Credit Risk via Lending Club

Bopei Nie

03/24/2023

Lending Club Analysis

Background

Lending Club is the world's largest peer-to-peer online lending platform that connects borrowers and investors. By eliminating traditional financial institutions, Lending Club provides higher returns for individual investors and lower interest rates for borrowers. As a result, Lending Club has grown exponentially in recent years and has become an attractive alternative for investors who are seeking higher returns.

However, investing in loans through Lending Club is not without risks. Borrowers may default on their loans, causing investors to lose their money. Therefore, it is important for Lending Club and its investors to identify the types of loans that are less likely to default and to build a portfolio that maximizes returns while minimizing risks.

In this report, we will apply machine learning techniques to identify the important risk factors that contribute to loan default and to build a classifier that can accurately predict the likelihood of default. Specifically, we will focus on the period between 2007-2011, during which we have around 39,000 observations and 38 attributes for each of these loans. These attributes include loan amount, home ownership status, interest rate on the loan, loan status, and grade of the loan, among others.

By addressing these questions, we aim to provide a classification rule that will help investors to identify the types of loans that should be included in their portfolio, while minimizing the risk of losing money due to loan defaults.

Data Summary

The cleaned data set has around 39k observations and 38 attributes. Attributes can be segmented into pre-funded loan data, borrower data, borrower credit data and post-loan data. The target variable is `loan_status` in post-loan data including 5468 `Charged Off` and 33503 `Fully Paid`. According to the definition, `Charged Off` means defaulted and there is no longer a reasonable expectation of further payments. Thus the loan status are separated into 5468 defaulted (`Charged Off`) and 33503 non-defaulted (`Fully Paid`). However, it is unbalanced that non-defaulted accounts for a large percentage. Among the four categories, post-loan data will not be used to classify `loan_status`. Thus, we look into the other three categories.

From Pre-funded loan data, we have quantitative variables including loan_amnt, int_rate, installment and factor variables including grade, sub_grade, purpose, term. We will not use sub_grade in our case. There are in total 7 grades distribution shown as following that most grade is B and have a decreasing trend on the grades. There are two terms, including 28301 `36_months` and 10670 `60_months`. We will turn factor variables (grade and term) into dummy variables for future modeling. There are 14 purposes among which debt_consolidation accounts most. Other detailed distribution will be shown in the Appendix.

From Borrower basic information, we have quantitative variables annual_inc and qualitative variables including emp_title, emp_length, home_ownership, zip_code, addr_state, verification_status. Among those qualitative variables, home_ownership contains 17427 MORTGAGE, 18456 RENT, 2992 OWN, and 96 OTHER. Verification_status contains 16165 Not Verified, 9992 Source Verified, and 12814 Verified.

From Borrower credit data, we have quantitative variables including dti, delinq_2yrs, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, pub_rec_bankruptcies and date variables earliest_cr_line.

After having a clear understanding of what data looks like, we replace `Charged Off` with 1 and `Fully Paid` with 0 for future logistic regression and classification.

Additionally, we split them into 80% training dataset to train the model and 20% testing dataset to test the result.

Modeling – risk factors

Since we have target value 0 and 1, we will use direct Logistic Regression to fit our data and find out the important risk factors.

Based on some basic filtering, we have the following logistic regression to test factor importance. From the omitted results, we can conclude that int_rate, (term)60_months, annual_inc, inq_last_6mths, revol_bal, revol_util, and some purposes including medical, moving, small_business are significantly important at level 0.001. Some other important factors contains loan_amnt, installment, open_acc, pub_rec, home_ownership, and some purposes including debt_consolidation, educational, home_improvement, house, renewable_energy. Detailed results are in the appendix.

```
fit1 <- glm(loan_status~loan_amnt+int_rate+installment+factor(term)+factor(grade)
            +factor(purpose)+annual_inc+factor(home_ownership)+factor(emp_length)
            +factor(verification_status)+dti+delinq_2yrs+inq_last_6mths+open_acc
            +pub_rec+revol_bal+revol_util+total_acc+pub_rec_bankruptcies
             , loan_train, family=binomial(logit))
summary(fit1, results=TRUE)
```

We can also get the p-value of individual model terms below. Combined with previous modeling result, we can conclude some risk factors making the loan to be defaulted, including loan_amnt, int_rate, installment, factor(term), factor(grade), factor(purpose), annual_inc, dti, delinq_2yrs, inq_last_6mths, pub_rec, revol_bal, revol_util.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: loan_status
##
## Terms added sequentially (first to last)
##
##
##                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                         31176      25447
## loan_amnt                     1    70.24     31175      25377 < 2.2e-16 ***
## int_rate                      1  1134.58     31174      24242 < 2.2e-16 ***
## installment                   1   109.27     31173      24133 < 2.2e-16 ***
## factor(term)                  1    42.67     31172      24090 6.468e-11 ***
## factor(grade)                 6    23.28     31166      24067 0.0007093 ***
```

