A Case Study on Predicting Financial Crisis in Two Years

Nam Nguyen

A large national bank had a dataset of customers with credit history and would like the analytics team to build a model to predict individuals that would go into a financial distress, defined as incurring a 90-day past due delinquency in the next two years.

Objective: To build an optimal model to predict people who might get into a financial crisis in the next two years.

Summary

- This project explores several machine learning approaches to a classification problem: random forest, boosting, logistic regression, and stacking
- In any classification machine learning projects where models are compared to find the "best" model, each model needs to be considered by two elements: a key performance indicator by which to compare with other models, and the cut-off probability whereby the optimal KPI is achieved.
- F-beta score is a flexible model KPI that balances sensitivity (i.e. complement of false negative rate, revealing how effective the model predicts positive observations), and specificity (complement of false positive rate, revealing how efficient a model predicts positive observations). Beta denotes how many times sensitivity is more important than specificity, or how many times false positive is more detrimental than false negative.
- In our case, losses from a defaulted loan would eat up profits from two customers. Therefore, we need a KPI that reflects false negatives as twice more costly as false positives, or sensitivity as twice more important than precision. The beta value of F-beta score is thus set to 2. It is important in any classification modelling project to work out the judgment of relative importance between false positive sand false negatives.
- Models predict the possibility of an observation being positive. Each cutoff probability above which an observation is determined as positive will yield a different set of predictions, from which a different sensitivity, specificity, and F-score arrive.
- Sometimes (like in our case), there is a highest allowable cutoff probability, above which business is not viable, as the model would predict too many positives. In other times, there is a lowest allowable cutoff, below which too few positives are predicted for operations to be viable.
- Models are compared by the highest F-beta score possible, derived either by the cutoff probability associated with the maximum F-beta score, or the highest (or lowest) allowable cutoff probability.
- The “Learning” in machine learning also means what insights we as human learn from the model, not just what the model predicts.

Code details on GitHub: https://github.com/namofvietnam/machine_learning_classification_r

Pre-Modelling Judgments and Theoretical Framework

As all classification models considered in this project predict the probability each individual would incur a serious delinquency in the next two years, two judgments in evaluating the models' performance are (1) the cut-off probability above which an individual is likely to incur a serious delinquency and (2) the key performance indicator.

If the probability of an observation being "true" is higher than this cut-off probability, we shall assign that observation a "true" prediction. The cut-off, therefore, depends on the importance of false positives (i.e, predicting someone to incur serious delinquency when they would not) relative to false negative (i.e., predicting that someone would not incur delinquency, but they would). If a false negative costs more or has more adverse impacts than a false positive, for example, we would want the cut off to be higher to allow more negative (false) predictions. Assuming that the entire dataset represents the population, an average individual had a 6.68% chance of serious delinquency in 2 years. Therefore, the cut-off should not be much higher than this number, as the bank should not invest in those who have a higher chance of delinquency. To account for the importance of not incorrectly predicting someone as a potential delinquent (which will be discussed next), one standard deviation above the average probability of 6.68% can be allowed for the cut-off. The maximum chance of delinquency an individual can have to be categorized as a potential delinquent in two years is thus 8.35%.

Our objective of predicting serious delinquencies in the next two years presents an interesting problem of quantifying the relative importance between false positives and false negatives. It is intuitive that a false negative prediction would be of dire consequence, because if the bank accepts a loan based thereon, it would lose both monthly payments and principal. To the bank, bad debts might build up to a dangerous level that might result in a major debt crisis, albeit an unlikely event. From an individual perspective, a person's "chance" of getting into a financial crisis might be low but has serious impacts. To both individuals and the financial institute, therefore, the individual's financial crisis is a "tail" risk, very unlikely—at the tail of the distribution of events—but highly impactful.

On the other hand, a false positive would mean losing out on not only interest payments but also a profitable or loyal customer. Make enough false positives within the two-year horizon, and the opportunity costs to the bank would be close, if not equivalent, to the impact of a once-in-a-while bad debt crisis. The bank must consider both the probabilities and impacts of these outcomes to construct a risk portfolio, based on which to choose the cut-off probability.

