DS-GA 1001 Capstone Project

Group Name: Data Police
Tianyu Du, Nuo Lei, Chuan Shi, Yong Zhao

Introduction

The Peer-to-peer (P2P) lending market is a dynamic and growing sector that involves the borrowing and lending of money between individuals and small businesses. However, this market can also be risky due to the lack of collateral and credit history for many borrowers. In addition, the high volume of daily lending transactions can create a significant workload for risk evaluation. How to mitigate these risks is an interesting problem. Here we utilize a variety of statistical and machine learning techniques to analyze the financial history, credit score, income, and other demographic factors of borrowers in order to predict loan amount and default risk, which further support investment decisions in the P2P lending market.

Dataset explanation

Our dataset is collected from Lending Club, one large P2P lending platform, and is available for download on Kaggle. Our dataset consists of 396,030 lending records, each with 27 features. The features can be divided into two main categories. One is borrower information, such as loan grade, employment history, and annual income. Another one is transaction details, such as interest rate and loan amount. Our prediction target is the loan amount of every borrower, which gives us an estimation of loss given default (LGD). Our classification target is the loan status, which is labeled as either "Fully Paid" or "Charged-off." Our main aim is to classify the loan status by the borrower and transaction information. By statistical and machine learning models, we hope to gain insights into the factors that contribute to default risk in the P2P lending market. The specific 27 feature explanations are included in the appendix.

Preprocessing

In our prediction and classification questions, we handle missing values in different ways due to the use of different features. For the prediction question, we generally drop rows with missing values, as there are fewer missing values in numerical data. However, for the classification question, we also perform imputation on certain features because we want to have more data and thus more statistical power to achieve better classification accuracy.

We choose to retain extreme values in both models because they may not always represent noise in financial data. For example, a small group of people may have extremely high annual incomes. Therefore, after careful consideration, we decided not to drop these outliers because they reflect real-world situations. We will provide more information about our data processing methods in the relevant sections.

Inference

Question: Does the mean of the loan amount differ among locations?

We notice 10 different zip codes in the address(`address`). We would like to see whether the loan amount of each borrower(`loan_amnt`) differs among the borrower's location. Interested in whether there is a regional difference among borrowers and if the regional information potentially represents some characteristics of the borrower (e.g. social status), we would like to investigate the proportion of statistically significant results between each pair of the locations of the borrowers.

Analysis approach

To answer this question, we first extract the zip codes from the address data. The size of missing data is small so we simply do row-wise elimination. Then, we convert the zip codes to categorical indicating 10 different location groups of borrowers. We are using the t-test for independent groups here. We check for the assumptions of the t-test for independent groups. First, the loan amount is numerical and the mean of the loan amount can be interpreted. Since the data size is large and contains almost all borrowers' information within a certain period, we can assume the data is representative and has normal distribution for the borrowers. The variance within each location group is similar, as we calculated, so it is reasonable for us to assume homogeneity of variance. We, in addition, assume the borrowers are randomly assigned to each location group and every participant is only in one group. We then proceed with a t-test for independent groups and calculate the proportion of significant results.

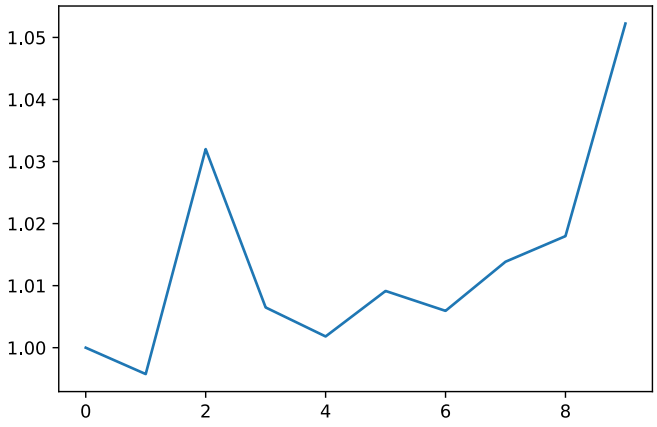


Figure 1: Varience of each location group

Result analysis

The plot for cohen-d, confidence interval of t-statistics, and p-value are listed below. The proportion of the significant test is 0.733 as we calculated.

Cohen-d for effect size

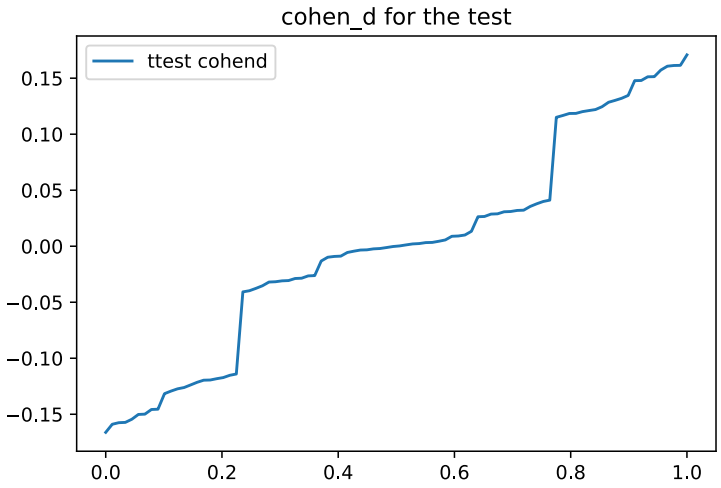


Figure 2: Cohen-d

We can see from figure 1 that all of the cohen-d are with magnitudes less than 0.2, and about half of the cohen-d is larger than 0.1. Considering the relatively large sample size, we can conclude that there are some differences in the mean between the location group pairs.

