1. Problem Definition and use cases

Credit risk arises when a corporate or individual borrower fails to meet their debt obligations. It is the probability that the lender will not receive principal and interest payments of a debt required to service the debt extended to a borrower. On the side of the lender, credit risk will disrupt its cash flows and also increase collection costs, since the lender may be forced to hire a debt collection agency to enforce the collection. The loss may be partial or complete, where the lender incurs a loss of part of the loan or the entire loan extended to the borrower.

There are several types of credit risks, however, our project will focus on credit default risk. Our aim is to predict the likelihood of a loan going default based on the profile of the borrower. This will help the lender decide on disbursing a loan to a particular borrower.

Previously, this prediction was done without the use of any machine learning algorithms. Lenders handed out loans based on a manual due diligence process driven by past experience and intuition. This process however, has two main drawbacks. Firstly, these approaches may be subjected to human biases. Secondly, lenders would have to go through all loan applications manually, which is a very tedious and time consuming process.

Hence, we aim to leverage machine learning algorithms to aid in the due diligence process, as a first filtering of loans, and potentially used to develop a custom credit score. Using the past behaviour of loans, our team aims to predict how the future loans will perform. Our features consist of the metadata and profile of the borrowers (i.e. age, education and so on), whereas our target variable is whether it’s good or bad credit. For this analysis, our team has framed the problem as a supervised, binary classification problem. In the future, the problem could also be framed as a regression problem such as to predict future income to determine a customer’s credit worthiness. Both techniques are commonly used in the credit risk modelling industry.

2. Dataset and Pre-Processing

The dataset is of historic loan repayment information from an Indonesian lender. The data contains 271,488 rows and 58 columns which came directly from Helicap’s database. The dataset was split as 80% for training the models, 10% for the validation of the models using 9 fold cross validation and the last 10% was to test the models for final evaluation.

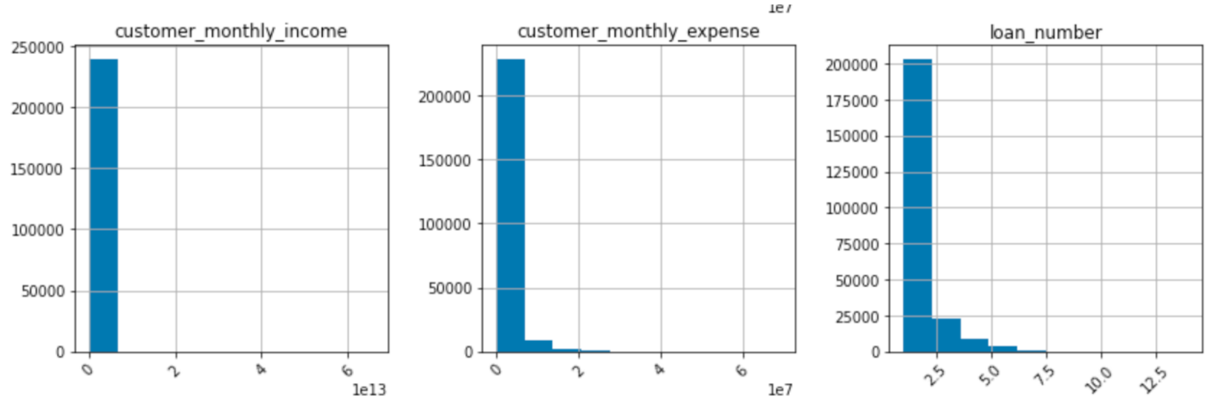
Outliers and Irrelevant Columns

Empty columns are insignificant to the size of the dataset, hence they were dropped. After multiple rounds of iteration, we learnt from the client that the dataset includes features from loan pre-approval and post-approval stage. We removed features from the post-approval stage so we can simulate the actual conditions of predicting loans as closely as possible. We learnt that this makes a critical difference to the choice of our models because there is a strong bias towards features in post-approval stages as predictors for bad credit.

