Capstone Project – Loan Default Prediction
Final Submission

Prepared for:
Great Learning - MIT
Applied Data Science – Jan22

By:
Richard Sim

April 22nd, 2022

Execute Summary

Key Takeaways

Decisions to extend credit to potential customers are complex and were made heuristically until a few decades ago, prone to human error and bias. This led to risky and potentially catastrophic decisions for the bank if they were not careful. Therefore, the ability to accurately predict the likelihood of default is important. To do so, this analysis will leverage the Home Equity dataset. The dataset has various characteristics concerning the number of applications, the number and type of attributes, the extent of missing values, and different ratios for bad loans/good loans. The dataset contains 5960 loan applications. About 20% of them defaulted on their loan. Except for LOAN, all the features are missing some values. About 69% applied for a loan to consolidate their debts, and about 42% have a job other than the main ones. There's also the fact that about 70% of loaners paid on time, and 11% faulted once, while 4%, 2% and 1% faulted thrice, four times and five times, respectively.

When looking at the relationship between the categorical and numerical features, many show no to little correlation. However, some feature pairs show a hint worth further exploration. For example, the relation between the number of scathing reports and delinquent credit lines is positively correlated with defaulters. There are also more recent credit inquiries for people who eventually defaulted. Loaners who defaulted also have a debt-to-income ratio much higher than those who repaid. There seems to be a relationship between the number of recent inquiries with people wanting to consolidate their debts. Whereas with home improvement, recent inquiries are lower. Because of this, the number of the existing credit line is also slightly higher for people wishing to consolidate their debts. The amount due on the mortgage varies greatly depending on the profession. This can be due to the difference in average salaries between types of employment. However, the debt-to-income ratio is the same regardless of the profession. There also seems to be a relationship between the type of employment and defaulters. Sales and self-employers tend to be more susceptible to default. After analyzing the relationship between all continuous features, only the amount due to the mortgage with the property's value seems to be correlated.

The analysis will consider these five classification methods: logistic regression (LR), decision tree (DT), random forest (RF), k-nearest neighbours (kNN), and support vector machine (SVM). The success metrics will be using precision-recall curves to obtain the classifications accuracies, precisions and recalls. The analysis showed that the SVM model performed substantially better than the other models, with an average overall accuracy of 96%. This was after hyperparameter tuning using the grid search method. The analysis has shown that the financial attributes of an applicant are more relevant than social, personal or employment attributes for accuracy. More precisely, the debt-to-income ratio, the number of delinquent credit lines and the age of the oldest credit line are among the essential features to consider when looking at an application.

Next Steps

There is still room for improvement. The first is the imbalance of the data. For this analysis, only standardization on all features was performed. It would be beneficial to perform a combination of normalization and standardization depending on the feature type to yield better performance from the models. There's also the problem of collinearity between independent features. As highlighted in Milestone I, the value of the property and mortgage amount were highly correlated at 0.88. It would be worth investigating the performance by dropping one of the two features. There are other methods,

such as under/oversampling the dataset. There's also the possibility of feature engineering, finding impactful features from combinations of other existing features. Then, other models are worth exploring, such as Naive Bayes and stochastic Gradient Descent. There's also the possibility to optimize the hyperparameters and look at different tuning techniques, such as random search or Bayesian optimization.

To make the best of the solution, it would be highly recommended to apply the same steps as performed in the analysis. This means using a standardization on the raw dataset and treating outliers by replacing them with the mode and median (for categorical and numerical features) if they're outside the interquartile range (lower than the 25th percentile and higher than the 75th percentile). It would also be essential to use the same hyperparameters during the tuning. These would be the hyperparameters:

```
Fitting 5 folds for each of 25 candidates, totalling 125 fits
SVC(C=1, class_weight={0: 0.2, 1: 0.8}, gamma=1, random_state=1)
```

And using an 'f1' scorer is more appropriate for imbalanced datasets.

