Loan Approval Classification Model

Using Binomial Logistic Regression

Summer Olmstead

CPSC 4180

Dr. Brendan Gressel

A loan approval prediction model aims to answer the key question: Should a loan with specific criteria be approved or denied based on the applicant's characteristics? While this question may be asked by banks mainly rather than the average person, there could be many benefits from deriving what key factors explain why or why not a loan is approved. A loan could make or break someone's dream to start that business or even buy their first home. It is beneficial to society to learn what patterns could help someone achieve success in building the life they want if finances are stopping them from taking that leap. By creating and fine tuning our model to try to explain why the loan status is approved or not approved based on the characteristic of the applicant, we could aim to answer this complex question.

The dataset used to create this model is found from Kaggle and is noted to be a synthetic dataset inspired by the Credit Risk dataset. The dataset was created to be synthetic to be enriched with additional variables based on financial risk loan approval data and utilized SMOTENC to stimulate new data points to make minority classes to make the dataset larger for training. There are 14 total columns in the original dataset including the applicant's age, gender, education level, income, years of employment experience, home ownership status, loan amount requested, loan intent, loan interest rate, loan percent income, length of credit history in years, credit score, and finally the loan status. The applicants age informs the age of the applicant when requesting the loan. The applicant gender is if the applicant is male or female. The education level is years of education someone has which has only the values of High School, Associate, Bachelor, Master, Doctorate. The income of the applicant is their annual income. The years of employment experience is the years the applicant has worked. The home ownership status is if the applicant rents, owns, has a mortgage, or other. The loan amount requested is the amount the applicant requested. The loan intent is any of the following: personal, education, medical, venture, home improvement, or debt consolidation. The loan interest rate is the interest rate on the loan in question. The length of credit history is in years of the applicant's credit history. The credit score is the credit score of the applicant. The loan status is if the loan got approved or not while 0 means no and 1 means yes. In the logistic regression model, we will have Y as our target dependent variable, loan status, and any of these other independent variables as potential x's when we explore different models to see which variables best explain loan status. The original dataset had 45,000 observations.

I cleaned the data by first checking for NaN values. No NaN values were present which is not surprising given the fact that all of the data for each person applying for a loan should have most of this information present generally in consideration and should not be blank. Next, I checked for the presence of duplicate values which there were none present as well. Then, I one hot encoded all of the categorical variables to be represented to equal to one if it was true for that category and zero if not true for that category. One hot encoding is extremely important to correctly understand the data rather than just mapping values to an index because the results will be unstable and flawed if doing so because it would think for instance the numbers one through four is a smaller scale than what each effect each variable has actually is. The variables I had to one hot encode was gender, the person's amount of education, person's home ownership status,

loan intent, and if there was a previous loan on file. For the loan on file, it was more changing the Yes or No to equal one or zero respectively. After one hot encoding, I had to remove one of the fields in the each one hot encoded category to not have multicollinearity in my model. Having every variable in one category creates multicollinearity because we cannot measure the actual effect of the categorical 'dummy' variables as they would be canceling each other out because the model would be measuring the effect of how they correlate to one another. This would be against the assumption of the model of no multicollinearity between variables which can lead to statistical error in the model itself. Therefore, I decided to check for outliers in the income variable as it seemed to make sense to only check for that one as that one could generally only skew the data so much if possible as many of the others were categorical. When deleting outliers in the income-based columns which deleted the entries it was in, this process deleted 286 entries in total leading us to a dataset of 44,714. Moreover, I checked for outliers in age for ages over 100 which was 4 observations interestingly enough and this left us with a total cleaned dataset of 44,710 observations.

Here is a visualization example of the dataset with two of the few variables that were not categorical, so I thought they would be interesting to visualize.

