Justin Madsen

DSC550/Brett Warner

Term Paper

Introduction

The problem I chose to tackle was if I could build a model to accurately predict who would take out a personal loan in the next year. The dataset I used was designed to help a bank target its advertising to those that were most likely to take out these loans. This project is critical to increasing return on investment for advertising campaigns. With properly targeted ads, the bank executives will have more freedom to allocate excess funds for other goals.

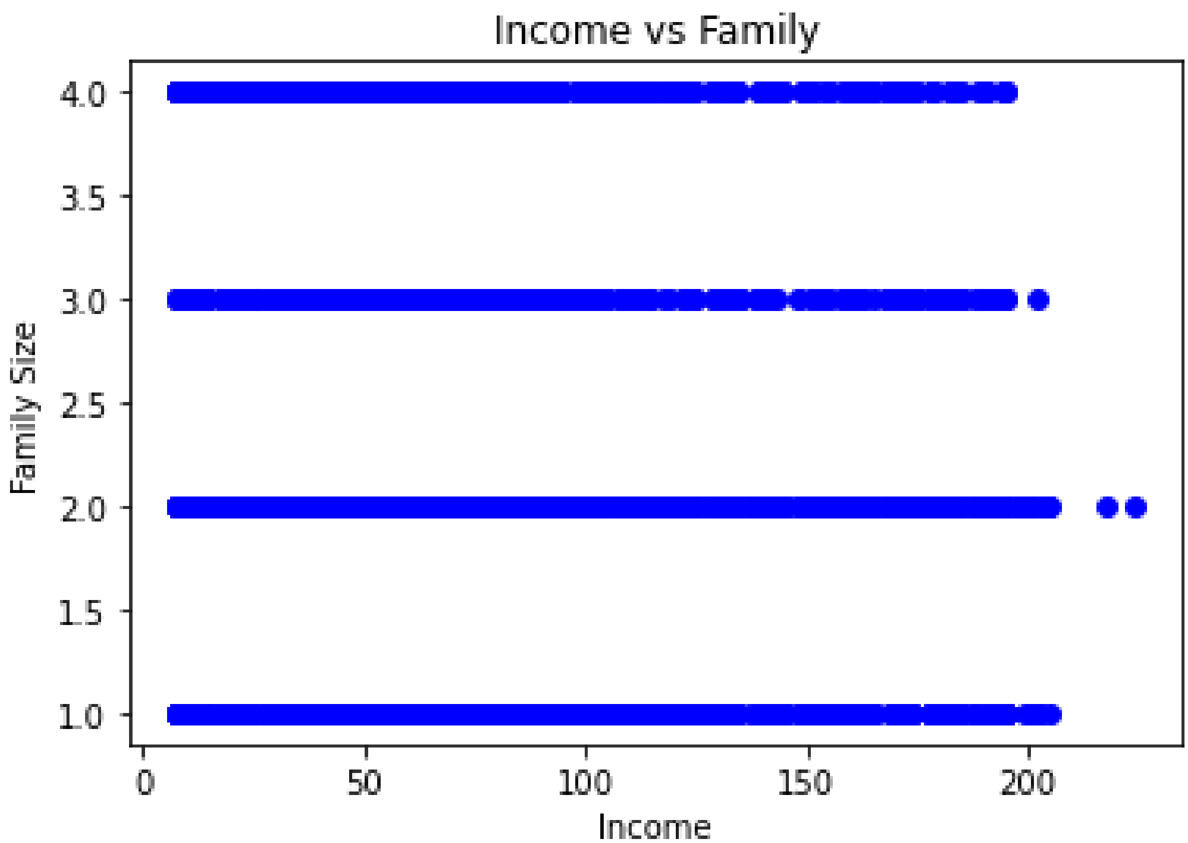
To appropriately pitch this idea, I would like access to previous and current budgets for advertising compared to previous profit margins from personal loans during the same timeframe. Comparing previous returns on investment alone will allow me to steer the conversation to a numbers game that the executives will easily understand. The dataset for this project was obtained through Kaggle, however, it is meant to mirror the data available for current customers at any financial institution that would be interested in this process. The data set is a record of 5000 bank customers and 14 variables describing their individual financial status.

