Credit Risk Assessment with Textual Analysis

Zachary Haroian

September 19, 2020

Abstract

Peer-to-peer (P2P) lending services, like LendingClub, are growing as an alternative to those who cannot or choose not to borrow from traditional institutions such as banks. Individuals who act as lenders fund a borrower and are self-deterministic in who they fund. It is imperative that this new class of lenders has the tools to accurately assess borrowers and their likelihood of defaulting. We tested to see if the model will see a statistically significant improvement in performance when the textual data are used in conjunction with financial data to predict loan default. A baseline model was created using Logistic Regression. The sentiment was extracted using the Bag-of-Words method with the Loughran-McDonald Word Sentiment list. This data was added to the original features and a new model was compared against the baseline model. We found that there is a significant difference when adding the sentiment analysis to the pre-existing model, given enough data. This agrees with current literature surrounding textual analysis and credit risk, though more advanced sentiment analysis may need to be used to glean a greater performance improvement.

A thesis presented for the degree of
Bachelor of Science in Data Science
Willamette University
United States

Contents

List of Figures

1 Introduction

Personal loans are one of the fastest-growing sources of debt within the United States, tripling from $49 billion in 2010 to an all-time high of $162 billion in Q1 2020 and is steadily growing 10% every year [13]. There are 20.9 million consumers with personal loans, and the average debt per borrower is $9,025. The Borrower-Level Delinquency Rate is at 3.39%, higher than Credit Card Delinquency Rate (1.97%), and more than double the Delinquency Rate of Mortgages (1.40%) and Auto Loans (1.37%). More than 65% of borrowers take out a personal loan to consolidate their debt or refinance credit cards [11]. But, many borrowers are unable to obtain financial help through traditional loan avenues due to poor or nonexistent credit history.

Financial Technology (FinTech) innovators have created a solution for those who would not otherwise qualify for traditional loans. FinTech loans now comprise 38% of all unsecured personal loan balances, whereas five years ago it only accounted for 5% of balances [12]. FinTech lenders are often quicker and cheaper than traditional lenders because they use machine learning algorithms to screen borrowers rather than using humans to sift through applications.

Peer-to-Peer (P2P) lending is one of the most successful lending models implemented by FinTech. The advantage of P2P lending is that borrowers and lenders are matched together. Borrowers request a small personal loan, and multiple lenders evaluate and crowd-fund loans, rather than soliciting a small group of investors [3]. LendingClub is the world's largest P2P lender and was the first P2P lender to register with the Securities and Exchange Commission (SEC) [24]. In their loan application, they supply an optional text field for borrowers to describe their situation in their own words, in addition to their financial information.

It is imperative that lenders can accurately assess the risk associated with the borrowers they are lending to. Credit risk assessments have traditionally been based on debt-to-income ratio, credit history, and other financial information that may be indicative of a borrower's future actions [8, 6]. Error rates in predicting default are usually 20-30%, as historical data is extrapolated multiple years into the future, during which unforeseen events may arise and data is rarely linearly separable [4, 2]. Even a marginal improvement in the prediction of loan default might result in significant savings in the future. Soft data, such as textual data, when used in conjunction with hard data, like the consumer's borrowing history, could give a more accurate picture of the borrower's situation and creditworthiness.

Most research has been focused on credit-risk scoring with financial data using machine learning [8, 10, 4, 9, 15]. There has been a growing amount of research done on textual analysis and its place within the context of personal loans. Research has found that with textual analysis and extensive financial history, text can improve loan default prediction [21, 26]. With this in mind, our hypothesis is the following: If the sentiment and reading grade from the text are used in conjunction with financial data to predict loan default, then the model will see a statistically significant improvement in performance.

In this paper, we will explore the relationship between text and financial data in loan default prediction through LendingClub's data on P2P loans from 2007 to 2018. Starting with a data cleanse, we will take a deeper look at the existing relationships between the features in the dataset in the Exploratory Data Analysis (EDA). Then, a baseline analysis will be performed on the financial data exclusively using logistic regression to create a comparison when the textual data is included. Once we have a baseline, we will use the Loughran-McDonald Word Sentiment List, a corpus built specifically for analyzing financial documents, to extract sentiment from the text and calculate the reading level and compare the models. Finally, we will explore the results and discuss their impact on the analysis. Figure 1 displays a visual representation of the process used to analyze the data.

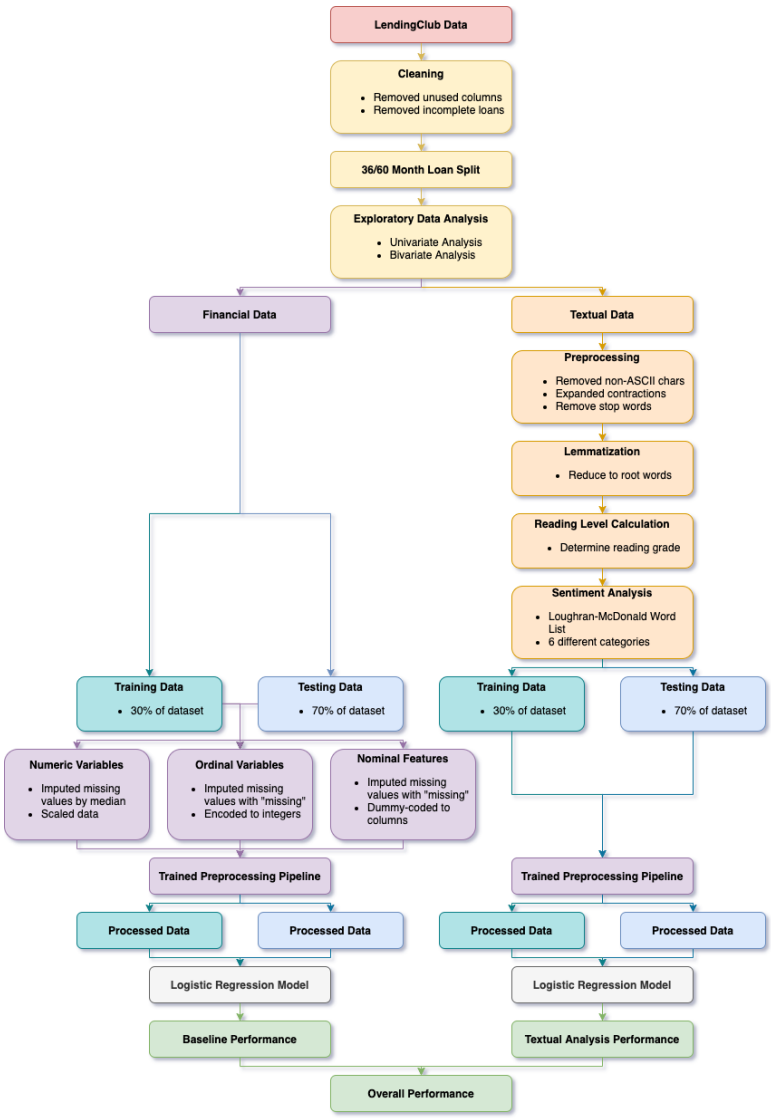


Figure 1: The complete path that data takes through the project.

2 Data Analysis

2.1 Data

2.1.1 Origin

LendingClub (LC) provides data on their consumer loans. The dataset consists of 295,891 loans funded between 2007 and 2018: 222,765 have 36-month long terms and 73,126 are 60-month terms. The dataset originally contained 33 columns, but only 12 columns were kept for various reasons. See Table 5 in the Appendix for a complete description of each feature that was kept and the reasons for removal.

2.1.2 Cleaning

Loans in progress were removed (1,306 loans across both 36 and 60 month terms). The data was split into two separate datasets by term, one for the 36-month loans and the other for the 60-month loans. Categories that contained ordinal data had their order entered manually, as with Sub-Grade, where A1 is higher than G5. See Table 5 in the Appendix for the specific processes for each feature.

2.2 Exploratory Data Analysis

Before creating the model, it is imperative to get an idea of what the data looks like and find any important relationships between features and any features that might be correlated with the loan defaulting. We will explore the numeric features first.

As is shown in Table 1, there are 4 missing values for Annual Income in the 36 month loans that will be imputed during preprocessing. The mean loan amount for 60 month loans, as seen in Table 2 is about $7,000 higher than the 36 month loans, while the FICO score distribution is incredibly similar, with the same mean and standard deviation. This implies that while borrowers are asking for more money, they are not necessarily more qualified to receive it. Interestingly, 60 month loans have a higher average annual income, but the maximum value for 36 month loans is nearly twice as high at $7,446,395.

Table 1: An unscaled description of the numeric features in the 36 month loans.

	Loan Amount	**Interest Rate**	**Install-ment**	**Annual Income**	**DTI**	**FICO Score**
Count	222,765	222,765	222,765	222,761	222,765	222,765
Mean	$12,089.11	12.84%	$405.54	$70,274.68	16.26%	700.14
SD	$7,359.05	3.86%	$249.76	$56,541.67	7.61%	31.78
Min	$500.00	5.42%	$15.67	$1,896.00	0.00%	612
25%	$6,625.00	9.91%	$221.52	$42,000.00	10.54%	677
50%	$10,000.00	12.92%	$343.63	$60,000.00	15.92%	692
75%	$15,850.00	15.31%	$526.99	$85,000.00	21.72%	717
Max	$35,000.00	25.99%	$1,409.99	$7,446,395	34.99%	847

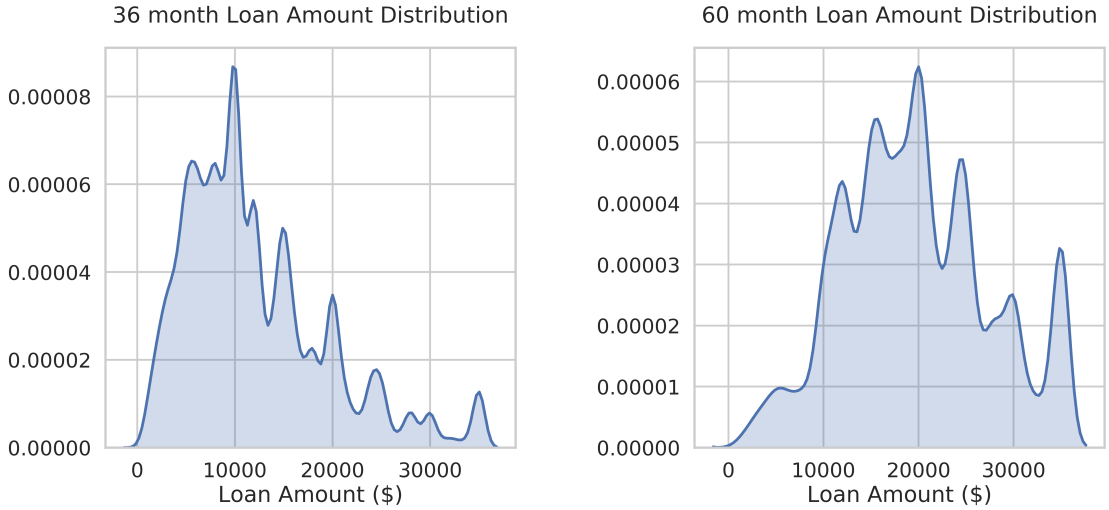
Table 2: An unscaled description of the numeric features in the 60 month loans.

	Loan Amount	**Interest Rate**	**Install-ment**	**Annual Income**	**DTI**	**FICO Score**
Count	73,126	73,126	73,126	73,126	73,126	73,126
Mean	$19,743.53	17.43%	$497.01	$79,155.17	17.65%	699.96
SD	$7,735.31	3.95%	$209.33	$48,073.52	7.39%	29.92
Min	$1,000.00	5.79%	$4.93	$4,800.00	0.00%	662
25%	$14,400.00	14.91%	$348.69	$52,000.00	12.27%	677
50%	$19,500.00	17.1%	$474.86	$70,000.00	17.55%	692
75%	$25,000.00	20.49%	$624.54	$93,600.00	22.95%	717
Max	$35,000.00	26.06%	$1,049.17	$3,900,000	34.99%	847

2.2.1 Univariate Analysis

In the Univariate Analysis we are looking at the distribution of each feature on its own. For quantitative variables, we compare the overall pattern of the data - the shape, center, and spread. For qualitative variables, we compare most frequent occurrences.

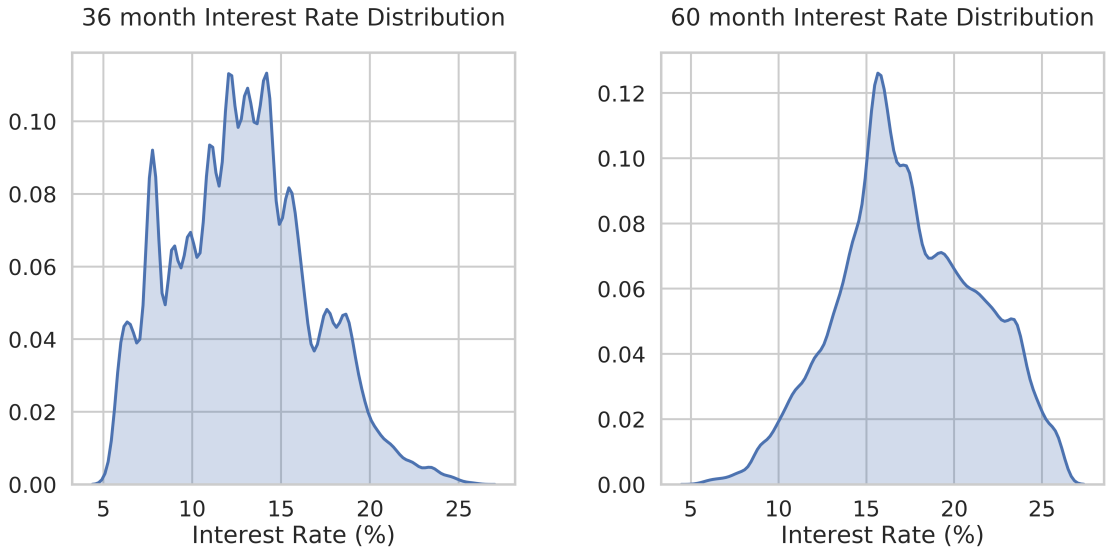
Loan Amount: As shown in Figure 2a, the 36 month loans are positively skewed with the center hovering around $10,000. Figure 2b shows the 60 month loans are closer to a normal distribution, with the center around $20,000 and close to symmetrical with a small peak around $35,000.