The best classifier is SVM, with an average accuracy of 96%; the top 10 features are ranked from 1 to 10; 1 is the most crucial feature.

Table 1: Top 10 features for SVM

Feature	Rank
LOAN	1
MORTDUE	2
VALUE	3
YOJ	4
DEROG	5
DELINQ	6
CLAGE	7
NINQ	8
CLNO	9
DEBTINC	10

However, if the customer wishes to leverage another classifier, the importance of the features changes. Here's a graph showing the rank of each feature depending on the model.

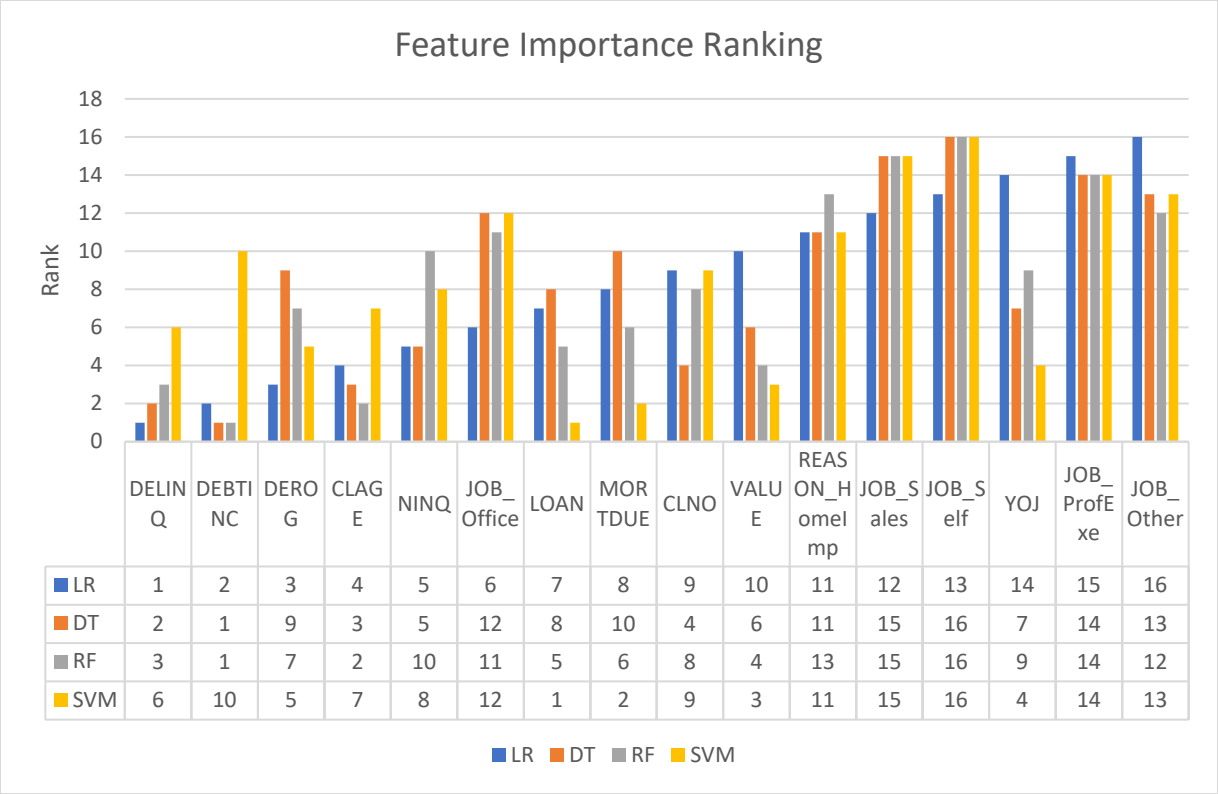

Figure 1: Feature ranking for LR, DT, RF and SVM

Among the top 10 features, the features affecting the SVM model most are the features affecting the least the other models. However, the top 10 features are almost the same across all models on average. Therefore, the table shown previously should be the features the bank should concentrate on gathering. I think having a sense of ranking is more than enough in terms of trade-offs between model interpretability and feature importance. Other models will give better interpretability because it precisely shows how much it affects the overall classification via coefficients. In practice, knowing which features are essential and their ranking relative to one another is usually sufficient.