Data Preparation

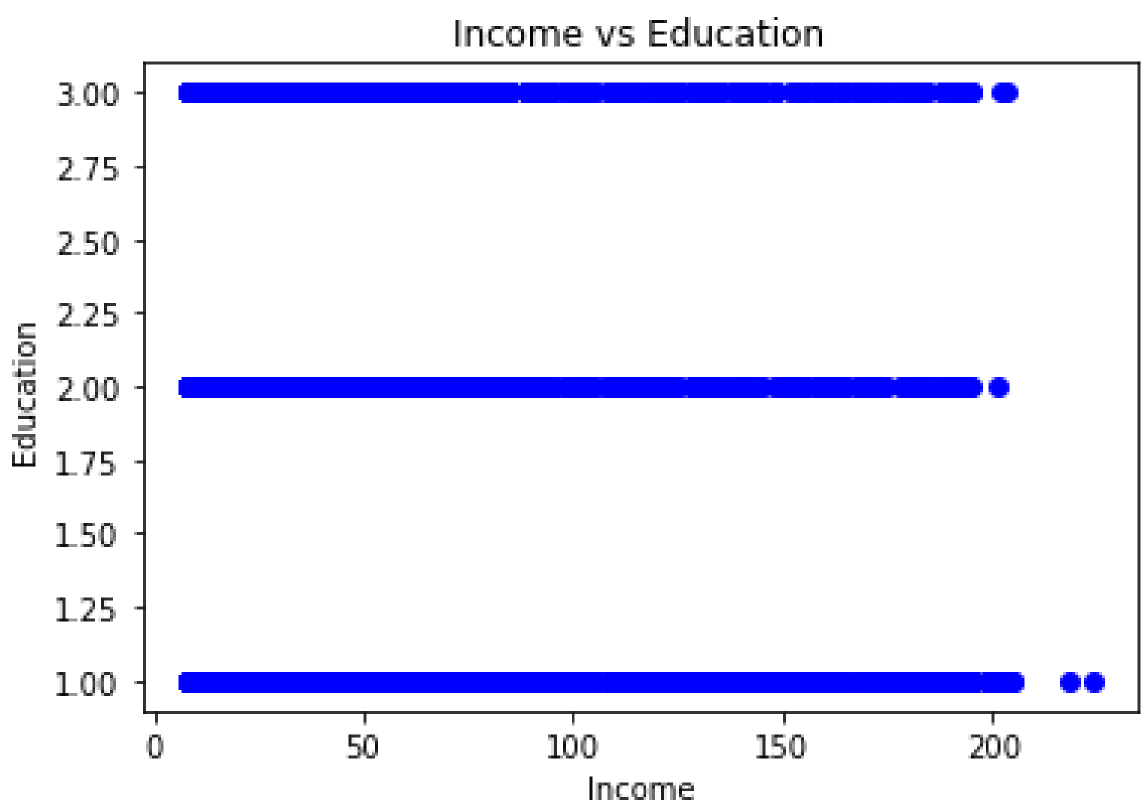
Once I had the data, I went through the basic steps for exploratory data analysis. This included looking for missing values or even erroneous values. There were a few entries for work experience that were negative values, and those were treated as erroneously put it in as negative, so the absolute value was taken to correct those errors. Once corrected, I looked at how income was impacted by various variables. For starters, I looked at Income vs Family Size, the highest incomes were in the married with no children category. As can be seen below, there were no indications that family size held a major impact on the family's income.

Next, I compared education level to a person's income. Consequently, I did find that the lower level of education had the highest incomes. I chalked this up to having more time in the workforce in lieu of being stuck in academia and improving their resume with a longer work history. Outside of that, income appeared even across the board independent of education levels. I then looked at how income compared to experience. This showed that most incomes fell below $100,000 a year, regardless of experience levels. The higher incomes were significantly more sparse, independent of experience levels. This indicates that most of the bank's customers fall in the same buckets of income levels, meaning that income will not be indicative of financial well-being on its own.

