DSCI 470 Project 1 Loan Prediction Data Set

Grant Dawson and Ryan Rosiak

September 2021

1 Introduction

Our data set was acquired from "kaggle.com": a machine learning and data science community page where people share data sets with one another. We chose a loan prediction data set. Each instance in the data set is a person with data consisting of income, age, professional experience, marital status, house ownership, car ownership, profession, residing city, residing state, and risk flag. Essentially, this data is to be looked at from the point of view of a bank. A bank cares about return on investment and must look at these specific factors to assess the risk of an individual. This process of assessing risk and determining loan status has been performed for many years. With data science and computers, we can mitigate the timely manner of assessing risk by hand and automate the task by processing large chunks of data at once. Furthermore, we can use machine learning to further assess risk using the computer rather than an actual person.

2 Attribute Analysis

- ID - Numerical - Quantitative
  - An arbitrary number that uniquely identifies each instance
- Income - Numerical - Quantitative
  - The person's yearly income, assumed all in Rupees (Rs.)
  - Data: Between 10k - 10m (Rs.)
  - Mean: 5.03 million Rs (68,454 USD)
- Age - Numerical - Quantitative
  - The age of the person physically
  - Data: Between 21-79 years
  - Mean: 50.1 years
- Experience - Numerical - Quantitative
  - The age of a person's career
  - Data: Between 0-20 years
  - Mean: 10.1 years

- Married/Single - String - Categorical
  - The marital status of a person
  - Data: Single/Married
  - 90% Single — 10% Married
- House Ownership - String - Categorical
  - The person's living situation
  - Data: Rented/Owned/norent_noown
  - 92% Rent — 5% Own a home — 3% Say no to both
- Car Ownership - String - Categorical
  - The person's personal vehicle situation
  - Data: Yes/No
  - 70% No — 30% Yes
- Profession - String - Categorical
  - The person's job title
  - Unique entries: 51
- City - String - Categorical
  - The city in which that person resides
  - 317 unique entries/places
- State - String - Categorical
  - The state in which that person resides
  - Unique entries: 29
- Current Job Years - Numerical - Qualitative
  - The number of years that a person worked at their current job
  - Mean: 6.34
- Current House Years - Numerical - Qualitative
  - The number of years a person has lived in their current living situation
  - Mean: 11.99
- Risk Flag - Numerical - Categorical
  - Whether or not a person was approved for a loan

3 Research Questions

Some research questions that we thought of are provided below to aid our analysis.

1. Does more experience mean higher income?
2. Does geographic location mean higher income?
3. Can you make a model to guess risk factor? Can prediction models help?
4. Are there certain jobs that reside in certain geographic locations?
5. Do certain geographic locations have higher rates of renting than others?
6. Do certain geographic locations have higher rates of loan acceptance?
7. Do certain geographic locations have lower rates of people without cars?
8. Does where you reside affect the loan process?
9. What is the difference in being considered risk/non-risk as it relates to being married/single?
10. Do older individuals have more professional experience?
11. What professions have the highest incomes?
12. Do people younger than 22 generally have relatively less job experience than people older (out of college)?
13. Does professional experience determine car ownership?
14. Does marital status have an affect your living situation?

4 Graphs/Analysis

After some basic analysis and observations, we concluded that the data set was fabricated using a uniform distribution. The first red flag we found from the data set is visible in Figure 2. As you can see clearly in the graph, there is no real distinction between age and experience. The correlation to be made should be that younger people will have a relatively smaller amount of professional experience than older people. We would expect people in their early 20s to have little to no professional experience compared to someone in their 40s or older. What's alarming is that the average 18 year old claimed to have approximately 9.4 years of professional experience. That would mean that they would have started professional work at 11 years old. There is a clear disconnection between us and the data collectors. We are either misunderstanding what the data set means by "professional experience" or this data was randomly generated with a uniform random distribution. A uniform random distribution seems highly likely when looking at how close every age bracket is in the graph. The results seem to be too far stretched from reality.

Profession	Income
Police Officer	$9999938
Librarian	$9999400
Drafter	$9999180
Aviator	$9998280
Secretary	$9998070
Designer	$9996946
Statistician	$9996861
Computer Hardware Engineer	$9996192
Surgeon	$9995445
Biomedical Engineer	$9994932

Table 1: Top Ten Highest Paid Professions (Max)

The second instance that questioned the data set's validity and reliability can be observed in table 1 above. Here we see the top 10 highest payed positions. While there is a wide range occupations, the range of pay is almost nonexistent. This is a large issue because in the real world, there are clear gaps in pay based off occupation. The reality of a librarian making as much or more than an engineer is astonishing. Regardless of this fact, the max professions are all paid within a couple $100 of each other. This creates a likely possibility that each data point had a max value of 1,000,000 rupees and the instances were selected from a uniform random distribution. Further analysis of the data set also revealed two instances with the occupation flight attendant, one with an income roughly over 50% higher than the other. Since their age was also within 5 years of each other, this seemed highly unusual. In total, this data does not make any sense other than it being sampled from a uniform distribution.