Defining Target Variable

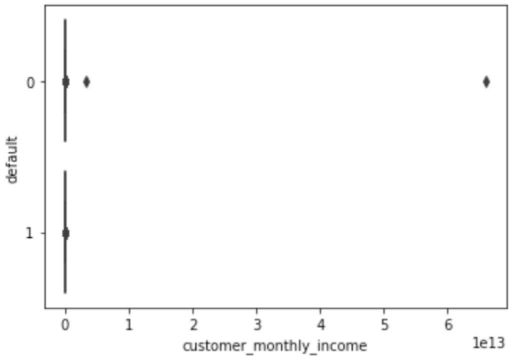
The target variable is derived from the feature “days past due”, based on a definition provided from Helicap, where loans that are 90 days past their due date are considered bad credit. This resulted in a 11:88 ratio between bad credit and good credit, signalling an unbalanced dataset. The remaining loans will be considered good credit. We proceeded with data exploration to perform feature selection for training our models.

Exploratory Data Analysis (EDA)



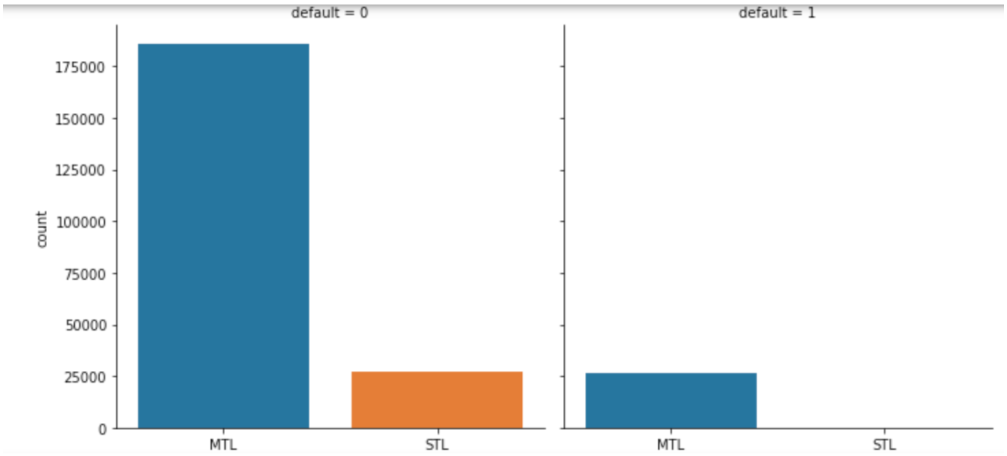
Histogram for numerical columns with skewed distributions

We plot histograms and box plots to explore our numerical columns. The key insight from the histograms is that the features customer monthly income, customer monthly expense and loan number are heavily skewed.



Box plot for customer monthly income

The box plots also reveal that most numerical values do not provide clear separation between good and bad credit because they follow similar distributions. Customer monthly income does reveal that only low income profiles cause defaults.



Comparison of counts of product type between good and bad credit profiles

For categorical columns, bar charts based on count were plotted. Product type does reveal potentially that only MTL product type results in bad credit profile because bad credit only contains MTL product type. The other categorical columns do not reveal obvious patterns between good and bad credits.

Feature Selection

We systematically selected the best subset of categorical and numerical columns based on the three following processes.

1. Pearson's Chi-Squared Test of Independence

This statistical test is used to determine whether the target variables is highly dependent on the categorical variables. A higher value with significant p-value will signal that the feature is a good predictor. We chose the top 4 categorical variables with the highest Chi-Square value and most significant p-values - Loan Product Type, Job Type, Job Industry and Last Education. One hot encoding was then used to create dummy variables for these categorical variables.

2. ANOVA F-Statistic Test

This statistical test was used to find the most important numerical variables that are significant to the target variable. We chose the top 5 numerical variables with the largest F scores and significant p-values - Loan Duration Amount, Loan Amount Request, Customer Monthly Expense, Number of Loans and Age.

