The Successful Loan Applicant

DATA 450 Capstone

Thomas Ortega

2/13/23

1 Introduction

Loans are a fundamental aspect of the United States economy and economies worldwide. Loans provide funding to people lacking the current resources to consolidate debt, buy a house, buy a car, open a business, and any other transaction that requires money. For loan companies to make a profit, loans are expected to be paid back with interest. Companies want to know which applicants that apply for a loan will pay it back. Selecting a string of clients that default would spell disaster for the lender. With teams of analysts, companies build models to identify candidates. People outside of these lending companies, including me, want to know what traits make a candidate more or less likely to receive a loan. This project will dig into a Kaggle dataset containing loan applicants to find insights and build visuals describing the populations of both accepted and rejected applicants. Also, the project will build a model using some of the features in the dataset to predict whether a customer is accepted or rejected. The original data belongs to a business called LendingClub. LendingClub is an FDIC-insured financial services company that provides peer-to-peer loan opportunities and other resources.

2 Dataset

The Kaggle dataset has two separate csv files: one for accepted applicants and one for rejected applicants (George (2019)). The accepted file contains over two million rows and one hundred and fifty-one columns. The rejected data contains over twenty-seven million rows and nine columns. The Kaggle posting does not efficiently provide a data dictionary for every column; however, another post in Kaggle provides information on each column (Chan (2018)). As stated in the introduction, the data is from LendingClub. The author of the Kaggle dataset periodically updated the Kaggle dataset; however, he stopped updating the dataset in 2019. The dataset contains LendingClub data from 2007 to 2018. Other datasets on Kaggle use LendingClub data, but this dataset contains a wider date range than most. LendingClub appears to no longer have the option to download this data.

2.1 Variables Under Examination:

2.1.1 Variables in Both Datasets

- Amount Requested: The total amount requested by the borrower (USD).
- Loan Title: The reason for applying for the loan (debt consolidation, business loan, etc.).
- Policy Code: Publicly available policy_code=1 new products not publicly available policy_code=2.
- Risk Score: The borrower's FICO score.
- Debt-To-Income Ratio: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- Zip Code: The first 3 numbers of the zip code provided by the borrower in the loan application.
- State: The state provided by the borrower in the loan application.
- Employment Length: Reported length of employment. Possible values are between 0 and 10 where 0 is less than one year and 10 means ten or more years.

2.1.2 Variables in the Accepted Dataset Only

- Interest Rate: The interest rate on the loan. Observations wrote as decimals (0.07 is 7%).
- Employee Title: The job title supplied by the borrower.
- Funded Amount: The total amount committed to that loan.
- Hardship Flag: Flags whether or not the borrower is on a hardship plan.
- Home Ownership: Homeownership status provided by the borrower.
- Loan Status: Current status of the loan (fully paid, current, charged off, etc.).
- Application Type: Indicates whether the loan is an individual application or a joint application with two co-borrowers.
- Issue Date: The month and year the loan was funded (May-2015, etc.).

3 Data Acquisition and Processing

3.1 Uploading the Data to GitHub

The data will be downloaded from the Kaggle dataset and extracted using 7-Zip. The data is too large to import into GitHub, so most of the data will be stored locally. Data will be processed into three separate datasets and imported into GitHub if the set is small enough. Attached in the GitHub repository will be a Jupyter Notebook file showing how the raw data was processed into the three tables.

3.2 Trimming and Cleaning the Data

All files will be inspected for missing values. Columns will be dropped with over thirty percent missing values. After trimming the columns down, columns in the rejected table will be matched with corresponding columns in the accepted table and placed into a separate data frame. Matching columns will have their names replaced with a single name to identify the match. Columns that have unclear names will be edited. A column labeling the applicant's result will be added (0 for rejected, 1 for accepted). Another dataset will be made including all of the zip code data and the applicant results. Finally, a table will be made that includes accepted applicant data.

3.3 Imputing and Cleaning

Missing values in the data will be imputed based on probability, mean, median, or mode techniques. Before each column is imputed, the data type of that column will be checked and changed if it does not appropriately match the observations. All columns appear to be tidy, so there is no need to alter column values besides for imputation.

3.4 Using the Data

The processed data will be used in three separate ways. The first path takes only the accepted applicant information and looks to create visualizations about those observations. The second uses accepted and rejected candidates to determine differences between the two classes. Visualizations will be made to compare rejected and accepted applicants. Summary statistics will be included describing every column. The third procedure will take the combined data frame and build a supervised machine-learning model. The model will attempt to predict the acceptance or rejection of a candidate. The modeling will be done with an 80-20 split of training and testing data.

