University of Toronto

Master of Mathematical Finance

Machine Learning

Model Development Individual Report:

Credit Risk Model

Lujia (Lucy) Yang

January 2022

Contents

1 Business Value and Optimal Strategy

The variables with the relatively high feature importance based on XGBoost Model are also strong when fitting the logistic regression model. The XGBoost Model is superior to the traditional logistic regression model because it requires fewer historical credit records and can capture the non-linear feature and hyper-parameters. Especially when certain variables are highly correlated in the dataset. Some of the decision-making rules are similar under these two models. Based on the findings under both approaches, the company can apply the following business strategy to maximize profit and minimize default risks:

- Increase the amount of maximum applicable loan to female applicants or provide extra facilitate during the application approval process.
- For loyal customers with a high total number of drawings within 1 year, the company can encourage them to apply for more loans.
- Reduce the amount of loan and carefully investigate applicants with a high score from external data sources 3 and 2.

For the model deployment, to further improve the estimation accuracy and served unsatisfied needs, the company can apply XGBoost Model for customers without enough credit history, such as students and startups, while using the traditional model as a supplementary to verify if the selected features using machine learning algorithms are appropriate and explainable.

1.1 Business Rational for Final XGBoost Model

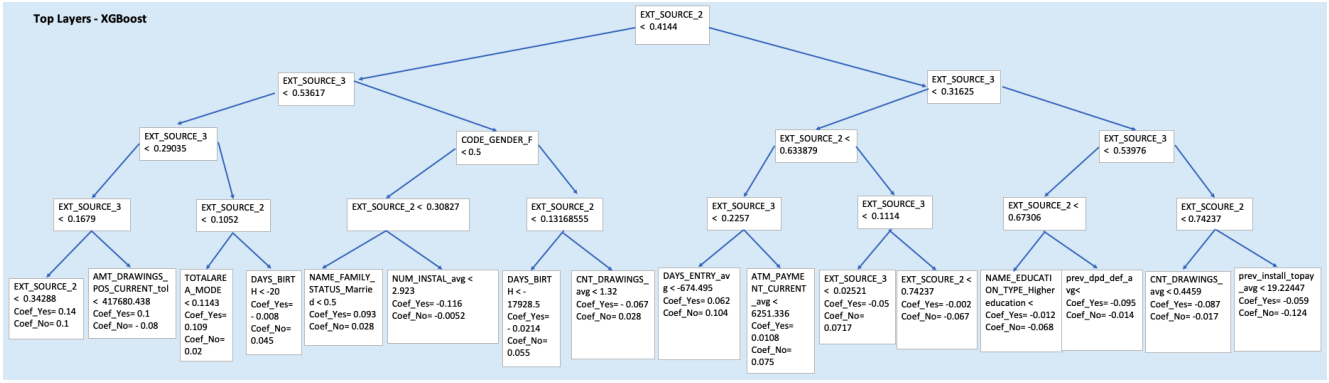

Figure 1: Model Top Layers: XGBoost

There are several findings and business insights based on the layers.

- The lower the total amount drawing or buying goods during the month of the previous credit, the higher the probability of default in the future.
- Younger applicants are more likely to default due to a lower financial ability and unstable source of income.

- Individual with lower education level seems more likely to default.
- Female with lower external score 2 and is single have higher default rate, so company can give more credits to a female who has already married. It is reasonable because obtaining extra financial supply for a married female is more meaningful, and they can use this money to support children and other families
- For a female with an external score 2 greater than 0.131685, the one with a lower average number of drawings have a lower default probability
- External score 3 and 2 are highly correlated and overall the lower the score the lower default probability. Specific numerical and categorical features linked to the default rate can be investigated under certain scenarios based on scores 2 and 3.
  - Individual who have score 2 less than 0.633879 and score 3 less than 0.2257, the higher the days' entry the higher the default probability.
  - Individual who have score 2 less than 0.633879 and score 3 greater than 0.2257, the higher the average current ATM payment amount the higher the default probability.
  - Individual who have score 3 less than 0.53976 and score 2 are less than 0.67306, the higher the average current ATM payment amount the lower the default probability.
  - Individual who have score 3 less than 0.53976 and score 2 greater than 0.67306, the higher the average DPD (days past due) during the month of previous credit the lower the default probability in the future.
  - Individual who have score 3 greater than 0.53976 and score 2 greater than 0.74237, the higher the average previous installment to pay the lower the default probability in the future.