(a) 36 month loan amount distribution (b) 60 month loan amount distribution

Figure 2: Loan Amount Distribution

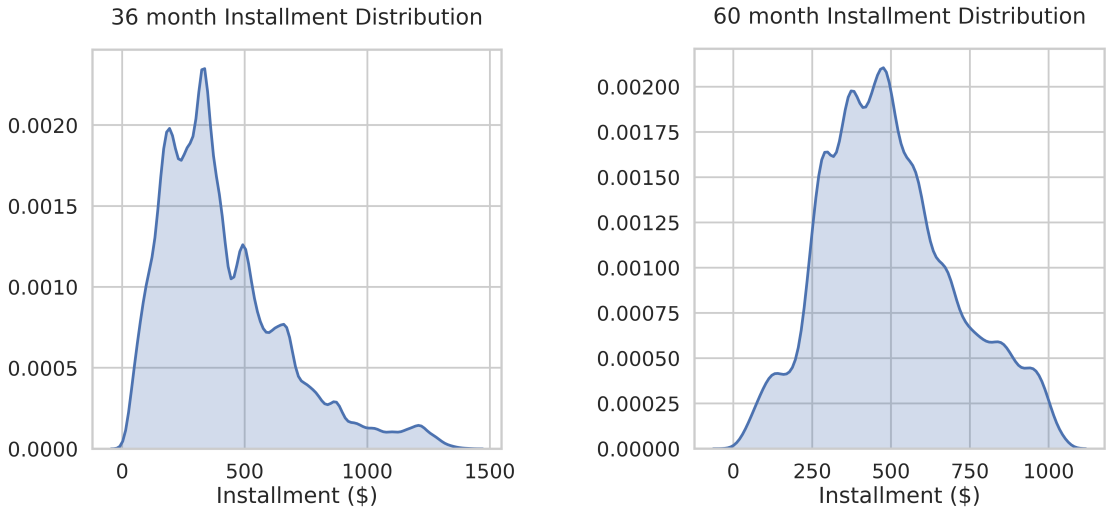
Interest Rate: Figure 3a displays that the 36 month loans have its center around 13%, making it slightly positively skewed and unimodal. It has a slightly wider spread than the 60 month loans, shown in Figure 3b. The 60 month loans are centered just above 15%, making it symmetrical and unimodal.



(a) 36 month interest rate distribution (b) 60 month interest rate distribution

Figure 3: Interest Rate Distribution

Installment: The 36 month loans are unimodal and heavily positively skewed, centered around $400 as shown in Figure 4a. As in Figure 4b, the 60 month loans are unimodal and closer to symmetrical, with its peak at $500.

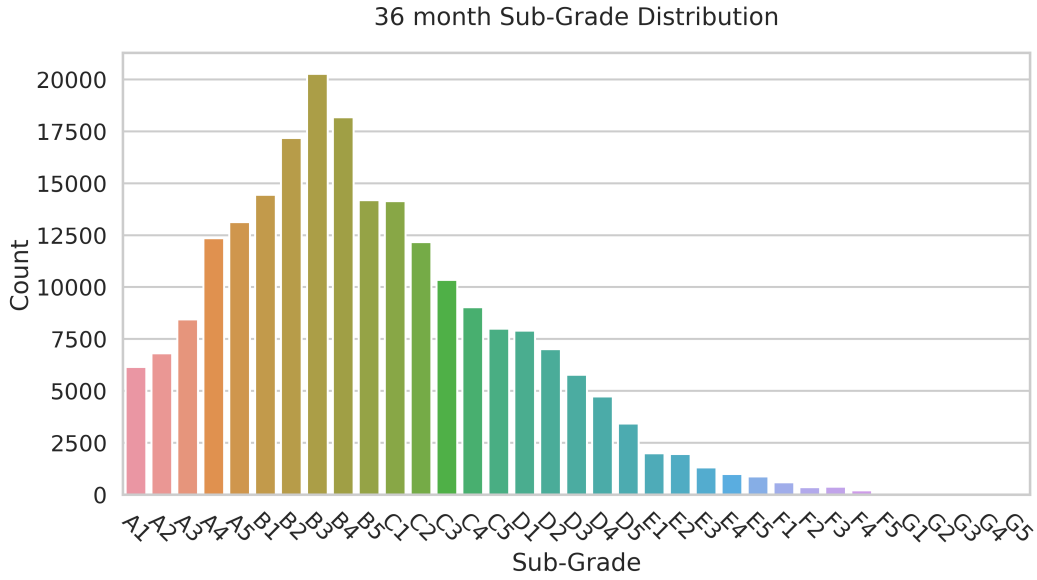


(a) 36 month installment distribution

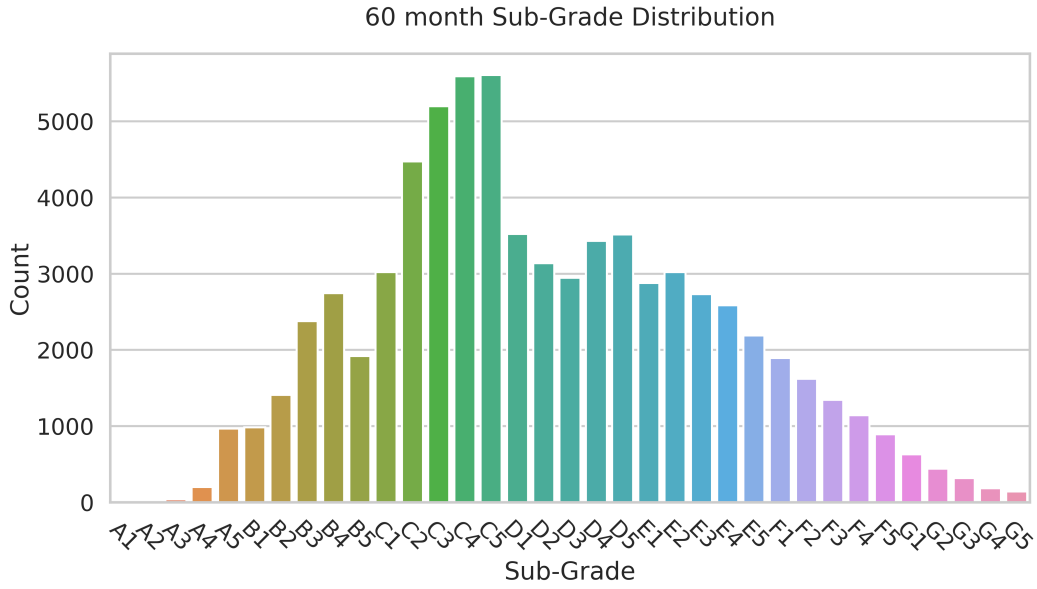
(b) 60 month installment distribution

Figure 4: Installment Distribution

Loan Sub-Grade: In Figure 5a, the most frequent Sub-Grade for 36 month loans is B3, whereas in Figure 5b for 60 month loans it is C4 and C5. We can also see that loan grades below F1 are rare in 36 month loans but are common in 60 month loans, and vice-versa for A grade loans.



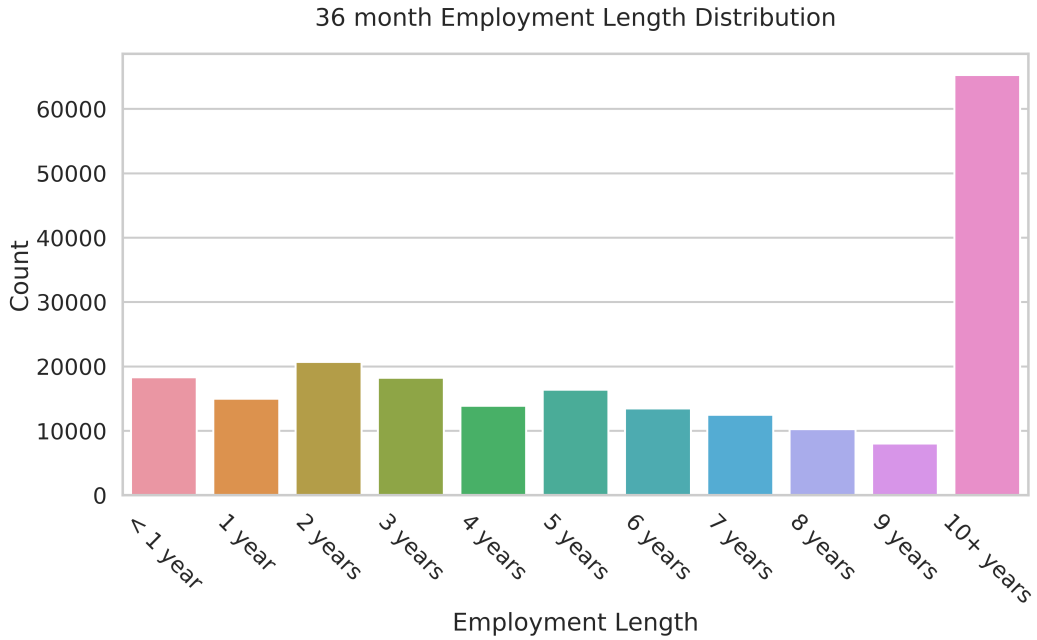
(a) 36 month sub-grade distribution



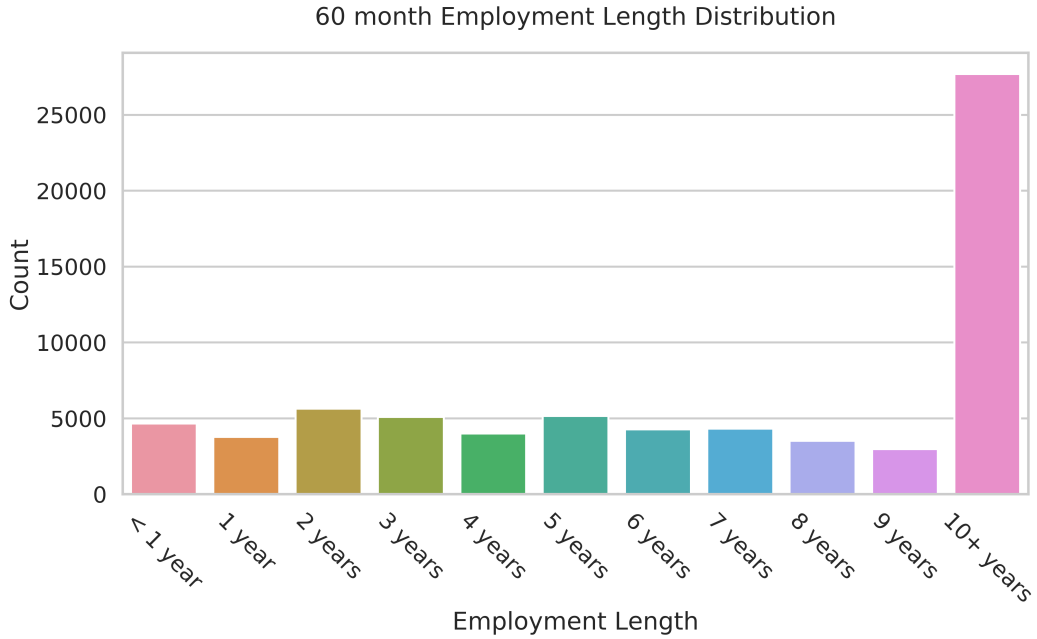
(b) 60 month sub-grade distribution

Figure 5: Sub-Grade Distribution

Employment Length: In Figures 6a and 6b, borrowers who have been employed for 10+ years is by far the most common for both terms.



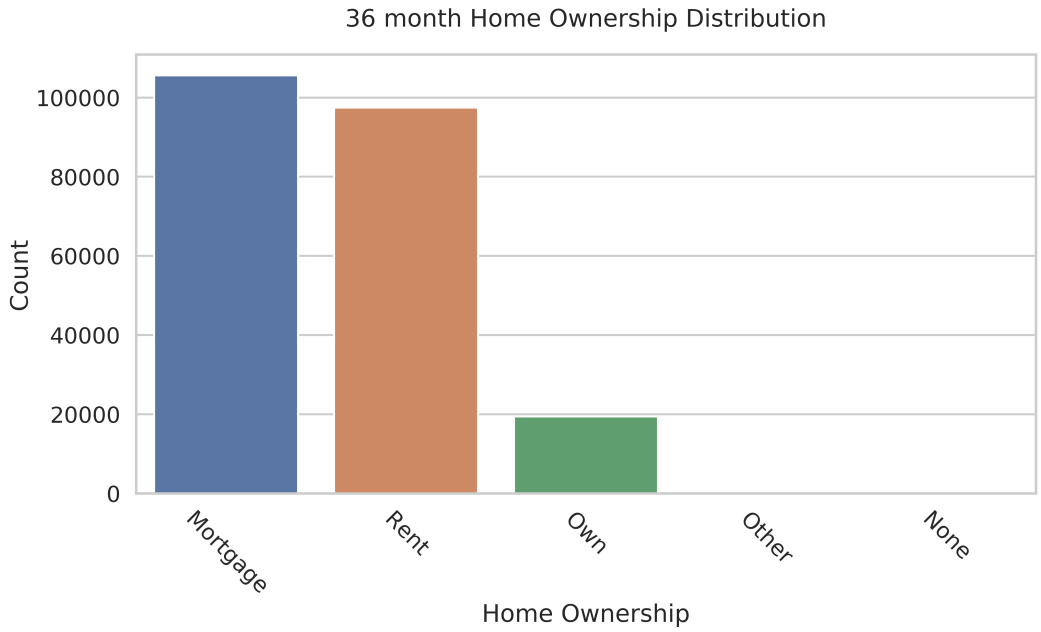
(a) 36 month employment length distribution



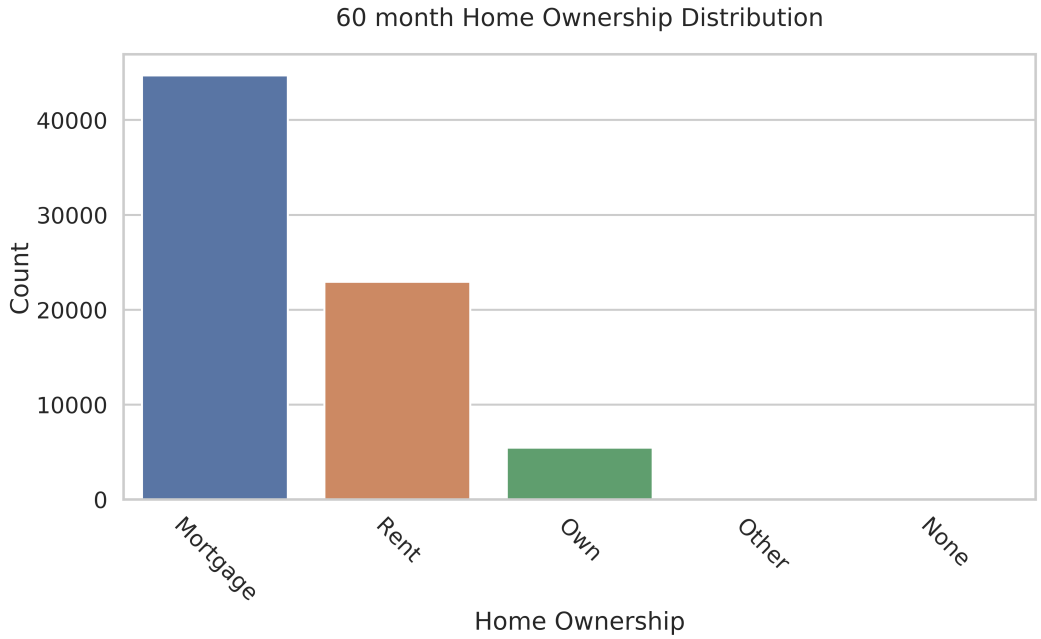
(b) 60 month employment length distribution

Figure 6: Employment Length Distribution

Home Ownership: As seen in Figures 7a and 7b, mortgage is the most frequent type of home ownership in both terms. Renting is more commonly found in 36 month loans compared to 60 month loans.