Confidence Interval

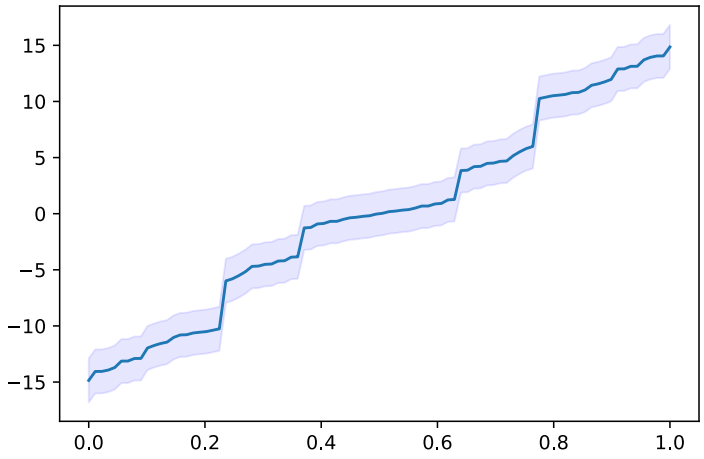


Figure 3: 95% CI of the t-statistics

Figure 2 indicates the confidence interval for our t-statistics. It shows the probability that the confidence interval contains our true t-statistics for each pair of tests. From this graph, we could predict the range of our proportion calculation.

P-value

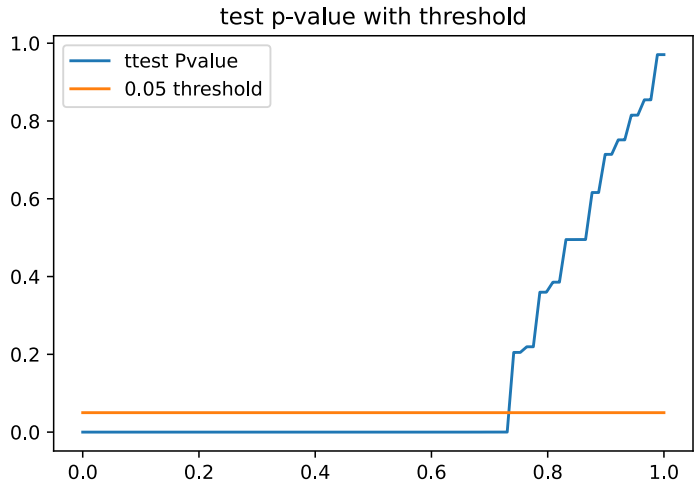


Figure 4: P-value with 0.05 threshold line

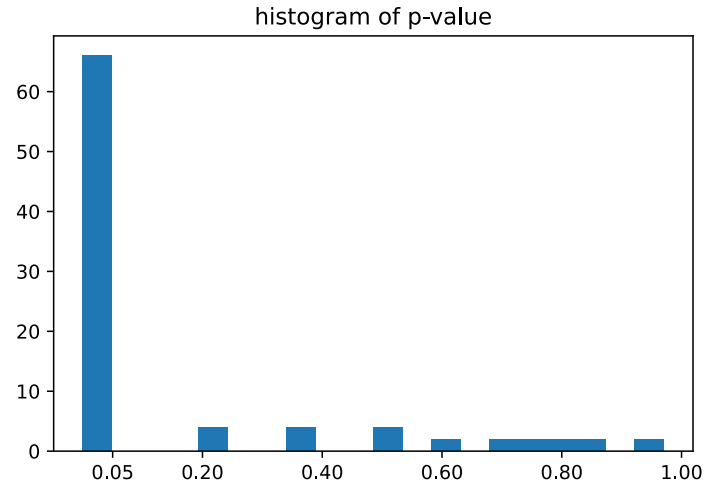


Figure 5: The histogram of P-value

As we calculated, 73% of the paired groups are tested with significant results. This is an incredible result that shows a strong regional difference between the borrowers on their loan amounts. We can observe mean differences in the loan amount between most pairs of two locations. We may expect some characteristics of the borrowers that may affect their loan amounts to be distributed or clustered among different locations.

Prediction

Question: Does annual income predict loan amount?

We are curious about the relationship between the loan amount of each borrower(`loan_amnt`) and their annual income(`annual_inc`). Our question is, can we use the borrower's annual income to pre-

dict its expected loan amount, while controlling for confounding factors like the credit revolving balance(`revol_bal`), the number of public record bankruptcies(`pub_rec_bankruptcies`), and etc.?

Analysis approach

To answer this question, we build multiple models including multiple regression, ridge regression, lasso regression, and artificial neural network model, trying to predict the loan amount and explore its relationship with annual income.

Here multiple regression model helps us to determine how annual income would influence the loan amount while controlling for other factors. We also employed other models including ridge regression, lasso regression, and artificial neural network model trying to improve prediction performance.

Feature selection

First, we use the Pearson correlation coefficient to check the potential multicollinearity between variables. The result below shows that features `loan_amnt` and `installment` are highly correlated with a coefficient of 0.95. Actually, they contain the same information about payment. The amount of the loan of each borrower determines the monthly installment. So we choose to delete `installment` from our predictor.

Also, `total_acc` contains `open_acc` with a coefficient of 0.68, `pub_rec` contains `pub_rec_bankruptcies` with a coefficient of 0.69. This is because that the total amount surely include the number of a specific type. So to avoid multicollinearity, we delete `total_acc` and `open_acc`.