The feature importance indicates that score 3 and 2 from the external resource is very strong, which is the same when we fit the logistic regression model. However, the XGBoost Model can consider hundreds of features and their correlation simultaneous instead of the limited variable selected in the regression model (around 10), which makes the model contains more meaningful features shown in the table 1.

Table 1: Other Meaningful features and Business application based on XGBoost Model

Variable	**Business Rationale**
AMT_CREDIT	Credit amount of the loan increase can raise the credit exposure and lead to higher credit risk
AMT_ANNUITY	Annuities are loans that are paid back over a set period at a set interest rate with consistent payments each period. The higher the annuities, the faster the credit institution collects the repayment from customers, resulting in less risk exposure.
DAYS_EMPLOYED	How many days before the application the person started current employment
AMT_GOODS_PRICE	The higher price of the goods for which the loan is given, the more difficult for a customer to repay the full amount and interest of the loan taken. Therefore, the higher probability of default.
AMT_INCOME_TOTAL	Income of the client higher means the ability to repay the loan is stronger.
DAYS_LAST_PHONE_CHANGE	If the client intends to change phone right before application, we might suspend his credit condition and do more investigation about the reason for the phone change.

1.2 Business Rational for Final Logistic Regression Model

During the feature engineering process, we create WOE to fit the logistic model, to overcome the data missing issue and outliers. Since WOE Transformation handles categorical variables so there is no need for dummy variables. Moreover, WOE can build a strictly linear relationship with log odds. Note that $WOE_i = \ln\left(\frac{f_G(i)}{f_B(i)}\right)$, and we expect the higher the default rate the higher the credit score $= \sum_{j,i=1}^{k,n}(-(WOE_{j,i}) \times \beta_i + \frac{a}{n}) \times factor + \frac{offset}{n})$. A positive (negative) model coefficient implies the underlying factors of the WOE variable and default rate is positive (negative) correlated. Based on the value of the coefficient, we can investigate the business rationale based on the magnitude of effect and the correlation of the explanatory and the default rate.

The summary for the final model are in the table below:

Table 2: Model Summary

Variable X_i	**Coefficient** β_j	**Std Err**	**Z**	**P>|z|**
REGION_RATING_CLIENT_W_CITY_woe	714.5484	98.582	7.248	0.000
CNT_DRAWINGS_0_1_year_woe	-2765.2940	290.931	-9.505	0.000
CODE_GENDER_F_woe	-1.406e+04	1624.123	-8.657	0.000
ORGANIZATION_TYPE_XNA_woe	1034.5720	184.194	5.617	0.000
NAME_EDUCATION_TYPE_Higher_education_woe	2319.7142	205.849	11.269	0.000
EXT_SOURCE_3_woe	5999.6226	186.631	32.147	0.000

We explain the model parameters and analyse the business insights in the table below:

