PREDICTING CREDIT DEFAULT 2018-2019

NOVEMBER 8, 2019

Authored by: Angelina Kiman

Overview

Goals and Challenges

PT. X provides credit to the poorest of the poor in Indonesia. We specialize in microfinance that focuses on non-collateralize lending for microbusinesses and collateralize lending for small (UKM – Usaha Kecil & Menengah) businesses. As of August 2019, it has 145,000 members, 90 percent of whom are women. With 130 branches, our company provides services in 3 areas –Java, Sulawesi, and Maluku Island.

Our business purpose is to provide an alternative for the community to obtain a source of venture capital funds in an easier and faster way in comparison to conventional banking system. As we will expand continuously across Indonesia, we are hoping we can enhance the economic well-being of a family through capitalization, education, and women entrepreneurs.

> ***"Cultivating a sense of entrepreneurial spirit that stimulates growth of social welfare and equitable economic growth in Indonesia"***

This project explores the average aging days in both sectors. We are eager to learn our clients' behaviors patterns in making their loan payments. There are three main aspects that we will deep dive: (1) Analyzing the factors that may influence aging, (2) Examining features of database performance tools, and (3) Predicting borrowers' aging in days.

At the same time, we are dealing with inconsistent large datasets due to trial-and-error experiments. Data incompleteness is another problem that we encountered which may cause a slight bias during wrangling and manipulating the datasets. By the end of this project, we will be able to comprehensively understand our clients' characteristics and yet curate meaningful recommendations to remedy these issues.

Microbusiness

Overview and Data Exploration

As August 2019, we have 123,000 members and all borrowers are women. Inspired by Dr. Muhammad Yusuf's philosophy, we are encouraging our borrowers to take initiatives in business or agriculture which provide earnings and enable them to pay off the debt. We believe making microcredit available to the rural poor will give them a chance to have a prosper life by stimulating new businesses and reduce the widespread rural poverty in Indonesia.

We are curious if our data features are sufficiently enough to predict an individual's aging in days. If not, what are other parameters that we need to include? We believe it is crucial to measure from many different facets. During data exploration, we learned on *Table 1* that Karawang (1C) has the highest number of clients and followed by Bekasi (1B) and Bogor (1A); whereas Lampung (6), Kendari (5B), and Makasar (5A) are the lowest three of number of clients.

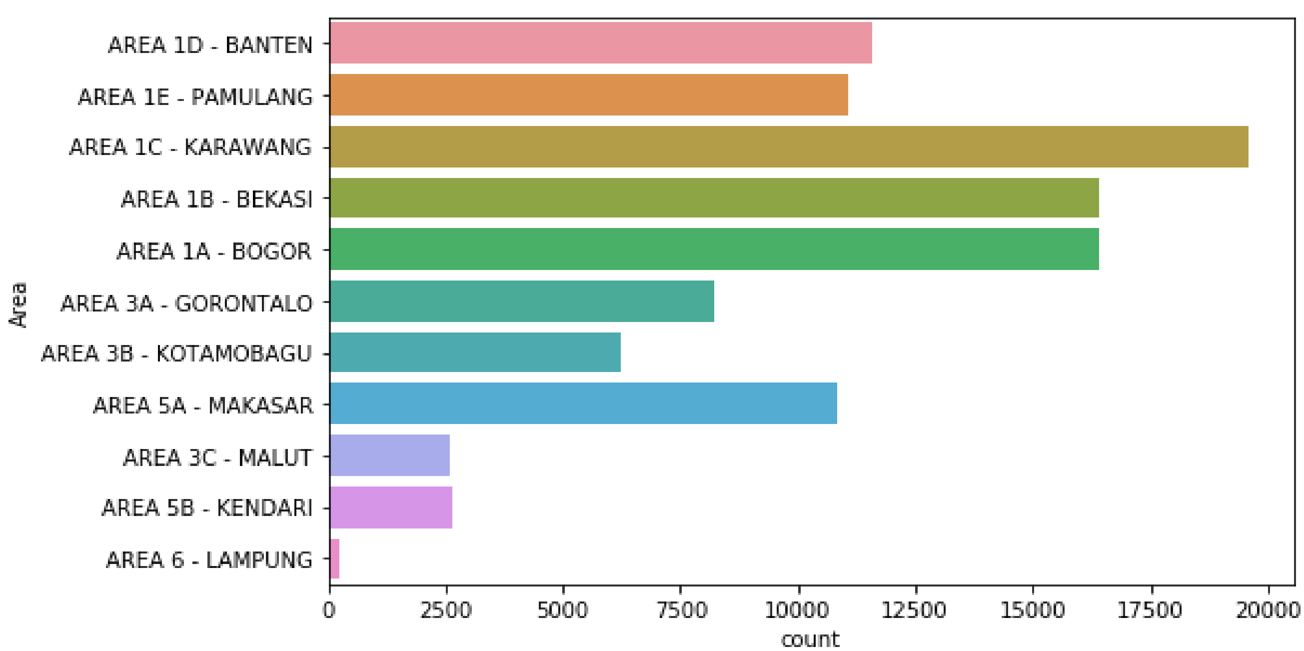

Table 1. Number of Clients based on Location

In *Table 2*, our clients are mostly homeowners. Less than one-third, our clients' houses are owned by their family members, and a little of them are still renting.

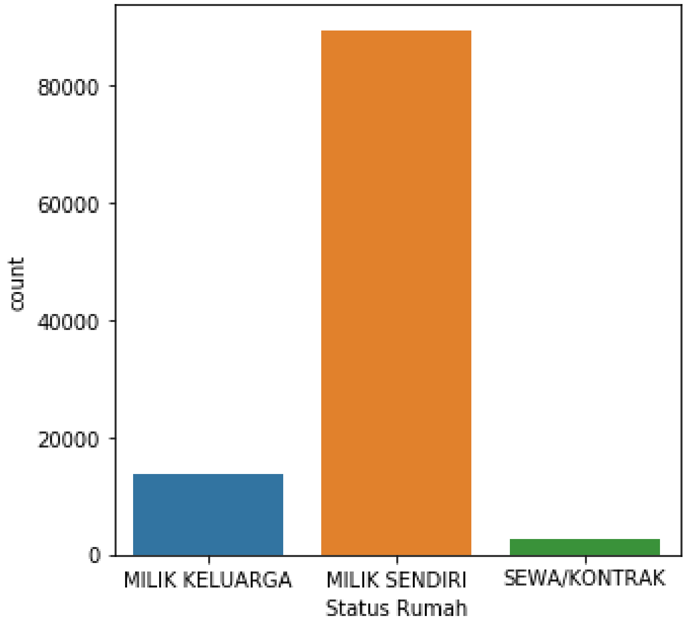

Table 1. House Ownership Status

In *Table 3*, the highest education of our clients is elementary school. The second highest is middle school (SMP), then followed by high school (SMA). It is quite surprising a little of them hold a diploma and bachelor's degree.