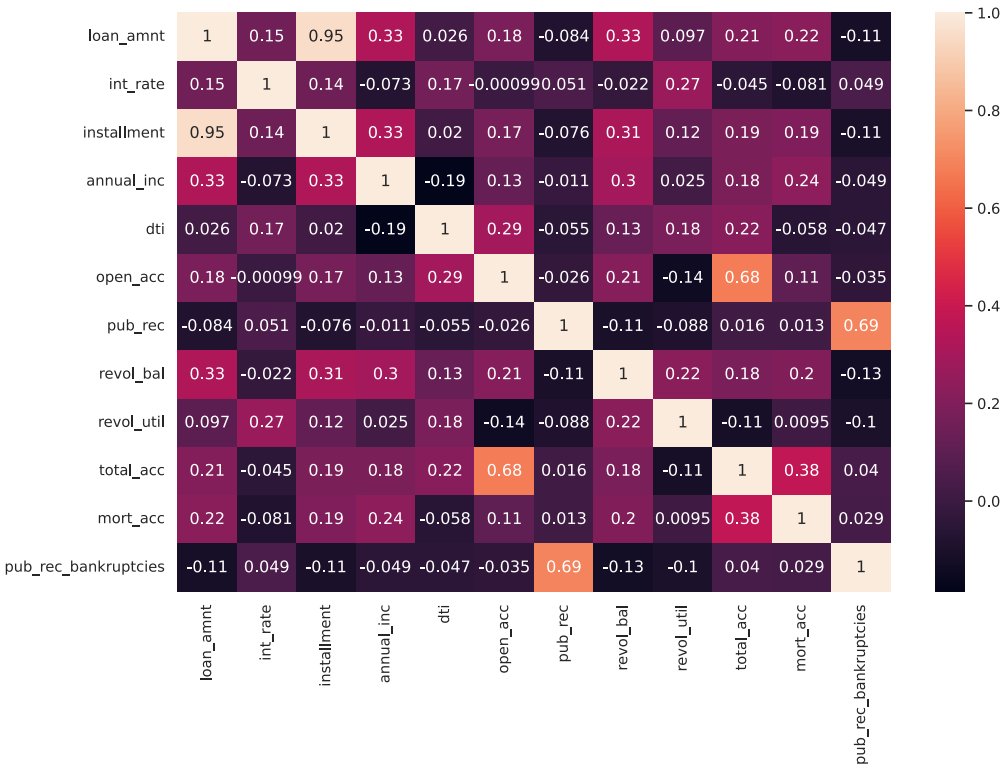


Figure 6: Pearson correlation coefficient of numerical features

Result analysis

We assess our model's performance by the following metrics: Residuals distribution, R^2, and $RMSE$. They give us an overall understanding of models' performance. After that, we look closer at the

multiple regression model's coefficient to see whether annual income can predict loan amount.

Residuals distribution

A good regression result should left white noise in the residuals, which means that the residual should follow a Gaussian distribution of zero mean and constant variance.

Here we plot the histogram of our four model's residual as follow. We find that linear models' residuals have zero mean but a left skewed distribution. This means that they do not follow a Gaussian distribution. On the contrary, ANN model's residual leads to a much more symmetric histogram with zero mean, which is quite like white noise.

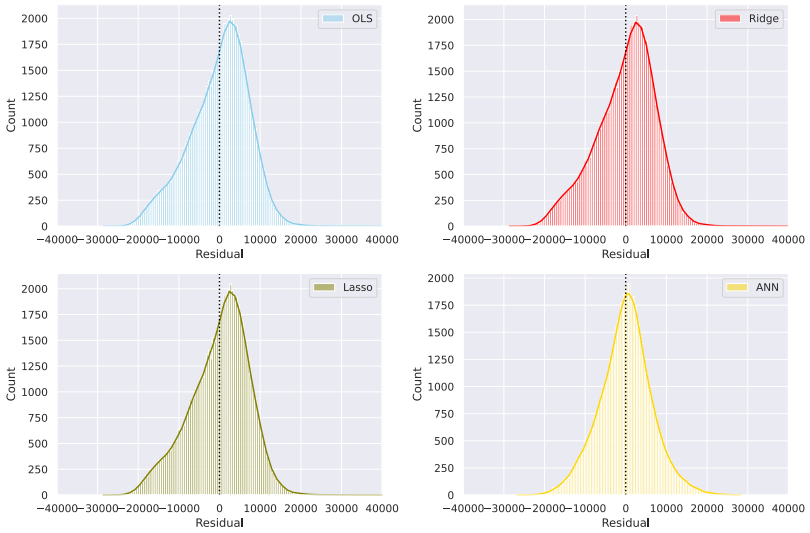


Figure 7: Residual distribution of each model

When we plot the four histogram together, we can see more clearly that the residuals of ANN is more symmetric and has smaller variance compared to linear models.

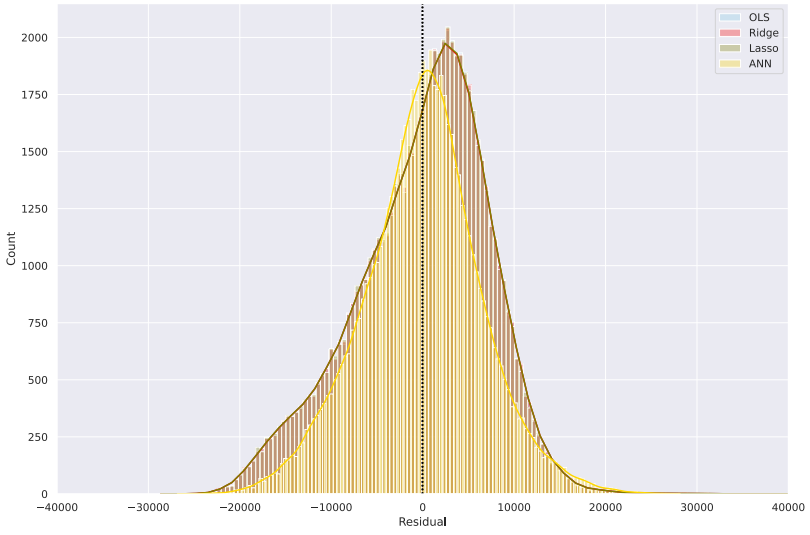


Figure 8: Residual distribution comparison

Comparison of R^2 and $RMSE$

We take a closer look at the regression R^2 and $RMSE$. Linear models have R^2 around 0.23. But ANN has R^2 around 0.41. And ANN also has a smaller $RMSE$ compared to linear models.

