A Work Project, presented as part of the requirements for the Award of a Master's degree in Management from the Nova School of Business and Economics.

THE ROLE OF CREDIT SCORE AND LTV IN PREDICTING BORROWER RISK: A LIGHTGBM APPROACH USING ELTV ACROSS COVID-19 PANDEMIC PERIODS

HEMALI SATISHCHANDRA

Work project carried out under the supervision of:

Professor Gonçalo Rocha

15/01/2025

Abstract

Assessment of borrower risk is essential in mortgage lending, with traditional mortgage metrics playing a critical role. However, the COVID-19 pandemic introduced unprecedented challenges, raising questions about their predictive power. This dissertation analyses the effectiveness of Credit Score and LTV in predicting borrower risk, using ELTV as a proxy across pre-pandemic, pandemic, and post-pandemic periods. Using a sample of approximately 1,047 loans from the U.S. mortgage market, LightGBM models were employed to evaluate these metrics. The findings highlight Credit Score's consistent superiority over LTV in predicting ELTV across all periods, underscoring its relevance in assessing borrower risk.

Keywords

Borrower Risk; Credit Score; Loan-to-Value Ratio; Estimated Loan-to-Value Ratio; Mortgage Lending; Predictive Analytics; Machine Learning; COVID-19 Pandemic.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DaaLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

Mortgage lending is a critical component of the housing market and the wider economy, driving socio-economic growth by providing households with access to homeownership (Bazilinska 2020). The mortgage sector alone accounts for more than 30% of GDP in numerous advanced economies, illustrating its pivotal role in wealth creation and economic stability (Federal Reserve 2022).

In this context, an accurate assessment of borrower risk is essential to predict loan performance. This enables institutions to take preventative measures, thereby maintaining financial stability and mitigating the risk of default. Historical imbalances in real estate markets have triggered severe financial crises, such as the 2008 global financial crisis, where widespread mortgage defaults led to the collapse of several financial institutions and to the subsequent economic downturns (Ospina and Uhlig 2018).

Despite empirical research identifying multiple factors influencing loan performance (Agarwal et al. 2015), such as negative equity and life events (Ganong et al. 2020), traditional metrics[1], such as Credit Scores[2] and Loan-to-Value (LTV) ratios play an important role in predicting borrower risk (Harrison et al. 2004).

LTV ratios have long been fundamental in evaluating the risk exposure of financial institutions, influencing capital stability during economic cycles (Avery et al. 2003). Similarly, Credit Scoring has become a widely used tool in loan origination, increasingly relied upon over the last few decades, driven by technological advances and improved information-sharing mechanisms (Altman and Saunders 1998). These metrics provide valuable insights into borrower reliability and market risk, enabling lenders to predict credit and prepayment risks with greater accuracy (Agarwal et al. 2015).

Nevertheless, their predictive power has been tested primarily under stable economic conditions (Avery et al. 2004). The COVID-19 pandemic introduced significant uncertainty that affected the mortgage market (Hari et al. 2020; Mansour et al. 2020), reshaping its landscape through large-scale economic interventions such as forbearance programs. These measures, which were unprecedented in scale, allowed borrowers to pause or reduce their payments, temporarily

[1] In this thesis, the terms 'metrics' and 'variables' are used interchangeably to refer to factors such as Loan-to-Value (LTV) and Credit Score. While 'metrics' emphasizes their quantitative nature, 'variables' reflects their role in predictive modelling. This interchangeable usage aligns with the context in which these terms are discussed.

[2] Credit Score is capitalized throughout this work to reflect its use as a formal mortgage risk metric studied in conjunction with LTV and ELTV for predicting borrower risk.

shielding them from default (Freddie Mac 2024). Such interventions raise critical questions about the predictive power of traditional metrics, such as Credit Score and LTV, during periods of economic upheaval.

Despite extensive research on these metrics, their effectiveness under these extraordinary conditions remains underexplored.

This work project seeks to fill these gaps by answering the following primary research question: How effective are Credit Score and LTV in predicting borrower risk, as measured by ELTV, using LightGBM (Light Gradient Boosting Machine) models, across three distinct pandemic periods: pre-pandemic, during-pandemic, and post-pandemic?

ELTV, a dynamic metric recalculated monthly, reflects the relationship between the outstanding loan balance and property value, providing ongoing insights into borrower equity and risk. Unlike the static Original Loan-to-Value (LTV) ratio, which remains fixed at loan origination, ELTV continuously updates to reflect changes in market conditions and borrower payments. These updates are based on property value estimates generated by Freddie Mac's Automated Valuation Model (AVM), Home Value Explorer (HVE), ensuring a more accurate and timely assessment of borrower risk.

Hence, using U.S. mortgage data from 2017 to 2023, this research examines the performance of these metrics, along with other potential variables influencing borrower risk, through LightGBM machine learning models with further categorization into higher, middle-lower, and lower importance groups.

In addition, this study explores the novel potential of using static loan origination metrics, such as Credit Score and LTV, to predict ELTV at the beginning of the loan lifecycle. By doing so, it extends the scope of traditional risk assessment frameworks by bridging short-term borrower reliability with long-term equity-based risk insights. This dual perspective provides a valuable tool for lenders to improve early risk assessment and adapt strategies during periods of economic uncertainty.

By linking static origination metrics with a dynamic, evolving measure such as ELTV through the LightGBM machine learning model, this work project aims to advance the understanding of borrower risk and contribute to a more adaptive and resilient mortgage risk management framework.

2. Literature Review

Section 2 provides a comprehensive review of the literature on borrower risk assessment in mortgage lending. It examines the role of traditional metrics, such as Credit Score and LTV, in predicting mortgage risk, highlighting their strengths and limitations. Additionally, it explores ELTV as a more dynamic alternative. Lastly, the section discusses the application of machine learning techniques, particularly LightGBM, in enhancing mortgage risk prediction.

2.1. Borrower Risk in Mortgage Lending

Borrower risk plays a central role in mortgage lending, as it determines the likelihood that borrowers will default on their obligations, which can lead to significant financial instability for lenders and the broader economy. Accurate assessment of borrower risk is essential for mitigating potential losses and ensuring the stability of financial institutions. Mortgage defaults have been a key driver of financial crises, such as the Subprime Mortgage Crisis or the 2008 global financial crisis, where inaccurate assessments of borrower risk contributed to widespread defaults and the collapse of numerous financial institutions (Ospina and Uhlig 2018; Duca and Muellbauer 2013).

Over the years, several theories have been proposed to explain mortgage default (Foster and Van Order 1984; Riddiough 1991; Goldberg and Capone 1998; Ganong et al. 2022). Similarly, numerous mortgage metrics have been employed, with Credit Scores and LTV historically serving as essential tools for evaluating borrower creditworthiness, offering a structured framework for categorizing applicants into risk tiers based on their financial history and the equity they hold in the collateral property (Galindo and Tamayo 2000). However, these static metrics are increasingly seen as insufficient in capturing the evolving dynamics of borrower behaviour due to an increasingly complex loan network, often restricting the assessment through traditional methods (Zhou et al. 2019; Liu et al. 2024). For instance, LTV, while valuable in stable markets, fails to reflect sudden changes in property values or shifts in borrowers' financial circumstances, leading to an underestimation of risk. Research highlights that high LTV, while indicative of greater borrower risk, do not account for broader economic factors such as market fluctuations or unemployment rates, which can exacerbate default risks (Makri et al. 2014; Avery et al. 2004; Bian et al. 2018).

Recent studies have suggested that incorporating borrower-specific variables, such as income volatility and debt-to-income ratios or macroeconomic factors, such as unemployment rates and

housing market conditions, can improve the prediction of defaults (Bierut et al. 2015). For instance, Bian et al. (2017) and Vasiliki et al. (2014) emphasize the importance of integrating these additional variables into borrower risk models, as they capture patterns of borrower delinquency that traditional metrics like Credit Score and LTV might overlook. During times of economic downturn, borrowers may face financial issues, such as job loss or health emergencies, which static models fail to account for (Farkas et al. 2020). This may lead to a growing recognition of the need for more dynamic and flexible approaches, such as Estimated Loan-to-Value (ELTV) and machine learning, which integrate real-time data and can adapt to changing economic conditions.

2.2. Credit Score

The Credit Score is a cornerstone of risk evaluation in mortgage lending, developed from historical financial behaviours, such as repayment history, credit utilization, and outstanding debt levels (Consumer Financial Protection Bureau 2024). Its primary utility lies in distinguishing between high- and low-risk borrowers, serving as a reliable benchmark for mortgage approvals (Chen et al. 2021). Nevertheless, as the financial landscape evolves, limitations of Credit Scores have become increasingly apparent. While effective in predicting borrower reliability during stable periods, Credit Scores rely on historical data and thus fail to account for sudden economic shocks, such as those brought on by the COVID-19 pandemic (Arya et al. 2011). Furthermore, non-traditional financial behaviours, such as timely rent or utility payments are not captured, disadvantaging individuals with limited credit histories, particularly, younger borrowers and immigrants (Goel et al. 2021; Buchanan and Wright 2021). These gaps highlight the need for a more dynamic and context-sensitive evaluation, which can be addressed by metrics like ELTV.

2.3. Loan-to-Value (LTV) Ratio

Loan-to-Value (LTV) Ratio is a widely used metric for assessing mortgage risk by comparing the loan amount to the appraised value of the property (Freddie Mac 2024). Higher LTV ratios generally indicate greater risk, as they reflect less borrower equity and increased lender exposure in the event of default (Hadžić 2016). Research has consistently shown that high-LTV loans are more prone to default, particularly during economic instability (Luis 2015). However, LTV's static nature is a key limitation. It fails to account for real-time fluctuations in property

values or borrower circumstances, which can result in inaccurate risk assessments during periods of market volatility (Demyanyk and Hemert 2011). Moreover, studies have shown that LTV can be skewed in overheated housing markets, where inflated property values mask the actual risk of default (Parlov and Wachter 2011). Recent studies also indicate that properties can be sold for amounts significantly higher than their collateral value, introducing bias into the LTV calculation and underestimating credit risk (Bian et al. 2018). These challenges emphasize the importance of dynamic metrics like ELTV, which incorporates real-time adjustments in both, property values and loan balances, to better assess borrower risk.

2.4. Estimated Loan-to-Value (ELTV) Ratio

Estimated Loan-to-Value (ELTV) Ratio is a dynamic metric updated monthly that reflects the ratio between the remaining loan balance and the property's value.

Thus, this metric represents an advancement over the traditional LTV metric by incorporating real-time adjustments to property valuations and outstanding loan balances, thereby enhancing its responsiveness to evolving market conditions. Unlike LTV, which remains fixed at the time of loan origination and does not account for subsequent market fluctuations, ELTV dynamically reflects changes in borrower equity (Freddie Mac 2024). This characteristic renders it a more precise indicator of risk, particularly during periods of economic instability, such as the global financial crisis or the COVID-19 pandemic.

Freddie Mac's Automated Valuation Model (AVM), Home Value Explorer® (HVE®), serves as the foundation for ELTV by employing an ensemble machine learning model to estimate current property values. This model synthesizes data from multiple sources, including loan records, public sales transactions, tax assessments, appraisal reports, and Multiple Listing Service (MLS) data (Freddie Mac 2024). By integrating these diverse datasets, ELTV offers a more comprehensive and timely representation of borrower equity fluctuations, thereby enhancing its utility as a risk assessment tool for financial institutions.

Despite its advantages, the widespread adoption of ELTV is contingent upon addressing challenges related to its accessibility and integration into predictive risk models. Further research is warranted to refine its methodological framework and evaluate its efficacy in mortgage risk assessment, as existing literature has yet to explore its full potential in this domain.

2.5. Machine Learning and LightGBM

Machine learning (ML) has become an essential tool in mortgage risk prediction, particularly in overcoming the limitations of traditional static metrics like Credit Score and LTV. Traditional models, which rely on simple linear relationships, often fail to capture complex borrower behaviour or market dynamics. ML models, especially ensemble techniques like Gradient Boosting Machines (GBM), excel at processing large, high-dimensional datasets, uncovering non-linear relationships between borrower characteristics, financial history, and market conditions that traditional models may miss (Chen et al. 2021). LightGBM, an implementation of gradient boosting, is particularly effective due to its efficiency, scalability, and ability to handle high-dimensional data with minimal preprocessing (Lundberg and Lee 2017).

Research has demonstrated that ML models outperform traditional models in predicting loan defaults (Chen and Guestrin 2016).