3. Pairwise Correlation with Heatmap

The correlation heatmap is used to find multicollinearity between the numerical input variables. The only strong correlation was 71% between loan amount request and loan duration request. However, we kept both variables as we determined that they would be useful in our predictions.

Oversampling

As the dataset was highly imbalanced, oversampling was done using SVM SMOTE. This creates synthetic observations of the minority class, to reduce dataset imbalance using SVM. Out of all the other modern oversampling techniques like BorderlineSMOTE or ADASYN, SVMSMOTE performed the best cross validation scores on our base models.

3. Best Model Chosen: Stacking

The chosen model is a stacking model with extreme gradient boosting (XGBoost) model as the base-model and a logistic regression meta-model.

Model	**Configuration**

Stacking Classifier
Base Estimator: XGBoost only
Meta Classifier: Logistic Regression
	XGBClassifier(objective="binary:logistic", subsample=0.5, gamma=0.4, scale_pos_weight=8)
	
	LogisticRegression(solver='liblinear', penalty='l1', C=0.001)

4. Justification

All candidate models were trained with their best hyperparameters from 9-fold cross-validation to get the scores in the table below. The three scores represent three independent facets of the model's performance. F2-Score is a threshold-centric performance measure that measures the model's balance between precision and recall with a focus on recall. Area under the precision-recall curve measures the discriminatory power and ability of the model to detect positive examples. Brier score is a probabilistic measure that measures the accuracy of probabilistic predictions. The stacking model performed the best overall assuming each score is weighted equally.

	F2-Score	**AUPR**	**Brier Score**
Logistic Regression	4	3	2
XGBoost	1	2 (tie)	3
Stacking	2	2 (tie)	1
MLP	3	1	4

The advantage of using stacking is to be able to combine heterogenous weak learners with a meta-model to harness the capabilities of a range of well-performing models. We trained a stacking classifier with only one base estimator and a single layer stacking architecture, but this can eventually be extended to more base estimators and multi-level stacking, as well as nested cross-validation to find the best parameters for all components in the stacking model.

