Credit Risk Assessment with Textual Analysis

Willamette University | Advisor: Haiyan Cheng

Zachary Haroian

Introduction

Personal loans are one of the fastest-growing sources of debt within the US, tripling from $49 billion in 2010 to an all-time high of $162 billion in Q1 2020 and is steadily growing 10% every year.

Peer-to-Peer (P2P) lending is one of the most successful lending models implemented by Financial Technology (FinTech) innovators. Borrowers and lenders are matched together, allowing lenders to "invest" in individual loans.

Credit risk assessments have traditionally been based on debt-to-income ratio, credit history, and other financial information. Error rates in predicting loan default are usually between 20%-30%, so even marginal improvements in predictive power could result in significant savings. Textual data, when used in conjunction with financial history, could give a more accurate picture of the borrower's individual situation and increase the model's accuracy.

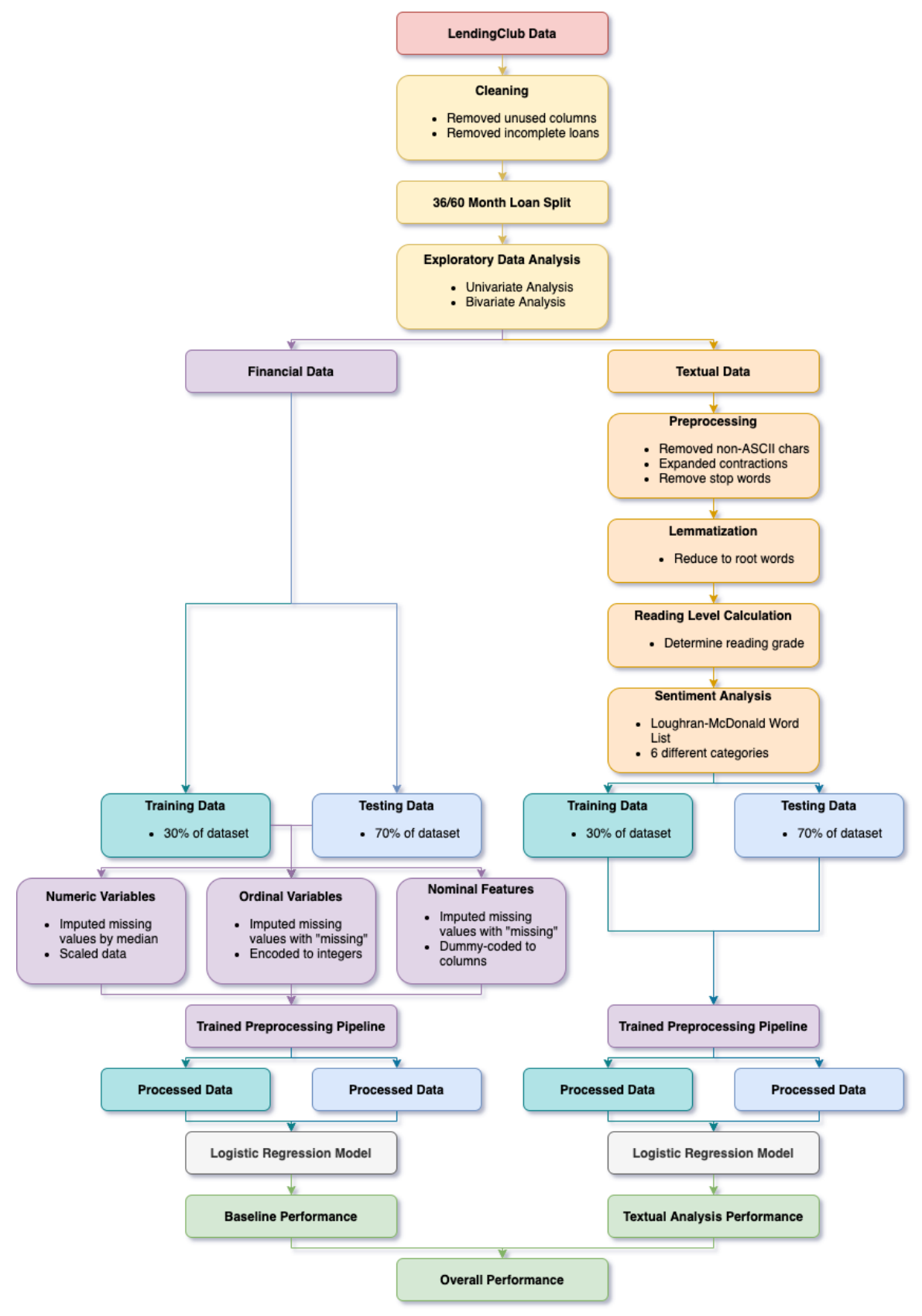
We find that there is a statistically significant improvement in accuracy when adding the sentiment analysis to the pre-existing model, given enough training data. Our finding agrees with current literature surrounding textual analysis and credit risk, though more advanced sentiment analysis may need to be used to glean a greater performance improvement.

