Jonathan Ling (Group members: Duncan Black, Fernando Machado, Hanson Qin)
STAT222
Write-Up #1

Credit Risk Prediction Using Supervised Learning

Problem Description

Credit risk is defined as the risk of loss from a borrower defaulting on a loan. Ideally, lending companies would like to minimize losses by rejecting loans for applicants who will default. While it is impossible to know exactly which loans will default in the future, it is feasible to estimate the likelihood of default so that lending companies can make better judgements when offering loans. In our project, we aim to use supervised learning methods to predict the probability of default when given characteristics of the borrower. Our data comes from LendingClub, a peer-to-peer lending company that connects individual borrowers with investors for a variety of loan types. In the dataset, we have instances of past loans that have been either fully paid off or defaulted, along with information on the borrower of each loan. By creating a robust model to predict credit risk using this data, we can answer the following questions:

How can we best estimate the probability of default?

What factors are relevant in predicting whether or not a loan will default, and to what degree?

What type(s) of models perform better on loan data?

If we can answer these questions, our results can be applied not only to LendingClub, but also to other financial institutions that are interested in the credit risk of their clients.

Data Description

The dataset we are using comes from Kaggle, uploaded by a user named Nathan George. Kaggle is a site for data scientists and machine learning engineers to publish datasets and build models. Nathan George obtained the data directly from LendingClub's website and compiled the many smaller datasets available into a single dataset of approximately 2 million loans. Since we are trying to predict if a loan will default, we narrowed down the observations corresponding to completed loans only (status = "defaulted" or "fully paid"). After this, our dataset contains about 1.3 million relevant observations with interspersed missing values (Fig. 1). Approximately 80% of the loans were fully paid, while the remaining 20% defaulted. In addition, the data includes numerous characteristics of the borrowers such as annual income, FICO score ranges, and information about other credit accounts.

Initially, we hypothesized that the FICO score, employment length, home ownership status, and loan amount would be the most significant predictors of the final loan status. FICO score is a type of credit score from a credit bureau representing an individual's credit worthiness, so it makes sense that those with a reliable credit history are less likely to default on a loan (Fig. 2). Similarly, individuals who own a house or have been employed for longer periods of time would likely have stable credit and be less likely to default compared to renters (Fig. 3). The size of the loan may also influence risk - larger loans are more difficult to repay, and are thus riskier.

