Summer Invitation Datathon 2022

Team 8

July 2022

1. Executive Summary

1.1. Problem Statement

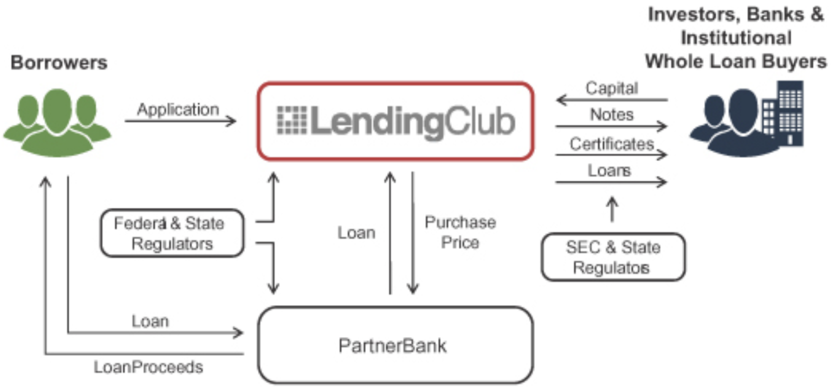


Figure 1: *Lending Club Business Model*

Since its first launch in 2007 to becoming the world's largest p2p lending platform in 2014, Lending Club has always made its mission to "provide financial health to Americans through leveraging technology and a marketplace model to seamlessly deliver access to fair and affordable credit"[1]. Starting from May 2017, Lending Club started offering hardship plans to borrowers which allowed for interest-rate only payments in light of unexpected life events, thus, providing additional flexibility to borrowers while securing returns for investors to otherwise charge-off loans[2].

There are numerous causes for hardship submitted by applicants. Nevertheless, a majority of them can be explained by three factors: macroeconomic condition, demographic distribution, and ex-ante credit-worthiness. As such, we seek to explore the relative importance of three factors in relation to hardship status. Specifically, we post two research questions.

Problem 1: Which features contribute more than others in identifying hardship conditions?
Problem 2: Can we predict whether a given individual should be in the hardship program?

1.2. Key Findings

We proceed to answer our questions by framing it into a binary classification problem. We took the accepted debtor dataset and expand it through years, so the expanded dataset is indexed by individual and year. Then, we select and engineered three groups of features: macroeconomic condition factors, demographic distribution, and ex-ante credit-worthiness. This would be the feature of our binary classification model, and the labels are whether a debtor at a specific year enters the hardship plan. After building two "good" models that reached certain out-of-sample accuracy, we would like to evaluate the feature importance in the two models. The two models are SVM and Random Forest. We computed feature importance using mean decrease in impurity and permutation importance metric for Support Vector Machine (SVM) and Random Forests, and developed a final ranking system.

We found the following five most important features that associate to debtor's hardship status

(a) Debt to income ratio (DTI): Demographic Feature

(b) Total Revolving Balance/Credit Limit (TRBL): Credit Feature

(c) County Unemployment Insurance Benefit amount (UIB): Inter-Temporal Economics Feature

(d) Annual Income (AI): Demographic Feature

(e) Number of open accounts/credit lines (OA): Credit Feature

At the end, we characterized three types of hardship debtors using above five features:

(a) Living in county where unemployment insurance benefit issuance is high and possess high debt-to-income ratio

(b) High Debt-to-income ratio, high number of open credit accounts and low revolving credit

(c) High number of open credit accounts and low income

2. Technical Exposition

2.1. Data Pre-Processing

2.1.1 Zip Code information

We noticed that the first three digits of the zip code accurately reflect the county where the debtor is located. We seek to convert these zip codes to Federal Information Process Standards (FIPS) code[3], which also represents the county - this allows us to better search for the economic and demographic information.

2.1.2 Inter-temporal Economic data selection

We seek to select representative economic data to study the macroeconomic trends, including and beyond GDP. This county-wise information is obtained from the Bureau of Economic Analysis based on ZIP code information of each debtor. Among all economic metrics, we found three particularly interesting metrics: GDP, unemployment insurance, and dividend rent income. debtor. Unemployment insurance could potentially be an extended indicator for measuring economic performance [4].

2.1.3 Demographic information selection

We also hope to explore the explainability of demographic information on initiating a hardship plan. Thus, we extracted county wise demographic data, such as gender and races from the U.S. census government site [5].