Data

Data Attributes

- Retrieved from LendingClub
- 295,891 loans from 2007 to 2018
- Split by term:
  - 36 month: 222,765 loans
  - 60 month: 73,126 loans
- Removed 1,306 incomplete loans
- 33 columns - 12 kept

See the flowchart below for a complete path the data takes.



LendingClub Data

Cleaning
- Removed unused columns
- Removed incomplete loans

36/60 Month Loan Split

Exploratory Data Analysis
- Univariate Analysis
- Bivariate Analysis

Financial Data

Training Data
- 30% of dataset

Testing Data
- 70% of dataset

Numeric Variables
- Imputed missing values by median
- Scaled data

Ordinal Variables
- Imputed missing values with "missing"
- Encoded to integers

Nominal Features
- Imputed missing values with "missing"
- Dummy-coded to columns

Trained Preprocessing Pipeline

Processed Data

Processed Data

Logistic Regression Model

Baseline Performance

Textual Data

Preprocessing
- Removed non-ASCII chars
- Expanded contractions
- Remove stop words

Lemmatization
- Reduce to root words

Reading Level Calculation
- Determine reading grade

Sentiment Analysis
- Loughran-McDonald Word List
- 6 different categories

Training Data
- 30% of dataset

Testing Data
- 70% of dataset

Trained Preprocessing Pipeline

Processed Data

Processed Data

Logistic Regression Model

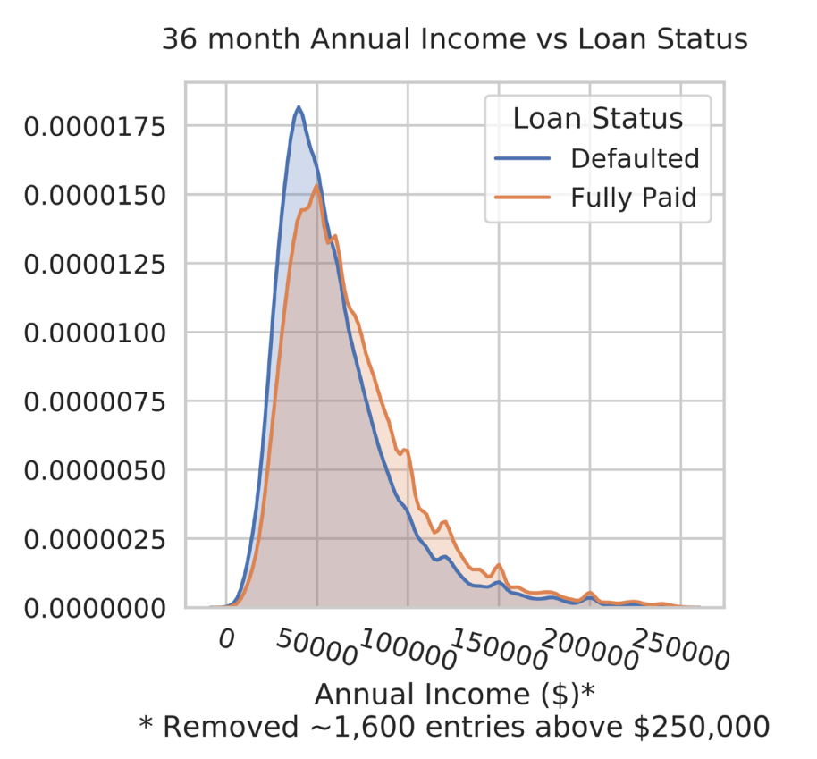
Textual Analysis Performance

Overall Performance

Exploratory Data Analysis

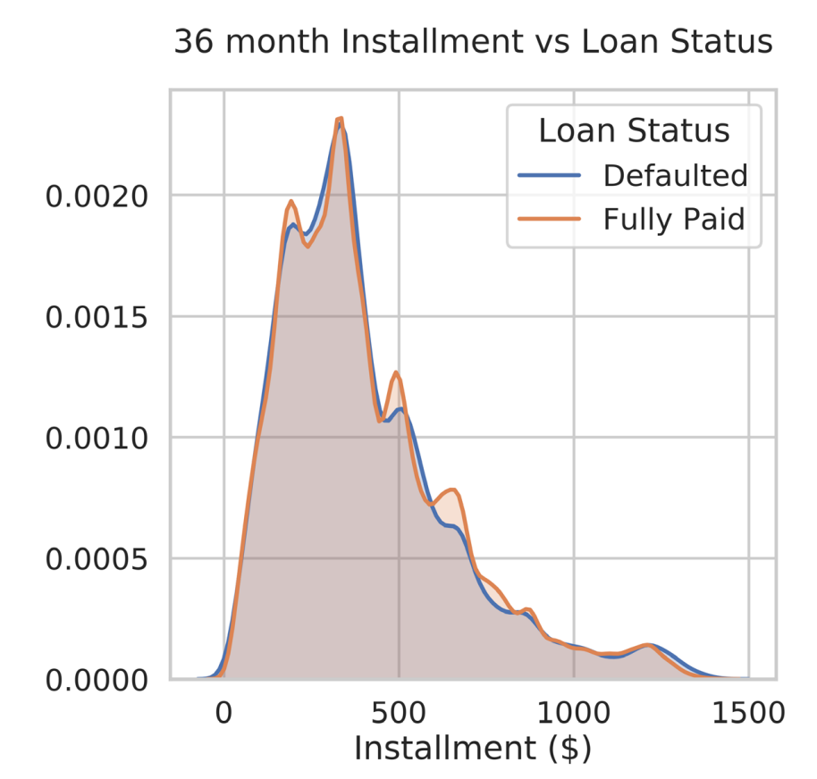
Before creating the model, it is imperative to get an idea of what the data looks like and find any important relationships within the features. Below are a few bivariate graphs from the 36 month loan set. We are specifically looking for a high degree of separation between the two classes (Defaulted and Fully Paid), which we can use later as a key predictor.

As you can see in the Annual Income density plot below, Defaulted loans are slightly more likely to have an annual income near $50,000. Installment has almost no separation between Defaulted and Fully Paid.



