Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods

Yiyun Liang (isaliang@stanford.edu)
Xiaomeng Jin (tracyjxm@stanford.edu)
Zihan Wang (wangzih@stanford.edu)

***Abstract*— Evaluating and predicting the repayment ability of the loaners is important for the banks to minimize the risk of loan payment default. By this reason, there is a system created by the banks to process the loan request based on the loaners' status, such as employment status, credit history, etc.. However, the current existing evaluation system might not be appropriate to evaluate some loaners repayment ability, such as students or people without credit histories. In order to properly assess the repayment ability of all groups of people, we trained various machine learning models on a Kaggle dataset, *Home Credit Default Risk*, and evaluated the importance of all the features used. Then, based on the importance score of the features, we analyze and select the most identifiable features to predict the repayment ability of the loaner.**

I. Introduction

Due to insufficient credit histories, many people are struggling to get loans from trustworthy sources, such as banks. These people are normally students or unemployed adults, who might not have enough knowledge to justify the credibility of the unidentified lenders. The untrustworthy lenders can take advantage of these borrowers by taking high interest rates or including hidden terms in the contract. Instead of evaluating the borrower based on their credit score, there are many other alternative ways to measure or predict their repayment abilities. For example, employment can be a big factor to affect the person's repayment ability since an employed adult has more stable incomes and cash flow. Some other factors, such as real estates, marriage status and the city of residence, might also be useful in the study of the repayment ability. Therefore, in our project, we are planning to use machine learning algorithms to study the correlations between borrower status and their repayment ability.

We found the dataset, Home Credit Default Risk from Kaggle.com, to be used in this project [1]. This open dataset contains 308K anonymous clients' with 122 unique features. By studying the correlation between these features and repayment ability of the clients, our algorithm can help lenders evaluate borrowers from more dimensions and can also help borrowers, especially those who do not have sufficient credit histories, to find credible loaner, leading to a win-win situation. [1]

[1]Code available at: https://github.com/Yiyun-Liang/loanliness

II. Related Work and Background

A. Loan Repayment Ability Prediction

In the lending industry, the lenders normally evaluate the repayment ability of the loaners and the risks of lending money to them. Based on the the repayment ability and risks, the lenders, especially the banks, can adjust the interest rates of the loans which are issued to the borrowers [2].

The research on the evaluation of repayment ability has been conducted for decades. Some of the research focus on finding useful metrics to quantitatively evaluate the repayment ability of the loaner, such as the residual income ratio and credit score [3] [4] [5]. Others target on finding the repayment ability of a group of people with similar status, such as students and farmers [6] [7] [8] [9] [10].

Furthermore, the financial crisis in 2008 made an impact on the repayment evaluation process. The term "ability to repay" was used in the book, *Dodd-Frank Wall Street Reform and Consumer Protection Act*, in 2010, which is used to describe one's financial capacity to make the payment to the debt [11]. It has been an requirement for a mortgage after the mortgage crisis in 2008. Before the financial crisis, the ability to repayment is not a hard requirement for the lenders to provide loans to the borrowers. The loaners, especially homebuyers, can get loans from the banks even their monthly income might not be able to cover the monthly mortgage payments [12]. In order to prevent and reduce the default rate of the loan payment, the Consumer Financial Protection Bureau (CFPB) came up a new set of rules and regulations to evaluate the ability to repay of the loaner. These rules and regulations include the loaner's [13]:

- Expected income or assets
- Employment status
- Expected monthly payment
- Monthly payment on the simultaneous loans
- Monthly payment of mortgage
- Current debt status
- Residual income
- Credit history

These factors become the rule of thumb to evaluate lenders' ability to repay.

However, these ability-to-repay rules might not fit for evaluating some types loaners. For example, the university students might not satisfy the rules to get a loan from

trustworthy resources, such as banks, since they are not employed and they have very limited credit history. Therefore, the untrustworthy lenders might take advantages of them. In order to prevent this to happen, the objective of our project is to discover more identifiable and useful features to evaluate the credibility and repayment ability of the loaners. Furthermore, we trained and tested machine learning models based on the features and find the best model to predict the repayment ability of the loaners.

