Applied Data Science Program Oct 2022 - Feb 2023

Loan Default Prediction
(Classification Capstone Project)

Submitted by Ami Kang

Feb 2023

Executive Summary

This report presents proposals for more refined lending decision criteria to bolster the efforts of the bank's consumer credit department in mitigating potential losses from their home equity line of credit. By adopting this new lending decision strategy, the bank can anticipate a shift from a negative return on investment of -5% to a positive return of +8%.

The key changes proposed in this study include:

- □ Calculation of a probability of default for each loan applicant using a recommended model. If the probability exceeds a predetermined threshold, the loan application will not be approved.
- □ Enhanced caution in cases where debt-to-income information is unavailable at the time of loan decision-making.
- □ Collection of additional information to support decision-making when debt-to-income information is not available, and consideration of historical performance data.
- □ A personalized approach to loan decision-making, based on the outcome of a complex model, which takes into account the unique circumstances of each applicant.
- □ Careful validation of the method proposed in this study to ensure compliance with fair lending laws, as each rejected loan application will have its own specific reasons for denial.

Problem Statement

A bank engages in lending to clients with the expectation of generating interest income and recovering the full loan principal at the end of the loan term. However, when a loan account becomes default, the bank experiences not only a loss of expected interest income but also a significant portion of the loaned principal, which is typically larger than the interest earned.

Analysis of historical data reveals that 20% of borrowers defaulted prior to the completion of their loan term. Such defaults can have a significant impact on the bank's financial performance, depending on the loan values involved. As such, it is imperative for the bank to establish a systematic and proactive approach for identifying potential clients with a high likelihood of default at the time of loan decision-making.