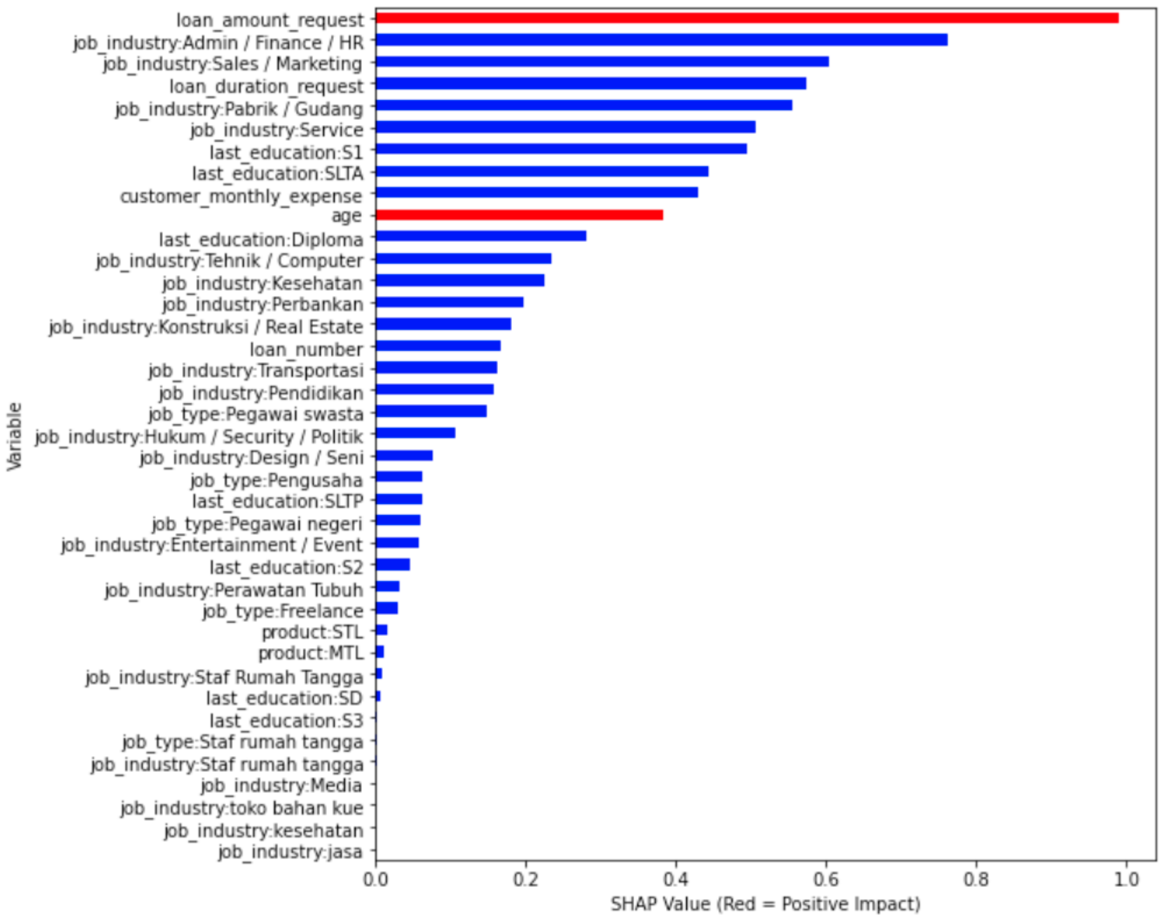
5. Model Explanations

Shapley additive explanations (SHAP) values summary plot is used to explain the global interpretability of the model. SHAP dependence plot is then used to explain possible interaction features that were suspected in the correlation table during data pre-processing.

Next, SHAP force plot is used to explain individual predictions. Lastly, the confusion matrix for the selected threshold is plotted for understanding the performance of the selected model.

SHAP Summary Plot

SHAP summary plot measures the average magnitude change in model output when a feature is “hidden” from the model. For our model, the SHAP output has log-odds units because it is a binary classification problem. Based on this, we can infer the most important features of the model.



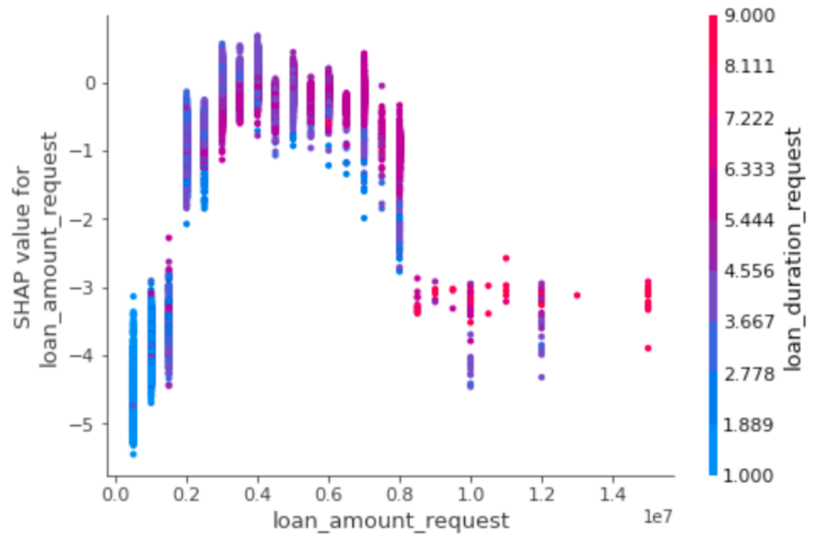
SHAP values from XGBoost Base Learner

Explanation for Most Important Features
The above SHAP summary plot reveals the most important features. Red bars indicate a positive correlation with bad credit, and blue bars indicate a negative correlation.