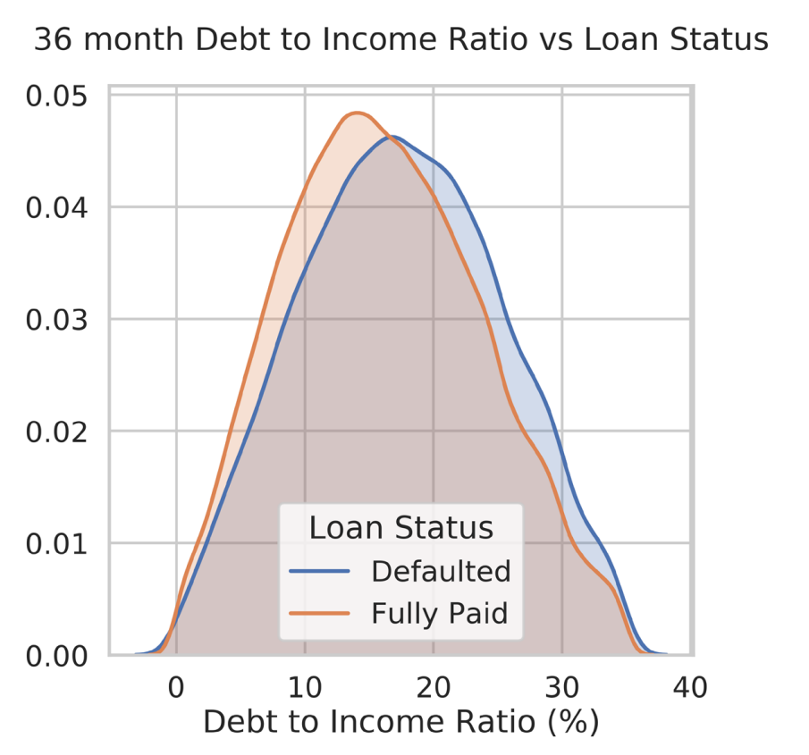
36 month Annual Income vs Loan Status

* Removed ~1,600 entries above $250,000

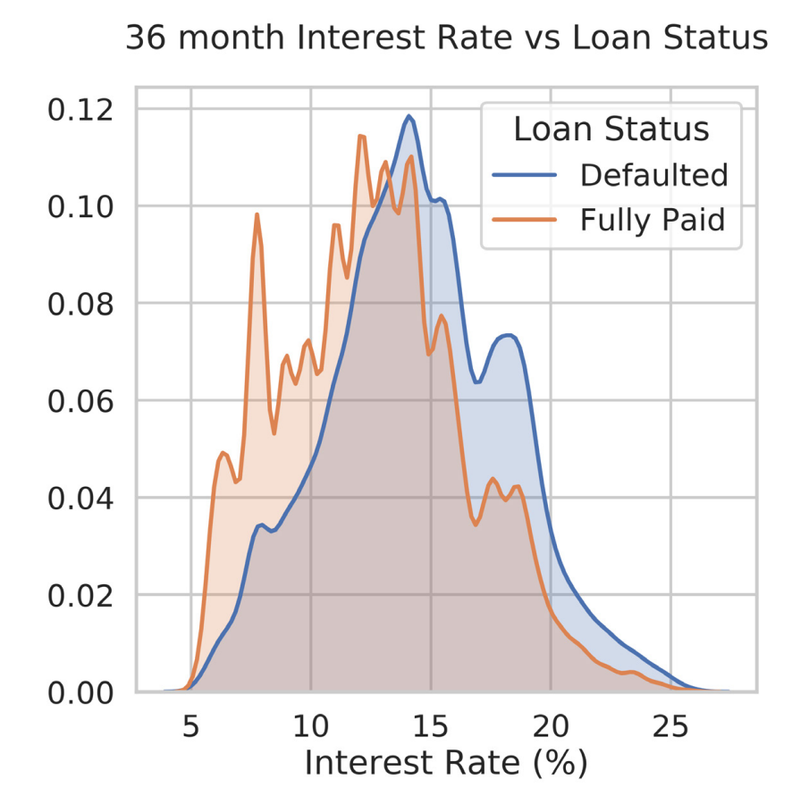


36 month Installment vs Loan Status

Debt-to-income (DTI) has a tiny amount of separation, with Defaulted loans being slightly more likely to have more debt compared to their income. Interest rate has a small amount of separation, with Fully Paid loans being more likely to have a lower interest rate.