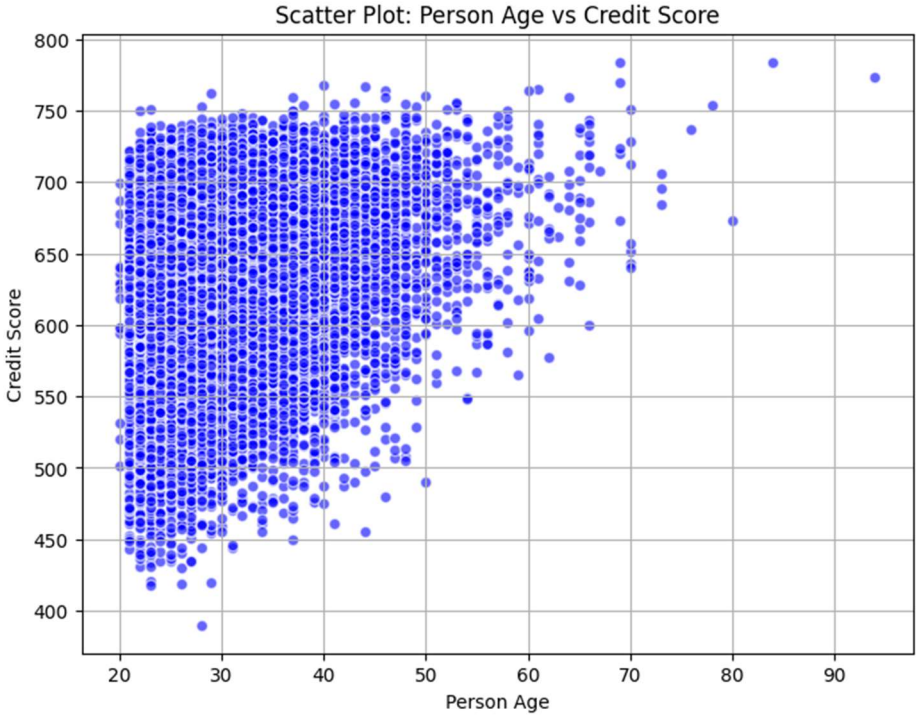

Figure 1. A scatterplot of a person's age and credit score.

Here we can see a trend that generally as someone gets older they tend to not have a lower credit score as they likely have more experience with their financials over time and gets better with budgeting among other things. Later, we will experiment to see a potential implication of this.

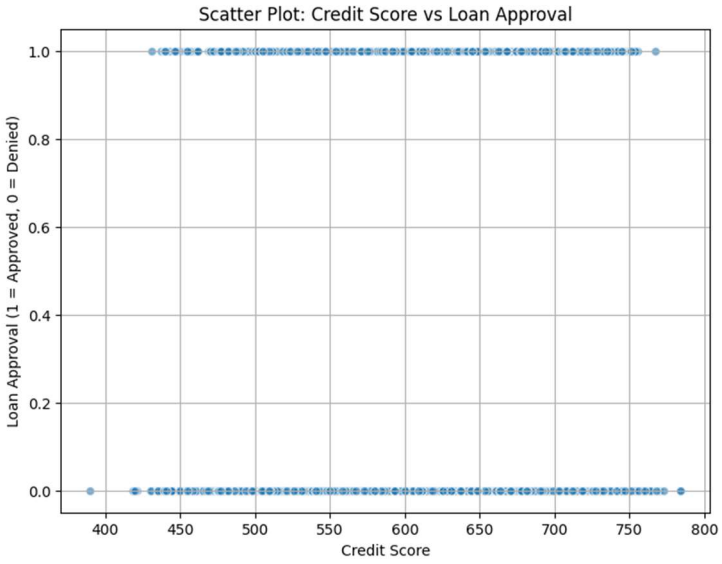

Figure 2. A scatter plot between credit score and loan approval.

I decided to also visualize one of my variables to see its individual relationship with the dependent variable, Loan Approval here in Figure 2. I wanted to include this to showcase the visualizations with many factors, predominantly categorical variables with a categorical output of loan status can be challenging to visualize general trends immediately.