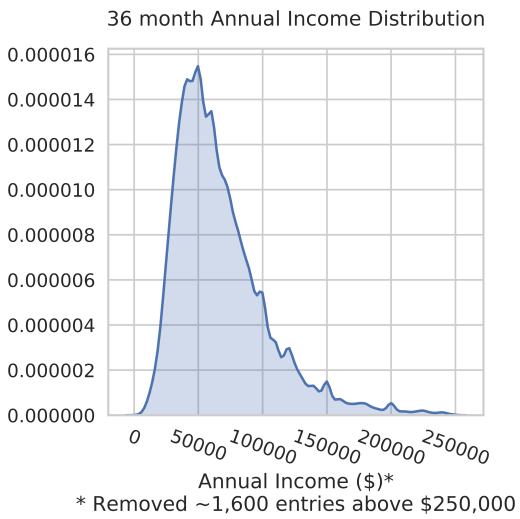
(a) 36 month home ownership distribution



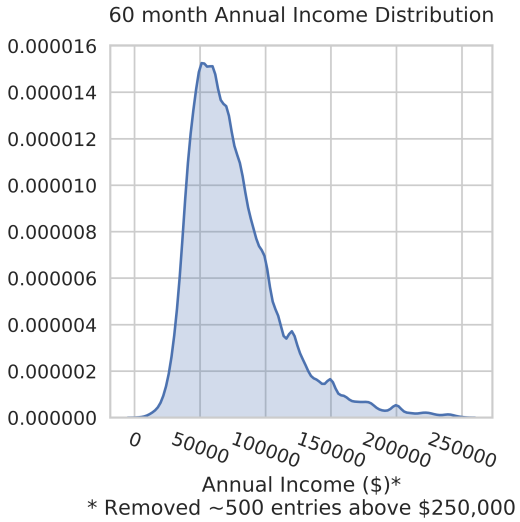
(b) 60 month home ownership distribution

Figure 7: Home Ownership Distribution

Annual Income: The distribution for both terms is relatively the same, both are unimodal and positively skewed with a tight spread, peaking at $50,000, as seen in Figures 8a and 8b.



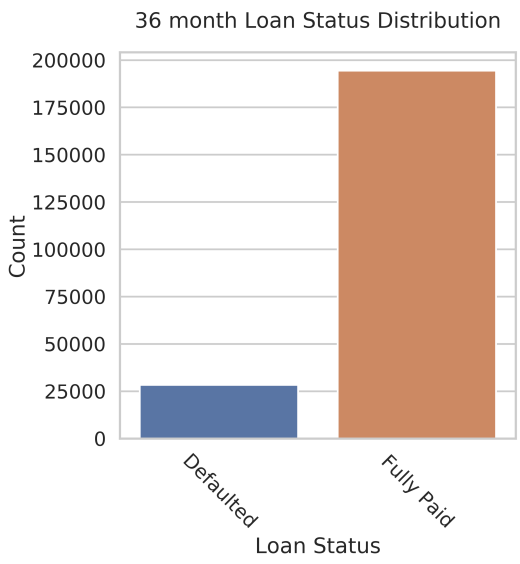
(a) 36 month annual income distribution



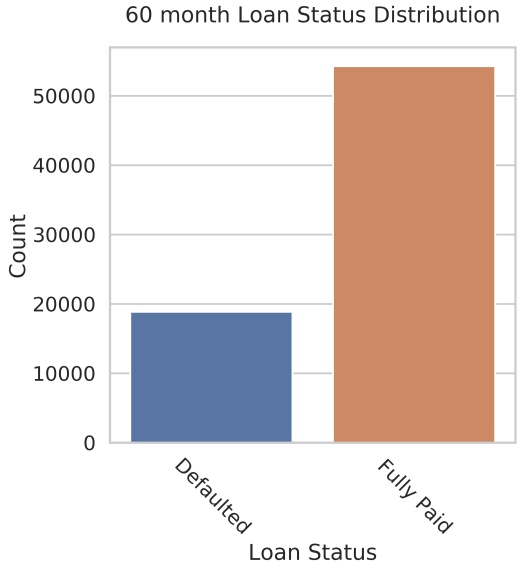
(b) 60 month annual income distribution

Figure 8: Annual Income Distribution

Loan Status: The 36 month loans (Figure 9a) have a lower ratio of defaulted to fully paid than the 60 month loans (Figure 9b), giving 60 month loans a larger relative sample to train and test on when predicting loan default.



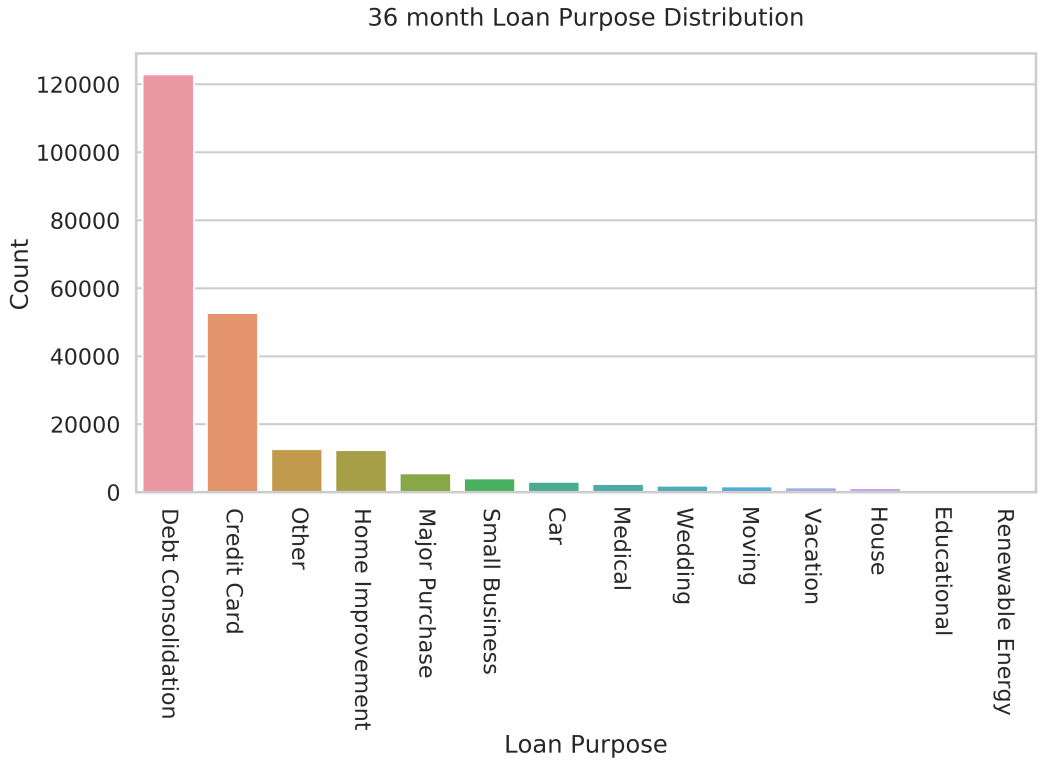
(a) 36 month loan status distribution



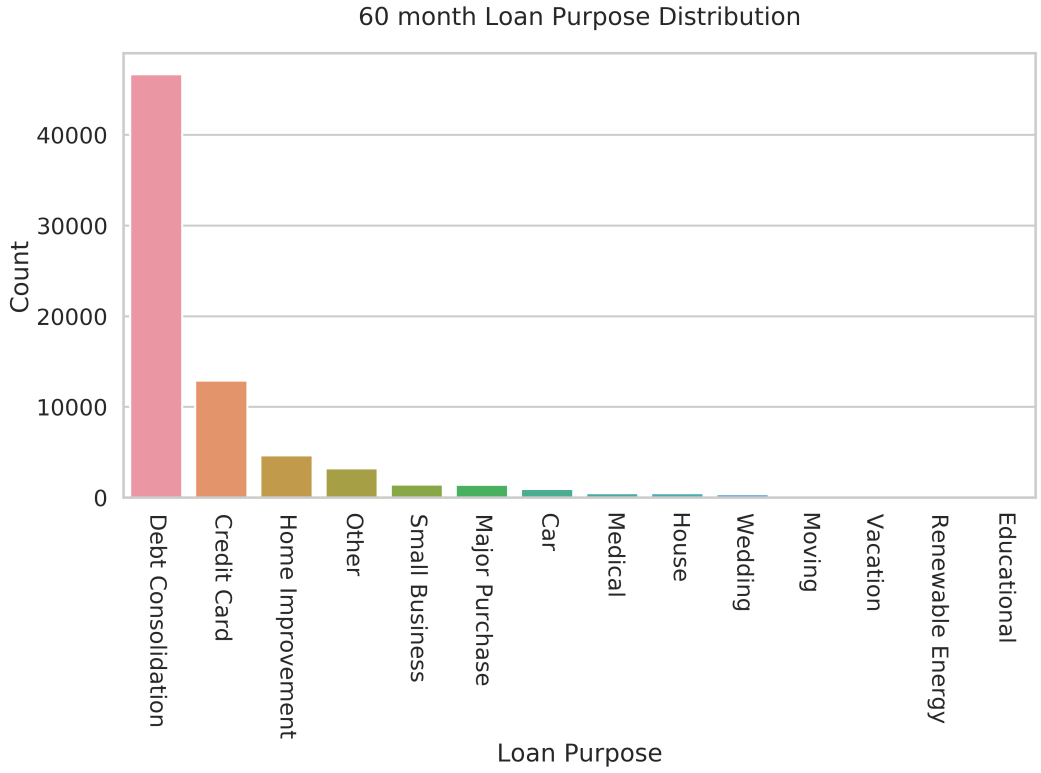
(b) 60 month loan status distribution

Figure 9: Loan Status Distribution

Loan Purpose: The most frequent purposes listed for both 36 month loans (10a) and 60 month loans (10b), with debt consolidation and credit cards being the two top reasons for borrowers requesting a loan.



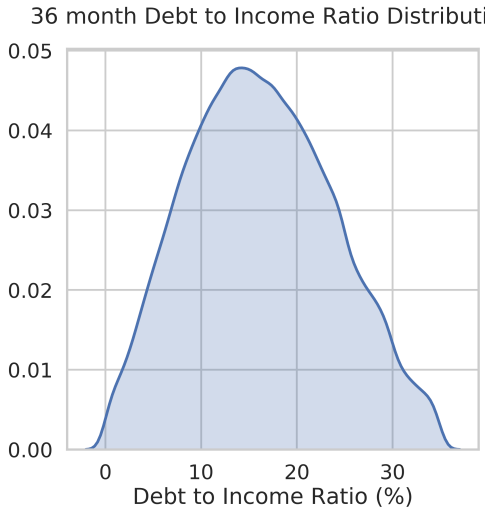
(a) 36 month loan purpose distribution



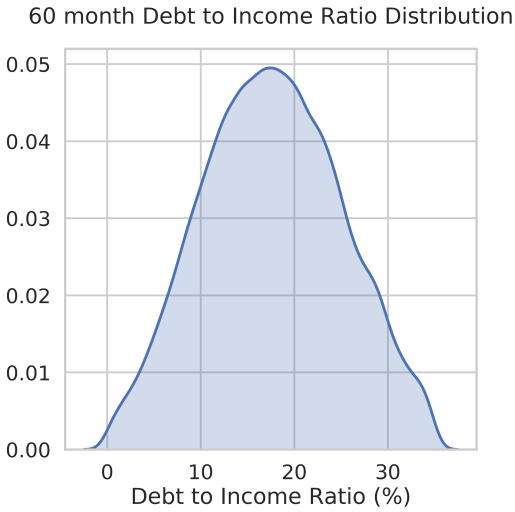
(b) 60 month loan purpose distribution

Figure 10: Loan Purpose Distribution

Debt-to-Income Ratio: As seen in Figures 11a and 11b, there is not much difference between the two terms, both are unimodal, symmetrical, and centered around 15%.



