Peer-Review

Nguyen, Chi Ton C. 2025. “Advanced Loan Default Prediction Models Using Machine Learning Boosting Algorithms.” *Journal of High School Science* 9 (3): 612–34. https://doi.org/10.64336/001c.144823.

A well designed, well written work. Congratulations. However, I have several concerns that need to be addressed.

1.was a stratified split performed? If not, then explain what procedure was used to achieve equal representation of good and bad loans in both the training and the testing data sets. Please discuss in the manuscript.

2.How was class imbalance between the good (90%) and bad (10%) loans addressed? SMOTE? data augmentation? over/under sampling? Please discuss in the manuscript.

3.How was the robustness of the model determined? K-fold cross-validation? adversarial training? Lipchitz constant? Please discuss in the manuscript.

4.The feature ‘inflation’ is among the first 7 in terms of imporatance for lightGBM and XGBoost. This feature will change with time. Explain how - assuming the inflation is significantly different from today - the model is expected to perform from a nearly constant inflation regime from 1999 to 2024 as compared to outliers such as > 13% in the 1970’s. This also ties back to point # 3 as to the robustness of the model.

5.I don’t understand what the feature “loan age” actually means. Does this mean the # of months into the loan that the default resulted? This does not make intuitive sense. Should I take this to imply that there is a ‘loan window’ between (say) 2 to 4 years when the loan has the greatest probability of default? Please explain this in the manucript.

6.Please provide an explanation for why the LightGBM improves significantly more than XGBoost in performance after optimization despite having similar feature importance. What do you mean by 'before optimization" and “after optimization” ?

7.you state “…..The paper found that XGBoost and Light GBM were the top-performers with 98% accuracy…….”, however the table shows Light GBM has 74% accuracy.

8.What percent of the defaulted loans were originated between 2007 and 2010 (the period of the subprime mortgage crisis) ? If a majority of defaulted loans origenated during this period, then the market’s faulty risk pricing methodology is more to blame rather than consumer centric features. Explain in the manuscript why just as good a prediction may not be achieved by including this faulty risk pricing strategy in the model, instead of consumer centric features.

Thank you for your comments! I appreciate your feedback. Please review my responses to your concerns. The manuscript has been updated to reflect the changes and address your concerns. I hope you will approve the changes and accept the paper as it contributes to the financial industry more robust and advanced machine learning models with boosting algorithms for credit risk modeling with remarkable accuracy. The abstract is also updated in the manuscript to reflect the changes and address your comments.

1. was a stratified split performed? If not, then explain what procedure was used to achieve equal representation of good and bad loans in both the training and the testing data sets. Please discuss in the manuscript.
   Yes, a stratified split was performed to achieve equal representation of good and bad loans in both the training and the testing data sets. Please see #2 for the detailed explanation added to the manuscript (Section 2.2.Data Preprocessing).
2. How was class imbalance between the good (90%) and bad (10%) loans addressed? SMOTE? data augmentation? over/under sampling? Please discuss in the manuscript.

To mitigate the problem of class imbalance between default and non-default loans (10.5% vs. 89.5%), a stratified data partition was used. To further balance the dataset and ensure equal representation of "good" and "bad" loans in the training and testing datasets, the Synthetic Minority Oversampling Technique (SMOTE) was applied to synthetically increase the minority class, in combination with under-sampling the majority class. Specifically, for each minority instance, SMOTE identifies its k nearest neighbors, randomly selects one neighbor to gather a synthetic sample; this process is repeated until the minority class reaches the desired level of representation relative to the majority class (Chawla et al., 2002). SMOTE addresses class imbalance by creating synthetic minority samples using an interpolation approach, which helps mitigate overfitting that can result from merely duplicating minority observations. As a result of applying SMOTE to increase the minority class and under-sampling the majority class, the final balanced dataset of more than 65,000 loans achieved an equal distribution of 50% loan defaults and 50% non-default loans across both training and testing data.

3. How was the robustness of the model determined? K-fold cross-validation? adversarial training? Lipchitz constant? Please discuss in the manuscript.
   Please see below for the description of K-fold cross-validation added to the manuscript (Section 4.1.Data Modeling and Cross-Validation).