When creating the model, I knew it would be a process of trial and error of substituting variables in and measuring their corresponding effects onto the model. For the general build of the model, I did an 80% training and 20% test split and then scaled the features. I scaled the features because I have variables of different scales and units so I did a standard normalization process. I predicted the model on the test data after training and calculated the accuracy, confusion matrix, and classification report. Moreover, I plotted a bar graph of the coefficients for each variable every time to visually see what most effects loan approval in the model. Furthermore, I would perform the logit regression model summary to see important statistics such as the pseudo R^2, z score, and the p value. The pseudo R^2 suggests if the model is a good fit for the data whereas the z score shows if the coefficient is statistically significant for the model utilizing Wald's test. If the absolute value of a z score is under 1.96 then the variable is insignificant in a 95% confidence interval. I am doing this process of understanding the variable and their effects on the explanation of loan status before deriving my final model which has the best explanatory power.

I went through a process of choosing different variables and finding their effects while cross referencing each coefficients z score to see if they were statistically accurate. Overall, out of five different models I created in this process of trial and error the best model with explanatory power was Model Four with a 86% accuracy score. Therefore, I will go over why Model Four was the best model in comparison and dissecting why the combination of factors chosen were perfect to conclude it was the best model.

In Model 1, I had the logistic equation of the following which was 83% accurate on the test data:

$$\begin{aligned} Loan\_Status = {} & 5.4435 + 0.0203 \cdot person\_gender\_male - 0.0085 \cdot credit\_score - 0.0066 \\ & \cdot person\_age - 1.3312 \cdot person\_home\_ownership\_MORTGAGE - 0.1266 \\ & \cdot person\_home\_ownership\_OTHER - 1.6886 \\ & \cdot person\_home\_ownership\_OWN + 0.1378 \cdot loan\_intent\_EDUCATION \\ & + 0.3228 \cdot loan\_intent\_PERSONAL + 0.7675 \cdot loan\_intent\_MEDICAL \\ & + 0.9119 \cdot loan\_intent\_DEBTCONSOLIDATION + 0.8196 \\ & \cdot loan\_intent\_HOMEIMPROVEMENT - 48.7168 \\ & \cdot previous\_loan\_defaults\_on\_file\_Yes \end{aligned}$$

The biggest take away from this model was that gender, previous loan on file, and home ownership other was both under 1.96 z score indicating their coefficients were not statistically significant under a 95% confidence interval.

In Model 2, here is the following regression equation which was 82%:

$$\begin{aligned} Loan\ Status = {} & -4.3968 - 0.0130 \cdot person\_age - 0.0004 \cdot credit\_score + 0.0000 \\ & \cdot loan\_amnt + 0.2783 \cdot loan\_int\_rate - 1.1797 \\ & \cdot person\_home\_ownership\_MORTGAGE + 0.1961 \\ & \cdot person\_home\_ownership\_RENT - 1.5714 \\ & \cdot person\_home\_ownership\_OWN + 0.1740 \cdot loan\_intent\_EDUCATION \\ & + 0.4114 \cdot loan\_intent\_PERSONAL + 0.8203 \cdot loan\_intent\_MEDICAL \\ & + 0.9905 \cdot loan\_intent\_DEBTCONSOLIDATION + 0.8324 \\ & \cdot loan\_intent\_HOMEIMPROVEMENT \end{aligned}$$

Notably we see some change of the effects of the coefficients from the previous model and we see that we should take out person home ownership rent as well as it also had a z score under 1.96 in absolute value. We also see that credit score became insignificant in this model for some reason which is strange so we can assume the adding and taking away of variables was capturing some of the significance of credit score's affect on loan status.

The following is Model 3's logistic regression with a 81% accuracy:

$$\begin{aligned} Loan\ Status = {} & -4.2023 - 0.0114 \cdot person\_age + 0.2893 \cdot loan\_int\_rate - 1.2399 \\ & \cdot person\_home\_ownership\_MORTGAGE - 1.7276 \\ & \cdot person\_home\_ownership\_OWN + 0.1638 \cdot loan\_intent\_EDUCATION \\ & + 0.4036 \cdot loan\_intent\_PERSONAL + 0.7860 \cdot loan\_intent\_MEDICAL \\ & + 0.9704 \cdot loan\_intent\_DEBTCONSOLIDATION + 0.8501 \\ & \cdot loan\_intent\_HOMEIMPROVEMENT \end{aligned}$$

Here we see all of the variables are statistically significant with a |z score| > 1.96. However, we see the worst accuracy. This is because of getting rid of a lot of information gain that while some independent variables were not significant they were accounting for some unknown effect not necessarily measured in the variable itself only. In simpler terms, the variables we got rid of

lessened the model's information gain and some of the variables gutted were also measuring some other effect in their variable than just the variable itself.