2.1.4 Ex-ante credit-worthiness features

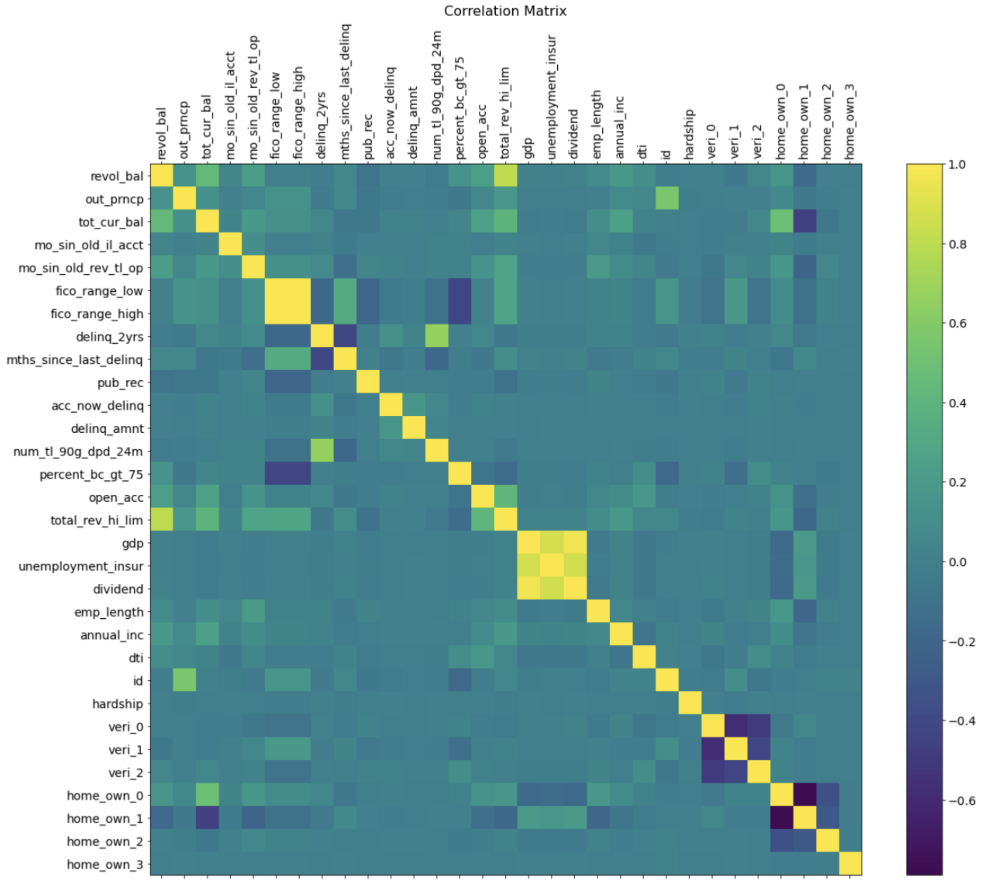


Figure 2: *Correlation Between Credit-worthiness Features*

We want to select contractual data that represents the credit history of debtors. Specifically, we hope to remove certain features that are highly correlated with each other. We give out a short example above: FICO_range_low and FICO_range_high indicate the lower and higher boundary for the FICO score for the debtor. However it is not of particular surprise that they are highly correlated–people usually have a similar range of credit scores. Therefore, we might only consider one of the data, or rather engineer an additional feature as the range. In essence, we want to select information that contains some unique information.

We, after analyzing the correlation matrix of our feature, selected 25 relatively uncorrelated features to represent credit information of debtors.

2.1.5 Adjustment to time

We took several measures to tailor our data to better answer our research question.

Expand each sample: To incorporate annual economic metrics into the data set, we expand each debtor data across the span of five years given the length of our data set, with an additional label indicating year. Therefore, the data points are indexed by each debtor at each year.

Hardship tags: We noticed that the hardship plan only starts in 2017 the earliest. Therefore, we only explore data entries in 2017 and 2018.

2.1.6 Categorical Variable encoding

One-hot encoding: To address categorical variables, we choose one-hot encoding to avoid introducing unnecessary bias in the encoding process, such as ranks, to maintain the ordinal relationship between classes.

2.1.7 Dealing with unavailable data

Among the features we selected, percent_bc_get_75 and mo_sin_old_il_acct are missing roughly 2% of input. For percentage of all bankcard accounts > 75% of limit, we chose to fill missing value with average. For months since oldest bank installment account opened, we chose to drop the missing feature, since in reality, these people don't have a bank installment account opened and it is not sensible to fill in a random value.

2.1.8 Normalization

We seek to normalize our data by z-score. We would want to keep our data on a similar scale for the purpose of clustering.

2.2. Data Downsampling

2.2.1 Reason for Data Downsampling

We found there is a need to downsample our data, mainly for two reasons: Expensive algorithms and Imbalanced data.

An initial exploration shows that only 10816 data points are labeled with 'hardship' and around 4 million data points are labeled with 'non-hardship'. The ratio of hardship debtors versus non-hardship debtors is somewhere around 200:1. This introduces a severe imbalanced problem in our dataset - an trivial majority vote algorithm would achieve 99.5% accuracy. To tackle this imbalanced dataset issue, the first model we considered was class-weighted SVM. However, the SVM implementation in sklearn is both computationally expensive, with a complexity of O(n^3), and memory heavy, a dataset with 100K data using SVM will take up roughly 80G of memory[6]. This enormously large dataset holds us back from employing cost-sensitive classification directly on the original dataset. Therefore, we seek to reduce our dataset through downsampling.