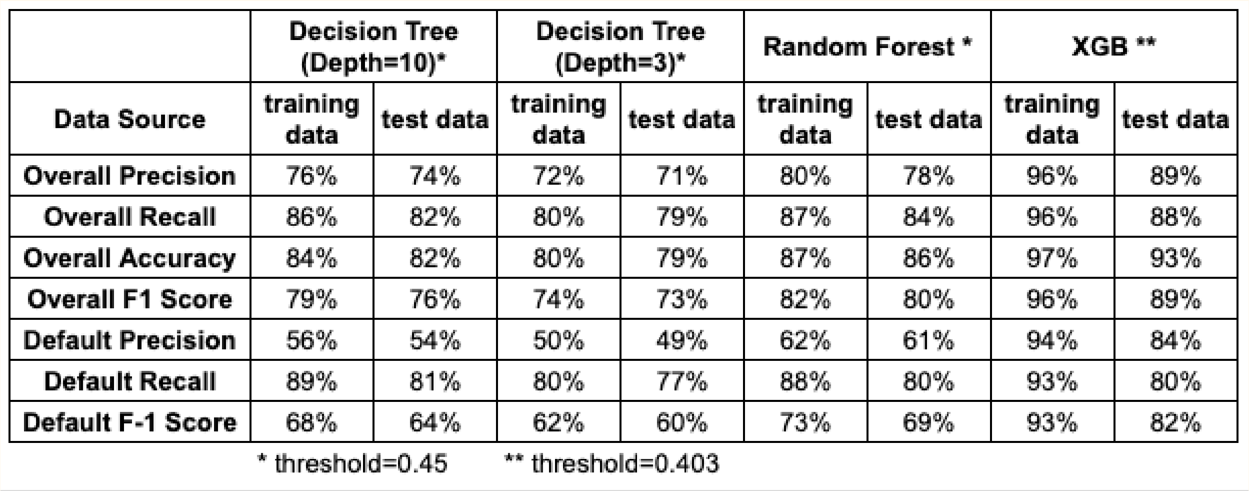
Solution Development Methodology

1. Conduct Extensive Exploratory Data Analysis: A comprehensive Exploratory Data Analysis (EDA) was performed to gain insights from the data and develop a treated dataset that is more suitable for model development than the raw data.
2. Prepare a Cleaned and Feature-Engineered Dataset:
    a. The original dataset had considerable cases of missing values for each variable. To address this issue, the following steps were taken:
        i. For categorical variables, a new category value labeled 'UNKNOWN' was generated.
        ii. For numerical variables, the following steps were taken:
            1. Binary tags indicating the presence of missing values were created.
            2. The mode was used to impute missing values for integer variables.
            3. For continuous variables, missing values were imputed with the average.
3. Standardize the Dataset: A clean and standardized dataset was prepared for models that are sensitive to distance information.
4. Feature Engineering: Various feature-engineered datasets were tested to improve model performance, and the best-performing dataset was used for the final model.
5. Explore Classification Methodologies: This study evaluated logistic regression, decision tree, random forest, ada booster, gradient booster, extreme gradient booster, KNN, linear discriminant analysis, and quadratic discriminant analysis to determine the most suitable modeling framework for the given dataset.

6. Model Evaluation Criteria: The loss from booking bad loans and losing the loaned principal (plus interest payments made to the National Central Bank) is significantly higher than the opportunity cost of not earning interest income from failing to book a profitable account. As such, a false negative error is deemed more critical than a false positive error. Therefore, the primary model evaluation metric used is recall rate. The final model should also demonstrate reasonable accuracy, precision, and F1-score.
7. Hyperparameter Optimization: The GridSearch method was applied to each model, including the default values of hyperparameters, to improve model performance.
8. Final Model Selection: The final model was selected from a few strong candidate models based on the recall rate, adverse action explainability (for rejected loan applicants) and the practicality of implementing new rules in real business operation.

Solution Development

The extensive evaluation of various classification models revealed that the decision tree model, random forest model, and extreme gradient boosting model, when adjusted with hyperparameters, demonstrated good performance. The performance results of these models are presented in a table, while the full collection of all tested models' performances is included in the accompanying python code submission.



	Decision Tree (Depth=10)*		Decision Tree (Depth=3)*		Random Forest *		XGB **	
Data Source	training data	test data	training data	test data	training data	test data	training data	test data
Overall Precision	76%	74%	72%	71%	80%	78%	96%	89%
Overall Recall	86%	82%	80%	79%	87%	84%	96%	88%
Overall Accuracy	84%	82%	80%	79%	87%	86%	97%	93%
Overall F1 Score	79%	76%	74%	73%	82%	80%	96%	89%
Default Precision	56%	54%	50%	49%	62%	61%	94%	84%
Default Recall	89%	81%	80%	77%	88%	80%	93%	80%
Default F-1 Score	68%	64%	62%	60%	73%	69%	93%	82%

* threshold=0.45 ** threshold=0.403

The decision tree model is favored for its relatively straightforward interpretability, especially when its decision rules are implemented across the board. To optimize operational efficiency, a decision tree model with a maximum depth of 3 was also developed. However, the comparison between the max-depth=3 and max-depth=10 models reveals a trade-off between the simplicity of implementing new rules and the prediction performance of the model.

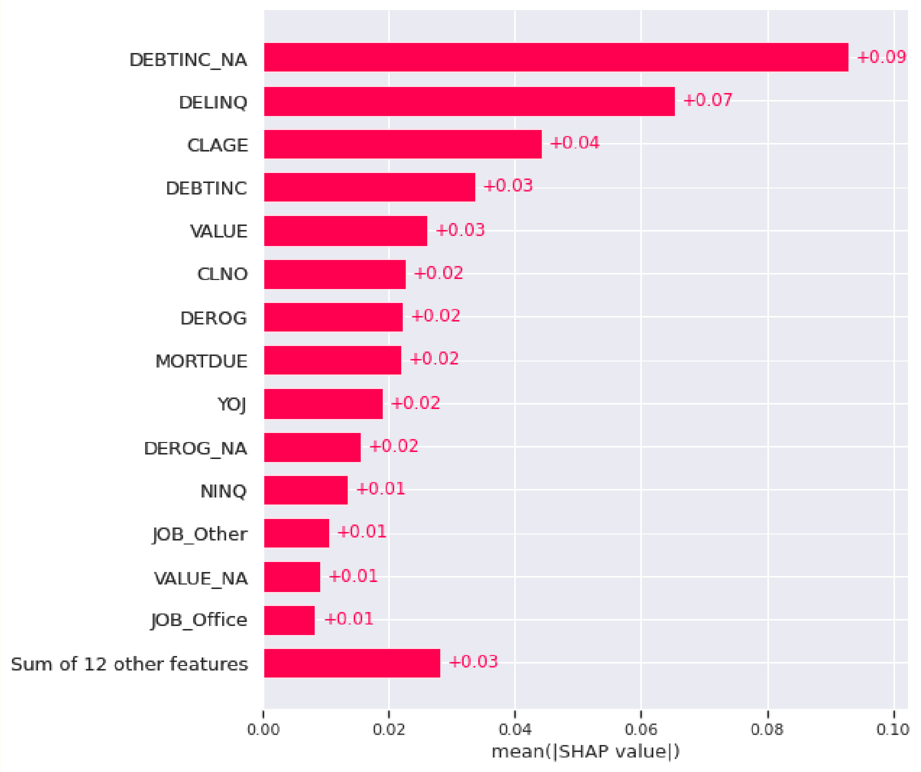
The random forest model demonstrated improved prediction capabilities compared to the decision tree models. However, it can be challenging to extract lending decision rules from the black-box random forest model and to explain rejections to home loan applicants. As a result, the next candidate, XGB, was considered and with a slightly altered decision threshold from 0.5 to 0.4, the XGB model displayed remarkable prediction capabilities. Therefore, XGBT was ultimately selected as the final modeling framework.

