Statistical Models for Data Science: Project Assignment 4

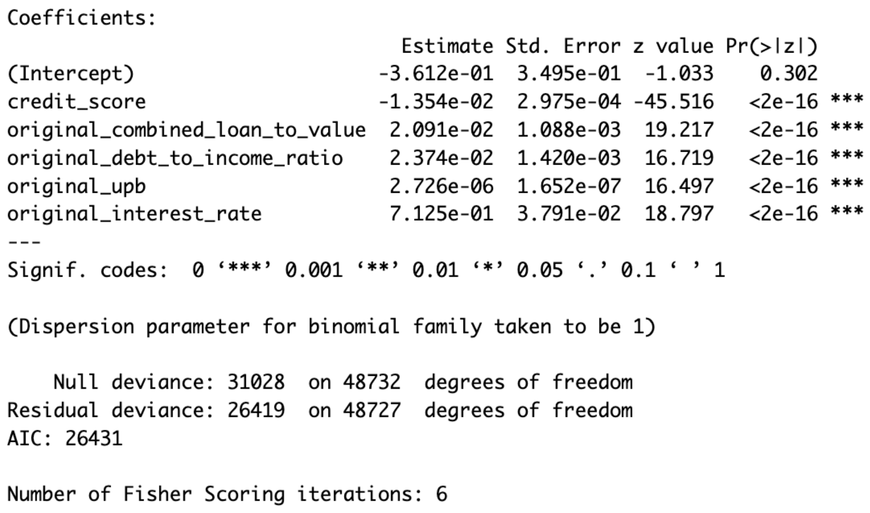
Ethan Chen, Joseph DiGiovanni, Kaushik Kannan, Zoe Toy

MS in Applied Data Science
University of Chicago
Professor Jonathan Williams
November 2024

1 Problem 2

We began our analysis by running a logistic regression to explain mortgage loan default as a function of credit score (FICO), combined loan-to-value (LTV) ratio, debt-to-income ratio (DTI), unpaid loan balance (UPB), and initial interest rate.

Below is the table of coefficients for this model:



```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.612e-01  3.495e-01  -1.033    0.302
credit_score                  -1.354e-02  2.975e-04 -45.516   <2e-16 ***
original_combined_loan_to_value  2.091e-02  1.088e-03  19.217   <2e-16 ***
original_debt_to_income_ratio    2.374e-02  1.420e-03  16.719   <2e-16 ***
original_upb                     2.726e-06  1.652e-07  16.497   <2e-16 ***
original_interest_rate           7.125e-01  3.791e-02  18.797   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31028  on 48732  degrees of freedom
Residual deviance: 26419  on 48727  degrees of freedom
AIC: 26431

Number of Fisher Scoring iterations: 6
```

Figure 1: Results Table for the Logistic Regression Default Model

Using the above summary table, we first looked into whether this model has more explanatory power than a model that simply assigns a flat default probability to every mortgage, which would be the null model in this case. We first inspected the null and residual deviances given by our model summary, seeing that the model that we built is much closer to the saturated model than the null model is, and the difference between the deviances is 4609 units. We then tested this difference between the deviances for statistical significance assuming a chi squared distribution with 5 degrees of freedom. We did this by subtracting the probability of a chi-squared distribution with a test statistic equal to the difference of the deviances and with degrees of freedom equal to 5 from the number 1, which denotes the saturated model, a perfect model fit, and the difference between these two quantities was equal to 0. This means that the changes in model fit experienced by adding the 5 new predictors to the logistic regression model is so large that it would be very unlikely to see it by random chance alone. This means that the logistic regression model does indeed have more significant explanatory power than the null model in this case.

We then assumed a 5% significance threshold to look into this model more, and we saw that all of the predictors used in this model factor heavily into changing the odds of a default on a home mortgage. Specifically, we see that a 1 unit increase in credit score decreases the odds of defaults by 1.35%, a 1 unit increase in the original combined loan to value increases the odds of defaults by 2.11%, a 1 unit increase in the original debt to income ratio increases the odds of defaults by 2.4%, a 1 unit increase in the original unpaid loan balance increases the odds of defaults by $2.73 * 10^{-4}$%, and a 1 unit increase in the original interest rate increases the odds of defaults by 104%.