Table 3: Feature Dictionary and Business Suggestions

Variable	**Business Rationale**
REGION_RATING_CLIENT_W_CITY	Our rating of the region where a client lives with taking the city into account (1,2,3). The positive coefficient = 714.5484 indicate the higher the rating the more likely the client default.
CNT_DRAWINGS_0_1_year	Total number of drawings in year 1 increase can decrease the default rate. It intuitively makes sense as a draw is a payment taken from construction loan proceeds made to material suppliers or contractors, meaning the borrower does not have to make the repay themselves while the project is ongoing.
CODE_GENDER_F	The coefficient indicate female is less likely to default, so credit lender may favor lending more to female.
ORGANIZATION_TYPE_XNA	The coefficient is 1034.572, which indicates a positive correlation with default. If the type of organization for the client is "XNA", they are more likely to default, so the company needs to carefully investigate its credit records and ability to repay.
NAME_EDUCATION_TYPE_Higher_education	The coefficient is 2319.7142. The applicants with higher education level seems to have a higher default rate
EXT_SOURCE_3	Normalized score from an external data source, which is one of the strongest variables. The higher the score 3, the higher the probability of default.

2 Model Assessment and Comparison

To measure and compare the performance of the XGBoost model and the logistic regression model, we use three performance measures to validate the selected model including the confusion matrix, the precision-recall plot, and the receiver operating characteristic curve (ROC curve). We first start by comparing the model prediction power using AUC and ROC curves. A precision-recall curve provides a visualization of the trade-off between precision and recall for different thresholds. The further away the precision-recall curve of the model is from that of the random guesses, the better the model is. Moreover, the ROC curve shows the model diagnostic ability and the higher the AUC means the higher the prediction accuracy. We select the model with greater AUC. Then, we further validate and compare the model using a confusion matrix, which measures the predictive power of a model by providing comparisons of values of correct and incorrect predictions, including True positive (TP), False-positive (FP), False-negative (FN), True negative (TN). An accuracy and F1 score can be calculated based on the confusion matrix. We first find an optimum threshold providing the best F1 Score, and then calculate the corresponding accuracy ratio. The best model selected will have the highest accuracy ratio.

2.1 Process of Model Selection

2.1.1 Raw Logistic Model

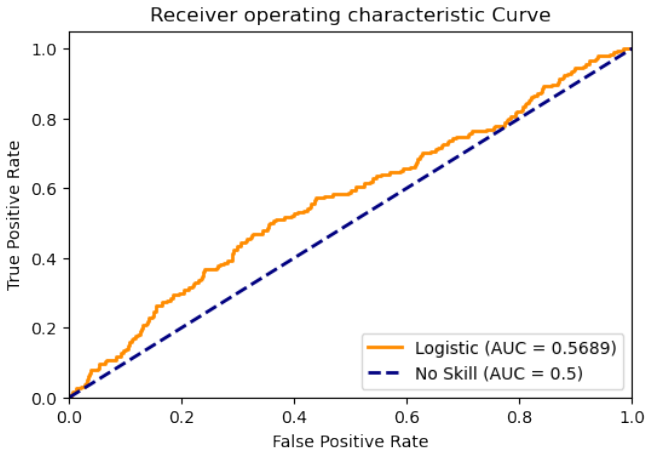

Figure 2: Receiver operating characteristic (ROC) Curve - Raw

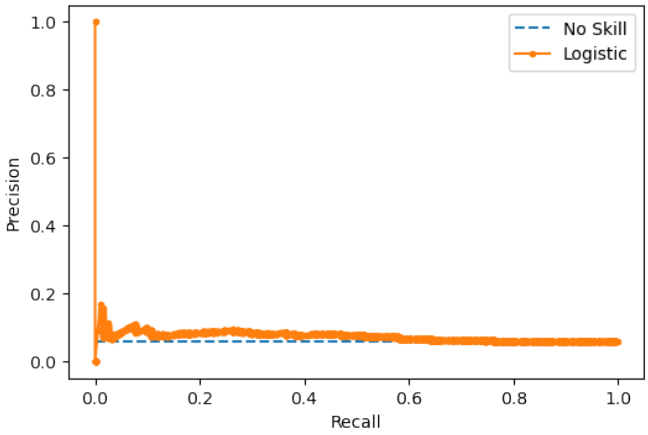

Figure 3: Precision-Recall Curve - Raw

We first start by fitting a full model using all features in the original dataset. Due to the multi-collinearity of explanatory variables, the raw model has very poor performance with small prediction power. We can see the ROC curve for the raw model is close to the random model, and AUC is only 0.5689. The precision-recall curve is close to the curve indicating no skill. Further investigating shows the variable correlation is high because there are hundreds of features in the linear model. The p-value for most of the variables is greater than 0.05, meaning the variables are not statistically significant.

