caption

SharpestMinds Project

Analysis of a Loan Dataset in an Irish Bank

Ben (Zhibin) Liu

Data Analyst

Supervised by

Mikhail Sidyakov

Data Scientist

May 19, 2020

Contents

1 Introduction

What major factors does the management team of a bank consider when they decide to grant a loan? Do they prefer to accept the loan application from clients with high income? These questions are preoccupying the minds of loan applicants around the world. Understanding this, I initiate to discover several factors which may play an important role in a bank's decision to approve a loan application, utilizing a mix of statistical analysis, visualization and machine learning model.

This project is based on a dataset that contains information about 1 million potential borrowers of a specific bank, which is a peer to peer lending bank based in Ireland. The complete dataset is from Lending Club, and it has been changed from Kaggle. The goal of this report is to extract business insights from the analysis of this large loan dataset. I am sharing this work to people who are interest in bank loans, to help them get a better understanding about the factors that may influence the outcome of a loan application.

2 Methodology

The purpose of this project is to drive insights from the loan dataset of the Irish bank. The complete loan dataset consists of 887,379 rows and 30 columns. To start with, I took 9 major features to conduct descriptive statistics, such as average, minimum and maximum. In addition, I applied various kinds of visualization (e.g. bar plot, pie chart and box plot) to display the distribution of each feature, to obtain preliminary understanding.

After researching each feature, I focused on their relationship. Firstly, the scatter plot and heat map are implemented to shown the pairwise correlation of all 9 features. Then, I chose some specific pairs of features to conduct further detailed analysis.

Last but not least, a machine learning method, KNN, is used to demonstrate how some features influence the evaluation of a loan application by the bank quantitatively. This classification model has a satisfactory predictive validity, so it can help us better understand the key factors of a loan approval.

3 Analysis

As mentioned in the last section, the target dataset contains nearly 1 million records and 9 major features, including 5 numeric features and 4 categorical features. To start with, I conducted descriptive analysis about these 9 features respectively, by their statistical values or distribution plots. Then I further explored the correlation between the features.

3.1 Descriptive statistics

To begin with, I analyzed the dataset by the first feature, annual income, which was generally regarded as the most significant factor for the approval of a loan. I extracted detailed statistics values in Table 1. Among nearly 1 million clients, the average annual income is around 75,000 euro. It is straightforward that the maximum value, 9,500,000 is an outliner, since the 75th percentile value is only 90,000. Furthermore, it is noteworthy that this bank classifies the clients by 3 categories, in terms of annual income. To be specific, clients with annual income less than 100,000 euro belong to category 1. For those who earn more than 200,000 euro every year, they are considered as category 3, and the ones with annual income between 100,000 and 200,000, belong to category 2. In others words, since the 75th percentile value is less than 100,000, at least 3 quarters of the clients belong to category 1.

Table 1: Statistics for annual income.

Statistics	**Value**
Count	887,379
Average	75,027
Minimum	0
25th percentile	45,000
50th percentile	65,000
75th percentile	90,000
Maximum	9,500,000

The second feature, known as the employment length by years, has possibles values from 0.5 to 10 years. We can take a look at the histogram of different categories from the following Figure 1. It is obvious that a major portion of clients have long employment length. This finding implies that this Irish bank probably prefers to grant a loan to the clients with relevantly long employment length.

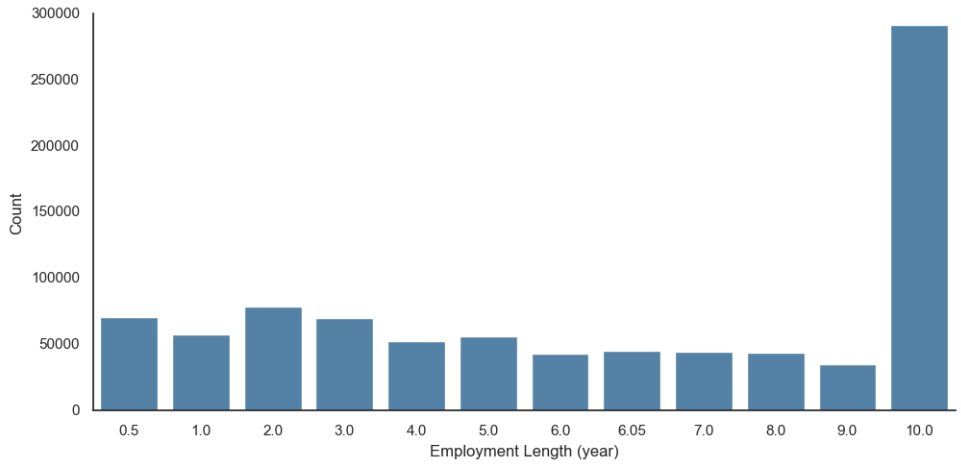

Figure 1: Histogram for employment length in years.