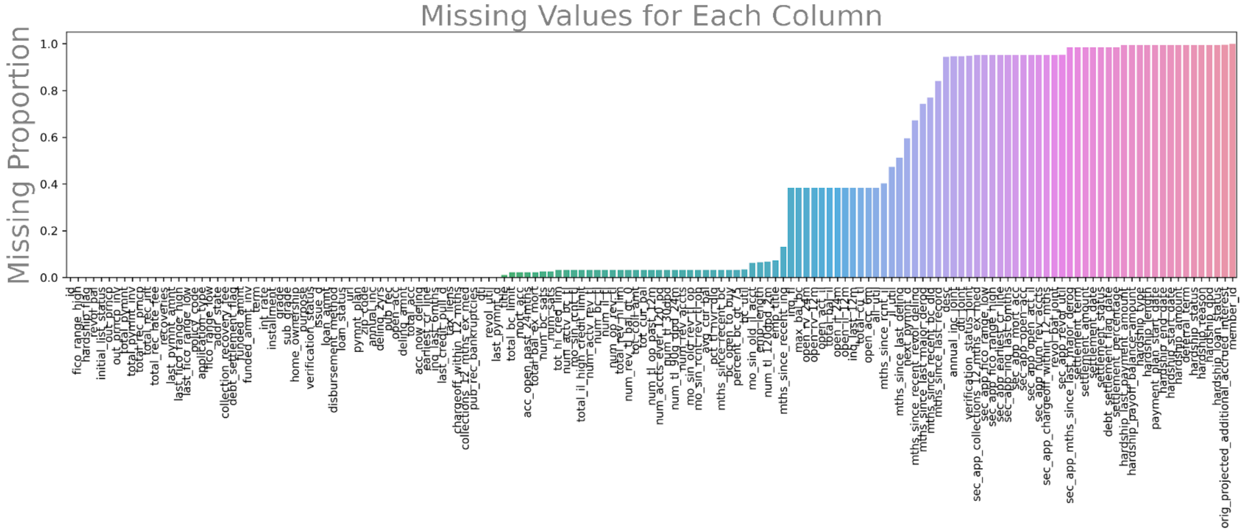


Figure 1: *The percentage of missing values in each column.*

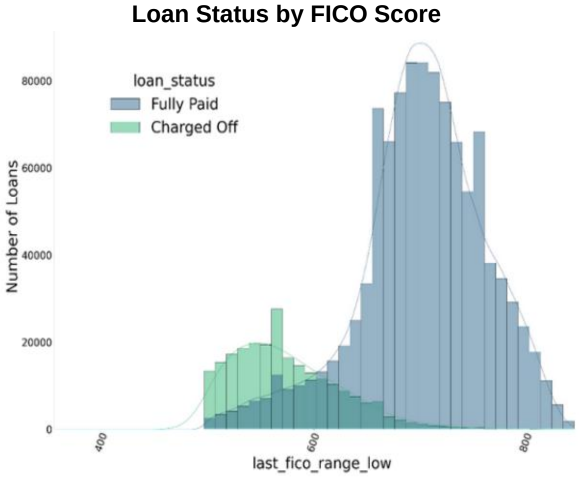
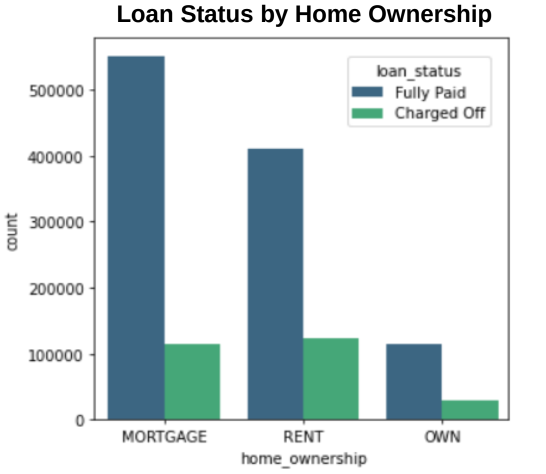


Figure 2: *Defaulted loans correspond with lower FICO scores.*



```
Default rate for home ownership = MORTGAGE: 0.16574626838522852
Default rate for home ownership = RENT: 0.22644589879058816
Default rate for home ownership = OWN: 0.19495008807985909
```

Figure 3: *Renters have the highest proportion of default versus homeowners and mortgagors.*

Methods

We first performed some data cleaning by removing irrelevant or highly correlated columns. Certain columns had over 90% of their values missing, so we removed those as well. For the remaining missing data, we analyzed the ratio of defaulted vs. fully paid loans before and after removing the observations with missing data, and found that they were roughly the same. This leads us to believe that the data is missing at random (MAR), meaning we can drop these observations. This leaves us with around 1 million observations in the cleaned dataset. Next, we split the dataset into training, validation, and test sets using a 80/10/10 ratio.

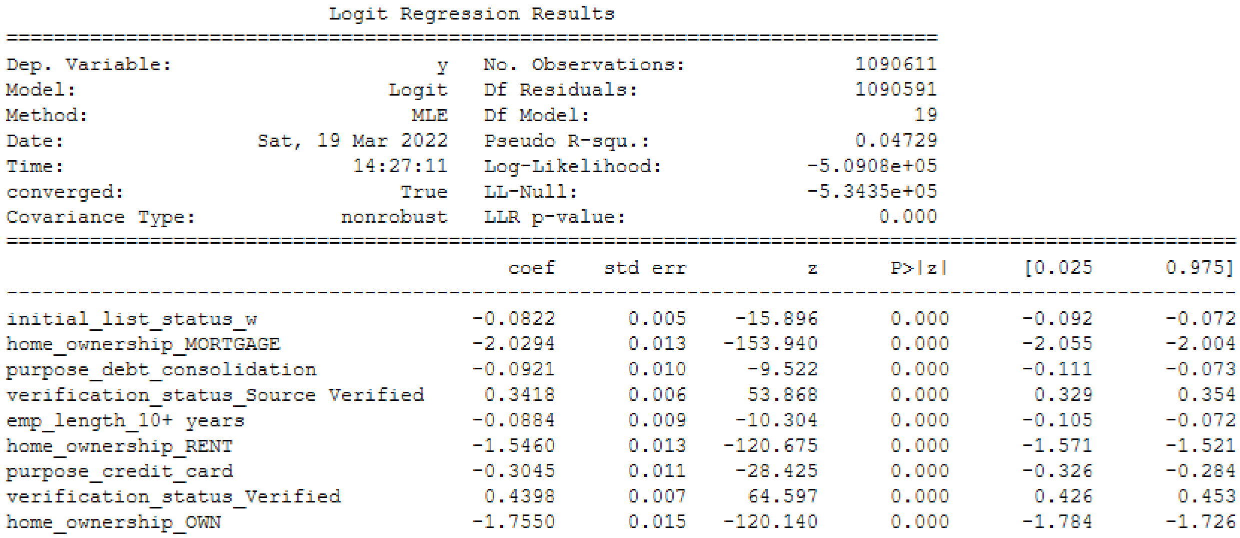
We started our modeling with logistic regression. Logistic regression allows for prediction on a binary outcome variable; in our case, we take *loan_status* = 1 to indicate a defaulted loan and *loan_status* = 0 to indicate a fully paid loan. The logit model allows us to estimate the expected value of the *loan_status* variable, which is equivalent to the probability of default denoted as *P* in the following equation:

