Predicting Home Loan Defaults
Jae Kum (Jackie) Kim

Executive Summary

This project proposes that the Fine-Tuned Random Forest model should be utilized for predicting the probability of applicants' loans defaulting in a retail bank. After comparing with other models, it was the most optimal at tracing applicants defaulting while simultaneously minimizing the scenario where the model predicts that applicants will not default when in fact they actually did. Overall, the Recall rate was at 73%. There exists a limitation such that the model is dependent on the applicants' Debt-to-Income Ratio or the number of Delinquent Credit Lines. Furthermore, the ratio does not fully explain Net Incomes or Debts that can potentially be better features for the model. Moreover, what will happen if the applicant is not honest with the application? There is definitely the human factor that the model may not be able to fully accommodate, so a role of underwriters will be important to do a background check for approved applicants. In the end, the model strongly recommends that there should be a more holistic application to capture applicants' backgrounds as much as possible, especially Net Incomes and Debts. If the applicant fails to provide either one of them, the company should decline the loan approval. Furthermore, if the Debt-to-Income Ratio is at least 44%, regardless of other features, the bank should decline his or her loan approval. Additionally, it is recommended that, if possible, the applicant is able to provide an information on delinquent credit lines. If the loan requestor fails to provide the detail, the bank should automatically decline the loan approval.

Problem Summary

Interests in the form of home loans have been a major proportion of retail bank profits. Concurrently, they are some of the perilous methods to gain profits because retail banks have to be keen of potential defaulters whose downfall naturally become losses for these companies. While they have to be judicious with the approval process, it has become substantially manual and time-consuming for banks such that it is not only inefficient but also error-prone as there is a manual, human element to continuously check-in the application. The focus of this project is to construct a predictive classification model that not only predicts which applicants will have the highest chance of defaulting but also minimizes the error of approving applicants when in fact they will default. Ultimately, the goal is to automate the process of the application process and alleviate the significant pressure for retail banks from manually checking each application and concentrate on background checks for each approved applicant.

Solution Design

Multiple models have been tested with the data provided by the Home Equity Dataset (HMEQ) such as Logistic Regression, Decision Tree, and Random Forest. Additionally, Fine-Tuning has been applied for the last two models accordingly. The primary focus was ensuring that the Recall Rate surpasses 70% though the minimum condition was that Accuracy had to be greater than 80%. Appendix 1 and 2 show the top 2 confusion matrixes of respective models that has shown the best result at Recall Rate level.

Figure 1 below shows a summary table that describes four key metrics: Accuracy, Precision, Recall, and F1 Score.

Model	Accuracy	Recall*	Precision	F1 Score
Random Forest (Tuned)	**87.1%**	**73.4%**	**67.5%**	**70.3%**
Decision Tree (Tuned)	85.2%	73.1%	62.2%	67.2%
Random Forest (Untuned)	91.0%	64.0%	89.8%	74.7%
Logistic Regression (Newton-CG)	76.2%	65.3%	44.8%	53.2%
Decision Tree (Untuned)	86.7%	59.9%	71.7%	65.3%
QDA	80.1%	39.5%	53.1%	45.3%
LDA	82.3%	28.0%	68.4%	39.7%

Figure 1: Results from different tested models. Notice that Tuned Random Forest shows the strongest Recall Rate at 73.4%

As the summary table above shows, the Recall Rate was the greatest with Tuned Random Forest model (73.4%), with the Tuned Decision Tree model coming close second (73.1%). One of the limitations was that the provided data had 20% missing data for Debt-to-Income Ratio, so these missing data have been replaced with medians. Consequently, tuning the model with Grid Search Cross-Validation may not have functioned as much as one has hoped because the missing data has been replaced with median values and may potentially impact the interpretability. However, one should not recklessly increase the Accuracy as that will result in overfitting the model.

Analysis and Key Insights

In order to understand why there has been an emphasis on Debt-to-Income Ratio, it will be important to comprehend the background of the data.

Figure 2 shows descending ordered Feature Importance as interpreted by Tuned Random Forest model.

Figure 2: Feature Importance interpreted by Tuned Random Forest

In this case, **DEBTINC** is an acronym for Debt-to-Income Ratio, and the chart shows this variable as the key feature that determines whether the retail bank should approve the applicant's home loan or not. According to the model, its importance was at around **60%**, compared to **DELINQ**, delinquent credit lines, that is listed at 10% importance.

Figure 3 below explains whether the applicant has defaulted or not based on the Debt-to-Income Ratio.

