A Machine Learning Approach in US Mortgage Risk Analysis

Zhan (Sam) Shi
University of Pennsylvania
Philadelphia, United States
samzshi@seas.upenn.edu

ABSTRACT

Mortgage loans, pivotal in real estate acquisition, necessitate meticulous risk analysis and forecasting, significantly influencing stakeholders and shaping future market trends. Accurate predictions in this area are crucial for evaluating the risk of Mortgage-Backed Securities (MBS) — collections of mortgages whose collapse, as demonstrated during the 2008 financial crisis, can trigger widespread economic repercussions. Contemporary public research in this field often suffers from a lack of explainability or depends heavily on voluminous data for precise forecasts. This independent study proposes a machine learning approach to improve mortgage risk assessment by: (1) creating interpretable models that identify the top 10 features escalating risk, (2) validating the effectiveness of a Two-Layer Long Short-Term Memory (LSTM) model in producing dependable predictions with less data, and (3) introducing a novel loan-level framework for mortgage risk evaluation.

1 INTRODUCTION AND BACKGROUND INFORMATION

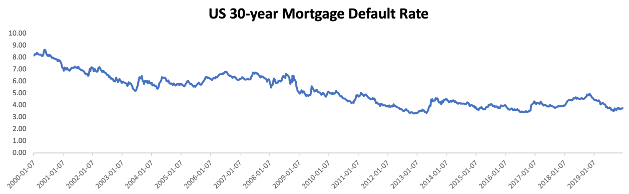


Figure 1: 30-Year Fixed Rate Mortgage Average in the United States [8].

During the 2008 financial crisis, the collapse of the housing market precipitated widespread defaults on subprime mortgages. These mortgages were often aggregated into Mortgage-Backed Securities (MBS), spotlighting the inherent risks in mortgage lending and investment practices. In response to this economic turmoil, several macroeconomic strategies were proposed to mitigate these risks and address the underlying vulnerabilities of the economy.

A pivotal legislative response during President Obama's administration in 2010 was the enactment of the *Dodd-Frank Wall Street Reform and Consumer Protection Act* [18]. Within this framework, the *Mortgage Reform and Anti-Predatory Lending Act* imposed stricter regulations on the mortgage system, enhancing oversight across both lending and securitization processes. The impact of this financial reform in reducing mortgage default risk is evident. For instance, as illustrated in Figure 1, the US 30-Year Mortgage Default Rate gradually decreased from 5.68% in January 2008 to 3.72% in January 2020.

1.1 Data and Resources

The source code and datasets of economic indexes used in this research are available at the following link: https://github.com/samzshi0529/A-Machine-Learning-Approach-for-US-Mortgage-Risk-Analysis.

1.2 Important Terms

This subsection explains about key terms important to mortgage finance and the broader financial context surrounding it. A comprehensive understanding of these concepts is vital for grasping the core content of this paper, particularly in interpreting the final results presented.

1.2.1 Mortgage Related Terms.

Mortgage. A mortgage is an agreement between you and a lender that gives the lender the right to take your property if you fail to repay the money you've borrowed plus interest[2].

Mortgage Backed Security. an MBS is a security collateralized by a discrete pool of mortgage loans, with payments based primarily on the performance of those loans[10].

Mortgage Charge-Off. A "mortgage charge-off" occurs when a lender removes a mortgage loan from their balance sheet as uncollectable, usually after the borrower defaults on payments and the lender forecloses. This process significantly contributes to the reduction of outstanding mortgage debt, affecting financial statistics and measures related to household liabilities and personal saving[9].

Mortgage Default. If you miss one or more payments on your mortgage loan, your loan is considered to be in default. [11]

1.2.2 Economic Terms.

GDP-Gross Domestic Product. According to the U.S. Bureau of Economic Analysis, GDP is the value of the goods and services produced in the United States. It serves as a key measure for Americans to gauge how their economy is doing and is also observed globally as an economic indicator. GDP represents the value and composition of the nation's output, the types of income generated, and how that income is used[15].

CPI-Consumer Price Index. As defined by the U.S. Bureau of Labor Statistics, the CPI measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. It provides indexes for the U.S. and various geographic areas and includes average price data for select utility, automotive fuel, and food items [16].

1.3 Literature Review

Extensive research has been conducted annually on mortgage loan default prediction. These studies range from forecasting immediate default status [3, 14] to projecting the ultimate loan outcome, such as charge-off status [13].