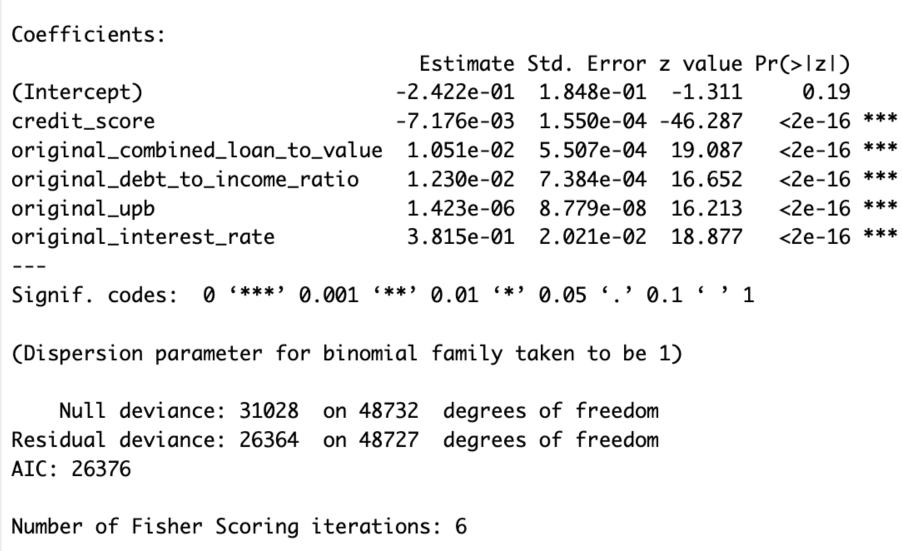
Using this information, we can see that the odds of someone defaulting on a loan are incredibly sensitive to 1 unit changes in the interest rate and that the odds of someone defaulting on a loan also go up in cases of increases in the combined loan to value, debt to income ratio, and unpaid loan balance, but that the percent change is not as large as the change experienced when interest rate is increased. Further, we can also see that as someone's credit score goes up, the odds of that person defaulting on a mortgage go down. This is all consistent with our knowledge of real world consumer finances and the logic behind home buying, which is that people will move out of a home when they have a higher debt to income ratio, a higher unpaid loan balance, a higher interest rate, or a higher combined loan to value ratio, as these indicate cases where the home buyer is less ”trustworthy” to finish out paying off the full home mortgage, whereas a higher credit score indicates a more trustworthy buyer who will tend to not default on the loan and be more responsible in paying the loan back.

2 Problem 3

2.1 Part A:

We then continued our analysis by running a probit regression with the same predictors to develop a basis of comparison for our logistic regression model from the prior portion.

Below is the table of coefficients for this model:



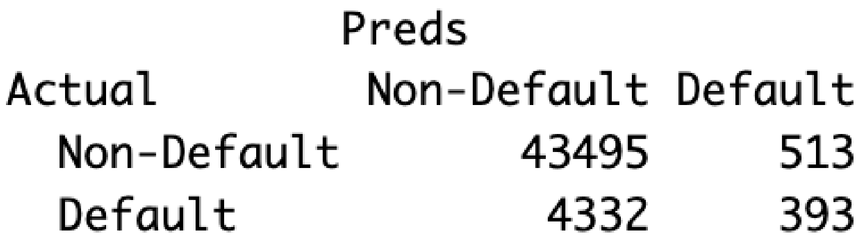
```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.422e-01  1.848e-01  -1.311     0.19
credit_score                  -7.176e-03  1.550e-04 -46.287   <2e-16 ***
original_combined_loan_to_value  1.051e-02  5.507e-04  19.087   <2e-16 ***
original_debt_to_income_ratio    1.230e-02  7.384e-04  16.652   <2e-16 ***
original_upb                     1.423e-06  8.779e-08  16.213   <2e-16 ***
original_interest_rate           3.815e-01  2.021e-02  18.877   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31028  on 48732  degrees of freedom
Residual deviance: 26364  on 48727  degrees of freedom
AIC: 26376

Number of Fisher Scoring iterations: 6
```

Figure 2: Results Table for the Probit Regression Default Model

In this case, we can see that the probit link function better explains the data than the logit link function, since the AIC of the probit regression model (26,376) is lower than the AIC of the logistic regression model (26,430). We know that a lower AIC is better, as the higher the AIC is, the worse is the balance between model fit and model size. Thus, using this, we can conclude that the probit regression model is going to better explain our data compared to the logistic regression model.

