A New Approach to Addressing Lending Disparities in Home Mortgages

Team 44: Ryan Gould, Shayla Nelson, Cesar Servin, Shane Traister, Leeann Warren

Introduction

The Home Mortgage Disclosure Act (HMDA) is a federal law that requires financial institutions to report on their mortgage lending activity. Since its enactment by Congress in 1975, the HMDA has made a wealth of data available, providing transparency into lending practices [15]. Disparities that exist in mortgage lending have been regularly identified and documented [1, 4, 7, 8, 9, 11, 12]. This project aims to use available HMDA data to build tools that prospective buyers can use to make data-informed decisions about their home loan applications.

Problem Definition

Our goal is to empower mortgage applicants with information that will give them the greatest chance of a successful application and at the most efficient value [H1], by analyzing, understanding, and visualizing HMDA mortgage application data to highlight disparities that exist in loan origination for applicants of different races, ages, and genders [A] [H3]. The modeled outputs will generate a user-level lending forecast with greater statistical accuracy than those accessible via mainstream sites and banking estimates while simultaneously providing more applicant-centric visualizations that rank order state counties through choropleth mapping [A & C].

Literature Review

Government policies such as monetary and fiscal policy helped the economy to boom on the number of home sales during the pandemic [13]. However, differences in outcomes for new homeowners and refinancing applicants persist between minorities and non-Hispanic white borrowers [9]. One study of interest modeled applicant denials using a logit model that incorporated the use of new HMDA variables added to the dataset in 2018 [14]. This analysis will help guide our selection of variables used in our own modeling and we can expand this research to loans other than conventional loans. People of color are more likely to be denied home loan applications in certain areas of the country, including Philadelphia. This state-focused analysis will help us begin our project on a smaller, most geographically-focused scale. The binary output in this logistic regression approach to application denials can be expanded to more areas of the country [H2][12]. In addition to differences in acceptance rates for specific groups of people, we've discovered significant heterogeneity in lending when it comes to lending institutions themselves. The denial rates of lenders can be influenced by factors such as size of the mortgage and credit score [2,10]. While this analysis can help us see a different perspective where other variables, unrelated to minority groups, can cause differences in lending, we plan to recognize more sides to lending disparities than presented in this one-sided approach [3].

Proposed Method

Our approach leverages logistic and random forest models, where they outperform traditional, modern models that often omit demographic attributes such as gender, age, and race, which are unable to be incorporated into institutional models due to legal constraints such as the Fair Lending Act [18] [A].

H1-H9: Heilmeier Questions
1-17: Literature Review Reference
A-C: Key Innovations

Despite the legal limitations, it is evident that statistical differences based on these attributes have been observed in lending practices. Building on the progress generated by Popick [14] and Submitter [15], our approach will utilize post-2018 HMDA data, which is rich in demographic data, and incorporate the latest data through 2023 [B]. Through additional data cleaning, updated variable selection, upgraded modeling, and user-centric visualizations, the results should set a new bar for state of the art forecasting within the mortgage lending community [A, B, C].

Detailed Description of Approaches:

Data modeling in this study is conducted using a logistic regression as this provides good interpretability and will help the deployment of the returns to the interactive visualization tool.

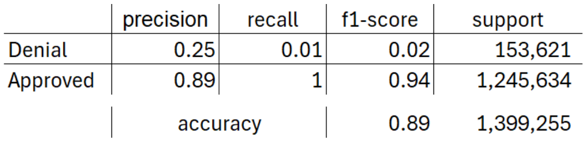
$$P\ (Approved\ Loan)\ =\ 1\ /\ 1\ +\ e^{-Features}$$

Features = β0 + Race Ethnicityβ1 + Loan Amountβ2 + Property Valueβ3 + Loan Termβ4 + Interest Rateβ5 + Stateβ6 + Yearβ7 + Occupancy Typeβ8 + Log Incomeβ9 + Loan to valueβ10 + Debt to Incomeβ11

Prior to modeling, the dataset underwent a rigorous data cleaning process to ensure all data was formatted correctly and possessed proper integrity. This process involved identifying and addressing missing values, imputing them if possible, and removing outlier errors that may have skewed the results. If a loan application was missing the applicant's information, the observation was dropped to keep our model focus on the data available. Categorical variables were transformed using one-hot encoding to enable their use in the modeling process. For example, categorical variables such as race and gender were encoded as binary variables (e.g., 0 or 1) to allow the algorithms to incorporate them into the analysis. This transformation ensured that categorical variables were appropriately represented in the models and did not introduce bias or inaccuracies.