2.1.2 Enhanced Logistic Model with Feature Engineering

As discussed in the group report, we create new features when aggregating different datasets, equal weight binning, and calculate WOE to use in the logistic regression model. We used both clustering and IV (importance value) based on WOE to reduce the number of variables. In the variable reduction process, we select the variable with the highest IV and lowest 1-R^2 within each cluster. The figure shows that the correlation between selected variables dropped significantly, which means there is no multicollinearity issue.

Figure 4: Variable Correlation

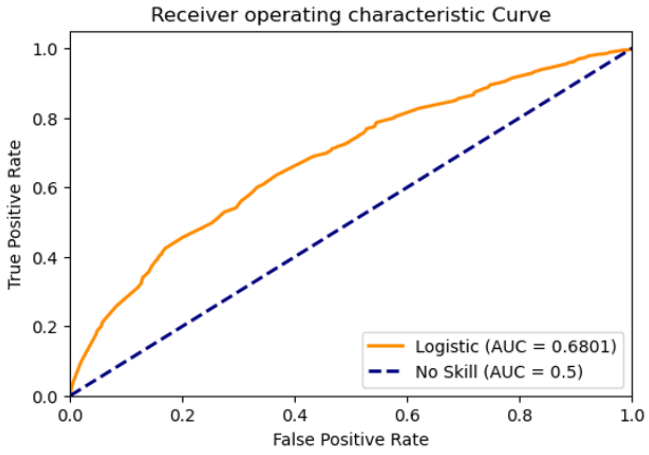

Figure 5: Receiver operating characteristic (ROC) Curve - Final

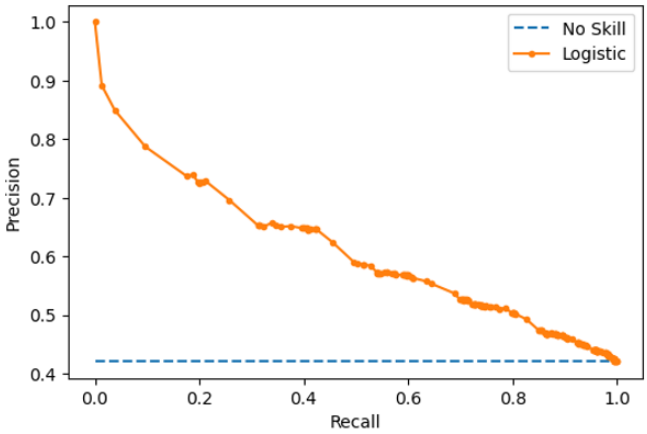

Figure 6: Precision-Recall Curve - Final

We conduct step-wise model fitting based on the list of the variable selected, based on IV and clustering. In each model fitting iteration, we compare the AUC, variable correlation matrix, and p-value for each candidate model, and select the best model. Comparing the Figure for the final version of an enhanced model to the raw model, we observe that the AUC has effectively improved from 0.5689 to 0.68. The distance between the precision-recall curve and no skill curve is much greater compared to the raw model. Therefore, under the logistic regression method, it is reasonable to select this model as the best final model. The final selected variables have information about region rating, the total number of drawings in the first year, gender, type of organization, education level, and external score 3.

2.1.3 Advanced Machine Learning model

Considering the logistic regression are unable to handle the non-linear relationship between explanatory and target variables, it is worth fitting the model using another methodology - Machine Learning XGBoost, to see if this method is more suitable to the given data. Boosting is an ensemble method to aggregate all the weak models to make them better and stronger. For feature selection, we implement recursive feature elimination (RFE) with cross-validation. This method fits a model and removes the weakest feature the entire process stop when a specified number of features is collected. One advantage of RFE is that it can reduce multicollinearity among explanatory variables. The features were reduced from 260 to 159 in the final model. The strongest features with high importance are in the Figure 7.