The most important feature is the loan amount requested. A higher loan amount requested will result in higher odds of loan defaulting. The next most important features are categorical variables which show a negative correlation with the odds of default. This indicates that borrowers in the job industry of “Admin / Finance / HR” and “Sales / Marketing” will result in lower odds of loan defaulting. In other words, borrowers in these industries can be trusted more to pay back their loans. Loan duration request is an interesting feature. Figure 1 shows that a higher loan duration requested will result in lower odds of loan defaulting. This could possibly indicate that a borrower is less likely to default if they request a longer payback period. This is insightful because some may believe that a longer payback period will result in an increased likelihood of defaulting.

SHAP Dependence Plot

SHAP dependence plot shows the marginal effect that one or two variables have on the predicted outcome. It allows us to infer if a relationship between the target and the variable is linear or more complex, and reveal interactions between features.



SHAP Dependence Plot of Loan_amount_request and Loan_duration_request

Explanation for Possible Interaction Features
Each dot is a single prediction from the dataset. The x-axis is the actual value of the loan amount requested. The y-axis is the SHAP value for loan amount requested, which represents how much knowing this feature's value changes the output of the model for that sample's prediction.

The color of the dot corresponds to a second feature that may have an interaction effect with the feature we are plotting. If an interaction effect is present between this other feature and the feature we are plotting it will show up as a distinct vertical pattern of coloring.

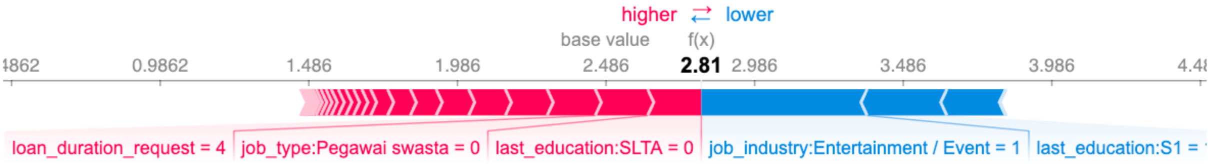
The above SHAP dependence plot reveals some clear vertical patterns of coloring in certain regions which means there may be possible interaction between the features, loan amount requested and loan duration requested. This lies in the regions where loan amount requested is low (below 0.2 * 10^7) and high (above 1.4*10^7).

SHAP Force Plot

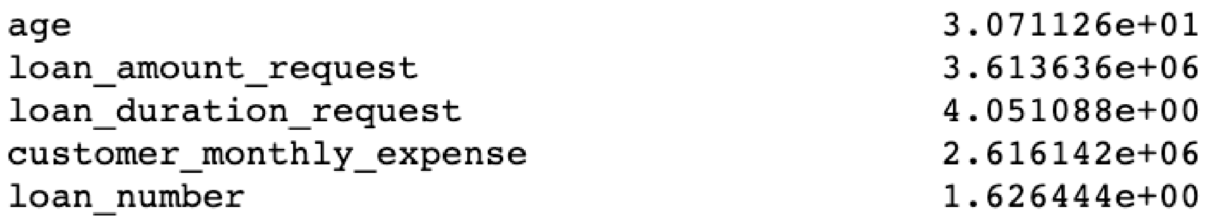
SHAP force plot reveals the top features that increase or decrease the odds of a borrower defaulting, aiding in our understanding of how and why the model makes a prediction.

```
True: 1 --> Pred: 1 | Prob: 0.6501620246019498
```

Ground truth label and prediction made by model for row 66 in test set



SHAP force plot for row 66 selected from test set, by log-odds



```
age                                         3.071126e+01
loan_amount_request                         3.613636e+06
loan_duration_request                       4.051088e+00
customer_monthly_expense                    2.616142e+06
loan_number                                 1.626444e+00
```