This conclusion leads us to Model 4 with an accuracy of 86%:

$$\begin{aligned} Loan\ Status = {} & 2.6788 - 0.0096 \cdot person\_age + 0.2739 \cdot loan\_int\_rate - 1.3015 \\ & \cdot person\_home\_ownership\_MORTGAGE - 1.7094 \\ & \cdot person\_home\_ownership\_OWN + 0.1956 \cdot loan\_intent\_EDUCATION \\ & + 0.3467 \cdot loan\_intent\_PERSONAL + 0.7791 \cdot loan\_intent\_MEDICAL \\ & + 0.9358 \cdot loan\_intent\_DEBTCONSOLIDATION + 0.8411 \\ & \cdot loan\_intent\_HOMEIMPROVEMENT - 0.0088 \cdot credit\_score + 0.0004 \\ & \cdot cb\_person\_cred\_hist\_length - 10.2895 \\ & \cdot previous\_loan\_defaults\_on\_file\_Yes \end{aligned}$$

Here all of the z scores are significant besides the person credit history length which also has pretty much a coefficient of 0 which means it probably did not necessarily need to be in the model. However, its presence increased the accuracy, and I originally created an interaction variable because of Figure 1. My interaction variable was a column I made of person's credit history length*credit score. The coefficient when added was extremely close to 0 was it was gutted and shown that the interaction variable which we hypothesized in figure 1. of age vs credit score was not as accurate as we believed unfortunately. So, I took it out and did not use it in the model due to its uneffectiveness. However, it is important to note that this model was the most successful in all of the runs.

Lastly, we have Model 5 with the following logistic regression equation with an accuracy of 85%:

$$\begin{aligned} Loan\ Status = {} & 2.4512 - 5.051 \times 10 - 5 \cdot person\_age + 0.2740 \cdot loan\_int\_rate - 1.3020 \\ & \cdot person\_home\_ownership\_MORTGAGE - 1.7075 \\ & \cdot person\_home\_ownership\_OWN + 0.1967 \cdot loan\_intent\_EDUCATION \\ & + 0.3470 \cdot loan\_intent\_PERSONAL + 0.7797 \cdot loan\_intent\_MEDICAL \\ & + 0.9358 \cdot loan\_intent\_DEBTCONSOLIDATION + 0.8404 \\ & \cdot loan\_intent\_HOMEIMPROVEMENT - 0.0088 \cdot credit\_score + 0.0006 \\ & \cdot cb\_person\_cred\_hist\_length - 8.6958 \\ & \cdot previous\_loan\_defaults\_on\_file\_Yes - 0.0101 \cdot person\_emp\_exp \end{aligned}$$

Here we added the persons employee experience in years as we never added it before, and you can see we ran into a multicollinearity issue as the person's age is measuring the same effect generally on credit score and it made the age coefficient extremely insignificant pretty much equal to zero. Therefore, we can conclude out of all five runs Model Four has given the best amount of success.

To begin visualizing and understanding our model 4, we can reflect on the following to see the relationship between the coefficients to the dependent variables, loan status.