These investigations have employed methodologies including Deep Neural Networks [3, 14], Convolution Neural Networks [14], and Long Short-Term Memory (LSTM) components [13, 14]. In terms of data sources, the research predominantly utilizes datasets from the two largest mortgage lending institutions in the United States: 1) Fannie Mae's single-family loan performance data and 2) Freddie Mac's single-family loan performance data.

A main challange across these studies is the of class imbalance within the datasets. Various approaches, including sampling methods and cost-sensitive learning, have been proposed to mitigate this issue. However, several gaps in the literature persist. Firstly, there is a notable lack of interpretability in these models. Due to the usage on complex deep learning algorithms, these models often sacrifice interpretability compared to traditional machine learning techniques like Logistic Regression or Random Forests. Secondly, the majority of these models are built on extensive datasets, often exceeding billions of rows [3], or utilize large-scale neural network architectures [13], which necessitate prolonged periods for data processing and model training. Thirdly, there is a deficiency in the development of practical, systematic solutions that can be readily implemented by companies for evaluating mortgage risk.

1.4 Methodologies

This study mainly utilizes the Freddie Mac single-family loan performance dataset[6] to address three identified gaps in current research.

1.4.1 Enhancing Interpretability. To enhance interpretability, often lacking in deep learning approaches, we employ traditional machine learning techniques such as logistic regression and XGBoost algorithms. The study focuses on extracting the top 10 risk-augmenting features from these models, based on normalized feature importances, to provide clearer insights.

1.4.2 Reducing Computational Resources. The deep learning models in current studies often require substantial computational resources, either due to large datasets or complex neural network architectures. This research proposes a more resource-efficient approach. By implementing under-sampling and cost-sensitive learning, we develop a simple two-layer Long Short-Term Memory (LSTM) neural network. This model aims to match the predictive performance of more complex systems but with significantly reduced data and structural complexity.

1.4.3 Introducing a New Loan-Level Risk Model. Existing mortgage risk models in the studies predominantly emphasize predictive accuracy but often overlook the operational challenges associated with need for regular updating. This research begins by building on the model presented by Huang et al. (2023)[13], which utilizes monthly updated probabilities of loan charge-offs as mortgage risks. Although this approach marks a significant advancement in precision, it still faces limitations in terms of model complexity and need for frequent updates. To address these shortcomings, our study introduces an innovative Loan-Level Risk Model, designed to achieve even higher precision and fewer updates with a simpler model structure, thus optimizing both predictive quality and operational efficiency. The detailed methodology, along with a comprehensive evaluation of the model's effectiveness and its broader implications, will be discussed in Section 6.

2 DATA GATHERING

The foundation of this study is the hypothesis that mortgage risk is linked to both individual loan characteristics and the broader financial status of borrowers, as reflected by macroeconomic indicators. Primary data for individual loan characteristics is sourced from the Freddie Mac single-family loan performance dataset[6], which includes essential variables such as Current Loan Delinquency Status—indicating the duration of borrower payment delinquency—and borrower credit scores.

To comprehensively assess the financial status of borrowers, the study integrates macroeconomic indicators. These indicators are assumed to be proxies for the borrower's financial environment, influencing their ability to meet mortgage obligations. They include the GDP-Based Recession Index[15], representing the overall economic climate; the Unemployment Rate[17], reflecting employment stability; the State-Level Housing Price Index[5], indicative of housing market conditions; and the Consumer Price Indexes[16], as a measure of inflation and purchasing power. A deterioration in these macroeconomic factors is assumed to correlate with worse US citizens financial stability and heightened mortgage risk.

The time scope of the study spans from 2011 to 2019 to avoid the influences of the financial crisis and the Covid-19 pandemic. For analytical expediency, only 30-year mortgages are considered.

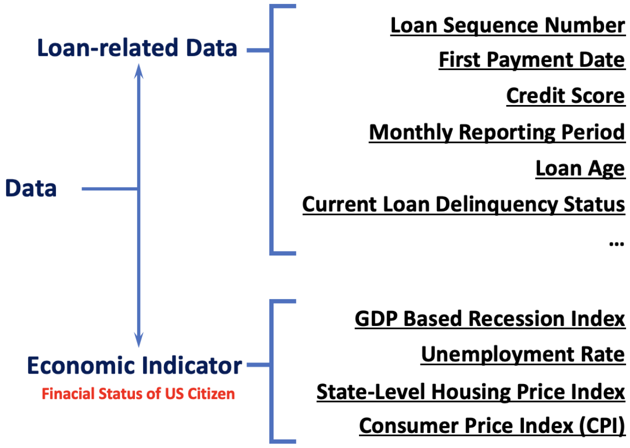

Loan-related Data
Loan Sequence Number
First Payment Date
Credit Score
Monthly Reporting Period
Loan Age
Current Loan Delinquency Status
...
Data
Economic Indicator
Finacial Status of US Citizen
GDP Based Recession Index
Unemployment Rate
State-Level Housing Price Index
Consumer Price Index (CPI)

Figure 2: Structure of Data Gathering.

