caption

SharpestMinds Project

Analysis of a Loan Dataset in an Irish Bank

Ben (Zhibin) Liu

Data Analyst

Supervised by

Mikhail Sidyakov

Data Scientist

June 9, 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 how a bank classify the loan application and decide whether to grant a loan or not, utilizing a mix of statistical analysis, visualization and machine learning models.

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 assessment of a loan application.

2 Methodology

The purpose of this project is to drive insights from the loan dataset of an Irish bank. The complete loan dataset consists of 887,379 rows and 30 columns. To start with, I took 10 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 show the pairwise correlation of all 10 features. Then, I chose some specific pairs of features to conduct further detailed analysis. Moreover, principal component analysis (PCA) is applied to reduce the dimensions and generate an intuitive visualization.

Based on the visualization, I chose relevant machine learning models, k-NN and Decision Tree, to demonstrate how some features influence the evaluation of a loan application by the bank quantitatively. These models have a satisfactory

predictive validity, so it can help us understand the key factors in classifying loan applications.

3 Exploratory data analysis (EDA)

As mentioned in the last section, the target dataset contains nearly 1 million records and 10 major features, including 5 numeric features and 5 categorical features. To start with, I conducted descriptive analysis about these 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.

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.

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

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

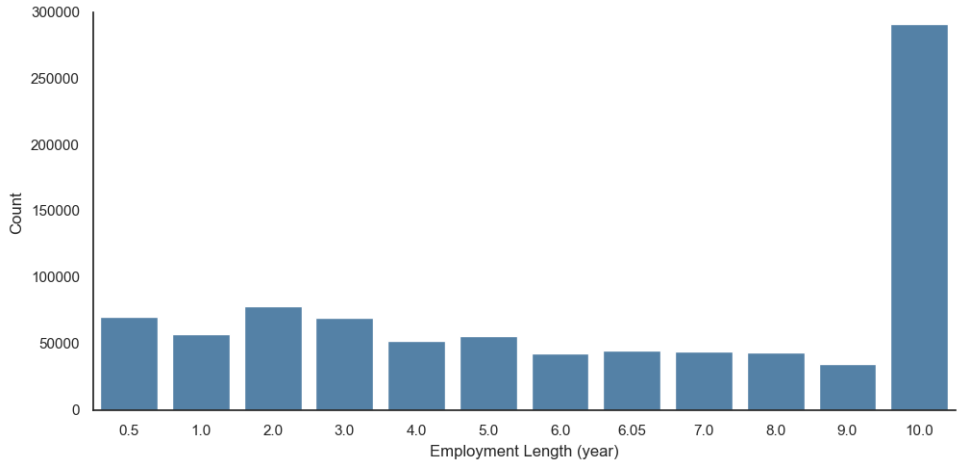

Figure 1: Histogram for employment length in years.

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.

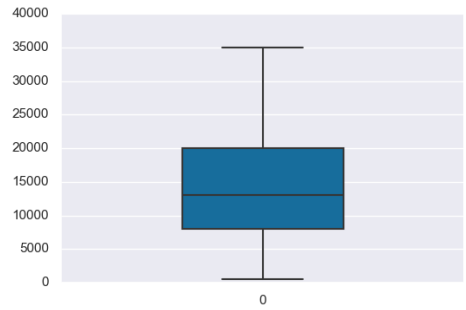

Figure 2: Box plot for loan amount.

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. From Figure 3, loan term with 36 months is predominant, taking up 70% of overall loan records.

After introducing the 5 numerical features, we now take a look at the first

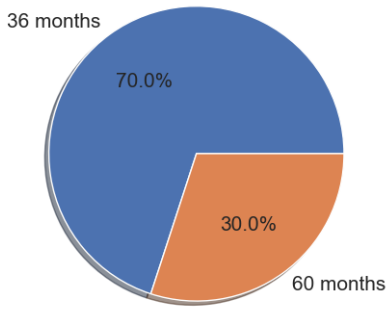

Figure 3: Pie chart for loan term in months.

categorical feature, which is grade. It is the ultimate evaluation for the loan application by the bank, consisting of a total of 6 categories, A, B, C, D, E, F and G. As expected, A represents a secure loan and G represents a risky loan. In addition, the grade is the response variable for the classification model in section 5. Based on Figure 4, a majority of loan applications are graded as "B" or "C".