```
## factor(purpose)              13   179.78     31153      23887 < 2.2e-16 ***
## annual_inc                    1   153.02     31152      23734 < 2.2e-16 ***
## factor(home_ownership)        3     4.92     31149      23729 0.1779182
## factor(emp_length)           11    49.60     31138      23680 7.382e-07 ***
## factor(verification_status)  2     0.03     31136      23680 0.9831239
## dti                           1     7.23     31135      23673 0.0071516 **
## delinq_2yrs                   1     5.86     31134      23667 0.0154470 *
## inq_last_6mths                1    59.19     31133      23608 1.434e-14 ***
## open_acc                      1     0.00     31132      23608 0.9817340
## pub_rec                       1    20.00     31131      23588 7.736e-06 ***
## revol_bal                     1    18.74     31130      23569 1.495e-05 ***
## revol_util                    1    31.39     31129      23537 2.111e-08 ***
## total_acc                     1     0.27     31128      23537 0.6011999
## pub_rec_bankruptcies          1     0.40     31127      23537 0.5264016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

final model

Based on important factors, we remove some variables including verification_status, emp_length, dti, total_acc, pub_rec_bankruptcies, and open_acc. After running several Logistic Regression on selected factors, we select the following model as our final model.

```
fit2 <- glm(loan_status~loan_amnt+int_rate+installment+factor(term)+factor(grade)
            +factor(purpose)+annual_inc+factor(home_ownership)+
            +dti+inq_last_6mths+pub_rec+revol_bal+revol_util
             ,loan_train, family=binomial(logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Among various purposes, borrower with the goal of small business are more likely to have a defaulted loan. 60-month term more likely results in a defaulted loan than 36-month term. Higher interest rate will lead to higher default rate.

Running the anova test below, we can reject the null hypothesis at 0.001 level.

Modeling – classifier