   K-fold cross-validation is a resampling technique for evaluating the generalizability of predictive models. The key benefit of cross-validation methods is to evaluate model performance on new data. One of its benefits is to reduce the overfitting risk that often arises in the presence of high-dimensional or noisy datasets. Using this approach, the data are randomly partitioned into K equal subsets, or "5 folds". In each iteration, (K -1) folds (i.e., four folds in 5-fold cross-validation) are used to train the model, while the remaining fold is held out for validation. This process is repeated K times. The results from all iterations are ultimately averaged to produce an overall estimate of the model. A 5-fold cross-validation procedure was employed to evaluate all four classification models in this study. The 5-fold approach provides a balance between computational efficiency and reliable performance estimation.
4. The feature 'inflation' is among the first 7 in terms of imporatance for lightGBM and XGBoost. This feature will change with time. Explain how – assuming+ the inflation is significantly different from today - the model is expected to perform from a nearly constant inflation regime from 1999 to 2024 as compared to outliers such as > 13% in the 1970's. This also ties back to point # 3 as to the robustness of the model.
   Thank you for the comment. In our dataset, **all loans are originated in 2020 with performance tracked through September 2023** (please see #8 below for the updated manuscript of overview of the data). Therefore, the inflation feature is not hypothetical but corresponds to the **realized local inflation rates in each loan's geographical location** during this observation window.

   It is true that inflation is a time-varying macroeconomic variable, and indeed the model is trained in a regime with relatively moderate variation (2020–2023). If inflation levels were to shift substantially outside the training distribution - for example, extreme values such as the >13% observed in the 1970s - the predictive performance of the model could be affected, since tree-based methods such as LightGBM and XGBoost rely on patterns present in the training data.
   However, two mitigating factors are worth noting:

1. **Model scope** – The current model is calibrated for **post-2020 loan cohorts**, making the observed inflation variation during that horizon **appropriate** for the prediction task at hand.
2. **Robustness** – While inflation is an important feature, the model also integrates a **wide set of loan-level and borrower-level characteristics** that remain relevant across regimes. **This reduces reliance on inflation alone**.

3. **Future Research** - For long-term deployment in different macroeconomic environments and real-world modeling in banks and financial institutions, the model should ideally be retrained or stress-tested using simulated or historical high-inflation scenarios (e.g., >10%) to assess generalization. This would ensure robustness if inflation dynamics deviate strongly from the 2020–2023 regime. Stress-testing environment to show forward-looking awareness can be considered in the future study, which is mentioned and updated in the section of future study (please see #8 below for future research in the manuscript).

However, considering the current economy, a forward-looking horizon of five years, and within the scope of this analysis, 1970s-type inflation may not generalize and not likely to happen, the current model results based on both training and testing data should be reliable.

5. I don't understand what the feature "loan age" actually means. Does this mean the # of months into the loan that the default resulted? This does not make intuitive sense. Should I take this to imply that there is a 'loan window' between (say) 2 to 4 years when the loan has the greatest probability of default? Please explain this in the manucript.
   Thank you for raising this point. In the dataset, "Loan Age" refers to the elapsed time since loan origination up to September 2023 or until the time the loans were tracked, or until the payoff date (in the case that loans are paid off), whichever comes first (please see Section 2.3. Feature Extraction for "loan age" description in the manuscript).
   It does **not** represent the number of months until a loan defaulted. Rather, it captures how long a loan has been active. In this case, loan age is not meant to imply a fixed default window. Instead, it provides the information about the loan's maturity stage, together with other factors (such as borrower characteristics and macroeconomic conditions) that can determine default risk.
   In the sample, some loans were active and paid on schedule; some loans were paid in full (loan payoff); and some loans were active and had missing payments. Loans may have 2 months of missing payments, 3 months of missing payments, 6 months of missing payments or 18 months of missing payments or even longer. The number of months active or in delinquency is not fixed. Therefore, loan age is not the number of months until a loan defaulted.
   More importantly, "Loan Age" was not used to define default outcome. Even loan defaults (i.e. missing payments for at least 6 months) can have a varied window of loan age. Non-default loans, including both active loans that were paid on schedule and payoff loans, also have a varied window of loan age. Default events, defined as missing payments for at least six months, can occur across a wide range of loan ages. Loan age is not the age of loans that missed payments for 6 months. It could also be the age of loans that missed payments for 20 months in September 2023 as long as the loan performance was tracked until September 2023.
   Similarly, non-default loans, including both active loans that remain current and loans that were fully repaid, also span varied loan age intervals. In this sense, loan age is analogous to human age: just as age correlates with survival outcomes but survival outcomes were not based on age

