LLOYDS DIGDATA CHALLENGE

Ali Mroweh – amroweh@gmail.com

using the "Data Strategy" approach

Part I – Data Exploration

This section considers a review of the data and some basic analyses which should point us towards a more directed approach. Not all data fields were inspected; only a subset of the data which was considered potentially impactful was chosen.

Income vs loan status

The first question one might ask is "Do people who receive more money pay off their loans more than those who receive less?". To answer this question, we looked at the "annual_inc" vs "loan_status" data columns. Considering this is a numerical vs categorical comparison, we need to edit the data slightly to be able to do the comparison. First, we needed to divide the "annual_inc" column into different categories. We used the US tax brackets (7 categories) for this division, which seemed like a good approximation for such categories. Of course, this categorization could be done differently (Washington, 2022). Next, we counted the number of paid and total loans for each category, then divided them to get a representative ratio per category. This parameter, called "percentage loan payback", was used in the remaining tests of Part I of this report. These ratios were then plotted (Figure 1). Note: a higher ratio means that more people paid off their loans.

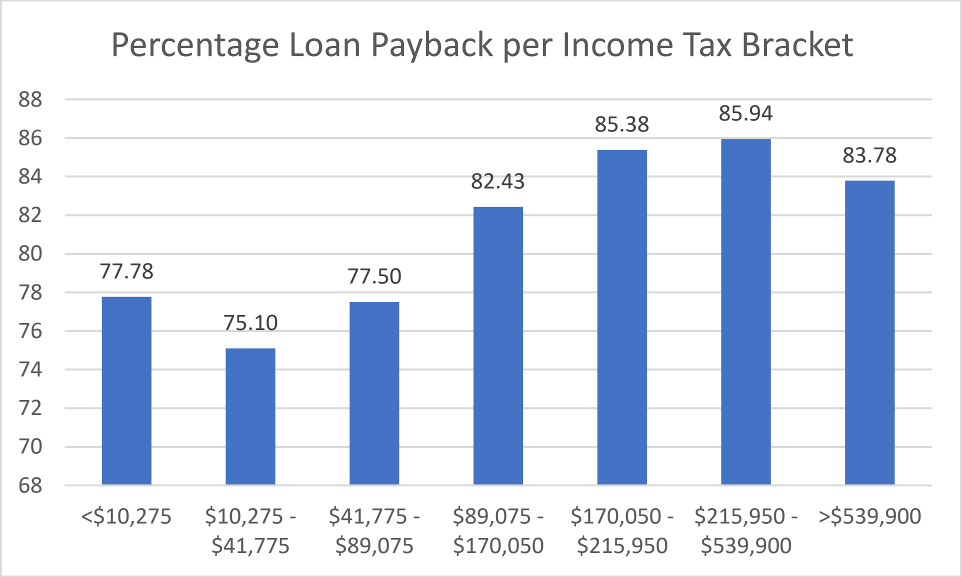


Figure 1 Paid to Charged Off client ratio per income (divided based on US Tax Brackets)

The chart indicates that a higher income generally correlates positively with a higher potential for loan payment.

Employment Title

Next, we decided to look at whether there are particular jobs that for some reason correlate with loan payment or charge off. We did so by aggregating the data based on "emp_title". This gave us how many people paid their loans and how many were charged off in each job. Considering the majority of these jobs had a small number of entries, most were not considered representative samples. In fact, we needed around 385 entries per "emp_title" to have statistically representative results. However, no "emp_title" had this many entries. So, we used a more lenient sample size of 100 as threshold under which the job entry was excluded from the analysis. This left us with 8 jobs containing enough data. For each of these jobs, we determined the percentage loan payback to get an estimate of how many people with such a job pay off their loans. The results are shown in Figure 2:

