**Loan Default Analysis**

Group 2

# **Member Contributions**

Ahmad As-Shodiqqul Amin B M T

* Slides, Report and Presentation

Dione Lim Yee Sze

* Exploratory Data Analysis

Kellie Chin Shu Wen

* Exploratory Data Analysis

Keith Tay Xiang Rui

* Data Processing and Modelling

Lua Jun An

* Data Processing and Modelling

Ni Hui Ling

* Data Processing and Modelling

Timothy Wong Hoey Pheen

* Slides, Report and Presentation

Tu Zhehao

* Data Processing, Modelling and Presentation

# **Table of Contents**

[**1. Context**](#_id6tiyl8ne68) **4**

[Defining Loan Default](#_x2o7ylg64q2w) 4

[Relationship between Loans and Commercial Banks](#_8opfcxcz3fpx) 4

[Relationship between Banks and Client](#_u158knn1n27l) 4

[**2. Objective**](#_oyksclnxrvok) **5**

[**3. Dataset**](#_pkdep6k105hr) **6**

[Cleaning Dataset and Feature Selection](#_3nfj2q7bornu) 6

[**4. Exploratory Data Analysis**](#_rbp3x7sghz9) **8**

[Insights](#_o9q1whna7dta) 8

[Remarks on Insights](#_a2wsn0pmlbec) 11

[**5. Feature Engineering**](#_jtuqe0cflikb) **12**

[Age\_Category](#_k751ql2xdh0i) 12

[Employment](#_2w0voqh62ihi) 12

[Credit\_Category](#_5al9u3375cqw) 12

[Credit Score](#_7nsd37hpdwy7) 12

[**6. Model Building**](#_p0bdmc7yrtvb) **14**

[Logistic Regression (LR)](#_ntdhijsz7ag2) 14

[Random Forest Classifier (RFC)](#_f38j4rf18nqe) 14

[XGBoost (XGB)](#_49grb0jfa17p) 15

[**7. Our Approach**](#_obr8j3f3b8j7) **17**

[Greater approval rate by lowering loan amount given to clients](#_1c893ya36wuu) 17

[Providing better repayment process for younger adults](#_d2hacumo75ey) 18

[**8. Evaluation**](#_tvgvf0l40yd9) **19**

[**9. Conclusion**](#_auptqyi8uvi) **20**

[**References**](#_k1wjs7yy4zgr) **21**

# **1. Context**

# **Defining Loan Default**

Before elaborating on the problem and approach, it is necessary to define what our topic of discussion - “Loan Default” - is. Loan default can be broken down into two words, ‘Loan’ and ‘Default’. The former means to borrow money while the latter can be denoted as the failure to fulfill an obligation. Hence, putting them together, loan default indicates the inability of an entity to repay the loan that it has undertaken. If loan defaults remain unpaid for a long period of time, they will be classified as Non-Performing Loans (NPLs).

## Relationship between Loans and Commercial Banks

For most commercial banks, loans are a major asset and source of income in the form of interest income. An accumulation of NPLs indicates a reduction in the bank’s liquidity and profits, making them unattractive to potential investors (Ojong Opa & Tabe-Ebob, 2019). As such, the decision to approve or reject a loan is extremely critical to commercial banks’ profitability and investor confidence.

## Relationship between Banks and Client

Loans mainly involve two key stakeholders - the bank and the client. To determine whether a loan will be defaulted, we require information on these two key stakeholders. Historically, loan transactions have been approved or rejected based on clients’ past information, including but not limited to family history and credit history.

There are two main decisions a bank can make when a client seeks approval for a bank loan and that is either approval or rejection. Likewise, the client can also cancel during the approval process of the loan. These actions by the client would denote that the bank has failed in its loan transaction.

**2. Objective**

Keeping the aforementioned drawbacks of loan defaults for banks in mind, and setting our clientele to be banks, our group’s objectives then are as follows:

1. To reject the individuals who are at very high risks of defaulting
2. To ensure that individuals capable of repaying are given the loan
3. To scale down the loans given to individuals at higher default risk to reduce potential losses while enabling potential gains

In short, banks consider an individual who was given the loan yet defaulted as a *false* *positive* whereas an individual who was not given the loan yet could have repaid on time is considered as a *false* *negative*.