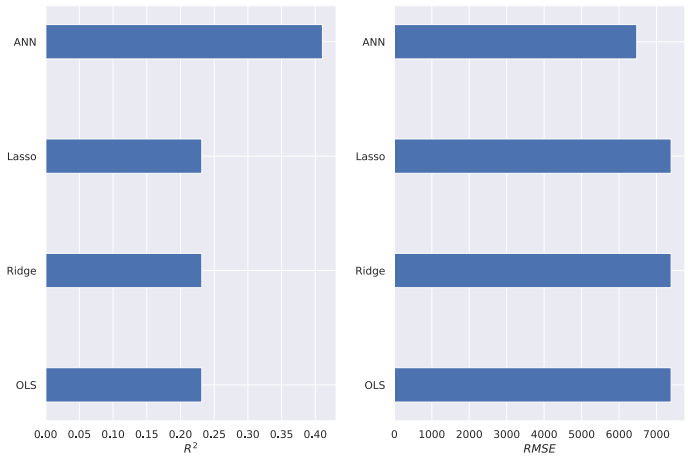


Figure 9: R^2 and $RMSE$ of regression

In conclusion, we find that ANN has a better performance than other linear models with white-noise residuals, higher R^2 and lower $RMSE$.

Coefficients of multiple regression

Although ANN has the best prediction performance, neural networks are hard to interpret. So to answer our questions about whether annual income can predict loan amount, we check the coefficient's significance of multiple regression. The coefficients' absolute t-values are plotted below.

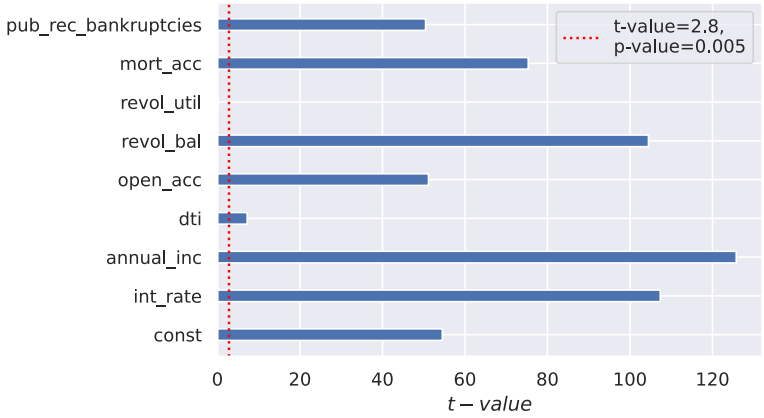


Figure 10: The absolute t-values of multiple regression

In this plot we find that annual income(`annual_inc`) has a coefficient t-value of 125.7 >> 2.8 (the corresponding t-value with p-value=0.005). Therefore, we conclude that we can use the borrower's annual income(`annual_inc`) to predict its expected loan amount, while controlling for eight confounding factors!

Classification

Question: Does clustering and/or the aforementioned zipcode feature improve our classfication of loan defaults?

The ultimate goal of this dataset and the project is to build a model predicting whether or not a loan would be into default. We want to investigate whether clustering data into groups help our prediction performance, and how we can improve our model by addressing the regional difference discovered in the Inference section.

Exploratory Data Analysis

Because we would be doing clustering before classification, we want to check the outliers for the numerical columns. We first scanned the table with mean, standard deviation, min, max, and quantiles of each column and drew box plots (see fig 11):

We can see that most of them have quite a lot of outliers. However, we think those outliers are actually from the real distribution. For example, we see that the annual income at the 75th quantile is 90000, and the max value is about 870000 dollars. This can happen because there are some people earning a lot of salaries. Therefore, we did not remove any outliers. Having many outliers in our data is detrimental to the model performance of K means, which is a problem we would discuss in detail in the clustering section.

Data Preprocessing

We applied feature engineering/dropping on the following features:

Feature	Method
`issud_d`	Dropped. It represents the month which the loan was issued. As this indicates whether the loan was accepted or denied, including this feature might lead to data leakage.
`installment`	Dropped due to collinearity with `loan_amnt`.
`emp_title`	Dropped due to being categorical feature with too many distinct values (because almost everyone has a different job title and a different reason to borrow loan), and one-hot this feature would be trivial.
`title`	Dropped with the same reason above.
`grade`	Dropped due to being subset of `subgrade`
`earliest_cr_line`	Dropped the month part and kept only the year.
`sub_grade`	Mapped from $A1, A2, \ldots, G4, G5$ to $0, 1, \ldots, 33, 34$
`emp_length`	Encode as the year of employment length. For entries with ¡1 years and 10+ years, we encode them as 0 and 10 respectively. We calculated the fully paid rate of people with 8, 9, and 10+ years and got values 0.80, 0.80, 0.82, and there is not much difference. Also for the fully paid rate of people with ¡1, 1, and 2 years, we got values 0.79, 0.80, 0.81, so we can see that it is reasonable to just convert them into 0 and 10 years.
`loan_status`	Encode Fully Paid as 0 and Charged Off as 1. This would be our target variable in classification.

For all other categorical features, we converted them to numerical ones using one-hot encoding.