2.1 Loan-Related Data

As mentioned earlier, Freddie Mac single-family loan performance dataset, which will be called Freddie Mac dataset in the rest of this paper, is used. The Freddie Mac dataset consists of two files for each quarter:

2.1.1 Origination Data File. Within each quarter, there is a origination data file in the format like "historical_data_2010Q1.txt" which includes the original loan information of the loans created at that quarter. As shown in Table 1, there 31 features in the origin data file. The data in these files won't change after its creation.

#	Column	Dtype
0	Credit Score	float64
1	First Payment Date	str
2	First Time Homebuyer Flag	object
3	Maturity Date	str
4	Metropolitan Statistical Area (MSA)	float64
5	Mortgage Insurance Percentage (MI %)	float64
6	Number of Units	float64
7	Occupancy Status	object
8	Original Combined Loan-to-Value (CLTV)	float64
9	Original Debt-to-Income (DTI) Ratio	float64
10	Original UPB	float64
11	Original Loan-to-Value (LTV)	float64
12	Original Interest Rate	float64
13	Channel	object
14	Prepayment Penalty Mortgage (PPM) Flag	object
15	Amortization Type (Formerly Product Type)	object
16	Property State	object
17	Property Type	object
18	Postal Code	float64
19	Loan Sequence Number	object
20	Loan Purpose	object
21	Original Loan Term	float64
22	Number of Borrowers	float64
23	Seller Name	object
24	Servicer Name	object
25	Super Conforming Flag	object
26	Pre-HARP Loan Sequence Number	object
27	Program Indicator	object
28	HARP Indicator	object
29	Property Valuation Method	float64
30	Interest Only (I/O) Indicator	object
31	Mortgage Insurance Cancellation Indicator	object

Table 1: Columns of Origin Data File

2.1.2 Monthly Performance File. Within each quarter, there is a file in a different format like "historical_data_time_2010Q1.txt" which includes the monthly updated loan performance information of the loans created at that quarter. As shown in Table 2, there are 11 features in the origin data file after the unrelated columns are removed. New data for loans created in the corresponding quarter will be appended in the files every month until the loans' lifetime ends.

#	Column	Dtype
0	Loan Sequence Number	object
1	Monthly Reporting Period	str
2	Current Actual UPB	float64
3	Current Loan Delinquency Status	object
4	Loan Age	float32
5	Remaining Months to Legal Maturity	float32
6	Defect Settlement Date	str
7	Zero Balance Code	float32
8	Current Interest Rate	float64
9	Delinquency Due to Disaster	object
10	Estimated Loan-to-Value (ELTV)	float32
11	Interest Bearing UPB	float64

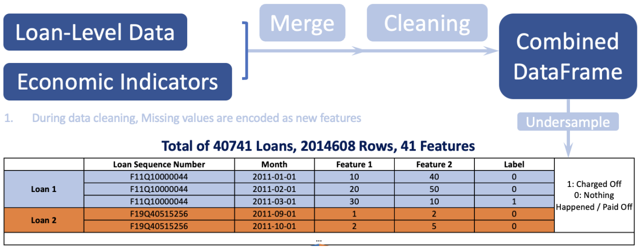
Table 2: Columns of Monthly Performance File

2.2 Macroeconomic Indicators

This study employs four macroeconomic indicators to represent the financial status of the average US citizen and to aid in predicting mortgage risks: the GDP-Based Recession Index[15], the Unemployment Rate[17], the State-Level Housing Price Index[5], and the Consumer Price Index[16]. The State-Level Housing Price Index is sourced from Freddie Mac, while the other three indicators are obtained from the Federal Reserve Economic Data (FRED). These sources are authoritative government institutions, ensuring both credibility and frequent usage in economic analyses. Utilizing these reliable data sources enhances the quality and replicability of the model developed in this research.

3 DATA PREPROCESSING

The data preprocessing, as depicted in Figure 3, comprises three primary stages: (1) concatenating and merging all loan-level data with economic indicators; (2) cleaning the dataset, which includes handling missing values; and (3) processing and preparing the final dataframe, involving tasks such as categorical variable encoding, creation of label/target vectors, and undersampling. This comprehensive preprocessing approach culminates in a refined dataframe for subsequent analysis, containing 40,741 loans, 2,014,608 rows, and 41 features per row, representing approximately 1% of the initial dataset size.