After training the Logistic Regression with the training dataset, we will now test on the rest testing dataset. From the ratio of picking up a bad loan to that of missing a good loan, which is 2:1, we have P_hat($Y = 1|x$) = 0.5/(1+0.5) = 0.333. Thus, we predict loans with percentage larger than 0.333 as bad loans and vice versa.

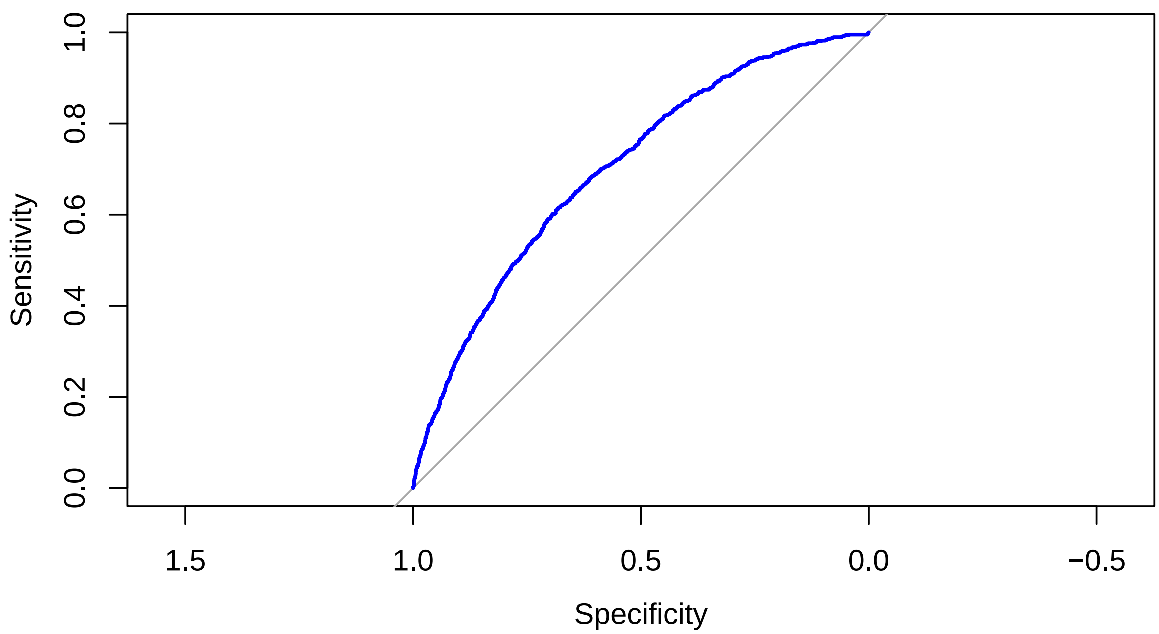
The confusion matrix below provides insight into the performance of the model. This means that the model predicted 6553 good loans correctly (TP), and 115 bad loans correctly (TN). However, it also predicted 934 bad loans as good loans (FP) and 192 good loans as bad loans (FN).

Based on the given testing result, the sensitivity of the model is 0.117, which means that only 11.7% of the true positive cases were correctly identified by the model. The specificity of the model is 0.97, which means that the model correctly identified 97% of the true negative cases. The false positive rate of the model is 0.0295, which means that 2.95% of the cases that were actually negative were incorrectly identified as positive by the model. The accuracy of the model is 0.8283, which means that the model correctly identified 82.83% of the cases.

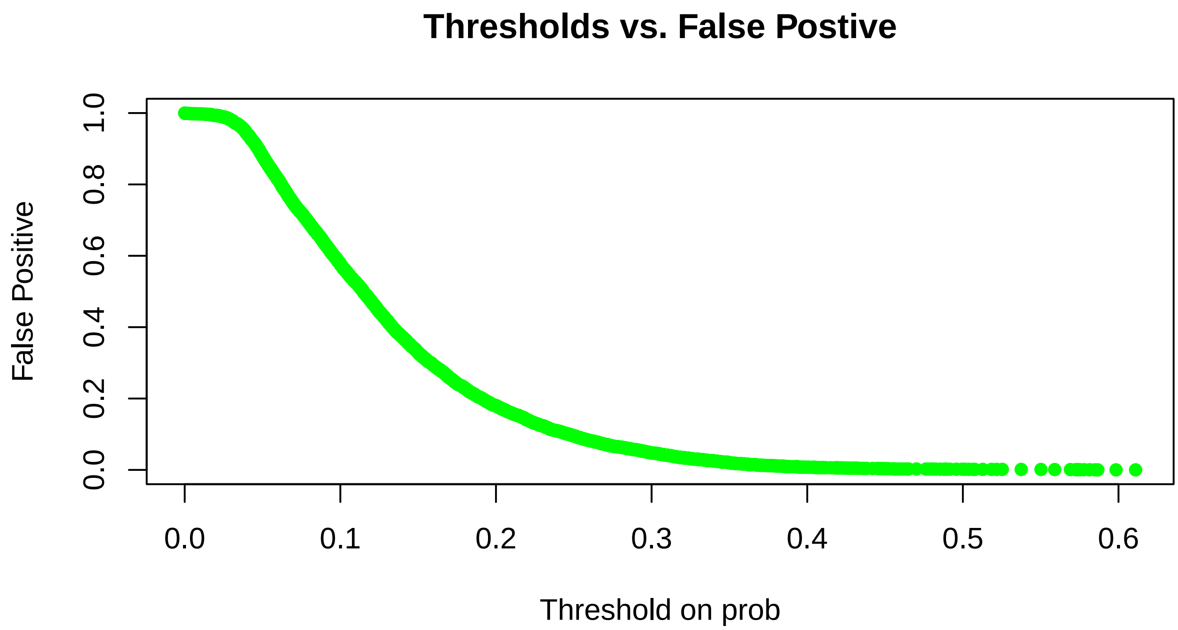
Overall, these metrics suggest that the model has a high specificity and a low false positive rate, which means that it is good at identifying true negative cases. However, the model has a low sensitivity, which means that it is not very good at identifying true positive cases.

Below is the ROC curve visualizing the performance of a classifier system across a range of thresholds. The Area Under the Curve (AUC) of the ROC curve is 0.702, suggesting that the classifier system is somewhat effective at distinguishing between positive and negative cases, but there is still room for improvement.

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



We can also plot a curve that shows the probability thresholds used and the corresponding False Positive rate. With a higher threshold, false positive decreases sharply.