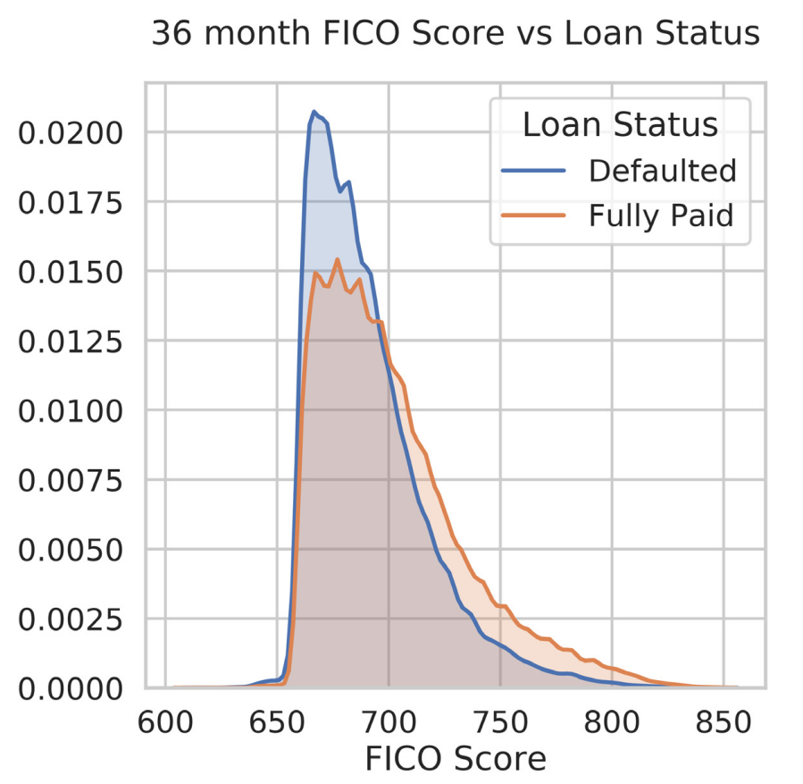


36 month Debt to Income Ratio vs Loan Status

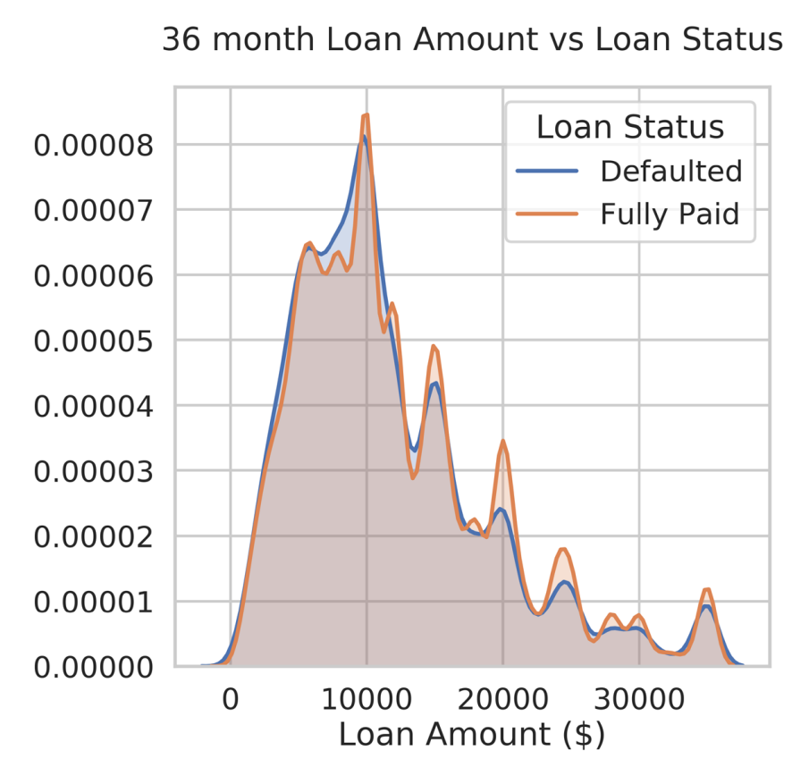


36 month Interest Rate vs Loan Status

FICO scores for Defaulted loans are more likely to be lower, with the highest FICO scores only being found in Fully Paid loans. The Loan Amount, like Installment, has no separation.