Furthermore, ML algorithms, including LightGBM, are particularly valuable in real estate price forecasting, offering real-time insights into market conditions, property value fluctuations, and borrower behaviour. This integration of multiple dynamic factors enables more accurate risk assessments (Hasan and Mahmood 2024). However, despite its effectiveness, LightGBM and other ML models face interpretability challenges, especially in regulated industries like finance, where transparency in decision-making is paramount. In mortgage lending, stakeholders often demand clear, understandable explanations for automated decisions, and the complex and opaque nature of many ML models poses a significant hurdle in ensuring regulatory compliance (Buchanan and Wright 2021). Despite these challenges, the ability of ML models like LightGBM to integrate complex, dynamic data sources, such as ELTV and borrower-specific factors, provides significant improvements in predicting borrower risk. ML's capacity to adapt to evolving market conditions and borrower behaviour makes it a crucial tool for future mortgage risk assessments, particularly in times of economic uncertainty.

3. Methodology

Section 3 provides a detailed description of the data collection and pre-processing techniques used to prepare the dataset for the three periods under analysis. Then, it introduces the two machine learning models employed to analyse the effectiveness of Credit Score and LTV in predicting ELTV, namely, the Standard LightGBM Model and the Optimized LightGBM

Model followed by the evaluation criteria applied to assess the model's reliability and performance. Lastly, a brief explanation of how the variable importance analysis was conducted is presented.

3.1. Data

3.1.1. Data Source and Description

The primary data for this study were retrieved from the Freddie Mac Single-Family Loan-Level Dataset, which provides comprehensive information on mortgages purchased or guaranteed by Freddie Mac between 1999 and 2024. This dataset includes a full Standard Dataset, covering approximately 53.6 million loans, as well as a Standard Dataset Sample, which provides a manageable subset of 50,000 loans per vintage year.

Given the study's focus on predicting ELTV using Credit Score and LTV, these two variables were identified as central to the research. Additional variables were incorporated into the LightGBM Model to enhance its predictive robustness but were treated as secondary for the purposes of this study. A comprehensive list of all variables included in the dataset, along with their descriptions, is provided in Appendix A.

For this research, to facilitate efficient analysis while retaining sufficient data to ensure reliable results, only the Standard Dataset Sample was considered. It provides a substantial amount of information for each borrower, including loan origination variables and monthly performance data. Given the focus of this research on predicting ELTV using Credit Score and LTV, these three variables were considered core to the study. Additional variables, although integrated into the LightGBM Model to enhance its robustness, were considered as secondary for the current research. A full list of all variables used in the dataset, along with their descriptions, is provided in Appendix A.

3.1.2. Data Segmentation

For further analysis across the different time periods, three specific timeframes were established: the Pre-Pandemic Period, spanning from December 2017 to November 2019; the During-Pandemic Period, covering December 2019 to November 2021; and the Post-Pandemic Period, extending from December 2021 to November 2023. The selection of these periods relied primarily, on the possible economic shifts caused by the COVID-19 pandemic and on the goal

of having a robust and sufficient data spanning two years for each period. Consequently, only the 37,623 loans falling within these timeframes were initially considered for the study.

3.1.3. Data Selection

For this research, to facilitate efficient analysis while retaining sufficient data to ensure reliable results, approximately 10% of the above-mentioned borrowers were randomly selected. This selection was made using Python, eliminating potential biases and ensuring a robust sample of 3,762 borrowers. Since the analysis of borrower behaviour across the different COVID-19 periods requires tracking the same borrowers throughout all three periods, only loans with a first payment as of January 2018 (despite the reporting period beginning in December 2017) were included in the study, excluding all other loans whose initial payment began after January 2018. This filtration process further reduced the initial sample of 3,762 borrowers, leaving only, 1,047 borrowers identified by their Loan Sequence Number.

3.1.4. Data Preprocessing

Before proceeding with model testing, a consolidated final data file was prepared, merging the borrowers' loan origination data with their respective monthly performance records, which had been initially stored in separate datasets. Given the complexity and substantial size of the raw data, several intermediate datasets were created before reaching the final version. All these processes were conducted using Python, ensuring efficient management and accurate integration of data for over 1,000 individual loans. Furthermore, all entries displaying either blank or missing information were removed to prevent interference in later stages of model construction. As a final step, the borrowers' data was categorized into three distinct segments, pre-pandemic, during-pandemic, and post-pandemic, resulting in the creation of three detailed datasets, each prepared to develop a specific model for each period.

3.2. Machine Learning Model

The study employed the LightGBM Model as the primary machine learning algorithm to predict the target variable ELTV (y). Credit Score, LTV and twenty-three additional variables were used as predictors variables of the model (x). LightGBM was chosen due to its considerable efficiency in handling large datasets and its ability to model complex, non-linear interactions

between variables effectively. Its capacity to train quickly and rank variable importance reinforced even more its usage.
Hence, two types of models were developed, a Standard LightGBM Model at a primary stage, followed by an Optimized LightGBM Model. The full Python code used to implement and train both types of models is provided in Appendix C for full transparency and reproducibility.

3.2.1. Standard LightGBM Model

The development of the Standard LightGBM Model was performed using the train-test split method where the dataset was divided into a training set (80%) and a testing set (20%), ensuring that the model was trained and evaluated on separate subsets of data to prevent overfitting. Categorical variables were clearly identified and treated using LightGBM's built-in technique to handle categorical variables. At this stage, the model was built upon the default hyperparameters, without implementing any additional model tuning.
Variables such as Loan Sequence Number and Monthly Reporting Period, which unrelated to ELTV, were excluded from the analysis, while all remaining variables were used as predictors of the target variable.

3.2.2. Optimized LightGBM Model

The construction of the Optimized LightGBM Model followed the same steps as the Standard LightGBM Model. Additionally, to find the best hyperparameters and achieve a high-performance model, a RandomizedSearchCV with cross-validation was performed. The complete list of the hyperparameters used can be found in Appendix B.

3.2.3. LightGBM Models Assessment

The evaluation of the models, namely their performance in predicting ELTV, was conducted using several evaluation metrics such as Mean Squared Error (MSE) and the Coefficient of Determination (R^2). MSE measures the average squared difference between predicted and actual values of ELTV, with lower values indicating better predictive accuracy. On the other hand, R^2 represents the proportion of variance in the target variable explained by the model, ranging from 0 to 1. Higher R^2 values suggest stronger model performance and greater predictive power. Together, these metrics assess both the error magnitude and the strength of

the models.

3.2.4. Variable Importance Analysis

This analysis was conducted to identify the most influential variables in predicting ELTV for each period, with a particular focus on the role of Credit Score and LTV in driving the model's predictions. The procedure was applied to each developed model, ranking the top ten variables based on importance scores generated by the built-in feature importance functionality of LightGBM.

4. Results

Section 4 presents the findings on the performance of the LightGBM Models in predicting ELTV across the three target periods: pre-pandemic, during-pandemic and post-pandemic. The results are structured as follows: first, the Standard and Optimized LightGBM Models were assessed for each period using performance evaluation metrics, MSE and R^2. Then, an analysis of variable importance was conducted. While LightGBM uses the term "feature importance", this study adopts "variable importance" to align with domain-specific terminology.

4.1. Standard and Optimized LightGBM Models

The performance of the Standard and Optimized LightGBM Models in predicting ELTV varied across the three periods: pre-pandemic, pandemic, and post-pandemic. This assessment was done using the MSE and R^2 values. Similarly, the corresponding percentage change (%) for both indicators was also computed. The full table with complete information can be found in Appendix D. Table 1 summarizes the results for each period.

Table 1. Performance of Standard and Optimized LightGBM Models across periods in predicting ELTV.

Period	Model Type	MSE	R^2	% Change in MSE for same period (Optimized vs. Standard)	% Change in MSE Standard (Pandemic vs Pre-Pandemic)	% Change in MSE Standard (Post-Pandemic vs Pandemic)	% Change in MSE Optimized (Pandemic vs Pre-Pandemic)	% Change in MSE Optimized (Post-Pandemic vs Pandemic)	% Change in R^2 Standard (Post-Pandemic vs Pandemic)	% Change in R^2 Optimized (Post-Pandemic vs Pandemic)
Pre-Pandemic	Standard	12465.99	0.872	-						
Pre-Pandemic	Optimized	8626.26	0.911	-30.80						
Pandemic	Standard	1913.97	0.951	-	-84.65	120.00	-86.99	111.61	-5.72	-3.04
Pandemic	Optimized	1122.63	0.971	-41.35						
Post-Pandemic	Standard	4210.80	0.896	-						
Post-Pandemic	Optimized	2375.61	0.942	-43.58						

Both models exhibited similar performance patterns with regard to ELTV prediction across the three analysed periods.

In the pre-pandemic period, the Standard LightGBM Model achieved an MSE of 12,465.99 and an R^2 of 0.872, indicating reasonable predictive accuracy of ELTV under stable economic conditions. The Optimized LightGBM Model registered a better performance, with an MSE of 8,626.26 and R^2 of 0.911, demonstrating a 30.8% lower MSE and a 4.5% higher R^2 compared to the previous model.

During the pandemic, both models demonstrated considerable improvements in predictive performance. This behaviour was explained by the approximately 85% and 87% reduction in MSE, comparing its corresponding pre-pandemic values for the Standard and Optimized Models, respectively.

Similar trends were observed in R^2 values, with the Standard Model achieving 0.951 and the Optimized Model 0.971, emphasizing the models' improved ability to predict ELTV during this period. Lastly, in the post-pandemic phase, the performance of both models declined compared to the peak observed during the previous period. Both models experienced an increase of over 100% in MSE, whereas R^2 values declined by less than 6%, indicating a moderate decrease in predictive accuracy compared to the pandemic period, though still an improvement over the pre-pandemic phase.

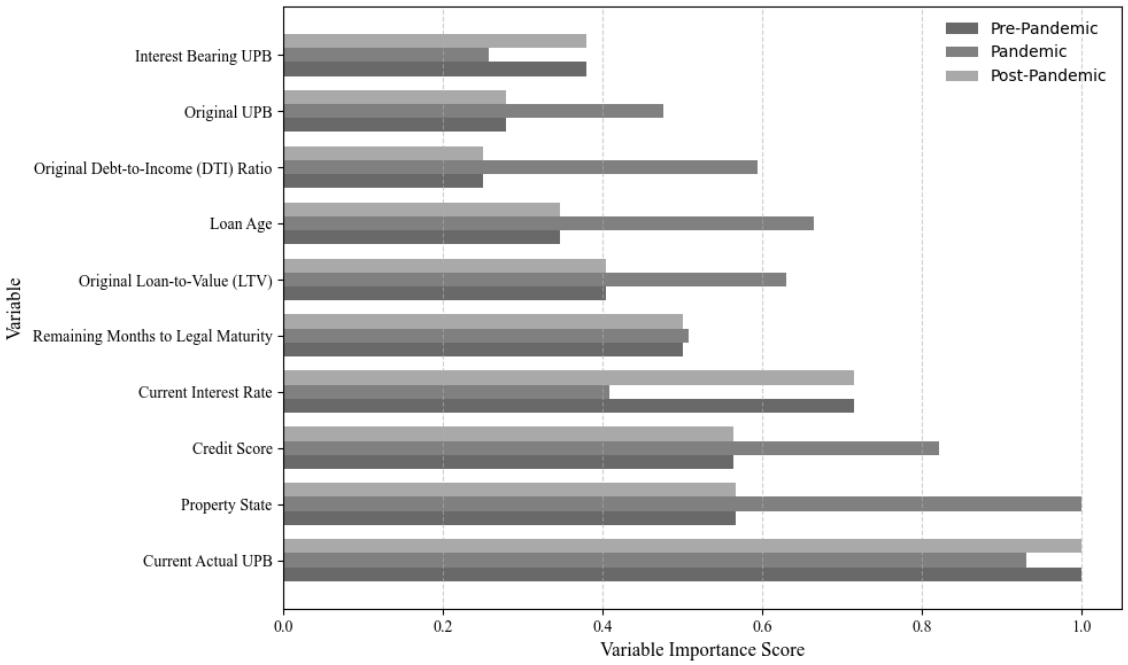
Overall, the relative performance difference between the two types of models became less pronounced during the pandemic rather than the other periods, as suggested by the MSE and R^2 values. The models followed a consistent pattern across the three periods concerning the prediction of ELTV, with their best performance occurring during the pandemic followed by

slight declines in the post-pandemic, though still outperforming pre-pandemic levels. The Optimized Model consistently demonstrated superior predictive accuracy compared to the Standard model across all periods, underscoring the benefits of hyperparameter tuning through GridSearchCV, cross-validation, and advanced optimization techniques. These enhancements enabled the Optimized Model to better adapt to data nuances, avoid overfitting, and improve generalization, while the Standard Model, reliant on default parameters, lacked the refinement to capture complex borrower behaviour patterns.

4.2. Variable Importance Analysis

The relative importance of Credit Score, LTV, and other influential variables in predicting ELTV was analysed for each period and model, using importance scores generated by LightGBM. The findings are presented in plots, showcasing the ten most impactful variables for ELTV prediction.

Figure 1. Variable Importance Analysis for Predicting ELTV Across All Periods Using the Standard LightGBM Model



Note: The scale used is a feature of the LightGBM model and does not have any broader significance in this context.