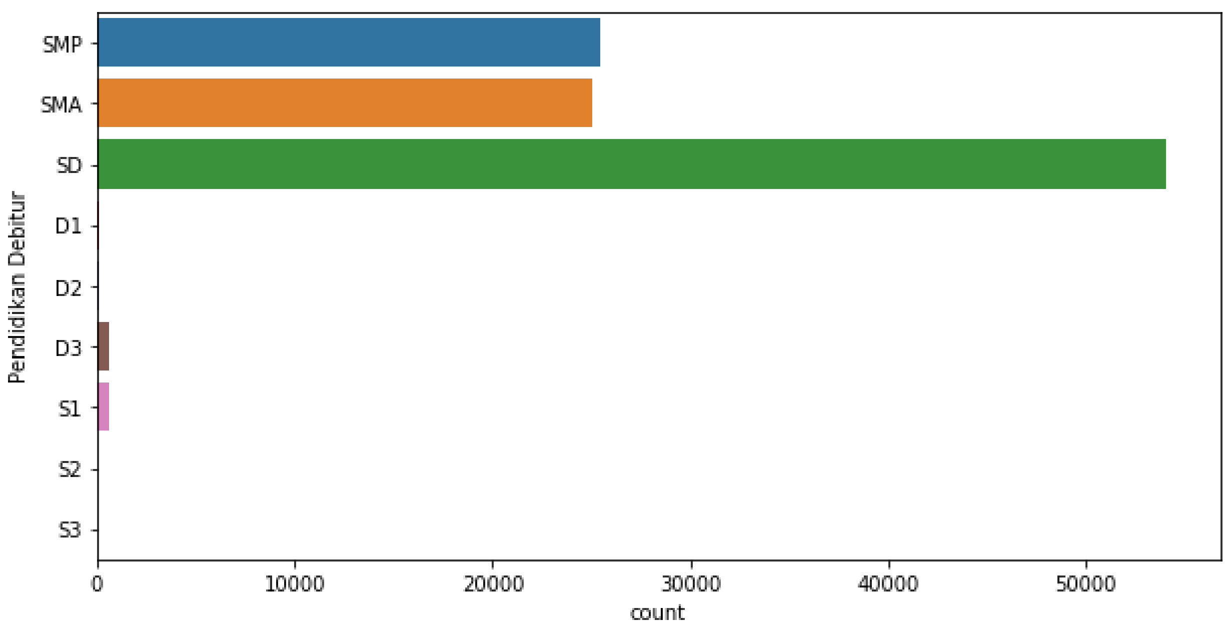

Table 3. Education Level

In *Table 4*, majority of our clients are married and around one-third are divorced.

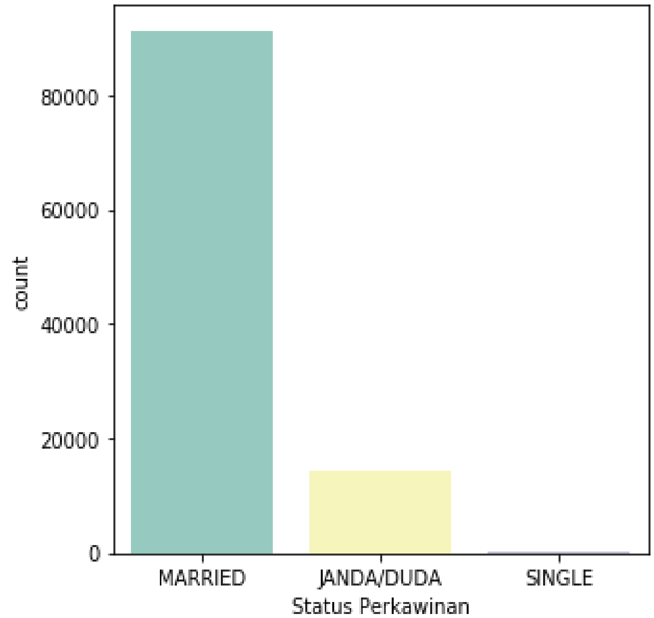

Table 4. Marital Status

In *Table 5*, 90 percent of our clients have trading business. This means they are purchasing goods from other source and reselling the goods to the public. Followed by home industry business as the second highest. As an example, list of activities that they are involved are creating homemade goods, recycling, and farming. Lastly, service business is the least one. Typically, they are providing service such as laundry, game rental, and massages.

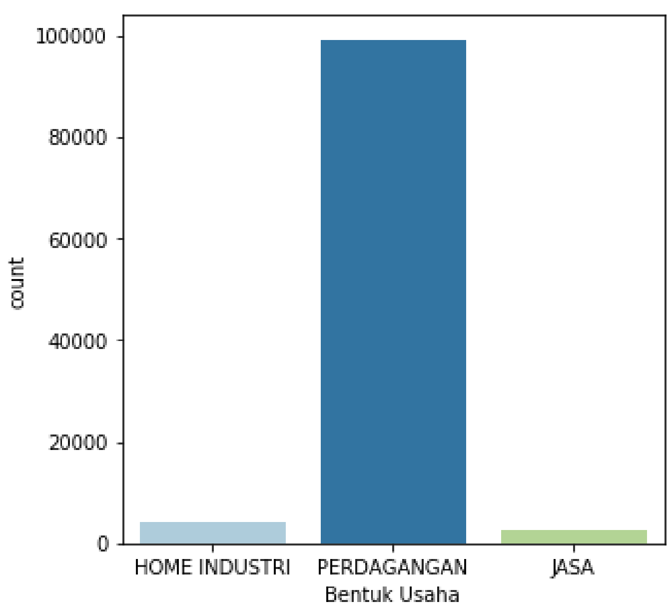

Table 5. Business Sector

In *Table 6*, it appears to be that our clients would like to raise capital to their current business and a little of them would like to start a new business.

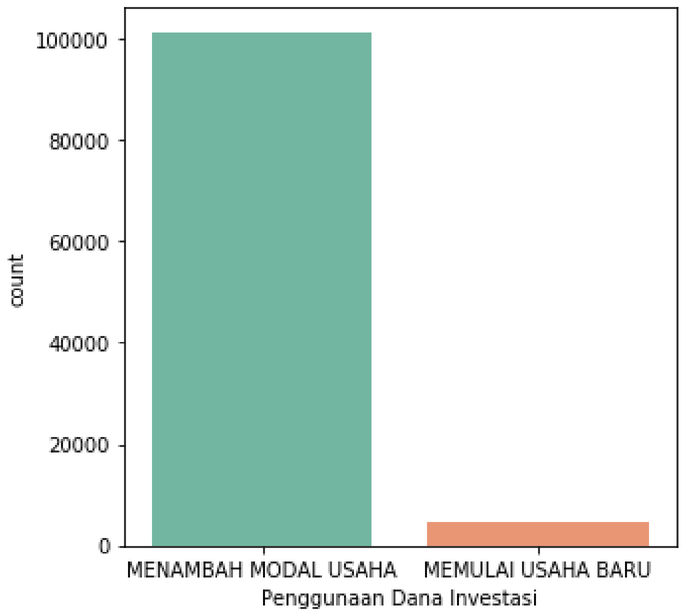

Table 6. Loan Purpose

In *Table 7*, even though the majority do not have other loans, more than half of our clients have an on-going loan from other institution.