Variable selection will continue to play a pivotal role in the modeling process to ensure that only the most relevant features are retained in the final models. So far we’ve employed a variety of techniques including drawing insights from similar studies in the field which provided valuable guidance and initial progress, shedding light on significant areas. To begin, variables with extremely low variance thresholds were eliminated. To address multicollinearity, we evaluated the correlation between variables and eliminated highly correlated ones to prevent redundancy in the model.

Based on our classification performance metrics we have quite robust values across multiple models. We decided to use this model with these features as a basis to potentially append counties to our model to meet our visualization tool. Below are the metrics showing our model has 89% accuracy with 89% precision on our approvals and close to 100% recall meaning is predicting approvals at a high rate. The sampling had a split of 80% in training and 20% for testing.



	precision	recall	f1-score	support
Denial	0.25	0.01	0.02	153,621
Approved	0.89	1	0.94	1,245,634
	accuracy		0.89	1,399,255

As part of our approach, we will utilize county-level choropleth mapping via JavaScript and the D3 library for the three case states to visualize the data and generate insights into areas with higher or lower likelihoods of successful mortgage applications based on users' specific inputs. These user inputs will include demographic variables that our model has identified as having a significant impact on loan application success (input variables yet to be determined). The user interface will include dropdown selections and the displayed map will change dynamically based on inputted values. This visualization technique of choropleth mapping offers several advantages, including providing users with a clear and intuitive representation of the data. By aggregating modeled mortgage application data onto geographic regions, users will be able to identify areas that may offer more favorable lending conditions for individuals with similar financial inputs as well as demographics. The choropleth mapping will utilize red and blue color gradients to highlight areas that are more and less favorable, respectively, for successful mortgage applications. Regions shaded in red will indicate higher application success rates, while those shaded in blue will signify lower success rates. This color-coded visualization allows users to easily identify areas with the most promising lending conditions and areas where additional support may be needed. Tooltips will also be enabled so that users can hover over a specific county to see its name and its predicted loan application success rate, based on the user’s selections. By presenting the data in this digestible format, users can gain valuable insights into the geographic distribution of lending disparities and make more informed decisions about their home loan applications.

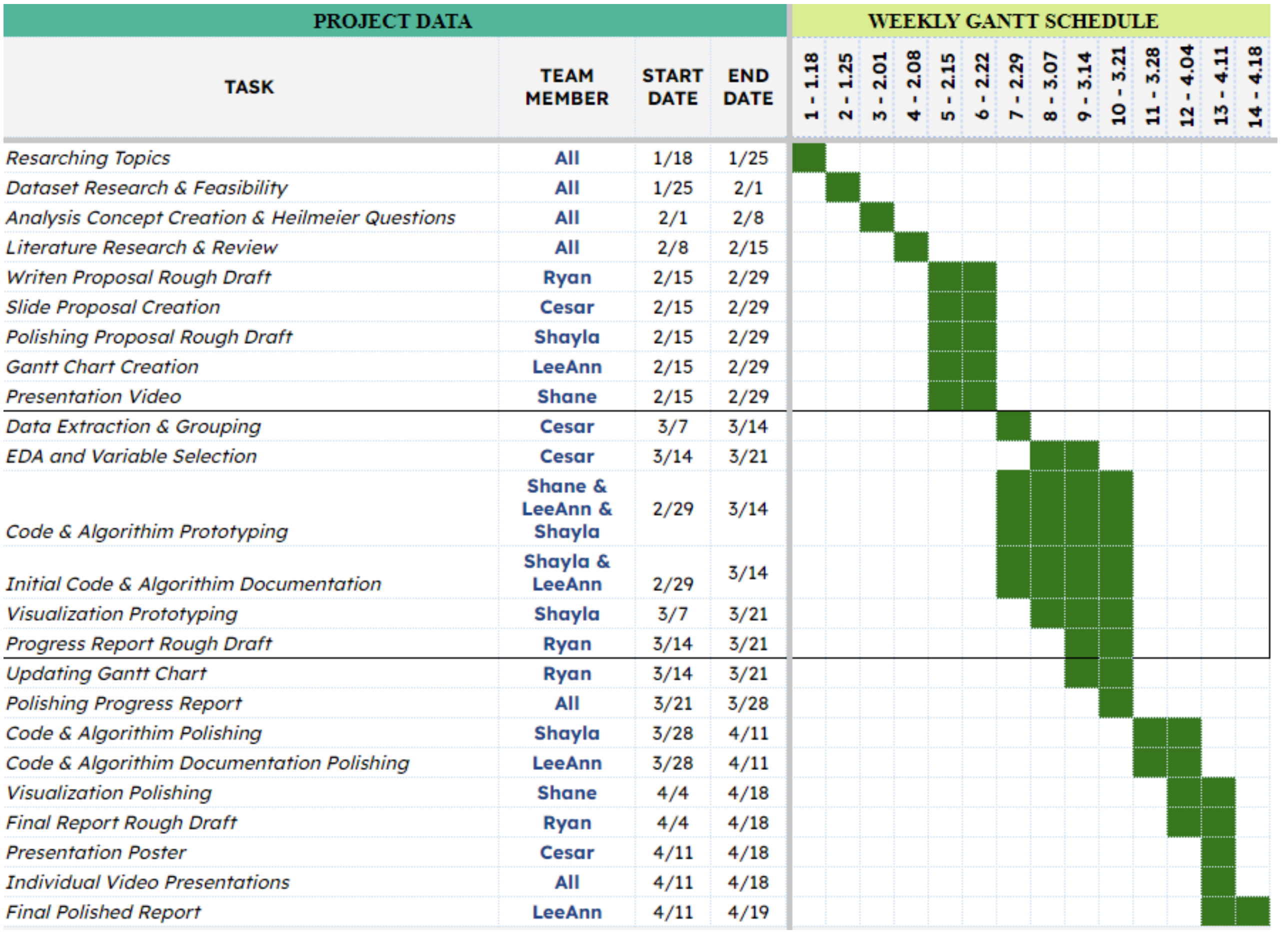
Experiments / Evaluation

In the experiments/evaluation phase, we will assess the proficiency of the proposed approach through three key evaluations: proficiency of the model forecasting, cross validations, and UI usability testing via sample users. Firstly, accuracy tests and confusion matrices will evaluate the predictive performance of the model outputs vs sampled actuals. Secondly, cross-validation techniques will validate the robustness and generalization capabilities. Lastly, usability testing of the visualization tool with sample users will assess its effectiveness and user-friendliness through test sessions.

Conclusion

In addition to a predictive model using logistic regression, we are exploring more sophisticated models for explaining feature importance, including XGBoost. We have tried initial XGBoost and SVM models but need to further explore virtualization for performance. We have experimented with random forest classification but we have concerts of overfitting and interpretability. Ultimately, our final production model will be decided by a combination of effectiveness and our ability to integrate it into our tool. Overall, the signs of the coefficient of our logistics regression make sense on the impact on loan application approvals and initial results suggest sufficient statistical power to continue exploring.

Plan of Activities



PROJECT DATA				WEEKLY GANTT SCHEDULE													
TASK	TEAM MEMBER	START DATE	END DATE	1 - 1.18	2 - 1.25	3 - 2.01	4 - 2.08	5 - 2.15	6 - 2.22	7 - 2.29	8 - 3.07	9 - 3.14	10 - 3.21	11 - 3.28	12 - 4.04	13 - 4.11	14 - 4.18
Resarching Topics	All	1/18	1/25	🟩													
Dataset Research & Feasibility	All	1/25	2/1		🟩												
Analysis Concept Creation & Heilmeier Questions	All	2/1	2/8			🟩											
Literature Research & Review	All	2/8	2/15				🟩										
Writen Proposal Rough Draft	Ryan	2/15	2/29					🟩	🟩								
Slide Proposal Creation	Cesar	2/15	2/29					🟩	🟩								
Polishing Proposal Rough Draft	Shayla	2/15	2/29					🟩	🟩								
Gantt Chart Creation	LeeAnn	2/15	2/29					🟩	🟩								
Presentation Video	Shane	2/15	2/29					🟩	🟩								
Data Extraction & Grouping	Cesar	3/7	3/14							🟩							
EDA and Variable Selection	Cesar	3/14	3/21								🟩	🟩					
Code & Algorithim Prototyping	Shane & LeeAnn & Shayla	2/29	3/14							🟩	🟩	🟩	🟩				
Initial Code & Algorithim Documentation	Shayla & LeeAnn	2/29	3/14							🟩	🟩	🟩	🟩				
Visualization Prototyping	Shayla	3/7	3/21								🟩	🟩	🟩				
Progress Report Rough Draft	Ryan	3/14	3/21									🟩	🟩				
Updating Gantt Chart	Ryan	3/14	3/21									🟩	🟩				
Polishing Progress Report	All	3/21	3/28										🟩				
Code & Algorithim Polishing	Shayla	3/28	4/11											🟩	🟩		
Code & Algorithim Documentation Polishing	LeeAnn	3/28	4/11											🟩	🟩		
Visualization Polishing	Shane	4/4	4/18												🟩	🟩	
Final Report Rough Draft	Ryan	4/4	4/18												🟩	🟩	
Presentation Poster	Cesar	4/11	4/18													🟩	
Individual Video Presentations	All	4/11	4/18													🟩	
Final Polished Report	LeeAnn	4/11	4/19													🟩	🟩

All team members have and will contribute equally.

H1-H9: Heilmeier Questions

1-17: Literature Review Reference

A-C: Key Innovations

References

1. Agarwal, S., Green, R., Rosenblatt, E., Yao, V., & Zhang, J. (2018). Gender difference and intra-household economic power in mortgage signing order. *Journal of Financial Intermediation*, *36*, 86–100. https://doi.org/10.1016/j.jfi.2018.01.001

2. Ali, S. E. A., Rizvi, S. S. H., Lai, F., Ali, R. F., & Jan, A. A. (2021). Predicting Delinquency on Mortgage Loans: An exhaustive parametric comparison of machine learning techniques. *International Journal of Industrial Engineering and Management*, *Volume 12*(Issue 1), 1–13. https://doi.org/10.24867/ijiem-2021-1-272

3. Bär, M., & Kakar, V. (2022). Lender Heterogeneity in Home Mortgage Lending Evidence from Hmda Data in Context of the Covid-19 Pandemic. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4074749

4. Bhutta, N., Hizmo, A., & Ringo, D. (2022). How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions. *Finance and Economics Discussion Series*, *2022–067*, 1–44. https://doi.org/10.17016/feds.2022.067

5. Cherian, M. (2014). Race in the Mortgage Market: An Empirical Investigation Using HMDA Data. *Race, Gender & Class*, *21*(1/2), 48–63. http://www.jstor.org/stable/43496959

6. Delis, M. D., & Papadopoulos, P. (2018). Mortgage lending discrimination across the U.S.: New methodology and new evidence. *Journal of Financial Services Research*, *56*(3), 341–368. https://doi.org/10.1007/s10693-018-0290-0

7. Diego Mendez-Carbajo, "Neighborhood Redlining, Racial Segregation, and Homeownership," Page One Economics®, September 2021

8. Faber, J. (2017). Segregation and the Geography of Creditworthiness: Racial inequality in a recovered mortgage market. *Housing Policy Debate*, *28*(2), 215–247. https://doi.org/10.1080/10511482.2017.1341944

9. Gerardi, K., Lambie‐Hanson, L., & Willen, P. S. (2021). Racial differences in mortgage refinancing, distress, and housing wealth accumulation during COVID-19. In *Discussion Paper*. https://doi.org/10.21799/frbp.dp.2021.02

10. Goldstein, E., & DeMaria, K. (2022). Small-Dollar Mortgage Lending in Pennsylvania, New Jersey, and Delaware. Federal Reserve Bank of Philadelphia. https://www.philadelphiafed.org/-/media/frbp/assets/community-development/reports/cdro-small-dollar-mortgage-lending-report.pdf

11. Loya, J. (2022). Ethno-racial stratification in the mortgage market: The role of co-applicants. *Social Science Research*, *106*, 102725. https://doi.org/10.1016/j.ssresearch.2022.102725

12. Martinez, E., & Glantz, A. (n.d.). *How Reveal identified lending disparities in federal mortgage data*. Retrieved February 22, 2024, from http://revealnews.org.s3.amazonaws.com/uploads/lending_disparities_whitepaper_180214.pdf

13. Newton, Natalie and Vickery, James Ian, The Pandemic Mortgage Boom (December 31, 2022). Economic Insights, Federal Reserve Bank of Philadelphia, 2022, Q3-Q4: 18-27., Available at SSRN: https://ssrn.com/abstract=4380493

14. Popick, S. (2022). Did minority applicants experience worse lending outcomes in the mortgage market? a study using 2020 expanded HMDA data. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4131603

15. Submitter, B. D. P. (2020). An Updated Review of the New and Revised Data Points in HMDA: Further Observations Using the 2019 HMDA Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3786976

16.Turiel, J., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, *7*(6), 191649. https://doi.org/10.1098/rsos.191649

17. Wheeler, C. H., & Olson, L. M. (2015). Racial differences in mortgage denials over the housing cycle: Evidence from U.S. metropolitan areas. *Journal of Housing Economics*, *30*, 33–49. https://doi.org/10.1016/j.jhe.2015.10.004

18. United States Department of Housing and Urban Development . (n.d.). *Fair Lending: Learn the Facts*. Fair Lending Guide. https://www.hud.gov/sites/documents/fair_lending_guide.pdf