Profession	Income
Petroleum Engineer	$5443309
Psychologist	$5357795
Drafter	$5336802
Scientist	$5282710
Surgeon	$5235358
Comedian	$5199538
Chemical Engineer	$5189804
Mechanical Engineer	$5175032
Artist	$5164765
Financial Analyst	$5145752

Table 2: Ten Highest Average Paid Professions

Table 2 above fully solidifies the fact that the data set was sampled from a uniform normal distribution. The top 10 means of each profession are very closely related. We see the same problem regarding a comedian making on average more money than an engineer. In reality, this just does not make sense. Since the maxes and the means are so closely related, all roads lead to the sampling of a uniform distribution.

Attribute	Mean
Income	$5029562
Age	50.05
Experience	10.12
Current Job Years	6.34
Current House Years	11.99

Table 3: Means of empirical attributes

In table 3, we can calculate the population mean and see that the data is uniform and centered. The income ranged from 10k to 10 million and the mean is 5 million, the age ranges from 21 to 79 and the mean is 50, the professional experience ranges from 0 to 20 and the mean is 10, and so on. This further enforces the fact that this data is just simply generated from the uniform normal distribution and would mean nothing to make a model for.

Our goal for generating figure 1 below was to answer the question; "What is the difference in being considered risk/non-risk as it relates to being married/single?" (#9). The results are slightly skewed since there are much more single people than married but we can notice that the single bars and married bars are very similar shapes. It looks as if the single person bars are just multiples of the married person bars.

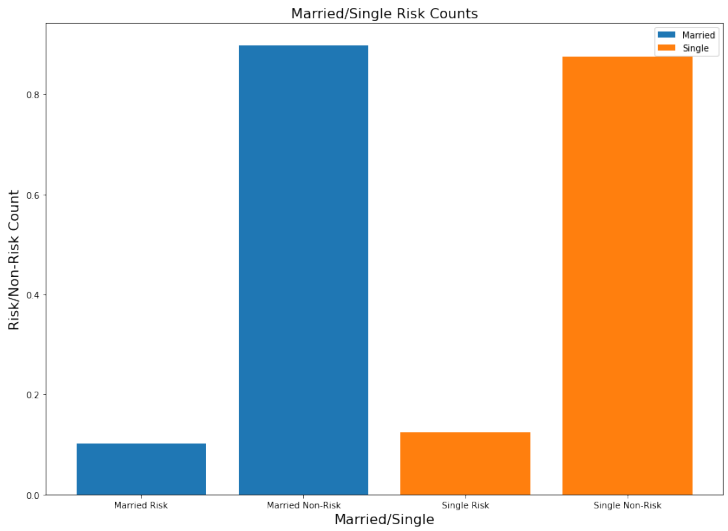

Figure 1: We are displaying marital status vs if that person would be a liability. The data is normalized to emphasize the quantity of each category. The correlation coefficient is 0.021.

In table 4 below, there is some interesting information about the data and what is to be expected from the data. This was truthfully misleading in the initial stages of analysis. Although there is a wide range of data types, the data is simply fabricated.

Attribute	# of Unique
Income	9916
Age	59
Experience	21
Married/Single	2
House Ownership	3
Car Ownership	2
Profession	51
City	317
State	29
Current Job Years	15
Current House Years	5