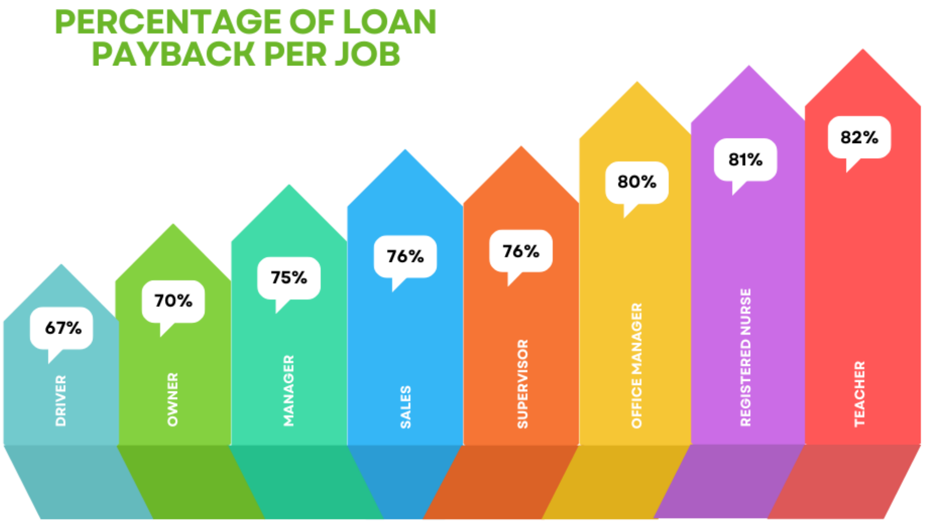


Figure 2 Percentage of Loan Payback per Job Title

It seems that teachers, registered nurses, and office managers pay back their loans more than drivers or owners.

Home Ownership

In a similar fashion, a close look at the different home ownership groups shows that people who rent (75.23% payment rate) are less likely to pay back their loans compared to those who own their homes (78.71%) or have a mortgage (81.54%) (Figure 3).

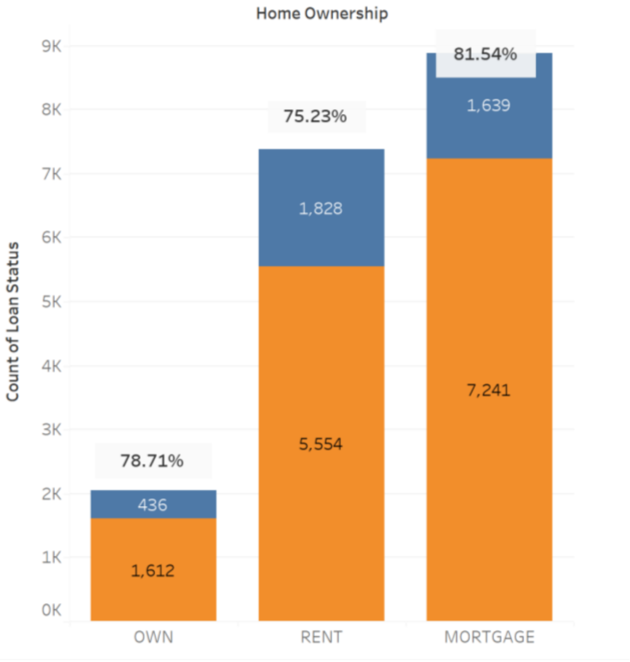


Figure 3 Loan payment breakdown based on home ownership

Loan and Instalment Amount

The loan amount also seems to affect whether people pay back the loans. It seems that the smaller the loan amount, the more likely that it is paid (Figure 4), with the exception of the last loan amount (40,000 USD). It is unclear why this group had a higher payment ratio compared to the previous groups.

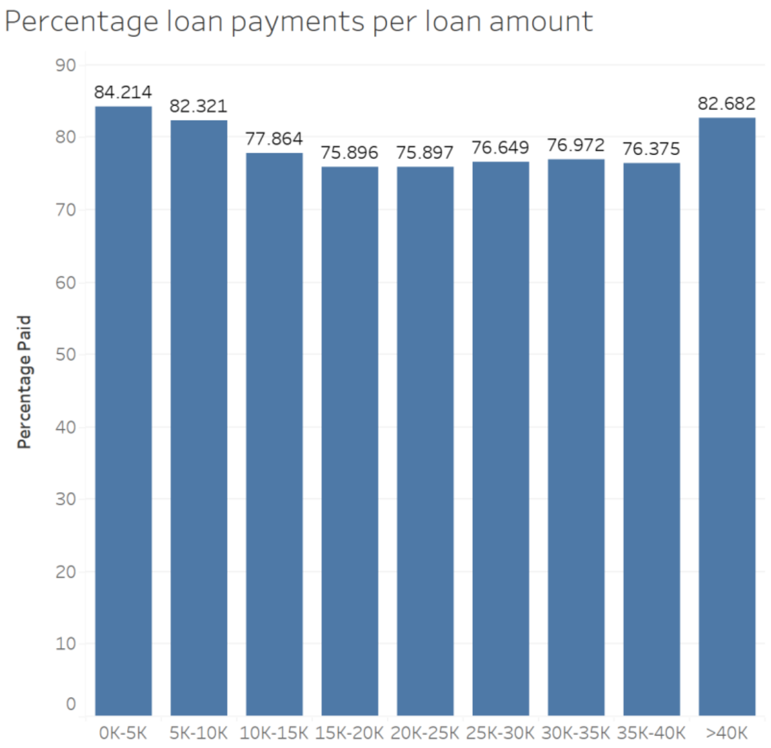


Figure 4 Percentage loan payment based on loan amount

The instalment amount shows a slightly different story (Figure 5), where the pay back starts high initially, decreases quickly for the next few categories, and then almost stabilizes. It is also interesting to see that the last category (1400-1500 USD) had a relatively elevated payment ratio compared to the previous categories. This is similar to the previous graph, indicating that there seems to be some kind of relationship between the loan amount and the instalment value at this range.