To determine the relative importance of false negatives and false positives, we must calculate the expected costs of a delinquency and a lost customer. According to Expected Utility Theory, the expected cost or impact of an event is a product of the event's impact when it actually happens and its chance of happening. For simplicity, we assume that the probability of a serious delinquency in the entire population mirrors that in the entire dataset, approximately 6.68%, and that the average of existing debts in the dataset, about $2050, also mirrors the population mean. If the bank's credit cards charge an average annual percentage rate (APR) of 24%, compounded monthly for the next two years, $2050 would grow at an effective interest rate of 60.84% into $3297 (with a total accumulated interest of $1247). We assumed that non-delinquent customers would pay principal and interest at the end of the two-year period, which is also a conservative assumption because it implies that the bank has lost all interests and principals accumulated throughout that period due to delinquent customers. Therefore, the expected cost of delinquency, according to Expected Utilities Theory, is $3297 × 6.68% = $220. Meanwhile, we assumed that profitable customers on average would pay off the $2050 debt within one year (i.e., pay interest plus 32% of the balance per month), which would translate to $126.53 of interest payment. The expected cost in two years of a missed customer incorrectly identified as a potential delinquent would be $126.53 × 93.32% = $118.08. Therefore, false negatives— i.e., failing to predict a delinquency, causing the loss of principals and interests— would be twice as costly as false positives—i.e., predicting a customer as a potential delinquent, resulting in the loss of a profitable customer.

Because a simple comparison in error rate (or accuracy) is not meaningful, due to the skewed distribution of positive and negative cases and the difference in impacts between the two cases, we decided to use other metrics. Typically, either *sensitivity* (the complement of false negative rate) or *precision* (the complement of false positive rate), is used as a metric to compare models in situations where either false positives or false negatives are clearly more concerning than the other. *Sensitivity* measures how well a model covers all observed positives. Higher sensitivity means more true positives are predicted but also allows for more false positives and reduces false negatives. *Precision* measures how efficient a model makes true positive predictions out of all positive predictions it makes. Higher precision means the model predicts more true positives with fewer positive predictions, but more false negatives are possible because more negative predictions are made, while reducing false positives. Because false negatives would cost twice as false positives in our case, as discussed above, we need a metric that combines *sensitivity* and *precision* in such a way that sensitivity is twice more important as precision, to reduce false negatives. Such a metric is the F-beta score, with beta value set to reflect the relative importance of sensitivity versus precision. When sensitivity is 2 times as important as precision, beta is set at 2.

The advantage of using F-beta score as the metric is two-fold. Firstly, the optimal F-beta score is a single value comparable across models. Secondly, readily calculated as components of the F-beta score, sensitivity and precision can provide complementary information about model performance. Managers with different viewpoints on market risk and internal risk tolerance can rely on sensitivity and precision to construct their own version of balancing. However, without a theoretical constraint on cut-off probability, F-beta scores tend to have an inverted U-shaped curve as the cut-off probability runs from very low to very high, which can be tricky when comparing model performance. F-beta score is only meaningful in performance evaluation before or at the inflection point of the U-shaped curve, which is often the optimal point.

The disadvantage of this approach is that the set beta is based on the judgment of the relative importance between false negatives and false positives, which in turn requires various assumptions. Different assumptions will support different judgments and choices of beta. Still, the process would remain the same for any choice of beta.

Models will be compared by their best value of the F-beta score, arrived at by applying different cut-off probability, from 0 to 1, passing 8.35%, which is the mentioned constraint according to the bank's risk management policy. We shall call this the *highest allowable cutoff*. The highest F-score a model can achieve, however, may correspond to a different cutoff probability, which we shall call the *theoretical optimal cutoff*. We shall compare model *performance* based on the unconstrained highest F-score and make a note on the *practicality* of each model, based on the highest F-score achieved with the highest allowable cutoff.

Data Management

Summary of the Data Set

The dataset has 150,000 rows of data. Except for the binary-outcome target named SeriousDlquin2yrs and an identification number "X", all other variables were numerical (integer or percentage).

Data management

The target variable—whether an individual goes into a serious delinquency in the next two years—was converted into a logical variable (i.e., consisting of "true" and "false"). The ordering variable X was eliminated.

Missing values required a cleaning process guided by theory. Nassim Nicolas Taleb[1] identified two areas for variables, mediocristan and extremistan. In mediocristan, values of a variable do not deviate significantly from the mean, and an additional observation would not change the standard deviation significantly. Examples include biometrics and physics-bound variables. In extremistan, values of a variable deviate widely from the mean, and an additional observation can significantly alter the standard deviation. These variables often have scalability, for example revenue, income, virus spread. In our data set, age and number of dependents likely fall into mediocristan. People have life expectancy and in general do not differ widely on the number of dependents. All fields in age were filled, but there was an instance of age 0, which we treated as a missing age value. Because mediocristan variables typically follow distributions found in nature, we fill the missing number of dependents values based on its existing distribution in the original data set. We wanted to separate numbers of dependents from financial variables in imputation, with the assumption that dependency situations can occur due to other factors than finance, but are still associated with age. Therefore, we used bootstrapping sampling to fill numbers of dependents in a separate data frame consisting of only age and number of dependents. This way, numbers of dependents are distributed according to age and vice versa.