cut-off. Similarly, loan age may correlate with loan default but was not used to define loan default.
To avoid confusion, the description of *loan age* has been added to the manuscript, stating simply **the elapsed time since loan origination up to September 2023 or until the time the loans were tracked, or until the payoff date (in the case that loans are paid off), whichever comes first,** not the number of months until default.

6. Please provide an explanation for why the LightGBM improves significantly more than XGBoost in performance after optimization despite having similar feature importance. What do you mean by 'before optimization" and "after optimization" ?
Optimization refers to "Hyperparameter tuning". The whole Section 4.3. Hyper-parameter Optimization was dedicated to explain how hyper-parameters (i.e., the number of estimators, learning rate, max features, max depth, min samples split, min samples leaf) contributed significantly to the robustness and accuracy of the LightGBM model.
"Before optimization" means random parameters were used. "After optimization" means parameters have been tuned. The optimal tuning is essential to ensure that the boosting algorithm generalizes well across new and unseen data and achieves the best performance.
7. you state "…..The paper found that XGBoost and Light GBM were the top-performers with 98% accuracy…….", however the table shows Light GBM has 74% accuracy.
The results do show that XGBoost and Light GBM were the top-performers with 98% accuracy (see Table 4 in the updated manuscript, which was Table 5 in the previous manuscript).
In the previous manuscript, Table 4 showed that Light GBM has 74% accuracy before optimization (see #6 for explanation of "before optimization"). However, Table 5 showed that Light GBM has 98% accuracy after optimization.
To avoid confusion, the updated manuscript only presents the final model performance results after optimization (see Table 4 in the updated manuscript).

8. What percent of the defaulted loans were originated between 2007 and 2010 (the period of the subprime mortgage crisis) ? If a majority of defaulted loans origenated during this period, then the market's faulty risk pricing methodology is more to blame rather than consumer centric features. Explain in the manuscript why just as good a prediction may not be achieved by including this faulty risk pricing strategy in the model, instead of consumer centric features.
All loans in this paper's dataset were originated in 2020 and their performance was observed until September 2023. Therefore, **<u>no</u> loans were originated between 2007 and 2010, which is the period of the subprime mortgage crisis**. *The issue of faulty risk pricing in that era, while important for historical context, is not applicable to the dataset in this study.* The data in this study instead reflects more recent underwriting standards and macroeconomic conditions. *The study also captures borrower behavior and loan performance under the current market environment, which is of greater relevance* for present and future risk assessment than retrospective patterns tied to past crises.
Moreover, because the dataset does not include loans from the crisis period, the research's prediction models rely on consumer-centric and macroeconomic features (such as borrower profile and local economic indicators) rather than the mispricing of credit risk that characterized the subprime era. This ensures that the study's findings are driven by borrower-level and environmental risk factors, which are more relevant for current and forward-looking risk management, rather than by structural pricing failures unique to the 2007–2010 market.
To avoid confusion and acknowledge your comments, the description of data overview and future research has been updated in the manuscript.
BEFORE (previous manuscript)

To achieve the goal of developing accurate mortgage loan default models, **the study employs data from Freddie Mac in 2020**. The data has more than 100,000 loans, which are mostly 30-year fixed-rate mortgages. The data **was extracted from the original** public loan-level dataset from Freddie Mac with 54.1 million mortgages originated between 1999 and 2024 (6).