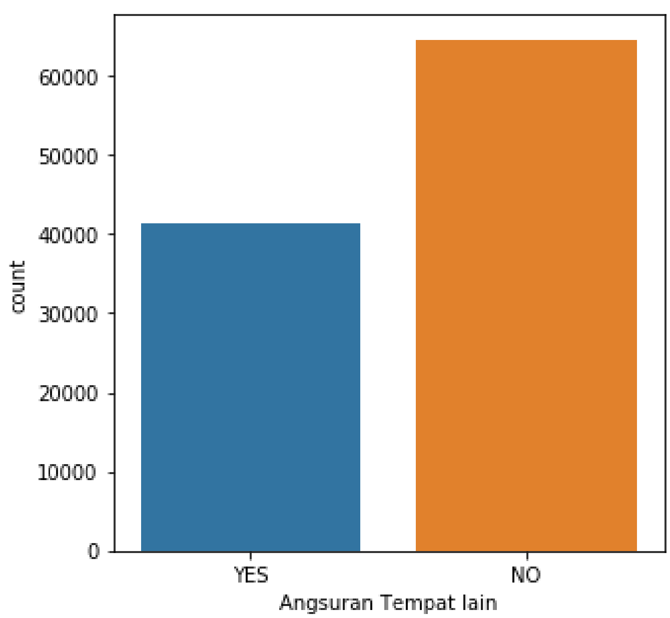

Table 7. Other Loans

In *Table 8*, we are performing a correlation graph to find in detail which variables are connected.

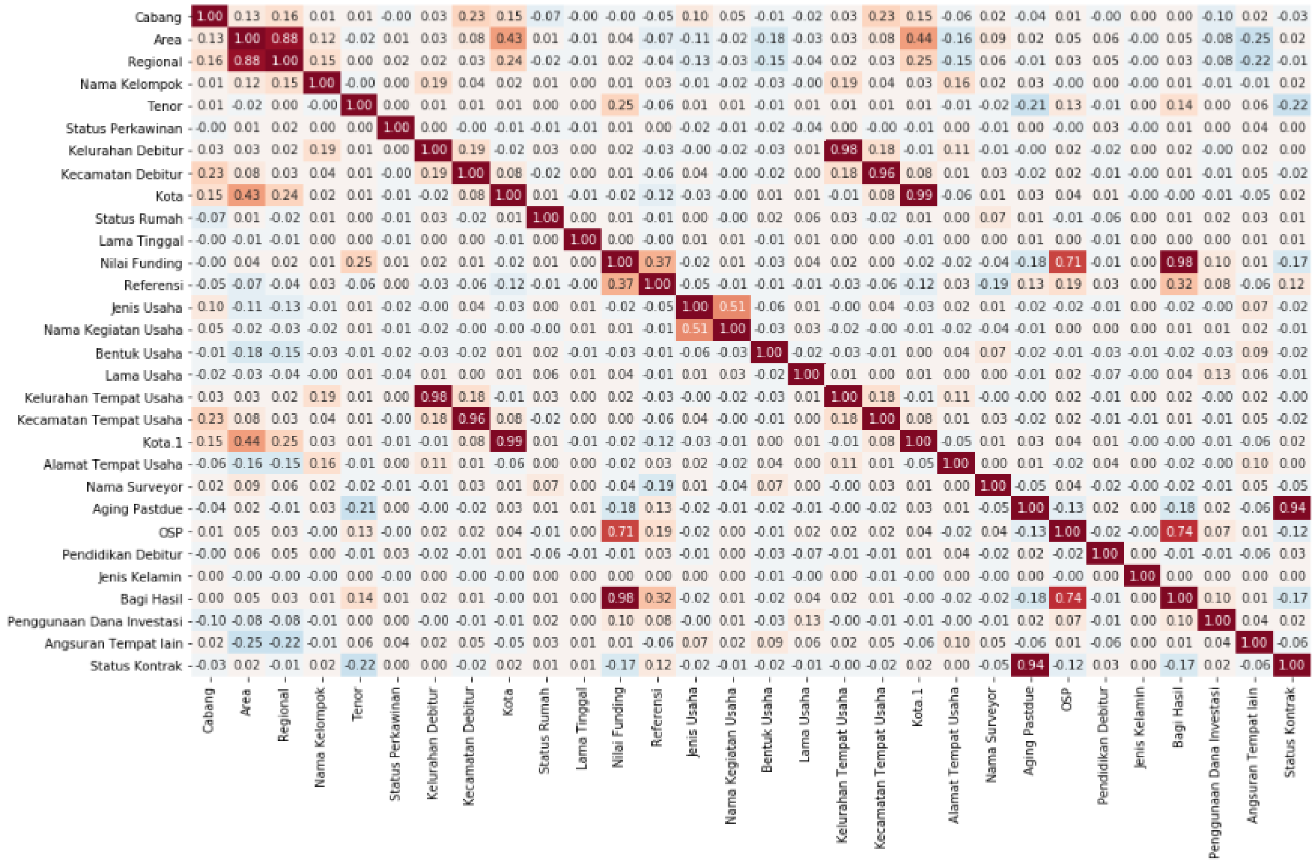

Table 8. Correlation Graph

In *Table 9*, we are ranking which variable is related to the target, 'Aging Past Due'.