III. Dataset and Challenges

A. Problems with existing models

We inspected the dataset and found that many entries contains invalid values such as nan(not a number). There are also three features 'EXT_SOURCE_1', 'EXT_SOURCE_2', and 'EXT_SOURCE_3' and we do not know what they represent. Existing models are evaluated using the three features and by removing the invalid values. These assumptions do not necessarily make sense.

Another problem with many existing models is that these models are not trained on a balanced dataset, so when making predictions, the model tends to achieve high accuracy by predicting all data with the majority label. This high accuracy does not tell us much since the accuracy is equivalent to the proportion of majority class data in the test set.

B. Challenges with the dataset

One challenge that rises in many finance-related machine learning problems is that the dataset is heavily imbalanced. Our dataset records borrower profiles and binary truth labels, the decision whether they should be accepted as a client by the lender. In reality, since only a small fraction of the loan applicants are eventually accepted, our dataset also suffers from the problem of being imbalanced.

The dataset we get from Kaggle is relatively large in terms of number of features as well as the amount of data(around 300k), so the training process is quite slow, especially when we build and apply more sophisticated machine learning models.

The two challenges above make the problem more interesting because they are problems frequently encountered by researchers and scientists.

IV. Methods

The goal of the project is to predict the repayment ability of the borrowers based on the factors other than the credit history. It can be framed as a classification problem with two classes. In the following section, we will introduce the methods we used to pre-process the data and the machine learning algorithms we will use to solve the problem.

A. Data Pre-processing

Due to the complexity of our raw data, we introduce some data pre-processing techniques to our dataset before it is used for training and testing.

- **Feature concatenation:** In the original data set, the features come from different sources. A brief summary of data files is shown in Table 1. Our first step of data pre-processing is to concatenate all the features together. The way to combine all the features together is to use each individual borrower's unique ID number. For example, the entries of bureau.csv file can be joined with corresponding rows in application_train.csv using SK_ID_CURR. In this way, we concatenated all the features together to construct the training and testing sets with the maximal usage of the given data. After feature concatenation, each data point has 217 features in total.
- **Feature Encoding and Normalization:** Our features come in a variety of format, eg. sentence strings, unbounded integers, floating numbers in the range of 0 to 1, boolean values, etc. This poses a challenge for us as the features cannot be directly used for training. To prevent classification biases towards certain features, we factorize these features using label encoding, that is, we map the string values to categorical values, each represented by an integer. However, for some features, the number of categories is too large and it is difficult to apply label encoding. Then, for these types of features, we use one-hot encoding to expand the single feature into multiple features where each expanded feature has values limited to only 0 and 1. In the end, we also normalize the feature values so that all features are evaluated on the same scale.
- **Invalid/Empty Entry Replacement:** Besides feature processing, invalid entries and empty entries is another problem that prevent us from training the machine learning algorithms properly. In the dataset, there is a noticeable amount of data with invalid entries (such as a very large number) or empty entries (e.g. Nan). One strategy we applied is to take the mean of the feature values and fill in invalid entries with this mean value. For entries with a large amount of invalid values, we use a strategy to remove columns or rows based on the percentage of invalid values that present in the corresponding column. We set a threshold value, which is initially set to 30% in our case. If the percentage of invalid values in the column is greater than the threshold, we mark the feature as an invalid feature and remove the column from the dataset. Otherwise, we just remove the row which contains the invalid value.
- **Polynomial feature transformation:** To gain the most out of our linear classifiers, we also performed polynomial transformation on our feature values to include a polynomial combination of the features.