In both cases, since banks missed out on an opportunity to profit, they are interested in **reducing the amount of *false* *positives* and *false* *negatives***.

# 

# **3. Dataset**

To learn from past mistakes (incorrect decisions by banks to approve or reject loans), we tapped on the vast collection of historical data on past loans approved by the bank. This includes the client’s personal information such as gender, income, owned assets and whether they defaulted or paid their loans on time.

## Cleaning Dataset and Feature Selection

At this point, we noticed that the dataset we have chosen to use was not cleaned beforehand as

there were a number of irrelevant columns: columns that lacked any link or have such a tenuous link to our target variable.

For example, most clients had not submitted flag documents of 4 to 21 as seen in *Fig. 1*. This skewness in data means FLAG\_DOCUMENT\_X would not be good predictors as they may not be representative of the true population. To avoid potential noise, we removed these variables.

|  |
| --- |
| Fig. 1. Low submission rate for FLAG\_DOCUMENT columns |

We further refined feature selection by removing features with high multicollinearity. We based this removal by using the following multicollinearity metrics:

* Point biserial correlation, Pearson correlation, Cramer’s V correlation

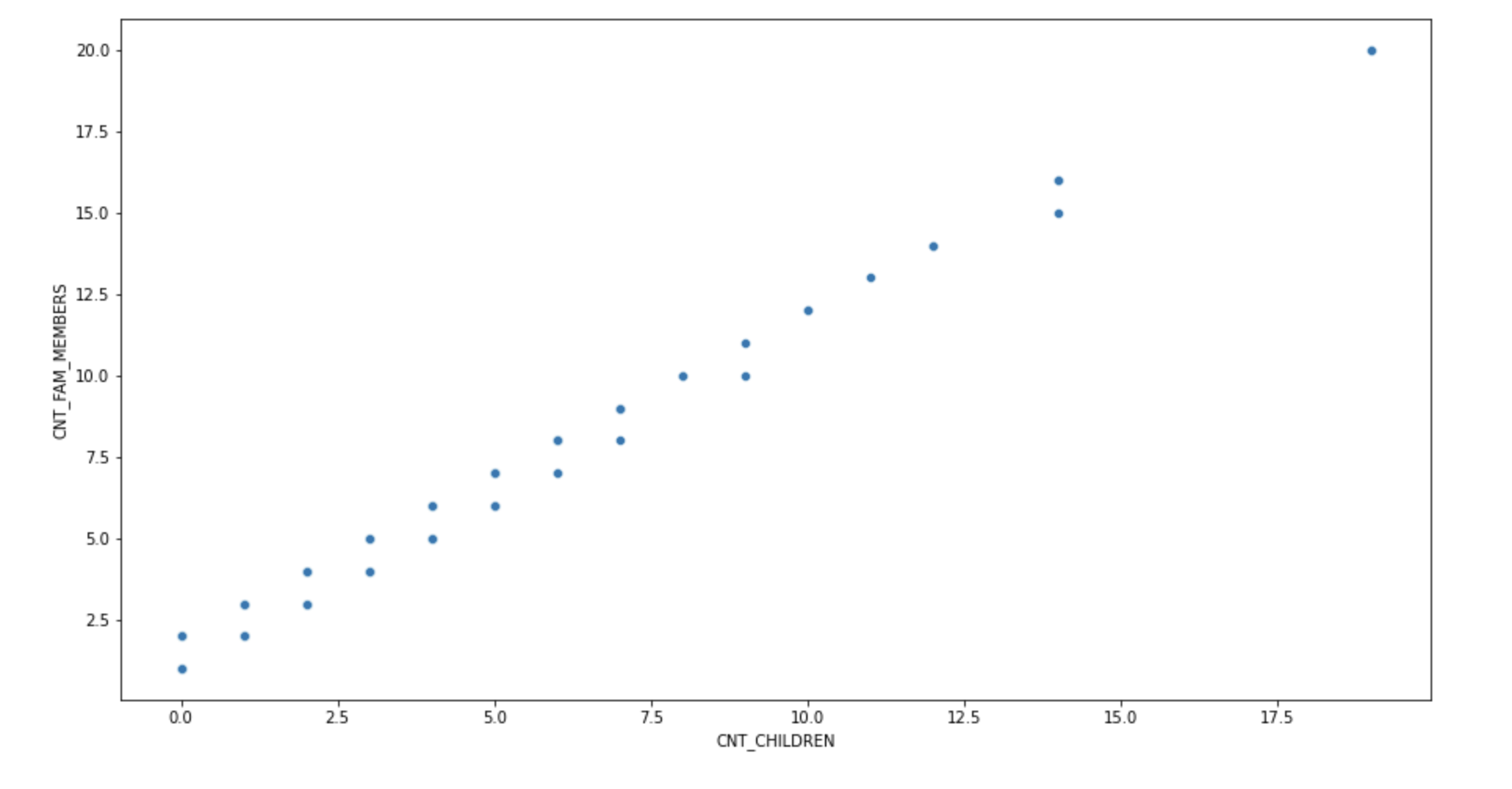


Fig. 2. Example of multicollinearity between two variables (CNT\_CHILDREN and CNT\_FAM\_MEMBERS)

Additionally, there were some normalized values for some of our columns. We removed these columns as they had undergone scaling prior to our examination and we are unable to determine exactly what kind of scaling has been done to these variables.

Lastly, the dataset had columns with missing entries. We decided to remove columns which were more than 25% empty to preserve the distribution of our data. For the remaining columns, we sampled the dataset using a normal distribution since our dataset is large and we can apply the Central Limit Theorem. We then padded the empty cells using the sampled values, making sure to round the obtained float values to the nearest integer for columns in integer denominations.

# **4. Exploratory Data Analysis**

## Insights

After cleaning the dataset, we implemented several data analysing techniques and discovered seven of the following insights, depicted below correspondingly by Figures 2 to 8.