$$ln(\frac{P}{1 - P}) \ = \beta_0 + \beta_1 x$$

Or equivalently,

$$P(loan\_status \, = 1) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

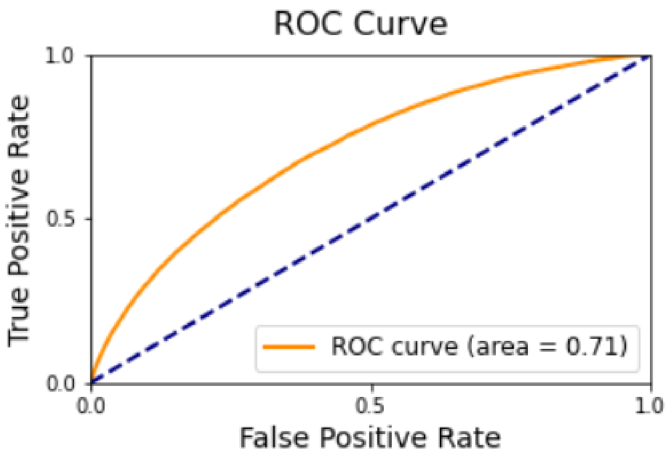
where x is the covariate vector or matrix. The coefficient β_1 represents the log odds ratio associated with a unit change in the predictor variable. From the regression results (Fig. 4), we can see which predictors are associated with higher odds (positive coefficient) and lower odds (negative coefficient) of loan default and the significance of each covariate. We performed univariate logistic regressions of *loan_status* against each covariate to judge their significance (Appendix Fig. A1). Then, we used the most significant covariates and fit them in a multiple logistic regression model to obtain the coefficients and predictions. We determined the number of covariates to use with our validation set.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:              1090611
Model:                          Logit   Df Residuals:                  1090591
Method:                           MLE   Df Model:                           19
Date:                Sat, 19 Mar 2022   Pseudo R-squ.:                 0.04729
Time:                        14:27:11   Log-Likelihood:            -5.0908e+05
converged:                       True   LL-Null:                   -5.3435e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
=======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
initial_list_status_w               -0.0822      0.005    -15.896      0.000      -0.092      -0.072
home_ownership_MORTGAGE             -2.0294      0.013   -153.940      0.000      -2.055      -2.004
purpose_debt_consolidation          -0.0921      0.010     -9.522      0.000      -0.111      -0.073
verification_status_Source Verified  0.3418      0.006     53.868      0.000       0.329       0.354
emp_length_10+ years                -0.0884      0.009    -10.304      0.000      -0.105      -0.072
home_ownership_RENT                 -1.5460      0.013   -120.675      0.000      -1.571      -1.521
purpose_credit_card                 -0.3045      0.011    -28.425      0.000      -0.326      -0.284
verification_status_Verified         0.4398      0.007     64.597      0.000       0.426       0.453
home_ownership_OWN                  -1.7550      0.015   -120.140      0.000      -1.784      -1.726
```

Figure 4: *Multiple logistic regression partial results. Full results can be seen in the appendix.*

From this model, we can obtain predictions of the probability of default for new observations. To convert the probabilities into binary predictions of default vs fully paid, we used the area under the receiver operating characteristic curve to determine the optimal threshold to separate default predictions and fully paid predictions. We obtained an optimal threshold of 0.183 (Fig. 5).