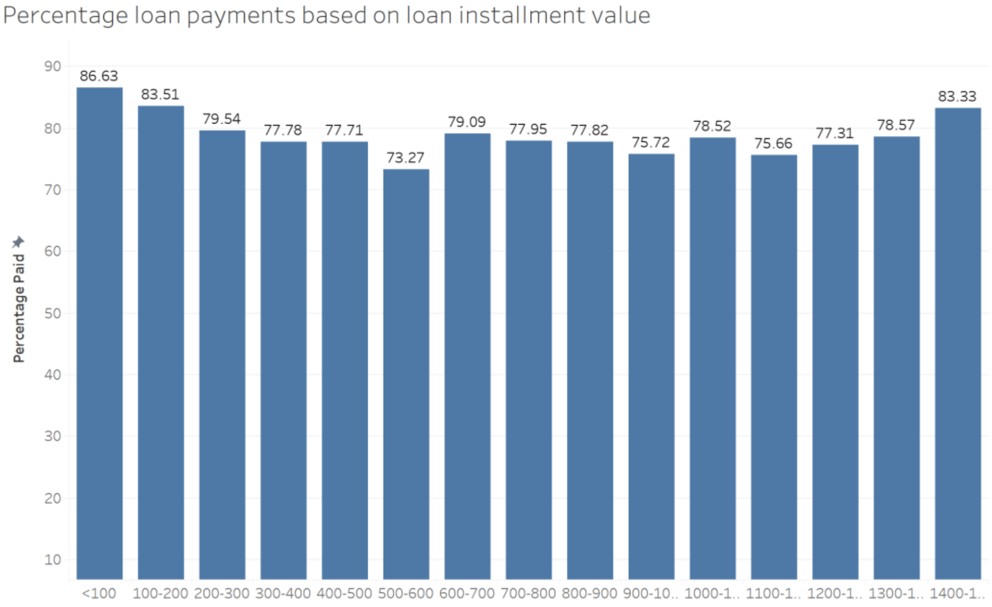


Figure 5 Percentage of loan payments based on loan instalment value

Loan Term

It also seems that loans paid over shorter periods are more likely to be paid back. Looking at the graph in Figure 6, around 83% of 36-month loans were paid back vs. around 68% for 60-month loans.

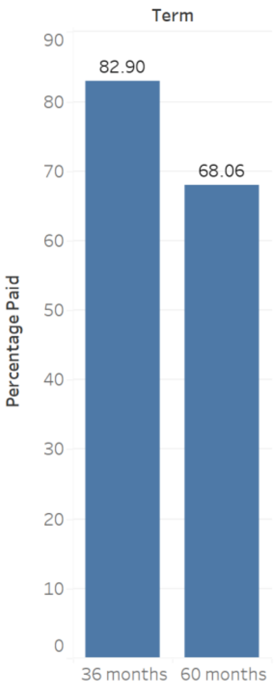


Figure 6 Loan payment percentage per loan duration

Interest rate

Interest rates seem to play a very important role in determining whether people pay back their loans or not. We have divided the interest rates into 5% bins and plotted the change in percentage paid (Figure 7).

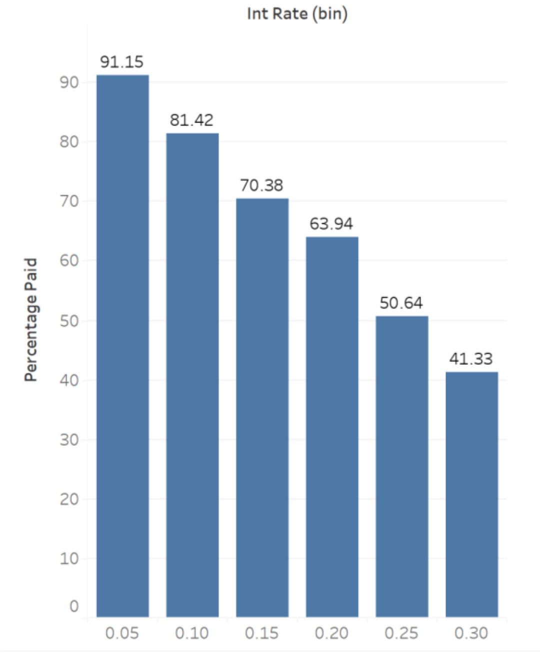


Figure 7 Variation of loan payment percentage as a function of increasing interest rates