Loan-Level Data
Economic Indicators
Merge
Cleaning
Combined DataFrame
1. During data cleaning, Missing values are encoded as new features
Undersample
Total of 40741 Loans, 2014608 Rows, 41 Features
1: Charged Off
0: Nothing Happened / Paid Off

	Loan Sequence Number	Month	Feature 1	Feature 2	Label
Loan 1	F11Q10000044	2011-01-01	10	40	0
	F11Q10000044	2011-02-01	20	50	0
	F11Q10000044	2011-03-01	30	10	1
Loan 2	F19Q40515256	2011-09-01	1	2	0
	F19Q40515256	2011-10-01	2	5	0
	...				

Figure 3: Pipeline of Data Preprocessing.

3.1 Data Importation and Undersampling

The importation process for loan-level data from Freddie Mac datasets selectively includes only those loans that have concluded their lifecycle, specifically those that are either paid off or charged off. To address the potential imbalance between the classes, paid-off loans are subjected to random undersampling to match the quantity of charged-off loans, ensuring a balanced dataset. Consequently, a total of 80,232 unique loans (40116 paid off and 40116 charged off loans) are imported for preprocessing.

3.2 Handling of Missing Values

#	Column	Missing Value
0	Metropolitan Statistical Area (MSA)	26.495663
1	Pre-HARP Loan Sequence Number	51.231429

Table 3: Missing Value Distribution

After addressing missing data in specific columns, such as *Super Conforming Flag* and *HARP Indicator*, missing values ('nan') and existing values ('Y') were transformed into binary representations (0 and 1, respectively). Despite these transformations, two columns continued to exhibit missing data as shown in Table 3. Owing to their categorical nature and the broad spectrum of potential values, these two columns were removed from the dataframe.

A particular focus was placed on the *Valid DTI Ratio* column. According to Huang et al. (2023) [13], the conventional approach when encountering a *Valid DTI Ratio* of 0—a sign of missing data—is to exclude this column from the analysis. However, an exploratory analysis identified a notable trend: loans labeled as *Charged Off* displayed a significantly higher incidence of missing *DTI Ratio* values compared to those labeled as *Paid Off* (24,901 vs. 5,840 instances, respectively). This pattern suggests a potential link between missing *DTI Ratio* values and mortgage risk, leading to the inclusion of this column in the current study.

3.3 Categorical Variable Encoding

Columns representing categorical variables with an excessively high count of unique values are excluded from the dataset. The remaining categorical variables undergo one-hot encoding, transforming them into distinct features. These include *Channel*, *Loan Purpose*, *First Time Homebuyer Flag*, *Occupancy Status*, *Property Type*, and *Property Valuation Method*.

3.4 Label / Target Feature Creation

#	Reason for Loan Termination	Zero Balance Code
1	REO Disposition	09
2	Short Sale or Charge Off	03
3	Third Party Sale	02
4	Prepaid or Maturity (Paid Off)	01

Table 4: Overview of Zero Balance Code

In accordance with the methodology for label generation employed by Huang et al. (2023) [13], the target feature is derived from the *Zero Balance Code* column. While the meaning of *Zero Balance Code* is presented in Table 4, a label of 1 is assigned to denote charge-off status of a mortgage if the *Zero Balance Code* is 02, 03, or 09, indicating that the mortgage has been charged off. Conversely, a label of 0 is assigned to all other instances, signifying that the mortgage has not been charged off or has been paid off.

3.5 Finalized Dataframe

#	Variable	Data Type
1	Current Actual UPB	float64
2	Current Loan Delinquency Status	int64
3	Loan Age	float64
4	Remaining Months to Legal Maturity	float64
5	Current Interest Rate	float64
6	Delinquency Due to Disaster	int64
7	Interest Bearing UPB	float64
8	Original UPB	float64
9	Mortgage Insurance Percentage (MI %)	float64
10	Original Loan-to-Value (LTV)	float64
11	Original Interest Rate	float64
12	Super Conforming Flag	int64
13	Credit Score	float64
14	Original Debt-to-Income (DTI) Ratio	float64
15	Number of Borrowers	float64
16	Number of Units	float64
17	Valid DTI Ratio	int64
18	Housing Price	float64
19	CPI	float64
20	Unemployment Rate	float64
21	Recession	float64
22	Channel_B	bool
23	Channel_C	bool
24	Channel_R	bool
25	Loan Purpose_C	bool
26	Loan Purpose_N	bool
27	Loan Purpose_P	bool
28	First Time Homebuyer Flag_N	bool
29	First Time Homebuyer Flag_Y	bool
30	Occupancy Status_I	bool
31	Occupancy Status_P	bool
32	Occupancy Status_S	bool
33	Property Type_CO	bool
34	Property Type_CP	bool
35	Property Type_MH	bool
36	Property Type_PU	bool
37	Property Type_SF	bool
38	Property Valuation Method_1.0	bool
39	Property Valuation Method_2.0	bool
40	Property Valuation Method_3.0	bool
41	Property Valuation Method_9.0	bool