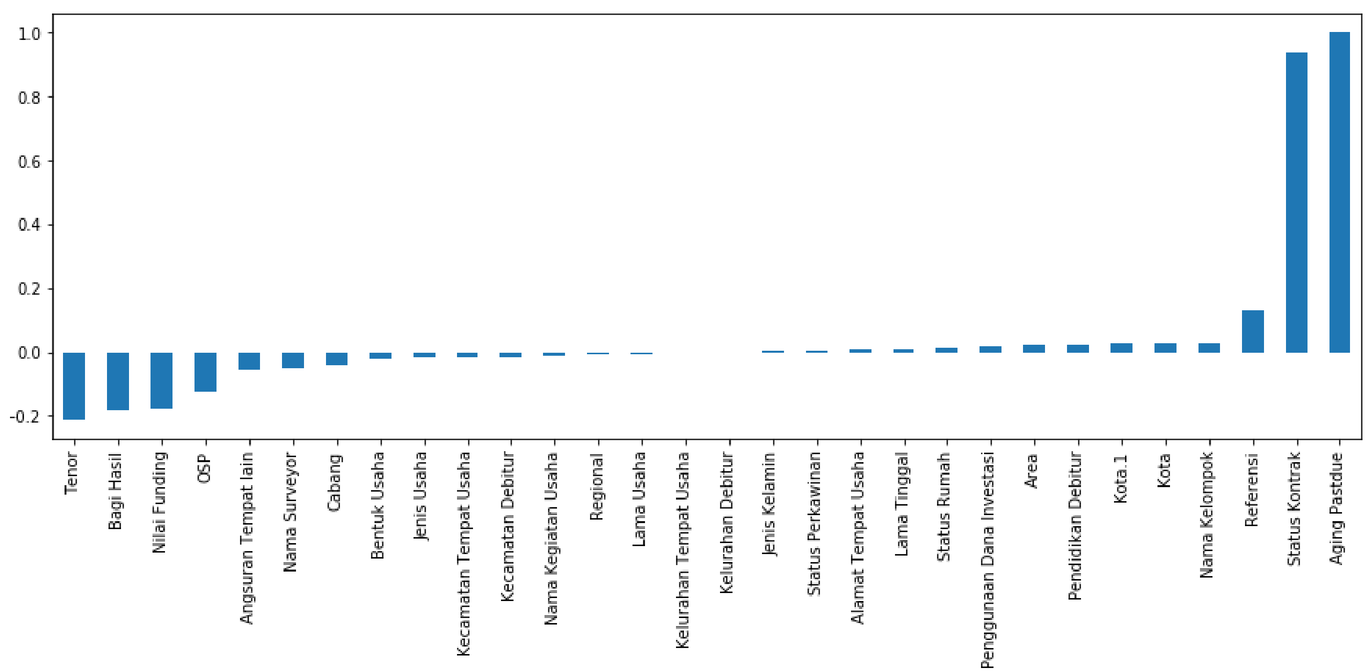

Table 9. Microbusiness Features Ranking

Machine Learning Algorithms

We are running two different algorithms, Random Forest and Gradient Boosting, to measure how well our parameters in estimating aging in days of our clients. We are using R^2, the goodness of fit test, to measure how well our linear regression models; and RMSE (Root Mean Square Error), to measure how much error between two datasets or prediction errors.

Gradient Boosting

This algorithm is mainly used by many data scientists because of the good results it yields on any given (unknown) problem. It is an example of boosting algorithm that uses sequential classifiers and greatly reduce bias and variance; however, the downside is a tendency to overfit the model.

	2018	**2019**
R^2	0.8157	0.7701
RMSE	30.8277	43.1336

Table 10. Score Results

In 2018, *Table 10* shows that 81.57% we can predict microcredit default among our borrowers. The average square of credit default error is 30.8277 days. Whereas in 2019, the performances go down by 4% and the square error increases by 13 days.

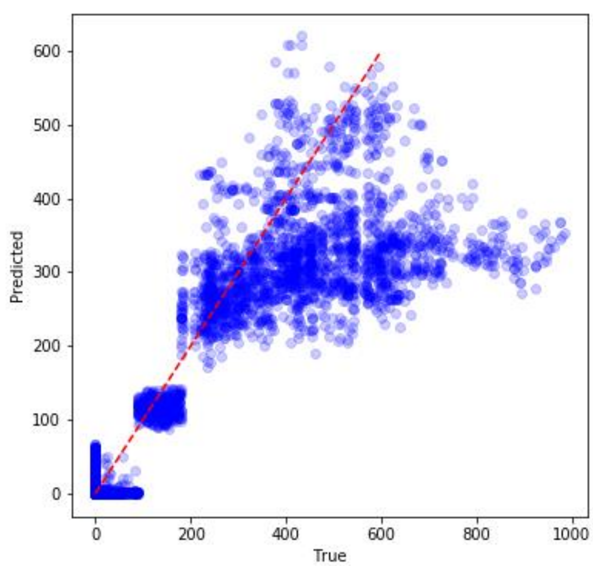

Table 11. Gradient Boosting: Aging in days (2018)

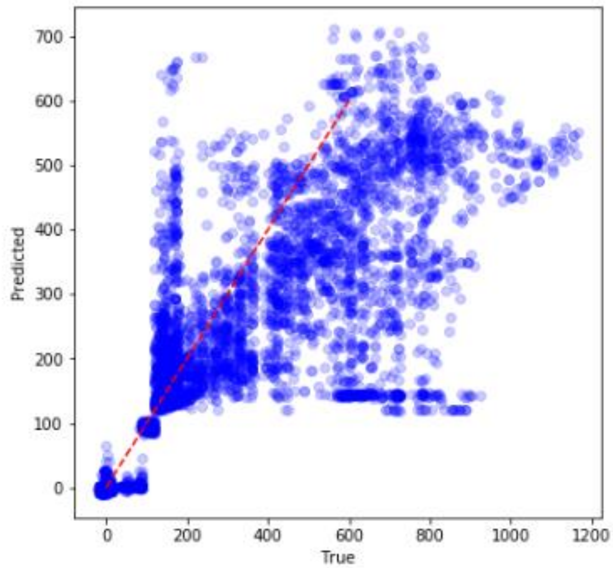

Table 12. Gradient Boosting: Aging in days (2019)

When projecting the prediction of aging in days into graph (see *Table 11 and Table 12)*, it is crystal clear that the dots in 2019 are sparse in comparison to the previous year. Let's say in 2018, a borrower's credit default is 400 days, but the machine learning predicts it would be 600 days. Another example, in 2019, a borrower's credit default is 200 days and it predicts ranging from 100 days to as nearly as 680 days.

Random Forest

It is a combination of number of decision trees –the fact that it is more robust than a single decision tree. Random Forest uses bagging technique in which it uses independent classifiers that can reduce variance and handle overfitting model.

	2018	**2019**
R^2	0.8763	0.8167
RMSE	25.2441	38.5210

Table 13. Score Results

Comparing to Gradient Boosting's score results, Random Forest has higher results by up to 6% and less average square of error by 5 days. We can say in 2018, we can predict the microcredit default among our borrowers is 87.63%; and the average square of credit default is 25.2441 days. While in 2019, the performances start to decline.

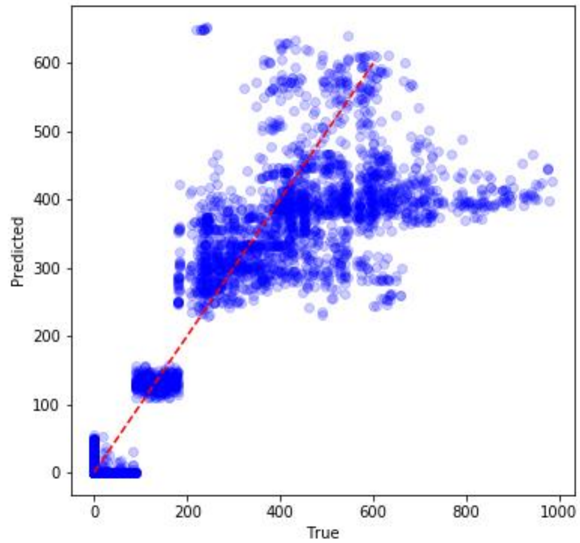

Table 14. Random Forest: Aging in days (2018)

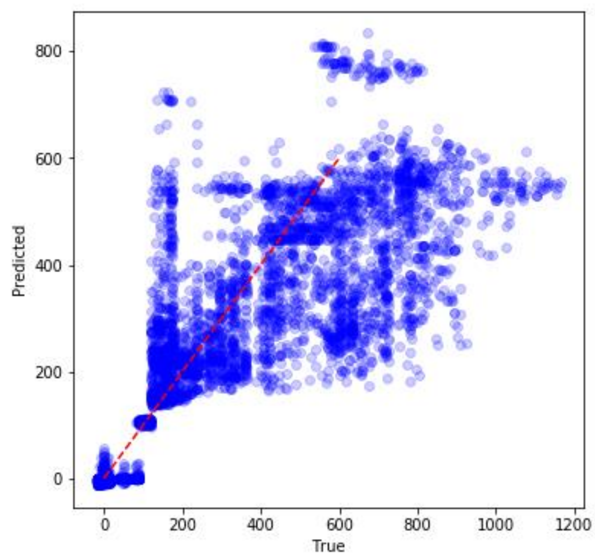

Table 15. Random Forest: Aging in days (2019)

In *Table 14 and Table 15*, we can see a huge difference of placement of the dots. Indeed, it does not fall along the red line. This means our data performances are getting weaker which may lead into inability to predict any future prediction of our borrowers' aging in days. As an example, in 2018, a borrower's aging in days is 200 days, the machine learning predicts ranging from 220 days to 350 days.