Being a convenient and liquid representation of value, money is highly scalable. Thus money-related variables, such as monthly income, likely fall into extremistan. As extremistan variables do not conform to normal distributions, we could not fill missing monthly income

[1] Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*.

values with a distribution. Predictive models are typically better at detecting possible relationships and associations in situations where a distribution of a target variable is unstable. We used a predictive model called "Random Forest" to predict missing monthly income values, based on other variables (more on Random Forest later).

Looking at the dataset, we noticed that debt ratio, the ratio of credit card debt amount over monthly income, had abnormal values (e.g., 477, 5710, instead of .4, .6, etc.) wherever monthly income is missing, because this ratio depends on monthly income. These abnormal debt ratios are most likely monthly debt payments, unconverted to debt ratios due to missing monthly income values as the denominators. Therefore, we kept the abnormal debt ratio values, because these would be used to estimate debt ratio after missing monthly income values are imputed. For the same reason, debt ratio cannot play a part in the prediction model to fill out missing monthly income values. The presumably monthly debt payments (not converted to debt ratio beforehand) would then be divided by the imputed monthly income to arrive at more sensible debt ratios.

After age and numbers of dependents were cleaned, we filled in missing values for monthly income using a random forest model with other variables as predictors, except for debt ratio. As another random forest model would also be built to predict the main target, more on this model will be discussed later. With the filled monthly income and the preserved amounts of debt payment, we could calculate the correct debt ratio.

The cleaned dataset was partitioned into a training set of 60% and test set of 40%, with a “seed” set at 1234, which stabilized the random process so that the same samples can be replicated at different times or by different teammates. This seed did not result in any abnormal partitioning of data (e.g. missing levels in variables). Besides the partition of data into training and test set, the 1234 seed was also set for all other random processes in this project.

Model Construction, Interpretation, and Evaluation

Random Forest

Theory and construction

A random forest is the aggregate of a large number of decision trees for classification. To build an understanding of random forest, we should first understand classification trees.

A Classification Tree Model determines a decision-making tree based on the position of the observations in a dataset on a certain “dimension” or variable. (There can be more than one predictor variable, and more predictors would mean more dimensions on which the branching or “splitting” of decision is possible). Following this decision tree, a person or machine is likely to make a correct prediction on the target variable of the observation, which has certain attributes on certain dimensions that fall into a certain position in the tree map.

Just as a forest consists of many trees, the Random Forest model aggregates the results of a large number of classification trees, in this case 500. The effect of this aggregation is two-fold. Firstly, aggregate predictions are typically more accurate than predictions of any single tree.

Secondly, by theory, the random selection of features on which trees split decisions reduces the likelihood of overfitting.