Concerning the findings of Figure 1, the Standard LightGBM Model identified Current Actual UPB, Property State, and Credit Score as the three most significant predictors of ELTV across all periods. These variables consistently achieved the highest importance scores, securing

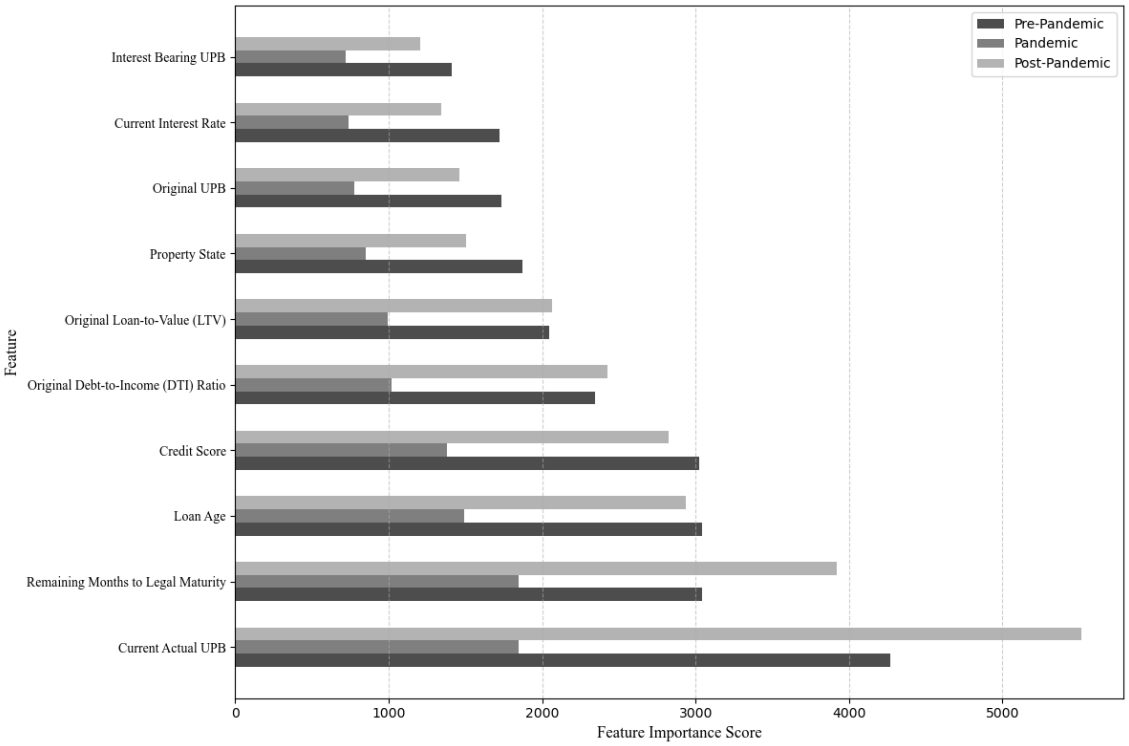
primary rankings throughout the analysis.

During the pandemic period, most variables, except for Current Actual UPB, Current Interest Rate, and Interest Bearing UPB, exhibited higher importance scores compared to the pre- and post-pandemic periods. In contrast, the remaining variables displayed the opposite trend, with lower scores observed during this period.

However, to facilitate the interpretation of the results, the variables were grouped into three categories based on their importance scores: Higher Score Variables, Middle-Lower Score Variables, and Lower Score Variables. The first group included the four most impactful predictors, the second group comprised the intermediate-ranked variables, and the last group contained the three least influential predictors.

In addition to the previously identified Higher Score Variables, the model also classified Original Debt-to-Income (DTI) Ratio, Current UPB, and Interest Bearing UPB as Lower Score Variables. Focusing on the primary target variables of this study, Credit Score and LTV, the findings revealed contrasting predictive power. While Credit Score was identified as one of the most influential predictors, consistently achieving a top ranking, LTV was categorized as a Middle-Lower Score Variable, reflecting its relatively lower ability to predict ELTV compared to the higher-ranked variables.

Figure 2. Variable Importance[3] Analysis for Predicting ELTV Across All Periods Using the Optimized LightGBM Model



Note: The scale used is a feature of the LightGBM model and does not have any broader significance in this context.

By employing the Optimized LightGBM Model, the trends in the results shifted in the opposite direction, with all ten variables displayed in Figure 2 having lower scores during the pandemic and higher scores in other periods. Nevertheless, following the above-mentioned categorization, the findings showed that Current Actual UPB, Remaining Months to Legal Maturity, Loan Age, and Credit Score were classified as Higher Score Variables, reinforcing their importance in predicting ELTV. On the other hand, Original UPB, Current Interest Rate, and Interest Bearing UPB were considered the least impactful variables, as they were classified within the Lower Score Variables category. With regard to Credit Score and LTV, these variables exhibited consistent behaviour, showing high and middle-to-lower predictive power, respectively. Overall, regardless of the model type, Credit Score consistently emerged as a key factor in ELTV prediction, while LTV played a minor yet supportive role in shaping the outcomes across all periods.

[3] The term 'variable importance' is interchangeable with 'feature importance,' a term more commonly used in machine learning. However, 'variable importance' was adopted in this study to align with the broader context of mortgage risk assessment and to prevent any misunderstanding, particularly when referring to predictors of ELTV such as Credit Score, which is conventionally described as a variable rather than a feature.

5. Discussion

This study investigated the predictive power of Credit Score and LTV ratio in assessing borrower risk across three distinct pandemic periods with ELTV serving as a key proxy for borrower risk. The findings revealed that while Credit Score consistently remained a crucial predictor of ELTV, the predictive ability of LTV was notably weaker, with both metrics showing varying effectiveness across the pre-pandemic, during-pandemic, and post-pandemic periods. Similarly, the Optimized LightGBM Model registered a better performance compared to the Standard LightGBM Model with both models demonstrating a significant improvement during the pandemic phase. Section 5 interprets these results in the context of prior literature, explores potential reasons for the observed patterns, and reflects on their implications for mortgage risk assessment. This section also outlines the limitations encountered throughout this work and suggests potential new research to be conducted in the future as well as a summary providing a synthesis of the key findings and their broader implications.

5.1. LightGBM Model Performance and Periodic Variations

The results of the Standard and Optimized LightGBM Models provide valuable insights into how machine learning can enhance the prediction of ELTV.

The models exhibited distinct performance trends across the three pandemic periods, with the best results observed during the pandemic phase. Specifically, the MSE for the Optimized Model decreased by approximately 87%, while that of the Standard Model decreased by approximately 85%, compared to their pre-pandemic levels.

Similarly, the R^2 values increased for both models during the pandemic, with R^2 values of 0.951 and 0.971 for the Standard and the Optimized Model, respectively. This notable enhancement in the model's performance may be partly attributed to shifts in borrower behaviour, such as increased refinancing or changes in repayment patterns. However, such instances were minimal, and the more plausible explanation lies in the implementation of Forbearance Programmes and other government interventions that reduced borrower distress during this phase. By allowing borrowers to pause or reduce payments, these measures temporarily shielded borrowers from default. As a result, borrower risk profiles appeared more stable during this period, which enhanced the models' predictive performance. These extraordinary measures effectively mitigated the financial shocks typically associated with periods of economic stress, aligning borrower behaviour more closely with pre-pandemic trends. This stability could have

reduced the variability and noise in the dataset, making it easier for the machine learning models to capture patterns in borrower behaviour. This finding aligns with other research that has pointed out the impact of forbearance programs during this phase (Freddie Mac 2024).

The LightGBM models, with its ability to account for complex nonlinear relationships between variables, is likely to have adapted better to these sudden changes in borrower behaviour, improving its predictive ability, in line with previous studies highlighting the importance of real-time data in financial risk modelling (Ganong et al. 2020). Likewise, it is also aligned with the idea machine learning methods are particularly versatile at capturing the dynamic nature of financial markets during periods of significant uncertainty (Feng et al. 2021). Hence, the improvements observed in model performance during the pandemic period underscore the importance of leveraging machine learning techniques that can dynamically incorporate changes in borrower behaviour, external economic factors, and government interventions into the risk prediction process since traditional metrics such as Credit Score and LTV may not fully capture the more fluid risk profiles exhibited by borrowers in the face of economic shock (Mansour et al. 2020). On the other hand, the results also demonstrated that the Optimized LightGBM Model outperformed the Standard LightGBM Model. This pattern verified across all periods is consistent with the notion that fine-tuned machine learning models, typically enhances their performance. As the model is adjusted to better capture the underlying data relationships, its ability to make more accurate predictions improves, regardless of borrower behaviour or other external factors (Zhou et al., 2019).

Conversely, the post-pandemic period saw a degradation in predictive accuracy, although the models still outperformed the pre-pandemic period. This decline can be attributed to the reduction of temporary government interventions, such as mortgage forbearance, which had previously stabilized borrower behaviour. Once these measures ended, borrower risk profiles became more variable, and economic factors, like unemployment, re-emerged, making predictions harder. However, despite this decline, the models still outperformed the pre-pandemic period due to the overall economic stabilization that took place, and the more predictable borrower behaviour compared to the uncertainty of the early pandemic phase.

5.2. Credit Score and LTV as Predictors of Borrower Risk

Discussing the central question of this work, Credit Score and LTV were identified as predictors of ELTV across all three periods under analysis. Nevertheless, the results of the LightGBM

modelling in this study revealed significant differences in the predicting power of Credit Score and LTV in the prediction process, with Credit Score consistently outperforming LTV, alongside other variables such as Current Actual UPB, whose predictive performance was still higher than that of Credit Score. These differences underscore the distinct roles these two predictors play in assessing long-term borrower risk, measured through ELTV in this context.

At the time of loan origination, both mortgage loan metrics are static measures, representing the initial financial profile of the borrower and the loan. In contrast, ELTV is a dynamic metric that evolves over time, estimated as the ratio of the actual outstanding loan balance to the estimated current property value. Thus, this disparity between static predictors and the dynamic nature of ELTV may explain, in part, the predictive performance of these metrics, particularly Credit Score. Additionally, the improved performance of the models when combining static predictors, such as Credit Score and LTV, with dynamic metrics like ELTV as targets, highlights the advantage of incorporating both types of variables. This combination allows the model to better capture the evolving borrower circumstances over time, thereby improving its ability to assess true borrower risk.

Regarding the observed differences in the predictive performance of both, while LTV remains fixed once the loan is originated, it does not adjust for subsequent changes in the borrower's financial behaviour or broader economic conditions, such as fluctuations in property values, interest rates, or unemployment. Across the analysed periods, these economic conditions experienced significant shifts, with property values and interest rates fluctuating, and unemployment rates rising sharply. These changes may have had a considerable impact on LTV's predictive ability, as the study does not account for these post-origination loan dynamics.

On the other hand, although Credit Score is also static at loan origination, it is generally regarded as a more holistic measure of borrower risk due to its broad incorporation of various financial behaviours, including payment history, debt management and credit usage. However, in this study, both Credit Score and LTV were treated as static numbers, each representing a single value at loan origination. This procedure limited the potential predictive power of Credit Score, as its underlying components, such as payment history and credit usage, which were not included in the model.

Despite this limitation, Credit Score still outperformed LTV in predicting ELTV, likely due to its ability to serve as a better proxy for long-term borrower behaviour. Even as a single static number, Credit Score provides a broader and more comprehensive financial picture of the borrower, offering greater predictive power than LTV. Nevertheless, Current Actual UPB

emerged as the highest-ranked predictor for ELTV in both, the Standard and the Optimized LightGBM Models, even surpassing Credit Score in predictive ability. Current Actual UPB reflects the mortgage's ending balance as reported by the servicer, including any scheduled or unscheduled principal reductions. While Credit Score captures the borrower's overall financial reliability, Current Actual UPB provides crucial insights into the ongoing changes in the mortgage balance over time. Together, these metrics can offer a more complete picture of borrower risk, significantly enhancing the predictive accuracy of machine learning models.

LTV, on the other hand, can be seen as a simple snapshot of the loan-to-property value at a specific point in time and does not adjust for changes in the borrower's financial situation or other dynamic factors. This narrow focus makes LTV less informative for predicting ELTV, especially when compared to Credit Score, which offers a broader view of the borrower's financial health and capacity to manage long-term financial risks. Additionally, another possible reason for Credit Score's outperformance in comparison to LTV lies in its combination with the other dynamic variables when running the models, which may provide a more robust and comprehensive measure of long-term borrower risk through ELTV prediction. The static nature of LTV, in contrast, limits its ability to provide meaningful insights over time, particularly in the face of changing economic conditions or borrower behaviour.

Moreover, the inclusion of twenty-three additional dynamic variables, particularly those related to monthly loan performance and borrower behaviour, further enhances the predictive accuracy of the model. Variables such as Current Actual UPB, Loan Age, and Remaining Months to Legal Maturity, which showed a predictive power even higher than that of Credit Score, capture more effectively how the borrower's financial behaviour evolves over time. These dynamic metrics reflect the borrower's repayment patterns, loan balance changes, and broader economic conditions, providing perhaps a more accurate prediction of ELTV.