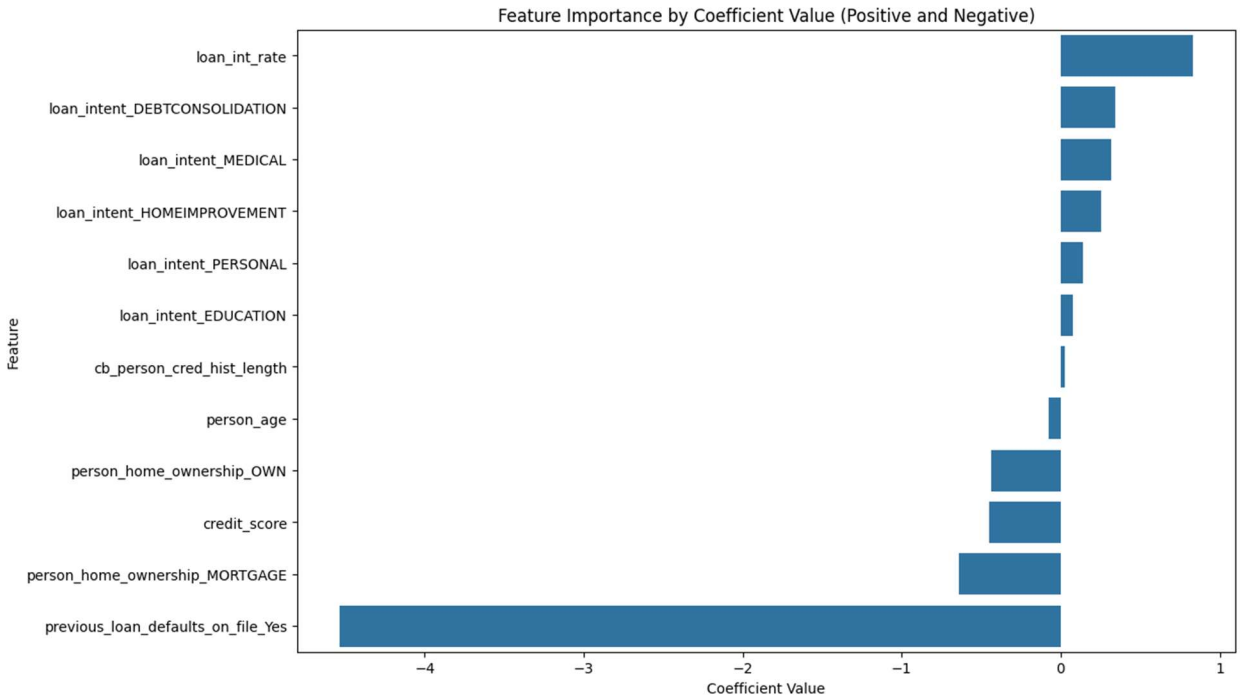

Figure 3. Feature Importance by Visualizing each feature onto a bar graph of Model Four.

Here we can see the individual effect of each variable in terms of positive and negative. We see that having a previous loan on file negatively effects the likelihood of getting approved for a loan; Furthermore, we see that as the proposed loan's interest rate increases it has a positive relationship with having the loan approved. This makes sense as it is likely more beneficial for the bank to get more out of their investment if it's a higher loan interest rate.