So, based on the feature ranking tables, the top 3 features affecting the SVM classification most are the amount of the loan, the remaining mortgage due and the value of the property. If the bank decides to opt for any other analyzed model, then the debt-to-income ratio, the number of delinquent credit lines and the age of the oldest credit line would be the essential features.

Problem and Solution Summary

Summary of the Problem

One of retail banks' profits comes from interest via home loans. Therefore, it is crucial for them to carefully select their loaners, as defaulters can compromise their profits. The approval process attempts to determine the applicant's creditworthiness by manually looking at different aspects of the application. However, this approach is labour intensive and susceptible to human error and bias. There have been attempts to automate the application process via heuristics. Still, with the advancement in

data science, there's a desire to build models that can learn the approval process while improving it by not repeating the same bias and error.

The objective is to build a classification model that can predict potential defaulters and provide banks with recommendations on the relevant features to consider from an applicant.

A bank's consumer credit department's goal is to simplify the decision-making process for home equity lines of credit to be accepted. They will adopt the Equal Credit Opportunity Act's guidelines to establish an empirical model for credit scoring. The model will leverage recently approved loan applications. The model will be built on predictive techniques, but it must remain interpretable to justify rejection.

Final Proposed Solution Design

Here's a table showing the overall performance of each model:

Table 2: Performance summary of the models

Model	Overall_Average_Accuracy	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
LR	0.8358	0.8461	0.8255	0.3341	0.2823	0.7358	0.7000
Tuned DT	0.8781	0.8837	0.8725	0.7821	0.7204	0.6755	0.6837
Tuned RF	0.9075	0.9202	0.8949	0.9045	0.7742	0.7435	0.7347
Tuned kNN	0.8978	0.9175	0.8781	0.5826	0.4194	0.9937	0.9873
SVM	0.8792	0.8938	0.8647	0.4651	0.3602	0.9845	0.9710
Tuned SVM	0.9591	0.9796	0.9385	0.9865	0.9247	0.9159	0.8075

The SVM model will be the proposed technique best suited for the bank's needs. It has the highest accuracy, and the feature rankings are easy to interpret. The interpretability of SVM is easy to understand, as you can plot the boundary lines, and it gives a clear visual representation of the classification.

Reasons why this will solve the problem

From a business standpoint, this high accuracy model will correctly identify potential defaulters 95.91% of the time, thus avoiding potential financial losses. The bank will also be more efficient with asking for the relevant features for the prediction model, e.g. saving employee hours, computational resources, etc.

Recommendations for Implementation

Recommendations to Implement the Solution

There has to be a means to measure and monitor performance to operationalize the solution continuously. The bank also has to set a baseline against which future iterations of the model can be

measured. Finally, the bank has to constantly iterate on different model aspects to improve the overall performance.

Model operationalization might include deployment scenarios in a cloud environment, at the edge, in an on-premises or closed environment, or within a closed, controlled group. Among operationalization, considerations are model versioning and iteration, model deployment, model monitoring and model staging in development and production environments. Depending on the requirements, model operationalization can range from simply generating a report to a more complex, multi-endpoint deployment.

Key Actionable for Stakeholders

This can be split into two categories: action to implement and deploy the model and efforts to improve the model. The latter is more long-term based and not necessarily immediate. However, the actions taken in the short term should consider the long term to avoid extra work.

An example of this would be to deal with missing values, especially for features of high importance.

Feature	%
BAD	0
LOAN	0
MORTDUE	8.69
VALUE	1.88
REASON	4.23
JOB	4.68
YOJ	8.64
DEROG	11.88
DELINQ	9.73
CLAGE	5.17
NINQ	8.56
CLNO	3.72
DEBTINC	21.26

Figure 2: Percentage of missing values in each feature

Given how some of these features are missing many values, it would help significantly improve the model by having a complete dataset. For example, the debt-to-income ratio is missing over 21% of its features, despite being one of the essential features outside of SVM.

