**BUFN 747 Asset-Liability and Nonfinancial Risk Management**

**Spring 2023**

**Case Study 1**

**Estimating Mortgage Prepayment & Implications for Interest Rate** **Risk**

*Submitted by Group 6*

*Ashmita Kaur, Elise Parker, Nitin Kailadath, Sarthak Rajani,*

*Siddharth Rao, Yekaalo Gabremeskel*

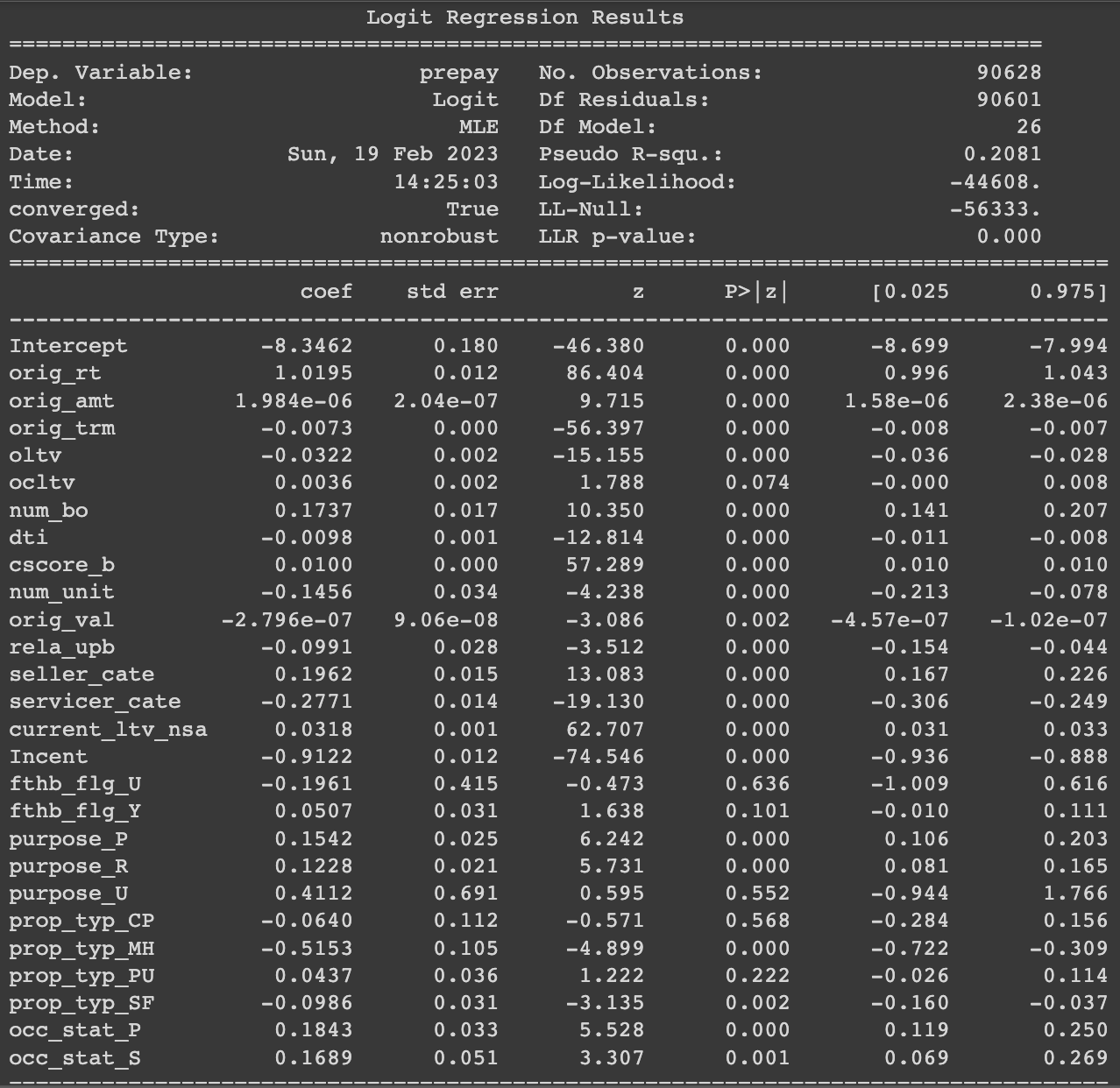
1. **Create a variable defined as INCENT (Borrower Note Rate – Market Rate) where Borrower Note Rate is orig\_rt from the loan data and Market Rate is the weekly Freddie Mac Primary Mortgage Market (PMMS) rate in the PMMS\_History.csv file. You will need to merge the loan level file in with the PMMS data. You can define the Market Rate as the PMMS rate with the closest weekly date to the last\_dte observed for the loan. Use the variable INCENT as an independent variable in your model.**

Variable “Incent” was created by using the formula ‘orig\_rate - market\_rate’. The ‘market\_rate was obtained from the PMMS rate file.