These findings align with prior research, where Credit Score has been shown to be a robust indicator of borrower reliability in both stable and volatile economic conditions (Agarwal et al. 2015; Altman & Saunders 1998). The consistency of Credit Score's importance across all periods emphasizes its consistent relevance in mortgage risk assessment, even in disrupted economic scenarios like the COVID-19 pandemic.

Given that, as far as we know, no prior study has specifically addressed the combined predictive power of dynamic variables like Current Actual UPB alongside traditional metrics such as Credit Score and LTV in forecasting ELTV, this research contributes a unique perspective to the field of mortgage risk assessment.

5.3. Conceptual Alignment of Credit Score and LTV with ELTV

The use of Credit Score and LTV to predict ELTV is grounded in their conceptual alignment with key components of borrower risk. Credit Score captures borrower-level risks, particularly relating to Probability of Default (PD). It reflects factors such as financial stability, debt management, and employment history, offering valuable insight into a borrower's likelihood of default. LTV, on the other hand, is associated with Loss Given Default (LGD). It measures the potential loss severity in the event of default by assessing the loan amount relative to the value of the underlying property at origination.

ELTV serves as a dynamic metric that integrates both PD and LGD, bridging the two components by accounting for changes in borrower equity and property values over time. This dynamic nature makes ELTV particularly relevant in fluctuating economic conditions, such as those seen during the COVID-19 pandemic, when both borrower financial health and property values experienced rapid changes.

During the pandemic, Credit Score's strong predictive power can be attributed to its sensitivity to borrower-level shocks, such as unemployment and financial distress. These factors were significant drivers of PD, making Credit Score a critical predictor of borrower risk in this period. However, it is important to note that widespread forbearance programs may have masked actual defaults, requiring caution in interpreting the results. On the other hand, LTV underperformed due to its static nature, which limited its capacity to account for rapid fluctuations in property values, key elements tied to LGD. These limitations suggest that during the pandemic, shocks to PD (such as unemployment) were more influential than those affecting LGD (such as housing price fluctuations).

The combination of Credit Score and LTV with ELTV in assessing borrower risk provides a more comprehensive and dynamic approach to evaluating long-term borrower performance. By integrating both static and dynamic metrics, such as Credit Score, LTV, and ELTV, this approach offers a more nuanced understanding of borrower risk, especially under conditions of economic stress and market volatility.

5.4. Implications for Mortgage Risk Management

The findings of this study have important implications for mortgage lenders, policymakers, and other stakeholders, particularly in relation to the use of traditional mortgage risk metrics and the incorporation of advanced machine learning models. The consistent importance of Credit

Score as a predictor of ELTV across all periods supports its continued use as a key metric in mortgage lending. However, the middle-lower predictive ability of LTV suggests that analysts should consider supplementing traditional models with additional borrower-specific factors and other key economic indicators. This could improve the relative importance of LTV and lead to the development of models that better capture the evolving nature of borrower risk.

Furthermore, the success of LightGBM modelling in predicting ELTV during the pandemic highlights the potential of machine learning to enhance risk management practices in the mortgage industry. As demonstrated by the improved performance during the pandemic of both the Standard and Optimized LightGBM Models, machine learning algorithms can identify complex, non-linear relationships between borrower characteristics, loan features, and macroeconomic variables, though these were not considered in the models. This ability introduces potential applications, such as real-time dynamic loan modifications, where lenders could proactively adjust loan terms in response to changes in borrower equity and financial distress.

Such applications can help lenders more accurately assess risk, particularly in periods of high uncertainty when the sole use of traditional metrics may fall short. Moreover, the findings underscore the importance of incorporating dynamic risk metrics, such as ELTV, into mortgage risk management strategies. While ELTV serves as a proxy for borrower risk rather than a direct measure, it provides a more nuanced view of borrower equity over time. This enables lenders to monitor changes in equity throughout the life of the loan, offering valuable insights into evolving borrower circumstances. Additionally, it facilitates an early evaluation of potential risk factors at the time of loan origination. By leveraging ELTV as part of broader risk assessments, lenders can refine their strategies and support the development of policies that enhance financial stability and sustainable lending practices. The integration of such dynamic metrics, like ELTV, which serves as a proxy for borrower risk over time, when combined with machine learning models, has the potential to establish a more adaptive and resilient mortgage risk management framework.

5.5. Limitations and Future Research Directions

Although the results of this study provide valuable insights into the predictive power of traditional risk metrics when using machine learning models, several limitations should be addressed in future research. First, this study focused on U.S. mortgage data from 2017 to 2023,

and the findings may not necessarily generalize to other countries or mortgage markets. To draw more robust conclusions, larger datasets with a broader timeframe would be required. Therefore, future research could explore the applicability of the LightGBM model and the ELTV metric in other regions, particularly those that experienced different economic interventions during the pandemic, by using a larger sample size, as this study only analysed a sample of 1,047 loans.

Second, while this study examined the predictive ability of Credit Score and LTV using only LightGBM, due to its computational efficiency and practical adaptability given the resource and time constraints, other advanced machine learning models, such as Neural Networks or Gradient Boosting Machines (GBM) like XGBoost or CatBoost, could be explored for comparison. These models may offer different advantages in handling more complex, large-scale data and capturing non-linear relationships. Such efforts could help further validate and refine the insights provided in this research.

The selection of variables used to predict ELTV was also critical, as many attempts were required to achieve a high-performance model. In addition, the study did not investigate the impact of more detailed borrower-level data, such as income or employment stability, on the predictive accuracy of the models. Similarly, it did not incorporate important macroeconomic factors, such as interest rates, inflation, housing price changes, or unemployment rates, among others. Future research could incorporate these variables into the LightGBM model or other machine learning models. This would enable a more detailed understanding of the factors driving borrower behaviour and provide more accurate risk assessments, though it is important to consider that ELTV serves as a proxy for borrower risk.

Furthermore, more studies could also explore directions such as integrating behavioural credit scoring and post-crisis risk calibration, which would require further model development. Behavioural credit scoring could involve analysing a borrower's ongoing financial behaviour and trends to assess future risk, offering a dynamic alternative to traditional credit scores. However, this analysis must be carefully conducted to avoid infringing on personal data privacy without consent. On the other hand, further research on post-crisis risk calibration would help adjust models to better predict borrower behaviour and risks following significant economic disruptions, such as the pandemic, improving their adaptability to changing economic conditions. In addition, unlike this study, which relied on static values of Credit Score and LTV, future research could incorporate their underlying components such as repayment history, credit utilization, and property appraisal details into predictive models to improve their performance

and predictive accuracy.

Lastly, although the data used in this study **was** retrieved from Freddie Mac and instances of default were relatively infrequent, data imbalance can still present challenges to the performance of machine learning models. While the impact may have been minimal in this study, future studies could address class imbalance using techniques like SMOTE or class weighting to further enhance predictive accuracy (He and Ma 2013; Chawla et al. 2002).

5.6. Summary

In conclusion, this study demonstrates the varying predictive power of traditional mortgage metrics, such as Credit Score and LTV, while highlighting the superior performance of the Optimized LightGBM Model, especially during the pandemic period. The findings underscore the importance of incorporating dynamic metrics like ELTV, which serves as a proxy for borrower risk, and leveraging machine learning techniques to better adapt to changes in borrower behaviour during times of economic uncertainty. While Credit Score remains a reliable predictor, this research advocates for supplementing it with additional borrower-specific and macroeconomic factors to enhance mortgage risk assessment and improve financial stability. Future studies should focus on broader datasets, advanced models, and the integration of behavioural credit scoring to refine risk predictions and adjust to post-crisis environments.

6. Conclusion

This study analyses the effectiveness of traditional mortgage metrics, namely Credit Score and LTV, in predicting borrower risk, using ELTV as a proxy. By employing LightGBM machine learning models to analyse borrower risk during pre-pandemic, pandemic, and post-pandemic periods, this research provides critical insights into the limitations of static metrics and underscores the necessity of integrating dynamic measures like ELTV to improve risk assessment approaches. The findings demonstrate that Credit Score, while static, remains a robust predictor of borrower risk, consistently outperforming LTV across all periods. Its predictive power lies in its widespread use as a core mortgage risk metric, reflecting a borrower's overall creditworthiness. Even when treated as a single static value in this study, Credit Score can be categorized as a Higher Score Variable given its superior performance alongside other variables. Additionally, the results highlight the lower accuracy of LTV in predicting ELTV across all periods. While it was also treated as a static value like Credit Score,

its consistently lower ranking placed it in the group of Middle-Lower Score Variables, emphasizing its limitations in the prediction process, likely due to its inability to account for post-origination changes.

Another key finding is the decline in the models' performance in the post-pandemic period. This degradation in predictive accuracy can be attributed to the reduction of temporary government interventions. As these measures ended, borrower risk profiles became more variable, and economic factors like unemployment re-emerged, challenging the models to maintain the same level of accuracy. However, the models still outperformed the pre-pandemic period due to overall economic stabilization and more predictable borrower behaviour.

Despite its contributions, this study has several limitations. The analysis was limited to U.S. mortgage data, with a sample size of 1,047 loans, and relied solely on LightGBM as the machine learning model. Future research should explore additional machine learning frameworks and incorporate a broader set of borrower-specific variables and macroeconomic factors to enhance predictive accuracy. Similarly, incorporating the underlying components of Credit Score and LTV, such as repayment history, credit utilization, and property appraisal details, could also improve the models. Furthermore, expanding the geographical scope and increasing the sample size would ensure greater generalizability of the findings.

The integration of dynamic metrics like ELTV, coupled with machine learning techniques, offers a powerful framework for advancing mortgage risk management. By bridging traditional static metrics with dynamic, evolving measures, this study contributes to the development of a more adaptive and resilient mortgage lending framework. This dual perspective provides actionable insights for lenders, policymakers and other professionals in the field to enhance early risk assessment, adapt strategies during periods of economic uncertainty, and support sustainable lending practices that promote long-term financial stability.

7. References

Agarwal, Sumit, Brent W. Ambrose, and Yildiray Yildirim. 2015. "The Subprime Virus." *Real Estate Economics* 43: 891–915.

Altman, Edward I., and Anthony Saunders. 1998. "Credit Risk Measurement: Developments Over the Last 20 Years." *Journal of Banking & Finance* 21: 1721–1742.

Avery, Robert B., Raphael W. Bostic, Paul S. Calem, and Glenn B. Canner. 2003. "An Overview of Consumer Data and Credit Reporting." *Federal Reserve Bulletin* 89 (2): 47–73.

Avery, Robert B., Kenneth P. Brevoort, and Glenn B. Canner. 2004. "Credit Scoring and Its Effects on the Availability and Affordability of Credit." *Journal of Consumer Affairs* 38 (3): 322–350.

Bazilinska, Olga. 2020. "Mortgage Lending in Europe: Challenges and Opportunities." *Journal of European Financial Studies* 15 (2): 101–118.

Federal Reserve. 2022. *Economic Well-Being of U.S. Households in 2021*. Washington, DC: Board of Governors of the Federal Reserve System.

Ganong, Peter, and Pascal J. Noel. 2020. "Why Do Borrowers Default on Mortgages?" NBER Working Paper No. 27585. National Bureau of Economic Research.

Hari, T. G., M. Smith, and A. Brown. 2020. "The Impact of COVID-19 on Mortgage Markets: Evidence from the United States." *Journal of Housing Economics* 48: 101–122.

Harrison, David M., David C. Ling, and Marcus T. Millan. 2004. "How Important Are Foreclosure Costs in the Default Decision?" *Journal of Real Estate Research* 26 (2): 153–188.

Mansour, Heidi, Christopher Meyer, and Simon Weiner. 2020. "Mortgage Forbearance and Housing Stability: A Policy Response to COVID-19." *Housing Policy Debate 31* (1): 16–35.

Ospina, Juan, and Harald Uhlig. 2018. "Mortgage-Backed Securities and Financial Crises." *American Economic Review 108* (2): 283–307

Tamayo, Jorge Galindo and Pablo. 2000. “Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications.” *Computational Economics 15* 107-143.

Duca, John, and John Muellbauer. "Subprime Mortgage Crisis." *Federal Reserve History*, 2013.

Accessed December 12, 2024.

Makri, Polyxeni, Antonios N. Tsagkanos, and Athanasios Bellas. "Determinants of Non-Performing Loans: The Case of Eurozone." *Panoeconomicus 61* 193-206.

Bian, Xun, Zhenguo Lin, and Yingchun Liu. 2018. "House Price, Loan-to-Value Ratio and Credit Risk." *Journal of Banking & Finance 92* 1-12.

