Anthony Partida, 3/17/25

Background

Predicting whether small business loans will default is challenging for banks due to the many variables that influence the outcome. Small business loans help stimulate the local economy, create jobs, and foster a sense of community and opportunity across the U.S. However, if loan officers approve loans that default, the bank or financial institution loses revenue and capital, ultimately harming its financial stability. Fortunately, advancements in analytics allow banks to make data-driven decisions by leveraging historical records and technology to improve loan approval accuracy. Developing an analytical model can help predict whether a business will default on its loan, enabling banks to make smarter investments. This approach increases revenue, reduces losses, and allows small businesses to thrive, ultimately strengthening local economies and job creation. The data that is used to make the logistic model used to predict if a loan will default is provided by the Small Business Administration (SBA), so we do not need to worry about data collection. Upon initial examination, I found a significant number of missing values (NAs) and some binary variables that require reformatting to be compatible with the analytical model. These preprocessing steps will be addressed in the next phase.

Data Analysis & Findings

After reviewing the descriptive analytics, I determined that we have a sufficient number of default observations to build a reliable model. Approximately 20% of loans in the dataset have defaulted (Figure 1.1). Additionally, when analyzing the UrbanRural variable, which classifies loans based on whether they were given to businesses in an urban area or a rural area, we observe that around 50% of the data originates from urban settings. This aligns with expectations, as cities typically have a higher demand for goods and services compared to sparsely populated areas. Some models that may be of interest to predict if a loan would default would be the following: **Model 1** uses term length and whether or not the business is in an urban setting as predictors for whether a loan will default. **Model 2** consists of the same variables that Model 1 uses, but also includes the gross amount disbursement and approved. **Model 3** uses the term length, gross disbursement amount if the loan was in the Low Doc program, and the gross approval amount (Figure 1.2). To create the variable that we would use to determine if the stock defaulted, I converted the ChrgOffDate variable into binary to determine if the loan defaulted instead of MIS_Status because some records were missing data and may have been inaccurate. I then selected variables that would be strong predictors of loan defaults. For example, a loan term could significantly impact whether a loan is repaid on time. The location of a business in an urban setting also seemed like a relevant predictor, as urban businesses have access to more potential customers but also face increased competition. Additionally, gross disbursement and approved loan amount could serve as key indicators, as larger loans may carry higher inherent risk due to larger interest payments and the greater revenue required for businesses to meet their repayment obligations.

Before constructing and testing the models, we needed to clean the data and make necessary modifications to prevent errors. This process involved multiple steps, but after implementing all the required changes, we reduced our dataset from 899,154 records to 628,502 records. I used a logistic model to predict the outcomes of loans. A logistic model uses variables to predict the outcome of an event (1 = Event happens & 0 = Event does not happen). The model would return a percentage of how likely the event will occur, and then we can determine at what threshold to approve the object or not. After constructing the models we get the following output (Figure 1.3:5). To put **Model 1's** outcome in context, a monthly increase in the term of a loan decreases the likelihood of that loan defaulting by 2.57% & if the term is in an urban setting that it is expected to increase the chance of it defaulting by 78.83%. **Model 2** can be interpreted as for every dollar increase in Gross Disbursement, the chance the loan will default decreases by .0001% (1% for every $10,000), and for every dollar increase in the gross amount approved we can expect an increase in the chance the loan will default by .0001%. Model 2's IsUrban and Term estimates are similar to Model 1's coefficients. **Model 3** introduced the Low Doc variables which decreased the likelihood of the loan defaulting by 131.50%, which meant that a loan that was approved for the Low Doc program was very unlikely to default. Model 3's other coefficients that it shares in common with the other two models also have similar coefficients; however, in Model 3, Gross approval is no longer as significant at predicting if a loan would default or not. After the models were constructed, I used a classification threshold of 35% to determine if a loan was at high risk of defaulting. For example, if a loan was predicted to have a 30% chance of defaulting it is considered low-risk but if the loan had a 40% chance of defaulting that loan was high-risk. Typically it is standard to use 50% as the threshold but given that we want to lessen a bank's risk when issuing loans we would want to err on the conservative side.

After evaluating each model's accuracy, sensitivity, and specificity I have determined that **Model 2** is the most accurate and effective model for determining if a loan will default and reducing the losses a bank will incur from defaulted loans. According to the confusion matrix in Figure 1.6 **Model 2** has the lowest proportion of False Negatives or observations the model predicted to be low-risk but were observed to default. It is crucial when assessing the model to limit this number because the main object of creating the model is to limit risk and the bank would want to reduce the amount of loans they approved to default. One downside with this model is that it does have a high proportion of False Positives or loans that the model predicted to default but in reality did not, this is bad for the bank because this a the loss of revenue.