Success of Lending Club

Lending Club's success can be attributed to several factors, including its ability to connect borrowers directly with individual investors, its use of technology to streamline the loan application and approval process, and its focus on providing competitive interest rates to borrowers while generating attractive returns for investors.

One key advantage of Lending Club's business model is that it removes the middleman in the lending process, thereby reducing the costs associated with traditional lending institutions. This allows Lending Club to offer borrowers lower interest rates and fees, which in turn attracts a large pool of potential borrowers. Additionally, Lending Club's platform provides investors with access to a diversified portfolio of loans, which can help to minimize risk and generate steady returns.

Another factor contributing to Lending Club's success is its use of technology to automate much of the loan application and approval process. By leveraging data analytics and machine learning algorithms, Lending Club is able to quickly and accurately assess borrowers' creditworthiness and determine the appropriate interest rates and loan terms. This not only speeds up the lending process, but also allows Lending Club to make more informed lending decisions and reduce the risk of default.

Furthermore, Lending Club's focus on transparency and customer service has helped to build trust among its users. The company provides borrowers with clear information about the loan terms and fees, and investors with detailed data on the performance of their investments. This has helped to create a sense of community among Lending Club users, which in turn has helped to drive the company's growth.

Overall, Lending Club's success can be attributed to its ability to leverage technology to reduce costs, automate processes, and improve decision-making, while also providing borrowers with competitive rates and investors with attractive returns.

Improvement