Refer to attached .ipynb file for detailed information and to look at the merged data.

1. **Estimate a logistic regression prepayment model using as candidate independent variables those noted in bold in the list above. Note – you will not want to include all variables, but rather a subset based on a model that results in variables that are statistically significant, carry the expected sign in predicting prepayment and have good overall model discriminatory power. You may need to iterate (testing different variable combinations) on the best model in terms of its ability to distinguish between loans that prepay from loans that do not (the c-statistic will be helpful here).**

After cleaning and preparing our data, we started by running a simple logistic regression on all independent variables. We were able to get a view of the variables that are significant in predicting prepayment. The following table shows all the coefficients for all the independent variables and their p-values, as we can see most variables are significant except for a few.



In order for our model to fit well, we started by standardizing all the data and followed by splitting the data between training and testing. In order to find the best fit model in our case we used a Recursive Feature Elimination for best feature selection. The RFE algorithm helped us select the features that are most relevant in predicting our target variable (prepayment). This allowed us to find the best fit model to predict prepayment. The following independent variables are the features that were selected.

1. orig\_trm
2. oltv
3. ocltv
4. num\_bo
5. dti
6. cscore\_b
7. rela\_upb
8. seller\_cate
9. servicer\_cate
10. current\_ltv\_nsa
11. Incent
12. purpose\_P
13. **Provide your hypothesis of how each of your variables affects prepayment and why it should be included in the model.**

According to us, and with the help of Recursive Feature Elimination method, the following variables affect prepayment and hence they should be included in the model.

1. **orig\_trm** - origination term for the loan

The longer the origination term for the loan, the higher chances of change in yield curve due to interest rate changes and thus the borrower is more likely to prepay and refinance.

1. **oltv** - origination loan value to property value ratio

The more portion of the total value of the property is given out as loan, more likely to default by the borrower and lesser prepayment chances.

1. **ocltv** - origination combined LTV

The higher the number of loans availed to buy a property, higher the chances of default and lesser chances of prepayment.

1. **num\_bo** - number of borrowers for mortgage

The higher the number of borrowers for a mortgage, the payment source is thus diversified resulting in lesser chance of default and higher chance of prepayment.

1. **dti** - debt to income ratio

Higher the debt issued as a proportion of the borrower’s income can lead to higher default rates and lesser prepayment chances.

1. **cscore\_b** - borrowers credit score

The better the credit score, the better the credit history of the borrower and thus more likely inclined to prepay.

1. **purpose\_P** - loan purpose for purchase (converted to numerical value)

If the purpose of the borrower is to live in the property bought as compared to renting it out or as an investment, higher the chances of prepayment.

1. **rela\_upb** - relative median unpaid principal balance

The higher the upb, higher the chances of default by the borrower and hence lower chances of repayment.

1. **seller\_cate** - seller category

The more regulated and organized the seller is, lower the chances of default and higher the chances of prepayment.

1. **servicer\_cate** - servicer category

The loans serviced by banks are less likely to default and have higher chances of prepayment as banks are highly regulated and have very efficient systems in place for credit risk analytics.