Buchanan, N., and R. Wright. 2021. "Financial Inclusion and Credit Scoring: The Exclusion of Non-Traditional Financial Behavior." J*ournal of Financial Inclusion* 22 (4): 203-217.

Chen, Q., Y. Li, and S. Liu. 2021. "Machine Learning for Mortgage Default Prediction: Comparing Traditional Metrics with Dynamic Models." *Journal of Mortgage and Real Estate Finance 45* (3): 152-167.

Consumer Financial Protection Bureau. 2024. "Credit Scores: How Lenders Use Them." CFPB. https://www.consumerfinance.gov.

Demyanyk, Y., and O. Van Hemert. 2011. "Understanding the Dangers of Loan-to-Value Ratios and the Real Estate Bubble." *Financial Stability Review 11* (1): 44-56.

Freddie Mac. 2024. "Understanding Loan-to-Value (LTV) and Its Importance in Mortgage Risk." *Freddie Mac Lending Resources.* https://www.freddiemac.com.

Farkas, E., R. Zhang, and J. Wilson. 2020. "The Impact of Economic Shocks on Borrower Default: Lessons from the COVID-19 Pandemic." *Journal of Financial Stability 12* (1): 42-60.

Goel, V., P. Kumar, and R. Jain. 2021. "Discriminating Credit Scores and the Exclusion of Non-Traditional Financial Behaviors." *Economic Policy Review 23* (2): 89-104.

Luis, L. 2015. "Evaluating the Effectiveness of the LTV Ratio in Predicting Mortgage Default Risks." *Journal of Risk and Financial Management* 8 (1): 70-80.

Parlov, A., and R. Wachter. 2011. "Subprime Lending and the Housing Bubble: The Role of Loan-to-Value Ratios." *Journal of Financial Economics* 85 (2): 185-204.

Seiler, M. 2017. "The Subprime Mortgage Crisis: Analyzing the Impact of High-LTV Loans on the Housing Market." *Real Estate Economics 45* (4): 1223-1240.

Liu, Yiting, Lennart John Baals, Jörg Osterrieder, and Branka Hadji-Misheva. 2024. "Network Centrality and Credit Risk: A Comprehensive Analysis of Peer-to-Peer Lending Dynamics." *Finance Research Letters* 63: 105308.

Ganong, Peter, and Pascal J. Noel. 2020. "Why Do Borrowers Default on Mortgages?" *National Bureau of Economic Research* Working Paper No. 27585.

Bierut, Beata, Tomasz Chmielewski, Adam Głogowski, Andrzej Stopczyński, and Sławomir Zajączkowski. 2015. "Implementing Loan-To-Value and Debt-To-Income Ratios: Learning from Country Experiences. The Case of Poland." *NBP Working Paper No. 212*. Narodowy Bank Polski.

Avery, Robert B., Kenneth P. Brevoort, Glenn B. Canner, Cheryl R. Cooper, Christa N. Gibbs, and Rebecca Tsang. 2008. "The 2007 HMDA Data." *Federal Reserve Bulletin*, December 23.

Aziz, Saqib, and Michael Dowling. 2019. "Machine Learning and AI for Risk Management." In *Disrupting Finance: FinTech and Strategy in the 21st Century*, edited by Theo Lynn, John G. Mooney, Pierangelo Rosati, and Mark Cummins, 33–50. Cham: Springer.

Chang, Victor, Sharuga Sivakulasingam, Hai Wang, Siu Tung Wong, Meghana Ashok Ganatra, and Jiabin Luo. 2024. "Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers." *Risks* 12: 174.

Liu, Yiting, Lennart John Baals, Jörg Osterrieder, and Branka Hadji-Misheva. 2024. "Network Centrality and Credit Risk: A Comprehensive Analysis of Peer-to-Peer Lending Dynamics." *Finance Research Letters* 63: 105308.

Arya, Shweta, Catherine Eckel, and Colin Wichman. 2013. "Anatomy of the Credit Score." *Journal of Economic Behavior & Organization* 95: 175–185.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. New York: ACM.

Chen, Weiyang, David Sondak, Randall Davis, and Andrew Lo. 2021. "Machine Learning for Mortgage Risk Prediction: An Application of Gradient Boosting Machines." J*ournal of Financial Data Science 3* (1): 15–28.

Hasan, Saif, and Shahid Mahmood. 2024. "The Role of AI in Real Estate Forecasting: A Case Study of LightGBM Applications." *Journal of Housing Market Dynamics 12* (1): 45–67.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 4765–4774. Long Beach, CA.

Buchanan, Ben, and Andrew Wright. 2021. The AI Governance Challenge: Regulation in an Algorithmic Age. Cambridge, MA: MIT Press.

Foster, Chester, and Robert Van Order. 1984. "An Option-Based Model of Mortgage Default." *Housing Finance Review* 3 (4): 351–372.

Riddiough, Timothy J. 1991. "Equilibrium Mortgage Default Pricing with Non-Perfect Capital Markets." *Real Estate Economics* 19 (3): 450–472.

Goldberg, Lawrence G., and Charles A. Capone. 1998. "A Dynamic Model of Mortgage Default." *Journal of Housing Economics* 7 (4): 267–288.

Zhou, Guang, Zhiwei Hu, and Jing Zhang. 2019. ″Network Analysis of Loan Default Risk: Insights from a Complex System Perspective.″ *Journal of Financial Stability 40:* 120–134.

Hadžić, Faruk. 2016. ″The Impact of Loan-to-Value Ratios on Mortgage Default Risk.″ Journal of Financial Risk Management 9 (3): 45–59.

He, Haibo, and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications.* Hoboken, NJ: Wiley-IEEE Press, 2013.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research 16* (2002): 321–357.

8. Appendices

Appendix A: Overview of Single-Family Loan-Level Dataset – Variables Characterizing Each Loan and Borrower

Table 2. Origination Data File

COLUMN POSITION[6]	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
1	**CREDIT SCORE** - A number, prepared by third parties, summarizing the borrower's creditworthiness, which may be indicative of the likelihood that the borrower will timely repay future obligations. Generally, the credit score disclosed is the score known at the time of acquisition and is the score used to originate the mortgage.	• 300 - 850 • 9999 = Not Available Credit Scores < 300 or > 850 will be shown as Not Available.	Numeric	4
2	**FIRST PAYMENT DATE** - The date of the first scheduled mortgage payment due under the terms of the mortgage note.	• YYYYMM	Date	6
3	**FIRST TIME HOMEBUYER FLAG** - Indicates whether the Borrower, or one of a group of Borrowers, is an individual who (1) is purchasing the mortgaged property, (2) will reside in the mortgaged property as a primary residence, and (3) had no ownership interest (sole or joint) in a residential property during the three-year period preceding the date of the purchase of the mortgaged property. With certain limited exceptions, a displaced homemaker or single parent may also be considered a First-Time Homebuyer if the individual had no ownership interest in a residential property during the preceding three-year period other than an ownership interest in the marital residence with a spouse.	• Y = Yes • N = No • 9 = Not Available or Not Applicable	Alpha	1
4	**MATURITY DATE** - The month in which the final monthly payment on the mortgage is scheduled to be made as stated on the original mortgage note.	• YYYYMM	Date	6
5	**METROPOLITAN STATISTICAL AREA (MSA) OR METROPOLITAN DIVISION** - This disclosure will be based on the designation of the Metropolitan Statistical Area or Metropolitan Division as of the date of issuance. Metropolitan Statistical Areas (MSAs) are defined by the United States Office of Management and Budget (OMB) and have at least one urbanized area with a population of 50,000 or more inhabitants. An MSA containing a single core with a population of 2.5 million or more may be divided into smaller groups of counties that OMB refers to as Metropolitan Divisions. If an MSA applies to a mortgaged property, the applicable five-digit value is disclosed; however, if the mortgaged property also falls within a Metropolitan Division classification, the applicable five-digit value for the Metropolitan Division takes precedence and is disclosed instead. This	• Metropolitan Division or MSA Code. • Null indicates that the area in which the mortgaged property is located is a) neither an MSA nor a Metropolitan Division, or b) unknown.	Numeric	5

COLUMN POSITION[6]	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
	disclosure will not be updated to reflect any subsequent changes in designations of MSAs, Metropolitan Divisions or other classifications.			
6	**MORTGAGE INSURANCE PERCENTAGE (MI %)** - The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan. Only primary mortgage insurance that is purchased by the Borrower, lender or Freddie Mac is disclosed. Mortgage insurance that constitutes "credit enhancement" that is not required by Freddie Mac's Charter is not disclosed. Amounts of mortgage insurance reported by Sellers that are less than 1% or greater than 55% will be disclosed as "Not Available," which will be indicated 999. No MI will be indicated by zero.	• 1% - 55% • 0 = No MI • 999 = Not Available	Numeric	3
7	**NUMBER OF UNITS** - Denotes whether the mortgage is a one-, two-, three-, or four-unit property.	• 1 = one-unit • 2 = two-unit • 3 = three-unit • 4 = four-unit • 99 = Not Available	Numeric	2
8	**OCCUPANCY STATUS** - Denotes whether the mortgage type is owner occupied, second home, or investment property.	• P = Primary Residence • I = Investment Property • S = Second Home • 9 = Not Available	Alpha	1
9	**ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)** – In the case of a purchase mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the lesser of the mortgaged property's appraised value on the note date or its purchase price. In the case of a refinance mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the mortgaged property's appraised value on the note date. If the secondary financing amount disclosed by the Seller includes a home equity line of credit, then the CLTV calculation reflects the disbursed amount at closing of the first lien mortgage loan, not the maximum loan amount available under the home equity line of credit. In the case of a seasoned mortgage loan, if the Seller cannot warrant that the value of the mortgaged property has not declined since the note date, Freddie Mac requires that the Seller must provide a new appraisal value, which is used in the CLTV calculation. In certain cases, where the Seller delivered a loan to Freddie Mac with a special code indicating additional secondary mortgage loan amounts, those amounts may have been included in the CLTV calculation. If the CLTV is < LTV, set the CLTV to 'Not Available.' This disclosure is subject to the widely varying standards originators use to verify Borrowers' secondary mortgage loan amounts and will not be updated.	2018Q1 and prior: • 6% - 200% • 999 = Not Available 2018Q2 and later: • 1% - 998% • 999 = Not Available HARP ranges: • 1% - 998% • 999 = Not Available	Numeric	3
10	**ORIGINAL DEBT-TO-INCOME (DTI) RATIO** - Disclosure of the debt to income ratio is based on (1) the sum of the borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the borrower is making at the time of the delivery of the mortgage loan to Freddie Mac, divided by (2) the total monthly income used to underwrite the loan as of the date of the origination of the such loan. Ratios greater than 65% are indicated that data is Not Available. All loans in the HARP dataset will be disclosed as Not Available. This disclosure is subject to the widely varying standards originators use to	• 0%<DTI<=65% • 999 = Not Available HARP ranges: • 999 = Not Available	Numeric	3

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
11	**ORIGINAL UPB** - The UPB of the mortgage on the note date.	• Amount will be rounded to the nearest $1,000.	Numeric	12
12	**ORIGINAL LOAN-TO-VALUE (LTV)** - In the case of a purchase mortgage loan, the ratio obtained by dividing the original mortgage loan amount on the note date by the lesser of the mortgaged property's appraised value on the note date or its purchase price. In the case of a refinance mortgage loan, the ratio obtained by dividing the original mortgage loan amount on the note date and the mortgaged property's appraised value on the note date. In the case of a seasoned mortgage loan, if the Seller cannot warrant that the value of the mortgaged property has not declined since the note date, Freddie Mac requires that the Seller must provide a new appraisal value, which is used in the LTV calculation. For loans in the non HARP dataset, ratios below 6% or greater than 105% will be disclosed as "Not Available," indicated by 999. For loans in the HARP dataset, LTV ratios greater than 999% will be disclosed as Not Available.	2018Q1 and prior: • 6% - 105% • 999 = Not Available 2018Q2 and later: • 1% - 998% • 999 = Not Available HARP ranges: • 1% - 998% • 999 = Not Available	Numeric	3
13	**ORIGINAL INTEREST RATE** - The interest rate of the loan as stated on the note at the time the loan was originated.		Numeric Literal decimal	6
14	**CHANNEL** - Disclosure indicates whether a Broker or Correspondent, as those terms are defined below, originated or was involved in the origination of the mortgage loan. If a Third Party Origination is applicable, but the Seller does not specify Broker or Correspondent, the disclosure will indicate "TPO Not Specified". Similarly, if neither Third Party Origination nor Retail designations are available, the disclosure will indicate "TPO Not Specified." If a Broker, Correspondent or Third Party Origination disclosure is not applicable, the mortgage loan will be designated as Retail, as defined below. Broker is a person or entity that specializes in loan originations, receiving a commission (from a Correspondent or other lender) to match Borrowers and lenders. The Broker performs some or most of the loan processing functions, such as taking loan applications, or ordering credit reports, appraisals and title reports. Typically, the Broker does not underwrite or service the mortgage loan and generally does not use its own funds for closing; however, if the Broker funded a mortgage loan on a lender's behalf, such a mortgage loan is considered a "Broker" third party origination mortgage loan. The mortgage loan is generally closed in the name of the lender who commissioned the Broker's services. Correspondent is an entity that typically sells the Mortgages it originates to other lenders, which are not Affiliates of that entity, under a specific commitment or as part of an ongoing relationship. The Correspondent performs some, or all, of the loan processing functions, such as: taking the loan application; ordering credit reports, appraisals, and title reports; and verifying the Borrower's income and employment. The Correspondent may or may not have delegated underwriting and typically funds the mortgage loans at settlement. The mortgage loan is closed in the Correspondent's name and the Correspondent may or may not service the mortgage loan. The Correspondent may use a Broker to perform some of the processing functions or even to fund the loan on its behalf; under such circumstances, the mortgage loan is considered a "Broker" third party origination mortgage loan, rather than a "Correspondent" third party origination mortgage loan. Retail Mortgage is a mortgage loan that is originated, underwritten and funded by a lender or its Affiliates. The mortgage loan is closed in the name of the lender or its Affiliate and if it is sold to Freddie Mac, it is sold by the lender or its Affiliate that originated it. A mortgage loan that a Broker or Correspondent completely or partially originated, processed, underwrote, packaged, funded or closed is not considered a Retail mortgage loan.	• R = Retail • B = Broker • C = Correspondent • T = TPO Not Specified • 9 = Not Available	Alpha	1