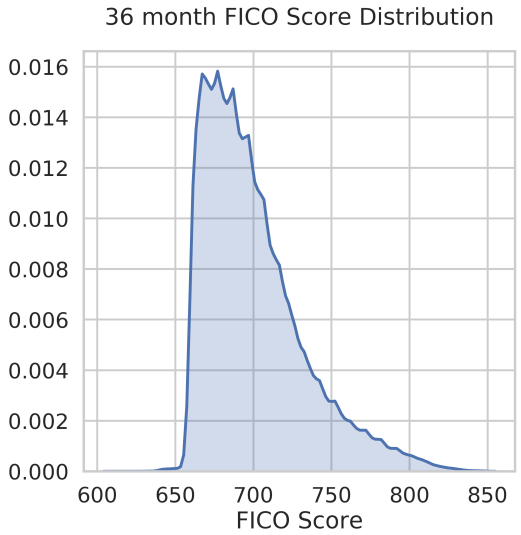
(a) 36 month DTI distribution



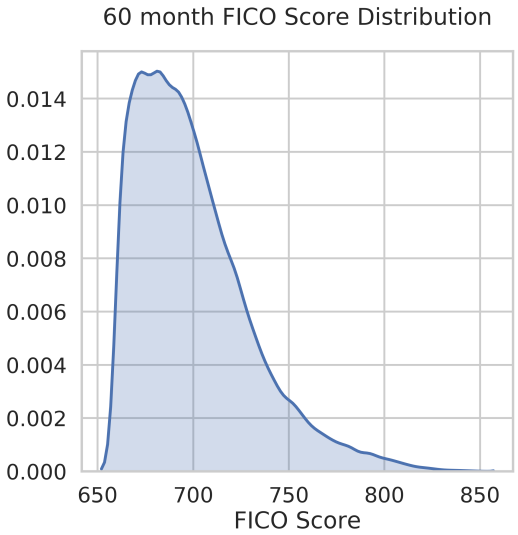
(b) 60 month DTI distribution

Figure 11: Debt-to-Income Ratio Distribution

FICO Score: Both terms are centered around 675, positively skewed with a tight spread. The 36 month loans in Figure 12a have some lower FICO scores, with very few appearing to be below 650 compared to the 60 month loans in Figure 12b.



(a) 36 month FICO score distribution



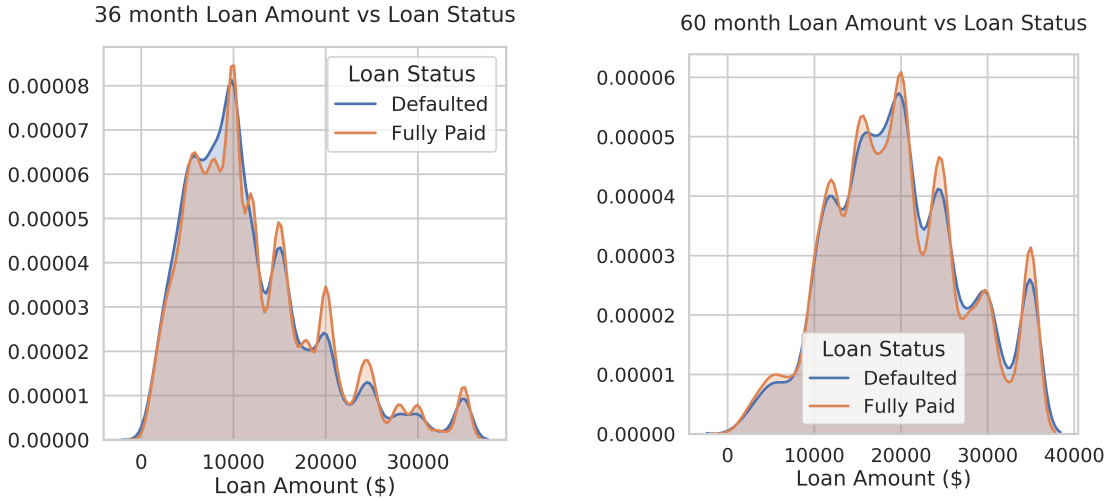
(b) 60 month FICO score distribution

Figure 12: FICO Score Distribution

2.2.2 Bivariate Analysis

In the Bivariate analysis, we're looking for large separation between the two loan statuses - defaulted and fully paid. If there is separation, then that variable will be critical in the model when trying to predict the loan status.

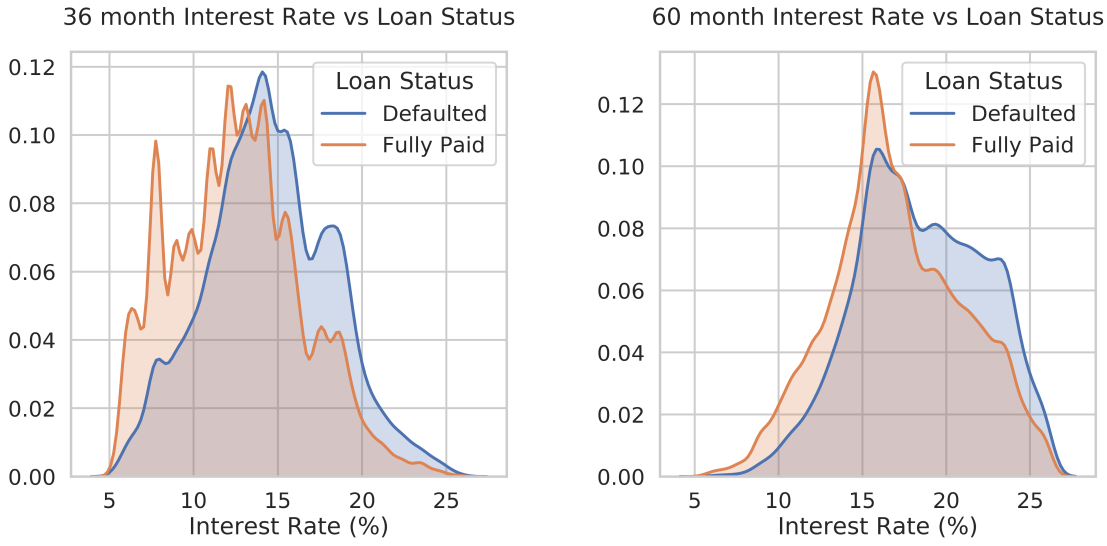
Loan Amount: There is very little separation between the classes for both terms, as seen in Figures 13a and 13b.



(a) 36 month loan amount vs loan status (b) 60 month loan amount vs loan status

Figure 13: Loan Amount vs Loan Status

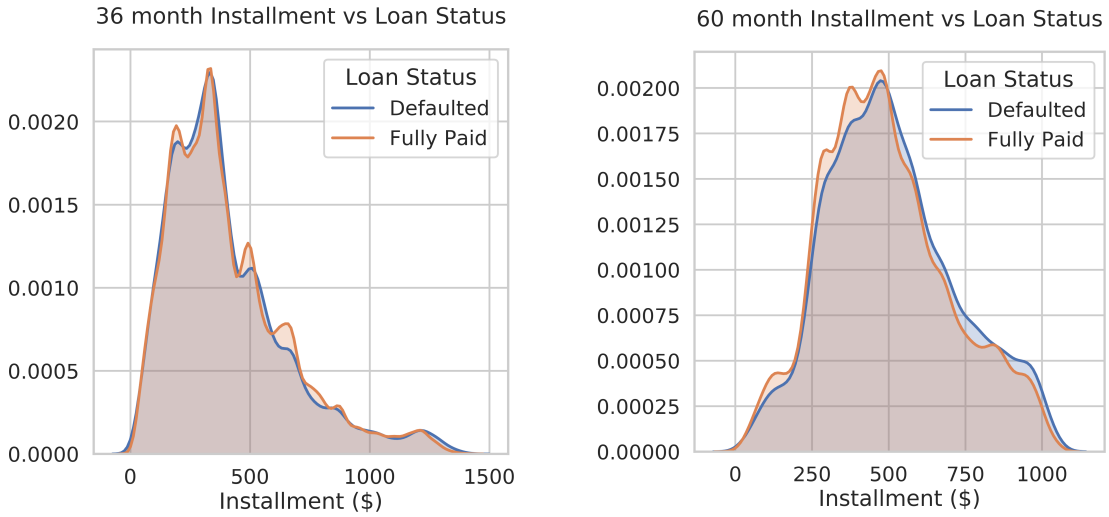
Interest Rate: There is a small amount of separation between the classes, as shown in Figure 14a and 14b. Defaulted loans are more likely to have a higher interest rate in both terms.



(a) 36 month loan interest rate vs loan status (b) 60 month loan interest rate vs loan status

Figure 14: Interest Rate vs Loan Status

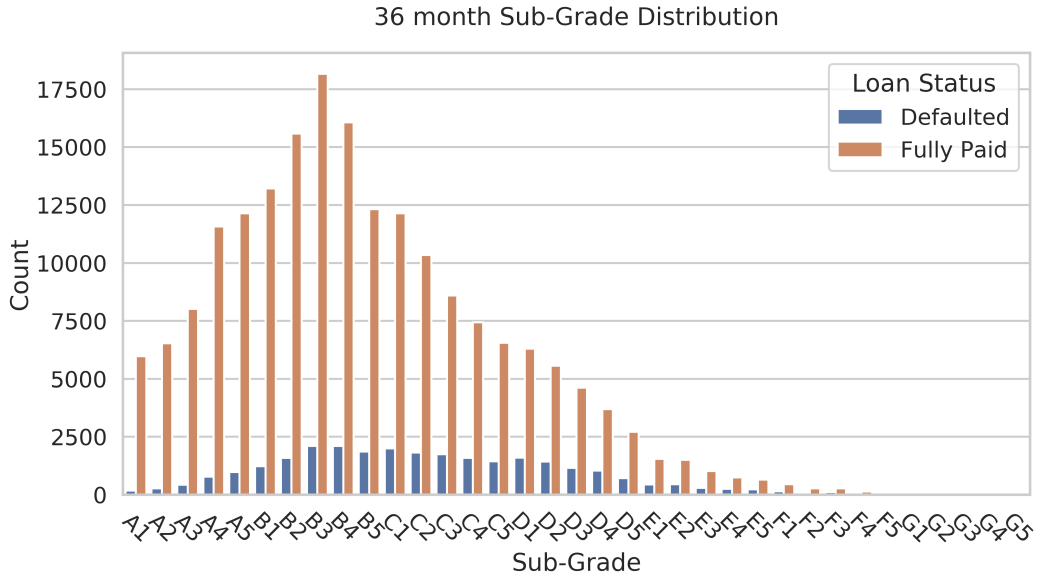
Installment: There is very little separation between the classes, as shown in Figures 15a and 15b.



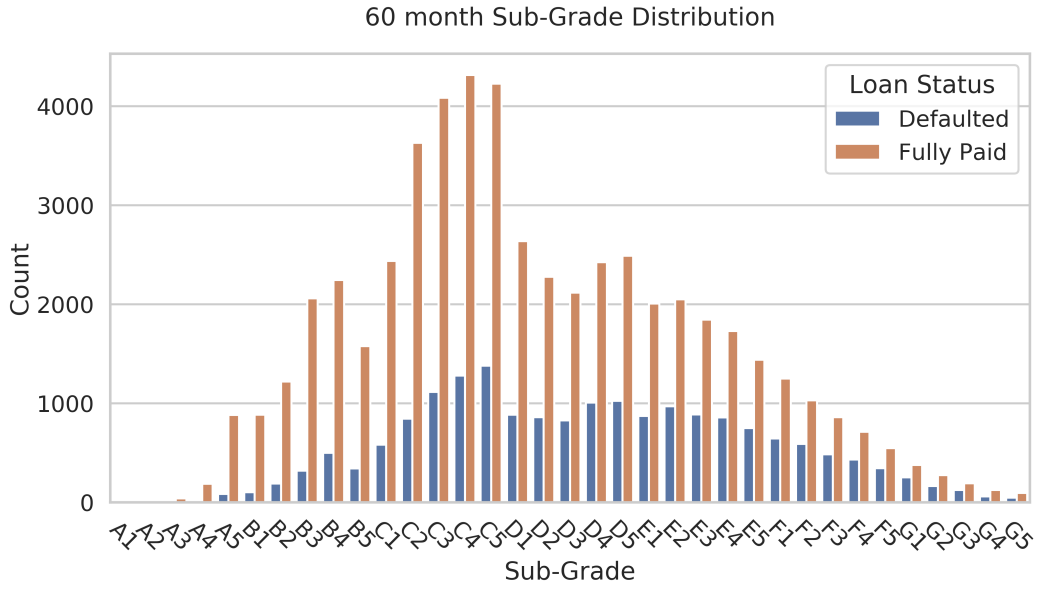
(a) 36 month loan installment vs loan status (b) 60 month loan installment vs loan status

Figure 15: Installment vs Loan Status

Loan Sub-Grade: It appears that as the sub-grade increases, the ratio between fully paid and default decreases so that grade A has a lower amount of default, and grade G has more defaults per fully paid. This appears in both the 36 month loans in Figure 16a and the 60 month loans in Figure 16b.



