Blue text on a white background

Description automatically generated

**IS453 – Financial Analytics**

**AY 2023/2024 Term 1**

**Section G2**

**Final Project Report**

**Credit Scorecard for Singles Housing Loan**

**Prepared for**: Professor Duran Randall

**Prepared by:** Group 3

|  |  |
| --- | --- |
| **Student Name** | **Student ID** |
| Edwin Chong Tong Tong | 01390827 |
| Lim Zhuo Yuan Elizabeth | 01409732 |
| Nur Hazlinda D/O Haja Maideen | 01395561 |
| Sneha Murali | 01410645 |
| Wei Zihao | 01366624 |

Table of Contents

[1. Introduction 1](#_Toc151125811)

[1.1 Business Opportunity Analysis 1](#_Toc151125812)

[1.2 Business Context and Target Customers 1](#_Toc151125813)

[1.3 Information from the Dataset 1](#_Toc151125814)

[2. Methodologies 1](#_Toc151125815)

[2.1 Pre-processing for Application data 1](#_Toc151125816)

[2.2 Pre-processing for Bureau Data 2](#_Toc151125817)

[2.3 Explorative Data Analysis (EDA) 2](#_Toc151125818)

[2.4 Data Preparation 2](#_Toc151125819)

[**2.4.1** **Missing Values** 2](#_Toc151125820)

[**2.4.2** **Outliers** 2](#_Toc151125821)

[**2.4.3** **Highly Correlated Variables** 3](#_Toc151125822)

[3. Analysis Approach and Insights 3](#_Toc151125823)

[3.1 Grouping & Screening 3](#_Toc151125824)

[3.2 Create Scorecard 4](#_Toc151125825)

[3.3 Analysis & Tuning 4](#_Toc151125826)

[4 Walkthrough 4](#_Toc151125827)

[5 Limitations of Analysis Approach and Opportunities for Further Improvements 5](#_Toc151125828)

[6. Challenges Encountered and Lessons Learned from Analysis 5](#_Toc151125829)

[References 6](#_Toc151125830)

[Appendix I](#_Toc151125831)

[Appendix 1: Final Credit Scorecard I](#_Toc151125832)

[Appendix 2: Model Performance II](#_Toc151125833)

# **1. Introduction**

## **1.1 Business Opportunity Analysis**

Singapore is experiencing a shift in demographics, with a rising number of individuals choosing to stay single or delaying marriage. This trend was notably prominent among those aged 25 to 29, as the proportion of singles rose 7% for single men and 15% for women (Ho, 2021). This continued trend over the past 10 years has led to a significant increase in the single demographic. However, many unmarried citizens still live with their parents, primarily due to the prohibitively high property prices in Singapore. This phenomenon is quite specific to Singapore, and not as observable overseas.

In this land-scarce nation, property prices consistently appreciate, making private properties unaffordable for many single citizens. Government housing schemes like Build-to-Order (BTO) offer a more affordable option for Singaporeans (PropertyGuru Editorial Team, 2023). As they were initially designed for married couples, they pose challenges for singles. Government subsidies and loans for BTO flats are more attractive than private banks but lack appeal for singles due to limited ballots and strict criteria (Yuit, 2022).

Despite recent enhancements in the CPF Housing Grant Scheme for singles, limitations persisted until 2019. However, as of 2022, government bodies such as the Housing Development Board (HDB) have expanded the scheme, allowing singles greater flexibility in choosing housing options (CNA, 2023). As such, more singles will have quotas for BTOs and do not have any restriction in purchasing resale flats. This will increase the number of singles purchasing public housing, and in turn increase demand for housing loans.

## **1.2 Business Context and Target Customers**

To tap on this increased demand, and the lack of specific loans available for singles in Singapore, our team believes that it is a good opportunity for our client, the Fintech lender, to provide housing loans to this untapped segment.

As HDBs are only available to singles above the age of 35, the target segment we are recommending would be singles between the age of 35 – 60 with no current property ownership in Singapore. Just from a business standpoint, we believe that they will likely have a higher income and more savings as they do not need to provide for children (assuming most singles do not have kids). As such, less likely to default as well.

## **1.3 Information from the Dataset**

The information in the dataset provided is separated into two datasheets: the Application Data, and the Bureau Data. The Application Data has 307,511 observations and 122 columns, mainly providing demographic information of the applicants, and the Bureau Data has 1,716,428 observations and 17 columns, providing information on each applicant’s credit history and financial behaviours.