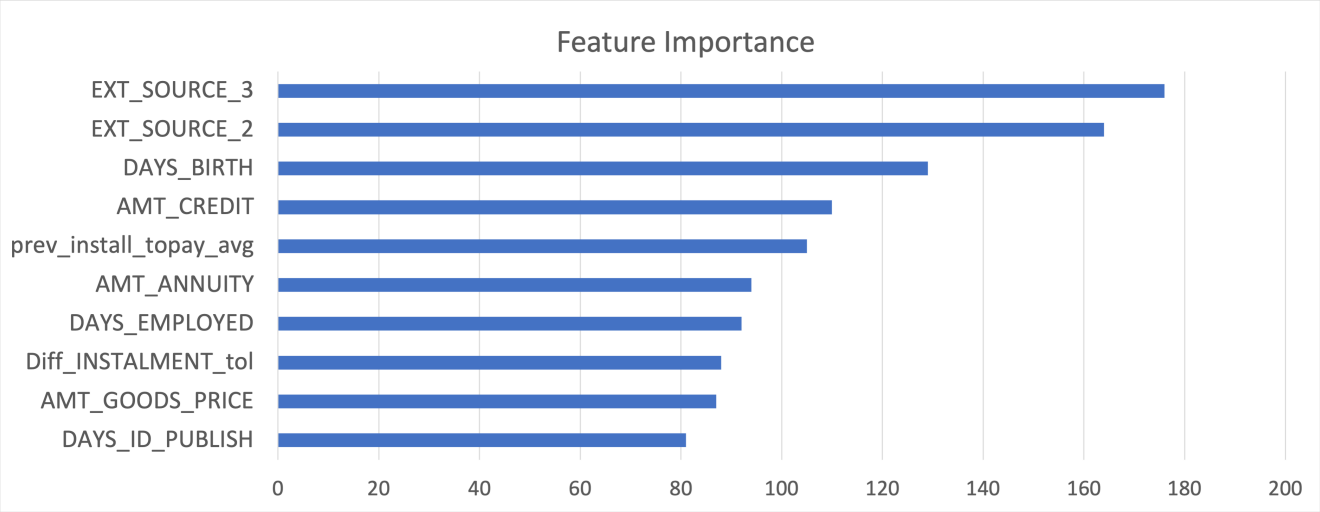

Figure 7: Feature Importance

Finally, we set the learning rate equal to 0.1 and maximum depth equal to 5 to tune the model. As XGBoost can include hyper-parameters, it can boost the results can measure the non-linear trend. The AUC for XGBoost is 0.77, which is 0.09 higher than the best logistic regression model.