Then we take a look at the third feature, interest rate. It contains thousands of unique values, so instead of a plot, statistics indexes are better to summarize the key information. According to Table 2, the average interest rate is 13.25%. In addition, based on the other percentile values, it is close to a normal distribution. However, I implemented the hypothesis test to justify, and it turned out that the interest rate column does not obey normal distribution.

Table 2: Statistics for interest rate.

Statistics	**Value(%)**
Average	13.25
Minimum	5.32
25th percentile	9.99
50th percentile	12.99
75th percentile	16.20
Maximum	28.99

As for the fourth feature, loan amount, I applied a box plot to visualize its distribution. A boxplot is constructed of two parts, a box and a set of whiskers shown in Figure 2. The lowest point is the minimum of the dataset, which is 0 in this case. Similarly, the highest point, 35,000 euro, is the maximum. The box is drawn from 25th percentile (about 8,000 euro) to 75th percentile (about 20,000 euro) with a horizontal line drawn in the middle to denote the median (about 13,000 euro). Similar to the last feature, although it looks like it follows a normal distribution, the hypothesis test proves that it does not obey a normal distribution.

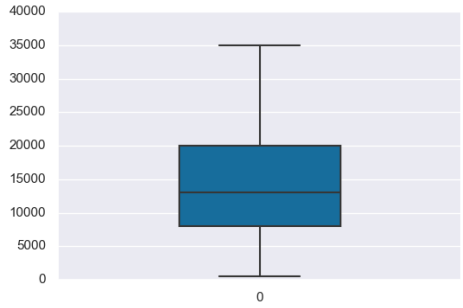

Figure 2: Box plot for loan amount.

The next feature is the loan term in months. There are only 2 possible values, which are 36 months and 60 months. Therefore, the pie chart is a suitable choice to show basic distribution of this feature. Loan term with 36 months is predominant, taking up 70% of overall loan records.

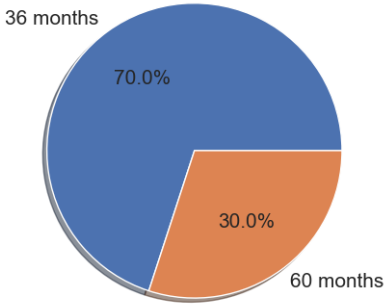

Figure 3: Pie chart for loan term in months.

After introducing the 5 numerical features, we now take a look at the first categorical feature, loan condition. It is the ultimate evaluation for the loan application, consisting of 2 categories, good loan or bad loan. In addition, it is the response/dependable variable for the classification model in section 5. According to Figure 4, 92.4% of loan application belong to the "good loan" category.

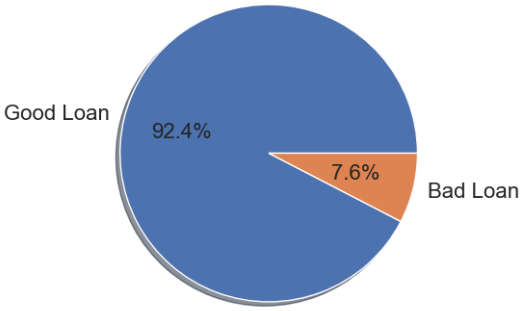

Figure 4: Pie chart for loan condition.

The second categorical feature, which is home ownership, consists of 6 categories, as shown in the bar plot below. A straightforward outcome is that "Rent", "Own" and "Mortgage" are 3 major categories of the home ownership. And we can exclude the other 3 types in the next step of analysis, due to their low instances.