The dataset provides us with a very comprehensive profile of the loan applicants. In total, we expanded the number of features to around 651 features for each loan applicant, as shown in TABLE II. A large number of features is helpful in training the model but can sometimes slow down our algorithms. We will also expand on our experiment with feature reduction in the next section. We will also try to make sense of the features to ensure our assumptions of the

TABLE I
SUMMARY OF FILES AND RAW FEATURES

File Name	Description	# of Features
application_train.csv	Information about loan and loan applicant when they submit the application	121
bureau.csv	Application data from previous loans that client got from other institutions reported to Credit Bureau	17
bureau_balance.csv	Monthly balance of credits in Credit Bureau	3
previous_application.csv	Information about the previous loan and client information at previous time	37
POS_CASH_balance.csv	Monthly balance of client's previous loans in Home Credit	8
instalments_payments.csv	Previous payment data related to loans	8
credit_card_balance.csv	Monthly balance of client's previous credit card loans	23

TABLE II
COMPARISON BETWEEN BEFORE AND AFTER INVALID/EMPTY ENTRY REPLACEMENT

	# of Features	# of Datapoints
Before	217	307511
After	651	102244

dataset are valid and that the feature importance outcomes match our expectations.

B. Machine Learning Techniques

In this section, we try some machine learning models on the task of making predictions on borrower repayment abilities. The machine learning algorihtms include: logistic regression, random forests, Naive Bayes, LightGBM and neural networks. Some of the algorithms are the algorithms we learned from class and some are we explored online which might have a good performance on the dataset. In this section, we present the reasons why we choose to use these algorithms. We will show the results of our initial experiments in section 3.

- **Logistic Regression:** Logistic regression is often a great baseline model to try on machine learning problems because it does not enforce strong assumption on the distribution of the underlying data. For our classification problem, logistic regression is a great model to try as our first step.
- **Random Forests:** Random Forests usually performs well on imbalanced dataset. Another benefit of running a random forest classifier on our dataset is that it provides us with an intuitive way of looking at our features by listing individual feature importance, which gives us intuitions into the factors that affect a person's loan repayment ability. Moreover, we want to see if introducing some degree of randomness into the classification problem would help in improving accuracy of our results [14].
- **Naive Bayes:** Naive Bayes uses the "naive" assumption on the features, which means that the features are conditionally independent with each other given the class variable. As we learned in class, Naive Bayes is a generative method, which is different from the previous two algorithms. Then, we can compare the performance between the basic discriminative methods and this generative method.
- **LightGBM:** LightGBM is a highly efficient gradient boosting decision tree algorithm for classification [15]. It is an improved version of Gradient Boosting Decision Tree (GBDT) algorithm, which is over 20 time faster. GBDT is a machine learning algorithm wildly used in multi-class classification and click prediction. Since it is a augmented tree-structure classifier, we can make comparisons between itself and the Random Forest classifier to get a sense of differences in performance between standard algorithms and more advanced algorithms.
- **Neural Networks:** Neural networks, or multi-layer perceptrons, are one of the most popular methods to be applied on classification problems. It is a function approximator, which can not only model the distribution of linear data, but can also classify data with non-linear decision boundaries due to the non-linearity added by the activation functions. In our project, we fit a multi-layer perceptron by carefully selecting hyper-parameters, such as the number of layers and the number of neurons.

V. EXPERIMENTS AND RESULTS

A. Training and Test Data Split

By observing the data in the processed dataset, we found that the negative data and positive data are imbalanced. The number of data points with negative labels is much larger than the data points with positive labels. There are many potential issues and risks related to having an imbalanced dataset. If we train the classifier on an imbalanced dataset, the classifier may classify all the minority data with majority labels which can also result a high test accuracy. However, the classifier might be meaningless since it does not have the capacity to recognize data in the minority class. Therefore, in order to prevent this problem, we tried several techniques. They include balancing the dataset with up-sampling, down-sampling techniques, as well as using class weights to penalize predictions on the majority class.