Clearly, the higher the interest rate the more likely it is for the client not to pay back the loan. The relationship between these two factors seems to be very linear and strong, and indicates that for loans with an interest rate higher than 30%, clients are more likely not to be able to pay back their loans.

Current Balance

The client's current balance is also a good indication of whether they will be able to pay back their loans. As seen in Figure 8, the percentage paid increases as the current balance increases. It can also be seen that clients with a balance of more than 100,000 USD are almost certain to pay back their loan. Note that there weren't enough entries to create meaningful deductions for groups over 140,000 USD. Thus, they were not included in the graph.

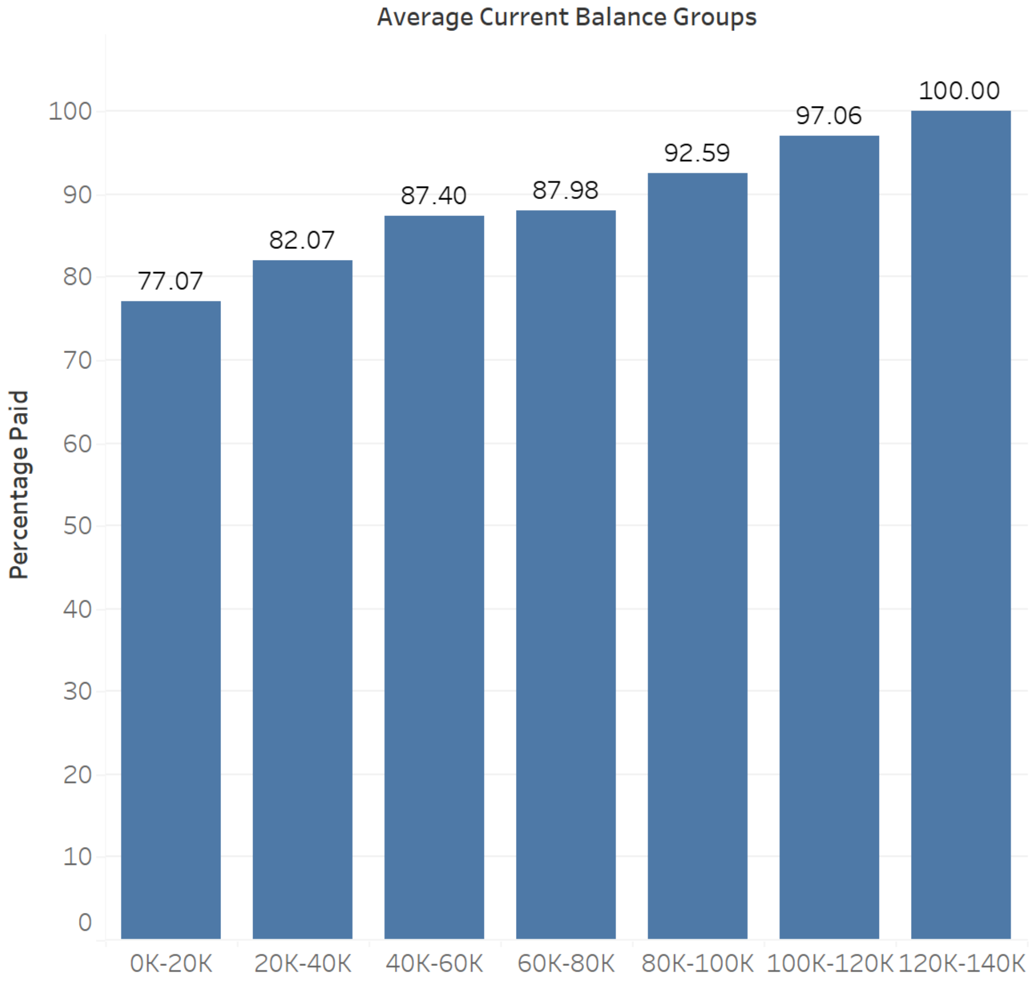


Figure 8 Percentage loan payback by average current balance groups

Maximum Current Balance

When looking at the maximum current balance owed on all revolving accounts "max_bal_bc", the graph shows a distribution that increases then starts to decrease (Figure 9). This is quite an interesting observation. It could be that those who are indebted in small quantities may be those with limited financial capabilities in the first place, and hence aren't able to payback their debt. Additionally, it could be that those who heavily owe on their credit cards or overdraft may be over spenders who take higher risks than others. People who are in between could be both financially capable and less of risk takers.

