Current Economic Analysis and Predicting Loan Defaults

Fairfield University Capstone Final Report

Enadi Pasholli

Summer 2022

Faculty Advisor: Dr. Yaqin Sun

Subject Matter Expert: Patryk Krakowski

Table of Contents

Project Background and Inspiration……………………………………………………………………………………………2

Project Executive Summary ……………………………………………………………………………………………………….3

Problem Statement……………………………………………………………………………………………………………………4

Analytical Pipeline: CRISP DM ……………………………………………………………………………………………………4

Next Steps ……………………………………………………………………………………………………………………………….13

Lessons Learned ………………………………………………………………………………………………………………………14

Final Signoff …………………………………………………………………………………………………………………………….15

Appendix / Linked Files ……………………………………………………………………………………………………………16

I. Project Background and Inspiration

Today’s economic environment is a much different one than we have seen in years before. In 2020 the United States shut down the economy in hopes to reduce the spread of the Covid-19 pandemic which has had rippling effects through our economy through interrupted supply chains, changing labor market, a large influx of cash in the system and many other factors. Today the economy sits at a crossroads between high prices and a looming recession. Gasoline, housing, food, and used cars are at all time high prices in contrast with stagnant wages. In these times how is the US economy going to adapt to the higher prices or will we go through a deflationary period? What industries are going to see price reductions? These are all questions I would love to answer given more time, but to start I will use the housing market for this analysis.

Background of the Housing Market:

In 2008 the US experienced the largest shock to the banking system since the great depression. Banks were giving out mortgages to people who did not qualify for mortgages of the size they were offering. These loans were called sub-prime mortgages. This additional population that was now eligible to enter the housing market increased demand for homes and thus drove prices higher. The way that banks would justify their actions was to offer ARM or adjustable-rate mortgages to these borrowers at a low entry rate to fit in the borrower’s budget and increase over time. As home prices increased so did interest rates from the FED to cool off the spending. Once the interest rates began to climb, the ARM mortgages began to see an increase in their interest rates and borrowers on the brink of affordability, could no. longer make their monthly dues, and thus defaulted. Banks that owned a large amount of these subprime mortgages either got bailed out, bought out, or went bankrupt.

Today the landscape for mortgages is much different. From the past event, banks are now required to hold certain cash for every dollar they loan out and have higher quality loans. The average credit score for a mortgage in 2008 was under 680 while today it is about 750. The other main difference today is that in 2008 high real estate prices were driven by additional demand offered to people who should not afford the prices of the time while today there is a shortage of homes in the market driven by material and labor shortages. Since new homes are not being built, or being built with expensive materials, many people who decide to move see a benefit of a higher sale price which the higher purchase price may even out, but it is not a good market for a new home buyer. Many people who may have saved for a house might be priced out of the market which lowers demand, yet the average home listing is taken down in about 2 weeks. There is still a lot of demand for the current supply. Another issue that is suppressing supply is that most family’s sell their home to upgrade, and with these higher prices many families would rather wait than buy a new home which again reduces the supply since they would not be willing to sell their homes.

II. Project Executive Summary

This project is designed to look at all the economic indicators available on FRED that pertain to the US economy and housing market. The data shows historical data and how it aligns with previous recessions and helps us determine is we should expect a deflationary period in the housing market. The second portion of the project uses a dataset that contains different loan accounts to use a machine learning model to determine what attributes most likely predict an account which might default.

**Problem**: The US economy is headed for a recession which is usually followed by rising unemployment, lost wages, and loan defaults. Banks loan out money in hopes to later receive it back over time with interest. If an entity defaults on a loan, then the bank is out the money that the borrower has not paid back. The bank will usually seize any collateral for the loan and sell it to recoup for its losses. This puts both the borrower and the bank at risk during a recession.