- **Down-sampling:** Since the dataset is not balanced, we can use the down-sampling method to reduce the number of negative examples for training purposes. First, we take the minority examples, here they are the data points with positive labels. Then, we randomly select the same amount of data with the majority labels. However, it might also be ideal to process the dataset in different ways. The reason is that a large number of

TABLE III
PERFORMANCE OF MACHINE LEARNING ALGORITHMS

Machine Learning Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	69.34%	0.66/0.75	0.81/0.58	0.72/0.65
Random Forest	63.51%	0.58/1.00	1.00/0.27	0.73/0.43
Naive Bayes	52.11%	0.51/0.71	0.97/0.07	0.67/0.13
Multi-layer Perceptron	69.15%	0.67/0.71	0.73/0.65	0.70/0.68
LightGBM	57.47%	0.54/1.00	1.00/0.15	0.70/0.26

TABLE IV
PERFORMANCE OF K-MEANS CLUSTERING WITH CLASSIFICATION ALGORITHMS

Machine Learning Model	Accuracy	Precision	Recall	F1 Score
Cluster 1	72.24%	0.63/1.00	1.00/0.47	0.77/0.64
Cluster 2	67.01%	0.62/1.00	1.00/0.28	0.77/0.43
Cluster 3	82.34%	0.77/1.00	1.00/0.56	0.87/0.72
Cluster 4	70.78%	0.56/1.00	1.00/0.53	0.72/0.69
Overall	71.57%	0.63/1.00	1.00/0.43	0.77/0.59

negative data was removed from the dataset when we perform down-sampling, which is a waste of resources.

- **Up-sampling:** Another technique we explored is the use of up-sampling techniques to up-sample the minority data in the training set. We can up-sample positive data using techniques such as SMOTE [16]. By using up-sampling techniques, we can take advantage of the negative data more effectively.
- **Class weights:** We experimented with class weight, which penalizes the classifier when it predicts with the majority class label. The ratio of the class weight is calculated as the inverse to the frequency of the class label. The weight of class i is calculated below, where n is the total number of examples, and n_i is the number of examples in class i.

$$w_i = \frac{n}{2n_i} \tag{1}$$

B. Performance of Implemented Algorithms

The best performance comes from our down-sampled dataset. The performance of each algorithm is shown in TABLE III.

As we can see, the logistic regression classifier has the best performance on both accuracy, reaching a value of around 0.69, followed by the random forest and MLP algorithms, both of which has an accuracy of above 0.6.

We can also look at the precision and recall scores of our classifiers more closely. From TABLE III, we see that our models all yields a good precision score on the positive class (second column of precision). This precision score is calculated as the number of true positives over the sum of true positives and false positives. A high value tells us that our model is confident in its result in classifying a borrower as trustworthy. This aligns with our goal, where we want to provide another criteria to evaluate trustworthy borrowers who may not have enough credit scores.

The ROC curves and corresponding area under the ROC curve for each classifier is shown in Figure 1. ROC curves tell us about the classifier's ability to distinguish between the two classes. Based on the figure, MLP achieves the best area under the curve, followed by random forest and logistic regression.

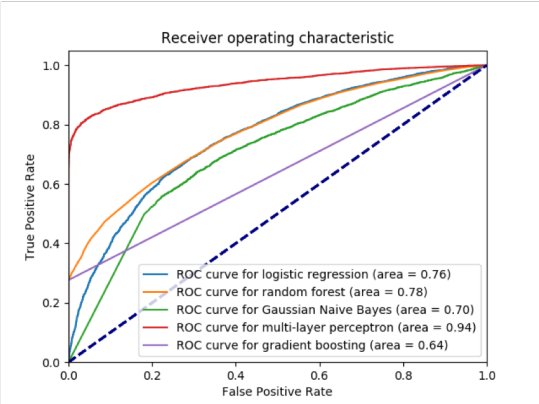

Fig. 1. ROC and area under the ROC for each classifier

C. Performance of K-means clustering and Classifications

The goal of the unsupervised learning part of the project is to get some meaningful insights into the structure of data and to potentially categorize the various types of loan applicants in our dataset. We wanted to see if there exist distinct characteristics among different groups of borrowers, if so, we could build different prediction models for different groups. The unsupervised learning technique we tried is K-means clustering.

