Professor: Jairam Avinash

Student's name: Hagay Ringel

Project #2 - Credit Card Approvals- methodology and findings

This project engaged with the prediction of credit card approvals. The goal is to predict who qualified for a credit card. The first step of this project was cleaning the dataset, in this case the dataset was full and there were no Nans or nulls. The next step is the visualization part, at this part, I showed some demographics data for example the proportion between those who got approved for credit card and those who got denied. We can see that there is a bit more then 300 people who got denied while almost 400 people who got approved (figure 1). Then, I added the column of marital status and according to the graph there is no significant difference between the married people, but it is seen that non married people are more likely to get denied in a ratio of 1/3 (figure 2). The next demographic data I plotted into a graph was the income per race and it's clear that white and Asian have a higher income compared to black and Latino (figure 3). The next and last graph came from the previous graph and show the distribution of debt between the races (I took the first 30 rows) and I could not find any clear trend, so I assume the debt rate is equally distributed (figure 4). Then, I defined the target (y= Approved) and the features variable (X= Gender, Married, Debt, Employed, Income).

Accuracy- the accuracy in this project was calculated by using all the six algorithms we learned: logistic regression, naive bayes, decision trees, random forests, support vector machines (4 types) and K-Nearest Neighbors, and for each algorithm I performed a 5-fold cross validation. The range of the accuracies of all the algorithms is between 0.66 (SIG_SVM) and 0.75 (RBF_SVM) while most of the accuracies are closer to 0.75 which leads to the conclusion that the accuracy of the model good but not excellent, which mean that the question of the credit card approval is predictable by this model.

Precision- the precision in this model is similar in each algorithms used. We can see that most of the precision scores are between 0.70 and 0.75, the worse precision score is 0.68 calculated by the SIG_SVM algorithm while the best precision score (0.75) was calculated by the Naïve Bayes algorithm. The precision is important as it means that 75% of my positive samples are classified as positive samples and the rest of the positive samples are classified incorrectly- this is not a perfect performance of precision but still good for prediction.

Recall- the recall scores in our model are almost the same when most of the scores are between 0.71 and 0.73. only one result is far from this range, and this is the result of the SIG_SVM algorithm- 0.68. All the other results are consistent and present a good recall level, when the algorithms of POLY_SVM and the Logistic Regression have the highest

scores of 0.73. Why is it important for our model? The recall shows us how many total relevant results correctly classified by my algorithm- in this model around 72% of all the relevant result are correct.

F1- score- the f1- score was calculated by all the mentioned algorithms as well. The results for the f1- score range from 0.68 to 0.73. same as the previous results, the lowest score made by the SIG_SVM algorithm and the highest score made by POLY_SVM and the Logistic Regression algorithms. The f1-score is necessary for our model since it's harmonized the precision and recall and after calculating it the f1 score is above 70% which again, not perfect but a good makes our model predicable and valuable.

According to the findings above, either one of the models presented above which is not SIG_SVM or Naïve Bayes will be appropriate for the prediction of the credit card approvals factors. As we can see, most of the models perform a consistent and stabilized results. There are not significant differences between the factors, for example, we don't see a big difference in the precision or recall in most of the models. However, in the Naïve Bayes model, we could see a higher score of precision compared to the other models and a lower score for the f1-score compared to other models -it's clear when looking at the combined graph of the weighted Avg for each one of the models (figure 5). As for the model SIG_SVM, as presented above, the score for each factor is significantly low compared to the average of the other models, therefore, I believe this model is not accurate.

To conclude, this project, including all the models in it, is a good tool to predict whether a person's application for credit card will approved per the features of Gender, Married, Debt, Employed, Income. I believe that the results for the all the factors in the model could be higher if I would add more features, but the point is that I wanted to add just the basic features to predict the credit card approvals. Also, I thought of continuing project to predict which feature has more impact on the application than the others. In a personal point of view, because of this project, I developed more technical skills such as plotting graphs, coding different types of models, analyzing the result and more. In the non-technical side, I learn a lot about the process of approving credit card, what parameters are more important and what parameters are less important. Also, I learn how to organize data and when I did it by order it made the technical part easier for me.