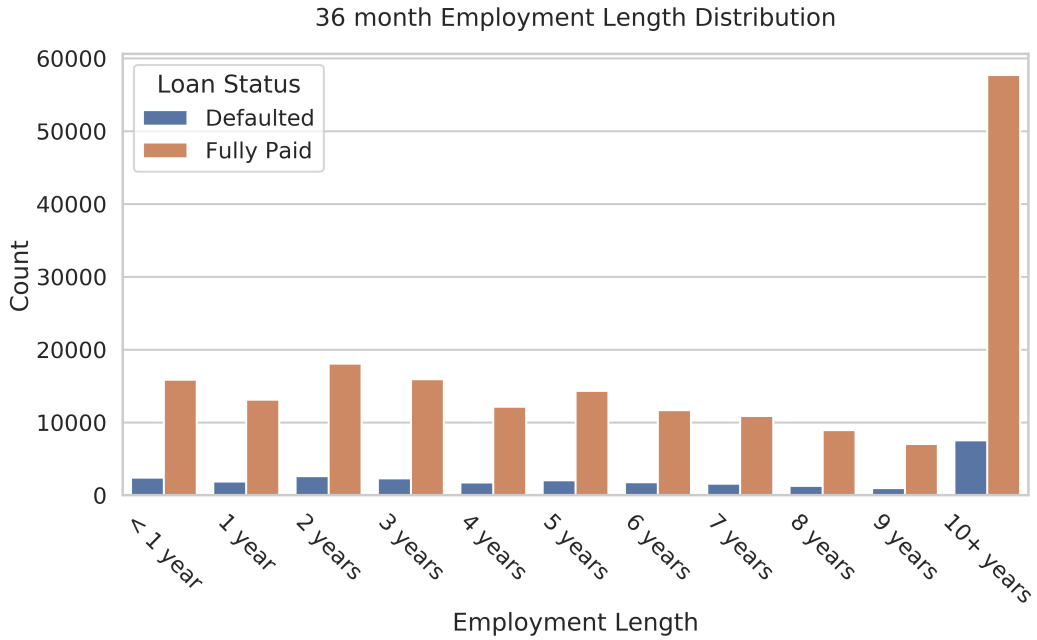
(a) 36 month loan sub-grade vs loan status



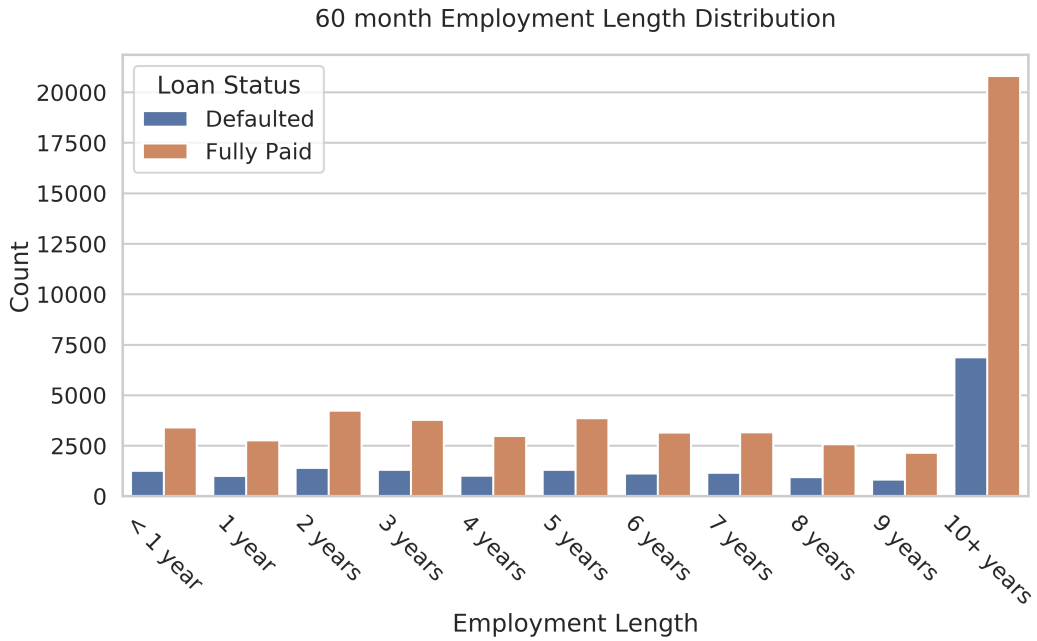
(b) 60 month loan sub-grade vs loan status

Figure 16: Loan Sub-Grade vs Loan Status

Employment Length: The default to fully paid ratio doesn't appear to change much between employment lengths. 10+ years has the most defaults, but also the most overall loans in both terms. See Figures 17a and 17b for more information.



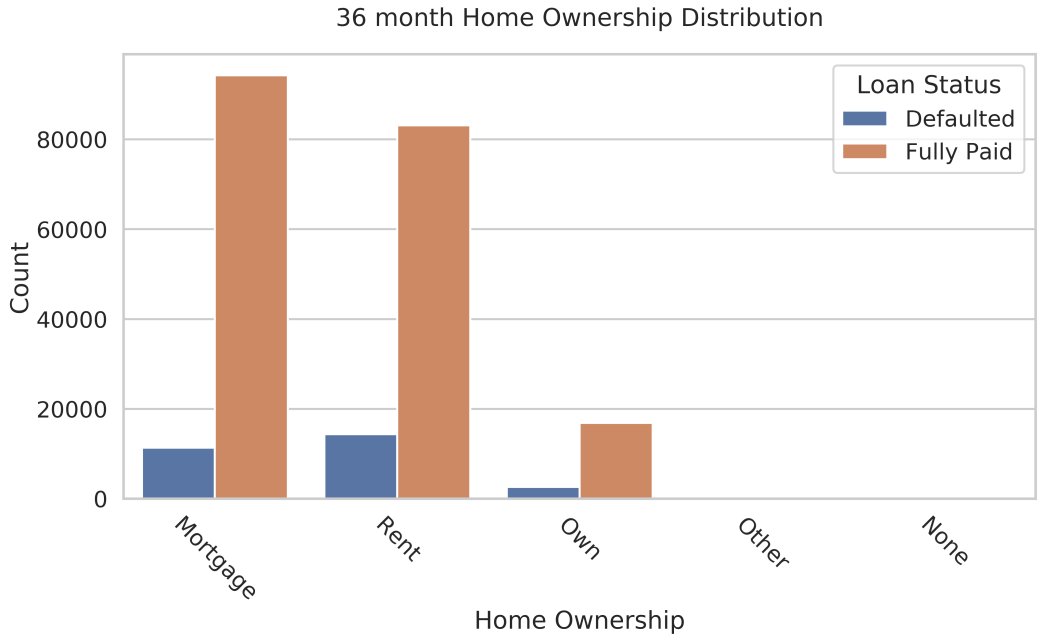
(a) 36 month loan employment length vs loan status



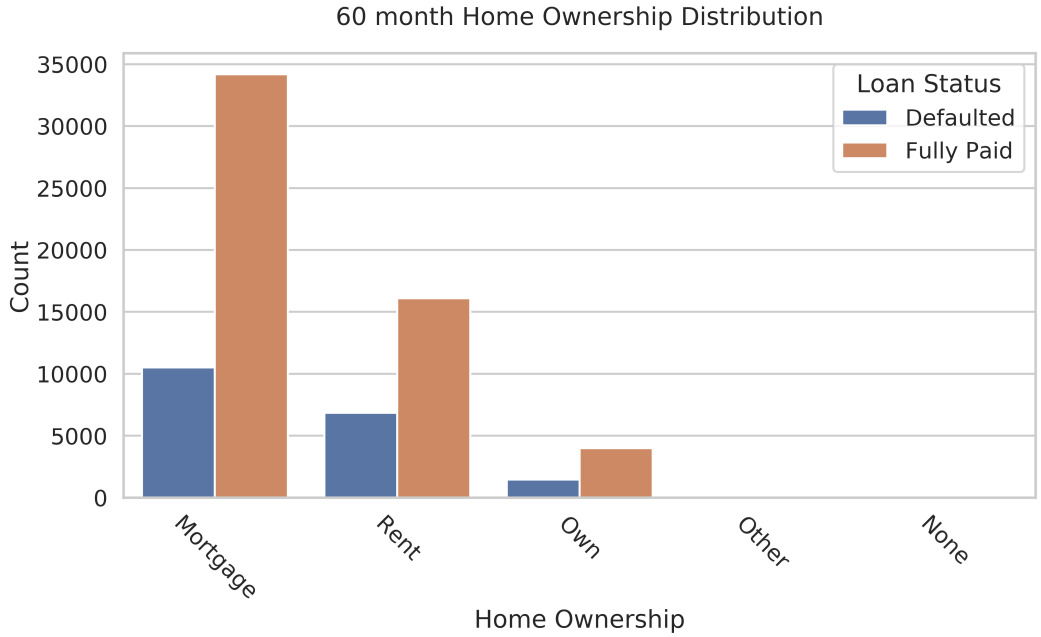
(b) 60 month loan employment length vs loan status

Figure 17: Employment Length vs Loan Status

Home Ownership: In Figure 18a, the amount of defaulted 36 month loans is higher in rent than mortgage in 60 month loans as shown in Figure 18b.



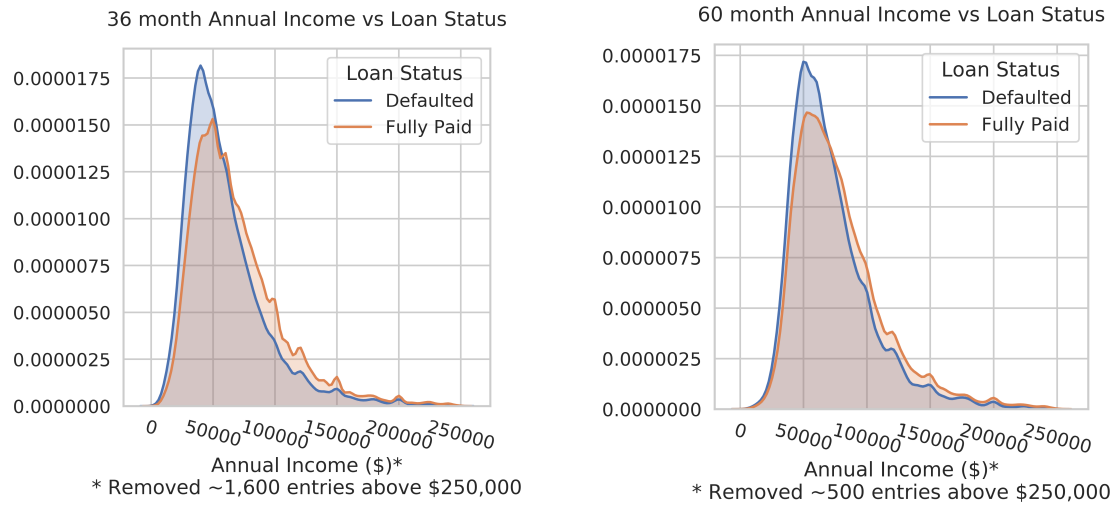
(a) 36 month loan home ownership vs loan status



(b) 60 month loan home ownership vs loan status

Figure 18: Home Ownership vs Loan Status

Annual Income: There is little separation between the classes, except that those that defaulted are slightly more likely to have a lower income, as the distribution's spread is a little tighter. This is seen in both Figure 19a and Figure 19b.

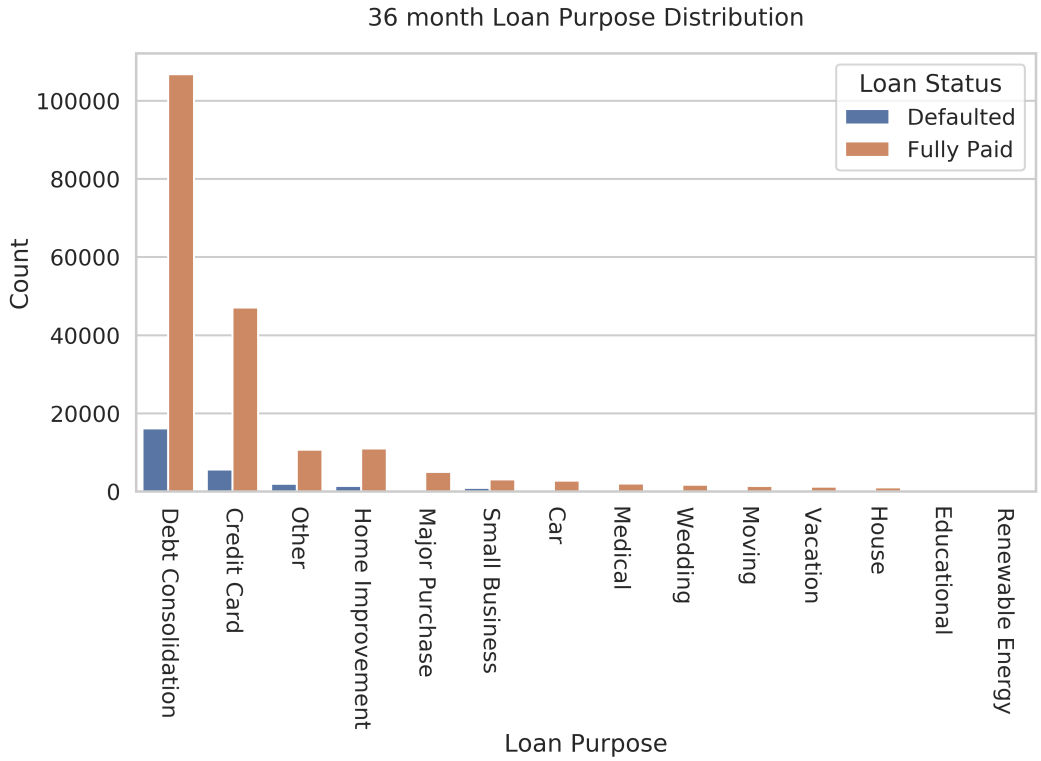


(a) 36 month loan annual income vs loan status

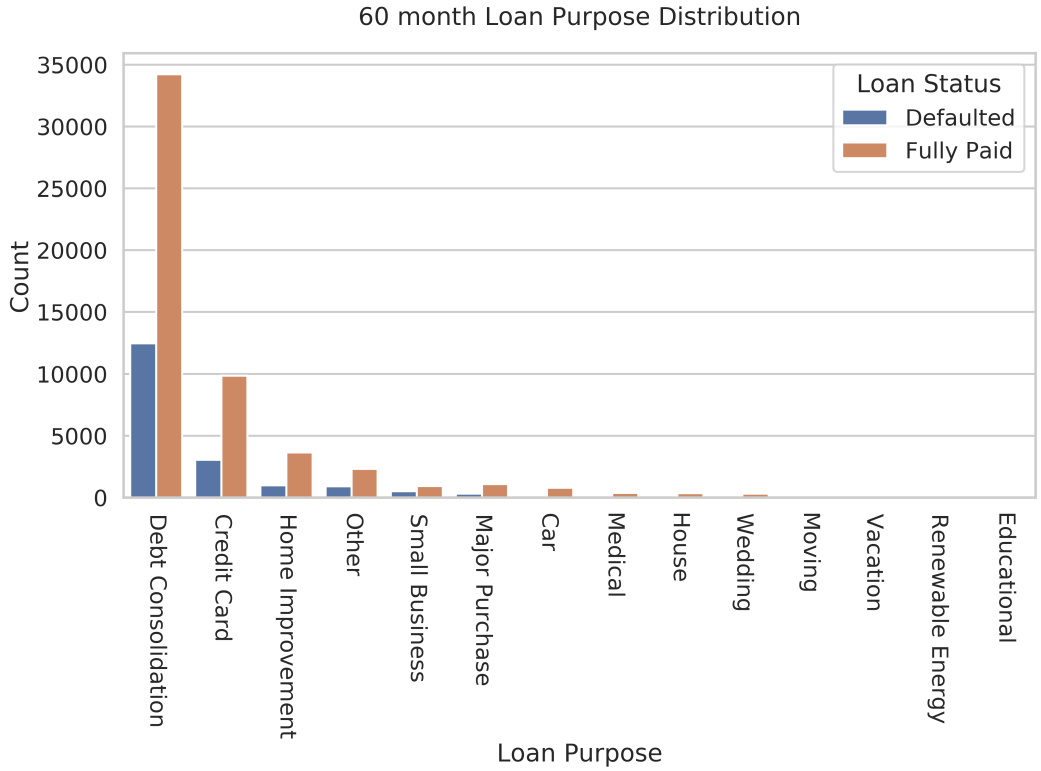
(b) 60 month loan annual income vs loan status

Figure 19: Annual Income vs Loan Status

Loan Purpose: For the 60 month loans in Figure 20b, debt consolidation seems to have a higher ratio between default and fully paid loans than the 36 month loans in Figure 20a.



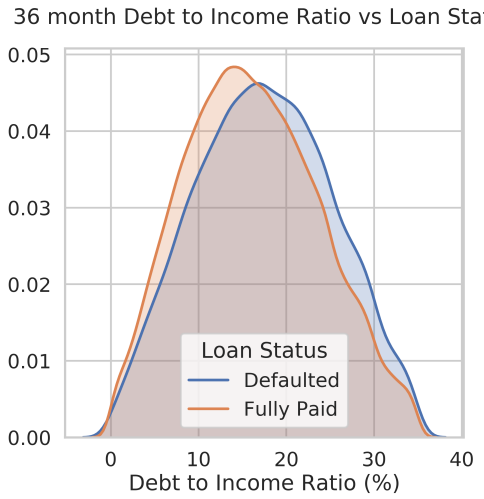
(a) 36 month loan purpose vs loan status



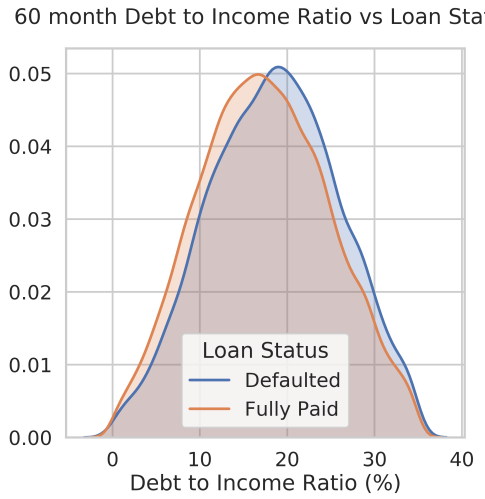
(b) 60 month loan purpose vs loan status

Figure 20: Loan Purpose vs Loan Status

Debt-to-Income Ratio: There is a small amount of separation between the classes in Figures 21a and 21b, with the defaulted loans being slightly more likely to have a higher debt-to-income ratio.



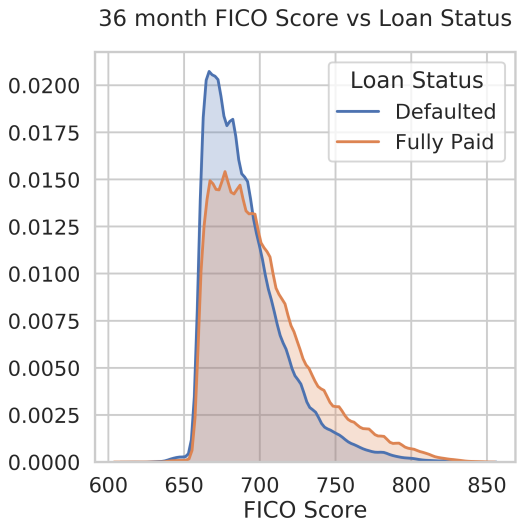
(a) 36 month loan DTI vs loan status



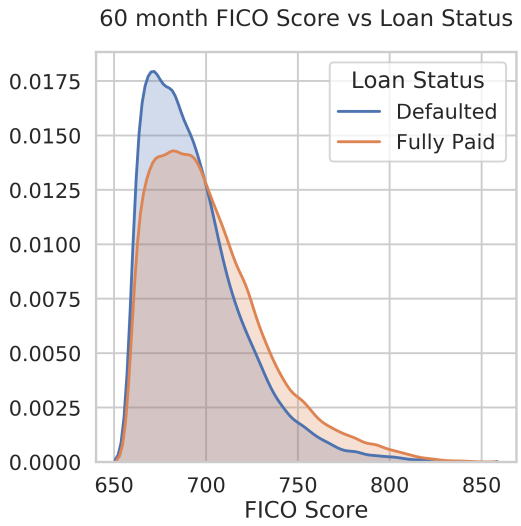
(b) 60 month loan DTI vs loan status

Figure 21: Debt-to-Income Ratio vs Loan Status

FICO Score: Defaulted loans have a higher density of borrowers with lower FICO scores, as is shown in the tighter spread in both terms, as displayed in Figures 22a and 22b.



(a) 36 month loan FICO score vs loan status



(b) 60 month loan FICO score vs loan status

Figure 22: FICO Score vs Loan Status

2.3 Baseline Analysis

To begin, a baseline analysis of the financial data was performed to see how well a model could predict loan default without the aid of the textual analysis.

2.3.1 Data Preprocessing

The data was randomly split into a test and a training set, with 70% of data going to the testing set. Figure 23 gives an overview of the path data takes during prepreocessing.

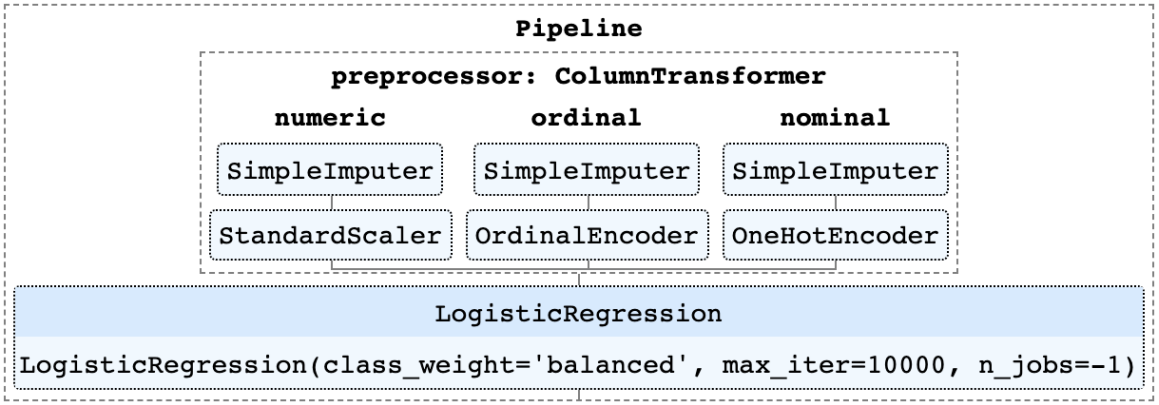


Figure 23: Pipeline used to preprocess the LendingClub data.

There are three different paths the data could take based on the column it resided in - numeric, ordinal, or nominal. Ordinal and nominal data are both different types of categorical data. Ordinal data has an inherent order, while nominal categories don't have a hierarchy applied to them. Loan Amount, Interest Rate, Installment, Annual Income, Debt-to-Income Ratio, and FICO score were all processed as numeric data. The missing values were imputed with the median value for the column and scaled. Employment Length and Sub-Grade were processed as ordinal features, which imputed missing values as a separate category, "missing." Then categories were replaced with integers, with "missing" coded as zero. Home Ownership and Purpose were processed as nominal features and were imputed the same way as ordinal features. Categories were dummy-coded with the One Hot Encoder, where each unique category is replaced with a column containing a binary indicator of whether that category is present in the sample.

2.3.2 Model

Logistic regression was used to train and classify the data. One of the most widely used and simpler statistical techniques used for credit risk classification, logistic regression has very good performance while having comparable results to more complex non-linear classification models such as Support Vector Machines [1, 2, 4, 22]. It is used mainly to model binary dependent variables, which is the case for the data here. The independent variables may be continuous, discrete, binary, or a combination of all three. The logistic function itself is a sigmoid function that receives real numbers and outputs a value between zero and one, displayed in Figure 24. The standard logistic function $\sigma : \mathbb{R} \to (0, 1)$ is defined as

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(B_0 + B_1 x_1 + \ldots + B_m x_m)}}$$

Where $p(x)$ is the probability of the dependent variable belonging to the success case.

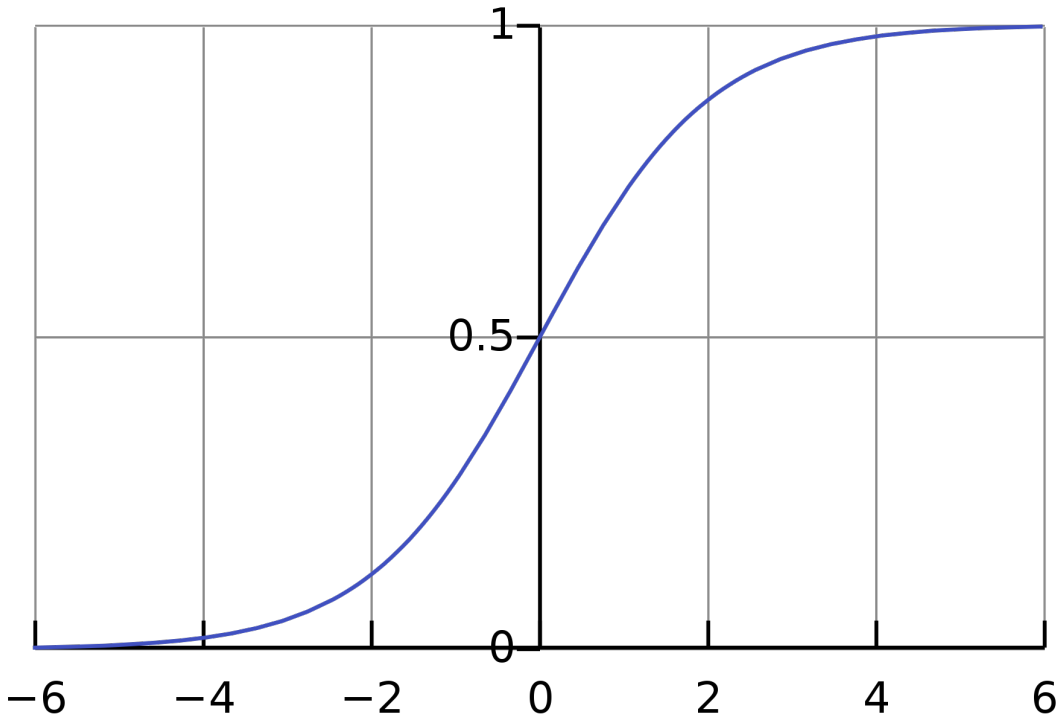


Figure 24: A graph of a sigmoid function.

There are 5 assumptions made when performing logistic regression [23]:

1. **Assumption of Appropriate Outcome Structure:** The dependent variable must be binary and ordinal.

2. **Assumption of Observation Independence:** The observations must be independent of each other, or measurements should not be related or repeated.

3. **Assumption of Absence of Multicollinearity:** There should be little to no multicollinearity within the independent variables, such that none of the variables are highly correlated with each other.

4. **Assumption of Linearity of Independent Variables and Log Odds:** The independent variables and log odds should be linearly related. The Log Odds are defined as the $\log\left(\frac{\text{Probability}}{1-\text{Probability}}\right)$

5. **Assumption of a Large Sample Size:** Logistic regression requires a large sample size, usually at least 10 cases for the least frequent outcome for each independent variable in the model.

Testing the assumptions on the model:

1. **Assumption of Appropriate Outcome Structure:** This assumption is valid. The dependent variable - whether the loan will be paid or default - is binary but is not ordinal

2. **Assumption of Observation Independence:** This assumption is valid. Each observation represents a separate loan. Though unlikely, it is possible that a single borrower may have multiple loans from LC - but it is impossible to know from the data provided.

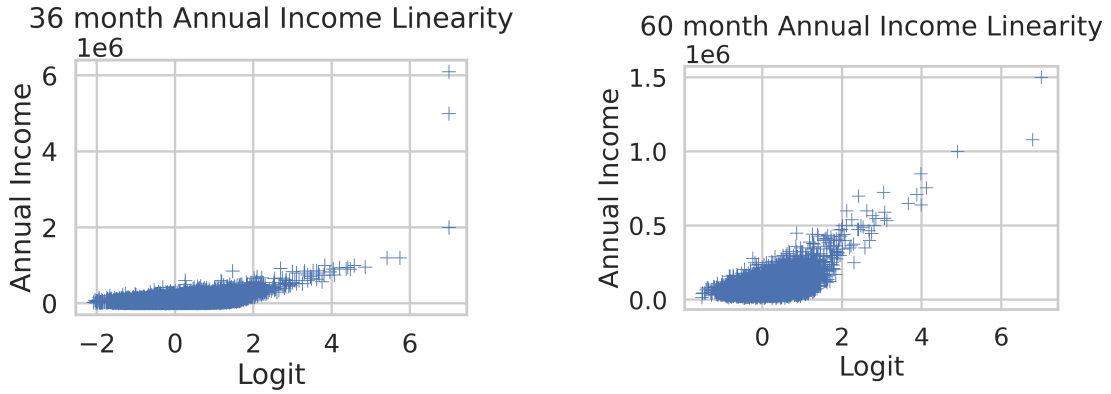
3. **Assumption of Absence of Multicollinearity:** This assumption is not valid. Table 3 shows the Variance Inflation Factors (VIF), which is a measure of multicollinearity. Any measure exceeds 5 or 10 indicates a problematic amount of collinearity. Interest Rate, Debt-to-Income Ratio, and FICO Score all have some amount of collinearity. This follow intuition, as DTI and FICO both are contributing factors in determining your interest rate.

VIF Factor	**Features**
4.8	Loan Amount
10.5	Interest Rate
3.3	Annual Income
6.2	Debt-to-Income Ratio
14.7	FICO Score

Table 3: Variance Inflation Factors for numeric data.

4. **Assumption of Linearity of Independent Variables and Log Odds:** This assumption is not valid. Below are the numeric variables versus the Log Odds, and only 3 of the 6 numeric features are linearly related.

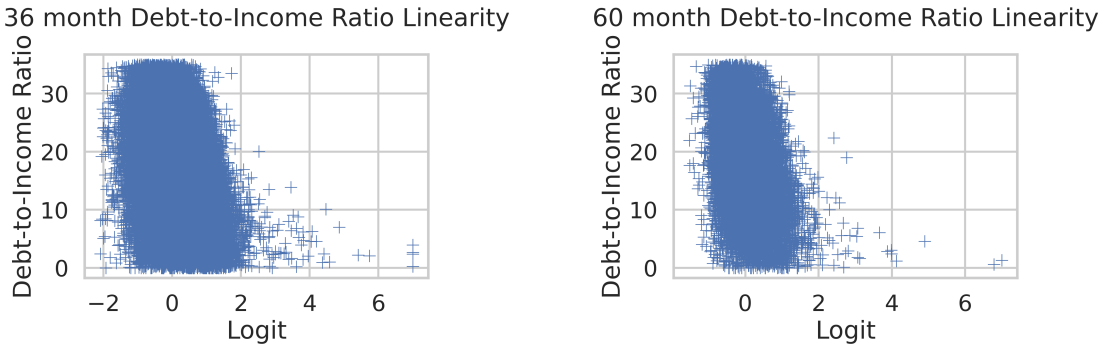
- **Annual Income:** Figures 25a and 25b show a moderate amount of linearity, with a moderate amount of correlation.



(a) 36 month loan annual income vs log odds (b) 60 month loan annual income vs log odds

Figure 25: Annual Income vs Log Odds

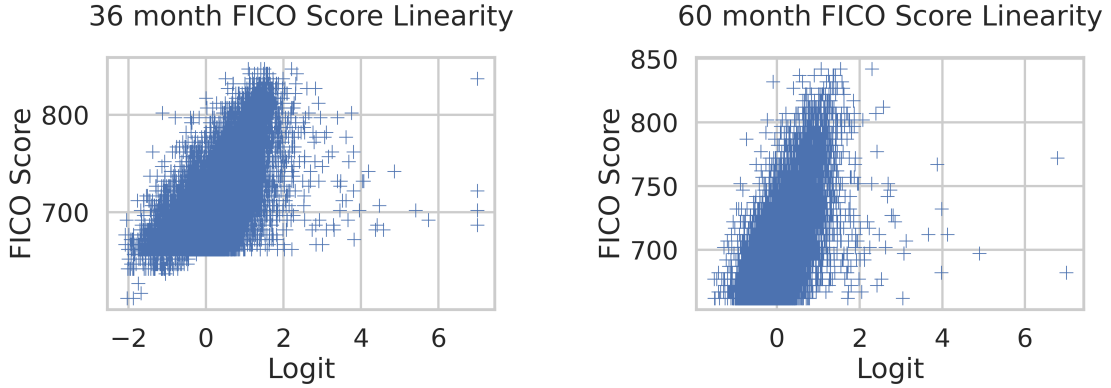
- **Debt-to-Income Ratio:** Figures 26a and 26b show almost no linear correlation.



(a) 36 month loan DTI vs log odds (b) 60 month loan DTI vs log odds

Figure 26: Debt-to-Income Ratio vs Log Odds

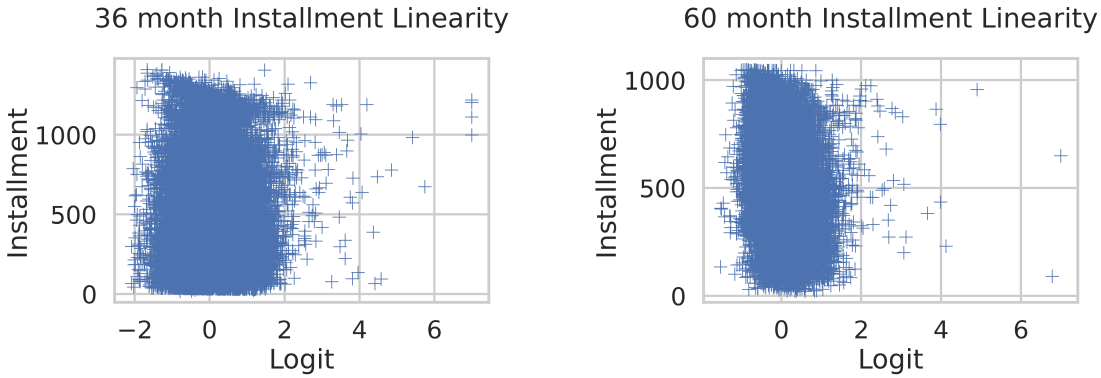
- **FICO Score:** Figures 27a and 27b show a weak linear correlation.



(a) 36 month loan FICO score vs log odds (b) 60 month loan FICO score vs log odds

Figure 27: FICO Score vs Log Odds

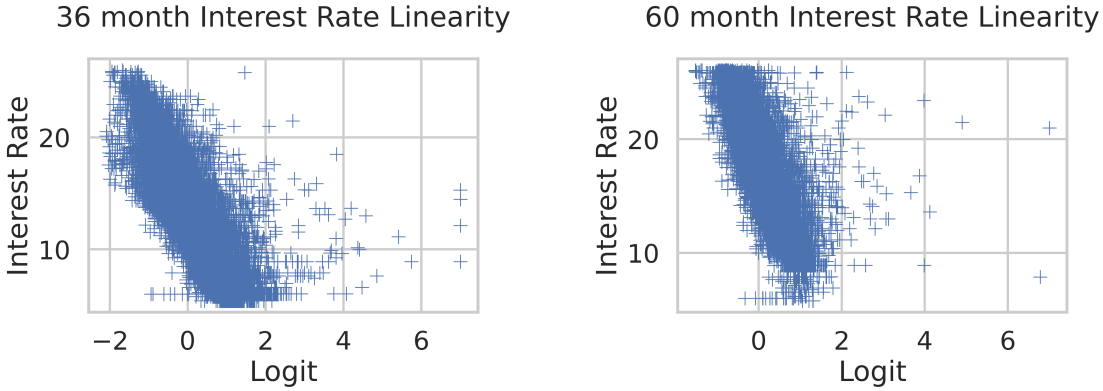
- **Installment:** As shown in Figures 28a and 28b, there does not appear to be a linear correlation.



(a) 36 month loan installment vs log odds (b) 60 month loan installment vs log odds

Figure 28: Installment vs Log Odds

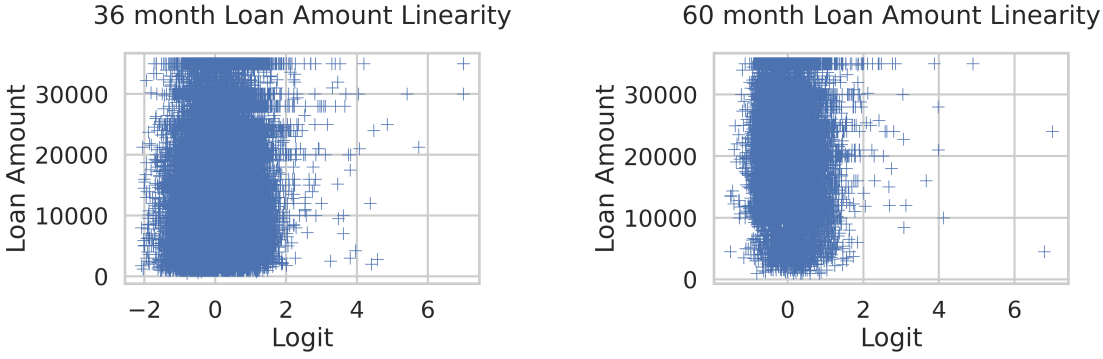
- **Interest Rate:** As shown in Figures 29a and 29b, there is a moderate linear correlation.



(a) 36 month loan interest rate vs log odds (b) 60 month loan interest rate vs log odds

Figure 29: Interest Rate vs Log Odds

- **Loan Amount:** As shown in Figures 30a and 30b, there does not appear to be a linear relationship.



(a) 36 month loan amount vs log odds (b) 60 month loan amount vs log odds

Figure 30: Loan Amount vs Log Odds

5. **Assumption of a Large Sample Size:** This assumption is valid. For 12 variables, and with a least frequent outcome of 0.12, there has to be at least $(10 * 12/0.12) = 1,000$ loan default samples. There are 47,170 defaulted loans between the two terms.

Many models have hyper-parameters that require tuning before the model can be used effectively. Logistic regression has C, tolerance, solver, and penalty as hyper-parameters. C is the inverse regularization parameter. It controls the strength of regularization - the lower the C value, the stronger the regularization, which prevents overfitting. The tolerance determines when the algorithm has converged and when it will stop iteration. Smaller values will result in longer run times and approach closer to actual convergence. The solver is the algorithm that is used when optimizing the problem. SAGA and LBFGS were selected for this study.

The SAGA solver is a variant of the SAG solver, which uses Stochastic Average Gradient descent. It is faster than other solvers for large datasets. SAGA adds support for the L1 penalty [5]. The LBFGS solver is an optimization algorithm that uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm, which approximates the Broyden-Fletcher-Goldfarb-Shanno algorithm with a limited amount of memory. It is the default setting for logistic regression in the Scikit-Learn Python package and is recommended for small datasets as its performance suffers on larger datasets.

The penalty specifies the norm used in penalization. LBFGS only supports L2 penalties, whereas SAGA supports L1 and L2. L2 regularization, or Ridge Regression, is a technique for analyzing multiple regression data that suffer from multicollinearity by introducing bias to reduce variance [7]. It penalizes logistic regression based on minimizing this cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\big(\exp\big(-y_i(X_i^T w + c)\big) + 1\big)$$

.

L1 regularization, or Least Absolute Shrinkage and Selection Operator (LASSO) Regression, is similar to Ridge Regression, but rather than penalizing the sum of squared coefficients, it penalizes the sum of their absolute values [25]. It minimizes the following cost function:

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log\big(\exp\big(-y_i(X_i^T w + c)\big) + 1\big).$$

The model was created using Python and the Scikit-Learn package on Jupyter Notebook. A 5-fold cross-validation grid search was performed to find the best combination of the hyper-parameters, scoring on balanced accuracy. The class weight was set to "balanced", and the maximum iteration was set to 10,000.

2.4 Textual Analysis

Once the baseline analysis was performed with the logistic regression model, the textual data from the description field of the loan application was parsed for sentiment and summed using the Bag-of-Words method and the Loughran-McDonald Sentiment Word list.

2.4.1 Data Preprocessing

To begin, missing values were filled with empty string literals, and the text was changed to lowercase. LC included a prefix whenever a borrower edited their description in the following format that was removed - "User name/User ID/Borrower added on date ¿ text." All HTML line breaks, carriage returns, tabs, URLs, and ¡/¿ signs were removed from the text. The Python package Unidecode was used to replace accented characters with their underlying ASCII

characters. Once the text was encoded in UTF-8, the Contractions package was used to expand contractions such as "wasn't" to "was not" to prevent accidental removal of apostrophes that might provide more content. This was performed before the removal of stop words because the package incorporates context-aware scanning to replace contractions with the correct word where there might be multiple expansions. For example, "He's" might be expanded to "he is" or "he has" depending on the surrounding context. While it is relatively trivial for a human to decode these contractions, there is not a hard rule to expansion, and thus machines are not 100% accurate when expanding them.

The text was then scrubbed to only contain alphabetic characters and spaces. Since Bag-of-Words was the selected form of analysis, sentence structure was not needed and punctuation was removed. Finally, each word was lemmatized with the WordNetLemmatizer, stop words defined by the Natural Language Toolkit (NLTK) Python package were removed, and repeated whitespace was removed. The final cleaned text was saved in a separate column, and the original text was also kept.

Lemmatization is the process of reducing words such as "ran," "running," and "runs" to their root word, "run." This improves the performance of Natural Language Processing, as they are all branches of the same word, but would not otherwise be interpreted as the same word. Lemmatization is a more complex version of stemming, which removes the suffixes to find the root word. It is a faster operation, but the root may not be an actual word, whereas lemmatization will find the language root word. Lemmatization is slower because it has to use the content around the word to derive the correct meaning and determine if a word is already in its root form. For example, "universal" and "university" both have the root "univers," but this would be an incorrect stem which lemmatization would catch [17].

2.4.2 Model

From the original text, the Textstat Python package was used to calculate the average score from the Flesch-Kincaid Grade Level, the Fog Scale, SMOG Index, Automated Readability Index, the Coleman-Liau Index, Linsear Write Formula, and Dale-Chall Readability Score, which returns a float representing what grade level student would be able to understand the text. Because many of the tests include word count as a variable, it was performed on the unclean text to include words that were not important for analysis.

Sentiment Analysis was performed using the Loughran-McDonald Sentiment Word List, which is a corpus built specifically for textual analysis within a financial context [16]. The list contains 6 different sentiment categories: Negative, Positive, Uncertainty, Litigious, Strong Modal, Weak Modal, and Constraining. Negative, Positive, and Uncertainty are self-explanatory, Litigious indicates a propensity for legal contest, including words such as "claimant" and "deposition." Strong and Weak Modal words express the level of confidence within the text, with Strong Modal words such as "always" and "must," and Weak Modal words such as "could" or "depending." Each word that appeared in a given category was counted and summed in a column. The Litigious category was not used in this analysis.

Once the text was analyzed and the results were stored, the best performing model from the grid search was trained on the same training data with the new columns included. The model was evaluated on the same measures as the original model.

3 Results and Discussion

3.1 Results

We will report the precision, recall, F_1-score, support, and balanced accuracy of the 36 and 60 month loan models, with their respective baseline and textual analyses. Each is broken down further by class, as our model aims to predict defaulted loans.

Precision is defined as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}} + \text{False Positives},$$

where a True Positive is a correctly identified positive condition, and a False Positive is a negative condition marked as positive [19]. Recall is defined as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}},$$

where a False Negative is a positive condition marked as negative [19]. More concisely, precision is the number of correct identifications within the selected items, and recall is the number of correct identifications out of all of the items that should have been identified. Recall is trivial to achieve 100% if you simply classify all loans as a specific class, so it must be used in conjunction with precision. The F_1-score is the harmonic mean of precision and recall [20], and thus gives a single statistic that factors in both, defined as

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Support is simply the number of loans that were used in the testing set to produce the given result.

Normally, accuracy is used to evaluate models. It is defined as

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}.$$

This is a misleading metric for imbalanced classes, as we have here. If the majority class is predicted for all items, it will achieve an 88% accuracy. Instead, we use balanced accuracy, which is defined as

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.$$

Sensitivity, or the True Positive Rate, is another name for recall and has the same formula. Specificity, or True Negative Rate, is defined as

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}.$$

Balanced accuracy takes both the minority and majority class into account, and gives a more accurate picture of the model [18].

Table 4 displays the final results. The bold numbers indicate the better performing model between the baseline and textual analysis models, and the underlined numbers indicate that the result is statistically significant. There were 100 trials run for each model, and a paired T-test was performed to calculate significance. An alpha level of 0.05 was used to test for significance.

Table 4: Logistic regression and textual analysis results, averaged over 100 trials.

	36 Month Loans				60 Month Loans			
	Default		Fully Paid		Default		Fully Paid	
	Base	Text	Base	Text	Base	Text	Base	Text
Prec	**18.81%**	18.80%	91.74%	**91.76%**	**33.15%**	33.14%	80.67%	**80.69%**
Recall	62.44%	**62.66%**	**60.75%**	60.59%	60.01%	**60.13%**	**57.97%**	57.87%
F1	28.91%	**28.92%**	**73.10%**	72.98%	42.71%	**42.73%**	**67.46%**	67.40%
Supp	19,824.39		136,111.61		13,195.58		37,993.42	
C	0.1				0.1			
Solver	SAGA				SAGA			
Pen	L1				L2			
Tol	0.0001				0.0001			
Time	Base: 103.89 sec — Text: 66.05 sec				Base: 18.52 sec — Text: 16.44 sec			
Acc	Base: 61.59% — Text: **61.62%**				Base: 58.99 — Text: **59.00%**			

3.2 Discussion

As we can see in Table 4, the 36 month model had a statistically significant improvement with the aid of textual analysis, while the 60 month model had a small improvement, but was not statistically significant. The 36 month model performed better overall. This is most likely due to the larger dataset - there are almost three times as many samples in the 36 month dataset. Returning back to the hypothesis: If the sentiment and reading grade from the text are used in conjunction with financial data to predict loan default, then the model will see a statistically significant improvement in performance. The hypothesis was confirmed for the 36 month loan model, but not for the 60 month model. The 60 month model gives cause for further exploration as it showed a slight improvement, just not enough to attribute to the textual data.

As is clear from the EDA, the classes are far from linearly separable. In the bivariate analysis, there is no one feature that is able to clearly separate default and fully paid loans. This is also embodied in the lack of linearity of the features to the log odds. Ideally, all of the features would have a strong correlation to the log odds. Additionally, there are extra measures we had to take to be able to work with the class imbalance, as 88% of the dataset is classified as fully paid for the 36 month loans. An imbalanced dataset is defined as "[a]n imbalance occurs when one or more classes have very low proportions in the training data compared to the other classes" [14]. As with most imbalanced classification problems, predicting the minority class is our top priority. But this makes it more difficult, as the majority of our training data will not provide information about what makes a default loan, giving us less training data to work with.

There were multiple steps taken to address the class imbalance. For the logistic regression formula, we are able to use the class weights parameter to adjust the weights inversely proportional to the class frequencies, expressed as

$$\text{Class Weights} = \frac{\text{Number of Samples}}{\text{Number of Classes} * \text{Number of Samples in Each Class}}.$$

For scoring, balanced accuracy was used, as it takes the average performance of both classes rather than overall performance, as discussed in Results.

For both datasets, we can see that the default recall and F_1-score improved with the textual analysis, but only the fully paid precision improved with the textual analysis. For the 36 month dataset, the default precision is quite low at 18.80%, but the recall is comparable to the fully

paid loans at 60% for each. Not suprisingly, SAGA was the most commonly selected solver as it is more efficient for larger datasets. Interestingly, the penalty was the hyper-parameter that varied between the two models. Perhaps with more data, the 60 month loan model would be able to penalize with the L1 penalty as well.

3.3 Future Work

A more thorough analysis and sentiment extraction may have lead to a larger improvement in performance. As discussed in the Textual Analysis section, we only analyzed sentiment through the Bag-of-Words method using the Loughran-McDonald Sentiment Word list. This inevitably will create an oversimplified model, as the context of the words within the loan application is discarded for their net meaning. A Parts-of-Speech analysis, or POS, might have extracted more information, as it separates sentences into their corresponding parts (nouns, verbs, adjectives, etc) and you are able to create sentence vectors to explore relationships between nouns and their modifiers. This type of analysis requires a thorough understanding of the English language and demands a large amount of computer power due to the size and complexity of the data.

The dataset itself was limited in its textual data. LendingClub originally made the description field required, but changed it to optional about halfway through the dataset. Descriptions were missing in nearly half of the entries, reducing the amount of data that could be learned by the model. If we were to restart this research, a different, more diverse and populated dataset could lead to more substantial findings.

Finally, a time-series analysis might be able to predict with more accuracy than without. One of the main reasons for removing columns was because they included time-series data, and it was not included in our analysis.

3.4 Specifications

The analysis was performed on the ThinkStation E31 Tower (2555) with a Xeon E3-1245V2 processor, 32 GB of RAM, Samsung NVMe storage, running Ubuntu v18.04. The code was run and compiled using a Jupyter Notebook server v6.0.3, running Python v3.8.3. The following packages were used:

- Pandas v1.0.5
- Numpy v1.18.5
- SciPy v1.5.0
- MatPlotLib v3.1.0
- Seaborn v0.11.0
- Scikit-Learn v0.23
- Natural Language Toolkit v3.5
- PyContractions v2.0.1
- TQDM v4.48.2
- IPython v7.16.1

4 Conclusion

The objective of this paper was to determine whether textual analysis can be used to improve the accuracy of loan default prediction. The final model performed at 61.62% and 59.00% accuracy for 36 and 60 month loans respectively. There was a significant improvement for the 36 month loan model with textual analysis, but not for the 60 month model. This is in line with other research done on textual analysis, showing an improvement in loan default prediction. Finally, this analysis shows that while textual analysis is important, it is certainly in no position to replace the financial history that is included in the application. It is much more useful when used in conjunction with other data, and at this point does not have predictive power on its own.

References

[1] Hussein A. Abdou, Marc D. Dongmo Tsafack, Collins G. Ntim, and Rose D. Baker. Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, 103:89–103, jul 2016.

[2] Bart Baesens, Tony Van Gestel, Stijn Viaene, M. STEPANOVA, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54, 06 2003.

[3] Paul Belleflamme, Thomas Lambert, and Armin Schwienbacher. Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, 29(5):585–609, sep 2014.

[4] Anthony Bellotti and Jonathan Crook. Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36:3302–3308, 03 2009.

[5] Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *CoRR*, abs/1407.0202, 2014.

[6] Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47(1):54–70, oct 2014.

[7] Marvin H. J. Gruber. *Improving Efficiency by Shrinkage*. Routledge, nov 1998.

[8] D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 160(3):523–541, 1997.

[9] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847–856, nov 2007.

[10] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, nov 2010.

[11] Elyssa Kirkham. Personal loan statistics: An overview of the changing loan landscape, January 2020.

[12] Matt Komos. Consumer credit origination, balance and delinquency trends: Q4 2018, March 2019.

[13] Matt Komos. Consumer credit origination, balance and delinquency trends: Q1 2020, June 2020.

[14] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer New York, 2013.

[15] S Li, W Shiue, and M Huang. The evaluation of consumer loans using support vector machines. *Expert Systems with Applications*, 30(4):772–782, may 2006.

[16] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, jan 2011.

[17] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[18] Jeffrey P. Mower. Prep-mt: predictive rna editor for plant mitochondrial genes. *BMC Bioinformatics*, 6(1):96, 2005.

[19] David L. Olson and Dursun Delen. *Advanced Data Mining Techniques*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[20] David Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Mach. Learn. Technol.*, 2, 01 2008.

[21] Gao Qiang and Lin Mingfeng. Lemon or cherry? the value of texts in debt crowdfunding. *Center for Analytical Finance*, July 2015.

[22] K B Schebesch and R Stecking. Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56(9):1082–1088, sep 2005.

[23] Deanna Schreiber-Gregory. Logistic and linear regression assumptions: Violation recognition and control. 2018.

[24] Schumpeter. Peer review - lending club, Jan 2013.

[25] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[26] Shuxia Wang, Yuwei Qi, Bin Fu, and Hongzhi Liu. Credit risk evaluation based on text analysis. *International Journal of Cognitive Informatics and Natural Intelligence*, 10(1):1–11, jan 2016.

5 Appendix

Table 5: The descriptions, cleaning, and removal of columns from the original LendingClub dataset.

Column	**Variable Type**	**Description**	**Removed**	**Reason for Removal**	**Cleaning Process**
ID	Numeric Integer	Unique LC identifier	Yes	Irrelevant Information	
Loan Amount	Numeric Integer	The listed amount applied for by the borrower	No		Read as an integer.
Funded Amount	Numeric Integer	The total amount committed to the loan at that point in time	Yes	Redundant Information	
Term	Ordinal Categorical String	The number of payments on the loan in months, either 36 or 60	No		Read as a string. Converted to category. Ordered to have "36 month" come before "60 month". Split datasets by this term, then removed from both datasets.
Interest Rate	Numeric Floating Point	Interest rate on the loan	No		Read as a float.
Installment	Numeric Floating Point	The monthly payment owed by the borrower	No		Read as a float.
Grade	Ordinal Categorical String	LC assigned loan grade	Yes	Redundant Information	
Sub-Grade	Ordinal Categorical String	LC assigned loan sub-grade	No		Read as a string. Converted to category. Ordered to have "A1" come before "G5".
Employee Title	String	The job title supplied by the borrower	Yes	Irrelevant Information	

Continued on next page

Table 5 – continued from previous page

Column	Variable Type	Description	Removed	Reason for Removal	Cleaning Process
Employment Length	Ordinal Categorical String	Employment length in years. Possible values are between < 1 year and 10+ years	No		Read as a string. Converted to category. Ordered to have "< 1 year" come before "10+ years".
Home Ownership	Nominal Categorical String	The home ownership status provided by the borrower. Possible values are Rent, Own, Mortgage, or Other	No		Read as a string. Converted to category. Renamed categories to be in Title Case.
Annual Income	Numeric Floating Point	The self-reported annual income	No		Read as a float.
Verification Status	Nominal Categorical String	Indicated if the income was verified by LC, not verified, or if the income source was verified	Yes	Irrelevant Information	
Earliest Credit Line	Date	The month the borrower's earliest reported credit line was opened	Yes	Time Sensitive	
Issue Date	Date	The month which the loan was funded	Yes	Time Sensitive	

Continued on next page

Table 5 – continued from previous page

Column	Variable Type	Description	Removed	Reason for Removal	Cleaning Process
Loan Status	Nominal Categorical Boolean	Current status of the loan. Possible values are Fully Paid, Charged Off, Current, In Grace Period, Late (16-30 days), or Late (31-120 days)	No		Read as a string. Removed "Does not meet the credit policy. Status:" prefix. Replaced "Charged Off" with "Defaulted". Removed rows that had the status "Current", "In Grace Period", "Late (16-30 days)", and "Late (31-120 days)". Converted to category.
Description	String	Description provided by the borrower	No		See Textual Analysis section.
Purpose	Nominal Categorical String	A category provided by the borrower for the loan request.	No		Read as a string. Converted to category. Renamed categories to be in Title Case.
Loan Title	String	The loan title provided by the borrower	Yes	Irrelevant Information	
Zip Code	String	The first 3 letters of the zip code	Yes	Irrelevant Information	
State Address	String	The state that the borrower lives in	Yes	Irrelevant Information	

Continued on next page

Table 5 – continued from previous page

Column	Variable Type	Description	Removed	Reason for Removal	Cleaning Process
Debt-to-Income Ratio	Numeric Floating Point	A ratio calculated using the borrower's total monthly debt payments excluding mortgage and the requested LC loan, divided by the monthly income	No		Read as a float.
Delinquencies in the last 2 years	Numeric Integer	The number of 30+ days past-due amount owed for the account on which the borrower is now delinquent	Yes	Time Sensitive	
FICO Range Low	Numeric Integer	The lower boundary of borrower's FICO score	Yes	Redundant Information	
FICO Range High	Numeric Integer	The upper boundary of borrower's FICO score	Yes	Redundant Information	
FICO Score	Numeric Integer	[Calculated] The average of the borrower's upper and lower FICO score boundaries	No		Read FICO Range Low and FICO Range High as integers. Averaged the two ranges together to create a new column, truncated to be an integer.
Inquiries in the last 6 months	Numeric Integer	The number of inquiries in the past 6 months (excluding auto and mortgage inquiries)	Yes	Time Sensitive	
Months since last delinquency	Numeric Integer	The number of months since the borrower's last delinquency	Yes	Time Sensitive	
Unpaid Principal Balance	Numeric Floating Point	The remaining balance on the loan if it was not completely paid off	Yes	Time Sensitive	

Continued on next page

Table 5 – continued from previous page

Column	Variable Type	Description	Removed	Reason for Removal	Cleaning Process
Default	Nominal Categorical Boolean	Whether the account has defaulted	Yes	Redundant Information	
Delinquent	Nominal Categorical Boolean	Whether the account is 31+ days delinquent	Yes	Redundant Information	
Time to Default	Numeric Integer	The time to default	Yes	Time Sensitive	
Loss Given Default	Numeric Floating Point	The amount of money lost when a loan defaults, expressed as a percentage	Yes	Time Sensitive	