2.2.2 Downsampling methods

We explored various down sampling methods and came up with the two following approaches that best suited our dataset: K-medoid Downsampling and Tomek Link Downsampling.

The gist of our approach is to first perform clustering on our dataset, then based on the clusters formed, randomly sample a selected number of data points proportional to the number of elements in each cluster. This process can be viewed as a special stratified sampling method is an optimal choice over random sampling, since under a majority of conditions, stratified sampling yields lower variance with equal expectation. Our assumption was further confirmed when we discovered that our predictive model when trained with downsampled data drawn with clustering is able to yield a higher test accuracy for after training.

We employed K-medoid clustering since it utilizes the L_1 norm for its objective function which makes it a robust algorithm against outliers, a good choice since our dataset contains 99 features and most features contain significant outliers.

Specifically, K-medoid is defined as follows:

Let C^i be center of clusters. We have our objective function as

$$Cost(C^1, ..., C^k, z^{(1)}, ..., z^{(k)}) = \sum_{j=1}^{k} \sum_{i \in C^j} d(x^{(i)}, z^{(j)})$$

the algorithm:

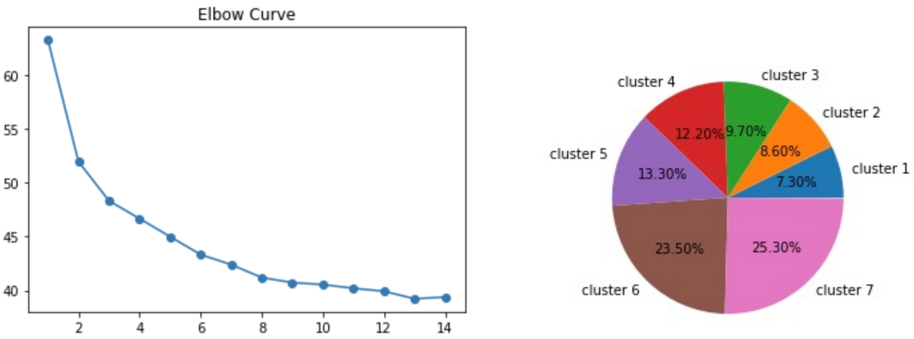
1. Initialized $z^{(1)}, ..., z^{(k)}$ as a subset of $x^{(1)}, ..., x^{(n)}$, where $x^{(i)}$ are our data points.
2. Define

(a) $d(x^{(i)}, z^{(j)}) = \sum_{k=1}^{M} |x_k^{(i)} - z_k^{(j)}|$

(b) $C^j = \{i | x^{(i)}\text{'s closest center is } z^{(j)}\}$

3. Adjust $z^{(i)}$ each time to minimize objective function. Repeat until cost converges to 0.[7]

Next, to obtain the most representative downsampled data through clustering, we used the elbow curve method to determine the optimal number of clusters to sample from. The K-medoid algorithm is run with clusters from size 1 to 14 with their relative fit score evaluated. Here, 7 was picked as our final cluster size.



(a) Elbow Curve

(b) Cluster Proportion

Figure 3: *Result From K-medoid Clustering*

With the cluster number determined, we start drawing random sample points from each cluster. The number of points drawn from each cluster is sampled proportionally to maximally preserve the characteristics of our large dataset.

Another problem we noticed is that we have a lot of boundary variables - those that are spatially closed while of different classes. We hope to further downsample the majority (non-hardship) entries using Tomek Link Downsampling. We define a Tomek Link as follows:

- x is the nearest neighbor of y
- y is the nearest neighbor of x
- x, y is of different class

After we find the Tomek Link, we seek to remove the majority class in this pair, because the majority datapoint is an ambiguous point, i.e. it is either a noisy or boundary instance. Iteratively find and break Tomek Link, we can draw a gain class boundary between two classes.

2.3. Evaluation of Feature importance

To capture features that explains a debtor's hardship status, we propose to analyse feature importance through binary classification and using importance measures include permutation importance and gini importance.

We implemented two models to solve the binary classification problem. The 99 features were divided into three sets of features, macroeconomic condition, demographic distribution, and ex-ante credit-worthiness. Before feeding the data to our models, we deployed Principal Component Analysis (PCA) as initial exploration to reduce dimensions and noise of each set of features separately. It was not surprising to find that in all three groups of features, only one principle component can explain nearly or more than 95% of variance. This thus provides us a good foundation to visualize how separable are the data points are and the label's joint distribution with the three groups of features.

After initial exploration, we fed the data after dimension reduction into classification models and got unsatisfying results (accuracy on training and testing dataset are both $< 55\%$). Therefore, we moved forward to implement two models that are fed with the complete dataset. The first model is Random Forest, which has a natural impurity-based feature importance measure. The second model is Support Vector Machine (SVM) with Radial Basis Function kernel. Both models reached an accuracy of closed to 70% on testing dataset after hyperparameter tuning through cross-validation. We used permutation-importance to measure the feature importance in this model.