Interpretation of results

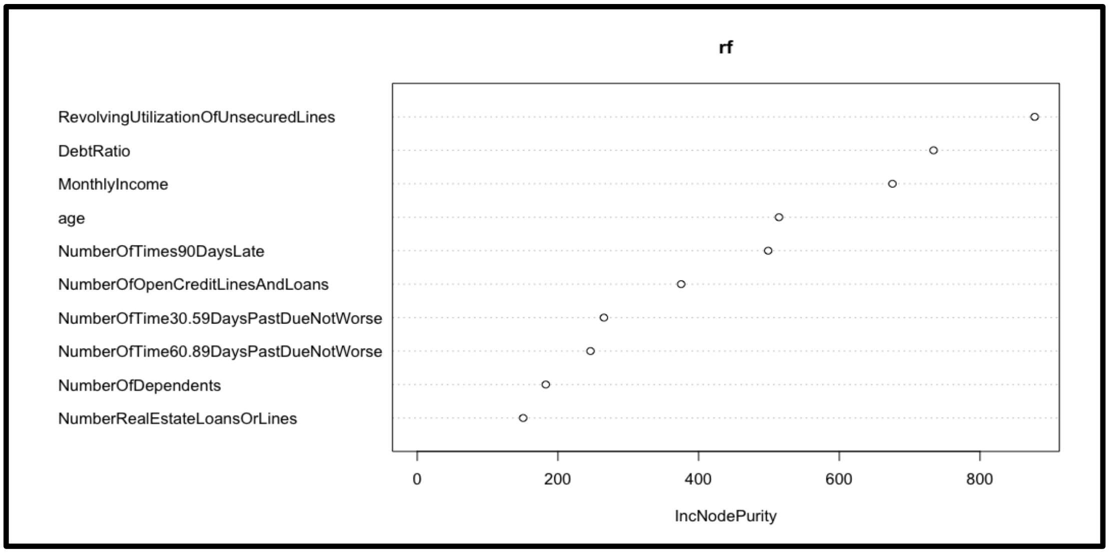
Model performance

Regarding the model's performance, had it not been for the bank's constraint, given the judgment that false positives are twice as costly as false negatives in the competitive environment, the highest F-score that the random forest achieved was .52 at the cut-off probability of 13.3%. This means that for a higher risk tolerance than the 8.35% constraint, or in a highly competitive environment where obtaining profitable customers is important, and with the expected cost of a serious delinquency twice that of a profitable customer, the bank can tolerate someone with a 13.3% chance of serious delinquency, according to this model. At the riskiest allowable cutoff probability of 8.35%, the random forest model's F-beta score was .51. However, to achieve .01 increase in the model's F-score, the bank would have to increase its constraint by 59%, which might be too risky.

At the optimal cutoff, sensitivity was .65 and precision of .29. This means that if the bank were to accept loans for individuals with serious delinquency risk up to one standard deviation higher than the population mean, on average the model could identify 65% of cases where these individuals would actually incur delinquency using 29% of its positive predictions—that is, at the expense of 71 individuals misclassified out of every 100 individuals labeled as potential delinquents. The highest allowable cutoff probability resulted in sensitivity of .74 and precision of .23.

Learning Insights

While it is impossible to lay out the decision-making mechanism each tree in the forest presents, we can learn the relative importance of each variable in the random forest's aggregated decision. Below is the chart ranking variables by their importance in the forest's decision:



In a random forest model, “Gini impurity” or “node impurity” is the measure of how frequently a randomly chosen feature would be mislabeled if it were randomly identified based on the random distribution of observations in that variable. *Im*purity increases with randomness, so a variable that increases in purity during the random selection process of the forest shall prove to be more important. According to the random forest model, based on increase in purity, revolving utilization of unsecured lines, debt ratio, and monthly income are the three variables with highest increase in purity and thus of greatest importance. Numbers of late payments of several durations were next in line, and number of dependents and number of real estate loan or lines came last.

A note of caution: The Gini Impurity method for variable importance has a bias: it gives more importance to variables that have many levels (most of all continuous variables), which have more possible split points than variables that have fewer levels. Therefore, the keyword "relative importance" should be interpreted with regards to the number of levels in each variable (for example: relative importance among continuous variables; relative importance among 5-level variables, etc.)

Boosting

Theory and construction

Boosting is relatable to random forest in that it also combines decision trees, but vertically instead of horizontally. The error of the prediction of the first tree in the chain would be predicted by the next tree, whose error would be predicted by the following tree, and so on, for a large number of trees (tens, hundreds, or thousands) in the chain. Because errors were so thoroughly predicted and accounted for by the large number of trees, the predictive accuracy of the entire chain is boosted. When the number of member trees is too large, this accuracy would be too good on the training set to be obtained when the model is applied to new data. Therefore, an optimal number of trees would need to be determined. We cannot try different numbers of trees to predict the test data set and compare the performances of different numbers of trees, because doing so would mean using the test data set in training and violate the integrity of training and test datasets in machine learning. However, we can use a technique called cross-validation, wherein the training set is be partitioned into a number of folds; each fold would be used to test the optimal tree number of the model built on the combined data points of the rest of the folds, and the optimal tree number of these trees (one for each fold) would be aggregated, as an estimate for the best number of trees for the model built on the entire training set. For the problem at hand, we started with 1000 trees and used cross-validation to find the optimal tree number of 620.

