Predicting Credit Card Approval

In today's day and age, banks receive a high demand for credit cards from people with various needs. While they try to get approval for the issuance of their credit cards, some of their applications may get rejected for a variety of reasons and investigating the cause of the rejection is crucial in helping applicants to predict their chance of success and ensure they can be approved with knowing what helps their chances. This credit card approvals dataset contains critical factors such as gender, age, debt, marital status, bank account ownership, job information, ethnicity, prior default, credit score, zip code, income, years of employment, and employment status. **The goal of analyzing this dataset is to predict whether future applicants will get approved for a credit card, given these key factors.** In general, it would be interesting to find the correlation of these factors (individually or in combination) with the outcome of the customers getting a new credit card and filtering out the ones that have the most significant impact. This prediction would give us a good understanding of the approval process to maximize the chance of getting credit cards in the future.

To achieve our goal, we chose four conjectures to work on, focusing on different factors to see what affects the approval of credit cards the most. **Our first conjecture is that employment status and more years of employment have a significant positive impact on the probability of getting a new credit card.** The reasoning behind this conjecture is that being employed may be an indicator of reliability since having work generally implies having a steady stream of income to pay the credit card bill. Similarly, the more years someone is employed, the more trustworthy said employee is since there is a higher chance that they will continue to have employment, putting them at a lower risk of unemployment. Through our analysis, we have found that this conjecture is true. **Our second conjecture is that the amount of debt someone has would negatively affect their chance of getting approved for a credit card.** We hypothesized this because we assume that the amount of debt a person has is one of the most significant factors in whether the person is able to pay their credit card bill. We know that credit score is key in credit card approval since a low credit score indicates untrustworthiness, and having unpaid debt is expected to contribute negatively to the credit score. For this conjecture, we found that we didn’t have enough evidence to prove this was true. **Our next conjecture is connected to our previous one; the approval of a credit card is positively affected and can be fairly accurately predicted by credit score.** As mentioned before, we knew that credit score is one of the biggest factors that affects credit card approval, and we knew that the higher the credit score is, the more trustworthy, responsible, and reliable the person would be expected to be and therefore, more likely to get approved for a credit card. We found this conjecture to be true. **Our last conjecture is that the previous citizenship of a customer does not have any impact on the probability of getting approved for a credit card.** We would like to explore whether there would be a difference in the approval of credit cards regarding whether the customer is a citizen by birth, a naturalized citizen, or a temporary citizen. We also found this conjecture to be true. Using the reasoning above and our analysis, we redefined some of our conjectures to explore the dataset further.

The **first conjecture (that employment status/length has a significant positive impact on the probability of getting a new credit card)** attempts to find the effect that some factors have over another factor. Thus, the best method is to use a heatmap, as shown in *figure 1*. It provides a correlation matrix to establish the relationship between variables so that we can focus on variables with a higher correlation with the approval of a credit card. As explained earlier, our first conjecture was proven to be accurate; thus, we can redefine it and create a new conjecture, which can be "If I know what a person has been employed, I predict they will most likely be approved for another credit card". To help us see if our new conjecture is also true, we can make a prediction model. When we look at the output graph of our new prediction model (*figure 2)* where 0 means not approved and 1 means approved, we see that our new conjecture is also proven true since employed people have a higher rate of getting approved for a credit card.