Learning from the Final Model

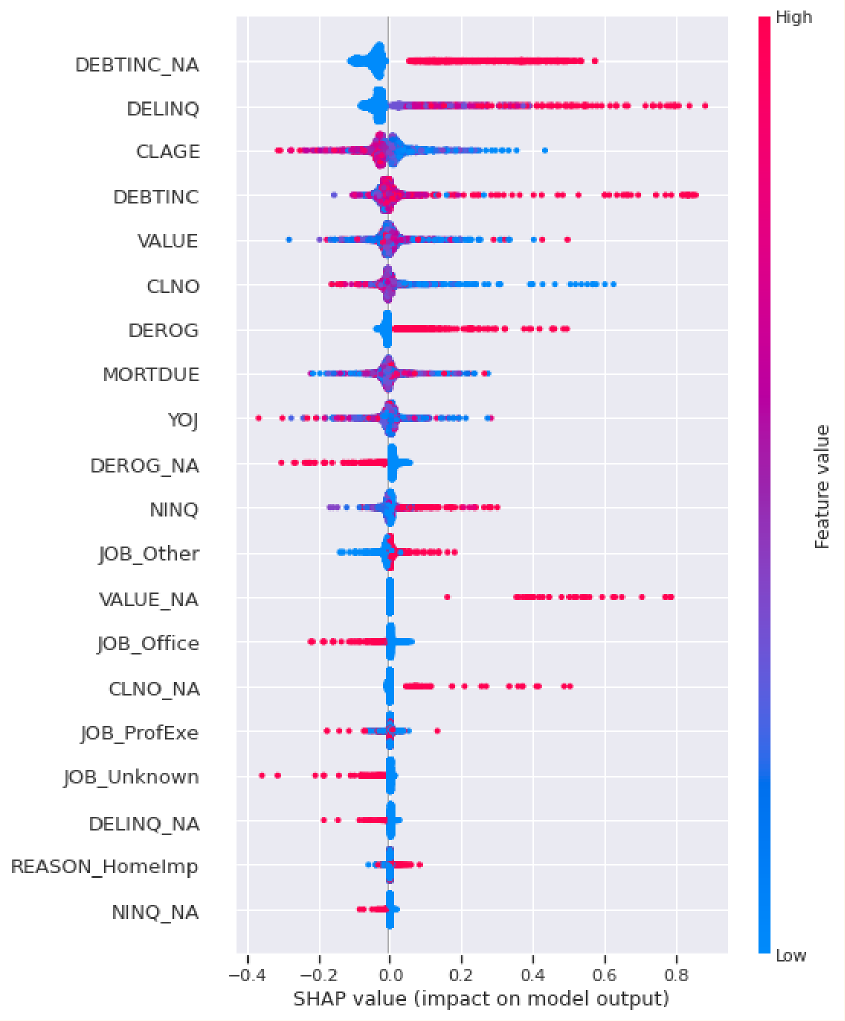
<Global Insight>

1. The lack of debt-to-income information can be a crucial predictor of loan applicants who may potentially default.
2. A history of delinquency is a robust predictor of future loan applicants who may default.
3. An extensive credit history is a dependable indicator that applicants are less likely to default.
4. A high debt-to-income ratio will raise the likelihood of default.
5. Derogatory history is another effective predictor of potential default, while the absence of such history can indicate a positive credit standing.