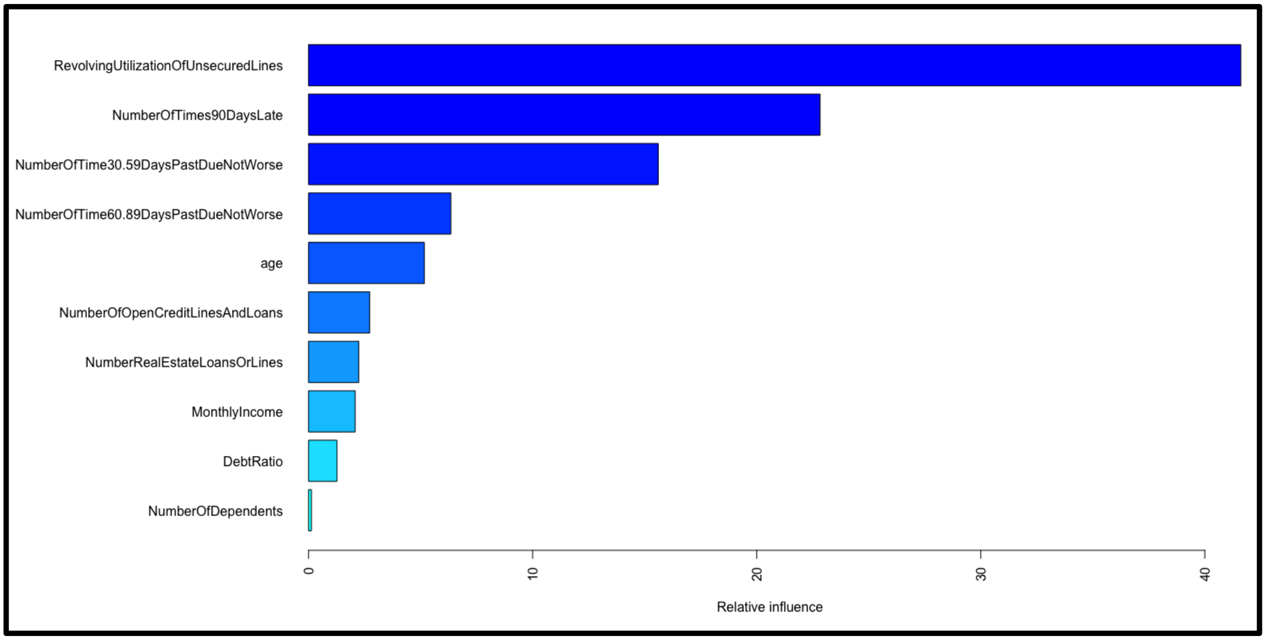
Interpretation and evaluation

Model performance

The boost model achieved the highest F-score at .52, rivalring that of the random forest, but at a lower optimal cutoff probability of 10%. Sensitivity was .64 and precision .29, fairly similar to those of the random forest at its optimal cutoff. At the highest allowable cutoff, the F-score was close at .515, *sensitivity* = .69, *precision* = .25. In practice, with the boost model, the bank does not need to raise its tolerance bar by 59% like in the random forest model to achieve the same level of balance between expected costs of a delinquency and a lost customer. Instead, with the boost model, the bank can increase the bar by about 20% if it wishes to achieve a better precision in predicting potential delinquency than at the 8.35% cutoff.

Learning Insights

Like random forest and any tree-based classification models, boosting also determines the relative influence of each variable in the classification. While in random forest, we used the increase in purity or decrease in impurity to determine relative importance of variables, in boosting we can use a similar concept called permutation. This method shuffles (permutes) the predictors to introduce randomness and calculate that variable’s prediction error (similar to impurity). With a similar method of determining relative importance to that used in the random forest, we arrived at the following bar chart of variables sorted by importance.



Both the random forest and boost model agreed that revolving utilization of unsecured lines are most important in predicting whether a person would go into financial crisis in the next two years. However, the boost model went on to list late payments by 90 days, 30 - 59 days, and 60 - 89 days as next in importance. Next in line were age, number of open credit lines, number of real estate loans or lines, and monthly income. Debt ratio, while second in importance in random forest, was “second” in unimportance in the boost model. Number of dependents was determined as not significant in both tree-based models.

Logistic Regression

Theory and construction

Logistic regression is relatable to linear regression in that both look for linear relationships between various factors and the target. However, instead of a numerical prediction for the numerical target itself like in linear regression, logistic regression predicts binary outcomes by analyzing how changes in one variable correlates with the odds that the positive case of the binary target happens.

Most regression models seek to fit variables into a best-fit line (be it linear or logistic), regardless of whether the variable fits in that line model. A statistically non-significant variable’s distance from the model might be accounted for, but the variable itself would still be included in the prediction, if an elimination rule is not specified for the model. Our team used stepwise elimination to strip the logistic regression model of statistically non-significant variables, one after another, until statistically significant variables are left, thereby accounting for overfitting to the training set due to statistically non-significant elements. This process rejected revolving utilization of unsecured lines as statistically non-significant in predicting the probability of someone incurring a serious delinquency. This posed a stark contrast to the random forest’s determination, wherein revolving utilization of unsecured credit lines is most important in predicting the same target.