Small Business (UKM)

Overview and Data Exploration

As August 2019, we have nearly 20,000 members and the growth rate within a year is about 20%. We believe this number is satisfactory. As the number of clients grow progressively, on the contrary, we must continuously review our data performances if they are performing adequately in order to predict the credit default accurately. Thus, we could mitigate risk of nonperforming loan (NPL) in the future.

We will further investigate of our clients' backgrounds as well as the parameters that are used. In *Table 1*, most of our clients are homeowners. Followed by owned by family members and the rest is still renting.

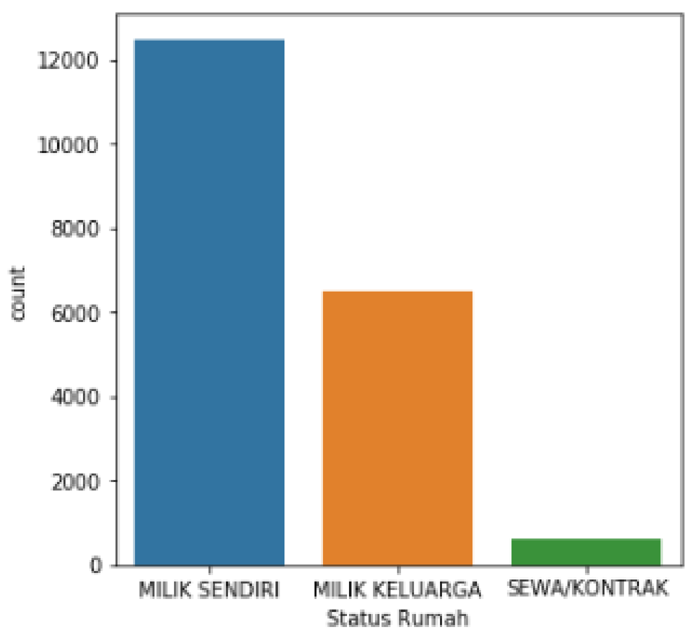

Table 1. House Ownership Status

In *Table 2*, majority of the clients are male, but it is sitting close to each other to the female ones.

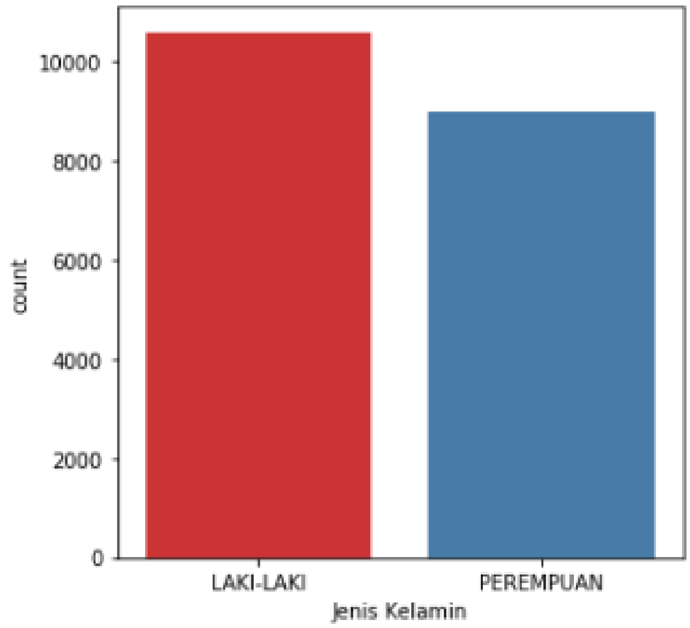

Table 2. Gender

In *Table 3*, the highest education of our clients is high school (SMA). Interestingly, the second highest is elementary school (SD), then followed by middle school (SMP). At the same time, there are some clients who hold a diploma (D1 & D3) and bachelor's degree (S1).

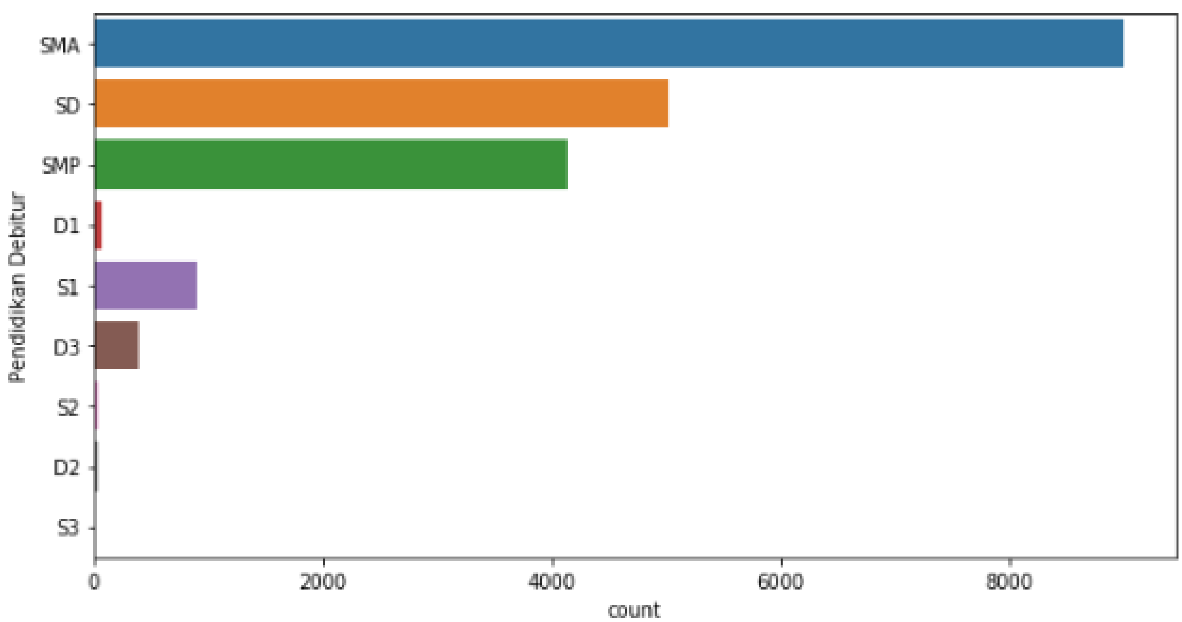

Table 3. Education Level

In *Table 4*, 80 percent of our clients are married. As nearly as 20 percent is still single, and the rest is divorced.

