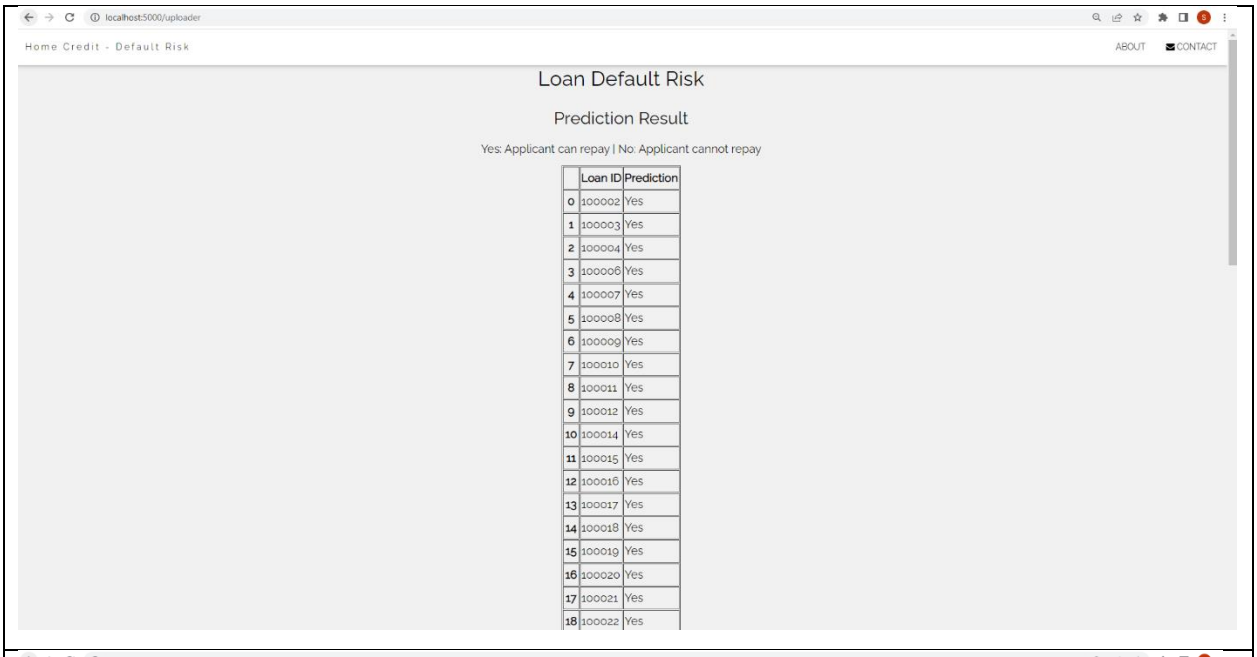
Training	Testing
	

3.4 WEB APPLICATION

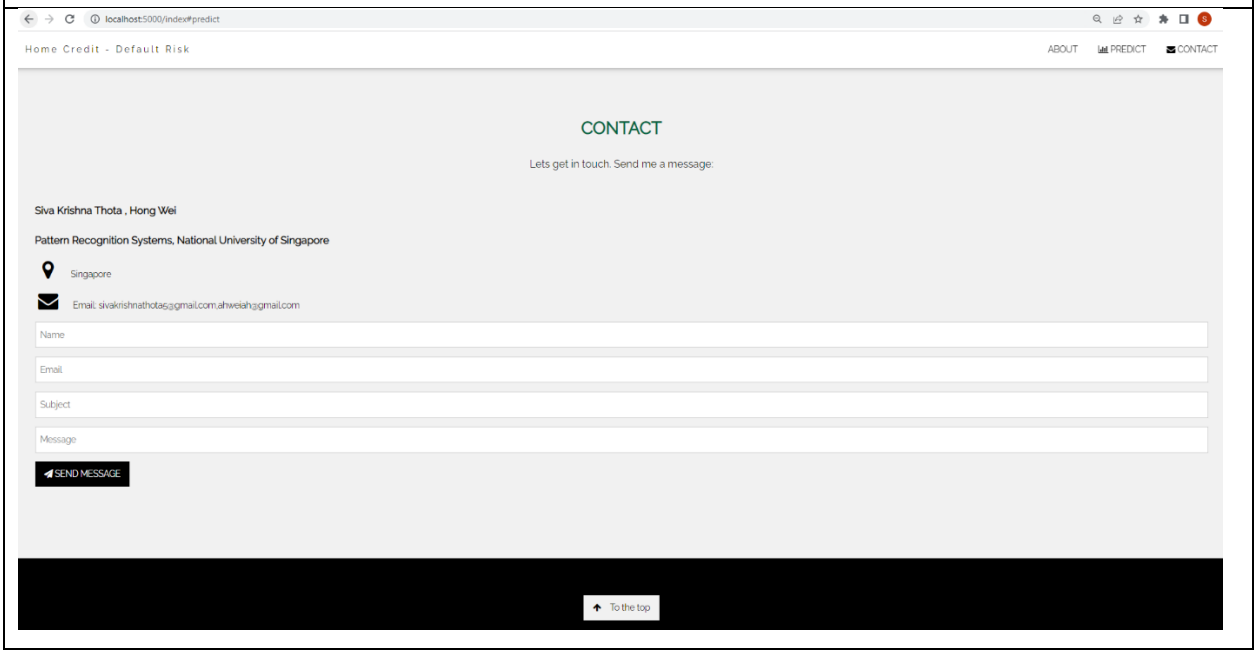
The Project team has also developed a Web Application using Python Flask, and this allows users to upload dataset for prediction, whereby prediction result = 'Yes' means that the borrower can repay the loan while prediction result = 'No' means that the borrower cannot repay the loan. Figure 14 demonstrates the usability and prediction functionality of the Home Credit Default Risk System (HCDRS). For testing the live system/demonstration, please refer to http://68.178.202.122:5000/login