2.3.1 Initial data exploration through PCA

Since our feature sets contain both categorical and continuous variable, and PCA doesn't work for categorical data, we simply dropped the the categorical features to proceed our analysis (there is actually an extension of PCA on categorical features called Multiple Correspondence Analysis, but we would like to focus on the the rough analysis of the three groups of features).

We found that in each group of features, only one principle component can explained close to 95% of variance in the dataset, so we reduce each group to a 1-D feature. The resulting dataset only contains three features representing features from each of the three groups of factors. This provides a good foundation for us to visualize how data points are distributed. Figure 4 plots the distribution of samples after PCA, where the two colors means different classes of the sample. It looks like the samples are naturally divided into several clusters by macro econ factors. However, the distribution of class labels has no strong connection with those clusters. Moreover, the samples are not linearly seperable from the graph. Therefore, we tried a Random Forest model on this dataset and reached a testing accuracy of a bit over 50%. The performance is close to random guess, so we decided to turn to use all features to build prediction models.

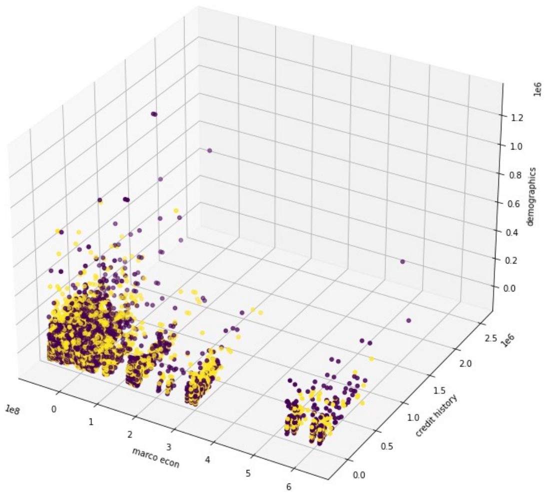


Figure 4: *Data Distribution After PCA*

2.3.2 Feature importance from random forests

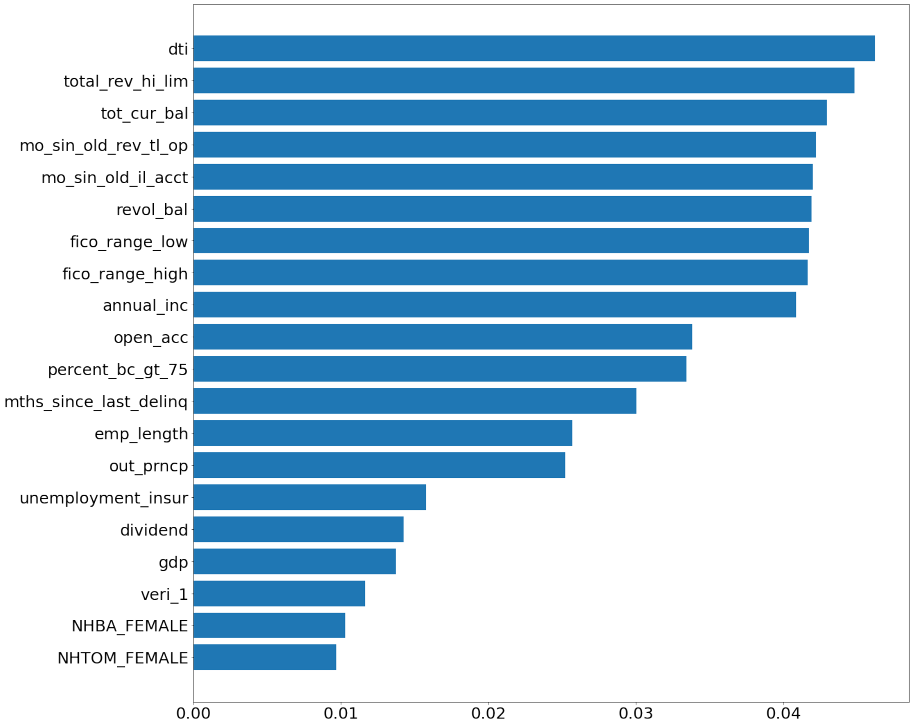


Figure 5: *Feature Importance From Random Forests*

Feature importance for random forest is ranked based on mean and decrease in impurity accumulated across trees. Mathematically, we used the following metric:

For each feature x, we define

S_x : The number of times that this feature is used to split the node

N_x : Number of total samples that this feature is splitted on

The score metric is then

$$\frac{S_x}{N_x}$$

The top 20 features for Random Forest are presented in Figure 5. Notably, features related to credit history such as the borrower's total revolving high credit and total current balance are among the essential features in explaining whether an individual is in the hardship plan, which matches with our expectation. However, credit history can only explain a portion of the reason behind why an individual is participating in the hardship plan, macroeconomic features such as GDP and unemployment insurance benefits by county together with demographic features such as debt-to-income ratio and female population percentage of certain races provides additional explanation to the situation.