Based on the analyses done so far, there are several recommendations that Lending Club could consider in order to modify their selection rules and increase returns for investors:

Implement more robust credit scoring models: While Lending Club currently uses FICO scores to evaluate creditworthiness, they could explore more advanced credit scoring models that incorporate additional variables such as income, employment history, and debt-to-income ratio. This could help to more accurately predict default risk and improve overall loan performance.

Improve borrower verification processes: Lending Club could enhance their verification processes for borrower information to reduce the risk of fraud and default. This could include verifying employment and income information, as well as conducting more rigorous identity verification checks.

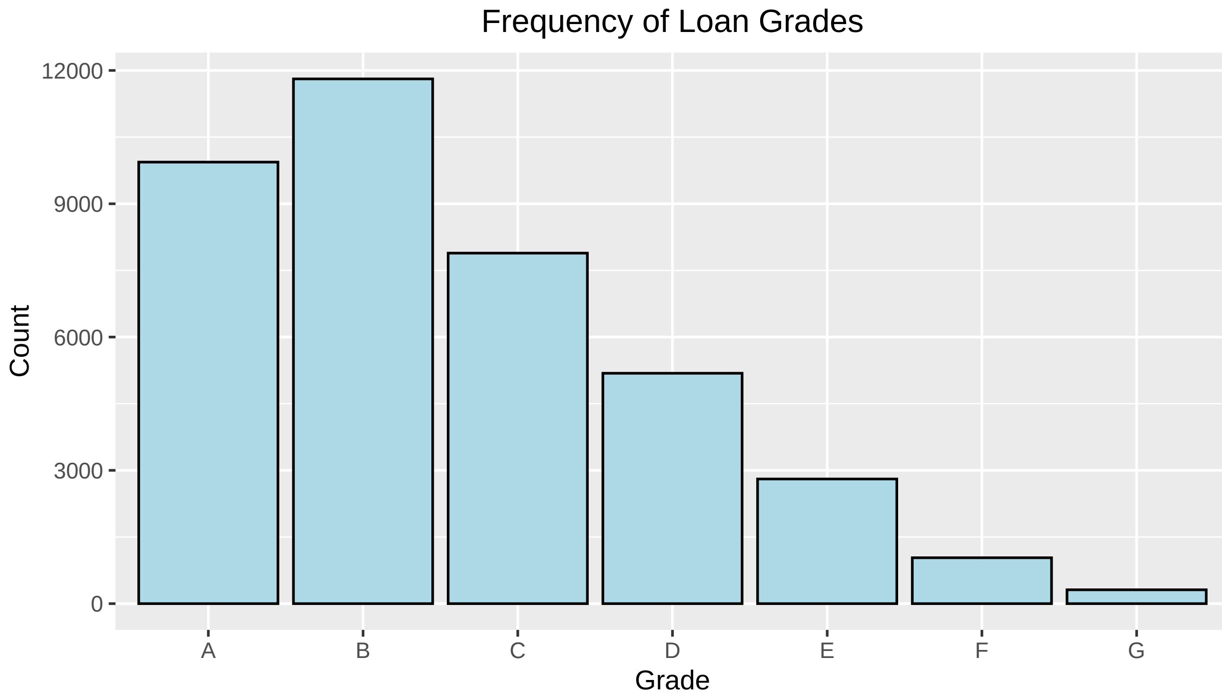
Expand loan product offerings: Lending Club could explore offering a wider range of loan products, such as secured loans or loans with longer terms, to attract a broader range of borrowers and increase the overall pool of available investments for investors.

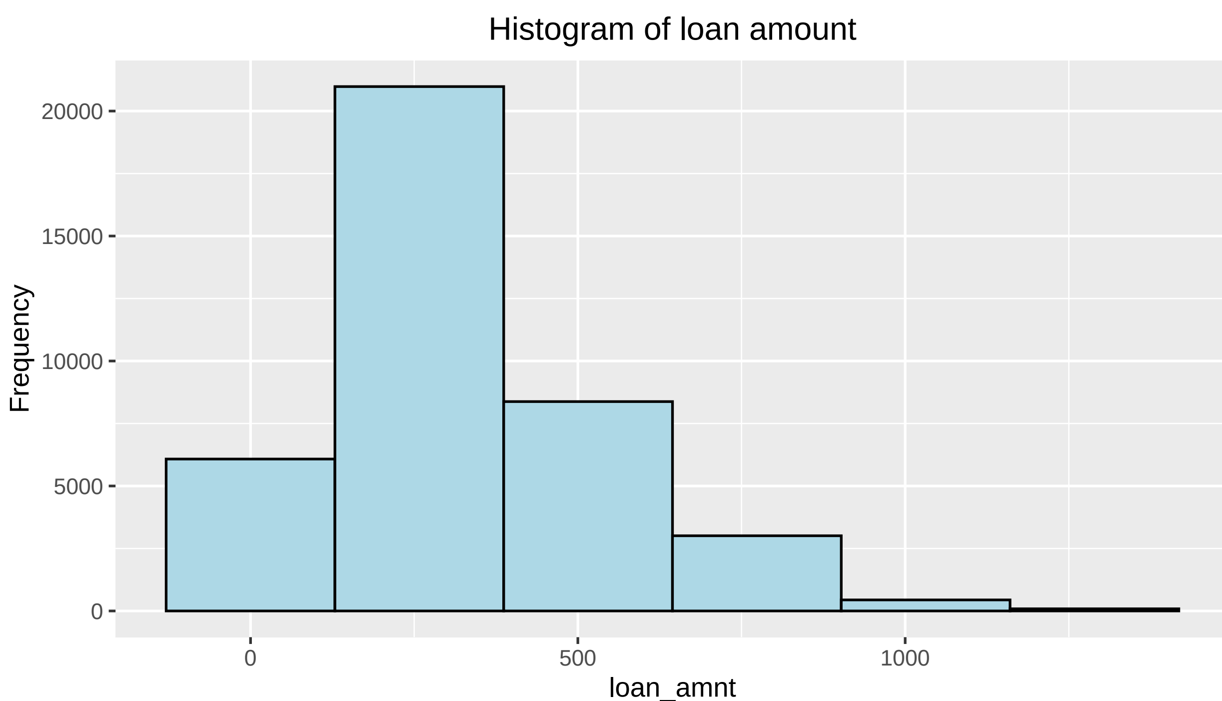
Increase transparency and communication with investors: Lending Club could improve transparency around their loan selection process and provide more frequent updates to investors regarding the performance of their loans. This could help to build trust with investors and encourage them to continue investing with Lending Club.

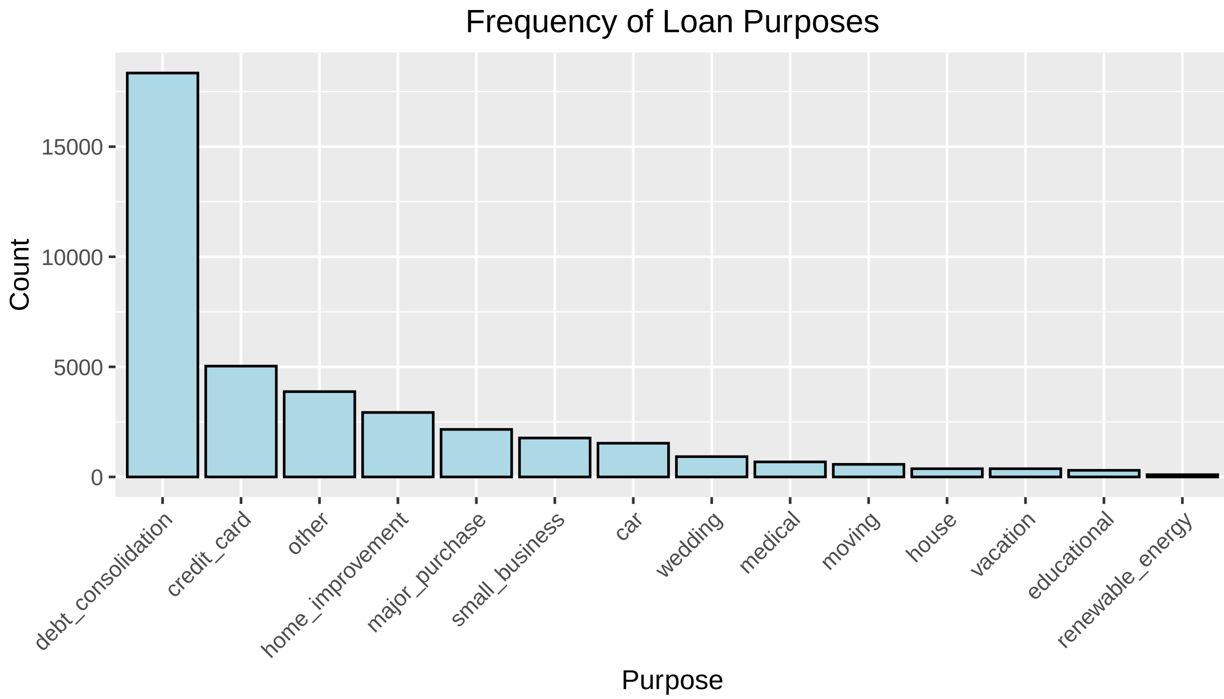
Implement more sophisticated risk management strategies: Lending Club could explore more advanced risk management strategies, such as diversifying their loan portfolios across a broader range of risk profiles or implementing hedging strategies to manage risk. This could help to reduce overall risk and increase returns for investors.

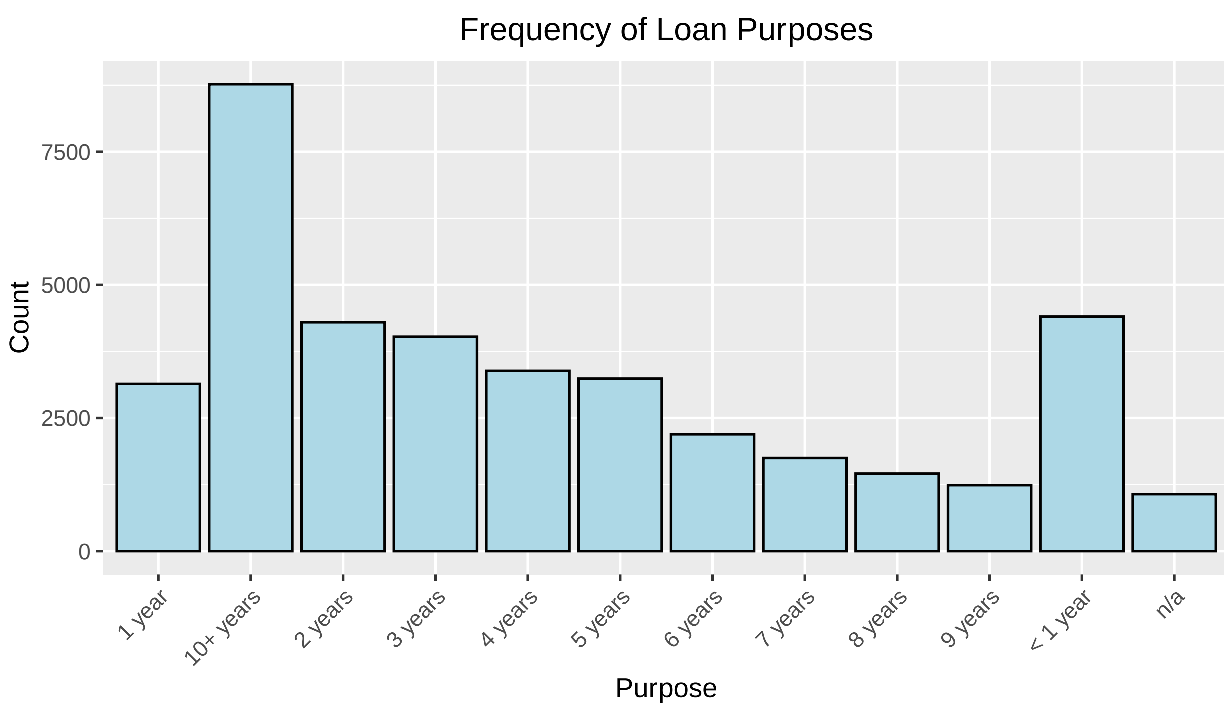
Appendix

Below are the distribution of data:









Below is the result of first Logistic Regression without filtering variables.