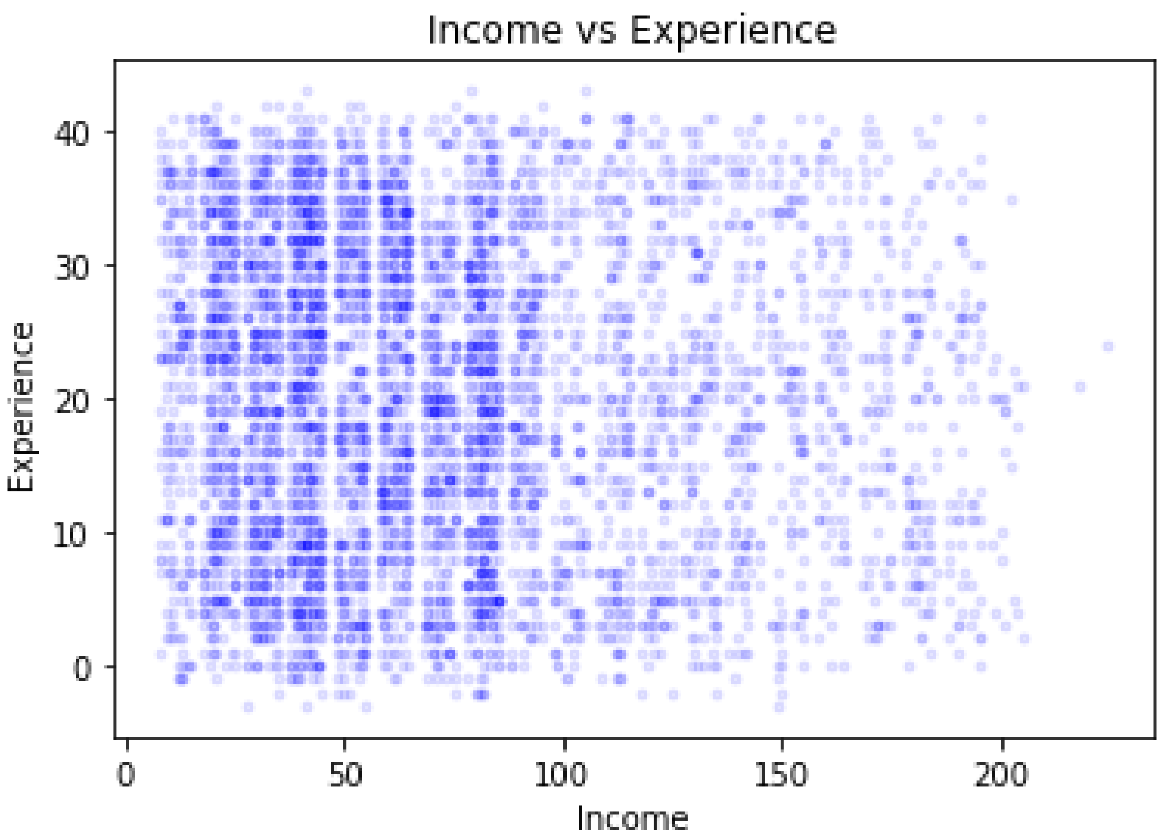
Finally, I plotted how an individual's average credit card balance compares to their income. As one would expect, we see a cluster of data points below the $100,000 income. However, we also see that as income increases there is a near-linear growth to average credit card balance as well. This graph indicates that most people try to stay within acceptable balances compared to their income, avoiding crippling credit card debt while they get by.