Recommendations

The models provide valuable insights to help loan officers assess business loans. However, officers should still rely on their judgment in cases where a business owner lacks a strong business model or market conditions are unfavorable. Despite its accuracy in predicting loan outcomes, the model should complement, not replace, human decision-making. Based on our cleansed dataset, the total amount lost due to defaulted loans is $13,523,866,925. Comparing this to the total amount lost from false negatives (loans the model approved that later defaulted), Model 2 reduces losses to $7,758,904,266, saving $5,764,962,659 (42%) of the pre-model losses. A limitation of the model is its false positives—loans it predicts will default but do not. If loan officers reject these loans solely based on the model, they risk losing potential interest revenue. By combining model predictions with sound managerial decisions, banks can reduce risk and improve loan approval accuracy while still considering the business's qualitative factors.

Appendix

Data Cleansing and Wrangling Steps

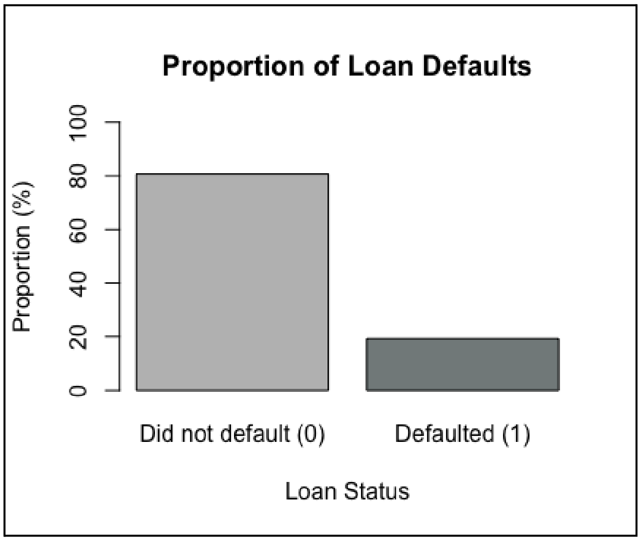


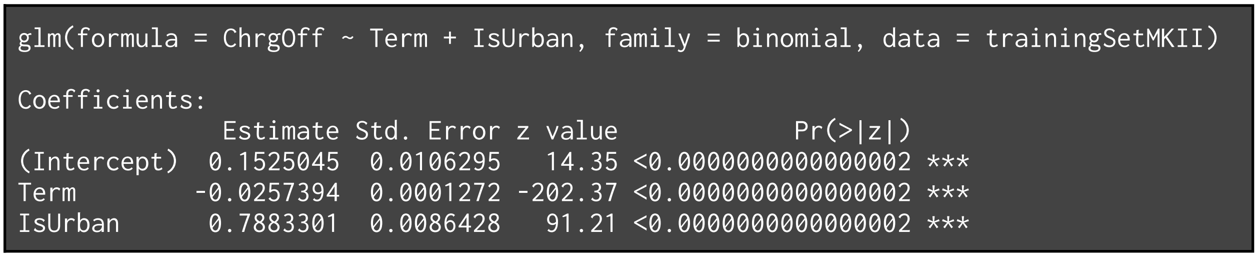
Figure 1.1

Model 1: ChrgOff = Term + IsUrban

Model 2: ChrgOff = Term + IsUrban + DisbursementGross + GrAppv

Model 3: ChrgOff = Term + DisbursementGross + LowDoc + GrAppv

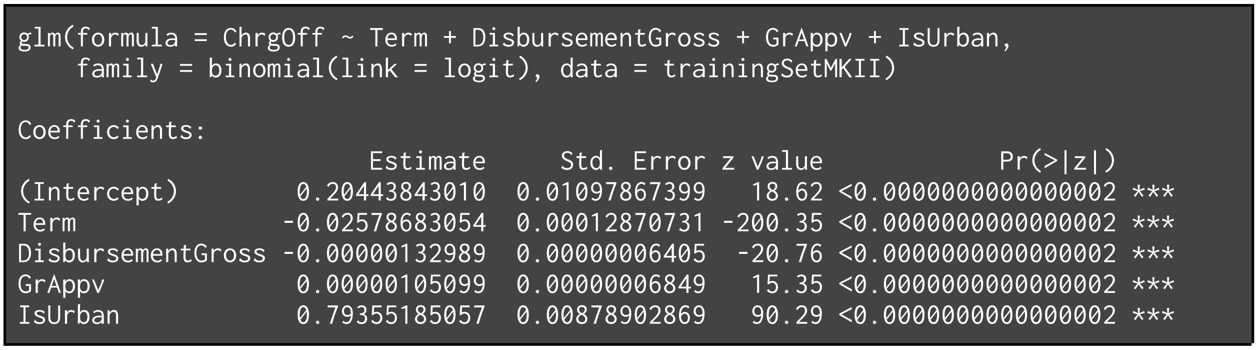
Figure 1.2



```
glm(formula = ChrgOff ~ Term + IsUrban, family = binomial, data = trainingSetMKII)

Coefficients:
              Estimate Std. Error z value            Pr(>|z|)    
(Intercept)  0.1525045  0.0106295   14.35 <0.0000000000000002 ***
Term        -0.0257394  0.0001272 -202.37 <0.0000000000000002 ***
IsUrban      0.7883301  0.0086428   91.21 <0.0000000000000002 ***
```