```
fit1 <- glm(loan_status~loan_amnt+int_rate+installment+factor(term) +factor(grade) + factor(purpose)
            + annual_inc+ factor(home_ownership)+factor(verification_status)+factor(emp_length)+
             dti+delinq_2yrs+inq_last_6mths+open_acc+pub_rec+revol_bal+revol_util+total_acc+pub_rec_ban
             , loan_train, family=binomial(logit))
summary(fit1, results=TRUE)
```

```
##
## Call:
## glm(formula = loan_status ~ loan_amnt + int_rate + installment +
##     factor(term) + factor(grade) + factor(purpose) + annual_inc +
##     factor(home_ownership) + factor(verification_status) + factor(emp_length) +
##     dti + delinq_2yrs + inq_last_6mths + open_acc + pub_rec +
##     revol_bal + revol_util + total_acc + pub_rec_bankruptcies,
##     family = binomial(logit), data = loan_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4278  -0.5895  -0.4558  -0.3225   4.3391
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                               -3.767e+00  1.766e-01 -21.333
## loan_amnt                                  1.052e-05  8.631e-06   1.219
## int_rate                                   1.119e+01  1.695e+00   6.606
## installment                               -3.420e-04  3.020e-04  -1.133
## factor(term)60_months                      4.748e-01  5.985e-02   7.933
## factor(grade)B                             1.154e-01  8.291e-02   1.391
## factor(grade)C                             8.908e-02  1.179e-01   0.756
## factor(grade)D                             6.712e-02  1.515e-01   0.443
## factor(grade)E                            -9.900e-03  1.835e-01  -0.054
## factor(grade)F                            -1.136e-01  2.230e-01  -0.510
## factor(grade)G                            -1.858e-01  2.722e-01  -0.683
## factor(purpose)credit_card                -7.139e-02  1.121e-01  -0.637
## factor(purpose)debt_consolidation          1.630e-01  1.019e-01   1.599
## factor(purpose)educational                 4.655e-01  2.053e-01   2.267
## factor(purpose)home_improvement            1.118e-01  1.192e-01   0.938
## factor(purpose)house                       2.096e-01  1.922e-01   1.090
## factor(purpose)major_purchase             -2.162e-02  1.277e-01  -0.169
## factor(purpose)medical                     3.219e-01  1.573e-01   2.047
## factor(purpose)moving                      4.052e-01  1.653e-01   2.452
## factor(purpose)other                       3.942e-01  1.098e-01   3.592
## factor(purpose)renewable_energy            5.971e-01  3.137e-01   1.903
## factor(purpose)small_business              9.259e-01  1.177e-01   7.868
## factor(purpose)vacation                    2.189e-01  2.043e-01   1.071
## factor(purpose)wedding                    -7.850e-02  1.578e-01  -0.498
## annual_inc                                -6.718e-06  6.308e-07 -10.650
## factor(home_ownership)OTHER                7.499e-01  2.924e-01   2.564
## factor(home_ownership)OWN                  6.100e-02  6.754e-02   0.903
## factor(home_ownership)RENT                 9.913e-02  4.014e-02   2.470
## factor(verification_status)Source Verified  1.426e-02  4.404e-02   0.324
## factor(verification_status)Verified        7.181e-03  4.455e-02   0.161
## factor(emp_length)10+ years                1.020e-01  7.021e-02   1.453
## factor(emp_length)2 years                 -6.421e-02  7.855e-02  -0.817
## factor(emp_length)3 years                  7.980e-03  7.934e-02   0.101
## factor(emp_length)4 years                 -4.221e-02  8.312e-02  -0.508
## factor(emp_length)5 years                  2.592e-02  8.357e-02   0.310
## factor(emp_length)6 years                 -7.551e-03  9.268e-02  -0.081
## factor(emp_length)7 years                  1.994e-02  9.860e-02   0.202
## factor(emp_length)8 years                  4.883e-02  1.060e-01   0.460
## factor(emp_length)9 years                 -8.422e-02  1.165e-01  -0.723
```