AFTER (updated manuscript)

To achieve the goal of developing accurate mortgage loan default models, the study employs a large dataset of **mortgage loans originated in 2020, with loan performance observed up to September 2023.** The data has more than 100,000 loans, which are mostly 30-year fixed-rate mortgages and extracted from Freddie Mac's public loan-level data (6). The dependent variable is Mortgage Loan Default Status, which is defined as 1 if the mortgage loan defaults with missing payments for at least 6 months and 0 otherwise.

Future Research

Future research could extend the current study by examining loan performance under historical or simulated economic stress scenarios. The dataset used in this study comprises loans originated in 2020, with performance observed through September 2023, and therefore does not include loans from earlier periods, such as the 2007–2010 subprime mortgage crisis. While historical risk mispricing during that period is informative for understanding past systemic vulnerabilities, it is not directly relevant to the more recent underwriting standards and macroeconomic conditions reflected in the current data. To enhance the robustness and generalizability of the proposed models, future work could add stress tests using simulated environments or historical scenarios characterized by extreme conditions, such as high inflation rates exceeding 10% or characteristics of subprime mortgage loans, to support more resilient risk assessment frameworks for banks, financial institutions, and regulators.

Thank you for addressing my comments. Could you please answer the first part of question/comment 6?

Thank you for your comments! I appreciate your feedback. Please review my responses to your concerns. The manuscript has been updated to reflect the changes and address your concerns. I hope you will approve the changes and accept the paper as it contributes to the financial industry more robust and advanced machine learning models with boosting algorithms for credit risk modeling with remarkable accuracy. The abstract is also updated in the manuscript to reflect the changes and address your comments.

9. was a stratified split performed? If not, then explain what procedure was used to achieve equal representation of good and bad loans in both the training and the testing data sets. Please discuss in the manuscript.

   Yes, a stratified split was performed to achieve equal representation of good and bad loans in both the training and the testing data sets. Please see #2 for the detailed explanation added to the manuscript (Section 2.2.Data Preprocessing).

10. How was class imbalance between the good (90%) and bad (10%) loans addressed? SMOTE? data augmentation? over/under sampling? Please discuss in the manuscript.

    To mitigate the problem of class imbalance between default and non-default loans (10.5% vs. 89.5%), a stratified data partition was used. To further balance the dataset and ensure equal representation of "good" and "bad" loans in the training and testing datasets, the Synthetic Minority Oversampling Technique (SMOTE) was applied to synthetically increase the minority class, in combination with under-sampling the majority class. Specifically, for each minority instance, SMOTE identifies its k nearest neighbors, randomly selects one neighbor to gather a synthetic sample; this process is repeated until the minority class reaches the desired level of representation relative to the majority class (Chawla et al., 2002). SMOTE addresses class

imbalance by creating synthetic minority samples using an interpolation approach, which helps mitigate overfitting that can result from merely duplicating minority observations. As a result of applying SMOTE to increase the minority class and under-sampling the majority class, the final balanced dataset of more than 65,000 loans achieved an equal distribution of 50% loan defaults and 50% non-default loans across both training and testing data.

11. How was the robustness of the model determined? K-fold cross-validation? adversarial training? Lipchitz constant? Please discuss in the manuscript.
    Please see below for the description of K-fold cross-validation added to the manuscript (Section 4.1.Data Modeling and Cross-Validation).

    K-fold cross-validation is a resampling technique for evaluating the generalizability of predictive models. The key benefit of cross-validation methods is to evaluate model performance on new data. One of its benefits is to reduce the overfitting risk that often arises in the presence of high-dimensional or noisy datasets. Using this approach, the data are randomly partitioned into K equal subsets, or "5 folds". In each iteration, (K -1) folds (i.e., four folds in 5-fold cross-validation) are used to train the model, while the remaining fold is held out for validation. This process is repeated K times. The results from all iterations are ultimately averaged to produce an overall estimate of the model. A 5-fold cross-validation procedure was employed to evaluate all four classification models in this study. The 5-fold approach provides a balance between computational efficiency and reliable performance estimation.
12. The feature 'inflation' is among the first 7 in terms of imporatance for lightGBM and XGBoost. This feature will change with time. Explain how – assuming+ the inflation is significantly different from today - the model is expected to perform from a nearly constant inflation regime from 1999 to 2024 as compared to outliers such as > 13% in the 1970's. This also ties back to point # 3 as to the robustness of the model.
    Thank you for the comment. In the dataset, **all loans are originated in 2020 with performance tracked through September 2023** (please see #8 below for the updated manuscript of overview of the data). Therefore, the inflation feature is not hypothetical but corresponds to the **realized local inflation rates in each loan's geographical location** during this observation window.