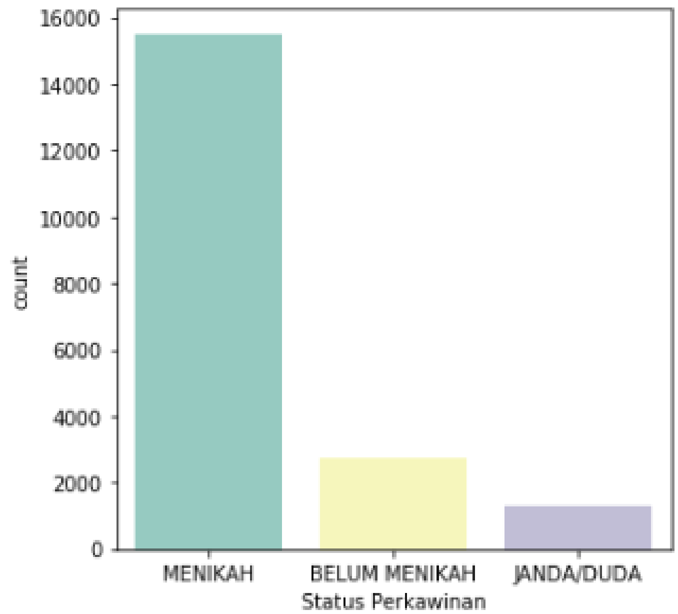

Table 4. Marital Status

In *Table 5*, this table has similar pattern to the micro's business sector. Most of our clients have trading business, followed by service business, and lastly home industry business.

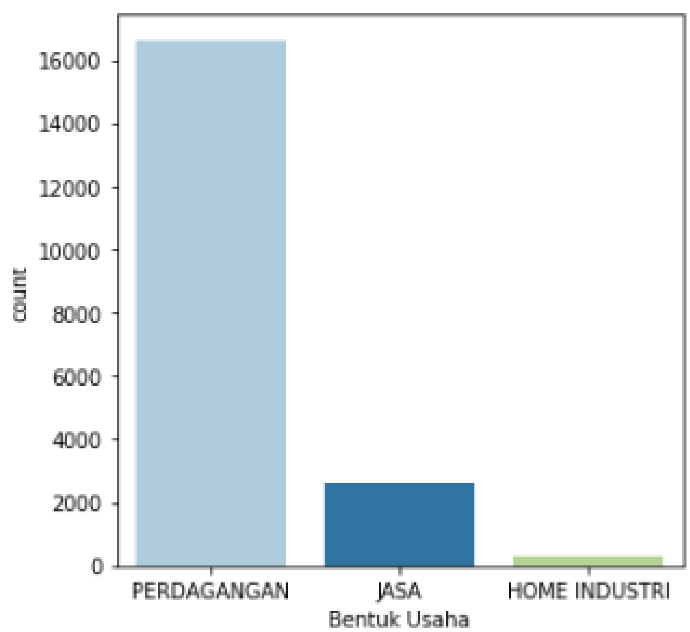

Table 5. Business Sector

In *Table 6*, a quarter of our clients are creating a new business and the rest of them are raising capital to their current business.

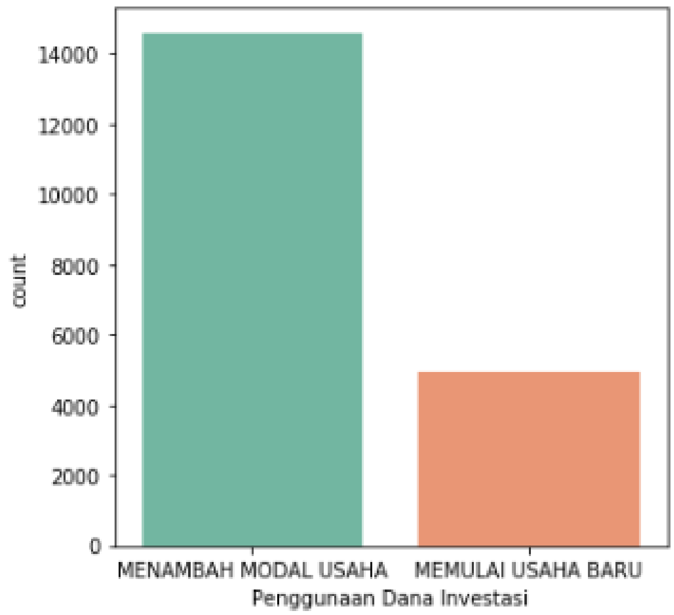

Table 6. Loan Purpose

In *Table 7*, majority of our clients do not have other loans from other banks. Yet, a little of them do.

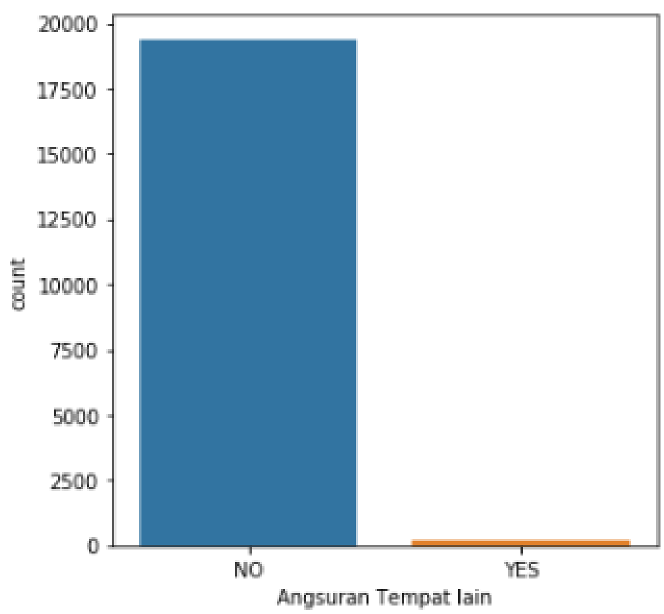

Table 7. Other Loan

In Table 8, we are performing a correlation graph to find in detail which variables are connected.