**1. Credit Amount has no significant effect on whether an individual will default**

|  | Fig. 3.  Boxplot of Credit Value against Target variable |
| --- | --- |

Median value for the credit amount is the same regardless whether the target is met. Also, the interquartile ranges for both boxplots are similar in size, signifying a similar distribution.

**2. Increasing number of children indicates more difficulty in returning the loan**

|  | Fig. 4.  Barplot of Target variable against Number of Children |
| --- | --- |

Visually, there is a clear increasing exponential trend of the target variable against the number of children, with the exception of 5 children, which is an outlier.

**3. Widows have the lowest likelihood of being unable to pay**

|  | Fig. 5.  Barplot of Target variable against Family Status |
| --- | --- |

The barplot for the target variable of the family status for ‘Widow’ can be deduced to be markedly lower than that of other family statuses, even after accounting for variance.

**4. There is a positive correlation between age and being able to repay loans**

|  | Fig. 6.  Barplot of Target variable against Age |
| --- | --- |

There is a clear decreasing trend of the target variable as the age variable increases; the older you are, the more able you are to pay for your existing loans

**5. Revolving loans have a lower default rate as compared to cash loans**

|  | Fig. 7.  Barplot of Target variable against Contract Type |
| --- | --- |

Revolving loans have a noticeable decrease in the target variable as compared to cash loans.

**6. Males find it harder to pay for their loans in general**

|  | Fig. 8.  Barplot of Target variable against Gender |
| --- | --- |

A marked increase in the target variables can be seen in males as compared to females.

**7. Highest proportion of defaulters are those that have loans of 300k - 600k**

|  | Fig. 9.  Barplot of Target variable against Credit Category |
| --- | --- |

## The barplot for the ranges between 300k to 600k have the largest proportion.

## Remarks on Insights

Amongst these 7 insights, two in particular stood out for our group - the ‘positive correlation between age and ability to repay loans’ (Fig. 5) and the ‘highest proportion of defaulters are those that have loans of 300k to 600k’ (Fig. 7). We shall incorporate these insights in our approach later on.