# **2. Methodologies**

## **2.1 Pre-processing for Application data**

We conducted pre-processing on the application data, filtering rows that align with our target customer segment, namely singles without property ownership. So, we refined our dataset to only the singles from the ‘NAME\_FAMILY\_STATUS’ column and those without property for the ‘FLAG\_OWN\_REALTY’ column. Subsequently, we selected variables which we believed were relevant based on our business judgment, resulting in a total of 28 variables. To further refine our dataset, we applied age restrictions (35-60) in compliance with HDB’s rules. Additionally, for improved readability, certain variables such as ‘DAYS\_BIRTH’, ‘DAYS\_REGISTRATION’, and ‘DAYS\_EMPLOYMENT’ were transformed into years and months. The newly refined dataset was placed in a new data frame ‘df’. This dataframe has 6102 rows with 26 columns, including the ‘TARGET’ column.

## **2.2 Pre-processing for Bureau Data**

The transformation performed in the Bureau Data was mainly to combine the different credit lines for each individual customer, as one customer could have many different past credits. We used the variable ‘SK\_ID\_CURR’ to filter out unique applicants, and subsequently got the average or sum of the values of the characteristics for all the past credits under the same applicant. The team first created new variables which are the calculated sum of the following variables ‘AMT\_CREDIT\_SUM\_OVERDUE', ‘AMT\_CREDIT\_SUM\_DEBT’, ‘AMT\_ANNUITY; and created a new variable for the average of the ‘CNT\_CREDIT\_PROLONG’ variable. In this case, the dictionary was unclear on whether variables like ‘AMT\_CREDIT\_SUM\_OVERDUE’ referred to the amount overdue to that individual credit or the applicant as a whole, so for this project we assumed the variables refer to just that particular credit line. Afterwards, we added the transformed variables into ‘df’ using left join to combine all transformed variables into the new cleaned dataset. The new combined dataset had 6102 rows with 29 columns, including the ‘TARGET’ column.

## **2.3 Explorative Data Analysis (EDA)**

Starting with the univariate analysis, for the numerical variables in the dataset, we plotted histograms to understand the distribution of numerical variables; and a correlation heat map to identify variables that are highly correlated. For continuous variables, we used boxplots to identify outliers (which will be explained later in the section). Lastly, we plotted bar charts to understand the distribution of the categorical variables.

In the bivariate analysis, we plotted KDE plots for numerical variables and stacked bar charts for the categorical variables in the dataset. Both visualisation plots are for the purpose of understanding the distribution of the events and non-events for each variable.

## **2.4 Data Preparation**

### **2.4.1 Missing Values**

When performing the data preparation for the dataset, the team started by identifying the percentage of rows with more than 5 missing values and the number of missing values for each column. Upon discovering that 9.6% of rows in our refined dataset had over 5 missing values, we initially wanted to remove them. However, upon further investigation, we discovered that most of the missing values originate from columns such as ‘EXT\_SOURCE\_1’, ‘EXT\_SOURCE\_3’, ‘OCCUPATION\_TYPE’ and ‘SUM\_AMT\_ANNUITY’. As features such as ‘EXT\_SOURCE\_1’ refer to credit rating given by other credit agencies, we decided that most of these missing values were ‘missing not at random’. Furthermore, as the binning process takes into account missing values, we decided not to remove the rows with missing values. Consequently, we retained all the rows to ensure a larger dataset for analysis, and thereby yield more meaningful results.

### **2.4.2 Outliers**

From the boxplot analysis done earlier to identify the outliers in the dataset, we identified an invalid observation in the variable ‘MONTHS\_EMPLOYED’, which had an outlier value less than -12,000. This translates to the employee being employed for more than 1,000 years, which is impossible. Upon further investigation, we discovered that there were 692 rows in which ‘MONTHS\_EMPLOYED’ was -12,175. As such, the team concluded that the value is likely to be a default for missing values or another segment. In order not to distort our analysis, we replaced all those values with ‘NaN’ to represent a missing value.

### **2.4.3 Highly Correlated Variables**

As mentioned earlier, the team used a correlation heat map to identify variables that are highly correlated to each other. We conducted the correlation check with a benchmark of 0.75 and identified that the following variables were highly correlated:

1. ‘AMT\_CREDIT’ and ‘AMT\_GOODS\_PRICE’;
2. ‘LIVE\_CITY\_NOT\_WORK\_CITY’ and “REG\_CITY\_NOT\_WORK\_CITY’;
3. ‘AMT\_GOODS\_PRICE’ and ‘AMT\_ANNUITY’;
4. ‘AMT\_CREDIT’ and ‘AMT\_ANNUITY’.