    It is true that inflation is a time-varying macroeconomic variable, and indeed the model is trained in a regime with relatively moderate variation (2020–2023). If inflation levels were to shift substantially outside the training distribution - for example, extreme values such as the >13% observed in the 1970s - the predictive performance of the model could be affected, since tree-based methods such as LightGBM and XGBoost rely on patterns present in the training data.
    However, two mitigating factors are worth noting:

    4. **Model scope** – The current model is calibrated for **post-2020 loan cohorts**, making the observed inflation variation during that horizon **appropriate** for the prediction task at hand.
    5. **Robustness** – While inflation is an important feature, the model also integrates a **wide set of loan-level and borrower-level characteristics** that remain relevant across regimes. **This reduces reliance on inflation alone**.

    6. **Future Research** - For long-term deployment in different macroeconomic environments and real-world modeling in banks and financial institutions, the model should ideally be retrained or stress-tested using simulated or historical high-inflation scenarios (e.g., >10%)

to assess generalization. This would ensure robustness if inflation dynamics deviate strongly from the 2020–2023 regime. Stress-testing environment to show forward-looking awareness can be considered in the future study, which is mentioned and updated in the section of future study (please see #8 below for future research in the manuscript).

However, considering the current economy, a forward-looking horizon of five years, and within the scope of this analysis, 1970s-type inflation may not generalize and not likely to happen, the current model results based on both training and testing data should be reliable.

13. I don't understand what the feature "loan age" actually means. Does this mean the # of months into the loan that the default resulted? This does not make intuitive sense. Should I take this to imply that there is a 'loan window' between (say) 2 to 4 years when the loan has the greatest probability of default? Please explain this in the manucript.
    Thank you for raising this point. In the dataset, "Loan Age" refers to the elapsed time since loan origination up to September 2023 or until the time the loans were tracked, or until the payoff date (in the case that loans are paid off), whichever comes first (please see Section 2.3. Feature Extraction for "loan age" description in the manuscript).
    It does **not** represent the number of months until a loan defaulted. Rather, it captures how long a loan has been active. In this case, loan age is not meant to imply a fixed default window. Instead, it provides the information about the loan's maturity stage, together with other factors (such as borrower characteristics and macroeconomic conditions) that can determine default risk.
    In the sample, some loans were active and paid on schedule; some loans were paid in full (loan payoff); and some loans were active and had missing payments. Loans may have 2 months of missing payments, 3 months of missing payments, 6 months of missing payments or 18 months of missing payments or even longer. The number of months active or in delinquency is not fixed. Therefore, loan age is not the number of months until a loan defaulted.
    More importantly, "Loan Age" was not used to define default outcome. Even loan defaults (i.e. missing payments for at least 6 months) can have a varied window of loan age. Non-default loans, including both active loans that were paid on schedule and payoff loans, also have a varied window of loan age. Default events, defined as missing payments for at least six months, can occur across a wide range of loan ages. Loan age is not the age of loans that missed payments for 6 months. It could also be the age of loans that missed payments for 20 months in September 2023 as long as the loan performance was tracked until September 2023.
    Similarly, non-default loans, including both active loans that remain current and loans that were fully repaid, also span varied loan age intervals. In this sense, loan age is analogous to human age: just as age correlates with survival outcomes but survival outcomes were not based on age cut-off. Similarly, loan age may correlate with loan default but was not used to define loan default.
    To avoid confusion, the description of *loan age* has been added to the manuscript, stating simply **the elapsed time since loan origination up to September 2023 or until the time the loans were tracked, or until the payoff date (in the case that loans are paid off), whichever comes first,** not the number of months until default.
14. Please provide an explanation for why the LightGBM improves significantly more than XGBoost in performance after optimization despite having similar feature importance. What do you mean by 'before optimization" and "after optimization" ?