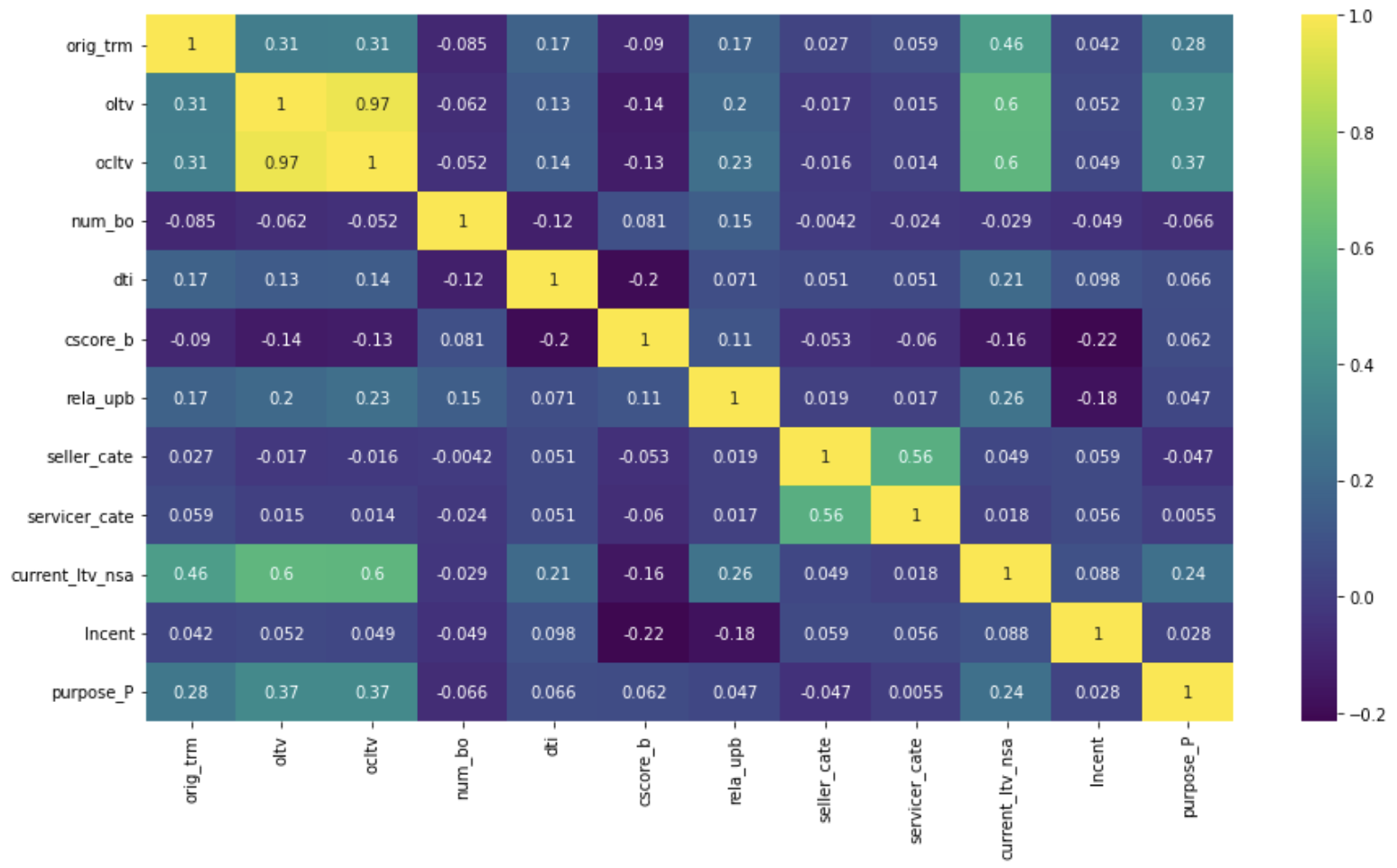
1. **current\_ltv\_nsa** - current loan to value ratio

Lower the current loan to value ratio, higher the chance of prepayment by the borrower.