To counter this, we decided to remove the highly correlated variables ‘LIVE\_CITY\_NOT\_WORK\_CITY’ and ‘AMT\_GOODS\_PRICE’ from the dataset to reduce redundancy. We removed these variables instead of their counterparts as they resulted in a lower Information Value (IV) when we did an initial WOE & IV calculation. After preparing the data, we were left with a total of 26 variables and 6102 rows.

# **3. Analysis Approach and Insights**

## **3.1 Grouping & Screening**

We then used *scorecardpy* to generate WOE and characteristic bins. As it automatically optimizes for maximum IV and performs One-Hot Encoding (OHE) as part of its binning process and groups attributes with similar WOE together, we did not do that as part of our data preparation to improve efficiency. We then sorted the variables in ascending order of IV. Most variables were in the weak predictive range of 0.02-0.1, and 6 variables had an IV lower than 0.02, which means that they were not useful for prediction. We also deemed them not important in terms of business context, as such, we decided to drop the following 6 variables:

‘REG\_CITY\_NOT\_WORK\_CITY’, ‘NAME\_HOUSING\_TYPE’, ‘CNT\_CHILDREN’, ‘REG\_CITY\_NOT\_LIVE\_CITY’, ‘CNT\_FAM\_MEMBERS’ & ‘DEF\_60\_CNT\_SOCIAL\_CIRCLE’.

Consequently, we plotted the WOE bins for all remaining variables. While *scorecardpy* optimizes for maximum IV, it disregards monotonicity. Hence, we analysed each variable to check for monotonicity across the bins. After the initial round of adjustments to the bins to achieve monotonicity, we identified that the variables ‘SUM\_AMT\_CREDIT\_SUM\_OVERDUE’, ‘MONTHS\_REGISTRATION’, ‘AMT\_CREDIT’, ‘SUM\_AMT\_CREDIT\_SUM\_DEBT’ and ‘AVG\_CNT\_CREDIT\_PROLONG’ either resulted in a non-monotonic and unexplainable trend, or single bins, which we deemed as unusable. Hence, we decided to drop these variables.

The remaining variables where their bins were adjusted are ‘AMT\_ANNUITY’, ‘SUM\_AMT\_ANNUITY’, ‘AMT\_INCOME\_TOTAL’, ‘AGE’, ‘AMT\_GOODS\_PRICE’, and ‘REGION\_POPULATION\_RELATIVE’.

We also gained several insights regarding the trend of the industry, such as how applicants with longer employment history had a lower probability of default as they have higher financial stability; and applicant with higher level of education had a lower probability of default due possibly because they have better paying jobs.

This allowed us to gain the following insights:

1. Older loan applicants have a lower probability of default due to longer working experience, and thus constant income, as compared to younger applicants.
2. Applicants with higher income have a lower probability of default as compared to applicants with lower income.
3. Applicants applying for a larger housing loan have a lower probability of default as the collateral charged is probably of higher value; and
4. Applicants living in areas of higher population density have a lower probability of default as compared to applicants living in areas that are of lower population density.

After adjusting for monotonicity in the variables, we ended up with a total of 15 variables. However, upon recalculating the IV, the variable ‘AMT\_ANNUITY’ fell below 0.02. Thus, we decided to drop it as we deemed it will not be useful in our prediction. Finally, we concluded this step with a final list of 14 variables, which can be seen in appendix 1.

## **3.2 Create Scorecard**

After finalizing the variables for the scorecard, we prepared the data for logistic regression by doing the train-test split of 70 - 30. We ensured a sufficient spread of events and non-events for both the train and test sets. Subsequently, we prepared a dataset with the WOE values for logistic regression training using the *woebin\_ply()* function.

We then created a function to loop through various class weights to identify the optimal class weight. We looped it through a weight range of 0.5-0.96, with an increment of 0.02. Eventually, the optimal class weight was determined to be 0.88 as it resulted in a higher percentage of defaulters being predicted and at the same time, did not predict too many false positives. We then used this weight to create our scorecard as seen in appendix 1.