For missing data, because we want to have as much statistical power as possible, we did imputation on some columns that we think reasonable. We noticed that `mort_acc` has 10% value missing, so dropping them would not be practical. Therefore, we filled the missing values with 0, as a histogram shows that 50% of the data falls in the bin [0, 1.36], and that the conjecture of NaN values is likely due to no mortgage accounts. We also imputed NaNs in `emp_length` to 0 because we believe the reason that most people do not fill in this information is because they are not employed (i.e. 0 years of employment). For the rest of the missing data, we did a row-wise removal.

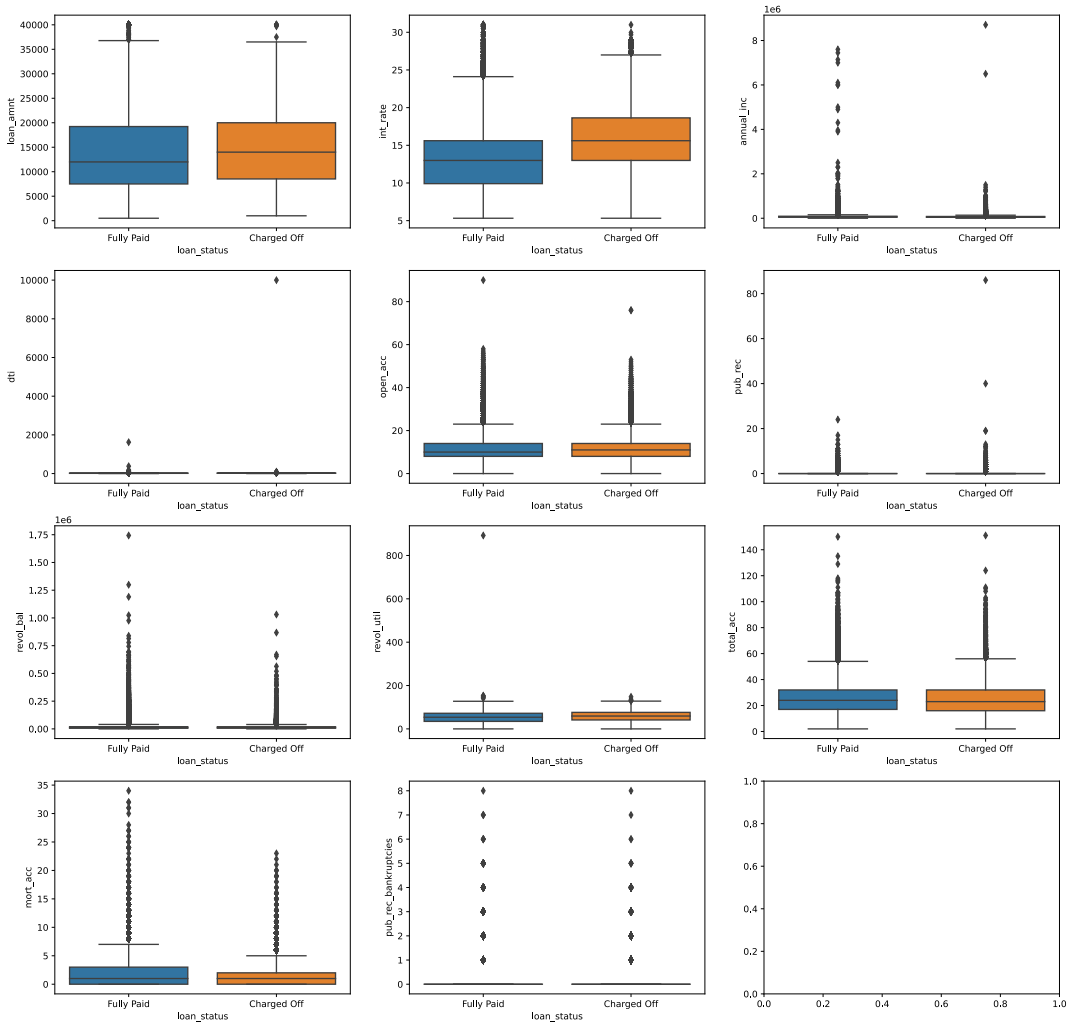


Figure 11: box plots of the numerical features

Clustering

We considered two clustering methods: K means and DBSCAN. Because we have many outliers, we expected DBSCAN would produce a better model. We found the best epsilon and minPoint using the k-dist method, but we would not explain it here because we ended up running out of memory when fitting the DBSCAN model. The code of hyperparameter tuning and model fitting of DBSCAN can be found in the code file.

Therefore, even though we have a lot of outliers that would worsen the model performance of K means, it is still a doable model given our computer memory, so we did it to see if adding the clusters as a new predictor would enhance our classification result. We applied the silhouette method to find the best k (number of clusters) for K means. We found that k = 14 gives us the best silhouette score, and we did K means with this best k.

To have a brief view of K means's performance, we drew a histogram of `loan_amnt` for each cluster and we thought there may exist some relation between the amount of money you loan and whether you fully paid the loan or charged off. We expected that if there exists some relation, then there would be some notable difference between the distributions (such as some clusters may contain mostly the people with small loan amount and some mostly contain large loan amount). The histograms are shown in figure 12.

Similarly, we drew the histograms of `sub_grade` in figure 13:

We found that for `loan_amount`, cluster 6 and 7 contains a lot of people with very high salary, and cluster 2 and 4 contains a lot of people with very low salary. The rest clusters show similar distributions. For `sub_grade`, all clusters show very similar distributions. We expect that `loan_amount` to be a possibly good predictor, but because of too many outliers in general, we think K means does not help us improve our classification accuracy much, and we will explore this in the classification session.