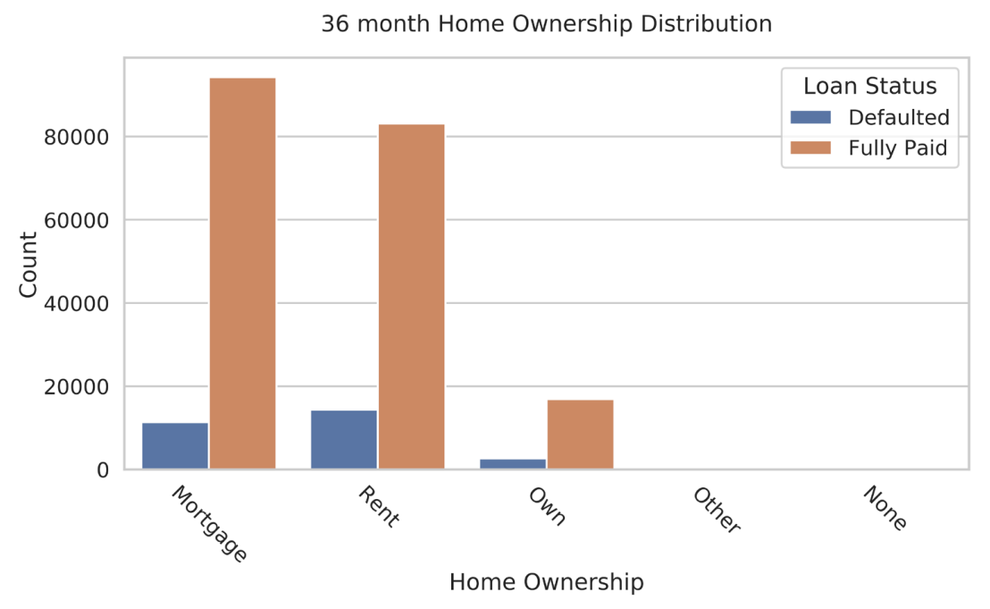


36 month FICO Score vs Loan Status

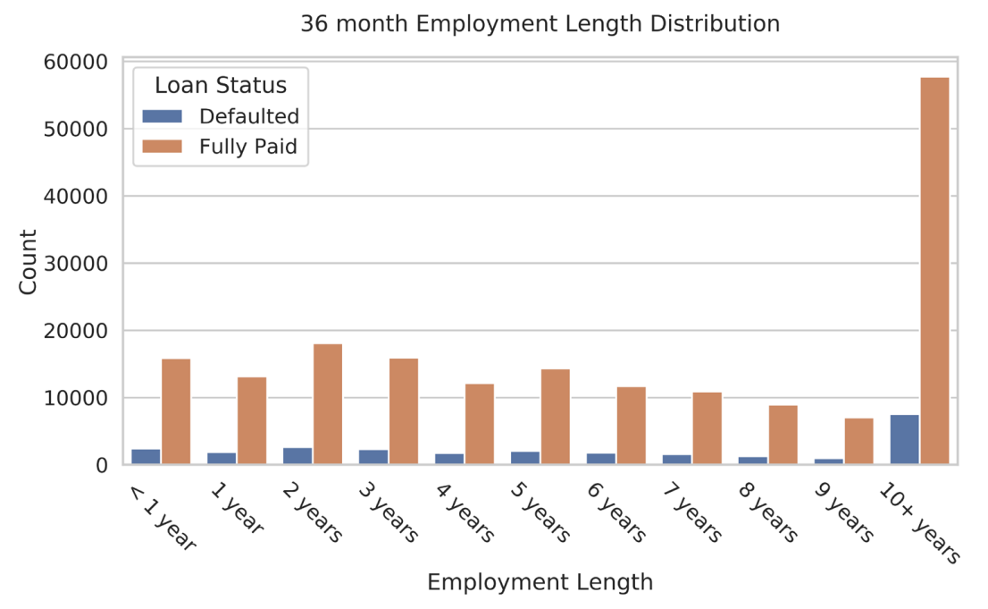


36 month Loan Amount vs Loan Status

Renters are slightly more likely to default than those with a mortgage. Those employed for 10+ years are the most common applicants.