## **3.3 Analysis & Tuning**

Upon running the scorecard with the test set, we generated the confusion matrix to analyse the performance of our model as seen in Appendix 2, along with the performance measures. Overall, 94% of the data was predicted correctly, of which 79% of the positive class was predicted correctly but only 53% of the negative class, the defaulters, was predicted correctly. The F-1 score was 0.86 which is also in the acceptable range.

We also plotted the train and test scores, which followed a normal distribution. Subsequently, we plotted a Receiver Operating Characteristic (ROC) curve which shows the diagnostic ability of the binary classifier system. For the test set, the Area Under Curve (AUC) was 0.7318.

We then wanted to improve the ability of the model in detecting defaulters, so we then attempted to include some variables into the scorecard such as ‘AMT\_ANNUITY’ and remove some variables. However, the result was not satisfactory, as the specificity value and the AUC for the initial set was still higher, hence we decided to keep the scorecard as it is.

# **4 Walkthrough**

An applicant has the following characteristics: The applicant is a 37-year-old pensioner who was accompanied by his family when applying for the loan. He is currently unemployed and receives an income of $112,500.00 from his pensions. He studied up until secondary school and is currently living in a region with a relative population of 0.025164 and a city rating of 2. From the bureau, the following information about the applicant was found: The total sum of annuity for all his credit amounts to $12163.50. His normalised score from External Data Source 2 is 0.30, External Data Source 3 is 0.12, and no score for External Data Source 1.

With reference to the scorecard in the appendix, this applicant will receive a score of 481, which is significantly lower than even the adjusted cut-off score of 520. Hence, this applicant’s loan will not be approved, and not considered manually as well.

# **5 Limitations of Analysis Approach and Opportunities for Further Improvements**

A potential limitation of our methodology may be that our assumptions regarding the chosen target group may not be entirely accurate or representative. For instance, we assumed that all singles applying for HDB loans would not own any other properties (including private properties) and hence immediately eliminated all several housing-related variables. Whilst we attempted to select the best out of the remaining variables, we acknowledge that there could have been useful information in the housing-related variables, had we not made the simplifying assumption.

Upon further research, we have found that it is legal for an individual in Singapore to own a private property for up to six months upon purchase of a resale HDB. Hence, it may be possible for members of our chosen target group to potentially own a property at the time of loan application. We believe this presents an opportunity for us to re-segment our target group and explore the variables we initially rejected to investigate if we are able to generate a better scorecard.

In addition, another limitation is the fact that we are uncertain regarding the recency and relevance of the dataset utilised. This dataset is from applicants in the US, hence there will definitely be differences in terms of behaviour. The real estate market in Singapore is also vastly different from the US, hence that might also be a limitation when housing loans are considered. If given a chance, we would propose comparing the sample to the current population in Singapore through PSI metrics to determine the appropriateness of the model we have generated. This would then allow us to tailor the model better for the current population representation of our target group.

# **6. Challenges Encountered and Lessons Learned from Analysis**

The main concern with this analysis was identifying characteristics that had sufficient IV to include in the scorecard. As stated above, the IVs of the chosen characteristics fall within the medium to weak categories, which offer some but insufficient predictive power. While only a possibility, there is a likelihood that this was due to the chosen customer segment, creating a subset of the original data that resulted in low WOE for the characteristics. In this case, it would be ill-advised to attempt data manipulation and force a characteristic to have high IV for the sake of the model. Instead, it would be better to present the findings as is, with the assumption that the data is an up-to-date and fairly accurate representation of the population and acknowledge that the proposed plan may not be as ideal as imagined. This would allow the business to make a more informed decision regarding the product and the risks involved.

Another challenge was the low specificity, with only 53% of defaulters being classified correctly. This can be detrimental to the FinTech lender as even a 2% default rate can result in significant losses being recognised. As mentioned earlier, with a more specific dataset and with more time where the team can explore more variables, it would allow us to improve the performance of the model, not just in predicting the positive class, but also in predicting defaulters.

In conclusion, we believe that singles housing loan is a good opportunity, and the model we generated performs decently with an appropriate AUC value and accuracy.