Hyperparameter Tuning

Our main task is to predict `loan_status` using the other features as predictors. We first built baseline models without using the zipcode/address feature or the cluster results. We chose the following 3 classification models: Random Forest, Gradient Boost/XGBoost, and Logistic Regression.

For Random Forest model, we perform a grid search on the following hyperparameters:

Hyperparameter	Grid
`n_estimators`	10, 50, 100, 200
`min_samples_split`	2, 5, 10, 20

After conducting a 10-fold cross validation, we found the best parameters to be `n_estimators`=200 and `min_sample_split`=10.

Similarly, for XGBoost model, we used the following grid:

Hyperparameter	Grid
`eta`	0.05, 0.1, 0.15, 0.2
`max_depth`	4, 5, 6, 7

And the cross validation resulted in `eta`=0.05 and `max_depth`=6.

For logistic regression, we apply the l_2 norm penalty, and tune the hyperparameter `C`, which is the inverse of the regularization parameter. The best `C` is found to be 21.54.

Baseline Models

Using these hyperparameters, we built the three classfication models. The ROC curves and P/R curves on the test set are shown in figure 14.

Incorporating Clusters

To incorporate the clustering calculated in the previous section, we added a new feature to the training set representing the cluster it belongs. Then we performed one-hot encoding to turn that categorical data to numerical data. We fit our three models using this new training set. As for predictions on the

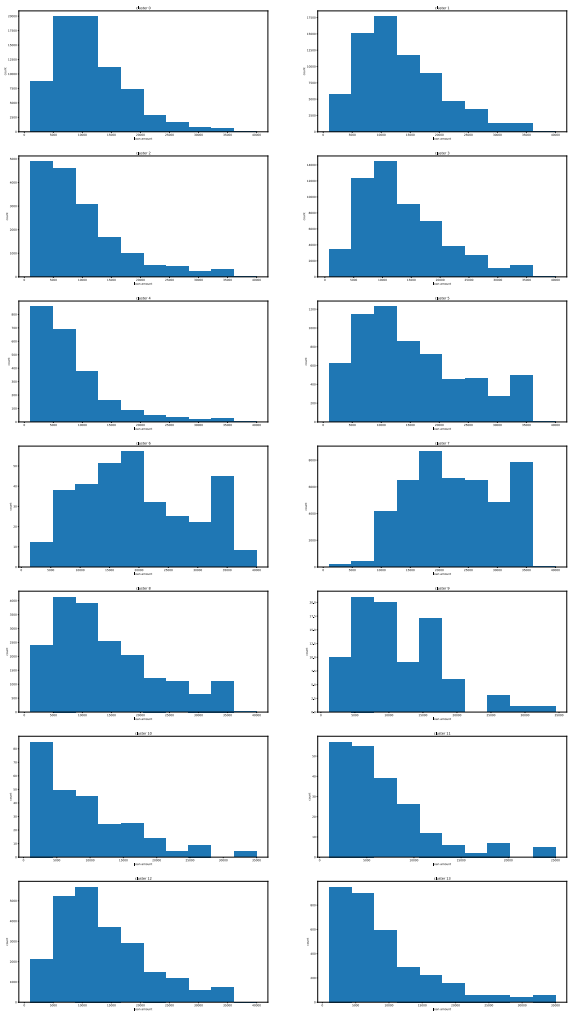


Figure 12: histograms of the loan amount of the 14 clusters

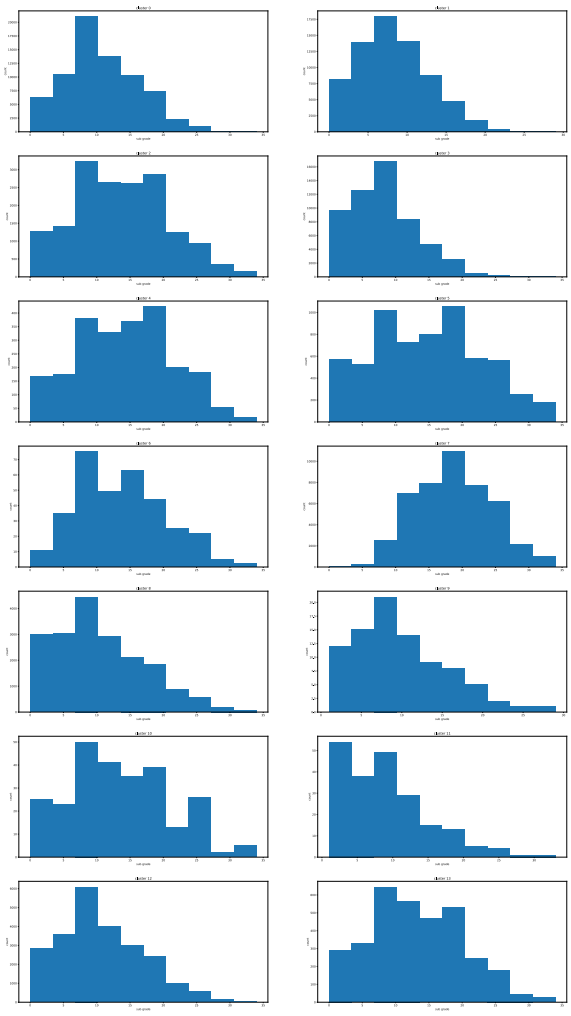
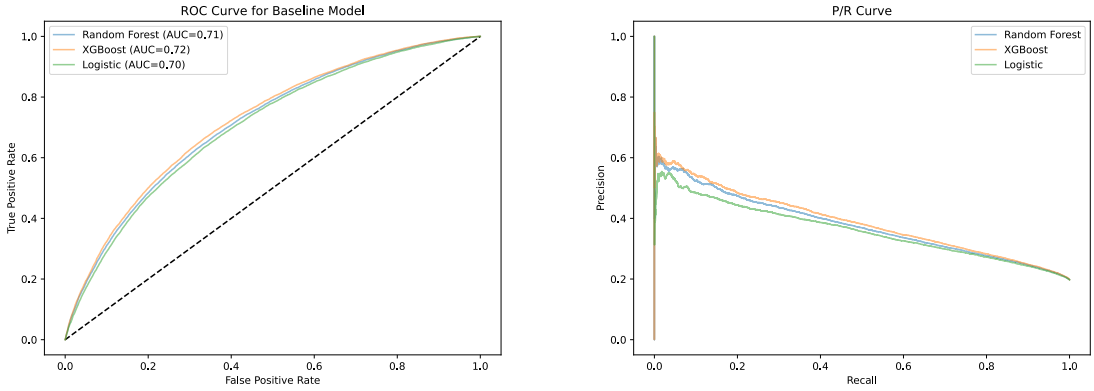
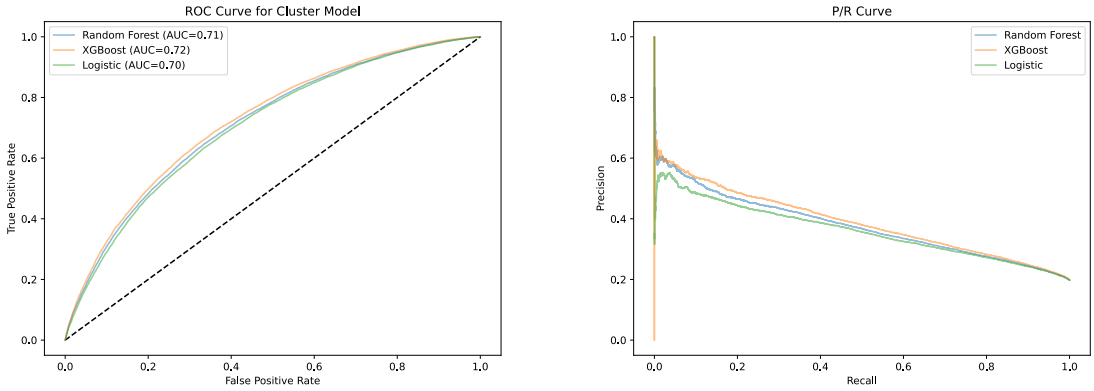