COLUMN POSITION*	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
	For purposes of the definitions of Correspondent and Retail, "Affiliate" means any entity that is related to another party as a consequence of the entity, directly or indirectly, controlling the other party, being controlled by the other party, or being under common control with the other party.			
15	**PREPAYMENT PENALTY MORTGAGE (PPM) FLAG** - Denotes whether the mortgage is a PPM. A PPM is a mortgage with respect to which the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal.	• Y = PPM • N = Not PPM	Alpha	1
16	**AMORTIZATION TYPE** - Denotes that the product is a fixed-rate mortgage or adjustable-rate mortgage.	• FRM – Fixed Rate Mortgage • ARM – Adjustable Rate Mortgage	Alpha	5
17	**PROPERTY STATE** - A two-letter abbreviation indicating the state or territory within which the property securing the mortgage is located.	• AL, TX, VA, etc.	Alpha	2
18	**PROPERTY TYPE** - Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single-Family home. If the Property Type is Not Available, this will be indicated by 99.	• CO = Condo • PU = PUD • MH = Manufactured Housing • SF = Single-Family • CP = Co-op • 99 = Not Available	Alpha	2
19	**POSTAL CODE** – The postal code for the location of the mortgaged property	• ###00, where "###" represents the first three digits of the 5-digit postal code • 00 = Unknown	Numeric	5
20	**LOAN SEQUENCE NUMBER** - Unique identifier assigned to each loan.	PYYQnXXXXXXX • Product F = FRM and A = ARM; • YYQn = origination year and quarter; and, • XXXXXXX = randomly assigned digits	Alpha-numeric	12
21	**LOAN PURPOSE** - Indicates whether the mortgage loan is a Cash-out Refinance mortgage, No Cash-out Refinance mortgage, or a Purchase mortgage. Generally, a Cash-out Refinance mortgage loan is a mortgage loan in which the use of the loan amount is not limited to specific purposes. A mortgage loan placed on a property previously owned free and clear by the Borrower is always considered a Cash-out Refinance mortgage loan. Generally, a No Cash-out Refinance mortgage loan is a mortgage loan in which the loan amount is limited to the following uses: Pay off the first mortgage, regardless of its age Pay off any junior liens secured by the mortgaged property, that were used in their entirety to acquire the subject property Pay related closing costs, financing costs and prepaid items, and Disburse cash out to the Borrower (or any other payee) not to exceed 2% of the new refinance mortgage loan or $2,000, whichever is less. As an exception to the above, for construction conversion mortgage loans and renovation mortgage loans, the amount of the interim construction financing secured by the mortgaged property is considered an amount [illegible] construction costs paid by the Borrower outside of the secured interim construction financing is considered cash out to the Borrower, if greater than $2000 or 2% of loan amount. This disclosure is subject to various special exceptions used by Sellers to determine whether a mortgage loan is a No Cash-out Refinance mortgage loan.	• P = Purchase • C = Refinance - Cash Out • N = Refinance - No Cash Out • R = Refinance - Not Specified • 9 =Not Available	Alpha	1

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
22	**ORIGINAL LOAN TERM** - A calculation of the number of scheduled monthly payments of the mortgage based on the First Payment Date and Maturity Date.	• Calculation: (Loan Maturity Date (MM/YY) – Loan First Payment Date (MM/YY) + 1)	Numeric	3
23	**NUMBER OF BORROWERS** - The number of Borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property. Disclosure denotes only whether there is one borrower, or more than one borrower associated with the mortgage note. This disclosure will not be updated to reflect any subsequent assumption of the mortgage note.	2018Q1 and prior: • 01 = 1 borrower • 02 = > 1 borrowers • 99 = Not Available 2018Q2 and later: • 01 = 1 borrower • 02 = 2 borrowers • 03 = 3 borrowers … • 09 = 9 borrowers • 10 = 10 borrowers • 99 = Not Available	Numeric	2
24	**SELLER NAME** - The entity acting in its capacity as a seller of mortgages to Freddie Mac at the time of acquisition. Seller Name will be disclosed for sellers with a total Original UPB representing 1% or more of the total Original UPB of all loans in the Dataset for a given calendar quarter. Otherwise, the Seller Name will be set to "Other Sellers".	Name of the seller, or "Other Sellers"	Alpha-numeric	60
25	**SERVICER NAME** - The entity acting in its capacity as the servicer of mortgages to Freddie Mac as of the last period for which loan activity is reported in the Dataset. Servicer Name will be disclosed for servicers with a total Original UPB representing 1% or more of the total Original UPB of all loans in the Dataset for a given calendar quarter. Otherwise, the Servicer Name will be set to "Other Servicers".	Name of the servicer, or "Other Servicers"	Alpha-numeric	60
26	**SUPER CONFORMING FLAG** – For mortgages that exceed conforming loan limits with origination dates on or after 10/1/2008 and were delivered to Freddie Mac on or after 1/1/2009	• Y = Yes • Space (1) = Not Super Conforming	Alpha	1
27	**PRE-RELIEF REFINANCE LOAN SEQUENCE NUMBER –** The Loan Sequence Number link that associates a Relief Refinance loan to the Loan Sequence Number assigned to the loan from which it was refinanced within in the Single-Family Loan-Level Dataset. Note: Populated only for loans where the Relief Refinance Indicator is set to Y. All other loans will be blank.	PYYQnXXXXXXX • Product F = FRM and A = ARM; • YYQn = origination year and quarter; and, • XXXXXXX = randomly assigned digits	Alpha-numeric	12
28	**PROGRAM INDICATOR –** The indicator that identifies if a loan participates in and of the Freddie Mac programs listed in the valid values. Note: The standard dataset discloses these enumerations for loans originated on or after March 1, 2015. The Non-Standard dataset discloses enumerations for loans originated under the HP program between 1999 and February 28, 2015. Underwriting standards for Home Possible prior to March 1, 2015 may be different than the current standards.	H = Home Possible F = HFA Advantage R= Refi Possible 9 = Not Available or Not Applicable	Alpha-numeric	1
29	**RELIEF REFINANCE INDICATOR –** Indicator that identifies whether the loan is part of Freddie Mac's Relief Refinance Program. Loans which are both a Relief Refinance and have an Original Loan-to-Value above 80 are HARP loans.	Y = Relief Refinance Loan Space = Non-Relief Refinance loan	Alpha	1
30	**PROPERTY VALUATION METHOD –** The indicator denoting which method was used to obtain a property appraisal, if any. Note: Populated for loans originated on or after 1/1/2017.	1 = ACE Loans 2 = Full Appraisal 3 = Other Appraisals (Desktop, driveby, external, AVM) 4 = ACE + PDR 9 = Not Available	Numeric	1
31	**INTEREST ONLY INDICATOR (I/O INDICATOR) -** The indicator denoting whether the loan only requires interest payments for a specified period beginning with the first payment date.	Y = Yes N = No	Alpha	1
32	**MI CANCELLATION INDICATOR:** The indicator denoting if the mortgage insurance has been reported as cancelled after the time of Freddie Mac's purchase of the mortgage loan. If a loan did not have mortgage insurance at the time of Freddie Mac's purchase of the mortgage loan, then this field will be disclosed as "Not Applicable."	Y = Canceled N = Not Canceled 7 = Not Applicable 9 = Not Disclosed	Alpha Numeric	1

Source: Freddie Mac

Table 3. Monthly Performance Data File

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
1	**LOAN SEQUENCE NUMBER** - Unique identifier assigned to each loan.	PYYQnXXXXXXX • Product F = FRM and A = ARM; • YYQn = origination year and quarter; and, • XXXXXXX = randomly assigned digits	Alpha-numeric	12
2	**MONTHLY REPORTING PERIOD** – The as-of month for loan information contained in the loan record.	YYYYMM	Date	6
3	**CURRENT ACTUAL UPB** - The Current Actual UPB reflects the mortgage ending balance as reported by the servicer for the corresponding monthly reporting period. For fixed rate mortgages, this UPB is derived from the mortgage balance as reported by the servicer and includes any scheduled and unscheduled principal reductions applied to the mortgage. For mortgages with loan modifications or payment deferrals, the current actual unpaid principal balance could include non-interest bearing "deferred" amounts. The Current Actual UPB will equal the sum of the Current Interest-Bearing UPB (the amortizing principal balance of the mortgage) and the Current Non-Interest Bearing UPB. If the Current Actual UPB is greater than $500 then the value will be rounded to the nearest $1,000 for the first six payment periods. The Current Actual UPB is not rounded after a modification occurs.	**Calculation:** (interest bearing UPB) + (non-interest bearing UPB)	Numeric Literal decimal	12
4	**CURRENT LOAN DELINQUENCY STATUS** – A value corresponding to the number of days the borrower is delinquent, based on the due date of last paid installment ("DDLPI") reported by servicers to Freddie Mac, and is calculated under the Mortgage Bankers Association (MBA) method. If a loan has been acquired by REO, then the Current Loan Delinquency Status will reflect the value corresponding to that status (instead of the value corresponding to the number of days the borrower is delinquent).	• 0 = Current, or less than 30 days delinquent • 1 = 30-59 days delinquent • 2 = 60 – 89 days delinquent • 3 = 90 – 119 days delinquent • And so on… • RA = REO Acquisition •	Alpha-numeric	3
5	**LOAN AGE** - The number of scheduled payments from the time the loan was originated up to and including the current period. For modified loans, the number of scheduled payments from the modification first payment date up to and including the current period. Note that Payment Deferrals are not considered modifications and therefore the loan age is not reset when they are completed.	**Calculation – Non-modified Loans:** ((Monthly Reporting Period) - Loan First Payment Date (MM/YY)) +1 month **Calculation – Modified Loans:** ((Monthly Reporting Period) - Modification First Payment Date (MM/YY)) +1 month	Numeric	3

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
6	**REMAINING MONTHS TO LEGAL MATURITY** - The remaining number of months to the mortgage maturity date. For mortgages with loan modifications, as indicated by "Y" in the Modification Flag field, the calculation uses the modified maturity date.	**Calculation – Non-modified Loans:** (Maturity Date (MM/YY) – Monthly Reporting Period (MM/YY) **Calculation – Modified Loans:** (Modified Maturity Date (MM/YY) – Monthly Reporting Period (MM/YY)	Numeric	3
7	**DEFECT SETTLEMENT DATE:** For underwriting defects, the date on which there is the occurrence of any of the following: (a) such mortgage is repurchased by the related seller or servicer, (b) in lieu of repurchase, an alternative remedy (such as indemnification) is mutually agreed upon by both Freddie Mac and the seller or servicer or (c) Freddie Mac in its sole discretion elects to waive the enforcement of a remedy against the seller or servicer in respect of such unconfirmed underwriting defect For servicing defects, the date on which there is the occurrence of any of the following: (a) the related servicer repurchased such mortgage or made Freddie Mac whole resulting in a full recovery of losses incurred ("Make Whole") or (b) the party responsible for the representations and warranties and/or servicing obligations or liabilities with respect to the mortgage that is determined to have an unconfirmed servicing defect becomes subject to a bankruptcy, an insolvency proceeding or a receivership.	YYYYMM	Date	6
8	**MODIFICATION FLAG** – A flag indicating if the loan has been modified in the current period or a prior period.	• Y = Current Period Modification • P = Prior Period Modification • Null = Not Modified	Alpha	1
9	**ZERO BALANCE CODE** - A code indicating the reason the loan's balance was reduced to zero.	• 01 = Prepaid or Matured (Voluntary Payoff) • 02 = Third Party Sale • 03 = Short Sale or Charge Off • 09 = REO Disposition • 15 = Whole Loan sales • 16 = Reperforming loan securitizations • 96 = Defect prior to other termination event	Numeric	2
10	**ZERO BALANCE EFFECTIVE DATE** - The date on which the event triggering the Zero Balance Code took place. The period in which the event triggering the Zero Balance Code occurred	YYYYMM Space (6) = Not Applicable	Date	6
11	**CURRENT INTEREST RATE** - Reflects the current interest rate on the mortgage note, taking into account any loan modifications.		Numeric Literal Decimal	8
12	**CURRENT NON-INTEREST BEARING UPB** - The non-interest-bearing portion of the UPB for a given mortgage	$ Amount	Numeric	12
13	**DUE DATE OF LAST PAID INSTALLMENT (DDLPI)** The due date that the loan's scheduled principal and interest is paid through, regardless of when the installment payment was actually made.	YYYYMM	Date	6
14	**MI RECOVERIES** - Mortgage Insurance Recoveries are proceeds received by Freddie Mac in the event of credit losses. These proceeds are based on claims under a mortgage insurance policy.	$ Amount.	Numeric Literal Decimal	12