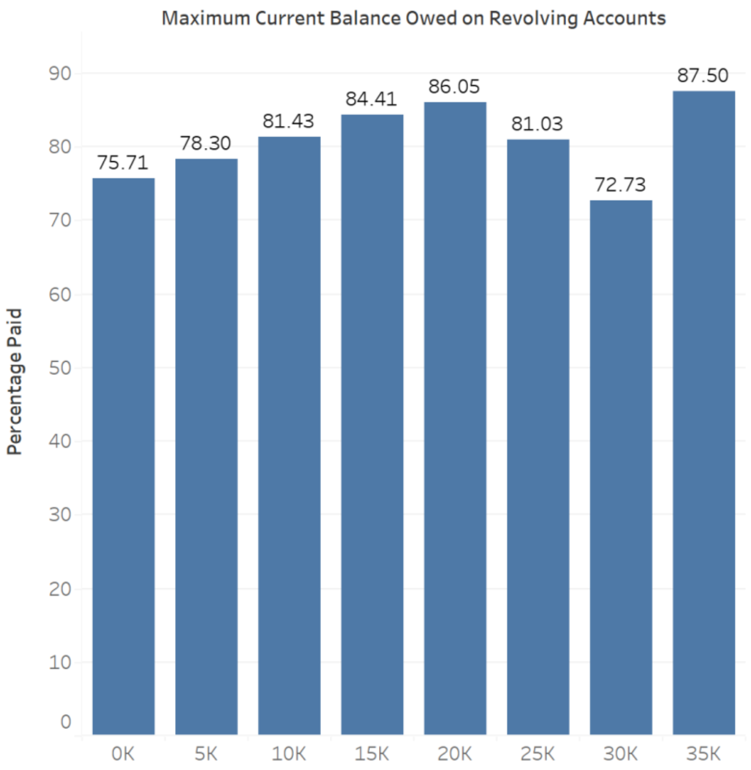


Figure 9 Percentage loan payback per revolving account debt

Public Record Bankruptcies

It is also important to consider that these clients may not only be individuals. They can also be companies or organizations. Interestingly, there doesn't seem to be a strong correlation between the number of public record bankruptcies and loan fulfilment (Figure 10). Note that there are only 2 accounts with 5 public bankruptcies and only 1 with 6. Hence, these entries were not considered enough to be included in the analysis and were discarded from the graph.

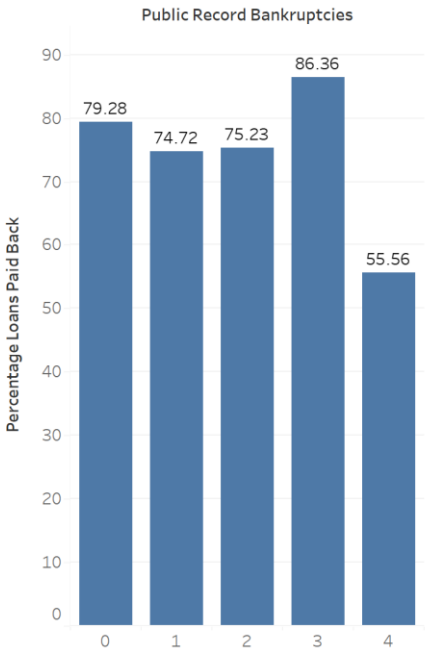


Figure 10 Percentage loan payback as a function of number of public record bankruptcies

Applicant Location

Finally, we looked at the addresses of the loan applicant. To determine whether this impacts loan pay back, we have visualized the percentage of loan payback (i.e. paid back loans / total loans) per state. The map showing this variation is shown in Figure 11 and Figure 12.