1. **Incent** - borrower note rate - market rate

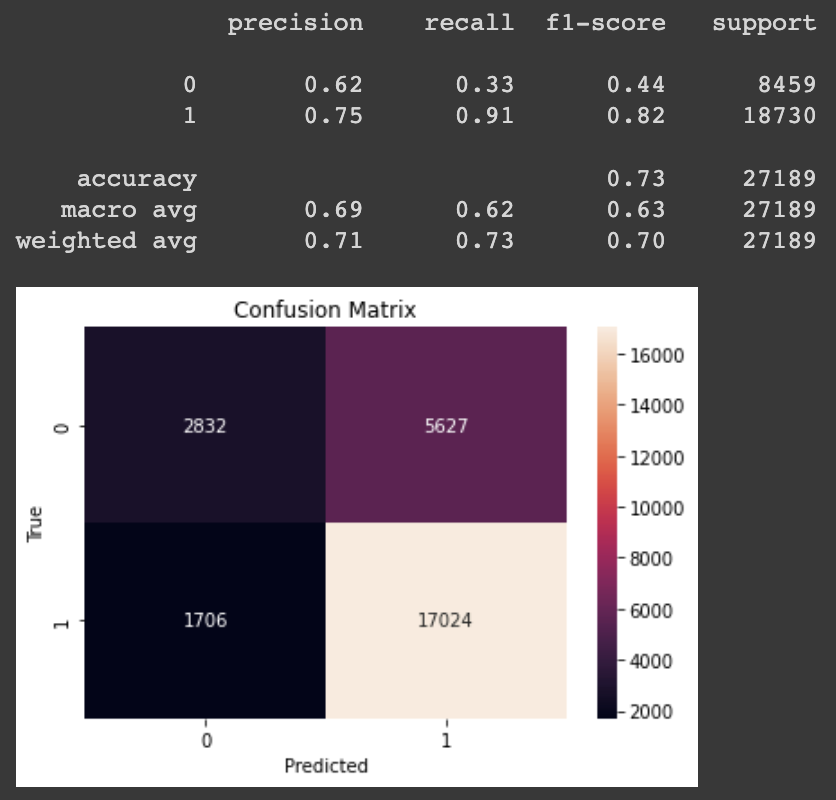
Higher the incentive, higher will be the chance of prepayment by the borrower.

1. **Provide the model output including estimated coefficients, statistical significance for each independent variable. Provide your overall assessment of the model’s discriminatory power (i.e., ability to distinguish between prepay and no prepay loans) – you may provide an area under the curve C-statistic for this.**



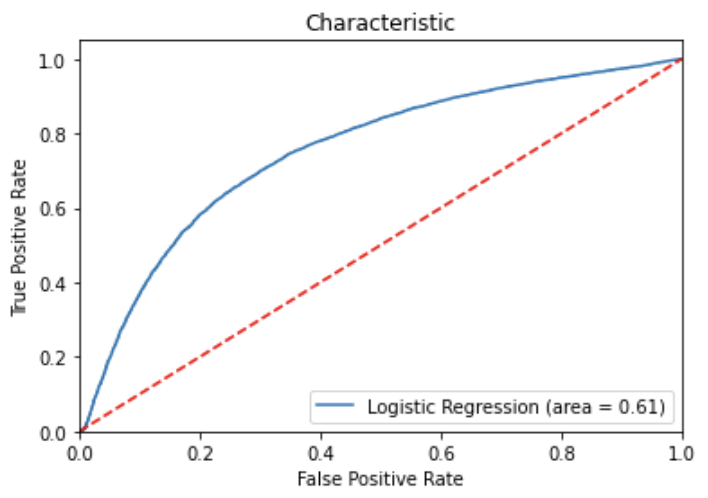
Correlation Heatmap

As we can see in the correlation matrix, we have some variables that are correlated, positively and negatively with one another. Some independent variables are correlated with one another some more than others. This may cause our model issues because of multicollinearity among our variables.



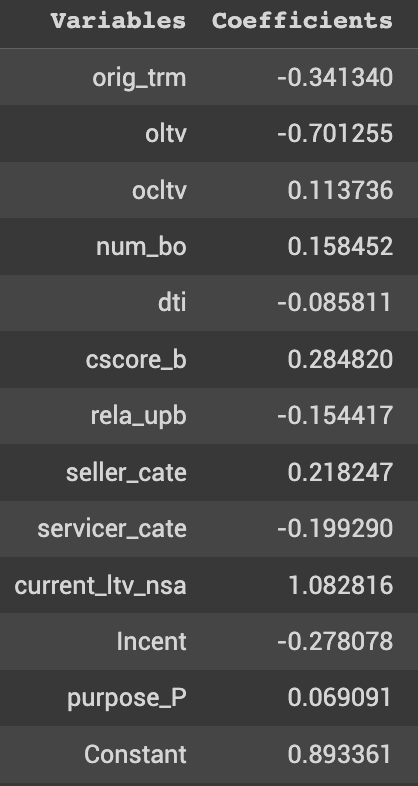
Confusion Matrix

As we observe the confusion matrix, our model had an accuracy rate of 71%. Our model was able to predict the prepayments better than the non-prepayments. This may be due to the fact that the dataset contained more prepayments rather than non-prepayments, and when splitting the data between training and testing, an unbalance might have made our model more sensitive to prepayments compared to non prepayments.