We first performed k-means clustering on the dataset. We experimented with several values of k. Then we build a prediction model for each one of the clusters. The results are shown in TABLE IV. The models achieved the best overall performance when $k = 4$. For each one of the four clusters, the LightGBM model performs the best out of all the machine learning models. If we compare the accuracy, precision and F1 scores in TABLE IV with that of LightGBM

in TABLE III, we observed that the model's performance improved significantly from an accuracy of 57.47% to 71.57%. We could infer from the result that each cluster identified by the k-means algorithm exhibit characteristics that could be picked up by the model when trained separately, but not when the model is trained on the entire dataset.

D. Visualization of high-dimensional data

- **Principle Component Analysis (PCA):** Since we have expanded our feature space to include around more than 800 features after data processing, despite the large amount of data we have, we do not yet have a good understanding of the relationship between each variable. To answer this question, we are looking for a dimension reduction technique that could tell us what the important features are. One technique we tried is the principle component analysis, a.k.a. PCA. Moreover, the dimension reduction technique could also be used as a convenient way to visualize our high-dimensional dataset.
  Since it is the easiest to perform visualizations on a 2D or 3D plot, in Figure 2, we applied PCA to get the first two principle components. As a sanity check, the first two principle components account for around 23% of the variation of our dataset. The scatter plot is based on the top two principle components and it colored the two classes differently. From the plot, we observed four distinct clusters. This aligns with our results in the previous section where k-means works the best when number of clusters represented by k is set to 4. Another thing we observed is that despite the fact that four clusters are formed, the data points from the two classes are still quite inseparable. The blue dots, representing the positive class, are mingled within the clouds of red dots. This result aligns with our experiment since it is quite difficult for the classifiers to distinguish between the two classes if they cannot be separated apart using existing attributes, as we have seen in the visualization.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** t-Distributed Stochastic Neighbor Embedding, a.k.a. t-SNE, is also another technique for dimension reduction. According to [18], t-SNE differs from PCA in that it "minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding." In other words, t-SNE takes an probabilistic approach to reduce dimensions, rather than a mathematical one that requires an eigenvector computation as in PCA.
  Similar to PCA, we compute the top two dimensions using t-SNE and plotted that in Figure 2. A similar problem exists in the plot since the two classes cannot be dinstinguished easily.
  Last but not the least, t-SNE works well when we first perform dimension reduction using PCA, and then run t-SNE again on the reduced data. The plot looks slightly better in terms of separating the two classes but the challenge is still present. This again confirmed that our dataset is quite challenging to work with. Despite of that, we gained several interesting insights into the dataset, and we are able to draw reasonable and meaningful conclusions from our results.

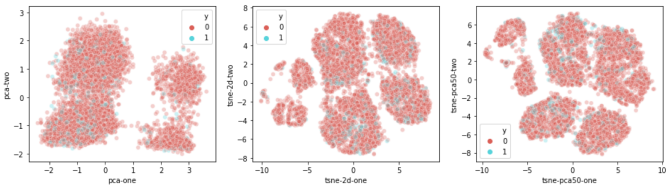

Fig. 2. Visualizations of high dimensional data. 1) PCA for the top two principle components. 2) t-SNE. 3) t-SNE after applying PCA

E. Feature Importance Analysis

If we look at the feature importance in Figure 3, we noticed that the top features are 'NUM_DAYS_EMPLOYED', 'DAYS_BIRTH', etc. The most important feature is number of days that this person has been employed. This is reasonable in that the more days the person gets employed, the more chance that he gets stable incomes. This shows the ability to keep a good credit and pay the loan on time. The second important feature is the days of birth, that is how old is this person. The older the person, the more chance that he can pay the loan more often and the lower risk of default.