Figure 14*: Web Application*





	Loan ID	Prediction
0	100002	Yes
1	100003	Yes
2	100004	Yes
3	100006	Yes
4	100007	Yes
5	100008	Yes
6	100009	Yes
7	100010	Yes
8	100011	Yes
9	100012	Yes
10	100014	Yes
11	100015	Yes
12	100016	Yes
13	100017	Yes
14	100018	Yes
15	100019	Yes
16	100020	Yes
17	100021	Yes
18	100022	Yes



4.0 LIMITATIONS

There are some possible limitations which could limit the effectiveness of the Project.

Firstly, the model's accuracy could be impacted if the source/raw data is not accurate, i.e "rubbish in rubbish out".

Secondly, the solution provided by this Project can only be implemented by personnel equipped with the relevant technical/data know-how. A company or staff with no basic understanding of Machine Learning techniques would not be able to appreciate the solution nor implement the solution.

5.0 CONCLUSION

In conclusion, the team has successfully created a Home Credit Default Risk System (HCDRS) with the following objectives in mind:
1) Reduce uncertainty from the home loan provider perspective
2) Reduce risk of lending to a client with high risk of default and
3) Ensure home loan can be provided to segment such as the unbanked population or people with insufficient or non-existent credit history

The team has achieved the following value-add in the course of the project.

- Performed data wrangling / cleaning, setting up the data for analysis and model building
- Dealt with data having anomalies
- Added Interaction variables
- Performed hyper parameters optimization
- Incorporated Domain Feature engineering
- Performed Exploratory Data Analysis
- Discovered patterns in data
- Built bagging based ensemble model

6.0 IMPROVEMENTS

Given additional time/resources, the Project can be further improved in the following areas: additional scope and platform delivery.

The scope of HCDRS can be expanded by overlaying the generated solution with additional business rules. For instance, if a loan applicant is working in a 'risky' function (e.g. sales) or risky industry (e.g. Banking during a period of Financial Crisis), the default rate of the applicant would be set higher than what the HCDRS has proposed.

The platform delivery of HCDRS can be expanded to include mobile app. If a user can obtain the prediction result via his mobile phone, it will be hugely beneficial as the user can have access to the prediction result anytime, anywhere and at the tip of his finger.

7.0 BIBLIOGRAPHY

Home Credit Default Risk Competition (2018). Kaggle. https://www.kaggle.com/c/home-creditdefault-risk/overview

Bagherpour, A. (2017). Predicting mortgage loan default with machine learning methods. University of California/Riverside.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machinelearning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24-39.

He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons

Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014, September). Optimal thresholding of classifiers to maximize F1 measure. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 225-239). Springer, Berlin, Heidelberg.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In ELPUB (pp. 87-90).

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

Poulos, J., & Valle, R. (2018). Missing data imputation for supervised learning. Applied Artificial Intelligence, 32(2), 186-196.

Kanter, J. M., & Veeramachaneni, K. (2015, October). Deep feature synthesis: Towards automating data science endeavors. In 2015 IEEE international conference on data science and advanced analytics (DSAA) (pp. 1-10). IEEE.

Bellman, R. (1966). Dynamic programming. Science, 153(3731), 34–37.

Wright, R. E. (1995). Logistic regression.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. R news, 2(3), 18-22.

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

Zhang H. (2004). The optimality of Naive Bayes. Proc. FLAIRS.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. Update, 1(1), 2007.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Bergstra, J., Yamins, D., & Cox, D. (2013, February). Making a science of model search: Hyper parameter optimization in hundreds of dimensions for vision architectures. In International conference on machine learning (pp. 115-123).

A Gentle Introduction. (2018). Kaggle. https://www.kaggle.com/willkoehrsen/start-here-a-gentleintroduction

Introduction to Automated Feature Engineering. (2018). Kaggle. https://www.kaggle.com/willkoehrsen/automated-feature-engineering-basics