<mean SHAP plot from the final XGBT model>



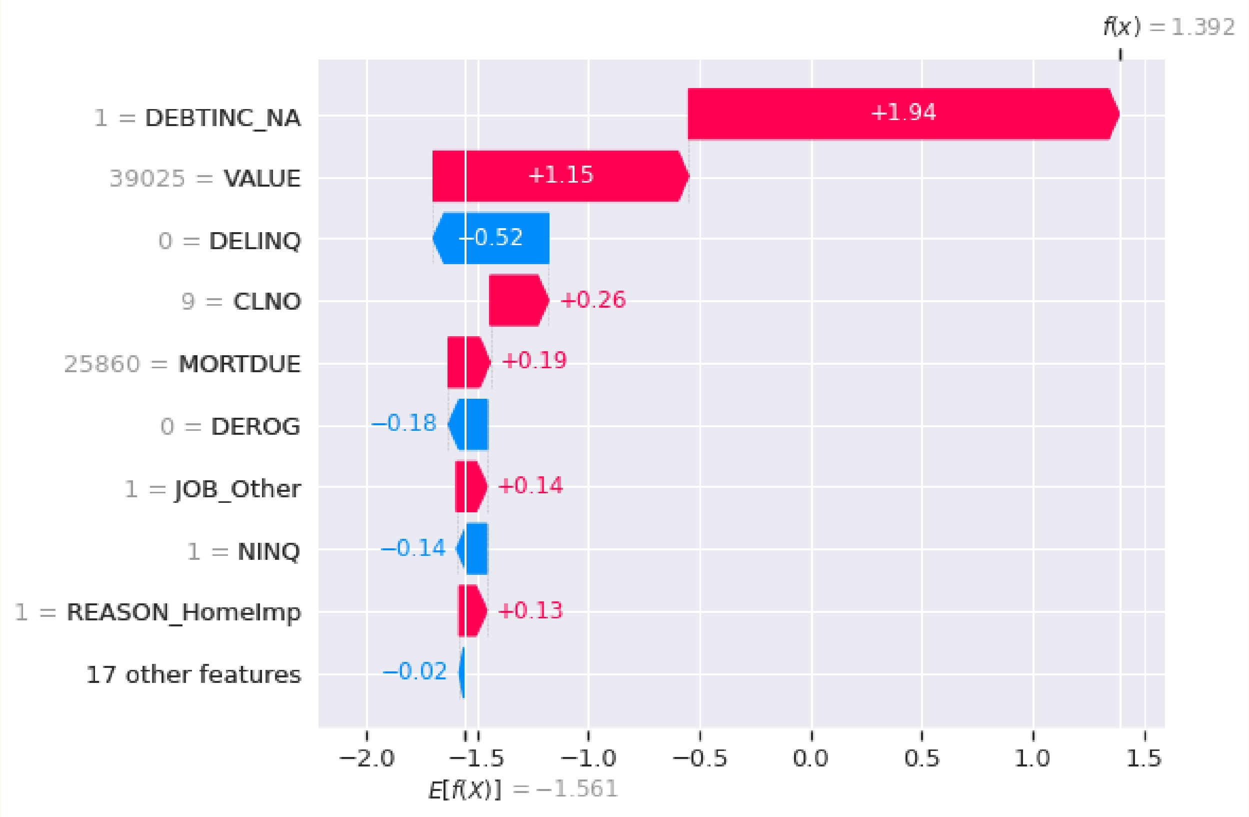
<Aggregated SHAP values in beeswarm graph>



<Local Analysis and Its Application to a Reject Explaining Letter>

Consider a loan application in the given dataset that resulted in default. If the banker had access to the XGBT model presented in this study, the applicant would have been rejected due to the high probability of default. This study utilized a decision threshold of 40% and the default probability of that account was calculated to be 80%, resulting in a rejection decision.