```
Optimal Threshold: 0.18303220407856516
```

Figure 5: *Receiver operating characteristic curve example.*

After trying logistic regression, we decided to compare the results with a more modern machine learning algorithm called XGBoost, or Extreme Gradient Boosting. This algorithm has risen in popularity recently due to its high performance in machine learning competitions. It uses several decision trees in tandem in order to predict the outcome variable by minimizing a loss function. We applied the algorithm on our training set, used the validation set to tune the hyperparameters, and obtained predictions on the test set. The most significant predictors can be seen below (Fig. 6).

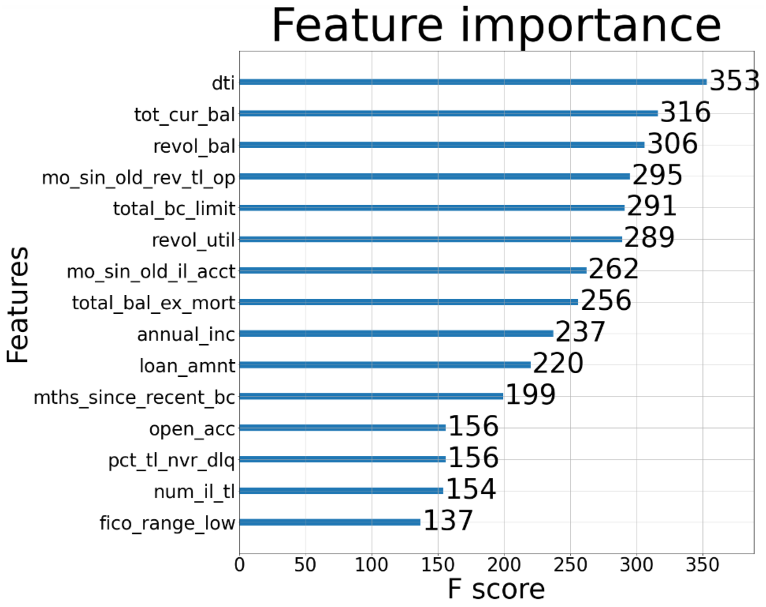
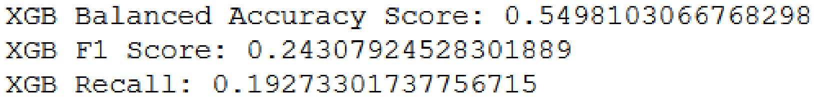


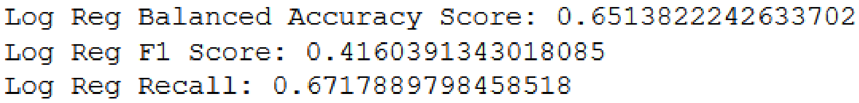
Figure 6: *Feature importance for the top 15 variables in the XGBoost model.*

Results

In order to compare our results effectively, we scored each model's predictions primarily using the balanced accuracy score metric. This metric is effective when dealing with a skewed outcome variable since it rebalances the accuracy score to be between 0.50 and 1. In our dataset, about 80% of the observations are fully paid and 20% are defaulted, so the simple accuracy score would be misleading, as guessing "fully paid" for every prediction would lead to a 0.80 accuracy score despite clearly being inaccurate. Other useful metrics include the F1 Score, which also accounts for skewed distribution of the outcome variable, and the recall, which is the ratio of true positive predictions to the sum of true positives and false negatives. Intuitively, this is the ability of the model to correctly predict the positive results (defaulted loans). The scores are shown below:



```
XGB Balanced Accuracy Score: 0.5498103066768298
XGB F1 Score: 0.24307924528301889
XGB Recall: 0.19273301737756715
```