# 

# 

# 

# **5. Feature Engineering**

Based on the state of the data after the merging and feature selection process, we were still unable to devise an appropriate model as certain attributes were not good representatives of the dataset. Feature engineering was applied in order to acquire better accuracy and insights when fitting our models.

## Age\_Category

In the original dataset, the entries were denominated in days and set to negative values. Hence, we converted the original DAYS\_BIRTH attribute into an age category attribute for readability as well as to study the default rates within each category.

## Employment

The entries were denoted in a similar manner to the age\_category attribute. Furthermore, when converted to years, some entries indicated that applicants had been unemployed for a 1000 years, which was not possible. For these entries, we assumed that they have retired and so, we set their years employed to 0.

To ease readability and modelling, the values were also categorised into various employment categories i.e. “Unemployed”, “10 or Less”, “10-20” and more.

## Credit\_Category

Since the range of AMT\_CREDIT was large and difficult to use, we decided to bin them into several categories with intervals of 100k i.e. ‘100k - 200k’, ‘200k - 300k’, ‘> 1 million' and more.

## Credit Score

The dataset did not include any information about interest rates. Hence, we created a Credit Score attribute which was based on each client’s record, devised in the following manner:

To sieve out the important features, we pick attributes which have an information value of more than 0.02. After which, we tabulate the weight of evidence for each attribute and scale the credit score accordingly. The final credit score would then be a sum of all the scores across all attributes.

Methodology:<https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03>

We then made use of the final credit score to determine the interest rate that was applied to them, assuming a lower bound of 5% and an upper bound of 15%. Next, we scaled the interest rate based on their credit scores and applied them to better quantify the gains and losses of the company.

# 

# **6. Model Building**

We fed the cleaned and engineered dataset into 3 different models and compared our results.

**Logistic Regression (LR)**

Logistic regression is a machine learning classification algorithm that is used to predict the probability of a categorical dependent variable. Since we were trying to predict a binary output [0,1], we decided to try out the Logistic Regression model.

Using the logistic regression model, we obtained the following results for these metrics:

| F1 | 66.3% |
| --- | --- |
| Recall | 56.7% |
| Precision | 86.8% |
| Accuracy | 56.7% |

As seen from above, out of the four scores, the logistic regression model has poor scores in three of them - f1, recall and accuracy. This might be due to the large number of categorical variables in our dataset. Hence, we have decided to adopt a tree-based model.

## Random Forest Classifier (RFC)

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Using the classifier for our model, we obtained the following results for these metrics:

| F1 | 74.6% |
| --- | --- |
| Recall | 67.4% |
| Precision | 86.1% |
| Accuracy | 67.4% |

As compared to the logistic regression model, the RFC performs better. All the scores have improved with the exception of the precision score though, it is still above 80%.

## XGBoost (XGB)

The XGBoost algorithm is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Errors are minimized using gradient descent algorithm to an extreme degree, hence, the term ‘gradient boosting’. Using this algorithm, we obtained the following results when ran on a cross-validated set:

| F1 | 74.6% |
| --- | --- |
| Recall | 77.6% |
| Precision | 71.8% |
| Accuracy | 73.5% |

Having compared the performances between the three models, we chose to go along with XGB for our final choice of model as it had the best overall performance across all the metrics.

We went on to further fine-tune the XGB model by finding the optimal depth of the tree. By stopping the tree early, we reduce the possibility of overfitting as well as training time. This allows the model to perform better on unseen testing data.

The process of finding the optimal depth was done iteratively and the model’s accuracy was plotted against the depth of the tree.

|  | Fig. 10.  Plot of Accuracy against MAX\_DEPTH  (training data) |
| --- | --- |

As seen from Figure 9, any increase from a maximum depth of 20 led to minimal improvements in the overall accuracy of the model.

Therefore, a maximum depth of 20 was chosen. Any further increase of the depth will result in greater model complexity and can potentially result in higher variance when used on unseen testing data.

# 

# **7. Our Approach**

Based on our model, we have chosen 2 sets of recommendations to suit the client.

## Greater approval rate by lowering loan amount given to clients

Previously, if our model predicts that our client will default, their loan application will be rejected. However, we shall adopt a more liberal approach when deciding on their loan application. In essence, instead of rejecting clients, we will grant them approval by lowering their loan amount by 20%.