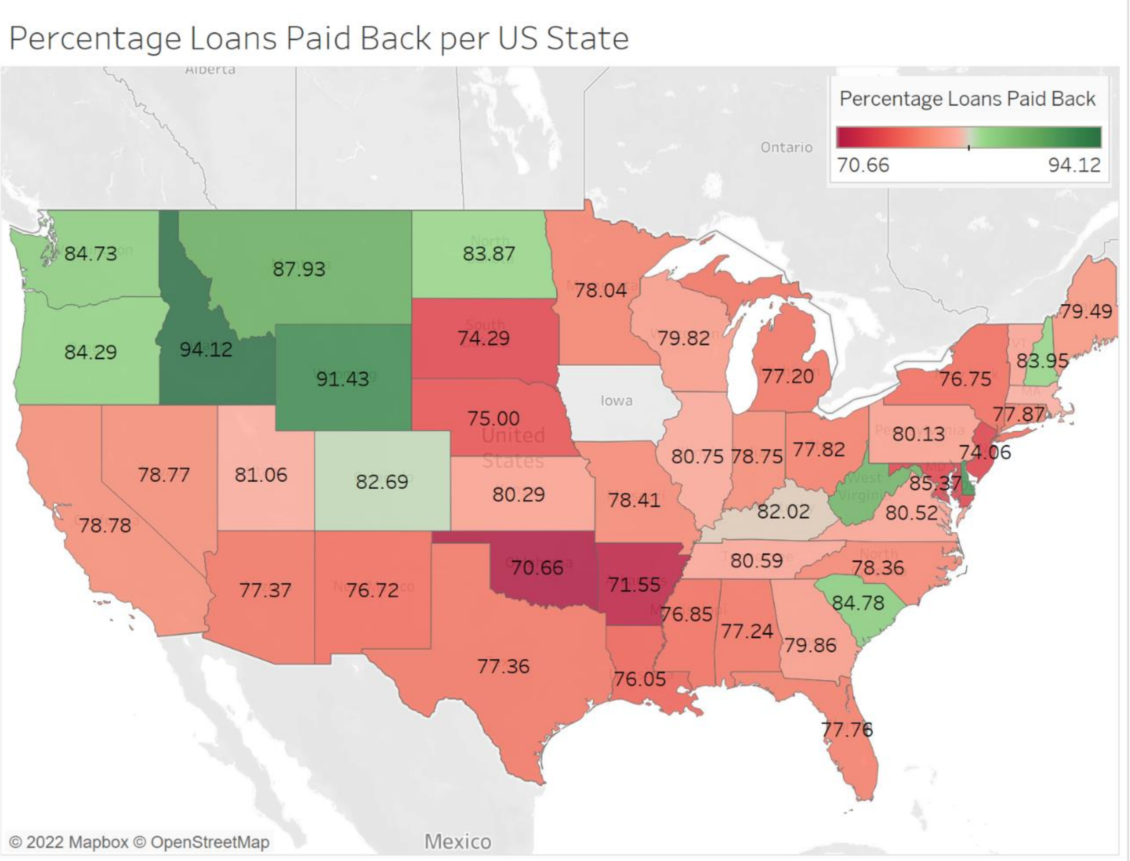


Figure 11 Percentage loan payback per US state (exc. Alaska)

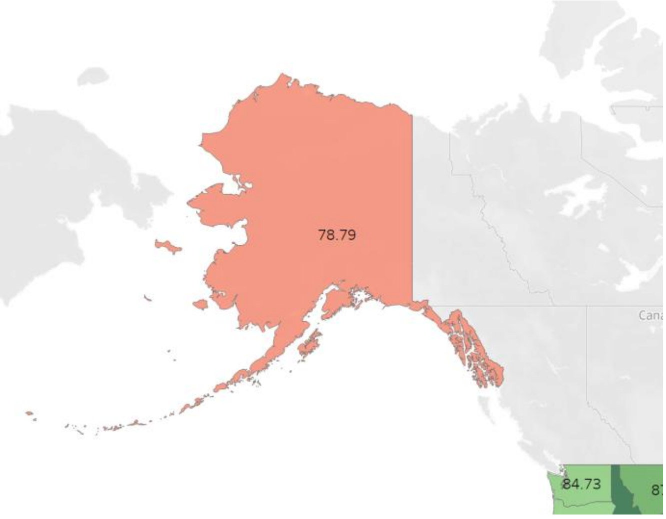


Figure 12 Percentage loan payback in Alaska

In general, it seems that north-western states have a higher loan payback ratio compared to other states. We have tried to match this map with other maps including those that show the “poor” vs “rich” states in the US. However, there doesn’t seem to be a direct correlation between the two. There does appear to be such correlation for many states in the western part of the US, but the eastern part shows high variability. Note that Iowa only had 1 entry, and therefore was discarded from the analysis and graph.