Table 5: Data Type Specification for Loan Features

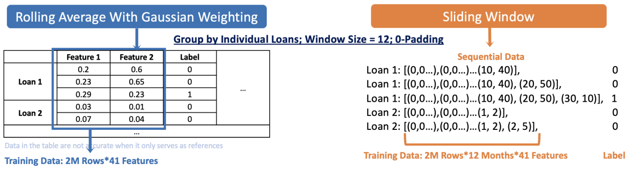
Following the aforementioned data processing steps, the finalized dataframe comprises 2,014,608 observations across 44 variables. Of these, three key columns are *Loan Sequence Number*, *Monthly*

Reporting Period, and *Label*. For modeling purposes, *Loan Sequence Number* and *Monthly Reporting Period* are excluded from the feature set, while the *Label* column is designated as the target variable, denoted by y. The remaining 41 attributes serve as the feature matrix, detailed in Table 5.

4 TOP 10 RISK-AUGMENTING FEATURES

This study aims to provide the key factors that amplify mortgage risks by identifying the top 10 influential features. To achieve this, we employ traditional machine learning algorithms, Logistic Regression and XGBoost, known for their efficacy in feature importance extraction.

4.1 Training Data Reconstruction


Rolling Average With Gaussian Weighting
Sliding Window
Group by Individual Loans; Window Size = 12; 0-Padding
Data in the table are not accurate when it only serves as references
Training Data: 2M Rows*41 Features
Sequential Data
Training Data: 2M Rows*12 Months*41 Features
Label

	Feature 1	Feature 2	Label	
Loan 1	0.2	0.6	0	...
	0.23	0.65	0	
	0.29	0.23	1	
Loan 2	0.03	0.01	0	
	0.07	0.04	0	
	...			

```
Loan 1: [(0,0...),(0,0...)...(10, 40)],                          0
Loan 1: [(0,0...),(0,0...)...(10, 40), (20, 50)],                0
Loan 1: [(0,0...),(0,0...)...(10, 40), (20, 50), (30, 10)],      1
Loan 2: [(0,0...),(0,0...)...(1, 2)],                            0
Loan 2: [(0,0...),(0,0...)...(1, 2), (2, 5)],                    0
```

Figure 4: Feature Engineering for Part 4 and Part 5

As illustrated in Figure 4, the training dataset is subject to a systematic re-processing procedure. This entails the implementation of a sliding window technique, defined by a window size of 12 and a stride of 1. In addition, zero-padding is applied to ensure consistent window dimensions across the dataset. The data is further segmented on the basis of individual loans. For each window, a rolling average is calculated, employing Gaussian weighting to derive a new data point. The parameters of the sliding window—namely the window size, stride, padding technique, and weighting method—have been meticulously optimized to achieve the most effective combination.

4.2 Modeling

The modeling process incorporates three methodologies based on the processed dataset. These include: (1) logistic regression utilizing all 41 features, (2) a reduced logistic regression model employing backward elimination, and (3) the XGBoost algorithm. The choice of backward elimination for dimensionality reduction over more common techniques such as Principal Component Analysis (PCA) is deliberate. This approach preserves the interpretability of the model, a crucial aspect potentially compromised by PCA.

4.3 Evaluation of Model Performance

The comparative analysis of the three models reveals a notable similarity in their performance metrics. This observation potentially suggests that more complex models may not yield significant improvements in this context.

4.4 Results

The normalization of feature importances across all three models, including coefficients from logistic regression models and the 'gain' from the XGBoost model (which represents the incremental improvement in accuracy contributed by a feature across the branches it influences), facilitates a comprehensive comparison. By consolidating these normalized feature importances into a singular plot and arranging them in descending order of their cumulative importance, we obtain Figure 5, as illustrated below.