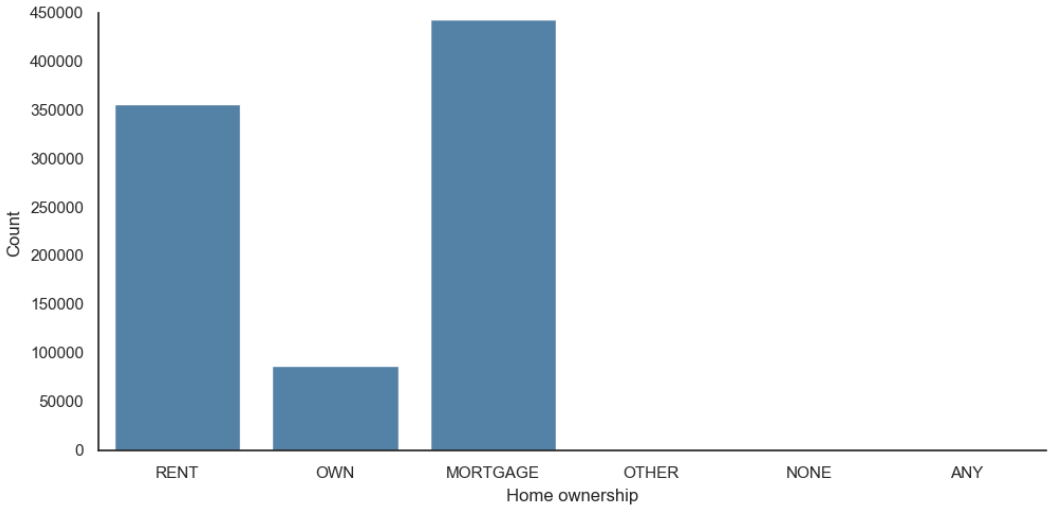

Figure 5: Histogram for home ownership.

The next feature, loan purpose, contains 14 total categories.To better demonstrate each type of loan purpose, the following bar chart is utilized. According to Figure 6, debt consolidation and credit card are the 2 main purposes for applying for a loan.

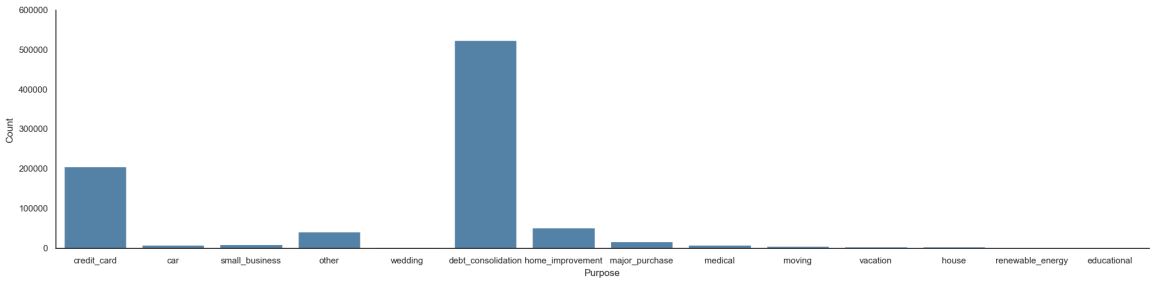

Figure 6: Histogram for loan purpose.

The last feature is the interest payments, which has only 2 categories, "low" and "high". So I used pie chart to display its basic distribution. Based on Figure 7, the percentage of loan application with low interest payment and those with high interest payment are both close to 50%, which implies this feature may follow an uniform distribution.

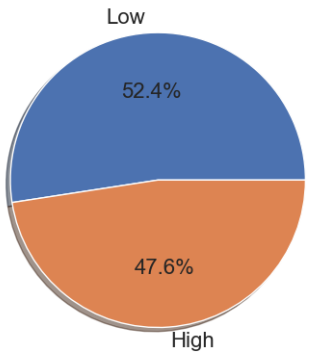

Figure 7: Pie chart for purpose.

3.2 Correlation

After the analyzing the descriptive statistics of 9 features respectively, I moved on to research their correlation. In statistics, a scatterplot is a popular way to show the correlation of different features vividly. From the following pairwise correlation plot, we can have a basic idea of the correlation of 9 features. It is noteworthy that plots in the diagonal from upper left to lower right are the density plots of each feature, while the ones not in the diagonal are the scatterplot of 2 various features. Intuitively, these 9 features do not have significant correlation.