Interpretation and evaluation

Model performance

The logistic regression model achieved the highest F-score of .35 at the optimal cutoff of 8.79%. Note that this optimal cutoff probability is not too far from the bank's constraint. While the logistic regression model did not perform as well as the random forest model in balancing the weighted importance of false negatives and false positives, it might have practical values because the bank might not have to adjust the constraint by too much to take advantage of the optimal performance of the logistic regression model. At the optimal cutoff, the logistic regression model could detect 48% of those who actually incurred a serious delinquency and left 83 individuals misclassified out of every 100 labeled as potential delinquents. At the riskiest allowable cutoff probability of 8.35%, the logistic regression model's F-beta score was .35, with a sensitivity of .50 and precision of .15.

Learning Insights

The table below lists the logistic coefficients and their statistical significance determined by the logistic regression model.

	Estimated Coefficient (β)	***P-value***	
(Intercept)	-1.31E+00	< 2e-16	***
age	-2.91E-02	< 2e-16	***
NumberOfTime30.59DaysPastDueNotWorse	5.19E-01	< 2e-16	***
DebtRatio	-1.28E-04	0.01844	*
MonthlyIncome	-3.97E-05	< 2e-16	***
NumberOfOpenCreditLinesAndLoans	-8.91E-03	0.00685	**
NumberOfTimes90DaysLate	4.89E-01	< 2e-16	***
NumberRealEstateLoansOrLines	8.71E-02	1.59E-10	***
NumberOfTime60.89DaysPastDueNotWorse	-9.79E-01	< 2e-16	***
NumberOfDependents	1.03E-01	< 2e-16	***

*Significance codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1*

To interpret this table, each coefficient β represents the change in the odds that a positive case happens by an average multiplication factor of e^{β}, for every unit change in the associated variable, all other variables equal. Accordingly, on average, all else equal, the odds that an individual incurs a serious delinquency in the next two years would:

a. decrease by 3% for every 1 year increase in age,
b. increase by 68% for every additional time 30 - 59 days past due,
c. decrease by .01% for every 1% increase in debt ratio,

d. decrease by 4% for every $1000 increase in monthly income (.004% for every dollar increase in monthly income),
e. decrease by 1% for every additional open credit line or loan,
f. increase by 63% for every additional time 90-day late payment,
g. increase by 9% for every additional real estate loan or line,
h. decrease by 62% for every additional time 60 - 89 days past due,
i. increase by 11% for every additional dependent.

Findings (a), (b), (d), (f), (g), (i) converge with intuition and provide a numerical change in odds, which is often helpful for decision-making scenarios that involve calculating expected impacts of events, as described in the pre-modeling framework section. Findings (c), (e), and (h) are somewhat counterintuitive. Findings (c) and (e) statistical significance are relatively lower than other findings ($p_{(c)}$= .02, $p_{(e)}$ = .007, compared to much lower p-values observed for other statements), so the prediction error might still play an albeit small role in this finding. However, (h) is both statistically significant and counterintuitive and requires further investigation.

Stacking Models

A battery of models is better than a single one, thanks to both computational power and the aggregation effect. For combinations of models, we use a technique called stacking, wherein a manager model predicts the target based on both the existing predictors and the predictions of three other models. To figure out the best-performing manager model for stacking, we let each stack test with a subset of the test dataset and compare the managers' F-scores on these test rounds. Because a randomly sampled subset of the test set is essentially different from the test set itself, we could avoid, to some extent, leaking the test data to the training process. The best-performing stack would then predict the original test set, and its performance will again be compared with that of the individual model.

Training Manager 1: Logistics Regression

When a stepwise logistic regression model based on the stacked train set was used to predict the target in a subset of the stacked test set, the highest F-score was .49 at a surprisingly low optimal cutoff of .05%. More interestingly, this low cutoff resulted in a healthy sensitivity of .61 and precision of .29. At the bank's highest allowable cutoff, *F-score* = .36, *sensitivity* = .34, *precision* = .42.