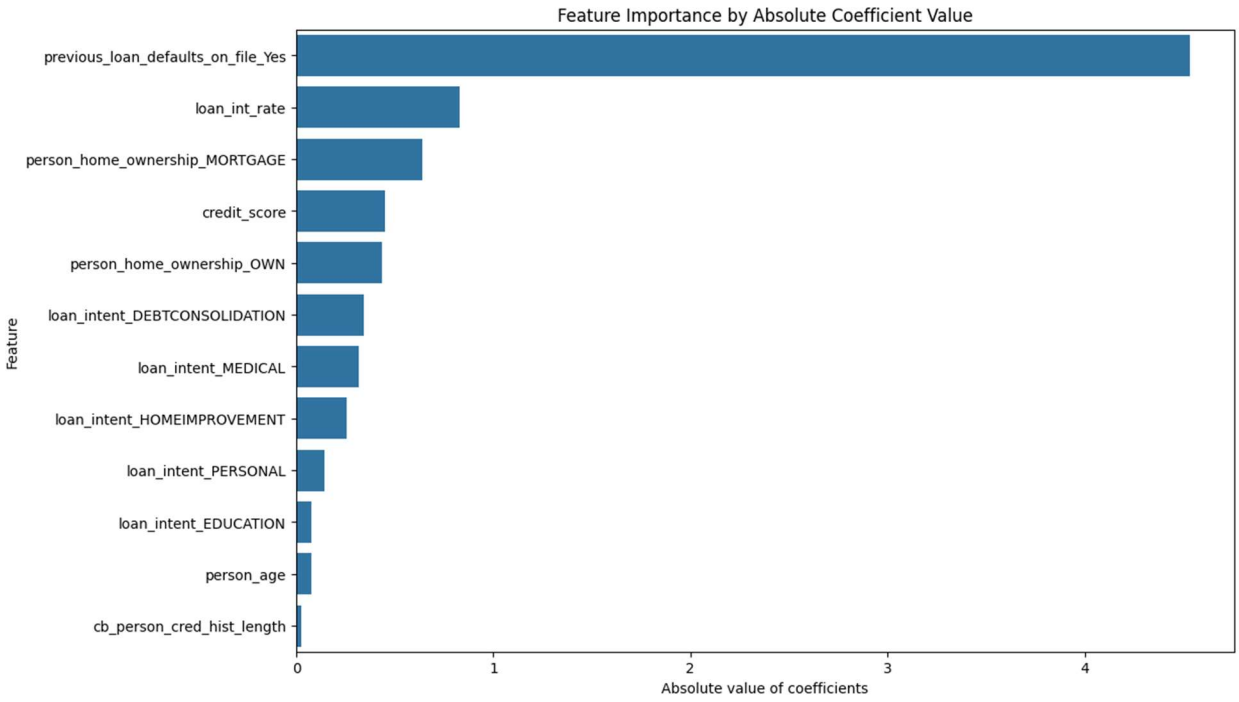

Figure 4. Absolute Value of Each Coefficient for Each Variable of Model 4.

Here we can reflect that having a previous loan on file, the loan interest rate, and if the person's home ownership status is having a mortgage that these are the most effective factors into explaining loan status. Whereas the least impactful factors are the persons credit history length and age.

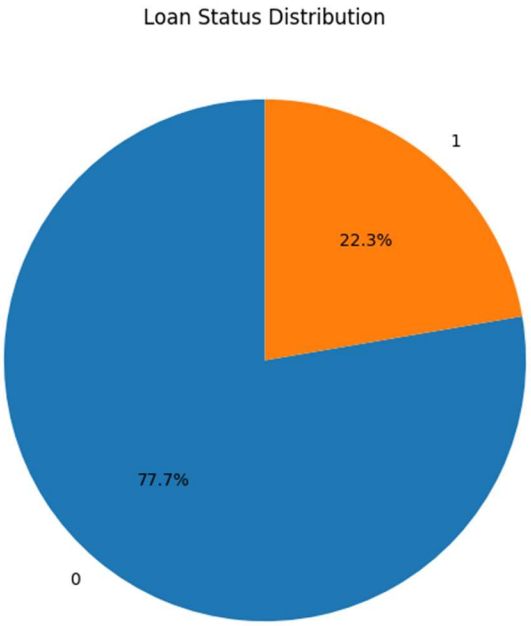

Figure 5. A Pie chart of the model's No and Yes predictions from the testing data.

From Figure 5, we can gather that the model predicted on the test data set No 77.7% and Yes 22.3% of the time from its predictions. Now, let's look to see how many times it was actually accurate though.