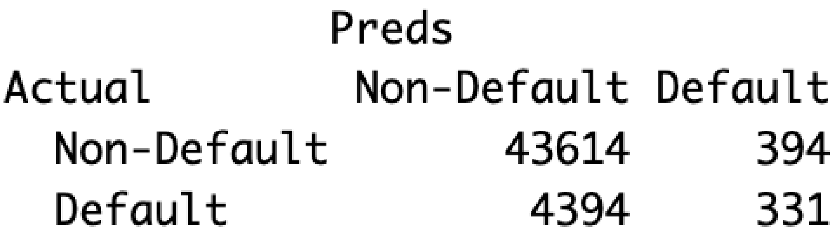
2.2 Part B:

We then created the confusion matrices for both the logistic and probit regression models using a classification boundary of 0.5, and these matrices are given below:



```
                 Preds
Actual           Non-Default Default
   Non-Default         43495     513
   Default              4332     393
```

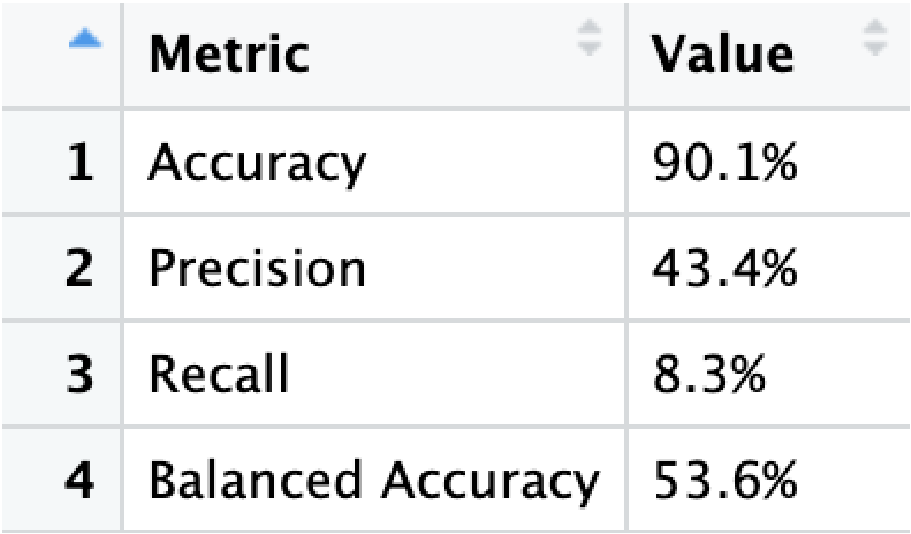
Figure 3: Confusion Matrix for the Logistic Regression Model



```
                 Preds
Actual           Non-Default Default
   Non-Default         43614     394
   Default              4394     331
```

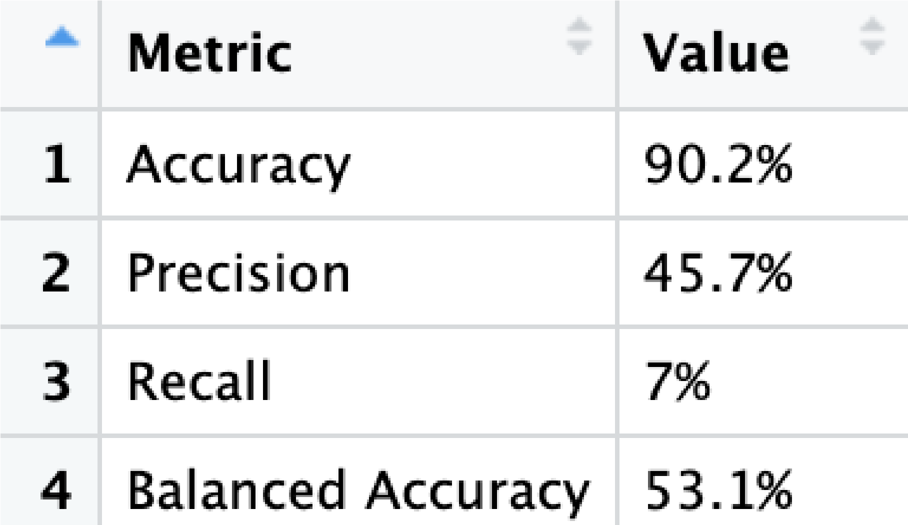
Figure 4: Confusion Matrix for the Probit Regression Model

We then tried to compare both models using a table of each model's respective metrics, including accuracy, precision, recall, and balanced accuracy in the calculation. Below are these tables:



	Metric	Value
1	Accuracy	90.1%
2	Precision	43.4%
3	Recall	8.3%
4	Balanced Accuracy	53.6%

Figure 5: Table of Model Metrics for the Logistic Regression Model



	Metric	Value
1	Accuracy	90.2%
2	Precision	45.7%
3	Recall	7%
4	Balanced Accuracy	53.1%

Figure 6: Table of Model Metrics for the Probit Regression Model

We can see from the above tables of metrics that the models share nearly identical metrics, with the logistic regression model having an edge over the probit regression model in terms of both recall and balanced accuracy, but the probit model having the edge over the logistic regression model in terms of both accuracy and precision. Considering the business context that was given to us in this case, which is that the mortgage backed securities industry went down following the extending of mortgages to borrowers who could not pay for them, we want to ensure that the model we build is as bulletproof as possible when it comes to correctly identifying the borrowers who will default to minimize losses to Freddie Mac and the investors of its mortgage-backed securities products. Thus, we will prefer a model with a higher percentage recall, since recall measures the proportion of actual defaults that were correctly identified by the model. So in this case, since the logistic regression model does have a higher recall, we would prefer it over the probit regression model.

3 Problem 4

After analyzing both models, we then wanted to dive deeper into a comparison of the actual and expected losses for Freddie Mac on its prime loans in 2007 using the logistic regression model. We started by creating a new dataset containing just the mortgage loans with a defect, and for these loans, we also pulled in their initial balance amount, their fitted probabilities of default as calculated from the logistic regression model, and their respective default score values, with 1 denoting a default and 0 denoting a non-default.

To get the expected loss, we first multiplied the initial loan balance by 0.25 and also the fitted probability of default, which accounted for the fact the loan is not certain to default, but rather that there is a chance of the loan defaulting. We did this for all loans that had a defect, regardless of whether they actually defaulted or not, and summed these up to get the aggregated total expected loss on 50,000 loans. We then multiplied this number by 20, since we wanted to scale this up to 1,000,000 loans instead of 50,000 loans, and since 50,000 * 20 = 1,000,000, so we felt that multiplying the total loss by 20 would account for this difference in scale. This gave us a total expected loss of $124,546,004 for Freddie Mac on all of its prime loans in 2007.

Next, to get the observed loss, we first multiplied the initial loan balances for the loans with defects that did default by 0.25, and did not need to account for fitted probabilities in this case, since we knew that these loans certainly did default (so probability of default = 1 for each of these loans). We then summed up all of these losses and multiplied by the same scalar factor 20 as we used when calculating the expected loss. This gave us a total observed loss of $417,670,000 for Freddie Mac on all of its prime loans in 2007. Thus, in conclusion, we see that loans with defects actually cost Freddie Mac's prime business $417,670,000 in 2007.

4 Problem 5

4.1 Part A

We then developed our own generalized linear model for predicting customer defaults that improved on the simple linear regression on both the model fit and classification metrics. The refined logistic regression model includes many of the same predictors as the original model, such as credit score, combined loan-to-value (CLTV), DTI, UPB, and interest rate, but also incorporates new categorical variables such as loan purpose and occupancy, as well as an interaction term between credit score and CLTV. These additional variables and the interaction term provide deeper insights into borrower risk by capturing more nuanced characteristics. For instance, loan purpose differentiates between cash-out refinancing/home purchases, while occupancy status identifies whether the property is occupied by the owner or an investment, both of which are critical indicators of loan default likelihood. The inclusion of the interaction term between credit score and CLTV is particularly important, as it highlights the compounded risk of borrowers with low credit scores taking on high-leverage loans. These borrowers are more likely to default because poor credit history combined with a lack of equity in their properties will present a bigger risk.

The reason loan purpose and occupancy status were added was because history has shown that borrowers seeking cash-out refinancing usually take on more financial risk, while investment properties are less likely to be prioritized for payments during financially difficult times in comparison to homes occupied by owners. By including these categorical predictors, the model accounts for more underlying risk factors.

The reason we included the interaction term was because of both financial reasoning and empirical evidence. Financial history has consistently highlighted the joint impact of credit score and loan structure on the risk of defaulting. Borrowers with low credit scores and high CLTV often lack financial stability and sufficient equity, making them vulnerable to default. By including this interaction term, we are able to quantify this risk and present better explanatory ability than we would with individual predictors.