Figure 13: histograms of sub grade of the 14 clusters



(a) ROC Curve (b) P/R Curve

Figure 14: Performance of Baseline Models



(a) ROC Curve (b) P/R Curve

Figure 15: Performance of Cluster-Improved Models

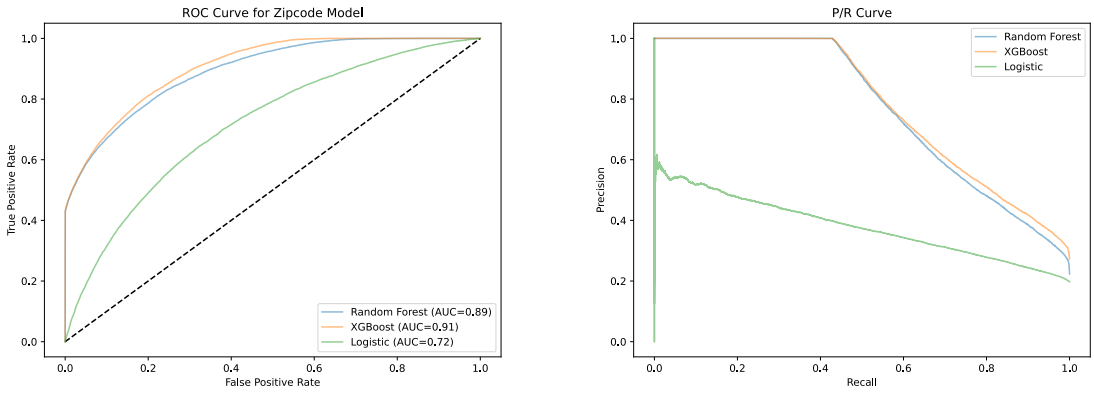
test set, we assign the cluster for each row by performing the KNN algorithm. The ROC curves and P/R curves on the test set are shown in figure 15.

From both curves and the AUC value, we can tell that incorporating the information given by the clustering algorithm does not seems to improve the model performance much. One possible reason is that whatever information given by the clustering has already been addressed in the baseline model.

Incorporating Zipcode

In this section, we will readdress and explore further of the problem of the `zipcode` feature as previously discussed in the Inference section. We performed one-hot encoding on this categorical feature, and reran our three models. The results are shown in figure 16

As shown in the figure, our results for the Random Forest model and Gradient Boosting/XGBoost model has greatly improved from the baseline model, with an increase of nearly 0.2 in AUC. We can therefore conclude that the `zipcode` feature is an important feature. However, as stated in the Inference section, we cannot verify the authenticity of this feature. We will discuss this later in the Summary and Conclusion section. It is also worth-noting that the Logistic Regression model does not seem to improve a lot compared to the baseline model, as expressed in both the ROC and P/R curves.



(a) ROC Curve

(b) P/R Curve

Figure 16: Performance of Zipcode-Improved Models

Conclusion

Overall Conclusion

In this project, we performed quantitative analysis on a lending record dataset from the P2P lending platform Lending Club. We focused on two features of the dataset: loan amount (`loan_amnt`) and loan status(`loan_status`). By conducting statistical tests and regression, we found that the loan amount differs significantly in different locations/areas, and annual income (`annual_inc`) is a good numerical predictor of this feature. We have also discovered that while clustering did not have much impact on predicting loan status, the zip code, extracted from `address`, did turn out to be an important feature that greatly improved our classification models. Further exploration showed an interesting result that will be included in the Extra Findings section.