**Methods**: To best understand today’s situation, I have created a tableau dashboard to go over all the important economic indicators, what they mean, and what could happen going forward. The second portion of this project is a machine learning model on over 1 million loans to determine what factors best predict a loan default.

**Data**: The data for the economic indicators originated from the Federal Reserve Economic Data website, or FRED. This data is reliable and available to the public for any use. The second dataset is a dataset of 2.2 million loans given out by the lending club between 2007 and 2015. These loans can be used for anything from funding a wedding to buying a home. The data has over 150 features from information about the loan such as amount and interest rate and information about the borrower such as previous defaults through a credit check. The target variable is ‘loan\_status’ which I made binary to indicate if the loan was defaulted or not. There were other variables in the target such as delinquent, but for a banks bottom line, a default is more important and more serious for a borrower and institution.

**Results**: Based on a few models I was able to run, the naïve bayes model was able to be 87% accurate while the decision tree model was able to predict with 99% accuracy. Logistical Regression did worse than the naive bayes with only 85% accuracy. Many of the models chosen to use for the project have taken a much longer time than my computer is capable of handling to provide results. Eventually with more allotted time and a better computer results for XGboost and hyper tuning model TPOT will be able to be assessed.

III. Problem Statement

Banks are businesses that are closely tied with the government and the US economy. Banks take money from the government and are regulated by law to fulfill their duty to protect the US and other local economies. This money is then given to borrowers so that the bank can then make money from the interest on the loan over time. For a bank to be profitable that means that banks must give out loans that they can be sure that they will be paid back. When borrowers do not pay their loans back, banks lose money and borrowers usually lose assets. It is important for banks to know who to give loans to protect themselves and borrowers in the worst-case event. As the economy heads for trouble and peoples incomes are put at stake, banks should know what loans to keep and what loans to sell to other institutions to protect themselves from loan defaults.

Main Analytical Questions:  
1. What is the current economic standing compared to previous recessions?

2. What factors most contribute to predicting loan default?

3. How accurately can we predict who is going to default on a loan?

IV. Analytical Pipeline: CRISP DM



1. Business Understanding

Lend Card is a peer-to-peer loan solution that allows investors to directly lend money to others for a higher return while the borrowers can on average pay less in interest than a traditional bank. Like a bank though, Lend Card does not want their partners to default on loans. A loan default would lose money for its bottom line and credibility to its investors.

As the economy gets worse and businesses begin to consider extending layoffs through the company, many people are at risk of losing their income and defaulting on their loans. Banks do not want to lose money in these situations but are also understanding of their customers problems. Local banks, like the one my SME works at are interested at solving their customers problems by offering payment deferments or interest only payments until the customer hardship is over. Other institutions on the other hand may sell their risky loans to other banks or investors, or just simply not approve their loans. Traditionally, a community bank would does not have the resources to invest in a data department so uses a general application to assess a applicant’s credit worthiness using data supplied by the applicant or from a credit bureau. Having a tool that can generally understand a customer’s outcome before approving a loan is helpful for banks, credit unions, or other lending platforms to increase profitability and losses on loans.

Although many loans are attached to some form or collateral that banks will take if the borrower does not pay back the money the bank that has not been repaid, the institution is still at risk of holding assets that are worth less than the loan and other fees incurred while seizing the assets. Investing the money in creditworthy and reliable borrowers is more profitable than the majority of defaulted accounts.

1. Data Understanding

The data for this project came from Kaggle and FRED. Most of my understanding and interpretation from the FRED data comes from prior experience in finance classes and outside research. Data collected by the Federal reserve such as GDP, interest rates, unemployment, and many others are all economic indicators that help us understand and benchmark the health of the US economy.

The other dataset came from a peer-to-peer lending service that is named Lend Club. Lend Club posted a large dataset to Kaggle for others to build models from. This dataset came with a data dictionary which helped determine what each feature was and its importance to the project. A link to the data features will be linked at the end of the project.