Training Manager 2: Random Forest

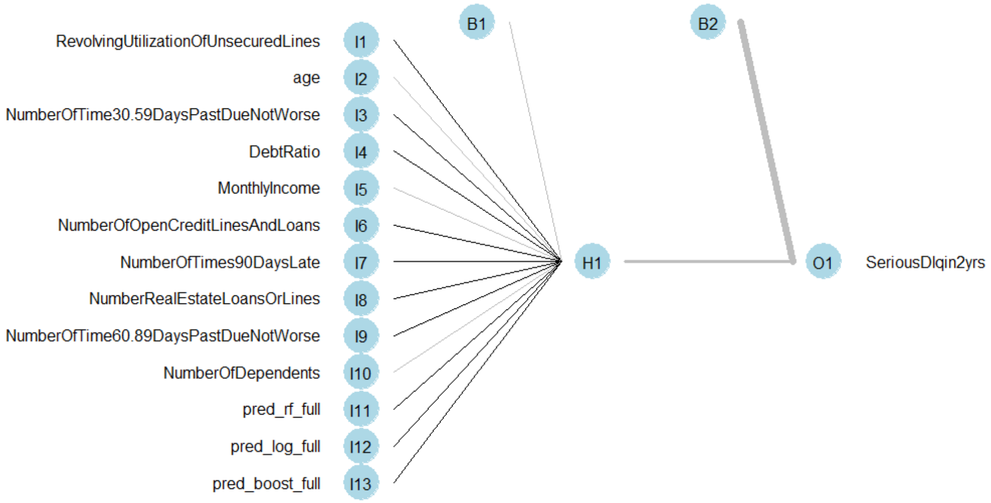
When a random forest model based on the stacked train set was used to predict the target in a subset of the stacked test set, the highest F-score was .48, also at a low optimal cutoff of

0.11%. This low cutoff resulted in a healthy sensitivity of .60 and precision of .27. At the bank's highest allowable cutoff, *F-score* = .41, *sensitivity* = .40, *precision* = .44.

Training Manager 3: Neural Network

The appeal of neural networks is that it imitates how neurons in the brain work, with nodes that connect layers, just as synapses connect concepts to one another. These layers consist of an input layer, output layer, and hidden layers in between. Just as neurons that receive more signals become stronger over time, nodes that receive more signals become more significant in predicting the adjacent layer. This approach allows learning from examples. Neural networks often predict with unexpected accuracy and are the basis of more advanced deep learning methods. However, because the in-between layers are hidden, neural networks might give little learnable insight into the workings of phenomena. Because we have tried several learnable models, the reputation of neural networks's performance in prediction made it seem appealing as a manager model in stacking.

Running neural networks models with one layer of different unit sizes several times, we found that the single-layered model, regardless of unit size in the layer, did not make any positive prediction, which was as good as a best guess based on the dominant negative cases. Therefore, we rejected the single-layered neural networks model as a manager. However, unlike with a best guess approach based on dominant cases, one can peek into the nodes of the neural network model to see which "neurons" have been activated more during the "learning" process.