Optimization refers to “Hyperparameter tuning”. The whole Section 4.3. Hyper-parameter Optimization was dedicated to explain how hyper-parameters (i.e., the number of estimators, learning rate, max features, max depth, min samples split, min samples leaf) contributed significantly to the robustness and accuracy of the LightGBM model.
“Before optimization” means random parameters were used. “After optimization” means parameters have been tuned. The optimal tuning is essential to ensure that the boosting algorithm generalizes well across new and unseen data and achieves the best performance.

15. you state “…..The paper found that XGBoost and Light GBM were the top-performers with 98% accuracy…….”, however the table shows Light GBM has 74% accuracy.
The results do show that XGBoost and Light GBM were the top-performers with 98% accuracy (see Table 4 in the updated manuscript, which was Table 5 in the previous manuscript).
In the previous manuscript, Table 4 showed that Light GBM has 74% accuracy before optimization (see #6 for explanation of “before optimization”). However, Table 5 showed that Light GBM has 98% accuracy after optimization.
To avoid confusion, the updated manuscript only presents the final model performance results after optimization (see Table 4 in the updated manuscript).

16. What percent of the defaulted loans were originated between 2007 and 2010 (the period of the subprime mortgage crisis) ? If a majority of defaulted loans origenated during this period, then the market’s faulty risk pricing methodology is more to blame rather than consumer centric features. Explain in the manuscript why just as good a prediction may not be achieved by including this faulty risk pricing strategy in the model, instead of consumer centric features.
All loans in this paper’s dataset were originated in 2020 and their performance was observed until September 2023. Therefore, **<u>no</u> loans were originated between 2007 and 2010, which is the period of the subprime mortgage crisis**. *The issue of faulty risk pricing in that era, while important for historical context, is not applicable to the dataset in this study.* The data in this study instead reflects more recent underwriting standards and macroeconomic conditions. *The study also captures borrower behavior and loan performance under the current market environment, which is of greater relevance* for present and future risk assessment than retrospective patterns tied to past crises.
Moreover, because the dataset does not include loans from the crisis period, the research’s prediction models rely on consumer-centric and macroeconomic features (such as borrower profile and local economic indicators) rather than the mispricing of credit risk that characterized the subprime era. This ensures that the study’s findings are driven by borrower-level and environmental risk factors, which are more relevant for current and forward-looking risk management, rather than by structural pricing failures unique to the 2007–2010 market.
To avoid confusion and acknowledge your comments, the description of data overview and future research has been updated in the manuscript.
BEFORE (previous manuscript)
To achieve the goal of developing accurate mortgage loan default models, **the study employs data from Freddie Mac in 2020**. The data has more than 100,000 loans, which are mostly 30-year fixed-rate mortgages. The data **was extracted from the original** public loan-level dataset from Freddie Mac with 54.1 million mortgages originated between 1999 and 2024 (6).
AFTER (updated manuscript)
To achieve the goal of developing accurate mortgage loan default models, the study employs a large dataset of **mortgage loans originated in 2020, with loan performance observed up to September 2023.** The data has more than 100,000 loans, which are mostly 30-year fixed-rate mortgages and extracted from Freddie Mac’s public loan-level data (6). The dependent variable

is Mortgage Loan Default Status, which is defined as 1 if the mortgage loan defaults with missing payments for at least 6 months and 0 otherwise.

Future Research

Future research could extend the current study by examining loan performance under historical or simulated economic stress scenarios. The dataset used in this study comprises loans originated in 2020, with performance observed through September 2023, and therefore does not include loans from earlier periods, such as the 2007–2010 subprime mortgage crisis. While historical risk mispricing during that period is informative for understanding past systemic vulnerabilities, it is not directly relevant to the more recent underwriting standards and macroeconomic conditions reflected in the current data. To enhance the robustness and generalizability of the proposed models, future work could add stress tests using simulated environments or historical scenarios characterized by extreme conditions, such as high inflation rates exceeding 10% or characteristics of subprime mortgage loans, to support more resilient risk assessment frameworks for banks, financial institutions, and regulators.

_____

Thank you for addressing my comments. Accepted.