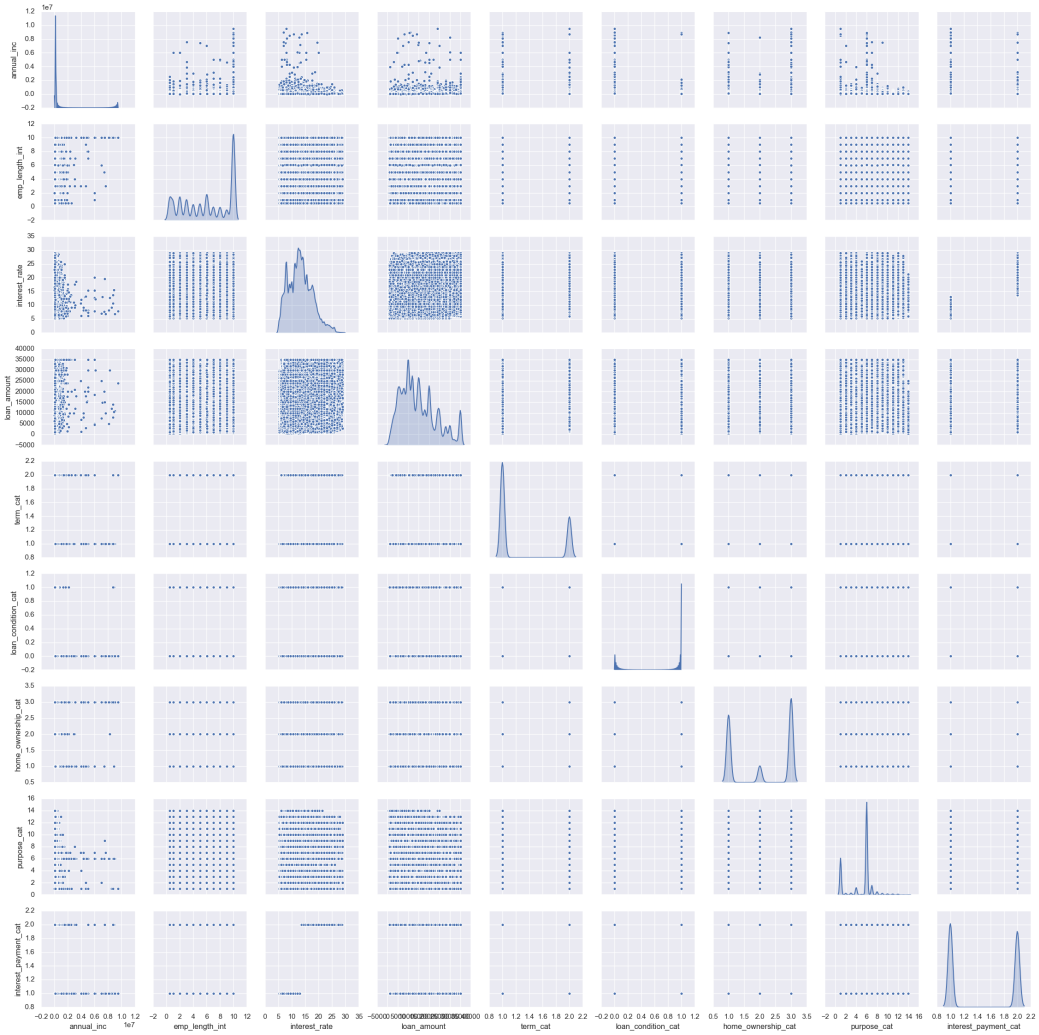

Figure 8: Scatterplot of pairwise features.

Therefore, I additionally constructed a heat map, so as to gain a more straightforward view of the correlation of 9 features. The following table is a 9 x 9 table, The value of each cell, represents a correlation of corresponding 2 features. The higher the value, the closer the cell color is to red. According to Figure 9, except the diagonal values, there are only 2 relevant ly large values in this table, which are 0.8023. By the symmetry of this table, this value implies that the interest rate and interest payment have a high correlation. This finding is reasonable, since the interest payment is supposed to be proportional to the interest rate.