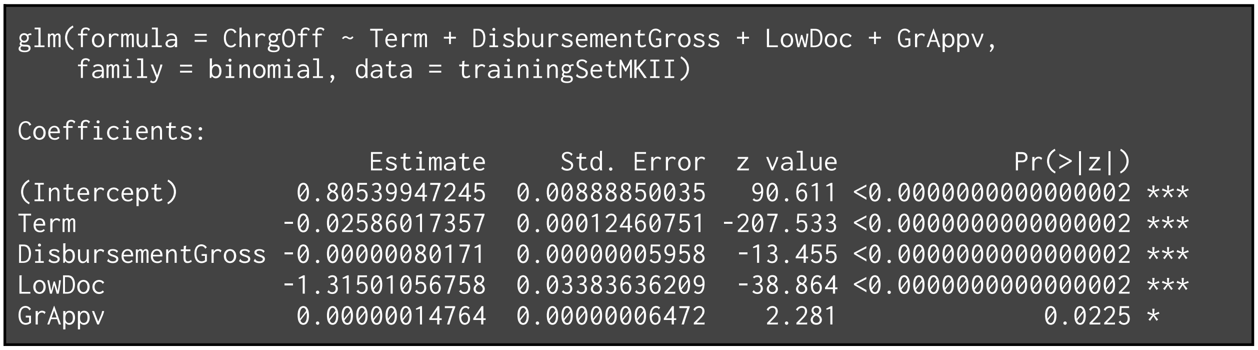
Figure 1.3



```
glm(formula = ChrgOff ~ Term + DisbursementGross + GrAppv + IsUrban, 
    family = binomial(link = logit), data = trainingSetMKII)

Coefficients:
                        Estimate     Std. Error z value            Pr(>|z|)    
(Intercept)        0.20443843010  0.01097867399   18.62 <0.0000000000000002 ***
Term              -0.02578683054  0.00012870731 -200.35 <0.0000000000000002 ***
DisbursementGross -0.00000132989  0.00000006405  -20.76 <0.0000000000000002 ***
GrAppv             0.00000105099  0.00000006849   15.35 <0.0000000000000002 ***
IsUrban            0.79355185057  0.00878902869   90.29 <0.0000000000000002 ***
```

Figure 1.4



```
glm(formula = ChrgOff ~ Term + DisbursementGross + LowDoc + GrAppv,
    family = binomial, data = trainingSetMKII)

Coefficients:
                        Estimate     Std. Error  z value           Pr(>|z|)    
(Intercept)        0.80539947245  0.00888850035   90.611 <0.0000000000000002 ***
Term              -0.02586017357  0.00012460751 -207.533 <0.0000000000000002 ***
DisbursementGross -0.00000080171  0.00000005958  -13.455 <0.0000000000000002 ***
LowDoc            -1.31501056758  0.03383636209  -38.864 <0.0000000000000002 ***
GrAppv             0.00000014764  0.00000006472    2.281             0.0225 *  
```

Figure 1.5

table(ValidSetMKII$ChrgOff, yHat1.0) **(.35)**		table(ValidSetMKII$ChrgOff, yHat2.0) **(.35)**	
TN: 81.99%	FP: 8.04%	TN: 82.44%	FP: 7.59%
FN: 3.52%	TP: 6.44%	**FN: 3.42%**	TP: 6.53%
table(ValidSetMKII$ChrgOff, yHat3.0) **(.35)**		**Comparing Models Props %**	
TN: 83.11%	FP: 6.92%		
FN: 3.87%	TP: 5.98%		

Figure 1.6

Data Wrangling Steps

Before	After	Why
Numerous NA results in the provided data set.	All quantitative NAs were omitted	These records may have been entered in incorrectly and to ensure the model is accurate the NAs were removed
Variables used Y & N	Changed the variables to binary 1 & 0 and omitted any other entry that wasnt Y & N before the conversion	The logistic model will not understand Y & N and needs the categorical variable to bee in binary in order to make an meaning full analysis
UrbanRural used 1, 2, & 0 to categorize variables	Changed to IsUrban and made 1 = Urban and 0 = Not Urban	The logistic model would also not be able to make an effective amnlysid by having the variable formatted this way