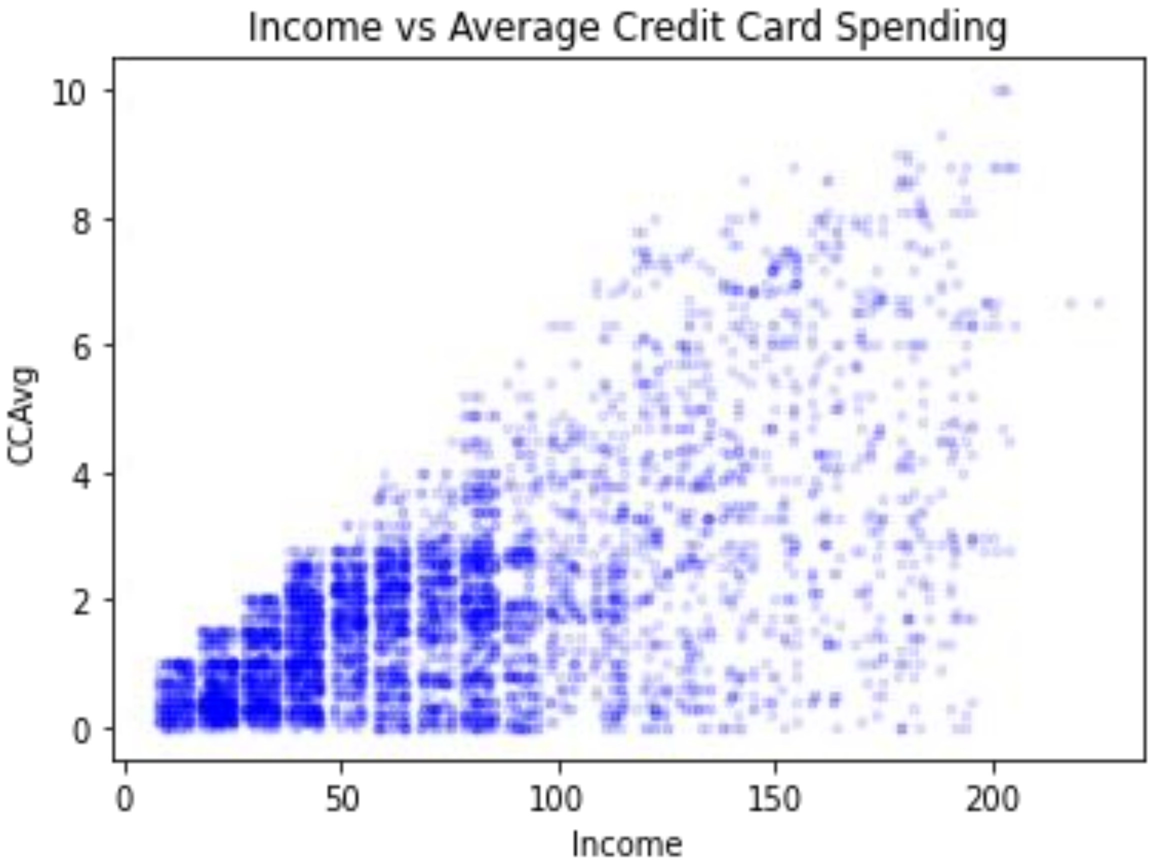
Income vs Family
Family Size
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
50
100
150
200
Income



Income vs Education
Education
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0
50
100
150
200
Income



Income vs Experience
Experience
Income
40
30
20
10
0
0
50
100
150
200



After analyzing these charts, I started working on preparing the model for this dataset. The goal was to model the likelihood of someone taking out a personal loan, so that was my target, whereas the rest of their financial portfolio was used as the features. These features included age, experience, income, ZIP code, family size, credit card average, education, mortgage, and various accounts through the bank. The previous year saw 9% of their clients taking out a personal loan, so this was used as our training data.

With 5000 clients, I felt an 80/20 split on training and test data was sufficient. When selecting the model, I decided to select a Decision Tree Classification model since this was a logistic problem set and not a linear set. The model ended up being 90% accurate in predicting who would take the loan, while 99% accurate on who would not take the loan. This was not surprising considering that the sample size for those who took out personal loans was only 9% of the entire data set. The entire model was 98%

accurate, however this model would benefit from continued training with more historical data before being fielded to stakeholders.

Conclusion

I would have liked to spend more time diving into generational differences for a better breakout of how different generations are handling their finances and their overall job history. Studies have shown that changing jobs every few years is better for your salary than staying at the same company for as long as possible. This is a noticeable difference between how older generations have viewed company loyalty. I would also like to see student loan data in this dataset. Seeing a more rounded financial portfolio would benefit the model. With the shift in tone around student loans over the last few years, I would imagine those with student loans are less likely to pull out another personal loan unless it is for loan consolidation.

There are potential further ethical concerns when advertising based on economic status is being employed. Payday lenders are widely considered financial predators and choosing to target those in dire financial straits could negatively impact the public opinion of the bank. Providing loans to only their top clientele can also appear to be picking favorites if not marketed appropriately. I would also like to see how often someone who took a personal loan out takes out another one shortly after. Having this data could skew how accurately the model predicts the outcome.

In conclusion, this data set shows that it is feasible to advertise based on past loan usage and current economic standing. As with all investments, there is a concern that past performance will not be indicative of future results. A more robust data set with more history would be the preferred data for this sort of problem set.

References

Pelta, Rachel. “Job Hopping with Intention: Pros, Cons, and Considerations: FlexJobs.” *FlexJobs Job Search Tips and Blog*, FlexJobs.com, 11 Mar. 2022, https://www.flexjobs.com/blog/post/job-hopping-v2/#:~:text=Increased%20Salary,receive%20a%205.3%25%20salary%20bump.