Nevertheless, there still exists a fallacy behind random forest's approach in finding feature importance, being favorable towards continuous variables, resulting in bias towards categorical features such as house ownership and income verification[8]. As a result, we decide to perform additional comparison methods in determining feature importance.

2.3.3 Feature importance from Support Vector Machine

SVM was the first model that came to our mind before we dealing with the unbalanced dataset problem because of its ability to tackle this problem (only support vectors would affect the decision boundary). After reducing our dataset to a balanced dataset, we still deploy an SVM classifier since it works well with high dimensional data (naturally contains a regularization term in its loss function) and our features does preserve high-dimensionality. Moreover, we found our dataset contains a lot of outliers, and SVM does retain robustness when dealing with outliers. We ran the SVM with RBF kernel and turned the only hyperparameter(constant before the L2 norm term in loss function) by cross-validation. We used 5-fold crovss-validation on the training dataset, where the all the data is splitted into training (70%), validation(10%), and testing (20%). Due to the lack of a natural feature importance measure for SVM with RBF kernel, we adopted the method of permutation importance to measure the feature importance.

The top 20 features for the SVM are presented in Figure 6. The top 20 features only include demographic factors and credit history. The majority of the features are demographic data about the race component in that we extracted from the first three digits of zip code. Taking a quick look at the features, we found that most of them are percentage of minorities like Hispanic and Havaiian population. This means that the percentage of minorities in the a lender's living area provides insights on whether or not he or she enters the hardship plan. Other important features include annual_inc, emp_length, and tot_cur_bal, which provide an estimate of a lender's financial health. This complies with our expectation because those features does directly provide information regarding lenders' financial situations.

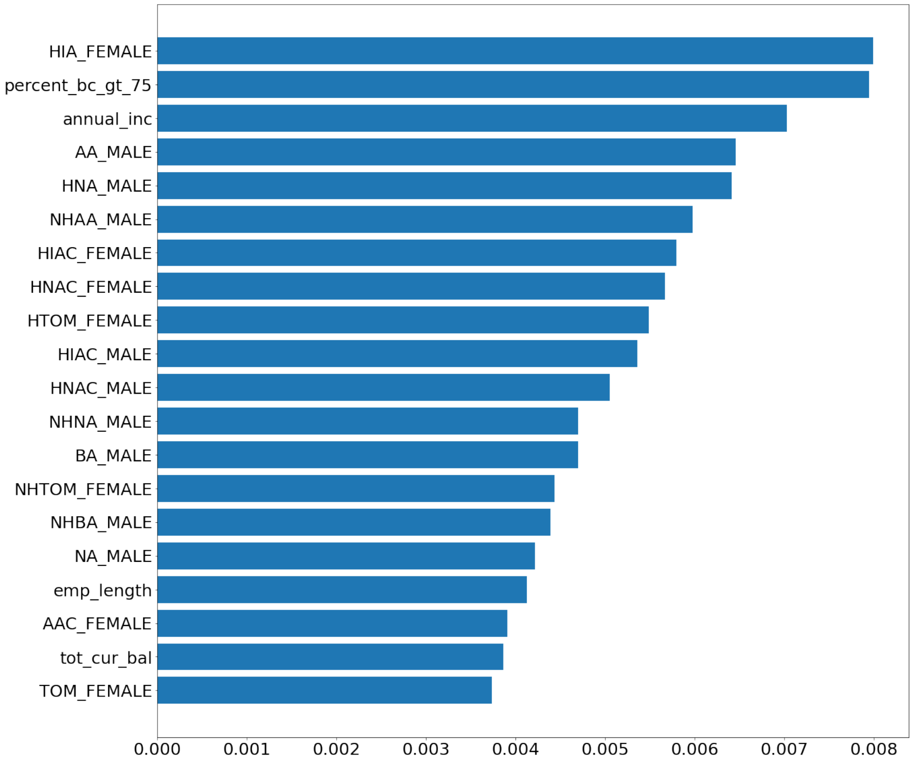


Figure 6: *Feature Importance from Support Vector Machines*

Remark. We have tested the accuracy of the trained SVM and Random Forest model on samples in the original dataset that are not selected in the downsampling process as a out-of-sample check. The results was consistent with a mean accuracy of 68.55%.

2.4. Analysis of important features

In this section, we seek to analyze the first five most important features by comparing the feature distribution between those who entered hardship loans versus those who did not. At the end, we seek to produce a decision tree classifier that captures the decision boundary of our algorithm, which offers a clear mathematically defined boundary for us to view.

2.4.1 Selecting top 5 most important features

From the two feature importance rankings, we seek to produce a metric that calculates the final ranking. We define our scoring metric as follows:

(a) Select the top 30 important features from both rankings based on percentage of explained ratio of variance, our importance metric.