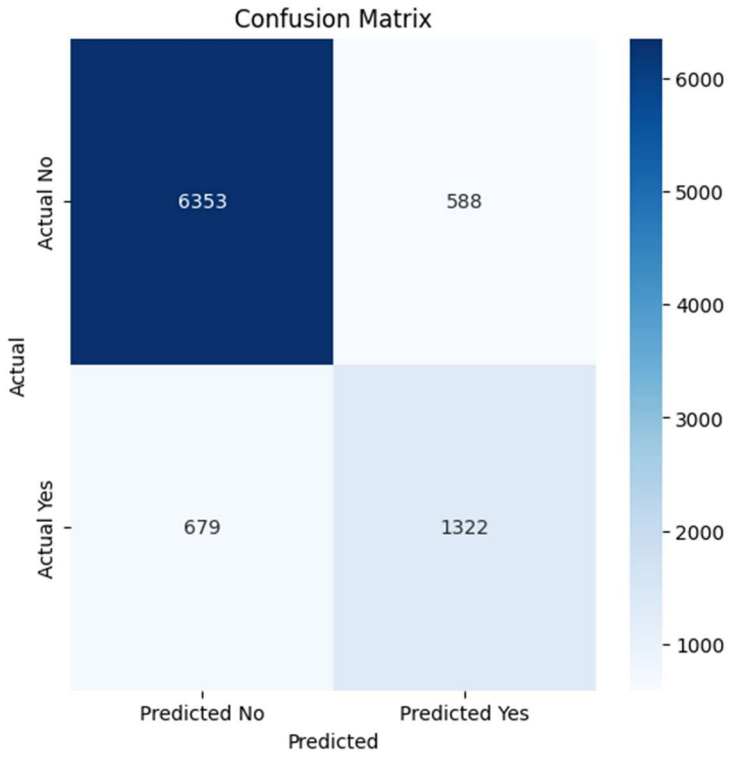

Figure 6. A heatmap of the confusion matrix of Model 4's test results.

Here in Figure 6, we can see our true and false negatives and positives. Here is some of the following statistics we can derive:

True Negative: 6353 / 8942 = 71.05%

False Negative: 679 / 8942 = 7.59%

True Positive: 1322 / 8942 = 14.78%

False Positive: 588 / 8942 = 6.58%

So we do see that our trues added together gives us a percentage of 85.83% and our false added total is 14.17% . This gives us more insight into our model and also aligns as our accuracy was 85.83 for this model which rounds to 86%.

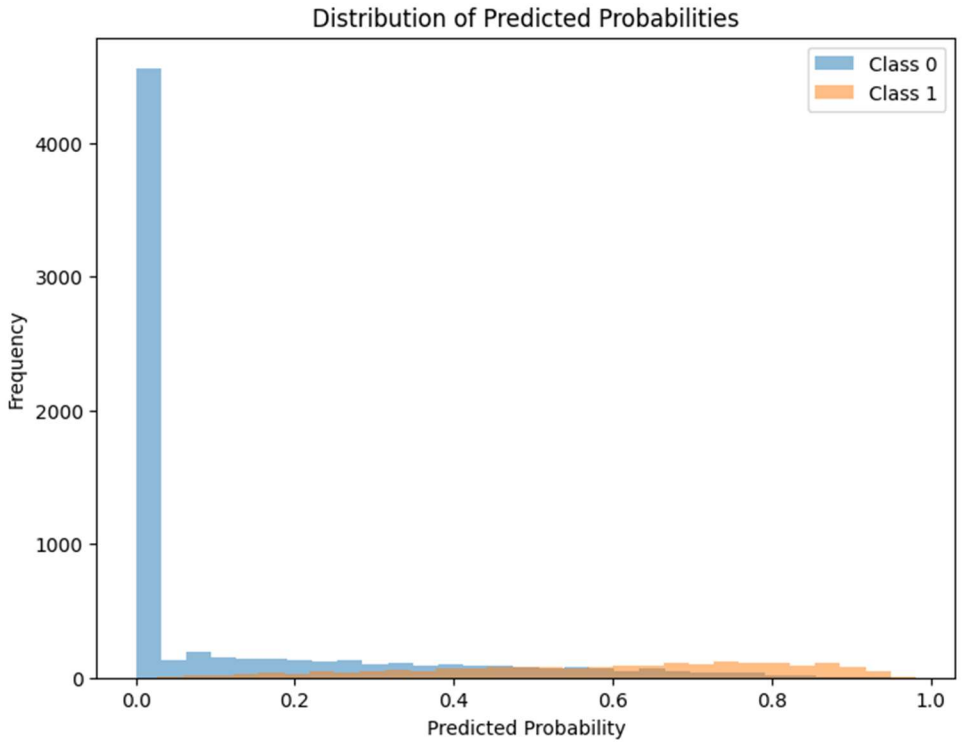

Figure 7. A bar chart of the distributions for each class: Class 0, No and Class 1, Yes.