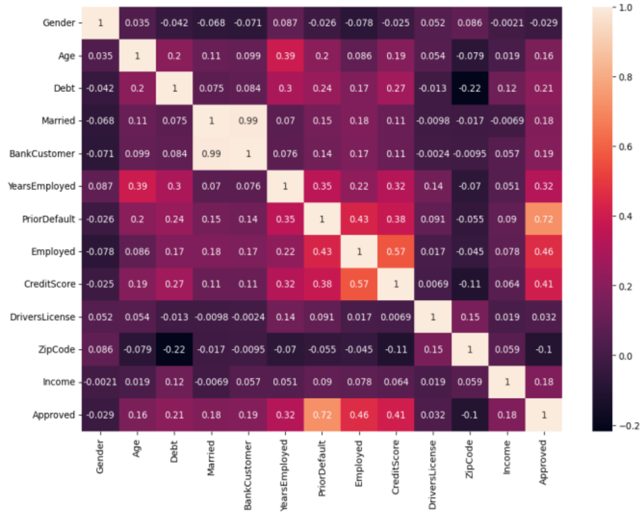

Figure 1

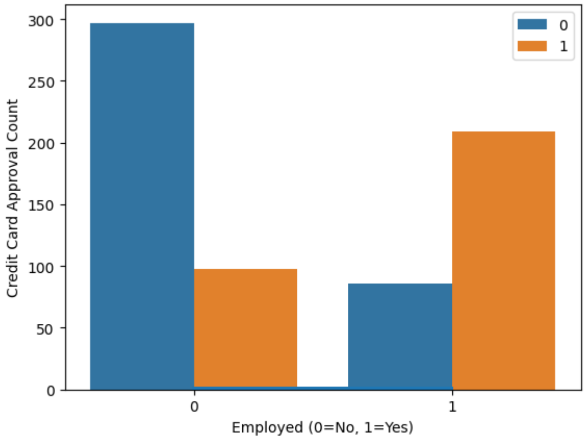

Figure 2

The **second conjecture (the amount of debt someone has would negatively affect their chance of getting approved for a credit card)** is a binary classification problem. Since the output would be 0 or 1, a logistic regression algorithm was used to test the accuracy of the model, or the method, as shown in *figure 3*, where the green points represent the training data and the blue points represent the testing data. Although the accuracy of the model (59.4%) is not high enough for us to make a definite conclusion on whether the amount of debt affects the credit card approval rate, from the previous part we can still see that the potential correlation is too weak to conclude that the amount of debt a person owes affect the chance of getting credit card approval. Since we could not prove this conjecture to be true, we could not redefine it.

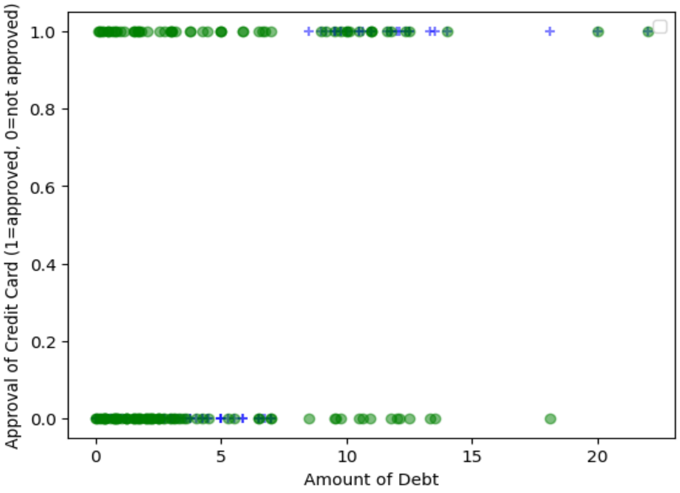

Figure 3

In our **third conjecture (approval of a credit card positively affected and can be fairly accurately predicted by credit score)**, we also used KNN to analyze the relationship between the two variables (credit score and credit card approval) since KNN is effective in binary categorical prediction. We visualized the result produced by KNN using the graph shown in *figure 4,* where the green points represent the training data and the red points represent the testing data. The model used in the third conjecture produces a higher accuracy of 78%. The graph shows that people with higher credit scores tend to get approval for a credit card. Because our third

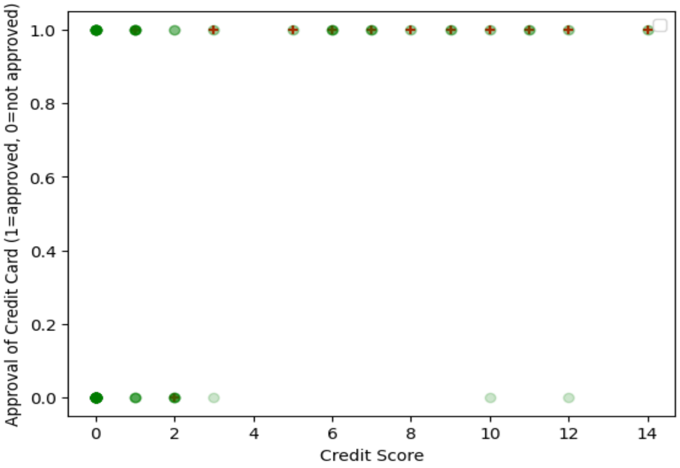

Figure 4

conjecture was found to be true, we can now redefine it. We know that a person's credit score can determine whether or not they are approved for a credit card in the future; thus, our new conjecture can become, "If I know the customer's credit score is on the higher side, they will most likely be approved for another credit card". To help us predict whether this is true, we can use the KNN model that we've previously used. Based on the model, our new conjecture also proves to be true.