(b) Use min-max standardization to standardize the importance metric of each feature on both rankings (reference picture here) such that the importance metrics are on the same scale.

(c) Sum up the importance metric for each feature from both rankings since the same feature may appear on both rankings, then sort features based on their new importance metric, thus obtaining our final ranking.

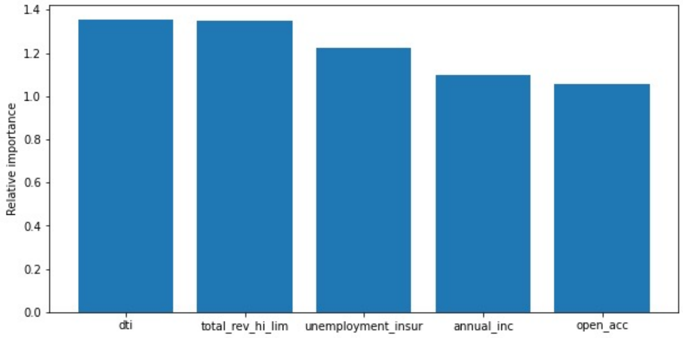


Figure 7: *Final Feature Importance*

2.4.2 Debt-to-income ratio (DTI)

According to our metric, DTI is the most important feature across the total of 99 features. We think of DTI as a demographic feature that characterizes the personal information regarding a specific debtor.

We noticed that DTI for those entered hardship is generally more condensedly distributed around a higher median. Revealing that for people participating in hardship plan, they generally suffer from higher debt pressure than non-hardship plan individuals.

2.4.3 Total revolving high credit / credit limit (TRHL)

Total revolving high credit is a very important credit-related feature. A debtor's credit limit could be an informative indicator of financial well-being. Looking at the total revolving high credit distribution, we observed that those entered hardship plans are more densely populated in the low-credit limit distribution. Boushey and Christian also has a similar finding that credit, along with economical factors, could be influential in determining hardship [9], which cross-validated our idea.

2.4.4 Unemployment Insurance Benefits by county (UIB)

UIB amount by county, in thousands of dollars, is an important economic indicator in our dataset. From the histogram below, we can observe for individuals in the hardship plan, the county they reside in provides lower unemployment insurance than the county of those not inside the hardship plan. Therefore, it is recommendable for local government to provide higher unemployment insurance to secure their resident's financial resilience in face of unexpected life events.

2.4.5 Annual income (AI)

AI is a demographic feature from the original Lending Club dataset. Those with high and steady income are much more likely to survive from Black Swan Events that results in financial hardship.

2.4.6 Open Account (OA)

Number of credit lines is an indicator of credit history for debtors. It is understood that those with more credit accounts opened might have more financial hardship problems - they need more money and hardship loans to increase personal liquidity.

2.5. Forward looking prediction

We hope to conclude our report by offering predictive model that determines whether a debtor, given his demographic, credit, and inter-temporal economical data, would be likely to encounter financial hardship. We train a decision tree algorithm based on the five most important feature selected.

We found some very interesting decision rules described by the tree output. For example, most upper nodes are split based on the debtor's county unemployment insurance benefit. The root splits on whether the county-wise unemployment insurance benefit is greater or less than 13,189,000 dollars. This resonates with our analysis in the previous section, where unemployment insurance benefit by region sheds light on a debtor's hardship status.

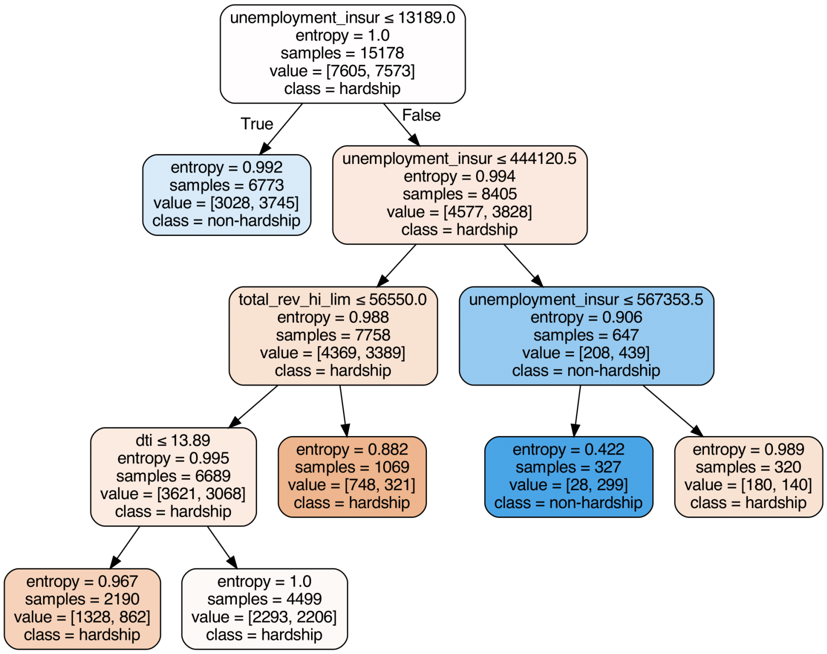


Figure 8: *One Branch Of Decision Tree Visualization*

We present one branch of our tree above. From our tree output, we summarized three types of a hardship debtor in terms of our five important features