Based on the recommendations to implement the solution, the bank must determine the most critical metrics to measure the model's performance. It could be the overall accuracy of the model, or the recall and precision, depending on what type of error you're trying to minimize.

They also have to determine how to deploy the solution. Each method has its pros and cons, depending on its current processes.

Expected benefits/Costs

The measurement of success for choosing the best model will be based on the average accuracy in classifying loaners. The highest will be deemed the best model. This will allow the bank to predict defaulters more consistently, thus avoiding heavy losses more accurately.

Even though the accuracy correctly predicts if the applicant will repay their loan or not, precision and recall are also necessary. Here's the summary of SVM's performance on the unseen data:

	precision	recall	f1-score	support
0	0.98	0.94	0.96	1416
1	0.81	0.92	0.86	372
accuracy			0.94	1788
macro avg	0.89	0.93	0.91	1788
weighted avg	0.94	0.94	0.94	1788

Figure 3: Precision-Recall Summary for SVM on an unseen dataset

The precision, in this context, aims not to mistake anyone as a payer. For example, if the model classifies 94/100, applicants will repay their loan. The precision indicates that out of the 94, the model predicts that 98% repaid their loan. It also suggests that of the remaining 2% who did not repay, 81% were confident that they didn't repay. Similar to precision, the recall aims not to mistake anyone as defaulters. With the SVM model, 94% are payers, and out of the remaining 6%, 92% are actual defaulters. To put it simply, precision is the ability to get the highest ratio of loan payers, or the hit rate, and recall is the ability to not miss out on loan payers or the capture rate of loan payers.

The other aspect of choosing the best model is determining the most important features when looking at an application. The identified features affect the predictions the most. This will ensure that loan representatives capture the most meaningful attributes of an applicant, making the overall process more efficient and saving time (for the rep helping the applicant, the applicant filing the forms, and less computational time).

For example, suppose only the top 5 features are considered in the model. This would reduce the number of questions in an application by more than half, saving time for the employee assisting the applicant. It would also reduce the overall operational cost of running the model to classify the

application. The exact cost of analyzing an outcome can vary significantly based on other factors; it's easy to see how it can reduce the overall cost of person-hours (over 50% in this example). The process can be even more streamlined if the features are easy enough that applicants can submit online without any assistance unless necessary. This would save even more time and headaches for both the customer and the bank.

Key Risks and Challenges

The higher the precision, the less likely it is to recruit defaulters, but the potential client pool becomes smaller. The higher the recall, the larger the potential pool of clients but the higher the risk of recruiting defaulters. The balance of recall and precision levels is a matter of risk appetite. Are we willing to accept more risk of recruiting bad clients to capture more potential clients? Crucially these are the key points that businesses are focusing on.

The strategic objective of businesses could be anything from identifying potential clients, capturing fraud, maximizing campaign effectiveness or reducing churn. To communicate prediction model results effectively, we should align with the metrics business leaders are looking at: conversion rate, churn rate, fraud incident rate, capture rate, hit rate, etc.

Further analysis to be done or Other Problems that need Addressing

Despite all of this, there is still room for improvement. The first is the imbalance of meaning of the data. For this analysis, only standardization on all features was performed. It would be beneficial to perform a combination of normalization and standardization depending on the feature type to yield better performance from the models. There's also the problem of collinearity between independent features. As highlighted in Milestone I, the value of the property and mortgage amount were highly correlated at 0.88. It would be worth investigating the performance by dropping one of the two features. There are other methods, such as under/oversampling the dataset. There's also the possibility of feature engineering, finding impactful features from combinations of other existing features. Then, other models are worth exploring, such as Naive Bayes and stochastic Gradient Descent. There's also the possibility to optimize the hyperparameters and look at different tuning techniques, such as random search or Bayesian optimization