```
Log Reg Balanced Accuracy Score: 0.6513822242633702
Log Reg F1 Score: 0.4160391343018085
Log Reg Recall: 0.6717889798458518
```

It appears that the logistic regression model has much higher scores than the XGBoost model. This is likely due to the high sensitivity of the XGBoost model to changes in the data and parameters, or possibly the threshold used to determine if a prediction will default or not.

Conclusions

So far, we have applied two supervised learning models to the data with varying results. We will continue to adjust our models and implement new models such as LASSO and ridge regression to try to increase our prediction accuracy. We believe LASSO may be a more appropriate method, as we have many covariates with near-zero impact on the outcome variable. Our next step is to create a metric for comparing our predictions to LendingClub's loan "grade" system, which is a score from A-G accompanied with a number (e.g., A2, C5, F1) that represents LendingClub's confidence in the loan (Fig. 5).

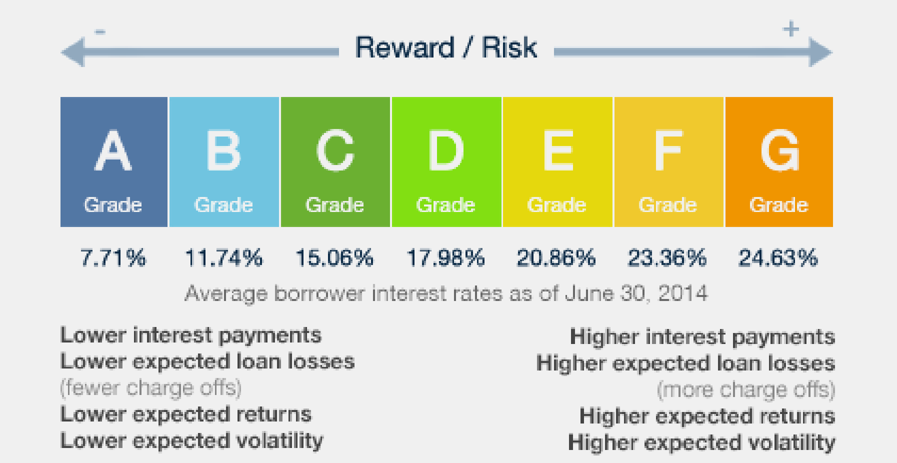
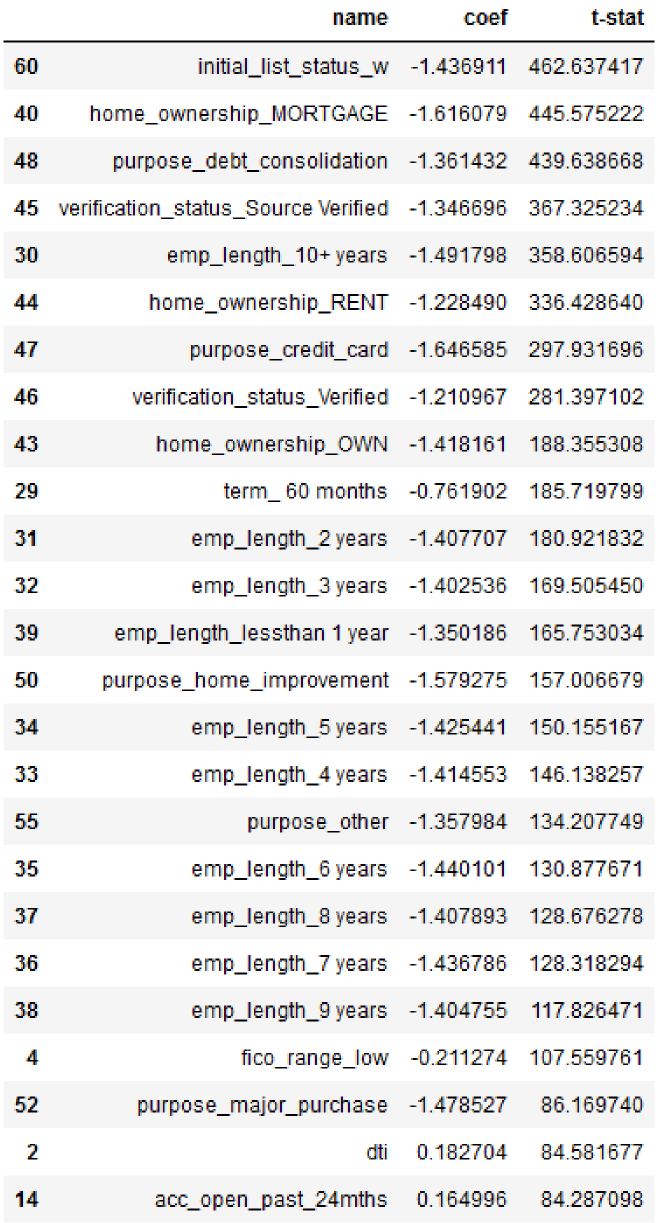


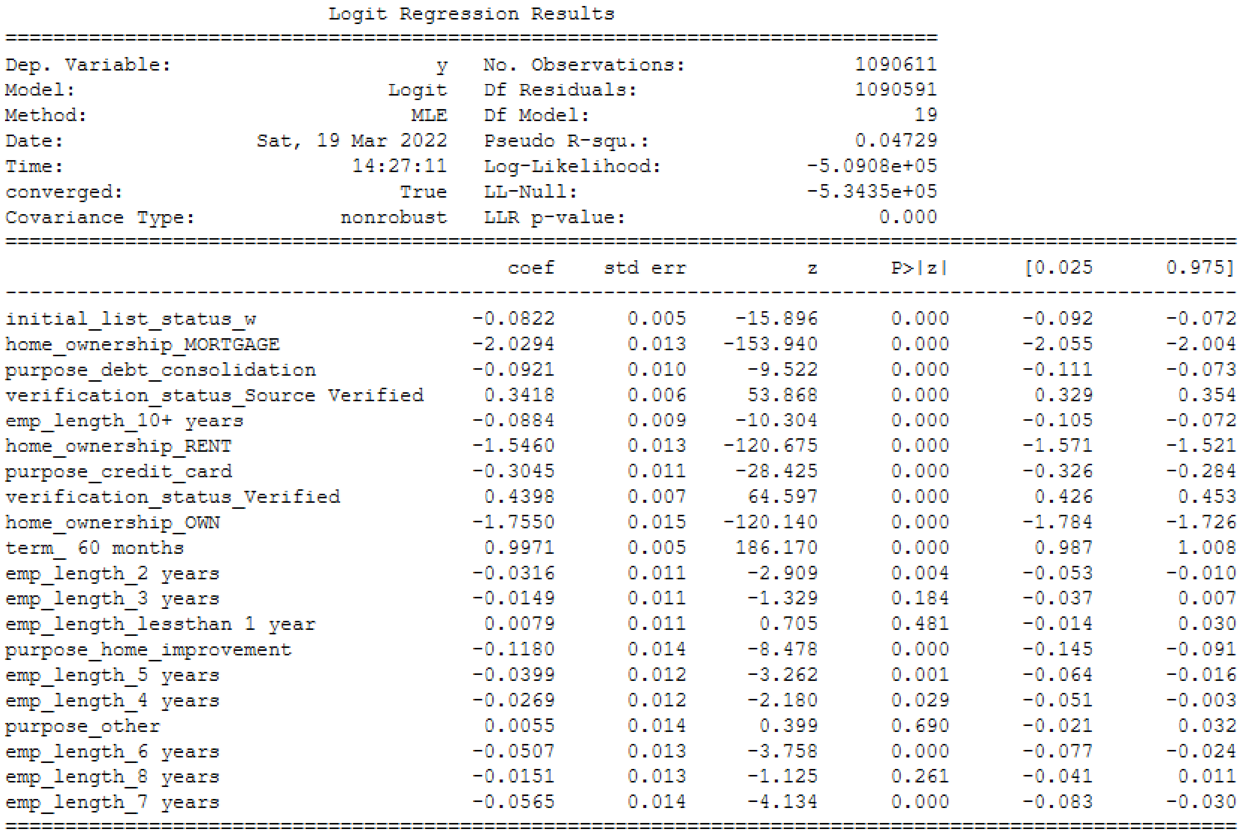
Figure 5: *LendingClub grading system and interpretation.*

Appendix



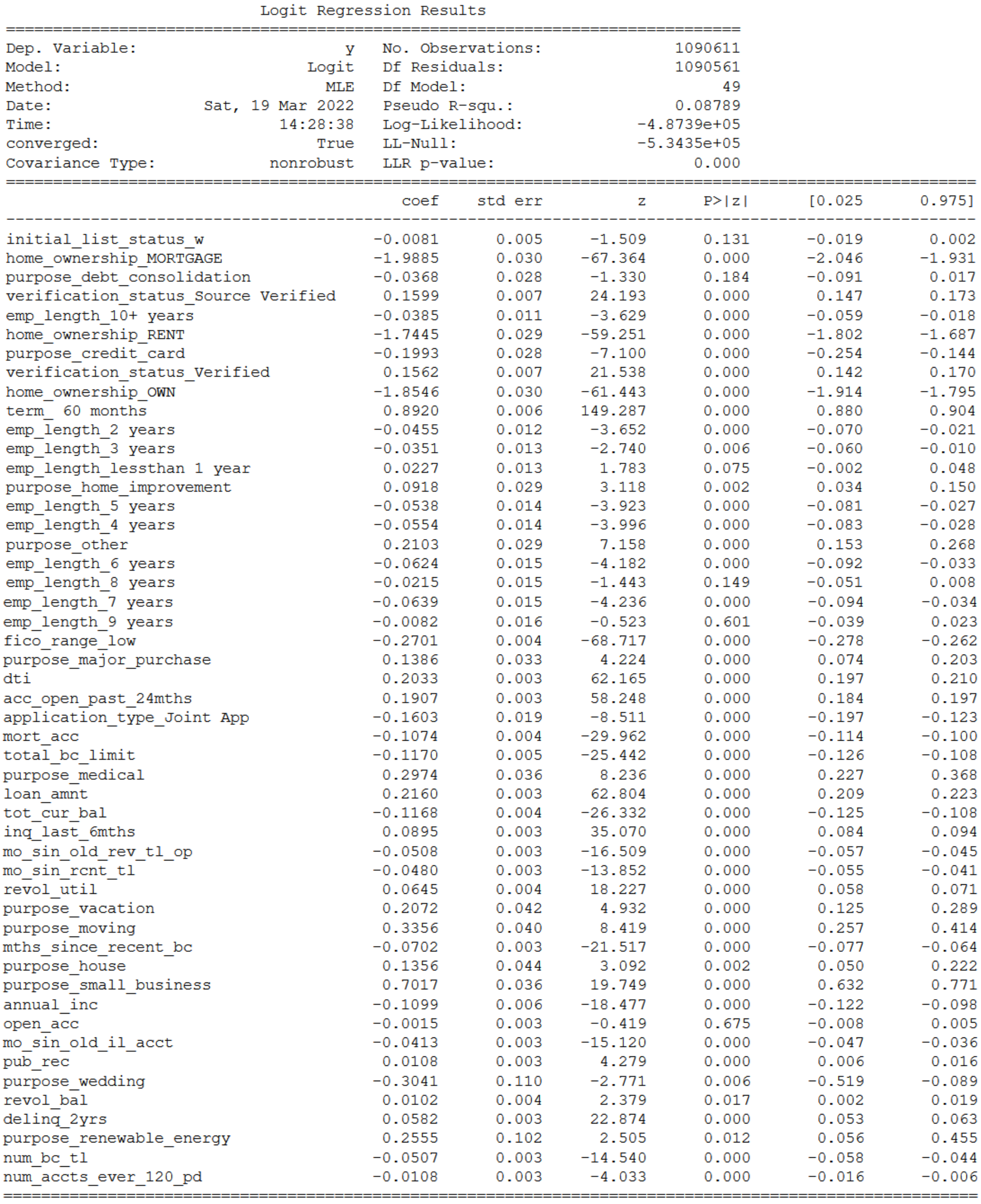
	name	coef	t-stat
60	initial_list_status_w	-1.436911	462.637417
40	home_ownership_MORTGAGE	-1.616079	445.575222
48	purpose_debt_consolidation	-1.361432	439.638668
45	verification_status_Source Verified	-1.346696	367.325234
30	emp_length_10+ years	-1.491798	358.606594
44	home_ownership_RENT	-1.228490	336.428640
47	purpose_credit_card	-1.646585	297.931696
46	verification_status_Verified	-1.210967	281.397102
43	home_ownership_OWN	-1.418161	188.355308
29	term_ 60 months	-0.761902	185.719799
31	emp_length_2 years	-1.407707	180.921832
32	emp_length_3 years	-1.402536	169.505450
39	emp_length_lessthan 1 year	-1.350186	165.753034
50	purpose_home_improvement	-1.579275	157.006679
34	emp_length_5 years	-1.425441	150.155167
33	emp_length_4 years	-1.414553	146.138257
55	purpose_other	-1.357984	134.207749
35	emp_length_6 years	-1.440101	130.877671
37	emp_length_8 years	-1.407893	128.676278
36	emp_length_7 years	-1.436786	128.318294
38	emp_length_9 years	-1.404755	117.826471
4	fico_range_low	-0.211274	107.559761
52	purpose_major_purchase	-1.478527	86.169740
2	dti	0.182704	84.581677
14	acc_open_past_24mths	0.164996	84.287098

Figure A1: *Univariate logistic regression results for the 25 most significant covariates.*



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:                      y   No. Observations:              1090611
Model:                          Logit   Df Residuals:                  1090591
Method:                           MLE   Df Model:                           19
Date:                Sat, 19 Mar 2022   Pseudo R-squ.:                 0.04729
Time:                        14:27:11   Log-Likelihood:            -5.0908e+05
converged:                       True   LL-Null:                   -5.3435e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
=======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
initial_list_status_w                  -0.0822      0.005    -15.896      0.000      -0.092      -0.072
home_ownership_MORTGAGE                -2.0294      0.013   -153.940      0.000      -2.055      -2.004
purpose_debt_consolidation             -0.0921      0.010     -9.522      0.000      -0.111      -0.073
verification_status_Source Verified     0.3418      0.006     53.868      0.000       0.329       0.354
emp_length_10+ years                   -0.0884      0.009    -10.304      0.000      -0.105      -0.072
home_ownership_RENT                    -1.5460      0.013   -120.675      0.000      -1.571      -1.521
purpose_credit_card                    -0.3045      0.011    -28.425      0.000      -0.326      -0.284
verification_status_Verified            0.4398      0.007     64.597      0.000       0.426       0.453
home_ownership_OWN                     -1.7550      0.015   -120.140      0.000      -1.784      -1.726
term_ 60 months                         0.9971      0.005    186.170      0.000       0.987       1.008
emp_length_2 years                     -0.0316      0.011     -2.909      0.004      -0.053      -0.010
emp_length_3 years                     -0.0149      0.011     -1.329      0.184      -0.037       0.007
emp_length_lessthan 1 year              0.0079      0.011      0.705      0.481      -0.014       0.030
purpose_home_improvement               -0.1180      0.014     -8.478      0.000      -0.145      -0.091
emp_length_5 years                     -0.0399      0.012     -3.262      0.001      -0.064      -0.016
emp_length_4 years                     -0.0269      0.012     -2.180      0.029      -0.051      -0.003
purpose_other                           0.0055      0.014      0.399      0.690      -0.021       0.032
emp_length_6 years                     -0.0507      0.013     -3.758      0.000      -0.077      -0.024
emp_length_8 years                     -0.0151      0.013     -1.125      0.261      -0.041       0.011
emp_length_7 years                     -0.0565      0.014     -4.134      0.000      -0.083      -0.030
=======================================================================================================
```

Figure A2: *Multiple logistic regression full results for the 20 most significant covariates.*



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:              1090611
Model:                          Logit   Df Residuals:                  1090561
Method:                           MLE   Df Model:                           49
Date:                Sat, 19 Mar 2022   Pseudo R-squ.:                 0.08789
Time:                        14:28:38   Log-Likelihood:            -4.8739e+05
converged:                       True   LL-Null:                   -5.3435e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
=====================================================================================================
                                            coef   std err          z     P>|z|     [0.025     0.975]
-----------------------------------------------------------------------------------------------------
initial_list_status_w                    -0.0081     0.005     -1.509     0.131     -0.019      0.002
home_ownership_MORTGAGE                  -1.9885     0.030    -67.364     0.000     -2.046     -1.931
purpose_debt_consolidation               -0.0368     0.028     -1.330     0.184     -0.091      0.017
verification_status_Source Verified       0.1599     0.007     24.193     0.000      0.147      0.173
emp_length_10+ years                     -0.0385     0.011     -3.629     0.000     -0.059     -0.018
home_ownership_RENT                      -1.7445     0.029    -59.251     0.000     -1.802     -1.687
purpose_credit_card                      -0.1993     0.028     -7.100     0.000     -0.254     -0.144
verification_status_Verified              0.1562     0.007     21.538     0.000      0.142      0.170
home_ownership_OWN                       -1.8546     0.030    -61.443     0.000     -1.914     -1.795
term_ 60 months                           0.8920     0.006    149.287     0.000      0.880      0.904
emp_length_2 years                       -0.0455     0.012     -3.652     0.000     -0.070     -0.021
emp_length_3 years                       -0.0351     0.013     -2.740     0.006     -0.060     -0.010
emp_length_lessthan 1 year                0.0227     0.013      1.783     0.075     -0.002      0.048
purpose_home_improvement                  0.0918     0.029      3.118     0.002      0.034      0.150
emp_length_5 years                       -0.0538     0.014     -3.923     0.000     -0.081     -0.027
emp_length_4 years                       -0.0554     0.014     -3.996     0.000     -0.083     -0.028
purpose_other                             0.2103     0.029      7.158     0.000      0.153      0.268
emp_length_6 years                       -0.0624     0.015     -4.182     0.000     -0.092     -0.033
emp_length_8 years                       -0.0215     0.015     -1.443     0.149     -0.051      0.008
emp_length_7 years                       -0.0639     0.015     -4.236     0.000     -0.094     -0.034
emp_length_9 years                       -0.0082     0.016     -0.523     0.601     -0.039      0.023
fico_range_low                           -0.2701     0.004    -68.717     0.000     -0.278     -0.262
purpose_major_purchase                    0.1386     0.033      4.224     0.000      0.074      0.203
dti                                       0.2033     0.003     62.165     0.000      0.197      0.210
acc_open_past_24mths                      0.1907     0.003     58.248     0.000      0.184      0.197
application_type_Joint App               -0.1603     0.019     -8.511     0.000     -0.197     -0.123
mort_acc                                 -0.1074     0.004    -29.962     0.000     -0.114     -0.100
total_bc_limit                           -0.1170     0.005    -25.442     0.000     -0.126     -0.108
purpose_medical                           0.2974     0.036      8.236     0.000      0.227      0.368
loan_amnt                                 0.2160     0.003     62.804     0.000      0.209      0.223
tot_cur_bal                              -0.1168     0.004    -26.332     0.000     -0.125     -0.108
inq_last_6mths                            0.0895     0.003     35.070     0.000      0.084      0.094
mo_sin_old_rev_tl_op                     -0.0508     0.003    -16.509     0.000     -0.057     -0.045
mo_sin_rcnt_tl                           -0.0480     0.003    -13.852     0.000     -0.055     -0.041
revol_util                                0.0645     0.004     18.227     0.000      0.058      0.071
purpose_vacation                          0.2072     0.042      4.932     0.000      0.125      0.289
purpose_moving                            0.3356     0.040      8.419     0.000      0.257      0.414
mths_since_recent_bc                     -0.0702     0.003    -21.517     0.000     -0.077     -0.064
purpose_house                             0.1356     0.044      3.092     0.002      0.050      0.222
purpose_small_business                    0.7017     0.036     19.749     0.000      0.632      0.771
annual_inc                               -0.1099     0.006    -18.477     0.000     -0.122     -0.098
open_acc                                 -0.0015     0.003     -0.419     0.675     -0.008      0.005
mo_sin_old_il_acct                       -0.0413     0.003    -15.120     0.000     -0.047     -0.036
pub_rec                                   0.0108     0.003      4.279     0.000      0.006      0.016
purpose_wedding                          -0.3041     0.110     -2.771     0.006     -0.519     -0.089
revol_bal                                 0.0102     0.004      2.379     0.017      0.002      0.019
delinq_2yrs                               0.0582     0.003     22.874     0.000      0.053      0.063
purpose_renewable_energy                  0.2555     0.102      2.505     0.012      0.056      0.455
num_bc_tl                                -0.0507     0.003    -14.540     0.000     -0.058     -0.044
num_accts_ever_120_pd                    -0.0108     0.003     -4.033     0.000     -0.016     -0.006
=====================================================================================================
```

Figure A3: *Multiple logistic regression full results for the 50 most significant covariates.*