Averages of numerical columns

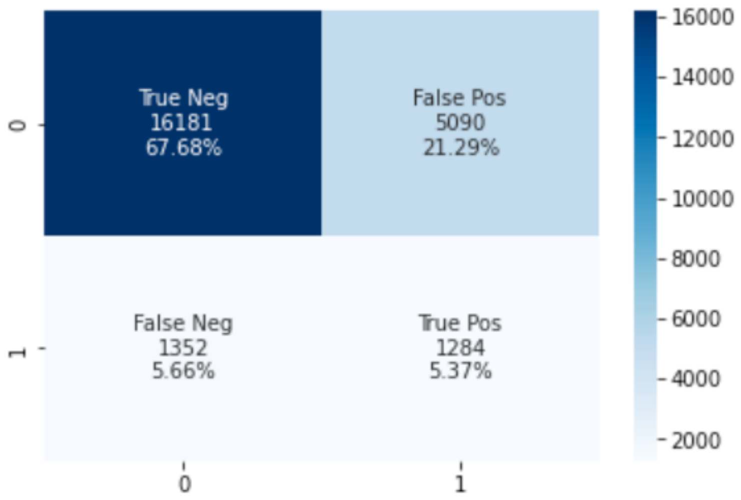
Explanation for Randomly Selected Prediction in Test Set

The force plot above is for row 66 in the test set, where the model correctly predicted a default. The axis of the plot is in log-odds. The base value is the average log odds of the test set. f(x) is the predicted log-odds value for this row, which is greater than the base value, indicating there are higher odds that this borrower would default.

The force plot reveals that the borrower's job industry and last education were the top features that reduced it's chances of default. We can infer that the loan duration request of 4 is one of the top 3 contributing features that increased the odds of the loan defaulting. This is close to the average of the loan duration requested. It is possible that loan durations requested should be much greater than the average to reduce the odds of defaulting.

Confusion Matrix

The confusion matrix is used to visualise the performance of the selected model, and more importantly easily see where the model fails exactly.



Confusion Matrix for selected model with threshold=0.52

We can see that the model finds it difficult to detect true positives well, with a relatively higher false positive rate, while having a similar false negative rate and true positive rate. This is achieved because we optimised the model for prioritising reducing false negatives at the expense of a higher false positive rate, due to the higher costs associated with false negatives compared to false positives. The trade-off is inherent but can be tuned if the model's priority changes in the future. One of the main reasons for such performance is the unbalanced dataset that gives a small number of examples of true positive classes to learn from.

6. Future Issues

Inability to Generalise for Lenders in Different Context
Because the dataset is from an Indonesian lender, it may not be directly relevant when used in the context of a Singaporean lender, who is subjected to vastly different micro and macro conditions. The importance of features are likely to depend on the internal and external conditions most relevant to the context of the lender. Therefore, we will be unable to extend our models and generalise them completely to other lenders that are dissimilar.

Inability to Capture Unforeseen External Circumstances
The models are unable to capture information from unforeseen circumstances or seasonality. Some borrowers might face unforeseen circumstances unrelated to their profile which our model captures, resulting in a prediction opposite of what our model expects. For example, a borrower might have faced the death of a sole breadwinner, which is the main reason for them not being able to pay their loan. In addition, there might be seasonal or cyclical external conditions that might result in higher loan defaults. For example, due to the COVID pandemic, it led to a sudden recession, job losses and pay cuts which is the main reason for some borrowers defaulting. These external factors all contribute to the borrower's probability of default, which the model currently cannot capture, leading to an incorrect prediction.

Issues with Dataset Growing Larger
The dataset can either grow deeper or wider. A wider dataset with more variables can lead to the curse of dimensionality, where the potential combinations of variables explode exponentially. Our key challenge then will be to choose the best features, or feature engineering to select the best predictors for our model. Regarding a deeper dataset, the required computational power to train the model might become too large. This is largely driven by the significant computation to perform SVM SMOTE oversampling and training the stacking model, assuming we have limited computing power and the proportion of bad credit profiles in our dataset remain the same. Furthermore, if the quality of our dataset does not improve, our models will overfit to the inaccurate data and deteriorate as the model becomes less representative of the population due to sampling bias.