(a) Living in county where unemployment insurance benefit issuance is high and possess high debt-to-income ratio

(b) High Debt-to-income ratio, high number of open credit accounts and low revolving credit

(c) High number of open credit accounts and low income

We Further iteratively remove the root feature and see the which feature does the root split on, knowing that first split select the most information-gain among all other features. We get the following order:

Unemployment Benefit → DTI → Total Revolving Balance→ Open Account → Annual Income

It is very interesting that most seemingly important feature - annual income - comes to last for information gain.

We hope our result of analysis could shed some lights on the properties of hardship loan debtors. It is unfortunate that we can only obtain the county information of the debtor. With a 5-digit zip-code, we will be able to characterize a debtor in a much more detailed and fruitful manner, thus drawing more significant and precise conclusion on the hardship status.

A final remark is that despite most hardships are caused by random Black Swan Event (see hardship application reason feature), those with particular macro-economical, credit-history, demographic features are more probable to be in financial hardship. An purposeful result of our study is that policy-makers could closely monitor the five features we selected to locate debtors who are more likely in need of hardship loans.

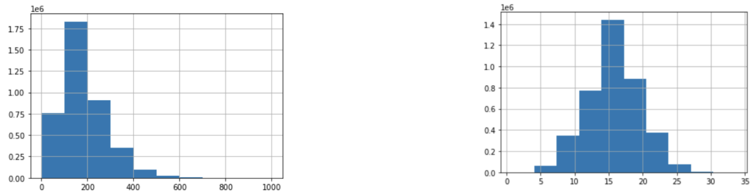
2.6. Hypothesis Testing and Cross Validation of Selected Features

Among selected Features, we want to cross-validate the fact that the features draw statistical significant implication to hardship. In the following we address two testing methods

For the purpose of testing , we hope to transform our dataset into approximately normally distributed. Therefore, we deploy Yeo Johnson Transformation to our data series, an exemplar result deployed on feature

$$y = \begin{cases} \frac{(x+1)^{\lambda}-1)}{\lambda} & \text{for} \quad x \geq 0, \quad \lambda \neq 0 \\ \log(x+1) & \text{for } x >= 0, \quad \lambda = 0 \\ -\frac{(-x+1)^{(2-\lambda)}-1}{(2-\lambda)} & \text{for } x < 0, \quad \lambda \neq 2 \\ -\log(-x+1) & \text{for} \quad x < 0, \quad \lambda = 2 \end{cases}$$

Where y is the transformed series, x is the original data series, and λ is the parameter that maximizes the MLE.



(a) mo_sin_old_rev_tl_op before norm (b) mo_sin_old_rev_tl_op after norm

Figure 9: *Z-Score Normalization*

2.6.1 Difference in sample mean

To test, we want to iteratively conduct the following z-test on each feature

For each feature x, we define $E_h[x]$ to be the mean of feature x for those who enter hardship plan, $E_{nh}[x]$ to be the mean of feature x for those who do not enter hardship plan. Mathematically, we consider the following hypothesis testing:

$$H_0 : E_h[x] - E_{nh}[x] = 0$$
$$H_a : E_h[x] - E_{nh}[x] \neq 0$$

We hope to give particular notice to the important features that are also successfully rejecting the above hypothesis testing using standard z-testing with following z-stats.

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{x}$ is the mean of sample data, and population variance is approximated by sample variance.

2.6.2 Problem of multiple testing and solution: Bonferroni Correction

Note that as we iteratively test through our data, we are subject to the problem of multiple testing. For example, with an α level (false positive probability) of 0.05, a set of $10,000$ hypothesis testing is guaranteed to have 500 significant results due to the false positive probability. In order to keep the same false positive probability (alpha) level for the whole set of testing, we deploy a Bonferroni Correction.

Bonferroni Correction refers to adjusting the new p-value rejection threshold to maintain the original desired alpha level (0.05). The new threshold is as follow:

$$\frac{\alpha}{m}$$

where α is the original false positive probability and m is the number of tests in total.

2.6.3 Chi-Square Goodness-of-Fit Test and Contingency table

Notice that our features contain categorical data. It is quite important that we employ another testing for our selected categorical table. We produce the following hypothesis testing

H_0 : there is no association between type of variable x and whether one gets the hardship

H_a : there is association

with following test statistics

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - E[\hat{Y}_{ij}])^2}{E[\hat{Y}_{ij}]} \sim X^2_{(r-1)(c-1)}$$

2.6.4 Validation Testing Result

Conducting above testing methods on the selected features, we observed that we are able to reject the null that there is no significant difference/association for all selected important features under the new corrected alpha level.