Table 4: Number of unique entries into the data set for each attribute

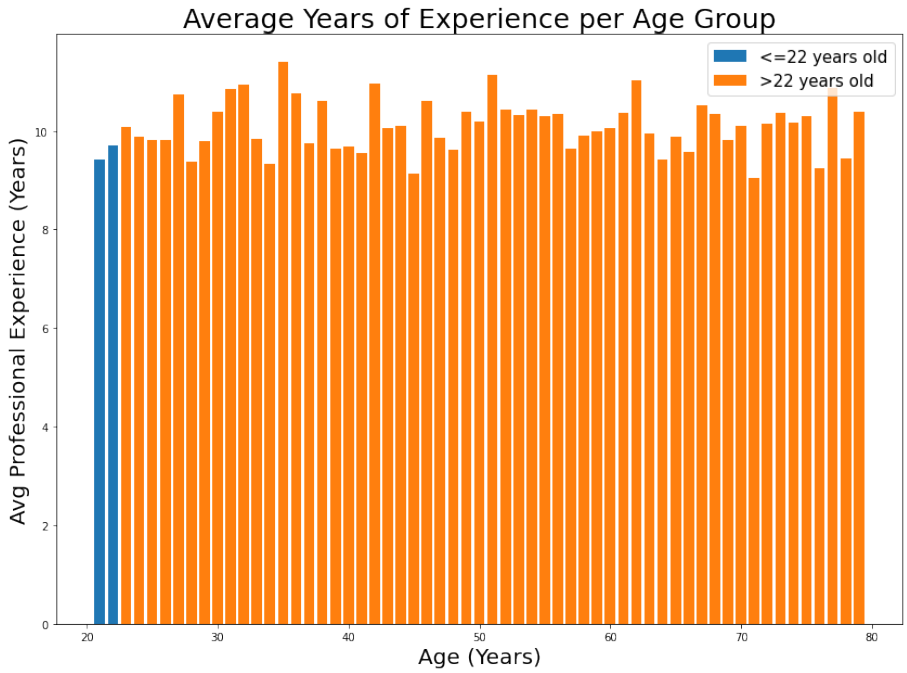

Figure 2: We compare age to average professional experience. There is no correlation between age of a person and the amount of experience they have. For instance, if a 21 year old had 9 years of experience like this data is saying they would have had to start their professional job at 12.

In Figure 3 below, we can see that the bars look like they are multiples of each other. Overall, this data does not provide much validity. Since the data is so skewed to rented households and the amount of single people is so much larger than married people, it is nearly impossible to draw conclusions because the data is so disproportional. It seems clear from the graph that most of the people who own some type of house are single. But, the data for married individuals is so small that it cannot even be compared to the single residents. The data is very off.

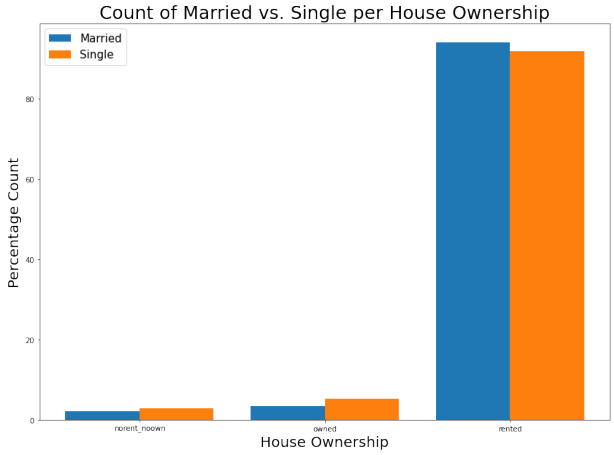

Figure 3: We compare marital status to whether someone owns a home. The figure would be better if we had normalized because there are more people single than people married but we can see that every house ownership answer is a scaled multiple of one another.

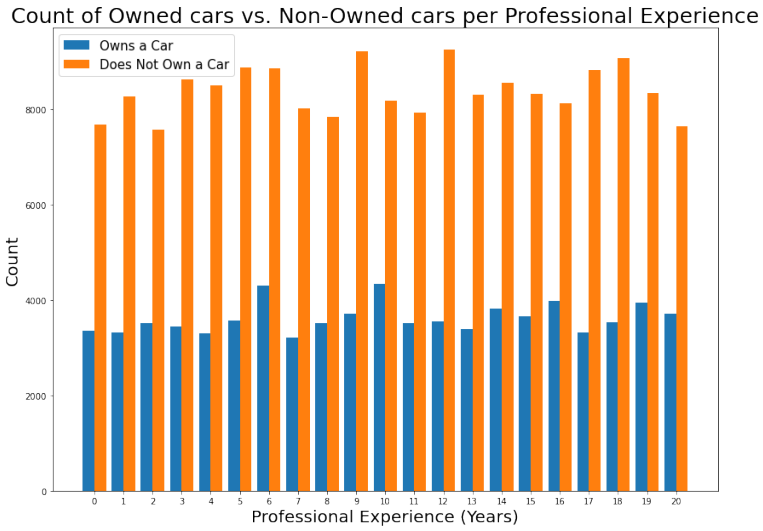

Figure 4: We compare the years of experience to whether or not a person owns a car. People at all experience levels tend to own a car at similar rates.

Figure 4 from the previous page answers the question "Does professional experience determine car ownership?" (#13). For the last time, we see the uniform normal distribution between years of experience and car ownership. The amount of cars owned and non-owned within each profession is almost uniform. There is clearly no variation in the data when there should be a significant difference in car ownership among younger people just starting their career in comparison to those farther along in their career. No clear cut conclusion can be drawn from the data.

5 Conclusion

The loan prediction data set analyzed in this paper looked to be a solid basis to model risk assessment among various individuals. However, the instances provided in the data proved to be sampled from a uniform distribution. For mock analysis, this data set can be used to perform basic visual analysis and modeling. For our analysis, we would like to build a model to assess the risk of an individual and this data does not provide real life meaningful relationships that can be used by a prediction model. There were many factors analyzed including professional experience by age, car ownership with professional experience, house ownership among married and single individuals, and years of experience based on age. Not a single correlation could be made amongst our analysis because of the uniformity in the data. We will use this data as an example to spring board off of and find a data set that provides us with the necessary relationships to assess the risk of an individual. We now know what type of data that will be used to assess risk and what type of relationships will aid us in creating a model for such a task. In conclusion, this data set did not provide what we needed but we will use it to find a data set that will.