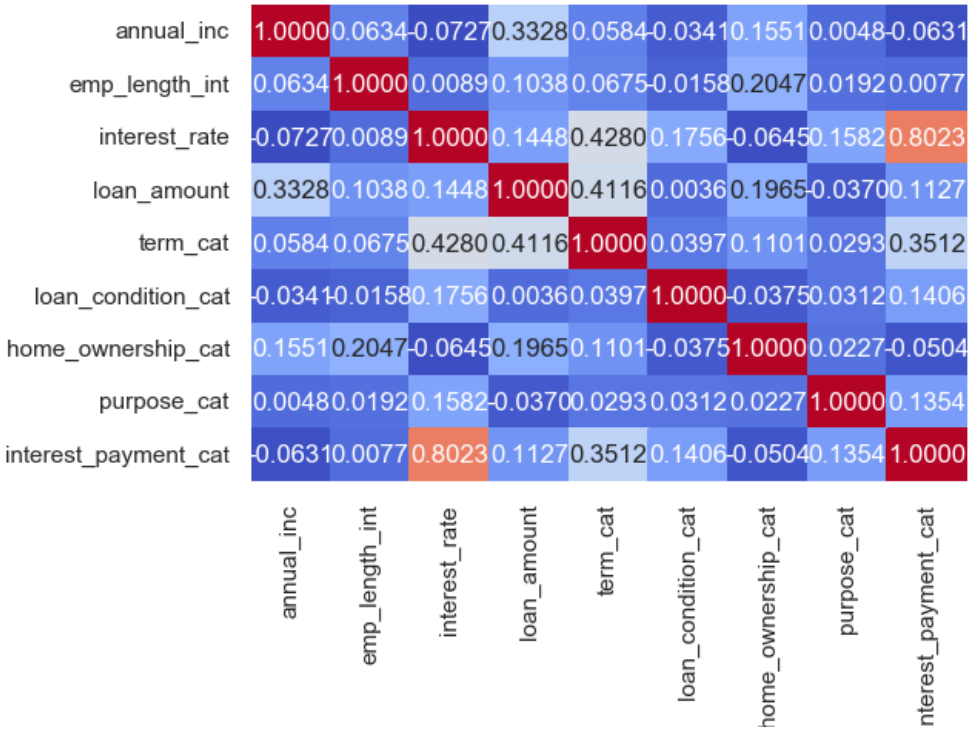

Figure 9: Correlation heat map.

4 Visualization

From the scatterplot and heat map in section 3.2, it seems that there are no significant correlation between these 9 features. In this section, I will take an in-depth analysis between some specific pair of features, in order to extract more business insights.

4.1 Income and home ownership

It is generally believed that clients who own a house have higher income than those who bought a house by mortgage. However, in this case, an interesting discovery from Figure 10 is that the average annual income of clients who bought a house by mortgage actually have a higher salary than those who already own a house, by a relevantly large margin.

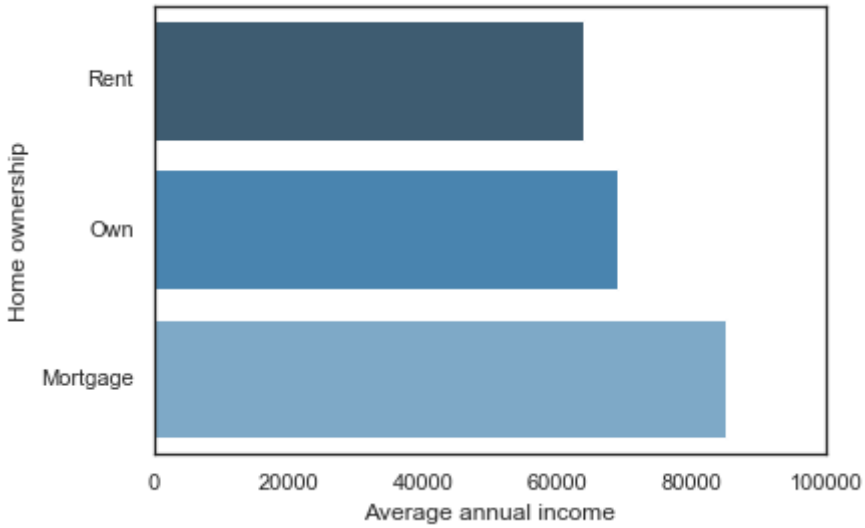

Figure 10: Income and Home ownership.

5 Model

6 Conclusions