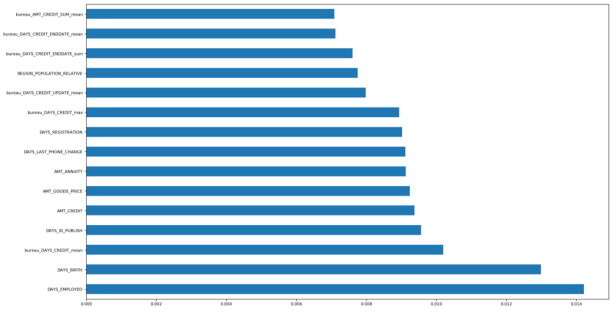

Fig. 3. Feature importance extracted from the random forest classifier

VI. Conclusions and Future Work

In our report, we demonstrated the use of machine learning algorithms on a very challenging dataset to predict loan repayment ability. To achieve the best performance, we showed that data pre-processing, a careful selection of techniques of balancing dataset and classification algorithms are all very important. Logistic regression and neural networks work quite well on our dataset, and the use of k-means is also effective. In the future, we want to continue exploring more sophisticated learning algorithms and dimension reduction techniques to further improve model performance on this important prediction task.

VII. CONTRIBUTIONS OF TEAM MEMBERS

All three team members have roughly equal contributions to the project. The main contribution of Yiyun is to implement various machine learning models for the loan repayment prediction and implement k-means for the clustering. The main contribution of Xiaomeng is to implement machine learning models and feature engineering. The main contribution of Zihan is to implement algorithms for data processing and machine learning model implementation. All team members contributed to the write-up of the report.

REFERENCES

[1] "Home Credit Default Risk." Kaggle, https://www.kaggle.com/c/home-credit-default-risk/data.

[2] Gorton, Gary, and James Kahn. "The design of bank loan contracts." The Review of Financial Studies 13, no. 2 (2000): 331-364.

[3] Langrehr, Virginia B., and Frederick W. Langrehr. "Measuring the ability to repay: The residual income ratio." Journal of Consumer Affairs 23, no. 2 (1989): 393-406.

[4] Kolo, Brian, Thomas Rickett McGraw, and Dathan Gaskill. "Systems and methods for using data metrics for credit score analysis." U.S. Patent Application 13/456,532, filed November 1, 2012.

[5] Çelik, Şaban. "Micro credit risk metrics: a comprehensive review." Intelligent Systems in Accounting, Finance and Management 20, no. 4 (2013): 233-272.

[6] Olivas, Michael A. "Paying for a law degree: Trends in student borrowing and the ability to repay debt." J. Legal Educ. 49 (1999): 333.

[7] Hesseldenz, Jon, and David Stockham. "National direct student loan defaulters: The ability to repay." Research in Higher Education 17, no. 1 (1982): 3-14.

[8] Flint, Thomas A. "Predicting student loan defaults." The Journal of Higher Education 68, no. 3 (1997): 322-354.

[9] Afolabi, J. A. "Analysis of loan repayment among small scale farmers in Oyo State, Nigeria." Journal of Social Sciences 22, no. 2 (2010): 115-119.

[10] Wongnaa, C. A., and Dadson Awunyo-Vitor. "Factors affecting loan repayment performance among yam farmers in the Sene District, Ghana." Agris on-line Papers in Economics and Informatics 5, no. 665-2016-44943 (2013): 111-122.

[11] Murdock, C.W., 2011. The Dodd-Frank Wall Street Reform and Consumer Protection Act: What Caused the Financial Crisis and Will Dodd-Frank Prevent Future Crises. SMUL Rev., 64, p.1243.

[12] Ivashina, Victoria, and David Scharfstein. "Bank lending during the financial crisis of 2008." Journal of Financial economics 97, no. 3 (2010): 319-338.

[13] Mierzewski, Michael B., Christopher L. Allen, Jeremy W. Hochberg, and Kevin Hall. "CFPB Finalizes Ability-to-Repay and Qualified Mortgage Rule." Banking LJ 130 (2013): 611.

[14] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[15] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems. 2017.

[16] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[17] Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

[18] Laurens van der Maaten, Geoffrey Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research, 2008.

[19] Visualising High-dimensional Datasets Using PCA and t-SNE. https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-