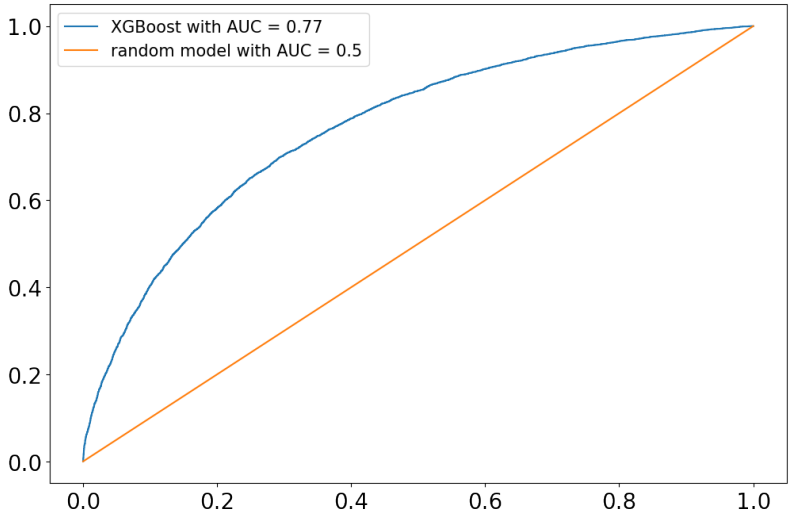

Figure 8: Receiver operating characteristic (ROC) Curve - XGBoost

2.1.4 Further Validation based on Confusion Matrix

To verify the previous conclusion, we calculated the best accuracy ratio, Precision, Recall, and the optimum F1 Score. The best accuracy ratio for XGBoost is 0.7043, which is 0.1821 higher than that of the final logistic regression model. The FP (false positive) rate for the logistic model is 0.75, which is very high. The FN (false negative) is only 0.1, indicating the business rule developed based on the regression model is not strict enough. In other words, approve decisions will be made for most applicants. If a company uses a logistic model, they are likely to lose profit by lending money to people that expected

to default in the future. On the other hand, the FP and FN (false negative) rates under XGBoost are lower. XGBoost model can predict the default rate more accurately.

Table 4: Accuracy Ratio Comparison

Criteria	**XGBoost**	**Logistic**
Accuracy	0.7043	0.5222
Precision	0.6970	0.4636
Recall	0.6501	0.8955
F1 score	0.6727	0.6109

Table 5: Confusion Matrix - XGBoost

		True	
		High(1)	Non-High(0)
Predicted	High(1)	TP = 0.65	FP = 0.25
	Non-High(0)	FN = 0.35	TN = 0.75

Table 6: Confusion Matrix - Final Logistic Regression

		True	
		High(1)	Non-High(0)
Predicted Class	High(1)	TP = 0.90	FP = 0.75
	Non-High(0)	FN = 0.10	TN = 0.25

Based on the further model validation, we conclude that our model selection process is reliable. Compare to the final model under logistic regression and XGBoost, the machine learning model have a better performance on the given dataset. If a company only tends to select one method, then it is wise to use the XGBoost model to achieve a more precise result.

2.2 Performance Gaps between XGBoost and Logistic Regression

We discuss the gap based on **modeling process, selected variables, and prediction accuracy**. XGBoost includes 159 variables after implementing recursive feature elimination with cross-validation. The model prediction result is more accurate because it contains hyper-parameters that solve the non-linear problem. This method does not have very demanding requirements for the amount of data and the characteristics of the variables. However, a lot of computational time and memory is required for feature selection and model fitting. Alternatively, the process of equal weight binning and calculating the weight of evidence is tedious in the logistic regression modeling process. Although the mathematical process is more complicated it requires less computational power and is time-saving when fitting the model.

There is a significant performance gap between XGBoost and the logistic regression model. XGBoost has better predictive power than the regression. Based on the ROC result, the AUC for XGBoost is 0.77, while the AUC for the logistic regression model is 0.68. In the confusion matrix, logistic regression has a higher true positive rate, whereas the true negative rate is around 25%. The company that uses the model to make a business decision is likely to lend money to a large proportion of customers unable to

repay the loan. The accuracy computed from the confusion matrix for XGBoost is 0.7043 whereas for logistic is only 0.5222.

Additionally, based on the model coefficient, most of the correlation relationship is consistent for both methods. For example, both models imply a positive correlation between external score 3 and the probability of default. They both conclude that the default rate for females may be lower. However, the result is inconsistent for some variables such as the EDUCATION_TYPE. The logistic model suggests applicants with higher education level is likely to default, which is a conflict with the common knowledge intuitively. On the contrary, XGBoost gives the opposite correlation result, and it is more in line with common sense. This may be because the external score 3 in the regression model is strong and likely correlated with other selected variables, including education level, and the number of drawings. However, we did not find this implicit relationship during the modeling process. The linear model does not include the hyper-parameter, and the correlation coefficient corresponding to the education level is negative, but it offsets the education-related components that may be included in the external score 3.