The refined logistic regression model achieved an AIC of 26,315, which was an improvement over the original model's AIC of 26,430. This reduction in AIC reflects a better balance between model complexity and fit, indicating that the additional predictors and interaction term provide additional explanatory power over the original logistic regression model.

For our classification metrics, the refined model also shows improvement:

- **Accuracy**: Improved slightly to 90.3%, reflecting a better ability to capture true defaults and true non-defaults compared to the original logistic regression model.

- **Precision**: Improved to 45.4%, reflecting better identification of true defaults among all predictions of a default.

- **Recall**: Declined noticeably to 2.7%, reflecting a weaker ability to correctly predict defaults among the actual instances of defaults compared to the original logistic regression model.

-**Balanced Accuracy**: Decreased slightly to 51.2%, likely impacted more by the large decrease in recall than by the very slight increase in precision compared to the original logistic regression model.

These metrics demonstrate that our refined logistic regression model improves upon the previous logistic regression model in its ability to mitigate false predictions of both defaults and non-defaults, as reflected in its higher model accuracy and precision values. While the recall shows a noticeable amount of decline, this is balanced out by the improvement in the model's accuracy, which allows us to see a better balance in predicting not only defaults, but also non-defaults, which can help Freddie Mac better identify more trustworthy customers to target with its loan offerings. Additionally, the improvement in these classification metrics translates into a better ability to identify and mitigate high risk loans.

4.2 Part B

When evaluating Freddie Mac's profit or loss, we considered two scenarios: purchasing all loans without applying any risk boundary (a probability of default threshold above which Freddie Mac would not buy the loan) at all and also by using an optimal risk boundary found using the logistic regression model. This optimal risk boundary was found by running an iterative loop between the values of 0 and 1 and by finding the value within this interval that maximizes the profit earned by Freddie Mac. Running through this algorithm, we found that this optimal risk boundary was 0.04, meaning that all loans that have a probability of default lower than this will be considered "good" loans for Freddie Mac and they will profit on these loans and all loans with a probability of default higher than this value will be considered "bad" loans for Freddie Mac and they will lose money on these loans. Thus, it is advised from our end that Freddie Mac not buy loans that have a probability of default above 4% in the interest of maximizing its profits.

Buying All Loans

If Freddie Mac purchased all loans without applying any risk boundary:

- **Total Loan Balance**: $183.13 billion (scaled from 50,000 loans to 1,000,000 loans, with an average loan balance of $183,129.80).
- **Predicted Default Rate**: 10%.
- **Defaulting Loan Balance**: $18.31 billion.
- **Non-Defaulting Loan Balance**: $164.82 billion.

We found this using the profit formula:

$$\text{Profit} = (\text{Non-Defaulting Loan Balance}\times\text{Profit per Loan})-(\text{Defaulting Loan Balance}\times\text{Loss per Lo}$$

we calculate:

$$\text{Profit} = (164,816,820,000 \times 0.01) - (18,312,980,000 \times 0.25),$$

$$\text{Profit} = 1,648,168,200-4,578,245,000 = -2,930,076,800\,(\$-2.93\,\text{billion}).$$

Optimal Threshold (0.04)

By applying the risk boundary as a constraint, Freddie Mac would exclude loans with a predicted default probability of 4% or higher:

- **Loans Retained**: 70% of the total pool.
- **Retained Loan Balance**: $128.19 billion.
- **Predicted Default Rate After Threshold**: 5%.
- **Defaulting Loan Balance**: $6.41 billion.
- **Non-Defaulting Loan Balance**: $121.78 billion.

Once again using the profit formula:

$$\text{Profit} = (\text{Non-Defaulting Loan Balance}\times\text{Profit per Loan})-(\text{Defaulting Loan Balance}\times\text{Loss per Lo}$$

we calculate:

$$\text{Profit} = (121,781,317,000 \times 0.01) - (6,409,543,000 \times 0.25),$$

Profit $= 1,217,813,170 - 1,602,385,750 = -384,572,580$ (\$−384.57 million per 50,000 loans).

Scaling for 1,000,000 loans:

$$\text{Total Loss} = -384,572,580 \times 20 = -1,076,145,160\,(\$-1.08\,\text{billion}).$$

Using the 0.04 risk boundary, Freddie Mac would reduce loss from −\$2.93 billion to −\$1.08 billion, improving its profit by approximately \$1.85 billion. This risk boundary maximizes Freddie Mac's profit by rejecting loans with higher predicted default probabilities and ensuring Freddie Mac is more aware of taking on only safe loans for its business.