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
	Note: The MI Recoveries field will be set to zero for loans with a Defect Settlement Date value populated.			
15	**NET SALE PROCEEDS** - The amount remitted to Freddie Mac resulting from a property disposition or loan sale (which in the case of bulk sales, may be an allocated amount) once allowable selling expenses have been deducted from the gross sales proceeds. A value of "U" indicates that the amount is unknown. Note: The Net Sale Proceeds field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount U = Unknown	Alpha-numeric Literal Decimal	14
16	**NON MI RECOVERIES**: Non-MI Recoveries are proceeds received by Freddie Mac based on confirmed defect/make whole proceeds, non-sale income such as refunds (tax or insurance), hazard insurance proceeds, rental receipts, positive escrow, and/or other miscellaneous credits. Note: The Non MI Recoveries field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal	12
17	**TOTAL EXPENSES** - Expenses will include allowable expenses that Freddie Mac bears in the process of acquiring, maintaining and/ or disposing a property (excluding selling expenses, which are subtracted from gross sales proceeds to derive net sales proceeds). This is an aggregation of Legal Costs, Maintenance and Preservation Costs, Taxes and Insurance, and Miscellaneous Expenses. Note: The Total Expenses field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal	12
18	**LEGAL COSTS** - The amount of legal costs associated with the sale of a property (but not included in Net Sale Proceeds. Note: The Legal Costs field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal	12
19	**MAINTENANCE AND PRESERVATION COSTS –**The amount of maintenance, preservation, and repair costs, including but not limited to property inspection, homeowner's association, utilities, and REO management, that is associated with the sale of a property (but not included in Net Sale Proceeds Note: The Maintenance and Preservation Costs field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal	12
20	**TAXES AND INSURANCE** – The amount of taxes and insurance owed that are associated with the sale of a property (but not included in Net Sale Proceeds	$ Amount	Numeric Literal Decimal	12
21	**MISCELLANEOUS EXPENSES** - Miscellaneous expenses associated with the sale of a property (but not included in Net Sale Proceeds Note: The Miscellaneous Expenses field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount	Numeric Literal Decimal	12
22	**ACTUAL LOSS CALCULATION** Actual Loss is calculated using the below approach for loans with Zero Balance Codes of 02, 03, 09, and 15:	$ Amount	Numeric Literal Decimal	12

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
	Actual Loss = (Zero Balance Removal UPB – Net Sale Proceeds) + Delinquent Accrued Interest - Expenses – MI Recoveries – Non MI Recoveries. Please note that the following business rules are applied to this calculation: a. For all loans, 35 bps is used as a proxy for servicing fee, servicing fee will be used if higher than 35 bps. b. The Actual Loss will be set to zero for loans with a Defect Settlement Date value populated. c. The Actual Loss will be set to zero for loans with Net Sales Proceeds = 'U" (Net Sales Proceeds are missing), or expenses are not available. d. The Actual Loss will be set to missing for loans disposed within three months prior to the performance cutoff date. a. e. Modification Costs are not included in the calculation of the Actual Loss field			
23	**CUMULATIVE MODIFICATION COST –** The cumulative modification cost amount calculated when Freddie Mac determines such mortgage loan has experienced a rate modification event or a UPB forbearance. This amount will be calculated on a monthly basis beginning with the first reporting period a modification event is reported and disclosed in the last performance record. For example: calculate monthly modification cost as (min(Origination ANY, (Original Interest Rate - 0.35))/1200 * Current Actual UPB) – (min(Current ANY,(Current Interest Rate-0.35))/1200 * (Interest bearing UPB)) ,and aggregate each month since modification through the Performance Cutoff Date into a cumulative amount. *For loans that go through a payment deferral program, cumulative modification cost includes interest foregone on the UPB deferred by the payment deferral until the last performance record.*	$ Amount	Numeric Literal Decimal	12
24	**STEP MODIFICATION FLAG** – A Y/N flag will be disclosed for every modified loan in their current period of modification, to denote if the terms of modification agreement call for note rate to increase over time.	• Y = Step Mod • N = Non-Step Mod • Null = Loan not modified in period	Alpha	1
25	**PAYMENT DEFERRAL** – A flag indicating a loan has been granted a Payment Deferral in the current or prior period.	• Y = Current Period • P = Prior Period • Null = Not Payment Deferral	Alpha	1
26	**ESTIMATED LOAN TO VALUE (ELTV) –** A ratio indicating current LTV based on the estimated current value of the property obtained through Freddie Mac's Automated Valuation Model (AVM). For more information on our proprietary AVM please visit https://sf.freddiemac.com/tools-learning/home-value-suite/home-value-explorer	• 1 – 998 • 999 = Unknown • Null = Data Not Available	Numeric Literal	4
27	**ZERO BALANCE REMOVAL UPB –** The amount of total UPB remaining on the loan immediately prior to the application of the Zero Balance Code.	$ Amount	Numeric Literal Decimal	12
28	**DELINQUENT ACCRUED INTEREST** – The amount of delinquent interest owed by the borrower at the time of default. This field will only be populated for Zero Balance Codes 02, 03, 09, & 15. Note: The Delinquent Accrued Interest field will be set to zero for loans with a Defect Settlement Date value populated.	$ Amount Delinquent Accrued Interest = (Zero Balance Removal UPB – Non Interest bearing UPB) * Min(Current Interest rate – 0.35, Current Interest Rate – Servicing Fee) * (	Numeric Literal Decimal	12

COLUMN POSITION	FORMAL NAME AND DEFINITION	VALID VALUES/ CALCULATIONS	TYPE	LENGTH
		Months between Last Principal & Interest paid-to date and zero balance date) * 30/360/100.		
29	**DELINQUENCY DUE TO DISASTER** – A flag indicating that the Servicer has reported disaster-related hardship as defined by the Freddie Mac Seller/Servicer Guide. Note: Only populated for January 2014 and following periods.	Y = Delinquency Due to Disaster Null = Not Delinquency Due to Disaster	Alpha	1
30	**BORROWER ASSISTANCE STATUS CODE –** Regardless of delinquency status, the type of assistance plan that the borrower is enrolled in that provides temporary mortgage payment relief or an opportunity to cure a mortgage delinquency over a defined period. Note: Only populated for January 2014 and following periods.	F = Forbearance R = Repayment T = Trial Period Null = No workout plan or not applicable	Alpha	1
31	**CURRENT MONTH MODIFICATION COST:** The current month modification cost amount calculated when Freddie Mac determines such mortgage loan has experienced a rate modification, principal forbearance, or Payment Deferral. This amount will be calculated on a monthly basis beginning with the first reporting period such modification event is reported and disclosed until the last performance record. For example: calculate monthly modification cost as (min(Origination ANY, (Original Interest Rate - 0.35))/1200 * Current Actual UPB) – (min(Current ANY,(Current Interest Rate-0.35))/1200 * (Interest bearing UPB)) . For a loan that is acquired by Freddie Mac Homesteps and subsequently disposed as REO, Current Month Modification Cost represents the monthly modification cost aggregated during the period between REO acquisition and REO disposition. For loans that go through a payment deferral program, modification cost is calculated as interest foregone on the UPB deferred by the payment deferral.	$ Amount	Numeric Literal Decimal	12
32	**INTEREST BEARING UPB:** The current interest bearing UPB of the modified mortgage.	$ Amount	Numeric Literal Decimal	12

Source: Freddie Mac

Table 4. List of Predictor (Independent) Variables Used in LightGBM for ELTV Prediction (Predictor Feature in LightGBM language)

Predictor
Credit Score
Original Loan-to-Value (LTV)
Original Debt-to-Income (DTI) Ratio
Current Actual UPB
Current Loan Delinquency Status
Loan Age
Remaining Months to Legal Maturity
Current Interest Rate
Current Deferred UPB
Interest Bearing UPB
Original Combined Loan-to-Value (CLTV)
Mortgage Insurance Percentage (MI %)
Number of Units
Original UPB
Original Loan Term
First Time Homebuyer Flag
Channel
Occupancy Status
Prepayment Penalty Mortgage (PPM) Flag
Amortization Type (Formerly Product Type)
Property State
Property Type
Loan Purpose
Seller Name
Servicer Name

Appendix B: Hyperparameter Tuning

Table 5. Hyperparameter Grid for Randomized Search

Hyperparamters
learning_rate: [0.01, 0.05, 0.1, 0.2], n_estimators: [100, 200, 300], num_leaves: [31, 50, 100], max_depth: [-1, 5, 10], min_data_in_leaf: [20, 30, 40], subsample: [0.7, 0.8, 1.0], colsample_bytree: [0.7, 0.8, 1.0]

Note: A total of 100 iterations were performed to find the optimal combination of hyperparameters that minimized the model's error on the validation set.