Part II – Major Loan Payback Factors

Based on the above, we have determined that the most important factors affecting loan payback are the following:

- Income
- Home ownership
- Loan duration
- Current balance

A relationship could be present between income and current balance. For instance, income could predict how much a person/entity have in their current balance. We decided to check if such a relationship exists. If a strong causational relationship exists, this could help us eliminate the current balance parameter as income would be sufficient to predict loan payback. To do so, we used a scatter plot and determined whether a linear relationship exists (Figure 13).

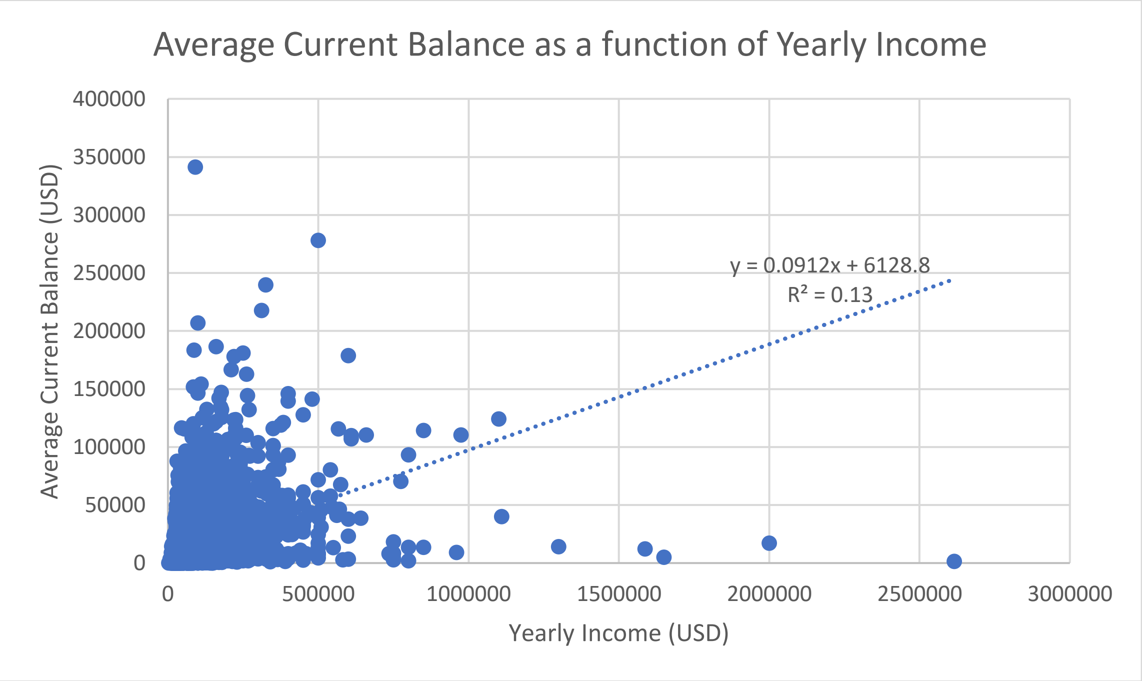


Figure 13 Relationship between average current balance and yearly income

For an R^2 equal to 0.13, this indicates a very low correlation between these two parameters. Hence, we must take both into consideration when analysing the profile of the loan applicant.

Part III – Proposed Model for Assessment of Loan Applicants

The preceding four factors can be divided into two categories: 1) Applicant factors: income, home ownership, and current balance, and 2) Loan factors: duration.

This division can help us create our model. The proposed model consists of the following steps:

1- Assign a coefficient for each of the major applicant factors
2- Divide each applicant factor into its sub-categories
3- Assign a coefficient for each sub-category equivalent to the percentage loan payback of that category