Here we can see that the model is definitely for stronger prediction for the Class 0, No predictions rather than confidentially predicting a correct Yes. We also see a very small probability distribution for the Yes class which is slightly due to banks giving out less yes loans anyway.

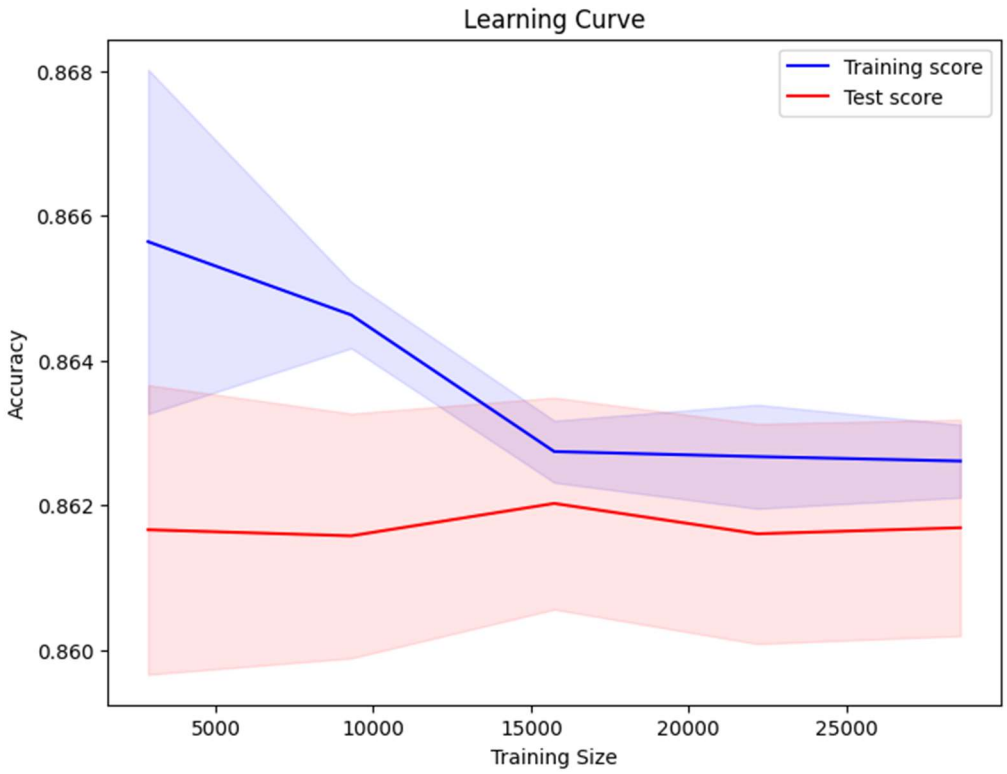

Figure 8. A learning curve plot showcases training and test score converging over time for Model 4.

From this learning curve plot, we see our Model 4 as it approaches the 25 thousands begins to converge the training and testing to an accuracy range of .861 - .864. This makes sense as our accuracy was close to this range generally but was about .3 percentage off from this range.

In conclusion, we find that our Model 4 was the best and most successful model comprised of the independent variables previous loan on file, age, credit score, loan intent, home ownership status, loan interest rate, and credit history length to explain if a loan was approved or not. We came to the conclusion model 4 with the highest accuracy and mix of statistically significant variables was the best model out of the 5 different attempts. I believe some limitations to this dataset was missing if race affected if a loan gets approved or not as discrimination is very prominent still to this day. I think if there were different race dummy variables it would explain more of the data better and hopefully increase the pseudo R^2 with this information if available. Moreover, this model could have been better for real world situations if it was not altered to generally be made for machine learning model training on Kaggle. However, I do believe this model is still a success regardless as it does help us see what does explain loan status generally as the original data was collected from real life data. Overall, this research is important to society to help explain the factors in ones life to help their cause of potentially getting their loan approved in aims to help the increase quality of peoples lives. Majority of the world was not born with a silver spoon and money from some sort of inheritance, so for the playing field to be equal it is important to understand why loans do and do not get approved.

References

https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data?resource=download