Appendix C: Python Code for ELTV Prediction

Table 6. Python Code for ELTV Prediction using the Standard LightGBM Model

Python Code

```
import pandas as pd
import lightgbm as lgb
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.metrics import mean_squared_error, r2_score

# Function to create and save combined feature importance plots
def plot_combined_feature_importance(pre_pandemic_scores, pandemic_scores, post_pandemic_scores,
feature_names):
    # Set up the figure and axis
    bar_width = 0.2
    index = np.arange(len(feature_names))

    # Set offsets for each period (Pre-Pandemic, Pandemic, Post-Pandemic)
    offsets = [-bar_width, 0, bar_width]

    # Define grey shades for each period
    greys = ['#4d4d4d', '#7f7f7f', '#b3b3b3']

    # Create a figure
    fig, ax = plt.subplots(figsize=(12, 8))

    # Plot the bars for each period
    ax.barh(index + offsets[0], pre_pandemic_scores, bar_width, color=greys[0], label='Pre-Pandemic')
    ax.barh(index + offsets[1], pandemic_scores, bar_width, color=greys[1], label='Pandemic')
    ax.barh(index + offsets[2], post_pandemic_scores, bar_width, color=greys[2], label='Post-Pandemic')

    # Labeling
    ax.set_xlabel("Feature Importance Score", fontsize=12, fontname="Times New Roman")
    ax.set_ylabel("Feature", fontsize=12, fontname="Times New Roman")
    ax.set_yticks(index)
    ax.set_yticklabels(feature_names, fontsize=10, fontname="Times New Roman")

    # Add a legend
```

```
    ax.legend()

    # Gridlines and layout adjustments
    ax.grid(True, axis='x', linestyle='--', alpha=0.6)  # Light gray gridlines
    plt.tight_layout()

    # Display the combined plot
    plt.show()

# Function to extract and return feature importances
def extract_feature_importances(best_model, X):
    feature_importances = pd.DataFrame({
        'Feature': X.columns,
        'Importance': best_model.feature_importances_
    }).sort_values(by='Importance', ascending=False)
    return feature_importances['Importance'][:10], feature_importances['Feature'][:10]

# Function for training the model, tuning it, and evaluating the performance
def train_and_evaluate_model(file_path, features, target, model_name, period):
    # Load dataset
    data = pd.read_excel(file_path, sheet_name='Sheet1')

    # Ensure all selected features and target exist in the dataset
    data.columns = data.columns.str.strip()  # Remove leading/trailing spaces
    missing_columns = [col for col in features + [target] if col not in data.columns]

    if missing_columns:
        raise KeyError(f"The following columns are missing from the dataset: {missing_columns}")

    # Define predictors (X) and target (y)
    X = data[features]
    y = data[target]

    # Convert categorical features to 'category' dtype for LightGBM to process
    categorical_columns = [
        'First Time Homebuyer Flag',
        'Channel',
        'Occupancy Status',
        'Prepayment Penalty Mortgage (PPM) Flag',
        'Amortization Type (Formerly Product Type)',
        'Property State',
        'Property Type',
        'Loan Purpose',
        'Seller Name',
        'Servicer Name'
    ]

    for col in categorical_columns:
        X[col] = X[col].astype('category')

    # Train-test split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Initialize LightGBM regressor
    model = lgb.LGBMRegressor()

    # Define the hyperparameters grid for tuning
    param_grid = {
        'learning_rate': [0.01, 0.05, 0.1, 0.2],
```

```
        'n_estimators': [100, 200, 300],
        'num_leaves': [31, 50, 100],
        'max_depth': [-1, 5, 10],
        'min_data_in_leaf': [20, 30, 40],
        'subsample': [0.7, 0.8, 1.0],
        'colsample_bytree': [0.7, 0.8, 1.0]
    }

    # Perform random search with cross-validation (3-fold)
    random_search = RandomizedSearchCV(estimator=model, param_distributions=param_grid, n_iter=100,
cv=3,
                                       verbose=1, n_jobs=-1, random_state=42, scoring='neg_mean_squared_error')

    # Fit the random search model
    random_search.fit(X_train, y_train)

    # Best parameters found
    print(f"Best parameters for {model_name}: {random_search.best_params_}")

    # Use the best model from random search
    best_model = random_search.best_estimator_

    # Make predictions
    y_pred = best_model.predict(X_test)

    # Evaluate the model
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    print(f"Mean Squared Error for {model_name}: {mse}")
    print(f"R² Score for {model_name}: {r2}")

    # Create and save the feature importance plots
    importance_scores, top_features = extract_feature_importances(best_model, X)

    # Save the results (optional)
    output_file_path = f"D:\\Msc Managmt\\Work Project\\My Work
project\\Data_manipulated_Beta\\{model_name}_LightGBM_model_results_{period}.xlsx"
    results = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
    results.to_excel(output_file_path, index=False)

    print(f"Results saved to {output_file_path}")

    return importance_scores, top_features

# Define the list of features and target for all three datasets
features = [
    'Credit Score',
    'Original Loan-to-Value (LTV)',
    'Original Debt-to-Income (DTI) Ratio',
    'Current Actual UPB',
    'Current Loan Delinquency Status',
    'Loan Age',
    'Remaining Months to Legal Maturity',
    'Current Interest Rate',
    'Current Deferred UPB',
    'Interest Bearing UPB',
    'Original Combined Loan-to-Value (CLTV)',
    'Mortgage Insurance Percentage (MI %)',
```

```
    'Number of Units',
    'Original UPB',
    'Original Loan Term',
    'First Time Homebuyer Flag',
    'Channel',
    'Occupancy Status',
    'Prepayment Penalty Mortgage (PPM) Flag',
    'Amortization Type (Formerly Product Type)',
    'Property State',
    'Property Type',
    'Loan Purpose',
    'Seller Name',
    'Servicer Name'
]
target = 'Estimated Loan-to-Value (ELTV)'

# Call the function for the three different datasets and capture importance scores and feature names
pre_pandemic_scores, pre_pandemic_features = train_and_evaluate_model(
    r"D:\Msc Managmt\Work Project\My Work
project\Data_manipulated_Beta\output_combined_filtered_201712_to_201911_2.xlsx",
    features,
    target,
    'Pre_Pandemic',
    'Pre_Pandemic'
)

pandemic_scores, pandemic_features = train_and_evaluate_model(
    r"D:\Msc Managmt\Work Project\My Work
project\Data_manipulated_Beta\output_combined_filtered_201912_to_202111_2.xlsx",
    features,
    target,
    'Pandemic',
    'Pandemic'
)

post_pandemic_scores, post_pandemic_features = train_and_evaluate_model(
    r"D:\Msc Managmt\Work Project\My Work
project\Data_manipulated_Beta\output_combined_filtered_202112_to_202311_2.xlsx",
    features,
    target,
    'Post_Pandemic',
    'Post_Pandemic'
)

# Ensure the feature names are consistent across all periods (using the top 10 features)
combined_features = pre_pandemic_features  # We can use one set of features for consistency

# Combine and plot all the feature importances in a single plot
plot_combined_feature_importance(
    pre_pandemic_scores, pandemic_scores, post_pandemic_scores, combined_features
)
```

Table 7. Python Code for ELTV Prediction using the Optimized LightGBM Model

Python Code

```
import pandas as pd
import lightgbm as lgb
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.metrics import mean_squared_error, r2_score

# Function to create and save combined feature importance plots
def plot_combined_feature_importance(pre_pandemic_scores, pandemic_scores,
post_pandemic_scores, feature_names):
    # Set up the figure and axis
    bar_width = 0.2
    index = np.arange(len(feature_names))

    # Set offsets for each period (Pre-Pandemic, Pandemic, Post-Pandemic)
    offsets = [-bar_width, 0, bar_width]

    # Define grey shades for each period
    greys = ['#4d4d4d', '#7f7f7f', '#b3b3b3']

    # Create a figure
    fig, ax = plt.subplots(figsize=(12, 8))

    # Plot the bars for each period
    ax.barh(index + offsets[0], pre_pandemic_scores, bar_width, color=greys[0], label='Pre-Pandemic')
    ax.barh(index + offsets[1], pandemic_scores, bar_width, color=greys[1], label='Pandemic')
    ax.barh(index + offsets[2], post_pandemic_scores, bar_width, color=greys[2], label='Post-
Pandemic')

    # Labeling
    ax.set_xlabel("Feature Importance Score", fontsize=12, fontname="Times New Roman")
    ax.set_ylabel("Feature", fontsize=12, fontname="Times New Roman")
    ax.set_yticks(index)

    ax.set_yticklabels(feature_names, fontsize=10, fontname="Times New Roman")

    # Add a legend
    ax.legend()

    # Gridlines and layout adjustments
    ax.grid(True, axis='x', linestyle='--', alpha=0.6)  # Light gray gridlines
    plt.tight_layout()

    # Display the combined plot
    plt.show()

# Function to extract and return feature importances
def extract_feature_importances(best_model, X):
    feature_importances = pd.DataFrame({
        'Feature': X.columns,
        'Importance': best_model.feature_importances_
    }).sort_values(by='Importance', ascending=False)
    return feature_importances['Importance'][:10], feature_importances['Feature'][:10]

# Function for training the model, tuning it, and evaluating the performance
def train_and_evaluate_model(file_path, features, target, model_name, period):
```

```
    # Load dataset
    data = pd.read_excel(file_path, sheet_name='Sheet1')

    # Ensure all selected features and target exist in the dataset
    data.columns = data.columns.str.strip()  # Remove leading/trailing spaces
    missing_columns = [col for col in features + [target] if col not in data.columns]

    if missing_columns:
        raise KeyError(f"The following columns are missing from the dataset: {missing_columns}")

    # Define predictors (X) and target (y)
    X = data[features]
    y = data[target]

    # Convert categorical features to 'category' dtype for LightGBM to process
    categorical_columns = [
        'First Time Homebuyer Flag',
        'Channel',
        'Occupancy Status',
        'Prepayment Penalty Mortgage (PPM) Flag',
        'Amortization Type (Formerly Product Type)',
        'Property State',
        'Property Type',
        'Loan Purpose',
        'Seller Name',
        'Servicer Name'
    ]

    for col in categorical_columns:
        X[col] = X[col].astype('category')

    # Train-test split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Initialize LightGBM regressor
    model = lgb.LGBMRegressor()

    # Define the hyperparameters grid for tuning
    param_grid = {
        'learning_rate': [0.01, 0.05, 0.1, 0.2],
        'n_estimators': [100, 200, 300],
        'num_leaves': [31, 50, 100],
        'max_depth': [-1, 5, 10],
        'min_data_in_leaf': [20, 30, 40],
        'subsample': [0.7, 0.8, 1.0],
        'colsample_bytree': [0.7, 0.8, 1.0]
    }

    # Perform random search with cross-validation (3-fold)
    random_search = RandomizedSearchCV(estimator=model, param_distributions=param_grid,
n_iter=100, cv=3,
                                       verbose=1, n_jobs=-1, random_state=42, scoring='neg_mean_squared_error')

    # Fit the random search model
    random_search.fit(X_train, y_train)

    # Best parameters found
    print(f"Best parameters for {model_name}: {random_search.best_params_}")
```

```
    # Use the best model from random search
    best_model = random_search.best_estimator_

    # Make predictions
    y_pred = best_model.predict(X_test)

    # Evaluate the model
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    print(f"Mean Squared Error for {model_name}: {mse}")
    print(f"R² Score for {model_name}: {r2}")

    # Create and save the feature importance plots
    importance_scores, top_features = extract_feature_importances(best_model, X)

    # Save the results (optional)
    output_file_path = f"D:\\Msc Managmt\\Work Project\\My Work
project\\Data_manipulated_Beta\\{model_name}_LightGBM_model_results_{period}.xlsx"
    results = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
    results.to_excel(output_file_path, index=False)

    print(f"Results saved to {output_file_path}")

    return importance_scores, top_features

# Define the list of features and target for all three datasets
features = [
    'Credit Score',
    'Original Loan-to-Value (LTV)',
    'Original Debt-to-Income (DTI) Ratio',
    'Current Actual UPB',
    'Current Loan Delinquency Status',
    'Loan Age',
    'Remaining Months to Legal Maturity',
    'Current Interest Rate',
    'Current Deferred UPB',
    'Interest Bearing UPB',
    'Original Combined Loan-to-Value (CLTV)',
    'Mortgage Insurance Percentage (MI %)',
    'Number of Units',
    'Original UPB',
    'Original Loan Term',
    'First Time Homebuyer Flag',
    'Channel',
    'Occupancy Status',
    'Prepayment Penalty Mortgage (PPM) Flag',
    'Amortization Type (Formerly Product Type)',
    'Property State',
    'Property Type',
    'Loan Purpose',
    'Seller Name',
    'Servicer Name'
]
target = 'Estimated Loan-to-Value (ELTV)'

# Call the function for the three different datasets and capture importance scores and feature names
pre_pandemic_scores, pre_pandemic_features = train_and_evaluate_model(
    r"D:\Msc Managmt\Work Project\My Work
```

```
project\Data_manipulated_Beta\output_combined_filtered_201712_to_201911_2.xlsx",
    features,
    target,
    'Pre_Pandemic',
    'Pre_Pandemic'
)

pandemic_scores, pandemic_features = train_and_evaluate_model(
    r"D:\Msc Managmt\Work Project\My Work
project\Data_manipulated_Beta\output_combined_filtered_201912_to_202111_2.xlsx",
    features,
    target,
    'Pandemic',
    'Pandemic'
)

post_pandemic_scores, post_pandemic_features = train_and_evaluate_model(
    r"D:\Msc Managmt\Work Project\My Work
project\Data_manipulated_Beta\output_combined_filtered_202112_to_202311_2.xlsx",
    features,
    target,
    'Post_Pandemic',
    'Post_Pandemic'
)

# Ensure the feature names are consistent across all periods (using the top 10 features)
combined_features = pre_pandemic_features  # We can use one set of features for consistency

# Combine and plot all the feature importances in a single plot
plot_combined_feature_importance(
    pre_pandemic_scores, pandemic_scores, post_pandemic_scores, combined_features
)
```

Appendix D: Models Performance

Table 8. Performance of Standard and Optimized LightGBM Models across periods in predicting ELTV (in detail)

Period	Model Type	MSE	R^2	% Change in MSE for same period (Optimized vs. Standard)	% Change in MSE Standard (Pandemic vs Pre-Pandemic)	% Change in MSE Standard (Post-Pandemic vs Pre-Pandemic)	% Change in MSE Standard (Post-Pandemic vs Pandemic)	% Change in MSE Optmized (Pandemic vs Pre-Pandemic)	% Change in MSE Optmized (Post-Pandemic vs Pre-Pandemic)	% Change in MSE Optmized (Post-Pandemic vs Pandemic)	% Change in R^2 for same period (Optimized vs. Standard)	% Change in R^2 Standard (Pandemic vs Pre-Pandemic)	% Change in R^2 Standard (Post-Pandemic vs Pre-Pandemic)	% Change in R^2 Standard (Post-Pandemic vs Pandemic)	% Change in R^2 Optmized (Pandemic vs Pre-Pandemic)	% Change in R^2 Optmized (Post-Pandemic vs Pre-Pandemic)	% Change in R^2 Optmized (Post-Pandemic vs Pandemic)
Pre-Pandemic	**Standard**	12465.99	0.872	-							-						
Pre-Pandemic	**Optimized**	8626.26	0.911	-30.80							4.52						
Pandemic	**Standard**	1913.97	0.951	-	-84.65	-66.22	120.00	-86.99	-72.46	111.61	-	9.02	2.79	-5.72	6.54	3.30	-3.04
Pandemic	**Optimized**	1122.63	0.971	-41.35							2.14						
Post-Pandemic	**Standard**	4210.80	0.896	-							-						
Post-Pandemic	**Optimized**	2375.61	0.942	-43.58							5.04						