Table 6: Evaluation Matrices for Three Models

Logistic Regression			
	Precision	Recall	F1-Score
0	1.00	0.92	0.96
1	0.09	0.90	0.17
Macro Avg	0.55	0.91	0.57
Weighted Avg	0.99	0.92	0.95
Reduced Logistic Regression			
	Precision	Recall	F1-Score
0	1.00	0.92	0.96
1	0.09	0.90	0.17
Macro Avg	0.55	0.91	0.57
Weighted Avg	0.99	0.92	0.95
XGBoost			
	Precision	Recall	F1-Score
0	1.00	0.92	0.96
1	0.09	0.91	0.17
Macro Avg	0.55	0.92	0.56
Weighted Avg	0.99	0.92	0.95

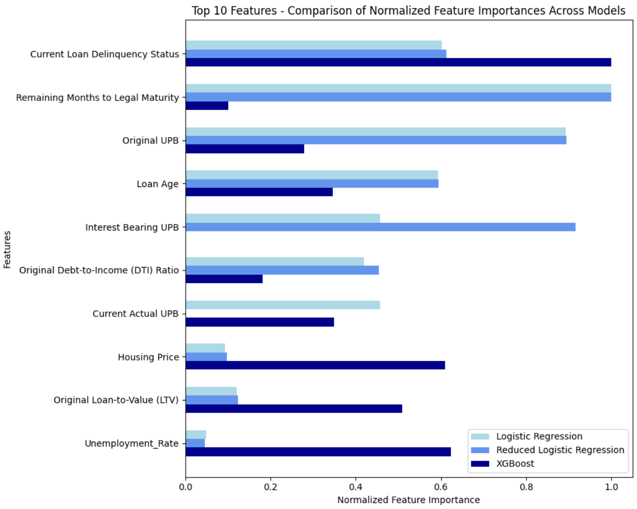


Figure 5: Top 10 Risk-Augmenting Features

5 LSTM-BASED PREDICTION MODEL

Addressing the challenge identified in previous studies, which involves the extensive data requirements, this section introduces an LSTM-based neural network characterized by a relatively small number of parameters. This network has been trained on a selectively undersampled dataframe, as previously discussed.

5.1 Training Data Reconstruction

As illustrated in Figure 4, the training dataset is subject to a systematic re-processing procedure. This entails the implementation of a sliding window technique, defined by a window size of 12 and a stride of 1. In addition, zero-padding is applied to ensure consistent window dimensions across the dataset. The data is further segmented on the basis of individual loans. Each window then becomes the new data point for our LSTM Neural Network. The parameters of the sliding window—namely the window size, stride, padding technique, and weighting method—have been meticulously optimized to achieve the most effective combination.

5.2 Modeling

The methodology encompassed the development of a two-layer Long Short-Term Memory (LSTM) neural network, incorporating cost-sensitive learning through class weighting, applied to the pre-processed dataset. This specific architectural choice exhibited enhanced predictive capabilities, surpassing those of alternative models evaluated in this study. A detailed schematic of the network's architecture is shown in Figure 6.

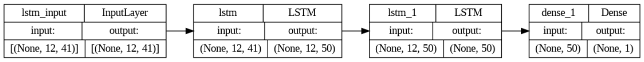


Figure 6: Two-layer LSTM Neural Network

5.3 Evaluation of Model Performance

Table 7: Evaluation Matrices for LSTM Neural Network

LSTM Neural Network			
	Precision	Recall	F1-Score
0	1.00	0.90	0.95
1	0.08	0.98	0.15
Macro Avg	0.54	0.94	0.55
Weighted Avg	0.99	0.90	0.94

As demonstrated in Table 7, there is a notable enhancement in the performance of the Long Short-Term Memory (LSTM) neural network, as evidenced by the increase in recall from 0.91 in the previously utilized Extreme Gradient Boosting (XGBoost) model to 0.98. The predictive performance of the LSTM model is further illustrated through the analysis of two specific loans, as depicted in Figures 7 and 8.

Concerning Figure 7, it presents an exemplar of a 'good' loan, characterized by the full payment of interests after 125 months. Throughout its duration, the predicted probability of the mortgage being charged off in the subsequent month, or the anticipated risk, consistently remained low, not exceeding 0.05. This observation corroborates the classification of this loan as 'good'.