```
## factor(emp_length)< 1 year                 -5.107e-02  7.813e-02  -0.654
## factor(emp_length)n/a                       5.298e-01  1.077e-01   4.921
## dti                                         3.729e-04  3.007e-03   0.124
## delinq_2yrs                                -2.717e-02  3.420e-02  -0.794
## inq_last_6mths                              1.357e-01  1.555e-02   8.727
## open_acc                                    6.160e-03  5.485e-03   1.123
## pub_rec                                     2.284e-01  1.113e-01   2.053
## revol_bal                                   3.641e-06  1.397e-06   2.606
## revol_util                                  4.468e-01  8.011e-02   5.577
## total_acc                                  -1.200e-03  2.270e-03  -0.529
## pub_rec_bankruptcies                        8.167e-02  1.295e-01   0.631
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## loan_amnt                                  0.222788
## int_rate                                   3.94e-11 ***
## installment                                0.257388
## factor(term)60_months                      2.13e-15 ***
## factor(grade)B                             0.164083
## factor(grade)C                             0.449862
## factor(grade)D                             0.657674
## factor(grade)E                             0.956974
## factor(grade)F                             0.610307
## factor(grade)G                             0.494841
## factor(purpose)credit_card                 0.524055
## factor(purpose)debt_consolidation          0.109817
## factor(purpose)educational                 0.023394 *
## factor(purpose)home_improvement            0.348297
## factor(purpose)house                       0.275525
## factor(purpose)major_purchase              0.865518
## factor(purpose)medical                     0.040703 *
## factor(purpose)moving                      0.014207 *
## factor(purpose)other                       0.000329 ***
## factor(purpose)renewable_energy            0.056994 .
## factor(purpose)small_business              3.60e-15 ***
## factor(purpose)vacation                    0.283990
## factor(purpose)wedding                     0.618788
## annual_inc                                  < 2e-16 ***
## factor(home_ownership)OTHER                0.010336 *
## factor(home_ownership)OWN                  0.366378
## factor(home_ownership)RENT                 0.013517 *
## factor(verification_status)Source Verified 0.746116
## factor(verification_status)Verified        0.871929
## factor(emp_length)10+ years                0.146127
## factor(emp_length)2 years                  0.413700
## factor(emp_length)3 years                  0.919888
## factor(emp_length)4 years                  0.611587
## factor(emp_length)5 years                  0.756393
## factor(emp_length)6 years                  0.935067
## factor(emp_length)7 years                  0.839769
## factor(emp_length)8 years                  0.645167
## factor(emp_length)9 years                  0.469523
## factor(emp_length)< 1 year                 0.513294
## factor(emp_length)n/a                      8.59e-07 ***
## dti                                        0.901320
```