Appendix

Figure 1:

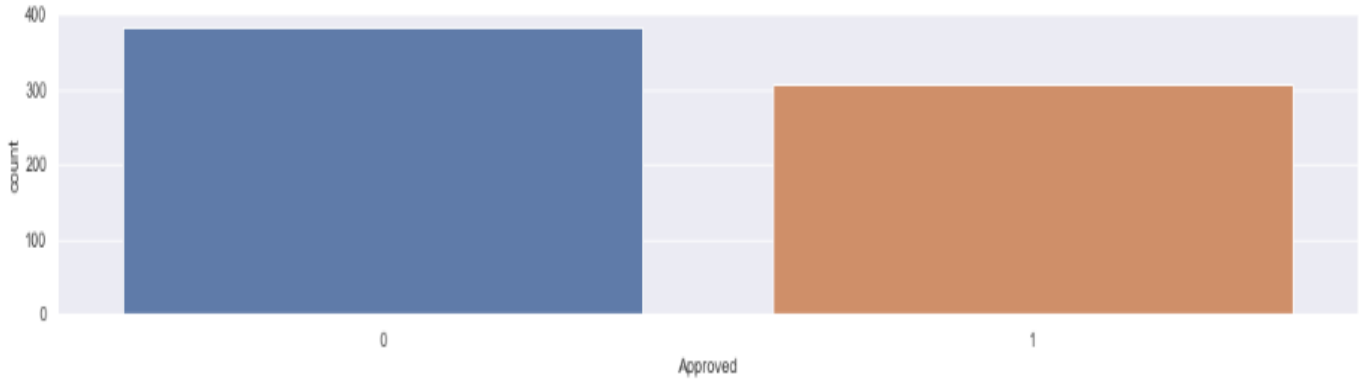
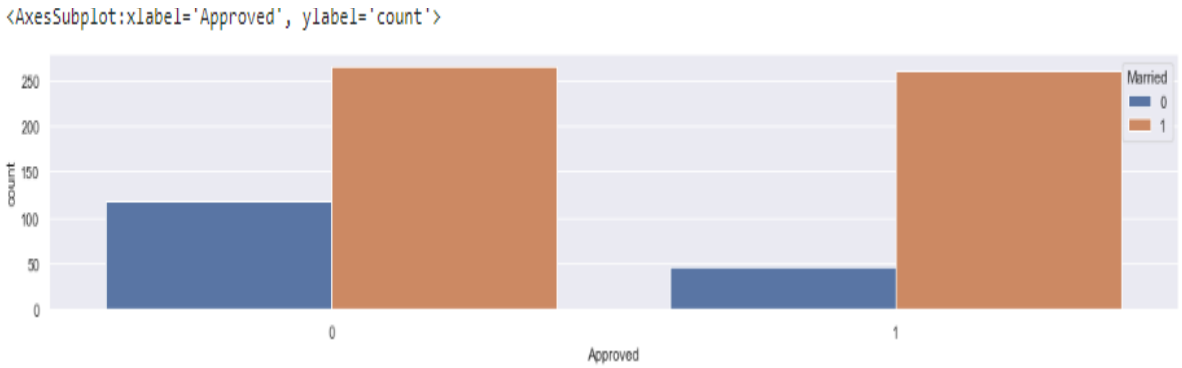


Figure 2:



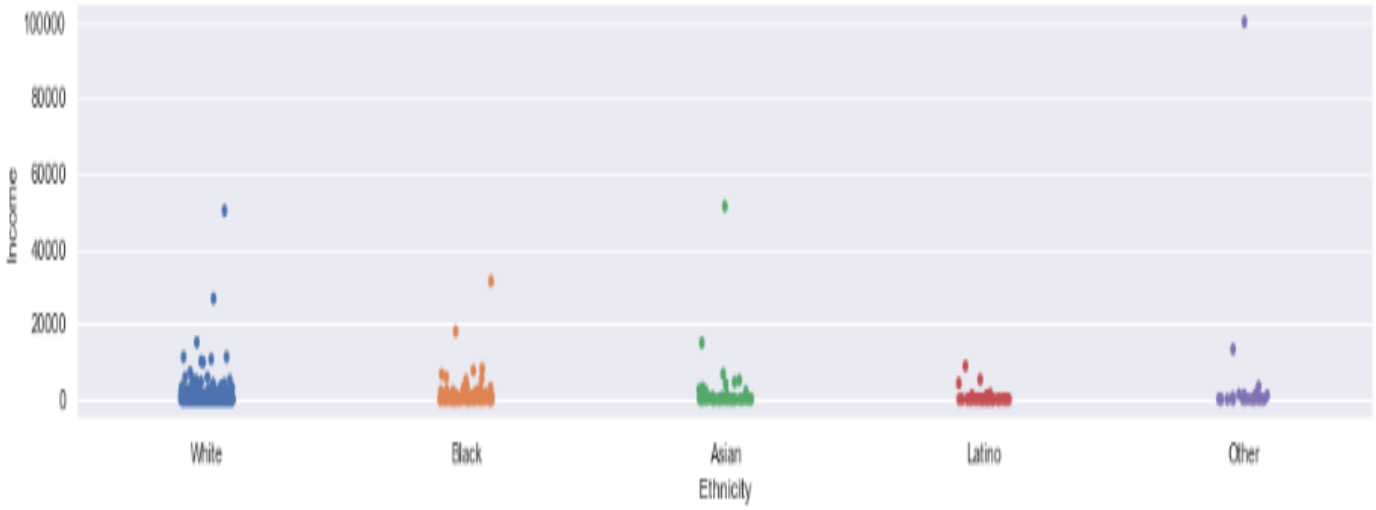


Figure 4:

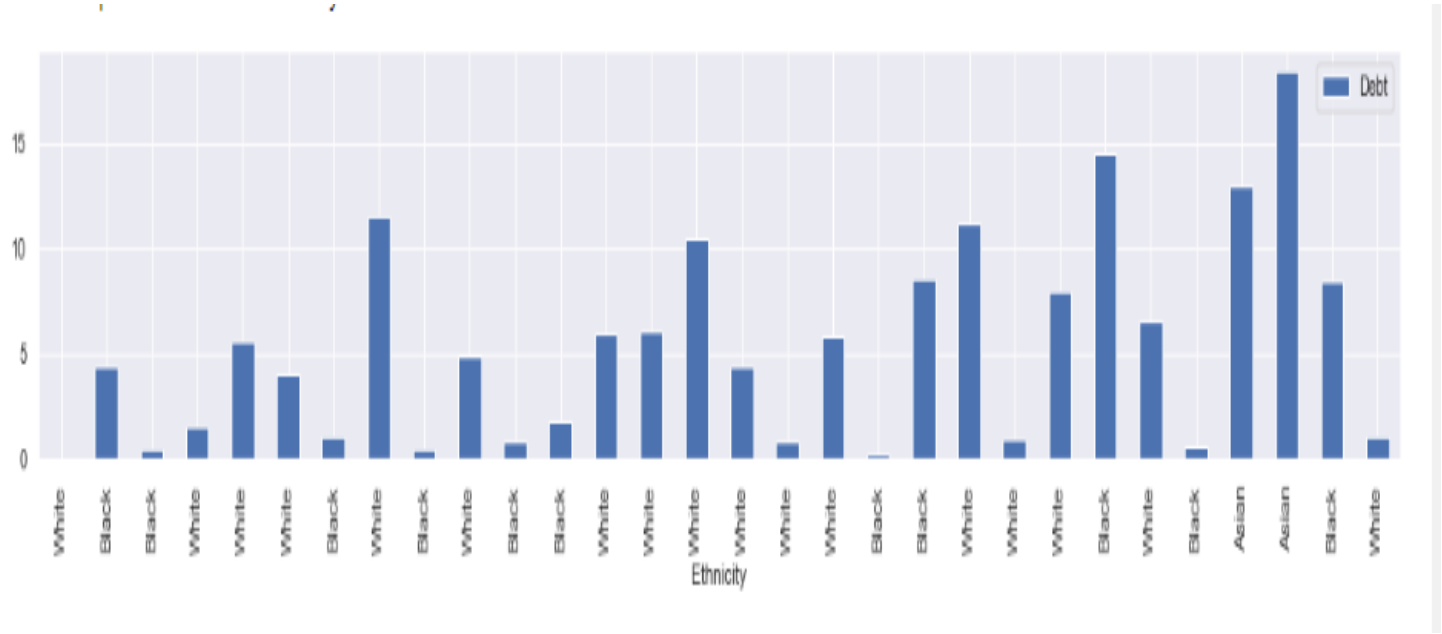


Figure 5: