Team 43 Proposal

Harris Ashraf[1], Sharmila Baskaran[2], Bo Chen[3], Travis Jefferies[4], Daniel Mower[5], Cody Nguyen[6], and Ryan Wong[7]

[1]Georgia Institute of Technology, hashraf3@gatech.edu
[2]Georgia Institute of Technology, sbaskaran30@gatech.edu
[3]Georgia Institute of Technology, bchen354@gatech.edu
[4]Georgia Institute of Technology, tjefferies3@gatech.edu
[5]Georgia Institute of Technology, dmower3@gatech.edu
[6]Georgia Institute of Technology, cnguyen311@gatech.edu
[7]Georgia Institute of Technology, rwong33@gatech.edu

1 Introduction (H1, H2)

(H1) Lenders such as banks provide borrowers with mortgage loans, where cash is given upfront to purchase a property. The bank uses the property as collateral, and the borrower makes payments over a specified period of time, typically 30 years. However, a borrower may default on their loan. This allows the lender to retain ownership of the property and the borrower to no longer make payments. The lender assumes the risk of selling the property at a value less than the remaining balance. It is in the bank's best interest to understand the risk of individual mortgage loans.

(H2) Currently banks employ large teams of analysts to create models based on mortgage portfolios, or collections of similar loans, to assess risk. However, with the Financial Accounting Standards Board (FASB) introducing a new regulation for banks in loan loss calculations called the Current Expected Credit Loss (CECL), traditional methods based on portfolio-level models will not comply. Banks now need models that estimate losses based on individual products or loans.

2 Literature Survey

In [2][4][5][9][12], logistic regression and regression trees have been popular to make non-linear, non-parametric predictions of mortgage loan default; however, these methods are limited to binary classification tasks. K-Nearest Neighbors was evaluated in [2][8][18] and resulted in poor loan default probability accuracy, although it surprisingly was fast at training. Many researchers applied Random Forest in [2][8][14][17][21] and found higher accuracy in the mid-90%. In [1][5], the researchers focused on feature selection extracted by deep learning models, and stressed the importance of model setup and industry knowledge. These points are also emphasized in [10], which evaluated mortgage loan features and their contributions to default prediction accuracy.

The authors in [6][12][20] tackled the issue through employing Survival Analysis techniques to predict the time until a mortgage will default, which is useful because our data contains loans that haven't defaulted within the recorded timeframe. We will evaluate whether the aforementioned models (e.g., logistic regression, random forest, survival analysis) will satisfy the constraints. [11][15] provide evaluation criteria that we will use to determine the best performing model at mortgage loan default prediction.

In [3][7], they provide an analysis of the geovisualization tools the authors developed, and specific features such as sliders and drop-down menus that allow users to interact with the tool to gain relevant insights. [16] analyzes various formats of the choropleth map to demonstrate how variations such as contiguous, non-contiguous, and demer cartograms provide very different information for the user. This paper also alludes to spatial treemaps, which the author in [19] has found to potentially provide an effective means of conveying geographically-defined information. Both the choropleth and spatial treemaps will be considered for our visualization based on the type of information we wish to easily provide to our user. We will also consider findings from [13], which emphasizes the need to consider Human-Computer Interaction (HCI) concepts such as user-demographic and colorblindness awareness when designing a geovisualization tool. In terms of evaluating the effectiveness of our tool, [3] presents user surveys based on cursor and map interaction behaviors to evaluate the effectiveness for a user to obtain the information they want.

3 Our Method (H3, H4)

(H3) Our approach will use historical loan data to assess the risk of loss based on individual loan characteristics to abide to the requirement of the CECL. We will expand the time horizon (2 years or more) and include borrower characteris-

tics such as FICO, geography, loan purpose, and other variables, while using the aforementioned data analytic models.

(H4) This new evaluation method will impact large banks, smaller regional lenders, and borrowers. Financial institutions can become CECL compliant, and smaller lenders could leverage new insight to grow their lending practices. Our proposed tool, provided in Figure 1, will allow a lender or borrower to evaluate their own information. The tool allows users to look at historical default behavior based on filters to evaluate risk based on loan parameters.

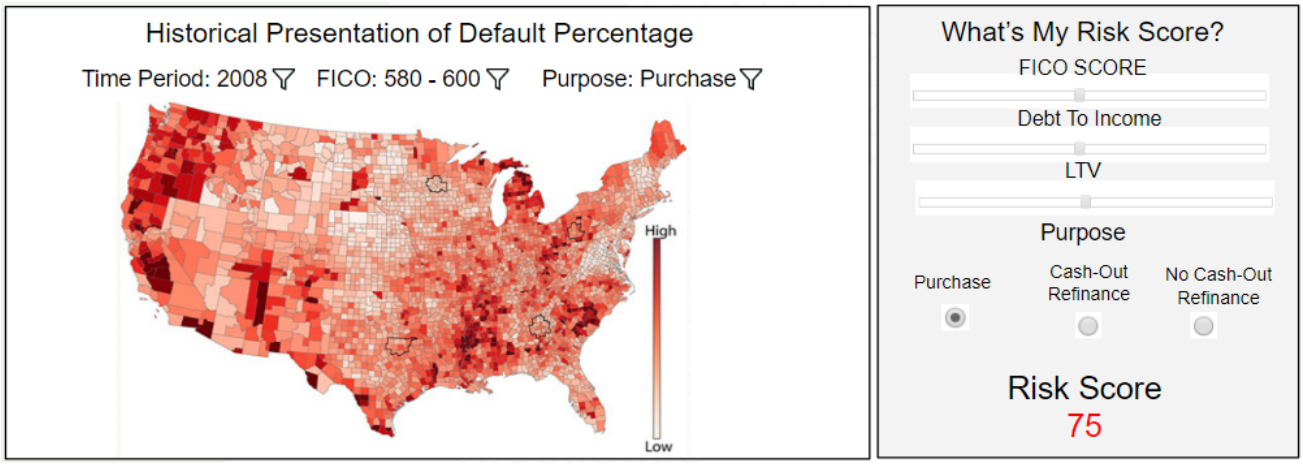

Figure 1. Proposed visualization of our tool; note that the heatmap used is a sample and is not a screenshot of our actual tool.

The tools we will be using are shown in Table 1.

Type of Tool	Tools
Clustering	Hadoop
Modeling	Python
Database Querying	SQLite
Interface	Javascript (D3, React), Tableau
Web Scraping	Selenium
Report Generation	LaTeX

Table 1. Tools that will be used in our project.

4 Impact (H5)

(H5) If successful, our tool would allow banks and lenders to comply with new CECL mandates. The changes in data requirements to CECL would make this tool necessary to cheapen and speed-up their processes.

5 Risk and Cost (H6, H7)

(H6) To measure the impact of these changes we would employ a variety of methods in the field such as user studies and experiments to determine the changes in behavior in both lending and borrowing. We would back test our methods and determine the expected outcomes with the newly derived loan practices these assessments would make. The project grade will define our quantitative success, and our increased understanding of data analytic and finance will define our qualitative success.

(H7) The current cost analysis is zero, based on the assumption of free credits from cloud providers such as AWS, Azure or GCP to run our models in a timely enough fashion for the project.

6 Plan of Activity (H8, H9)

Team member contribution will be equally distributed; however, we will assign responsibilities according to the sub-teams in Table 2. *Data Scraping* will research relevant datasets, and extract and consolidate the data for modeling. *Data Modeling* will design, prototype, and test machine learning models to predict loan risk. *Data Visualization* will develop the front-end interface. *Data Process Clustering* will evaluate methods for processing large datasets. *Project Management* will facilitate project progress to adhere to deliverables.

Table 2. Distribution of work through sub-teams

Team member	Data Scraping	Data Modeling	Data Visualization	Data Process Clustering	Project Management
Harris Ashraf				x	
Sharmila Baskaran	x	x			
Bo Chen			x	x	
Travis Jefferies	x	x			
Daniel Mower	x	x			
Cody Nguyen			x	x	
Ryan Wong			x		x

(H8) Table 3 shows the weekly goals for the *Modeling* (Data Scraping/Modeling/Process Clustering) and *Visualization* (Data Visualization) teams; the project should complete within ten weeks. **(H9)** At the "midterm" check, the modeling will consists of an initial model prototyped on a small sample of our dataset, with static visualizations that display historical loan data across the sample. At the "final" check, the model should encompass the entire dataset, with a front-end that enables users to dynamically query our dataset.

Table 3. Weekly goals for the modeling and visualization components of our tool, and deliverables

Dates	Modeling	Visualization	Deliverable
09/30 - 10/06	Form an understanding of the problem and initial dataset	Form an understanding of the problem and initial dataset	
10/07 - 10/13	Filter and clean the data Research other relevant datasets	Research methods to display geospatial data	Proposal document (10/10, 1400 UTC) Proposal presentation (10/10, 1400 UTC)
10/14 - 10/20	Research and prototype various models	Define a standardization of data inputs into the tool Complete a mock-up of the tool	
10/21 - 10/27	(Stage 1) Refine the model using a small segment of the dataset Begin researching best ways to cluster data processing	Begin creating a prototype tool based on static data parameters	
10/28 - 11/03	(Stage 2) Begin understanding how to dynamically tune model for variable input parameters using clustering methods	Begin creating the site that will integrate the tool Integrate the small dataset segment used for (Stage 1) into the prototype	
11/04 - 11/10	Implement and test the dynamic model		Progress report (11/07, 1400 UTC)
11/11 - 11/17		Integrate the small dataset segment used for (Stage 1) into the prototype Work with modeling team to integrate the rest of the dataset into the tool	
11/18 - 11/24	Begin summarizing findings	Begin summarizing findings	
11/25 - 12/01			Poster presentation video (11/28, 1400 UTC) Final report 11/28, 1400 UTC
12/02 - 12/08			Poster presentation video grading starts (12/03, 1400 UTC) Poster presentation video grading due (12/7, 1400 UTC)

7 Bibliography

[1] Addo, P. M., Guegan, D., Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risk, 6(38). doi:doi:10.3390/risks6020038. Retrieved from https://www.mdpi.com/2227-9091/6/2/38/pdf.

[2] Akindaini, B. (2017). Machine Learning Applications in Mortgage Default Prediction (Unpublished master's thesis). University of Tampere. Retrieved from http://tampub.uta.fi/bitstream/handle/10024/102533/1513083673.pdf.

[3] Aoidh, E. M., Bertolotto, M., Wilson, D. C. (2008). Understanding geospatial interests by visualizing map interaction behavior. Information Visualization, 7(3-4). doi:10.1057/palgrave.ivs.200824. Retrieved from http://journals.sagepub.com/doi/pdf/10.1057/IVS.2008.24.

[4] Bagherpour, A. (2017). Predicting Mortgage Loan Default with Machine Learning Methods. Retrieved from http://economics.ucr.edu/job_candidates/Bagherpour-Paper.pdf.

[5] Baldominos, A., Jose Moreno, A., Iturrarte, R., Bernardez, O., Alfonso, C. (2018). Identifying Real Estate Opportunities using Machine Learning. Retrieved from http://adsabs.harvard.edu/cgi-bin/bib_query?arXiv:1809.04933.

[6] Bhattacharya, Arnab P. Wilson, Simon Soyer, Refik. (2017). A Bayesian approach to modeling mortgage default

and prepayment. https://arxiv.org/pdf/1706.07677.pdf

[7] Dang, G., North, C., Schneiderman, B. (2001). Dynamic queries and brushing on choropleth maps. Retrieved from https://ieeexplore.ieee.org/document/942141/authorsauthors.

[8] Deng, G. (2016). Analyzing the Risk of Mortgage Default (Unpublished master's thesis). Retrieved from https://www.stat.berkeley.edu/ aldous/Research/Ugrad/Grace_Deng_thesis.pdf

[9] Hoaglin, David C., Frederick Mosteller, and John W. Tukey. Exploring data tables, trends, and shapes. New York: Wiley-Interscience, 1985. Print. Chapters 10-11

[10] Hwang, S., Park, M., Lee, H. (2013). Dynamic analysis of the effects of mortgage-lending policies in a real estate market. Mathematical and Computer Modelling, 57(9-10), 2106-2120. doi:10.1016/j.mcm.2011.06.023

[11] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. Loss models : from data to decisions. Hoboken, N.J: John Wiley Sons, 2008. Print. Chapters 16-17

[12] Li, M. (2014, October). Residential Mortgage Probability of Default Models and Methods (Rep.). Retrieved from https://www.fic.gov.bc.ca/pdf/fid/14-0877-sup.pdf

[13] Marsh, S. L. (2007). Using and Evaluating HCI Techniques in Geovisualization: Applying Standard and Adapted Methods in Research and Education (Unpublished master's thesis). Retrieved from https://pdfs.semanticscholar.org/c096/5d4427673401320ff927e860600f65d6e89a.pdf

[14] Moosavi, V. (2017). Urban Data Streams and Machine Learning: A Case of Swiss Real Estate Market. Retrieved from https://arxiv.org/abs/1704.04979.

[15] Servigny, and Olivier Renault. Measuring and managing credit risk. New York: McGraw-Hill, 2004. Print. Chapter 3: Credit Scoring

[16] Skowronnek, Alsino. "Beyond Choropleth Maps: A Review of Techniques to Visualize Quantitative Areal Geodata." Alsino.io, INFOVIS READING GROUP WS 2015/16, 2015 alsino.io/static/papers/BeyondChoropleths_AlsinoSkowronnek.pdf. https://alsino.io/static/papers/BeyondChoropleths_AlsinoSkowronnek.pdf

[17] Sonderby, S. (2014). Non-parametric survival analysis in breast cancer using clinical and genomic markers (Unpublished master's thesis). Technical University of Denmark. Retrieved from http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6779/pdf/imm6779.pdf

[18] Tan, P., Steinbach, M., Karpatne, A., Kumar, V. (2019). Introduction to Data Mining, 2nd Edition. Retrieved from http://www.mypearsonstore.com/bookstore/introduction-to-data-mining-9780133128901

[19] Wood, J. Dykes, J. (2008). Spatially Ordered Treemaps. IEEE Transactions on Visualization and Computer Graphics, 14(6), pp. 1348-1355. doi: 10.1109/TVCG.2008.165 http://openaccess.city.ac.uk/536/1/wood_spatially_2008.pdf

[20] Zhang, Q. (2015). Modeling the Probability of Mortgage Default via Logistic Regression and Survival Analysis (Unpublished master's thesis). University of Rhode Island. Retrieved from https://digitalcommons.uri.edu/cgi/viewcontent.cgi?referer=https://www.google.com/httpsredir=1article=1543context=theses

[21] Zhou, L., Wang, H. (2012). Loan Default Prediction on Large Imbalanced Data Using Random Forests. Indonesian Journal of Electrical Engineering, 10(6), 1519-1525. Retrieved from https://www.researchgate.net/publication/267864165_Loan_Default_Prediction_on_Large_Imbalanced_Data_Using_Random_Forests.