However, how would the bank explain the rejection to the applicant?



The above illustration can aid the bank in communicating the reasons for denying the home loan credit line application. Firstly, the applicant did not provide debt-to-income information, which is crucial for the bank to assess the borrower's ability to repay the loan. Secondly, the value of the property was not sufficient to serve as secondary collateral. Lastly, the high frequency of credit line inquiries (9 times) indicated that the applicant looked like a credit seeker rather than a financially stable individual.

Cost Benefit Analysis (with a conservative assumption)

The following assumptions were made in this analysis:

- Borrowers take out the full credit line available to them.
- An interest rate of 10% is applied (even though current home line equity loan interest rates are lower than 8%).

- It is assumed that the bank will incur a 70% loss on loan principal from defaulted accounts, although in reality the bank may incur an even greater loss.

	Investment	**Safe-loans**	**Lost-loans**	**Expected Profit**	**Expected Loss**	**Net Profit**	**ROI**
Current BAU	$110,903,500	$90,783,100	$20,120,400	$9,078,310	$14,084,280	-$5,005,970	***-5%***
Proposed Scenario	$91,049,900	$88,913,600	$2,136,300	$8,891,360	$1,495,410	$7,395,950	***8%***

These assumptions suggest that the business is currently incurring losses instead of profits. However, if the proposed rules from the final model in this study were applied to the current portfolio, the bank could see a significant increase in profits. The return on investment (ROI) would increase from -5% to +8%. This is because the bank would avoid 89% of future default loans, even though it would miss the opportunity to approve 2% of applicants who would not default. Additionally, by lending a total of $91 million instead of $111 million, the bank could reduce its interest costs by not having to borrow an additional $20 million (=$111 million - $91 million) from the central bank or could redirect those saved funds towards more profitable ventures, which are additional benefits that are not captured in this analysis.

Business Strategy Proposal

- □ Implement the XGBT model developed in this study to calculate the default probability of an applicant prior to lending decisions.
- □ Deny loan applications if a calculated default probability exceeds 40%.
- □ Utilize account-level SHAP analysis to draft a comprehensive and reasonable explanation for adverse lending decisions.
- □ Exercise caution when the debt to income information is unavailable and make a collective judgment based on past performance data.
- □ Make efforts to gather supplementary information whenever the debt to income information is missing.

- ☐ Conduct a thorough investigation into the reasons for the lack of proper collection of the critical Income to Debt ratio information, and how loan approvals were granted in its absence.
- ☐ Before implementing the proposed new rules, assess the potential for any intentional or unintentional biases affecting specific groups of individuals that may result from these changes.

Key Risks and Limitations

- ☐ The bank should check whether the current business practice itself is biased, and therefore the input data used for this study was already biased from which the final model is developed.
- ☐ The data provided does not have time stamp information and this study is built based on the assumption that all variables in the original dataset, except the loan amount, were collected BEFORE the lending decision was made for each loan.
- ☐ The final model is developed using a training dataset and evaluated using a test dataset. However, this model has not been evaluated using a different time period dataset yet.
- ☐ The bank may want to estimate the implementation and additional operation cost to execute the XGBT model in the system and re-run the cost and benefit analysis.

Future Analysis and Strategy Enhancing Ideas

The subsequent step is to formulate strategies aimed at enhancing Return on Investment (ROI).