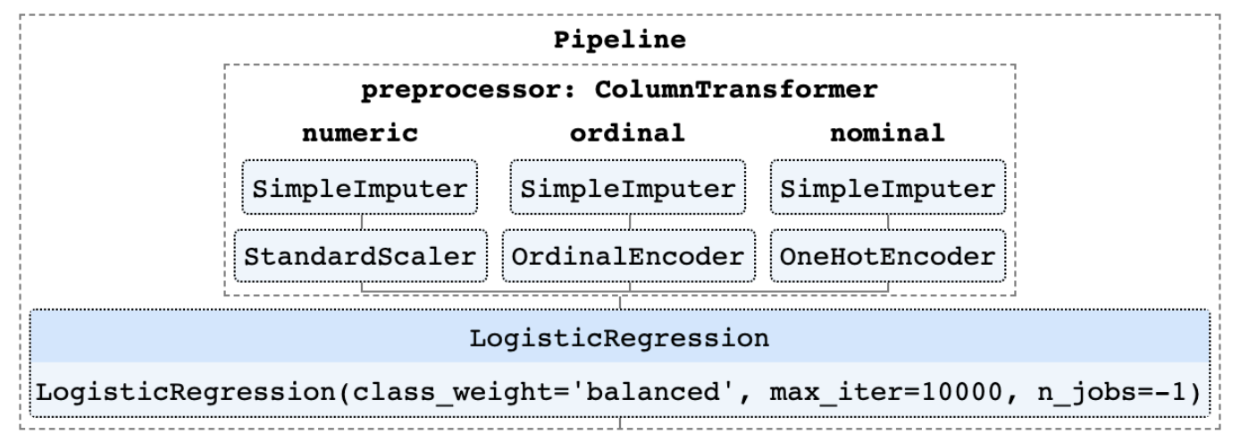
36 month Home Ownership Distribution



36 month Employment Length Distribution

Baseline Analysis

The data was randomly split 70% test and 30% training. The data was preprocessed through the pipeline below.



```
Pipeline
preprocessor: ColumnTransformer
numeric          ordinal          nominal
SimpleImputer    SimpleImputer    SimpleImputer
StandardScaler   OrdinalEncoder   OneHotEncoder
LogisticRegression
LogisticRegression(class_weight='balanced', max_iter=10000, n_jobs=-1)
```

Logistic Regression was used to train and classify the data. It is one of the most widely used models for credit risk classification. The features used were Loan Amount, Interest Rate, Installment, Sub-Grade, Employment Length, Home Ownership, Annual Income, Loan Purpose, DTI, and FICO Score, predicting Loan Status.

Python and the Scikit-Learn pckage were used to build the model, and hyperparameters were tuned using a 5-fold crossvalidated gridsearch. For Logistic Regression, these parameters are C, Solver, Penalty, and Tolerance. C is the strength of regularization to prevent overfitting. The solver is the optimization algorithm, SAGA and LBFGS were both used. The penalty specifies the norm to be used in penalization, L1 (LASSO) and L2 (RIDGE) were both used. The tolerance determines when the algorithm stops iteration.

Textual Analysis

The sentiment was extracted using the Loughran-McDonald Sentiment Word list, a corpus built for analyzing financial documents. There are six categories coded: Positive, Negative, Uncertain, Litigious, Strong Modal, and Weak Modal. Because text data is always messy, it has to be cleaned first. Unnecessary formatting was removed, such as the prefix added when the borrower edited their description, or non-alphabetic characters. Words were lemmatized to reduce them to their base word, such that "ran" and "running" would both be replaced with "run". Stop words, or words that don't contribute additional semantic meaning, are removed. The cleansing process is shown on an example text passage below. The original passage is then used to calculate a readability score, which is used in addition to the sentiment. These features are added to the original model to compare the accuracy.

Before Cleansing

Borrower added on 02/28/14 > This is to consolidate unexpected debt incurred while helping family with a medical emergency. This loan is so I won't have any balances on credit cards and I can pay it off in two years. I intend to pay $1,000 a month. I have never been late with any bills and will have no problem with this.<br>

Reading Score: 8.0

🟥 Removed with pattern
🟩 Stop word
🟦 Lemmatized

After Cleansing

consolidate unexpected debt incurred helping
Negative & Uncertain

family medical emergency loan balance credit card

pay two year intend pay month never late bill problem
Strong Modal Negative Negative

Results & Discussion

The results reported below are averaged over 100 trials, and report the Precision, Recall, F1-Score, Support, and the best model hyperparameters. Bold numbers indicate the better performing model between the baseline and textual analysis, and underlined numbers indicate that the result is statistically significant. Alpha = 0.05

	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%**			

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 sample size. It was almost three times larger than the 60 month dataset.

The classes are far from linearly separable, there were no features that clearly separated defaulted and fully paid loans. A more thorough extraction of sentiment may have lead to a larger improvement in performance. Analysis was performed using Bag-of-Words, which inheritly creates an oversimplified model. A Parts-of-Speech analysis might have extracted more information, as it separates sentences into their corrsponding parts (nouns, verbs, adjectives, etc.) and retains important context.

Conclusion

The objective of this research 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. 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.

Acknowledgements

I want to thank my advisor, Professor Haiyan Cheng, for her great help and encouragement, and Professor Wei Wu for his help with the financial background and research. Thank you to the Collins Fund for funding this research, and allowing me to contribute to the scientific community and learn through doing.

For a full, more complete picture of my research and the full analysis, please read my complete thesis (zach@haroian.net).