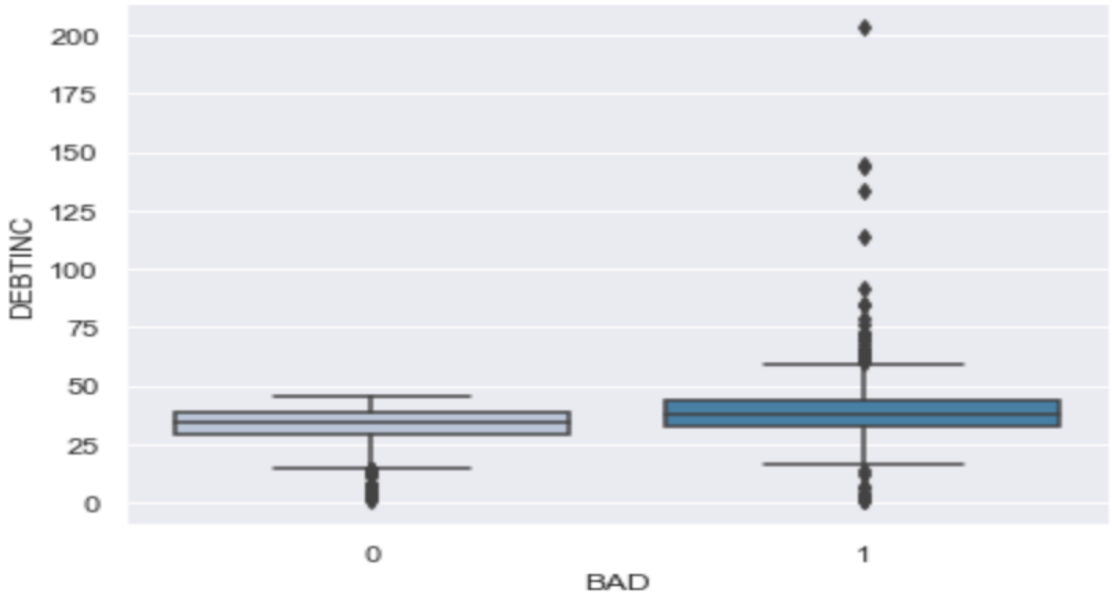


Figure 3: Whether the applicant has defaulted or not with respect to Debt-to-Income Ratio.

The percentage difference between the mean showed that the non-default applicant had 15% less Debt-to-Income Ratio compared to the defaulted applicant, showing that there exists much more stability on applicants who did not default.

It is worth to note that Delinquent Credit Lines also has some credits in determining the possibility of home loan defaults. While some applicants did default even though there was no delinquent credit line, the proportion significantly increases as the number increments by one.

Figure 4 below explains the significant change in the number of delinquent applicants if the number of credit lines increased by 1 from none.

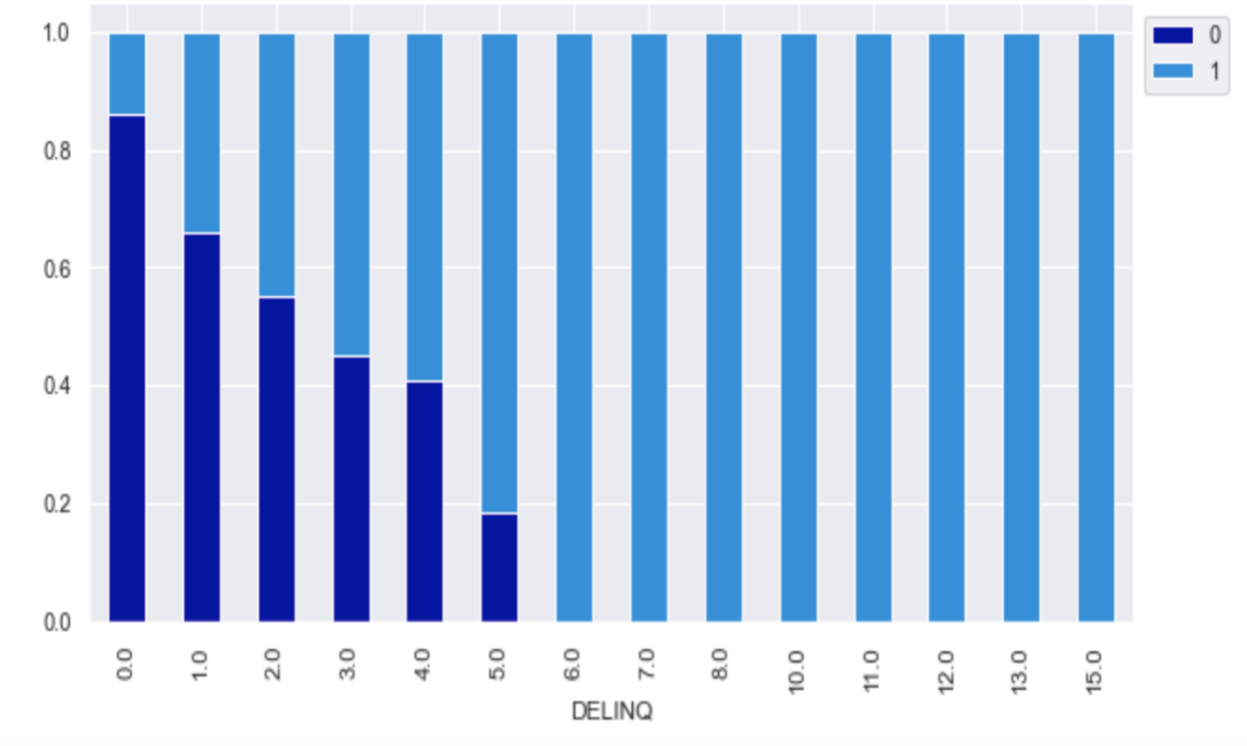


Figure 4: A proportion of applicants who defaulted or not for each delinquent credit lines

It is noticeable that the proportion of home loan defaults spikes significantly from around 18% to 35%. After the first delinquent credit line, the proportion does not change as much for 2 and so on. While its impact in predicting the home loan default may not be as impactful as Debt-to-Income Ratio, Delinquent Credit Line is another potential candidate to track the probability of the loan default.