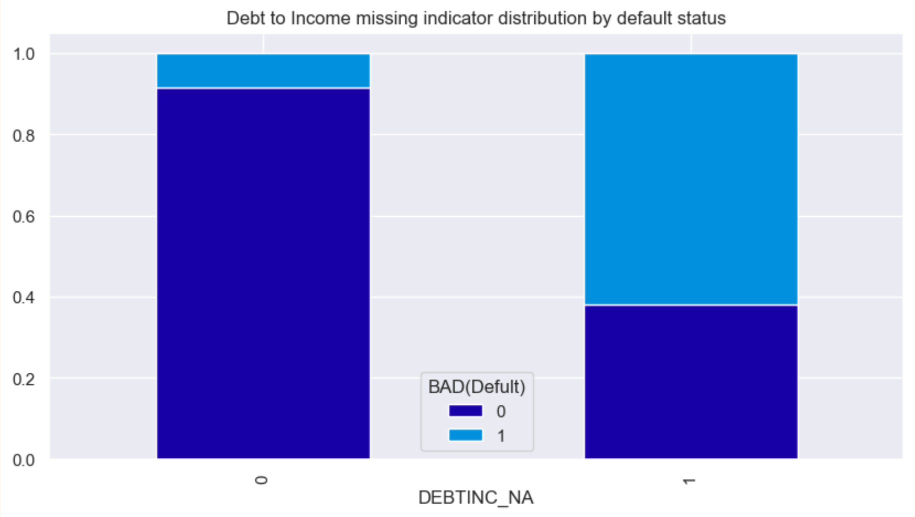
- ☐ The availability of actual loss figures from default accounts will facilitate the preparation of a more comprehensive and rational cost-benefit analysis.
- ☐ After predicting which accounts will default, it is critical to address the research question of the optimal loan amount for approved accounts. The data science team may create a separate model to optimize loan amounts or utilize the probability of default established in this study. For instance, if the probability of default is substantial (but below the loan rejection threshold), the bank may consider granting a lower credit line than their standard business practices to that particular applicant.

□ It is essential to continuously monitor the stability of the model and assess its performance using different time-period data. Regular calibration may be necessary to maintain optimal performance.

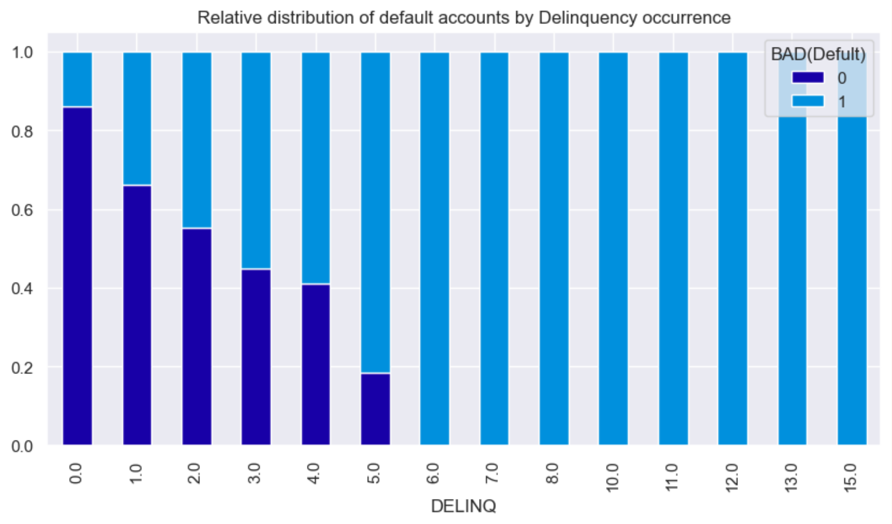
These additional efforts will enhance this research and significantly aid the bank in making more informed and rational decisions with home equity lending business.

Appendix:

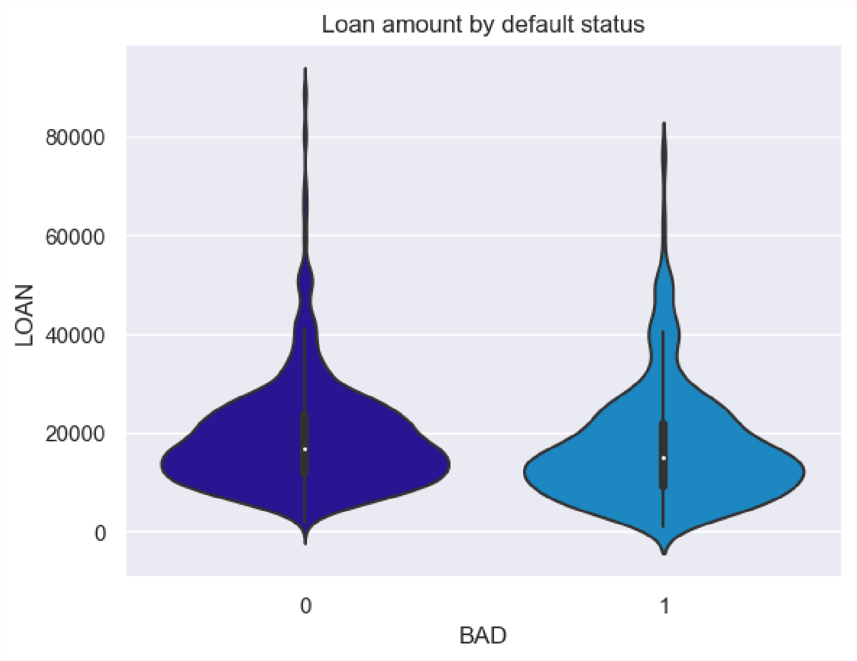
Highlight of EDA



- Finding: Much higher portions of default accounts miss the debt to income information.



- Finding: Defaulted accounts tend to have higher delinquent occurrence records.



	$ LOAN statistics by account's default status							
BAD	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
0	4,771	$19,028	$11,116	$1,700	$11,700	$16,900	$23,500	$89,900
1	1,189	$16,922	$11,418	$1,100	$9,200	$14,900	$21,700	$77,400

- Finding: The Loan amount for defaulted accounts is lower than non-defaulted accounts which shows that the bank might have some strategy already built on how much to lend to the risky population.

Final XGBT model hyperparameters setting:

- **Final XGBT model hyperparameter setting**:
- eval_metric=’error’
- n_estimators=500
- random_state=3
- subsample=0.7

Developed models performance with Test data set

More models' performance that were examined, but not displayed in the final report, in this study is included in the separately submitted python final code.

Model	Overall Precision	Overall Recall	Overall Accuracy	Overall F1 Score	focus_precision	focus_recall	focus_F1 _score
Adaboost with new threshold	0.100	0.500	0.200	0.166	0.200	1.000	0.333
Modified RF new threshold	0.734	0.830	0.805	0.754	0.507	0.871	0.641
Modified DT(new threshold)	0.713	0.806	0.781	0.728	0.473	0.846	0.607
KNN with a new threshold	0.901	0.889	0.934	0.895	0.848	0.815	0.831
Modified DT	0.753	0.821	0.836	0.776	0.563	0.796	0.660
Modified RF(new threshold)	0.871	0.866	0.917	0.869	0.797	0.782	0.789
Modified RF	0.778	0.826	0.859	0.797	0.618	0.770	0.686
Default RF with Weight and New Threshold	0.863	0.850	0.910	0.857	0.788	0.751	0.769
KNN	0.904	0.856	0.926	0.877	0.871	0.739	0.800
LDA with a new threshold	0.806	0.810	0.876	0.808	0.687	0.700	0.693
GBC with new threshold	0.859	0.823	0.902	0.839	0.792	0.692	0.738
XBG with new threshold	0.850	0.820	0.898	0.834	0.776	0.689	0.730
RF	0.893	0.829	0.915	0.856	0.860	0.686	0.763
Logistic Regreession	0.826	0.804	0.885	0.814	0.733	0.669	0.700
Default RF with Weight	0.881	0.814	0.907	0.841	0.842	0.658	0.739
LDA	0.831	0.799	0.886	0.813	0.747	0.653	0.697
XGB	0.876	0.789	0.899	0.822	0.844	0.605	0.705
GBC	0.880	0.788	0.900	0.823	0.853	0.602	0.706
Default DT	0.820	0.768	0.876	0.789	0.737	0.588	0.654
Default DT (new threshold)	0.820	0.768	0.876	0.789	0.737	0.588	0.654
Adaboost	0.871	0.777	0.894	0.812	0.839	0.583	0.688
Logistic Regreession	0.841	0.768	0.883	0.796	0.780	0.577	0.663
QDA with a new threshold	0.786	0.744	0.859	0.762	0.678	0.555	0.610
QDA	0.797	0.738	0.862	0.761	0.704	0.532	0.606
XGB with default threshold	0.912	0.869	0.932	0.888	0.881	0.765	0.819
XGB with new threshold (selected model)	0.894	0.882	0.930	0.888	0.837	0.804	0.820