2.7. Limitations

2.7.1 Rejected dataset

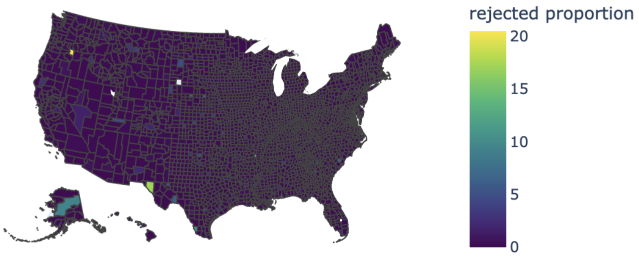


Figure 10: *Number of people rejected in each county, normalized by county population*

Looking at the demographics of the debtors who are in the rejected dataset, we found some debtors from certain regions are rejected more frequently than others, indicated by the choropleth map. We want to note that Lending Club rejects these debtors by a hard debt-to-income ratio and FICO threshold; therefore, the rejection dataset is not particularly fruitful. We felt pity that we did not have an opportunity to comprehensively study why people from these regions are more frequently rejected due to a lack in data described above.

2.7.2 Clustering method

Due to the existence of categorical features in our dataset such as homeownership and verification status, we should employ the K-prototype clustering algorithm proposed by Huang which enables clustering on mixed data instead of K-medoid method[10]. The K-prototype clustering algorithm is essentially a weighted sum of the K-mean and K-mode clustering method applied separately to numerical and categorical data.

Nevertheless, due to a limitation of computational power and time to perform K-prototype clustering on 4 million data points without GPU boosting, we are forced but to drop the two categorical features from our dataset.

3. Conclusion

For the three metrics we defined in the problem statement (Inter-temporal Economics, Demographic, and Credit), we were able to find features from these three metrics that show significant association to debtor's hardship status using both model feature selection and statistical testing. We further concluded important characteristics of hardship debtors through observing trained decision tree splitting rules based on information gain.

Our research finding pointed to certain changes that could bring about greater financial stability for members of our society. Policy-makers can provide additional social safety net to low-income and unemployed people, securing greater financial resilience for individuals in face of unexpected life events. Also, Policy-makers should pay additional attention towards underrepresented groups since our data show peak in hardship rates for these population.

Our predictive model with SVM gives a 69% training accuracy against test data, validating our model's ability to predict whether a person should be included in the hardship program given their status. With this knowledge, Lending Club may suggests loan applicants who fit the hardship condition to participate in the hardship program to relief financial stress of well-credit individuals in hardship condition.

Furthermore, if allowed with more time and resources, we could definitely tryout different feature importance metrics and computational expensive model like K-prototypes, which may yield more accurate results to our tasks.

References

[1] LendingClub Corporation Form 10-K. Document. (n.d.). Retrieved July 24, 2022, from `https://www.sec.gov/Archives/edgar/data/0001409970/000140997019000222/a201810-k.html`.

[2] Staff, deB. (2017, April 5). Lending club to beta hardship plans for borrowers (and protect returns for investors). deBanked. Retrieved July 24, 2022, from `https://debanked.com/2017/04/lending-club-to-beta-hardship-plans-for-borrowers-and-protect-returns-for-investors/`.

[3] Federal Communications Commission, `transition.fcc.gov/oet/info/maps/census/fips/fips.txt`.

[4] Stiglitz, Joseph E., et al. Beyond GDP Measuring What Counts for Economic and Social Performance. OECD Publishing, 2018.

[5] Bureau, U.S. Census. Explore Census Data, `data.census.gov/cedsci/`.

[6] Ns, Abdiansah. (2015). Time complexity analysis of support vector machines (SVM) in LibSVM. International Journal of Computer Applications. 128. 975-8887. 10.5120/ijca2015906480.

[7] MIT. 6.S064 Introduction to Machine Learning Phase 1: Clustering (Lecture 6). `piazza.com/class_profile/get_resource/jsqbxtsjtmsrp/jt7qpiwaqy24px`.

[8] Lewinson, Eryk. "Explaining Feature Importance by Example of a Random Forest." Medium, Towards Data Science, 26 Aug. 2021, `towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e`.

[9] Boushey, Heather, and Christian E. Weller. "Inequality and Household Economic Hardship in the United States of America." United Nation, Apr. 2006, `www.un.org/esa/desa/papers/2006/wp18_2006.pdf`.

[10] Huang, ZHEXUE. "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values." Data Mining and Knowledge Discovery, 1998, `citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&amp;type=pdf&amp;doi=10.1.1.227.696`.

A. File Organization

In the folder submitted via google drive, we have all data in `./data`, we put all our generated data in `./data/processed_data/` and we imported new outside data in `./data/demographics/` `./data/gdp/` and `./data/location_service/`.

The code are in `./src/`