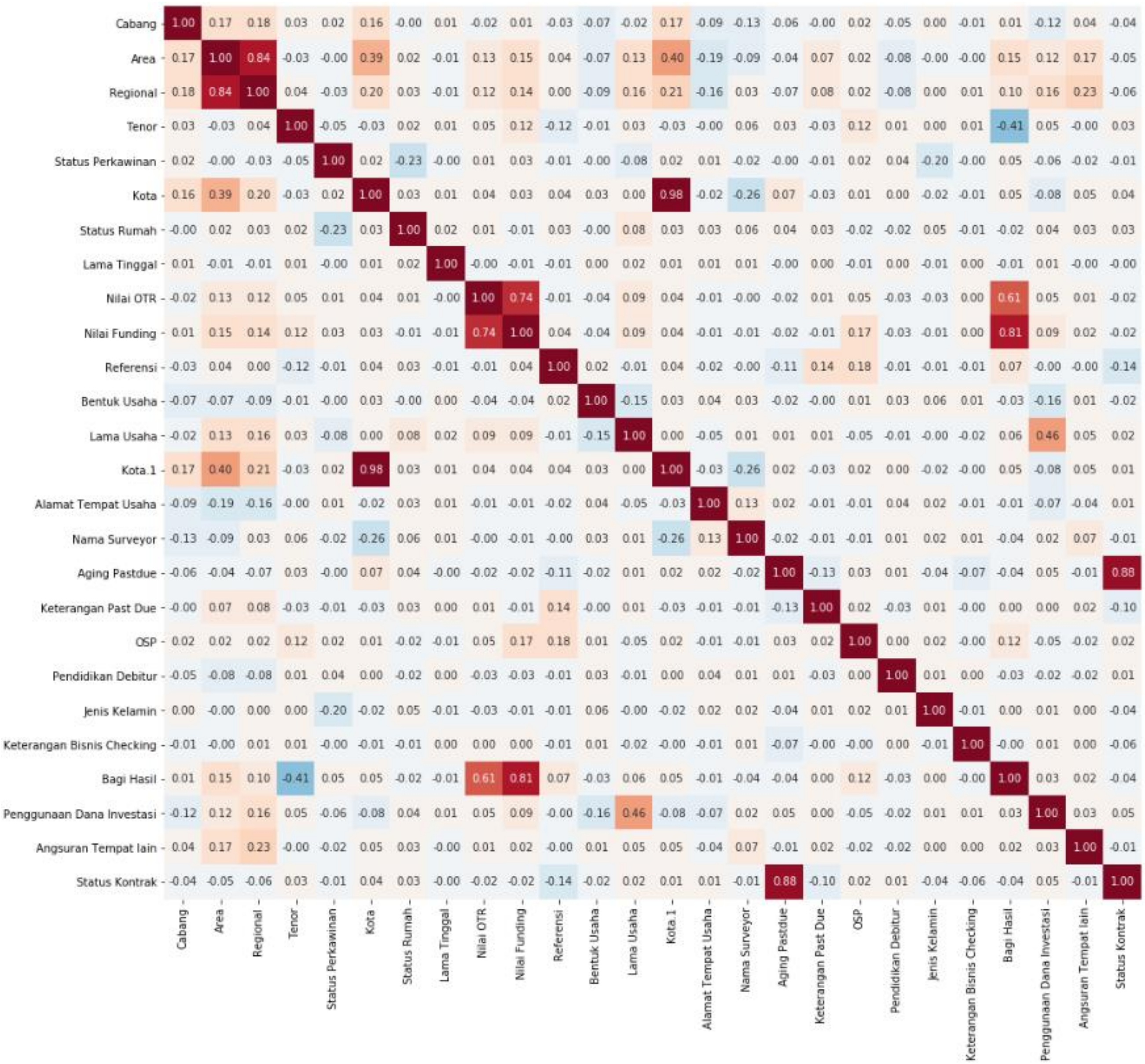

Table 8. Correlation Graph

In *Table 9*, we are ranking which parameter has strong correlation to the target, ‘Aging Past Due’.

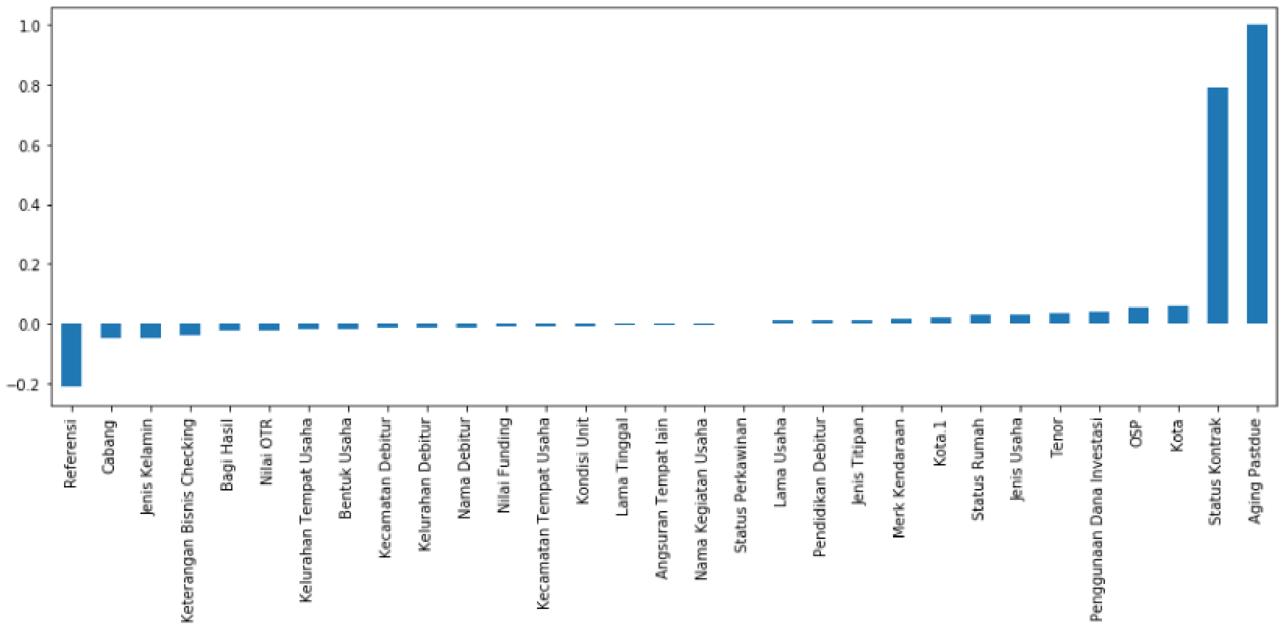

Table 9. Small Business Features Ranking

Gradient Boosting

We are surprised by the score results –it does not meet our expectation and standards as well. We can say, in 2018, 65.52% we can predict the credit default among borrowers and the average square of credit default error is 70.7721 days. In addition, there is a quite significant decline in 2019. The average square of credit error increases by 20.5711 days and yet the level to predict credit default falls under 60%. This score has raised our attention to improve and enhance our parameters.

	2018	**2019**
R^2	0.6552	0.5964
RMSE	70.7721	91.3432

Table 10. Score Results

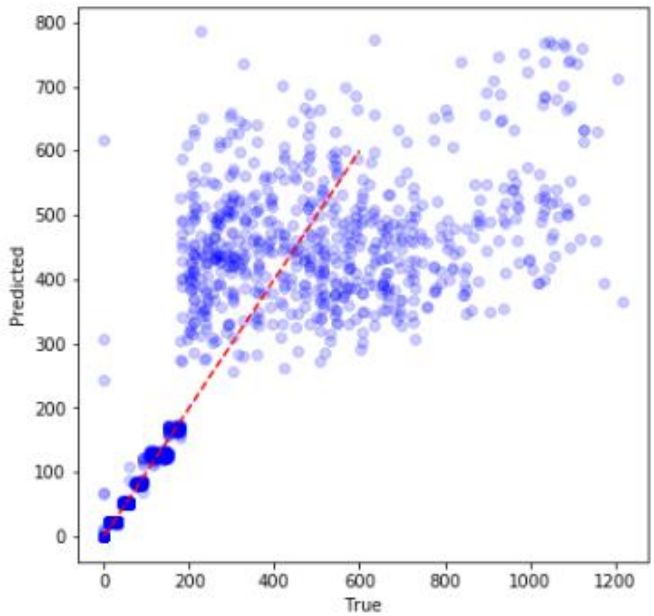

Table 11. Gradient Boosting: Aging in days (2018)