We shall implement this to the credit groups with the highest proportion of defaulters. From Insight 7, we know that the highest proportion of defaulters are from those with loans from $300,000 to $600,000, hence we will be targeting this group of defaulters with this approach.

|  |
| --- |
| Fig. 11. Reduction in Number of Defaulters against Reduction from Original Loan |

From the plot above, we noticed that the reduction in both loss and number of defaulters increases as we decrease the loan amount given to clients.

However, considering the bank’s interests, we intend to match the loan amount requested by the client as closely as possible since it is unlikely a person would take up a loan if the loan provided is only 50% or lesser than what he or she initially requested.

Furthermore, in our examination earlier, we observed that with a reduction of the loan from 20 to 30 percent, the slope of both plots became less steep than before. This indicates that further reductions in loan amount by increments of 10% will only result in correspondingly smaller decrements in the number of defaulters.

Hence, in an attempt to minimise the number of defaulters while maximising interest income, we will **reduce the loans of the predicted defaulters by 20%**. In other words, we will only grant them 80% of the loan amount they initially requested.

## Providing better repayment process for younger adults

Let us recall Insight 4 under our ‘Exploratory Data Analysis (EDA)’ section; there is a positive correlation between age and ability to repay a loan. This means a high default rate exists amongst younger adults since they have just entered the workforce and are more likely to face difficulty in repaying their loans. This means that, based on our model, we would have rejected their loan approval.

Nevertheless, considering the banks’ interests, we aim to maximise interest income. That is, we are keen to increase loans granted to younger adults. However, to ensure that they would have the means to repay their loans, we aim to devise the following for this specific age category:

1. extend payment period
2. impose a scaled interest on this extended period with reference to computed credit score

By extending their payment period, the younger adults will be given more time to enter the workforce and will be able to stabilise their finances. By then, they will be in a better position to repay their loans and avoid defaulting. At the same time, as they are being imposed a scaled interest rate during this extended period, banks will still be able to earn their interest income.

# **8. Evaluation**

**Approach 1**

|  |
| --- |
| Fig. 12. Reduction in Loss against Reduction from Original Loan |

Through this strategy, as seen in Figure 10, we have discovered that this **lowers the potential losses by $762,597,495.00 (or 18.23%)**. This is because clients, albeit the lower loan amount granted, will still have to pay interest. As such, the bank still gains a semblance of revenue. This is favoured rather than completely losing potential income by rejecting the clients upfront.

**Approach 2**

Based on our model, after increasing their loan period and accounting for interest compounded between 5% to 15%, we expect an **increase of $393,106,511.23 (or 8.03%) in interest income** gained from customers below the age of 25.

# **9. Conclusion**

To recap, we had proposed that our client retain the possible defaulters in two ways - firstly by lowering the loan amount provided and secondly, extending the repayment period for younger people.

Our proposed solution seems counterintuitive as it would appear that we are not maximising our potential interest income by decreasing loan amounts. However, by adopting the measures we have recommended, banks would take a calculated risk that is minimised as much as possible. This risk grants them the ability to avoid losing potential profits while simultaneously being able to minimize losses in the event of a default. In the long run, such a solution would ultimately allow the bank to have slightly higher overall profits, which is to their interests.

As data scientists and advisors to our client, we can only provide recommendations and ultimately, the choice of whether to perform the measures themselves is up to our client. Nevertheless, as can be seen from our report, we have done our utmost diligence by ensuring that our client, the bank, is able to make a wise and informed decision.

# **References**

1. Ojong Opa, Valentine & Tabe-Ebob, Wendy. (2019). The Effects of Loan Default on Commercial Bank Profitability in Cameroon, Case Study BICEC Limbe.
2. XGBoost Algorithm: Long May She Reign! (2019). Towards Data Science

URL:<https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>

1. Kinden Property. (2019). Intro to Credit Scorecard: Step by Step Guide on How to build a simple Credit Scorecard.

URL: <https://towardsdatascience.com/intro-to-credit-scorecard-9afeaaa3725f>