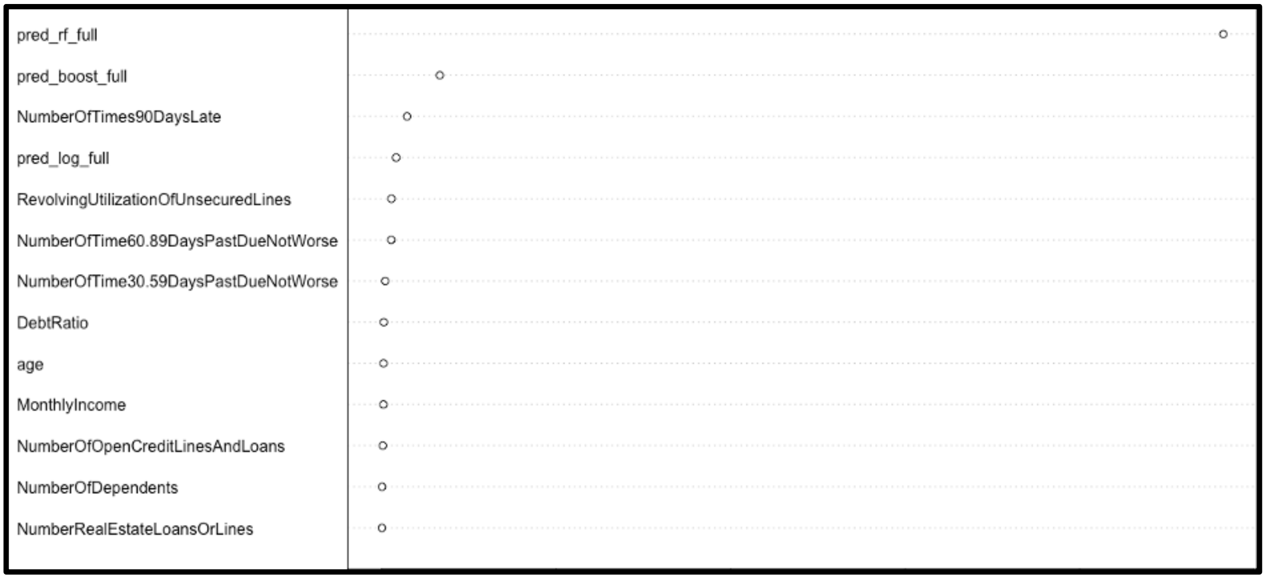

B1
B2
RevolvingUtilizationOfUnsecuredLines
I1
age
I2
NumberOfTime30.59DaysPastDueNotWorse
I3
DebtRatio
I4
MonthlyIncome
I5
NumberOfOpenCreditLinesAndLoans
I6
NumberOfTimes90DaysLate
I7
NumberRealEstateLoansOrLines
I8
NumberOfTime60.89DaysPastDueNotWorse
I9
NumberOfDependents
I10
pred_rf_full
I11
pred_log_full
I12
pred_boost_full
I13
H1
O1
SeriousDlqin2yrs

This graphic is the output of the one-unit single layer model. While we ran the model with several unit sizes for that one layer, the simplest model (*unit size* = 1) was chosen for visualization purposes. Bolder lines denote stronger connections with the single hidden layer, but that is as much as we could learn from the model.

Choosing manager and "stacking" up against the single models

Although the remaining managers had similar performance at their optimal cutoffs on their similarly high max F-score, when the practical highest allowable cutoff probability was considered, the forest-managed stack performed slightly better on F-score (weighted balance between false positives and false negatives), sensitivity (effectiveness in detecting positive cases), and precision (efficiency in detecting positive cases). When this random-forest-managed stack went on to predict the stacked test, it achieved its maximum F-score of .47 at the optimal cutoff of 0.13%, with *sensitivity* = .60, *precision* = .26. Because this optimal cutoff might be too low for the bank to implement, we considered the highest allowable cutoff. At this cutoff, the stack performed with *F-score* = .40, *sensitivity* = .39, *precision* = .42.

The chart lays out the relative importance (measured by increase in purity) of variables according to the forest-managed stack.



The fact that there were predictions by a random forest inside a stack managed by a random forest might explain the bias where the predictions of the singular forest were considered with such disproportionate importance. However, another hypothesis could be that the singular random forest had done such a good job in predicting the target that the random forest stack recognized that the singular forest's predictions increased in purity when more randomness was introduced. Other than that, predictions by the other two models were assigned high importance. The first variable in the list that was not a result of predictive modeling was numbers of times 90 days late, which is intuitive. The forest stack also generally ranked late payments of various durations as more important than the rest of the variables, which agreed with the singular *boost* model.

Recommendations and Discussions

If the assumption holds that the expected cost of false negatives is twice as high as that of false positives, then out of the four main models, both the singular random forest model and the boost model achieved the highest value of F-beta score (with *beta* = 2 representing the mentioned weights). However, the boost model's optimal cut-off probability (associated with its F-score) is closer to the highest allowable cutoff of the bank. At a theoretical level, the bank could choose either the boost model or the random forest model, depending on its risk tolerance. Depending on the preliminary judgment of the relative weights of false positives and false negatives, the optimal model might not be the same, but the process to determine such model is similar.

When the highest allowable cutoff probability is considered, the boost model still delivered the best balance of false negatives and false positives. However, the random forest model was most effective at identifying positive observations (*sensitivity of* .74), while the forest-managed stack was most efficient in making positive predictions ($\textit{precision}_{rf\_stack}$ = .42). Depending on the bank's risk tolerance, based on internal situation and external market condition, the highest allowable cutoff might change, and the weight between sensitivity and precision might also change.

The current study had some limitations. The most critical limitation is that, because the team attempted to uphold the weighted balance between sensitivity and specificity, models' performances were compared in relation to one another without a more stable benchmark applicable to all models for a more objective elimination rule. While the balance was upheld, both sensitivity and specificity were compromised to some extent, or discrepancy between a theoretical optimal cutoff and a practical tolerance point might arise (such as in the case of the stacked models, in which the "optimal" cutoffs were too low to be implemented in practice). Future studies might benefit from a superior metric of model performance—one that would consider both the discussed weighted balance *and* the accuracy of the models. Additionally, it is worth pointing out that sensitivity and specificity also depend on several other objective and subjective judgments, all of which can change. Therefore, this project should be treated more as a framework on methodology, and its recommendations ad hoc measures, rather than as a one-size-fit-all answer to all banks at all times.