# **References**

Auto, H. (2021, June 16). Fewer S'poreans marrying and having children: Population census. The Straits Times.  https://www.straitstimes.com/singapore/politics/fewer-sporeans-marrying-and-having-children-population-census

CNA. (2023, August 20). NDR 2023: Timeline of public housing options for singles through the years.  https://www.channelnewsasia.com/singapore/national-day-rally-2023-timeline-public-housing-hdb-option-singles-through-years-3711621

PropertyGuru Editorial Team. (2023, October 11). HDB Build-To-Order (BTO) Flats in Singapore.  https://www.propertyguru.com.sg/property-guides/hdb-bto-launches-68106

Yuit, C. K. (2022, June 25). Demand for housing loans from HDB expected to grow as bank rates rise. The Straits Times. https://www.straitstimes.com/business/property/demand-for-housing-loans-from-hdb-expected-to-grow-as-bank-rates-rise?close=true

# **Appendix**

## **Appendix 1: Final Credit Scorecard**

|  |  |  |
| --- | --- | --- |
| **Characteristic** | **Attribute Bins** | **Points** |
| AGE | [-inf,48.5) | 36 |
| [48.5,55.0) | 35 |
| [55.0,inf) | 32 |
| AMT\_INCOME\_TOTAL | [-inf,70000.0) | 38 |
| [70000.0,160000.0) | 36 |
| [160000.0,210000.0) | 36 |
| [210000.0,330000.0) | 34 |
| [330000.0,inf) | 32 |
| EXT\_SOURCE\_1 | missing | 33 |
| [-inf,0.38) | 25 |
| [0.38,0.54) | 36 |
| [0.54,0.62) | 44 |
| [0.62,inf) | 55 |
| EXT\_SOURCE\_2 | missing | 27 |
| [-inf,0.18) | 16 |
| [0.18,0.52) | 29 |
| [0.52,0.64) | 38 |
| [0.64,0.7000000000000001) | 45 |
| [0.7000000000000001,inf) | 54 |
| EXT\_SOURCE\_3 | missing | 29 |
| [-inf,0.22) | 8 |
| [0.22,0.4) | 31 |
| [0.4,0.6) | 38 |
| [0.6,inf) | 54 |
| MONTHS\_EMPLOYED | missing | 41 |
| [-inf,70.0) | 33 |
| [70.0,220.0) | 37 |
| [220.0,inf) | 41 |
| NAME\_EDUCATION\_TYPE | Academic degree%,%Higher education | 41 |
| Lower secondary%,%Incomplete higher%,%Secondar... | 34 |
| NAME\_INCOME\_TYPE | Unemployed%,%Pensioner | 35 |
| State servant | 35 |
| Commercial associate | 35 |
| Working | 35 |
| NAME\_TYPE\_SUITE | missing | 69 |
| Group of people%,%Other\_A%,%Other\_B%,%Family | 42 |
| Spouse,partner%,%Children%,%Unaccompanied | 35 |
| OCCUPATION\_TYPE | missing | 38 |
| HR staff%,%IT staff%,%Privateservice staff%,%... | 44 |
| Core staff%,%Highskilltech staff%,%Managers%... | 37 |
| Medicine staff%,%Waiters/barmen staff%,%Cleani... | 33 |
| Laborers%,%Secretaries%,%Cooking staff%… | 31 |
| ORGANIZATION\_TYPE | Industry: type 6%,%Industry:type 2%,%Transpor… | 51 |
| XNA%,%Kindergarten%,%University%,%School%.. | 42 |
| Industry:type 4%,%Other%,%Medicine%,%Legal Se... | 38 |
| Trade:type 2%,%Insurance%,%Industry: type 7%,... | 33 |
| Transport:type 2%,%Industry: type5%,% Transpo... | 27 |
| REGION\_POPULATION\_RELATIVE | [-inf,0.005) | 34 |
| [0.005,0.036) | 35 |
| [0.036,inf) | 39 |
| REGION\_RATING\_CLIENT\_W\_CITY | [-inf,2.0) | 42 |
| [2.0,3.0) | 36 |
| [3.0,inf) | 29 |
| SUM\_AMT\_ANNUITY | missing | 34 |
| [-inf,60000.0) | 36 |
| [60000.0,inf) | 36 |

## **Appendix 2: Model Performance**

Confusion matrix:

|  |  |  |
| --- | --- | --- |
| **Actual \ Predicted** | **0 (No Default)** | **1 (Default)** |
| **0 (No Default)** | 1323 | 342 |
| **1 (Default)** | 78 | 88 |

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **0 (no default)** | 0.94 | 0.79 | 0.86 | 1665 |
| **1 (default)** | 0.20 | 0.53 | 0.30 | 166 |
| **Accuracy** |  |  | 0.77 | 1831 |
| **Macro Average** | 0.57 | 0.66 | 0.58 | 1831 |
| **Weighted Average** | 0.88 | 0.77 | 0.81 | 1831 |