Limitations and Recommendations

The current model is dependent on the completeness of the Home Equity Dataset provided by retail banks. Consequently, it is not going to function properly if the applicant omits the detail in the application that can create a bias in the model due to an omitted data. Two of the key features that will be mandatory will be Debt-to-Income Ratio and a number of Delinquent Credit Lines. If neither is provided, the loan approval should automatically be declined.

Additionally, the Debt-to-Income Ratio is just a percentage measurement and does not fully explain the true net income and debt each applicant currently has. For instance, the applicant may have low debt amount but concurrently may not have enough net income to successfully repay the mortgage loan. In this case, there is a peril of approving the loan since the loan requestor may not be able to pay it back over time. While adding the two variables will definitely require additional model testing that will require more time to analyze, in order to further enhance the performance of the model without the potential bias coming from the ratio, both Net Income and Debts will be needed. If neither is provided, similar as above recommendation, banks should decline the applicant's loan approval.

Finally, the model is prone to the honesty of the loan requestor. For example, if he or she falsifies the application to be more advantageous to his or her chance of loan approval, the model will likely approve it since the detail looks

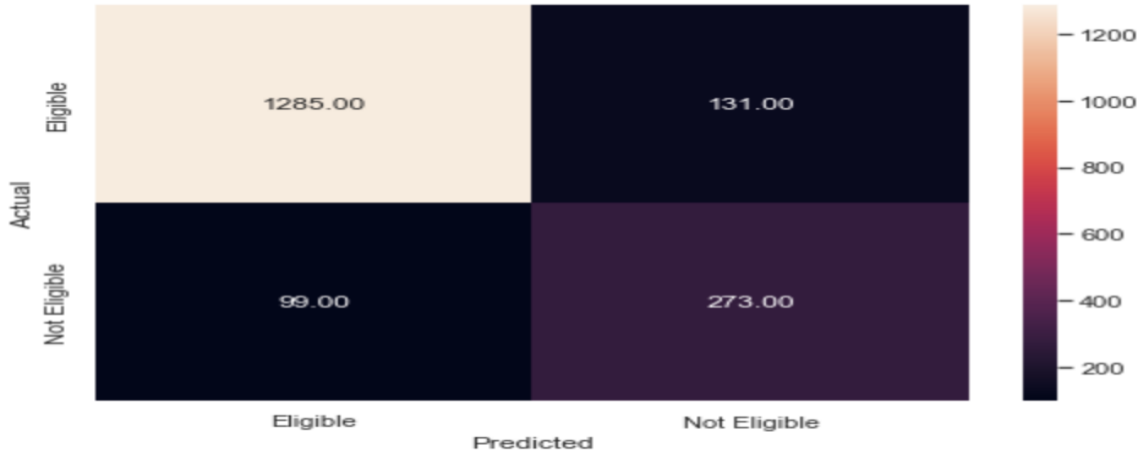
valid in terms of its perspective. Unfortunately, this crime has been a known case in the United States where around 11% of loan applications have been found false after background check. Hence, it will be essential underwriters will have to undergo additional background checks for approved applicants using credit histories. A collaborative effort between underwriters and the model will greatly minimize the risk that the loan will officially be lent to the respective applicant who falsely wrote his or her application. It is true that there is an overhead cost to use underwriters for background checks, but carrying the liability from defaulted loan is a greater loss in profit for retail banks than paying their employees for an effort to track approved applicants who lied in the application.

Bibliography

- Terri Williams 2021. What is Automated Underwriting? What is Automated Underwriting?
- Mike Cetera 2022. Lying on a personal loan is a bad idea Lying on a personal loan is a bad idea

Appendix 1: Tuned Random Forest (Recall: 73.4%)

	precision	recall	f1-score	support
0	0.93	0.91	0.92	1416
1	0.68	0.73	0.70	372
accuracy			0.87	1788
macro avg	0.80	0.82	0.81	1788
weighted avg	0.88	0.87	0.87	1788



Appendix 2: Tuned Decision Tree (Recall: 73.1%)

	precision	recall	f1-score	support
0	0.93	0.88	0.90	1416
1	0.62	0.73	0.67	372
accuracy			0.85	1788
macro avg	0.77	0.81	0.79	1788
weighted avg	0.86	0.85	0.86	1788