4 Research Questions and Methodology

1. Is employment length related to the acceptance of loans? To answer this, a grouped bar plot will be made, with each employment length category representing two bars. The matched dataset will be used. One bar will represent the count of accepted observations, and the other will be the count of rejected observations. The two groups (rejected and accepted) will have different colors. (2.5 hours)
2. Is there a tendency for specific loans to be given during certain months? To answer this, a rose plot will be constructed using the loan issue month column. The accepted dataset will be used. The legend will represent the top five most frequent loans. A few of the

loan titles are written responses. The written responses will have to be handled through NLP or other means. The top five will only be used because there are many different reasons for someone to take out a loan. Each loan will have a unique color. (2.5 hours)

3. What is the frequency of accepted and rejected loans throughout the United States? To answer this, a map using geospatial data from the U.S. will be plotted and matched with the first three digits of the applicant's zip code. The zip code dataset will be used. A map using the states can be built, but the main visual will use the zip code. One graph will represent the percentage of accepted applicants. The percentage will be used and mapped to a specific shade of color, with a higher percentage representing a darker gradient. A legend will denote the color intensity and have a number line to match the percent and shade. (3.5 hours)

4. Does the ratio of the amount funded over the amount requested change based on hardship flag, home ownership, and application type? To answer this, I will build a tree diagram that branches based on the options of the three columns. The accepted dataset will be used. Each level of the tree will be colored differently. The final leaf in each branch will represent the average ratio of all the candidates that fall to that leaf. If the tree becomes too complex visually, a table will be made to display the different branches. (3 hours)

5. Does an applicant's FICO score change the interest rate they receive? What about the funding-requested ratio? To answer this, an area chart will measure the count of applicants and split them by their FICO score grouping. The x-axis will be either the interest rate or the funding-requested ratio. The accepted dataset will be used. There will be two graphs, one for each x-axis variable. The FICO score groups will each have a specific shade of color, and a legend will specify the shade each score category belongs. The issue date will be used to select a range. If possible, a website will check U.S. economic rates and compare them to the interest rates in the data. (3 hours)

6. Is there a bound to the Debt-To-Income (DTI) ratio so customers below a limit cannot be accepted? To answer this, a histogram will represent the DTI ratio of both rejected and accepted applicants. The matched dataset will be used. The y-axis will represent the count of applicants, and the x-axis will show the DTI ratio. If there is a noticeable difference, it may indicate a minimum acceptance. A legend will separate rejected and accepted applicants by a distinct color. (1.5 hours)

7. Using the matching features between accepted and rejected data, can the data effectively predict whether a candidate will be accepted or rejected? Three supervised machine learning models will be used to find the accuracy, precision, and recall. The matched dataset will be used. With accepted being labeled ones and rejected zeroes, the goal is to identify the most applicants for acceptance while minimizing the false positives. The three models are KNN, Decision Tree, and Logistic Regression. A third ensemble method will use hard voting to classify based on the decisions of the three models. (7 hours)

5 Work plan

Week 4 (2/6 - 2/12):

- Data tidying and recoding (6 hours).
- Question 6 (1.5 hours).

Week 5 (2/13 - 2/19):

- Question 3 (3.5 hours).
- Question 1 (2.5 hours).
- Question 2 (2.5 hours).

Week 6 (2/20 - 2/26):

- Question 4 (3 hours).
- Question 5 (3 hours).
- Begin to outline the presentation, and assess the status of the project (1 hour).

Week 7 (2/27 - 3/5):

- Question 7 (7 hours).

Week 8 (3/6 - 3/12): *Presentations in class on Thurs 3/9.*

- Presentation prep and practice (4 hours).
- Presentation peer review (1.5 hours).
- Assess the status of the project. If time allows, add an ethical aspect about Zip Codes. (1.5 hours).
- Begin drafting blog (0.5 hours).

Week 9 (3/20 - 3/26):

- Clean up code, add any needed comments, and make sure the GitHub repository looks professional (3 hours).
- Poster prep (4 hours).

Week 10 (3/27 - 4/2): *Poster Draft 1 due Monday 3/27. Peer feedback due Thursday 3/30.*

- Continue writing the report (2.5 hours).
- Peer feedback (2.5 hours).
- Poster revisions (2 hours).

Week 11 (4/3 - 4/9): *Poster Draft 2 due Monday 4/3. Final Poster due Saturday 4/8.*

- Double-check the organization of files in the GitHub repository (1 hour).

- Research the possibility of embedding animations of graphs inside blog posts (2 hours).
- Poster revisions (4 hours).

Week 12 (4/10 - 4/16):

- Continue drafting the blog post (7 hours)

Week 13 (4/17 - 4/23):

- Continue drafting the blog post (7 hours).

Week 14 (4/24 - 4/30): *Blog post draft 1 due Monday 4/24. Peer feedback is due Thursday 4/27. Blog post draft 2 due Sunday 4/30.*

- Peer feedback (2.5 hours)
- Blog post revisions (2 hours)

Week 15 (5/1 - 5/7): *Final blog post due Tuesday 5/2.*

- Final presentation prep and practice (7 hours).

Final Exam Week (5/8): *Final Presentations during final exam slot, Monday May 9th 3:20-6:40pm.*

References

Chan, Jonathan. 2018. "Lending Club Data Dictionary." Kaggle. 2018. https://www.kaggle.com/datasets/jonchan2003/lending-club-data-dictionary?select=Lending+Club+Data+Dictionary+Approved.csv.

George, Nathan. 2019. "All Lending Club Loan Data." Kaggle. 2019. https://www.kaggle.com/datasets/wordsforthewise/lending-club?select=accepted_2007_to_2018Q4.csv.gz.