Regarding Figure 8, it provides an illustrative example of what we consider a 'bad' loan, as evidenced by its charge-off occurrence within just 17 months. Throughout the initial 14 months, the forecasted probability of the mortgage undergoing a charge-off in the subsequent month, representing the anticipated risk, consistently remained quite low, never exceeding 0.1. However, as we approach the 15th month, a noticeable shift occurs, with the risk beginning to rise above 0.5. This graphically demonstrates that our prediction not only accurately anticipates the eventual charge-off in the 17th month but also effectively signals potential issues in the preceding months leading up to the charge-off event. This observation strongly supports the classification of this loan as 'bad'.

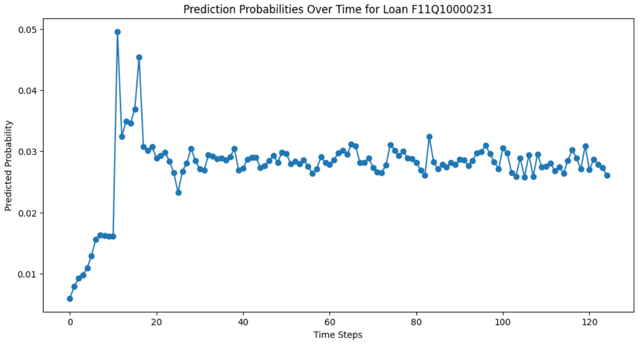


Figure 7: Loan F11Q10000231: Paid Off on month 125

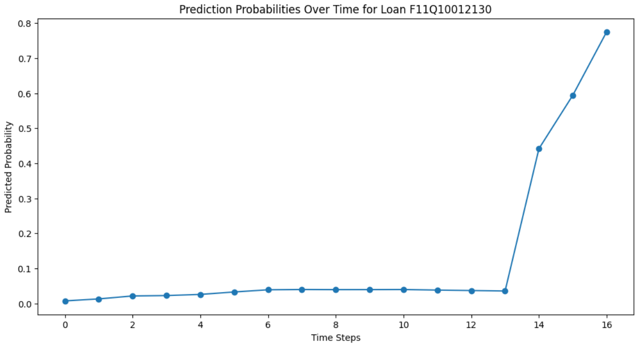


Figure 8: Loan F11Q10012130: Charged Off on month 17

6 ADVANCEMENTS IN MORTGAGE RISK MODELING: LOAN-LEVEL PREDICTION ANALYSIS

Table 8: Precision for Class 1 Across Various Models

Model	Precision (Class 1)
LSTM Neural Network	0.08
Logistic Regression	0.09
Reduced Logistic Regression	0.09
XGBoost	0.09

A critical problem in the models involved in this study right now is the uniformly low precision, as illustrated in Table 8. All models exhibit precision values below 0.1. This trend is attributable to the intrinsic nature of mortgage risk evaluation, where loans often exhibit a high risk of charge-off over several months before actual charge-off occurs. For instance, Figure 8 demonstrates that while our model accurately predicts risk in the 17th month, the precision significantly declines due to elevated risk levels in the 15th and 16th months.

To address this, we propose a new loan-level risk model on a threshold criterion. Specifically, a threshold of 0.8 is established. Loans with a predicted probability exceeding 0.8 of charge-off at the first time in their lifecycle are classified as high risk (1) and subsequently excluded from future computations; otherwise, they are deemed low risk (0).

This methodology yields remarkably accurate results, as evidenced by the following evaluation:

Table 9: Loan-Level Evaluation Outcomes

Label	Precision	Recall
0	0.96	0.99
1	0.98	0.95

In summary, the efficacy of this new risk model is proven by its evaluation results. By marking a loan as high-risk and excluding it from subsequent calculations upon prediction of a high charge-off likelihood, we achieve considerably enhanced precision. Furthermore, this approach reduces the need for ongoing risk assessments of high-risk loans, thereby decreasing computational resource requirements.

7 CONCLUSION

This study has successfully addressed the three previously identified gaps in mortgage risk analysis. First, the application of traditional machine learning models has shown the top ten features that impact mortgage risk. Notably, the influence of housing prices and unemployment rates on mortgage risk has also been revealed, offering valuable insights. Second, we have demonstrated the feasibility of training robust mortgage risk prediction models using limited datasets. This finding is significant as it suggests that future research can efficiently utilize subsampled data for model training, thereby saving substantial time. Third, we have introduced a new loan-level risk model that effectively overcomes the issue of low precision observed in existing models. This advancement is designed to further reduce computational resources required for mortgage risk assessment.

8 FUTURE RESEARCH DIRECTIONS

8.1 Data Enhancement for Mortgage-Backed Securities (MBS) Risk Prediction

As previously discussed, a primary application of the risk models in this study is to enhance Mortgage-Backed Securities (MBS) predictions. For future research, obtaining a comprehensive dataset that details the specific mortgages constituting each MBS is crucial. This will enable a more accurate and detailed analysis of MBS risks.