The **fourth conjecture (previous citizenship of a customer does not have any impact on the probability of getting approved for a credit card)** looks at a feature not included in the heatmap on whether previous citizenship status would affect the chance of getting approval for a new credit card. A correlation matrix is used in order to find the correlation between them, and the previous citizenship status has a low correlation with approval of credit cards (0.089); *Figure 5* shows that previous citizenship status does not have a notable impact since both approved and not approved occurs at about the same frequency in each group. Since we could not prove this conjecture to be true, we could not redefine it. Based on all of this knowledge, we can now share our results.

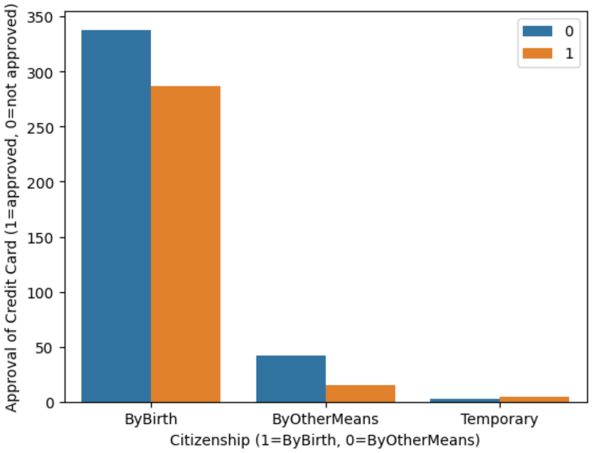

Figure 5

Based on the conjectures, the features selected would be "YearsEmployed," "PriorDefault," "Employed," and "CreditScore." The reason we chose a simpler algorithm was due to the relatively small dataset that we worked on. Simpler models avoid adding too much complexity to this prediction and would be ideal for a more straightforward prediction. Regarding the algorithm, logistic regression was chosen since the prediction is a binary classification problem that only outputs 1 or 0, or approval or not for a new credit card. A logistic model is very interpretable for these types of problems. The accuracy of the model was around 85%, which is fairly high, thus proving an optimistic prediction. To further assess the model, cross-validation with k=5 is calculated. Since all five values (0.78571429, 0.85714286, 0.75, 0.88888889, 0.92592593) are on the higher side, the model is proven to be a fair fit for the current dataset. Regarding the most impactful feature of the model, we calculated the logistic regression coefficient. As shown in *figure 6*, the prior default has the highest coefficient, meaning that it has the highest β value and thus, the highest odd ratio. Next, we can share some of the challenges we've encountered while finding the results of this project.

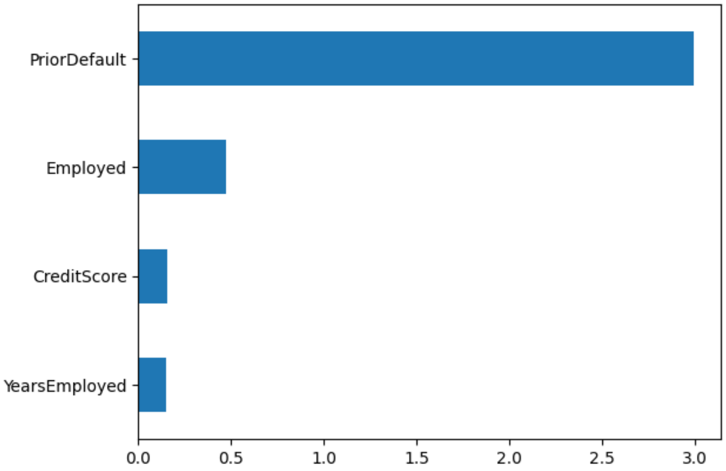

Figure 6

One of the challenges we faced was determining the model to use after selecting the feature. There were several classification models possible, some of the ones we considered included KNN, logistic regression, random forest, and SVM. As mentioned above, we selected the simple logistic regression algorithm to avoid over-complication. Another challenge we faced was after we redefined some of our conjectures, we needed to create new prediction models. We decided to use the KNN algorithm for both predictions, as that was the best option to depict how the data is grouped.

We corporate and divide work during meetings. Zilin worked on coding, idea organization, some research, idea proposal, visualization, model evaluation, conclusions, reports and slides, feature/label selection, and statistical summary. Pegah came up with ideas and brainstormed, researching, helping with conjecture ideas, some parts of the report, such as the motivation part, and some parts of slides, including adding the visuals with a good format to the presentation. Chloe helped come up with some of the conjectures/ redefining conjectures. She also contributed to some parts of the slides/report. Alisha worked on rewording and expanding on the conjectures, as well as helped with coming up with reasoning for those conjectures. She also played a part in refining the slides and report, proofreading the content, and editing where needed. Ashleigh focused on brainstorming and contributing to the slides, layout, organization, and writing the report.