```
## delinq_2yrs                                0.426990
## inq_last_6mths                              < 2e-16 ***
## open_acc                                   0.261392
## pub_rec                                    0.040118 *
## revol_bal                                  0.009168 **
## revol_util                                 2.45e-08 ***
## total_acc                                  0.596976
## pub_rec_bankruptcies                       0.528362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25447  on 31176  degrees of freedom
## Residual deviance: 23537  on 31127  degrees of freedom
## AIC: 23637
##
## Number of Fisher Scoring iterations: 5
```

Final model:

```
fit2 <- glm(loan_status~loan_amnt+int_rate+installment+factor(term)+factor(grade)
            +factor(purpose)+annual_inc+factor(home_ownership)+
            +dti+inq_last_6mths+pub_rec+revol_bal+revol_util
             ,loan_train, family=binomial(logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit2, results=TRUE)
```

```
##
## Call:
## glm(formula = loan_status ~ loan_amnt + int_rate + installment +
##     factor(term) + factor(grade) + factor(purpose) + annual_inc +
##     factor(home_ownership) + +dti + inq_last_6mths + pub_rec +
##     revol_bal + revol_util, family = binomial(logit), data = loan_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4742  -0.5903  -0.4573  -0.3244   4.3610
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.704e+00  1.649e-01 -22.468  < 2e-16 ***
## loan_amnt                     1.098e-05  8.547e-06   1.284 0.199103
## int_rate                      1.119e+01  1.686e+00   6.637 3.19e-11 ***
## installment                  -3.386e-04  3.012e-04  -1.124 0.260851
## factor(term)60_months         4.882e-01  5.909e-02   8.262  < 2e-16 ***
## factor(grade)B                1.038e-01  8.258e-02   1.257 0.208845
## factor(grade)C                6.984e-02  1.173e-01   0.595 0.551671
## factor(grade)D                4.913e-02  1.508e-01   0.326 0.744639
## factor(grade)E               -3.777e-02  1.828e-01  -0.207 0.836280
```

```
## factor(grade)F                    -1.377e-01  2.222e-01  -0.620 0.535421
## factor(grade)G                    -2.214e-01  2.712e-01  -0.817 0.414116
## factor(purpose)credit_card        -5.893e-02  1.119e-01  -0.527 0.598420
## factor(purpose)debt_consolidation  1.738e-01  1.018e-01   1.708 0.087648 .
## factor(purpose)educational         4.640e-01  2.050e-01   2.264 0.023591 *
## factor(purpose)home_improvement    1.272e-01  1.191e-01   1.069 0.285143
## factor(purpose)house               2.140e-01  1.923e-01   1.113 0.265725
## factor(purpose)major_purchase     -1.774e-02  1.276e-01  -0.139 0.889440
## factor(purpose)medical             3.379e-01  1.571e-01   2.151 0.031474 *
## factor(purpose)moving              4.195e-01  1.649e-01   2.545 0.010939 *
## factor(purpose)other               4.088e-01  1.097e-01   3.728 0.000193 ***
## factor(purpose)renewable_energy    6.589e-01  3.124e-01   2.109 0.034951 *
## factor(purpose)small_business      9.260e-01  1.176e-01   7.875 3.41e-15 ***
## factor(purpose)vacation            2.716e-01  2.034e-01   1.335 0.181876
## factor(purpose)wedding            -8.612e-02  1.576e-01  -0.546 0.584799
## annual_inc                        -6.907e-06  6.070e-07 -11.379  < 2e-16 ***
## factor(home_ownership)OTHER        7.190e-01  2.923e-01   2.460 0.013905 *
## factor(home_ownership)OWN          7.419e-02  6.709e-02   1.106 0.268758
## factor(home_ownership)RENT         7.522e-02  3.874e-02   1.942 0.052171 .
## dti                                9.963e-04  2.800e-03   0.356 0.721979
## inq_last_6mths                     1.364e-01  1.541e-02   8.851  < 2e-16 ***
## pub_rec                            3.218e-01  6.060e-02   5.310 1.09e-07 ***
## revol_bal                          4.328e-06  1.365e-06   3.171 0.001518 **
## revol_util                         4.343e-01  7.597e-02   5.717 1.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25447  on 31176  degrees of freedom
## Residual deviance: 23581  on 31144  degrees of freedom
## AIC: 23647
##
## Number of Fisher Scoring iterations: 5
```

Anova test:

```
anova(fit1,fit2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: loan_status ~ loan_amnt + int_rate + installment + factor(term) +
##     factor(grade) + factor(purpose) + annual_inc + factor(home_ownership) +
##     factor(verification_status) + factor(emp_length) + dti +
##     delinq_2yrs + inq_last_6mths + open_acc + pub_rec + revol_bal +
##     revol_util + total_acc + pub_rec_bankruptcies
## Model 2: loan_status ~ loan_amnt + int_rate + installment + factor(term) +
##     factor(grade) + factor(purpose) + annual_inc + factor(home_ownership) +
##     +dti + inq_last_6mths + pub_rec + revol_bal + revol_util
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1     31127      23537
## 2     31144      23581 -17  -43.911 0.000353 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```