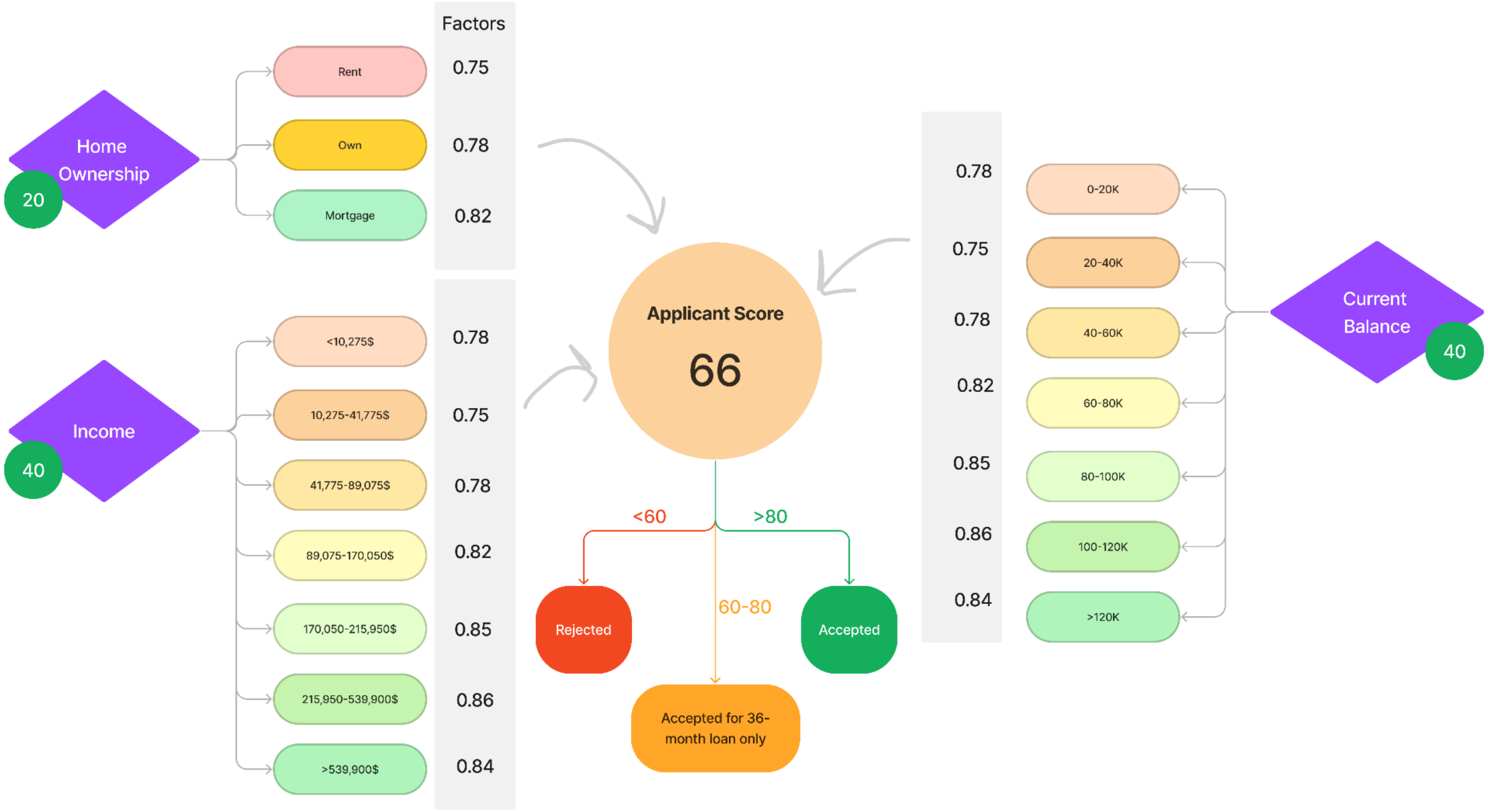
4- Calculate an aggregate score for each loan applicant by multiplying the coefficients of the subcategory that they belong to by the coefficient of the main category. This generates an applicant score.
5- Based on the applicant score, determine whether the applicant is given a loan, is refused a loan, or is given only a certain type of loan (i.e. the 36 month loan only and not the 60 month one)

This is illustrated in Figure 14 below.

Finally, the coefficients mentioned in these steps can be altered dynamically. Their values need to be inferred from additional data analyses to be performed on the major coefficients. This is outside the scope of this report, which relies on the “Data Strategy” approach. The diagram below shows only sample data.

Loan Application Proposed Model

Lloyds Digdata Challenge



Home Ownership 20

	Factors
Rent	0.75
Own	0.78
Mortgage	0.82

Income 40

<10,275$	0.78
10,275-41,775$	0.75
41,775-89,075$	0.78
89,075-170,050$	0.82
170,050-215,950$	0.85
215,950-539,900$	0.86
>539,900$	0.84

Current Balance 40

0-20K	0.78
20-40K	0.75
40-60K	0.78
60-80K	0.82
80-100K	0.85
100-120K	0.86
>120K	0.84

Applicant Score

66

<60

Rejected

60-80

Accepted for 36-month loan only

>80

Accepted

Figure 14 Proposed loan application model

References

Washington, K. (2022, 07/11/2022). 2022-2023 Tax Brackets And Federal Income Tax Rates. https://www.forbes.com/advisor/taxes/taxes-federal-income-tax-bracket/#:~:text=For%20the%202022%20tax%20year,filing%20status%20and%20taxable%20income