C-Curve

The ROC AUC score tells us the efficiency of our model. The higher the area, the better the model is doing in distinguishing between the positive and negative classes we are trying to predict. Our model has a fair AUC score of 0.61.



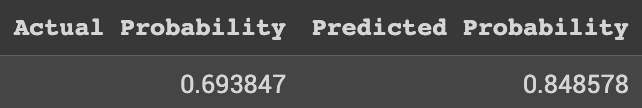
Prepayment = -0.34xorig\_trm - 0.70xoltv + 0.11xocltv

Variable - Coefficient Table

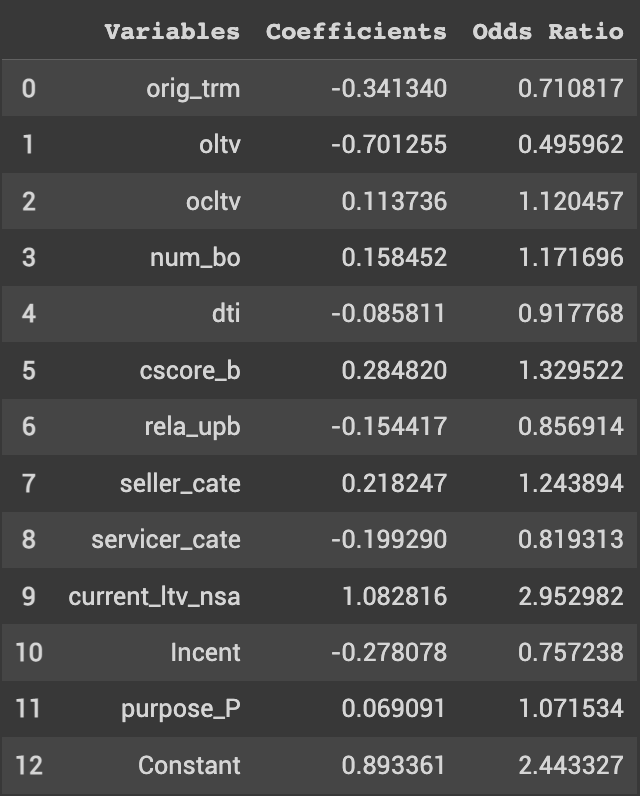
After running the logistic regression, the coefficients obtained for each independent variable are described above in the table. We can see that some variables have a negative sign and others have a positive sign helping understand their relationship to the prepay (constant) variable. Some variables have signs that we did not expect but it can be related to the multicollinearity.

1. **What is your model’s overall predicted probability of prepayment for the sample? How does that compare to the actual prepayment percentage?**

Our model’s overall predicted probability of prepayment for our sample is approximately 0.85. The actual prepayment probability in the data is approximately 0.69. The following results are expected, as our model may have contained more prepayment data rather than non-prepayment when splitting the data between training and testing.



1. **Provide the odds ratios for each variable (odds ratios are defined as exp(b) where b is the variable’s estimated coefficient. (see p. 158 of the textbook for an interpretation).**



Odds Ratio Table

The above odds ratio table shows which independent variables are more or less likely to predict prepayment based on our model. The independent variables with an odds ratio above 1, are more likely to predict prepayment and the variables with an odds ratio below 1 are less likely to predict prepayment.

1. **Are all of your variables consistent with your priors in terms of effect on prepayment?**

After obtaining the results of our analysis, we have derived the following conclusion as to which of our selected variables are consistent or inconsistent with our priors.

The inconsistency in our independent variables with our priors in terms of effect on prepayment may be due to multicollinearity, the relationship between the independent variables may be one of the reasons among others that make our variables inconsistent with our priors.

| **Consistent Variables** | **Inconsistent Variables** |
| --- | --- |
| cscore\_b | ocltv |
| purpose\_P | servicer\_cate |
| num\_bo | current\_ltv\_nsa |
| seller\_cat | Incent |
|  | dti |
|  |  |
|  |  |
|  |  |

1. **Assume that Market Rate declines by 1%. What impact on your model’s probability of prepayment would that have?**

When the Market Rate declines by 1%, our model’s probability of prepayment has a small decrease. We normally expected an increase in the predictability when Market Rates decrease as the Incentive variable would increase. But again, the relation between the INCENT variable and our constant is not positive in our model which may be due to the correlation with other variables and the unbalance between the split between training and testing.