Limitations

In preprocessing the data, we have dropped columns that might contain very useful but not as usable information. For example, `emp_title` and `title` contain information about the lendee's occupations. However, because titles are often too detailed, we cannot process them in a feasible manner in this project. We have also assumed linearity of the `sub_grade` feature by mapping them to consecutive integers, which might not be true, but how they calculate and assign this feature are unknown to us. We have also made assumptions that people with missing values for this feature have 0 mortgage accounts. This is based on the distribution of the non-missing data, but it may still be invalid. Besides, outliers have not been removed in the classification models, based on the assumption that they are part of the true population, which again can be false. What's more, in the prediction task, we assume that there is a linear relationship between our predictor and target. But our comparison of linear models and ANN shows that there are some non-linear relationships between the loan amount and other numerical factors. Therefore the regression coefficients may be influenced by the potential non-linear relationship.

In an ideal world, a dataset that does not have any of the limitations would be one that does not have missing values, and with features that are both precise and detailed, but also manageable in some ways, so it can be utilized directly, or after some encoding, by our machine learning algorithms. It should also be faithfully representing the underlying population so that outliers are handled. Gathering such a dataset would, of course, be close to impossible, as it requires a perfect profiling of the lendee. However, we can still improve the gathering process. For example, when banks/lenders gather information of the lendee, they should carefully design their forms so that there won't be any missing values. One way to do that is to make all fields required, and add "other" choice where people can elaborate. As for the over-complicated `title` columns, banks/lenders can also ask for the industry that the lendees are in. In short, one should look for quantifiable data, and avoid missing values as possible.

Extra Findings

Our result with the variable zip code, extract from the address(`address`) show impressive performance in terms of statistics in our parts of inference and classification. Nevertheless, the result looks so neat that we doubt whether it could be from real collected data. The categorical variable with only 10 labels seems to contain a lot of information. To resolve our doubt here, we searched for the zip code in the data and found that they are not the real existing zip codes. Instead, we surmise they are some kind of artificial labels integrated with the physical address (they might also be randomly generated) to increase the performance of the models for this set. It is very likely that it was generated with high correlation with the target variable so that it ends up causing data leakage, and improving classification performance in this way.

References

[1] lendingclub — online personal loans at great rates. https://resources.lendingclub.com/LCDataDictionary.xlsx. Accessed: 2022-12-20.

Appendix

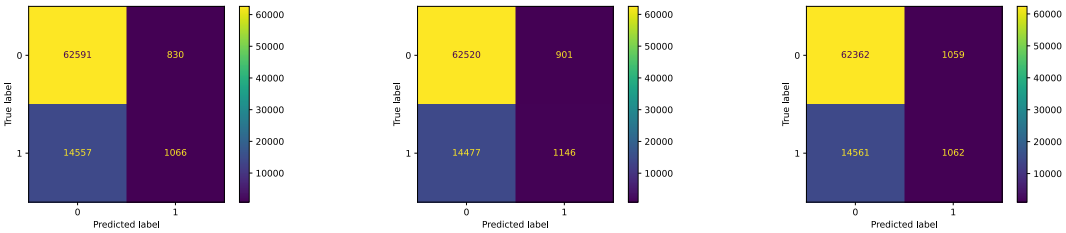
Feature Explanations

The specific meanings of 27 features are listed below, which come from the lending club website [1].

Table 1: Feature explanations

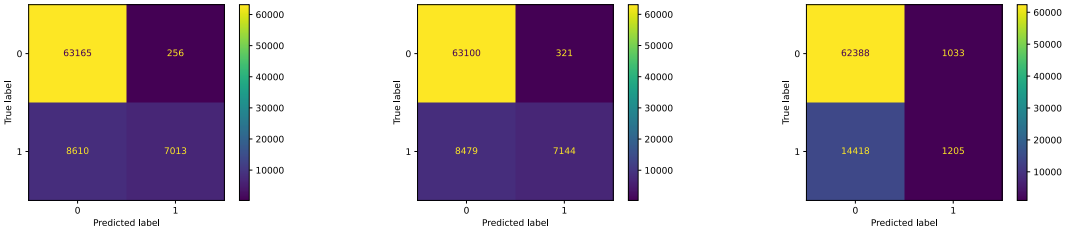
LoanStatNew	Description
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
initial_list_status	The initial listing status of the loan. Possible values are – W, F
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
mort_acc	Number of mortgage accounts.
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

Confusion Matrix



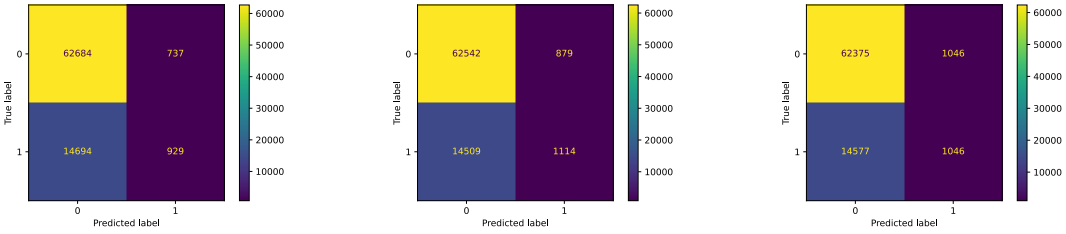
(a) Random Forest (b) XGBoost (c) Logistic Regression

Figure 17: Confusion Matrix for Baseline Models



(a) Random Forest (b) XGBoost (c) Logistic Regression

Figure 18: Confusion Matrix for Zipcode-Improved Models



(a) Random Forest (b) XGBoost (c) Logistic Regression

Figure 19: Confusion Matrix for Clustering-Improved Models