8.2 In-depth Interpretation of Key Features

This research has demonstrated a method for deriving interpretable models using logistic regression and XGBoost. However, further investigations are required to provide more profound economic or financial interpretations of these models. While the current focus is on the top 10 features, it is possible to explore additional insights. These may include examining feature correlations and conducting trend analyses on feature importance. Such analyses could involve retraining the models over various time ranges to facilitate comparative studies. This deeper exploration can greatly enrich our understanding and contribute to economic and finance research.

9 ACKNOWLEDGEMENTS

I extend my sincere gratitude to my advisor, Brandon L. Krakowsky, for his invaluable guidance and support throughout this independent study. Special thanks are due to Dr. Paul Shaman, whose expertise in methodologies greatly contributed to this research. I also wish to express my appreciation to Kevin A. Thomas, Business & Data Analysis Librarian at Penn, for his assistance in dataset collection.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 785–794.

[2] Consumer Financial Protection Bureau. 2023. What is a mortgage? Consumer Financial Protection Bureau. https://www.consumerfinance.gov/ask-cfpb/what-is-a-mortgage-en-100/

[3] M.J. Cooper. 2018. A Deep Learning Prediction Model for Mortgage Default. University of Bristol.

[4] D. R. Cox. 1958. The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–242.

[5] Federal Home Loan Mortgage Corporation (Freddie Mac). 2023. *House Price Index*. https://www.freddiemac.com/research/indices/house-price-index Freddie Mac Research.

[6] Federal Home Loan Mortgage Corporation (Freddie Mac). 2023. *Single-Family Loan-Level Dataset*. https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset Freddie Mac Research.

[7] Federal Reserve Bank of Atlanta. 2023. *Sticky Price Consumer Price Index less Food and Energy [CORESTICKM159SFRBATL]*. https://fred.stlouisfed.org/series/CORESTICKM159SFRBATL Retrieved from FRED, Federal Reserve Bank of St. Louis.

[8] Federal Reserve Bank of St. Louis. 2023. 30-Year Fixed Rate Mortgage Average in the United States. Federal Reserve Bank of St. Louis: Economic Research. https://fred.stlouisfed.org/series/MORTGAGE30US Retrieved from FRED.

[9] Federal Reserve Board. 2023. Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks. Federal Reserve Board. https://www.federalreserve.gov/releases/chargeoff/

[10] Federal Reserve Board. 2023. Mortgage-Backed Securities (MBS). Federal Reserve Board: Glossary. https://www.federalreserve.gov/

[11] Federal Trade Commission. 2023. Missing Mortgage Payments: Default and Foreclosure. Federal Trade Commission: Consumer Advice. https://www.consumer.ftc.gov/articles/your-rights-when-paying-your-mortgage#Missing_Mortgage_Payments:_Default_and_Foreclosure

[12] James Hamilton. 2023. *GDP-Based Recession Indicator Index [JHGDPBRINDX]*. https://fred.stlouisfed.org/series/JHGDPBRINDX Retrieved from FRED, Federal Reserve Bank of St. Louis.

[13] Mengzhe Huang. 2023. *Freddie Mac Mortgage Default Prediction using Monthly Performance Data*. https://sites.google.com/nyu.edu/mengzhehuang/data-science-projects/freddie-mac-mortgage-default-prediction-using-monthly-performance-data

[14] Riskcare. 2019. *Predicting Mortgage Loan Delinquency Status with Neural Networks*. https://www.riskcare.com/connect/research/predicting-mortgage-loan-delinquency-status-with-neural-networks Accessed: 2023-12-04.

[15] U.S. Bureau of Economic Analysis. 2023. Gross Domestic Product. U.S. Bureau of Economic Analysis. https://www.bea.gov/data/gdp/gross-domestic-product

[16] U.S. Bureau of Labor Statistics. 2023. Consumer Price Index. U.S. Bureau of Labor Statistics. https://www.bls.gov/cpi/

[17] U.S. Bureau of Labor Statistics. 2023. *Unemployment Rate [UNRATE].* https://fred.stlouisfed.org/series/UNRATE Retrieved from FRED, Federal Reserve Bank of St. Louis.

[18] U.S. Congress. 2010. Dodd-Frank Wall Street Reform and Consumer Protection Act. Public Law 111-203. https://www.congress.gov/bill/111th-congress/house-bill/4173/text