2.3 Future Improvements

We can improve the model performance by enhancing data quality. The original dataset contains 30% columns with greater than 50% missing value. Instead of dropping columns or filling the missing value using mean value, we can use a more advanced method such as KNN and random forest classifier. Additionally, we need to further investigate the meaning of some strong variables like EXT_SOURCE_3, including how the related third-party organization calculates the score 3. Although the project background might be based on customers with limited credit history, more detailed information, including the Number of 30+ dpd ratings in the past 12 months, and Average Credit Balance, and the Total number of trades ever 60 or more days delinquent or derogatory, are still worth to collect to consider more meaningful features and improve the interpretability of the selected model.

The model performance can be improved by a more comprehensive feature engineering. Based on the variable nature, analysts have to carefully investigate the meaning of the features and create appropriate features. For the aggregated variable by calculating the average, maximum, total, we can further conduct the normalization to scale these features. For categorical variables, instead of calculating WOE we can encode dummy variables or implement feature hashing. Additionally, for the XGBoost model process, we use k-means to turn spatial data into features. We can improve this process by using RBF SVM, GBT, or KNN. K-means only effect for real-valued bounded numeric features. We can define custom metrics to handle multiple data types and use the k-medoids algorithms. We can bin the categorical variables to further improve the feature engineering.

Moreover, based on the top layer in the XGBoost model, we see clients can be classified into different groups based on score 2 and 3 before the model start to consider other account level features like education and drawing numbers. Therefore, it is worth trying model segmentation based on the value of external scores 2 and 3. Under each population segment, we can fit a regression model and XGBoost model. By this process, we may get better results for the logistic regression model and machine learning model.

3 Potential 2-year Business Plan

To implement machine learning in the credit lending industry, the company needs to collect enough relevant data to measure credit risk. In the short run, based on the given data and the modeling result, it is worth collecting the current account-level data, including gender, region rating, education level, number of drawings, organization type, installment to pay, DPD (days past due), and cash balance. Previous account data also need to be collected, because the company can investigate the reason for previous rejection for the lending. Customers have been rejected before does not mean they cannot repay the loan in the future. Similar to developing a traditional credit scoring model, external credit bureau data can be collected.

We apply the XGBoost model to start to expand the lending business and use the logistic regression model as a supplementary tool. At the early stage, given limited data sources, we will develop one model and apply it to all types of customers, and keep collecting data by tracking their credit records once they start to use our products. In the middle stage, we will collect customer feedback and investigate the reason for customers being rejected for our system. Based on the latest dataset, we can rebuild the model and make further improvements to the XGBoost model. In the long run, once the dataset contains more comprehensive data and our loyal customers have developed their credit history with our company, we can further implement the model segmentation and improve the performance of both the XGBoost and regression model. Additionally, we can compare the XGBoost model with a more well-behaved traditional logistic regression model to enhance the model validation process of the machine learning model development.

With this innovative modeling methodology, the company can grab market share from traditional credit lending institutions. Forgiven principal, maturity, and repayment calendar, a customer that capable of repayment will not be rejected by the credit scoring system for our company, while they might be rejected from the bank's scoring system. By unlocking the full potential of our data, the company can create the following products:

- Credit cards: Company can issue cards to consumers in Canada, the United Kingdom, and the United States, including Visas and Mastercard options.
- Consumer banking: We can offer daily banking services to individual and small- to mid-sized business clients, including checking and savings accounts, loans, mortgages, and money market accounts.
- Commercial banking: This segment of the products portfolio serves commercial clients with banking, lending, real estate, and investment services.

Ongoing database management, model assessment, and product improvements will be implemented to maximize the profitability and minimize the credit risk of the company while ensuring customer satisfaction with the lending experience. In general, the business plan of the machine learning credit risk department can be well integrated with the company's business model.