1. Data Preparation

**Fred Data**: The data that comes from the Federal Reserve cannot be downloaded in bulk or be found up to date online without requiring some form of payment. To get around this, I downloaded all the important economic indicators separately and joined them based on date using excel. I had downloaded 16 indicators to build the Tableau dashboard component of this project. There was no need to impute or make any adjustments to the data since it was not used to do any modeling. The purpose of this component to the project is to tell a data story about the economy and why looking forward and prepare for worst case scenarios for lending institutions is important to maintain company health.

**Lend Club Data**: The Lend Club dataset is a large dataset that included over 2.2 million rows and 145 features. My computer could not handle the full dataset, so the first thing I had to do was remove just under half the rows of the original dataset to about 1 million rows. This was important to get the project done but not necessary since others may have different tools than me. The data was filtered such that each column kept as much data as possible, eliminating the instances with the most null values first.

**Machine Learning Data Operations Flow Chart**

**[](https://www.figma.com/file/AE8FDccUzzK2UJEJchk4vu/Untitled?node-id=2%3A126)**

(Click to view larger file)

**Data Quality Checks**

In the data audit report, the most important thing to look at was how many missing values each column had in comparison to the total number of rows there are in the data frame. There were many columns with over 95% missing data which could not be relevant to the project, so they were dropped. Having so many features to work with gave me the luxury of being able to set a rule that only columns with less than 50% missing data would be kept.

After this step I investigated what featured had too many options to bin for categorical or were duplicates of some other column. One feature for example, ‘funded’ had the same data as the ‘loan\_amount’ feature and ‘funded\_amnt\_inv’ feature. The other feature was Zip Code, which contained many different unique variables but are each contained in a state. This data contained another feature named ‘state’ which could be binned by region later in the project. Due to this relationship between the two features, I chose to drop zip code and only work with states. Lastly, the feature titled ‘next\_pymnt\_d’ contains information about the target so it needed to be dropped. If the account does not have a next payment date it means 1 of 2 things, either the account has been paid off, therefore there is no next payment date, or the account has been charged off and the bank expects no next payment. These are both part of the target and cannot be used to determine the probability of loan default.

After dropping the features, the new dataset consisted of 1,005,562 instances and 96 features.

**Exploratory Data Analysis**

The EDA portion of the project is to get a better understanding of the data and its characteristics. To understand the data correctly, I first split the EDA into 2 important categories, continuous and categorical data frames. The continuous data can be transformed into box plots and a correlation matrix.

Scatter chart

Description automatically generated with medium confidenceDiagram

Description automatically generated with medium confidence

From the analysis visualized above, we can see that much of the data is skewed. Although there are more data points that are included in the full data report, the general trend for the data in this project is very skewed or has distinguished values. An example of this is number of defaults in the past year, which is a feature in the dataset. The box plot for this feature has designated values at 0,1,2,3,4,5,6. Overall, the data has ranges and values that are expected for the data that is included in the dataset. The only outstanding feature is ‘il\_util’ which has an outlier at 999. This value may represent some other issue such as missing data. Unfortunately, this information cannot be found in the data dictionary that came with the file so we can leave it for outlier handling later in the project. For a more in-depth analysis of all the box plots, please refer to the data audit colab sheet linked at the bottom of this report.

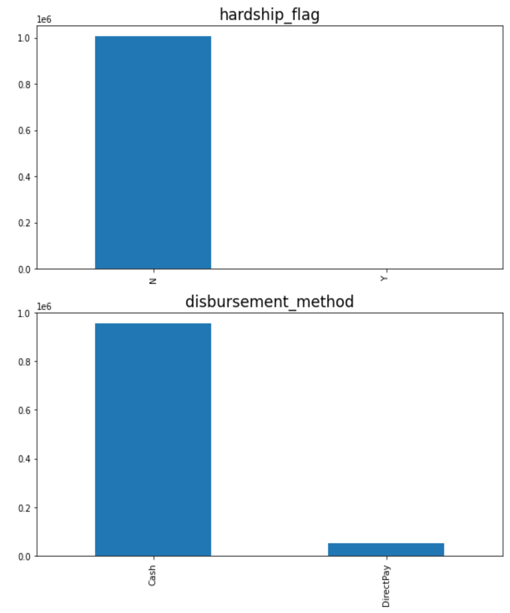
The categorical features can be made into bar graphs showing the most important variables in the feature by sorting the top 10 most frequent cases. The most important feature to visualize is the target. In this case our target is named ‘loan\_status’ which has multiple categories other than just default.

**Chart, histogram

Description automatically generated**

In our target variable there are 3 main categories that are relevant. First is current, then fully paid, and lastly charge off. Current means that the loan is active and does not have any late payments. Fully paid means the account is closed and no longer requires payments since the balance was paid off. Charge off is the equivalent to a default, but the main difference being that the bank no longer is expecting payment for the account and has written it off of the books. The data for the target is skewed where just around 10% of the values are charge off values. This will later require smote in order to balance the data out. This default rate is much higher than the national average over the past decade which has trended downward from a high of 10% in 2012 to 3% in 2022. This is presumably due to the fact that Lend Card has a business model that helps out those with less than great credit histories.

Chart, waterfall chart

Description automatically generatedChart, bar chart, histogram

Description automatically generated

From the categorical data everything is looking normal except for these features shown above. Last payment date and last credit pulled both are skewed heavily for 2019. This can mean either that the current opened account is the most recent account for the borrower or that there is something wrong with the data. In either situation, the most recent account being sought by the borrower being the Lend Card loan would be suitable but probably not very useful for the machine learning. These will be binned later. The existing binary categorical features are also quite skewed from the dataset and again may not be of use for the feature selector.

Interesting concepts to think about for the last set of graphs is that the majority of the loans given by Lend Card are debt consolidation loans and most of the users of the platform come from states with large cities. This is to be expected since banks charge high interest to consolidate debt since there is generally no collateral and the main value for Lend Card is a lower interest rate than a traditional bank. It is nice to see that the data reflects a good use case for the product they offer.

**Split the Data**

After the visual analysis of the data and understanding its workable shortcomings, the next steps are to split the data further and impute the data so that there are no missing values.

Graphical user interface, text, application, chat or text message

Description automatically generatedThe first step to splitting the data is to remove the target variable from the data frame. After the target has been removed then both the main data frame and the target variable were run through the train\_test\_split function at a 75/25 split ratio since there are over 1 million rows. Once the main data frame is split into test and train data frames, these data frames are then further split into categorical and continuous data frames, creating a total of 6 working data frames. This is necessary for the imputation and one hot encoding in the next step.

**Imputation**

Once the data is split into the 6 data frames the next step is to impute the data accordingly to remove any missing values in each of the remaining features. The categorical data is imputed by using the mode and the continuous data is imputed by using the median. The reason the data was split into each of the separate data frames before imputing was to ensure that each data frame is using its own local data to generate medians and modes instead of the generalized results from the macro dataset.

**Binning and OHE**

Binning the data is for the categorical features only. The features that need to be binned are the ones that have more than 5 different unique values. The features in this case are loan grade, subgrade, employment title, employment length, loan issue date, purpose for the loan, state, first credit line date, last payment received, last credit pulled, and the target.

The basic assumption for grade is that anything below a grade of D are all applicants that are very risky. Very risky debt in bond terms are called junk bonds, and in the same spirit any loan with a grade of d or less is put into a ‘junk’ bin. Subgrade is an extension of the letter grade broken up into 5 separate categories numbered A1, A2 ,A3 etc. From this I binned these into ranks 1 -5 since they were ranked through 5 for each letter grade.

Employment title was binned by similar jobs for example white collar jobs, blue collar jobs, medical professions and business backgrounds like C suite employees or business owners. Since there are many ways to spell out different professions, if the name was not in the top 100 unique values, then it was binned into ‘other’. The top 100 unique values for this feature contained just about 90% of the unique data. Employment length was binned based on phases of a person’s life. Assuming someone starts their first job at 18, then the first 2 years are generally not very lucrative, and a person steadily gains wealth from 2-8 years. The next phase would be marriage and having children, so it is binned from 9-18 years. After that phase a person is looking to send kids to school or college and is binned from 19-30 years. After 30 years of employment a person is looking to settle down, so it is binned to 30+ years.

Loan issue date was only in for 2019, 2018, and 2016 so it was binned by year.

Lend Card provides loans for many reasons so the most general terms to get a loan from them is for loan or credit consolidation, home purchase or renovation, life purchase like wedding or vacation, and lasty business loans.

State was binned by region and the last few variables were dates and binned by decades.

The target variable is an important one to bin and make binary. The binning for the target composed of using default and charge off values, setting them to 1 and making the rest 0.

**Outliers, Scaling and Normalization**

To do outlier handling, scaling, and normalization I created functions to apply to the continuous data frames. First, I did normalization, followed by outliers and then normalization.

Chart, histogram

Description automatically generated

Although there are still some outliers left over, the process mostly took care of all the issues in the data. Everything is scaled between 0 and 1 and almost all the features are balanced correctly with a few being slightly skewed.

Text, application, chat or text message

Description automatically generated**Merging Data Frames**

Once all of these steps to clean the data have been completed and the data reflects the changes, then the data frames all get merged back together to make 2 data frames and sent to the next step to have features selected and ranked. The data frames that are merged together are the test categorical, test continuous, and the test target which is called X\_test\_complete for the purposes of this project. The other data frame is the X\_train\_complete data frame which is composed of all the separate train labeled datasets.

**Feature Selection**

Following the step of merging the data frames back together, those are then saved as CSV files and moved to the last step of data processing which is feature selection. The feature selection process has 2 components to it, which is choosing the most important features to run models with but also ranks them to which ones are the most important. From the current way the data has been cleaned, the selected features with importance are pictured below. The model used to choose the features is the random forest classifier model.

Text, letter

Description automatically generated with medium confidence

These selected features were then used to create a new filtered data frame from each of the test and train data frames. I decided to also include the target in each of these data frames to split them out later before running the models on different colab sheets. Again, we are left with 2 data frames, a test and train data frame respectively for running the different models.

1. Modeling

The modeling for this project was done all on python using the different options we have learned from class. I used 5 different models to assess the performance of these selected features. The first model I used was the naïve bayes model to find out how well a model can figure out a charge off from simple guessing. The results set the benchmark to define a useful predictive model. The other models used to try and do better than the naïve bayes are logistical regression, decision tree, random forest classifier and XGBoost. The hyper tuning models used were TPOT and Random Forest Grid Search.

1. Evaluation

The evaluation of these models was based on F1 score and ROC\_AUC score. F1 is the measure of the model’s accuracy while ROC AUC is a measure of the usefulness/performance of the model.

Table

Description automatically generated

Based on the results given by each model, we can see that each mode does a very good job predicting the default outcome. These numbers all seem too high to trust the model and the data going through the system and may need some further evaluation to determine if the information about the target is slipping into the selected features. Overall, it is a good sign that logistical regression and decision tree are both doing better than the naïve bayes model.

1. Deployment

Although this model is not being deployed out into any real-world application it is worth the effort to create something that is probably already being used by many organizations across the globe. The data used to create these models are limited in scope due to data avaiibility but in real world practice for a larger organization I am confident that more data and more computing power would be able to solve for these factors with more granular detail and confidence.

V. Next Steps

During the course of this project and the limited amount of time I have to complete this, there are still a few other touches I want to go back and complete. The first thing is wait for the other models to finish running. These models take days to complete and my computer times out before it gets close to finishing. The next thing is to go back and double check all the features being sent into the feature selector and make sure that the target information is not seeking into the models. For outliers, normalization, and scaling, I would like to go back and rerun every possible order for those variables to figure out the best outcome for the selected features and verify there are no outliers in any of the features. During the data scaling and outlier handling there were some features that came back with having no range. In hindsight going over the project, I realized that these features should have been labeled as categorical features and have to go back and bin them accordingly instead of dropping them. Lastly, I would like to reconsider my binning and make sure that they all function correctly as well as are binned by similar attributes. The binning that I have done is functional, but I think it can be made better and more accurate given more time.

VI. Lessons Learned

From this project the lessons I have learned are invaluable for data analysis. Using tableau to tell a data story instead of investigative deep dives has taught me how to explain what is happening without much complexity with the tools themselves. It has grown me as a storyteller and the ability to explain why a specific topic matter. I have also learned how to source data from the internet and use it to the best of my ability even in adversity. Redfin has public tableaus that I had to import into my tableau and figure out how to reconstruct their data to give me the information I needed. I was not able to export the data from the tableau, but it was a wonderful lesson to learn how to be able to get it onto my desktop and find the gatekeepers of the data.

From the machine learning component of the project, I have learned how much time it takes to clean data and deal with the shortcomings of a desktop pc. My computer through the process has not been able to keep up with half of the dataset posted by Lend Card. Feature selection took almost 11 hours to complete and some of the models I could not get results to due to the amount of time it takes to get the results. My computer keeps timing out when I let my computer run overnight because I am not there to be active on the colab sheets. The other lesson that I learned is to save CSVs to your drive periodically through the data cleaning process. This was important for multiple reasons besides the reason that the RAM on my computer could not do many tasks at once. Being able to export the files built a nice paper trail that I can post to GitHub or any other link to express what I did at each step. Being able to pick up at different points in the sheet was invaluable in contrast of having to rerun hundreds of lines of code to pick up at the end of the sheet. On the same note, importing all the packages at the beginning of each section drastically reduced the time it took to pick up at different spots in the colab sheet instead of having to find each package and rerun them. Overall, this project enabled me to work on my own workflow and find the most optimal solutions to be as efficient as I can in editing my mistakes or overthought.

VII. Final Signoff

With his selected project subject, Enadi thoroughly demonstrated his interest in both finance and business analytics while also maintaining an ability to process and interpret massive amounts of data. His knowledge in the financial ecosystem allowed him to use the tools learned in his business analytics program to further expand the questions and explorations he attempted to answer throughout. I believe Enadi has met all of requirements for the completions of his capstone and I approve is project.

SME: Patryk Krakowski

Appendix / Linked Files

Kaggle Dataset:

<https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv?select=loan.csv>

Federal Reserve Indicators:

<https://fred.stlouisfed.org/>

Final Colab Sheet:

<https://colab.research.google.com/drive/1yw1mOEN9N_h9FVlQ2eqHfMhXmXMobR4S#scrollTo=SWgZ1DVzfnCG>

Data Dictionary:

<https://colab.research.google.com/drive/1KO74lIiWs0sFHxz1y7BdkJJSYHYRllNe#scrollTo=DMD-p6DCBcwY>

Data Audit Report:

<https://colab.research.google.com/drive/1JMAdXOMZ55Z8Rw0p9CHc2LVla5x14fCx#scrollTo=eBS2HJgYF1rS>

Model Results:

<https://colab.research.google.com/drive/1PF-NWGa7JioTeSks3L2bFC5BaGZmPHE_#scrollTo=DMD-p6DCBcwY>

GitHub repository:

<https://github.com/EnadiP/LoanDefaults>

Flow Chart:

<https://www.figma.com/file/AE8FDccUzzK2UJEJchk4vu/Untitled?node-id=2%3A126>