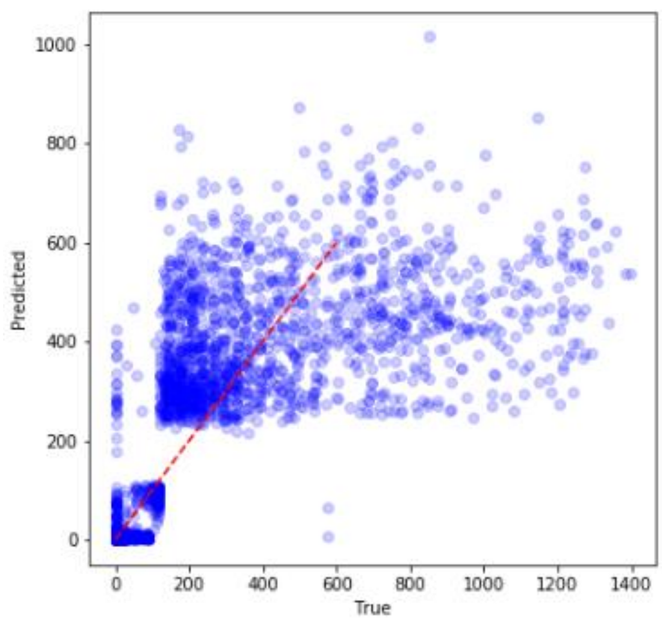

Table 12. Gradient Boosting: Aging in days (2019)

In *Table 11 and Table 12*, the dots are dispersed and do not fall under the red line. This graph shows that predicting credit default among borrowers are somehow uncertain. Not only we want to predict our clients' behavior in making their payments, it would be better if we could take things under control and yet we do would like to avoid nonperforming loan (NPL) in the future.

Random Forest

Comparing to Gradient Boosting, Random Forest's score results are better. However, this does not truly mean that we are in a good shape. In 2018, we can predict 80.38% credit default among borrowers with the average square error of 53.3808 days. Unfortunately, the performances start to decline within a year. We believe this is a critical point for us –the fact that, the number of borrowers are growing, our parameters are become less accurate to predict the credit default and the average square of errors increases. Thus, we must take action to prevent any further issues.

	2018	**2019**
R^2	0.8038	0.6854
RMSE	53.3808	80.6442

Table 13. Score Results

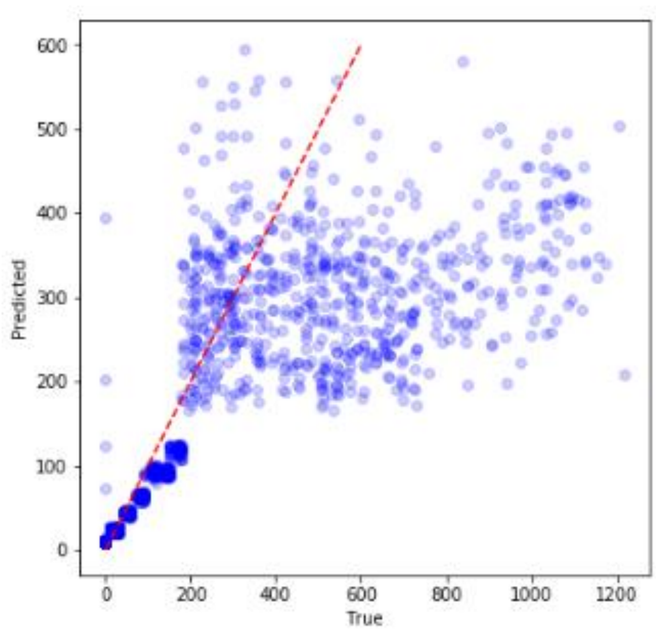

Table 14. Random Forest: Aging in days (2018)

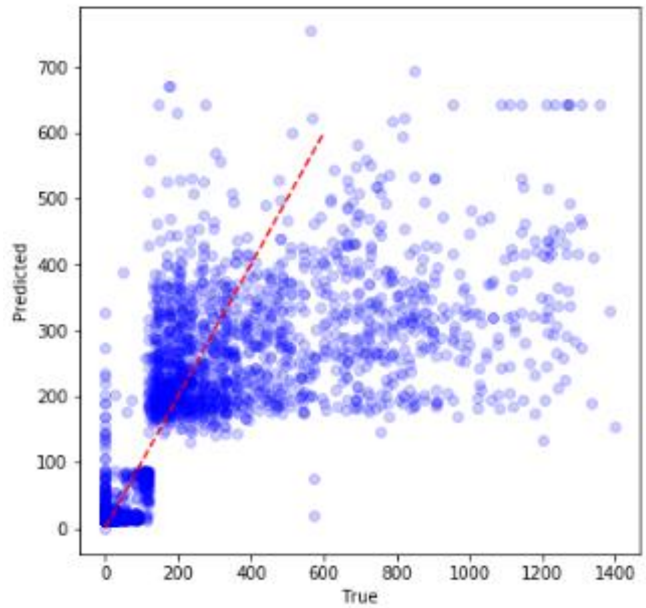

Table 15. Random Forest: Aging in days (2019)

Based on *Table 14 and Table 15*, the dots are a bit of everywhere. It does not show there is a certain pattern that they fall into. Both graphs give us an idea that our parameters need to be expanded in order to predict efficaciously. Therefore, by adding additional parameters will give us an idea which variable that influences the target the most.

Conclusion & Strategic Recommendations

The score results between the two sectors, microbusiness and small business (UKM), have given us a meaningful acumen to initiate changes within the features/parameters. Random Forest has better performance than Gradient Boosting in overall.

Microbusiness. Within a year, Random Forest's R^2 score decreases by 5.96% and the average square of credit default errors (RMSE) increases by 13.2769 days. Whereas, Gradient Boosting score decreases by 4.56% and the average square of credit default errors increases by 12.3059 days. In overall, the difference of R^2 score between the two algorithms is 1.4% and 0.971 days for the average square of error.

Small Business (UKM). Within a year, Random Forest's R^2 score decreases by 11.84% and the average square of credit default errors roses by 27.2634 days. Whereas, Gradient Boosting score decreases by 5.88% and the average square of credit default increases by 20.5711 days. In overall, the difference of R^2 score between the two algorithms is 5.96% and 6.6923 days for the average square of error.

The results do not meet our expectations, we believe there are three main points that we would like to address along with recommendations:

1. Data Incompleteness
2. Lacking Features/Parameters
3. Performance Decline

First, data incompleteness refers to missing data. Even though the percentage of missing data is relatively small, we may want to minimize in order to prevent any potential high risk of bias in the future. For example, when a data entrée fills out the borrowers' information, they should not leave it blank or enter 'N/A' (Not Available). We noticed that apparently our system allows that; thus, we need to set it as a mandatory field.

Second, lacking features/parameters. Even though there are more than 10 parameters are available, these parameters do not sufficiently explain what factors influence aging the most. We should add additional questions such as total annual income, monthly rent or mortgage, or any others that we believe might play contributions to the target. We may want to retest and retrain our data performance after we have collected a minimum of 10,000 data; so, we can get acuity of the new features.

A combination of data incompleteness and lack of features may generate lower performance results. It gives us ambiguity and forces us to speculate which factors that influence to the target the most due to poor correlation between the variables. It is critical to take this problem seriously –the fact that the average clients' growth is nearly 20% per year, and we need to take action to remedy these issues. Therefore, we can boost our model performance to minimize the risk of credit default among borrowers in the future.