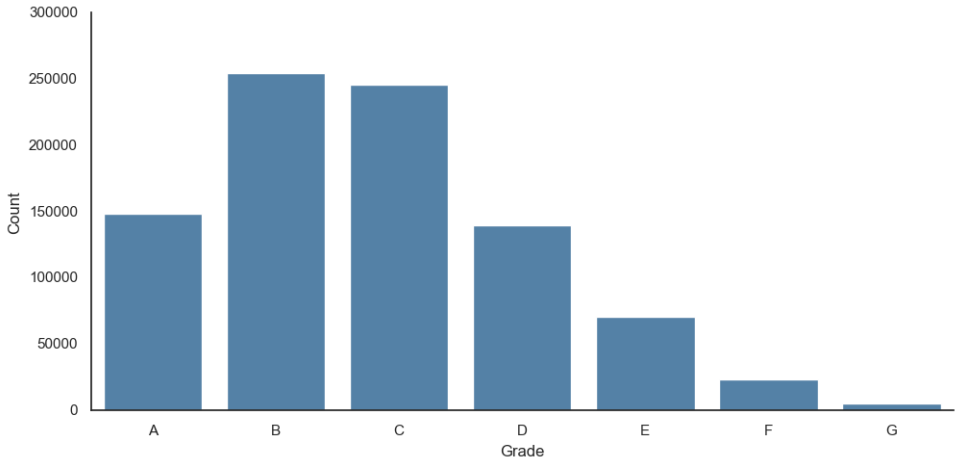

Figure 4: Histogram for grade.

The second categorical feature, loan condition, is the another type of evaluation for the loan application. But it only consists of 2 categories, good or bad loan. According to Figure 5, 92.4% of loan application belong to the "good loan".

The next categorical feature, which is home ownership, consisting of 6 categories, as shown in the bar plot below. A straightforward outcome is that "Rent",

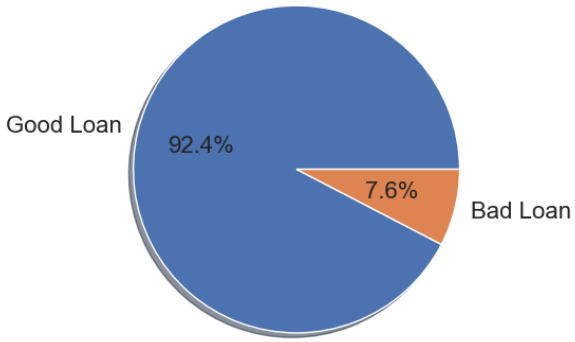

Figure 5: Pie chart for loan condition.

"Own" and "Mortgage" are 3 main categories of the home ownership. We can try excluding the other 3 types in the next step of analysis, due to their low instances.

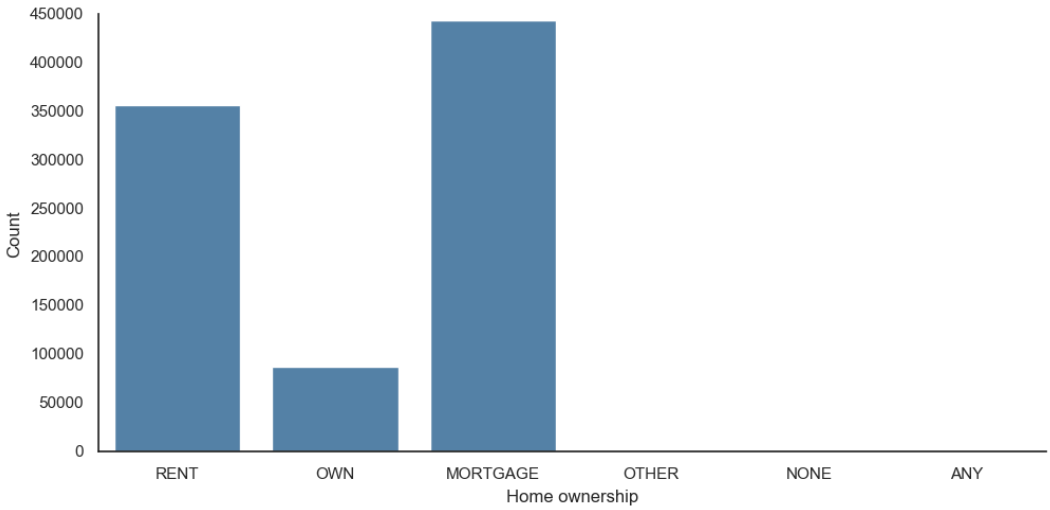

Figure 6: Histogram for home ownership.

The ninth feature, loan purpose, contains 14 total categories. To better demonstrate each type of loan purpose, the following bar chart is utilized. According to Figure 7, debt consolidation and credit card are the 2 main purposes.

The last feature is the interest payments, which has only 2 categories, "low" and "high". So a pie chart is good enough to display the basic distribution. Based on Figure 8, the percentage of loan application with low interest payment and

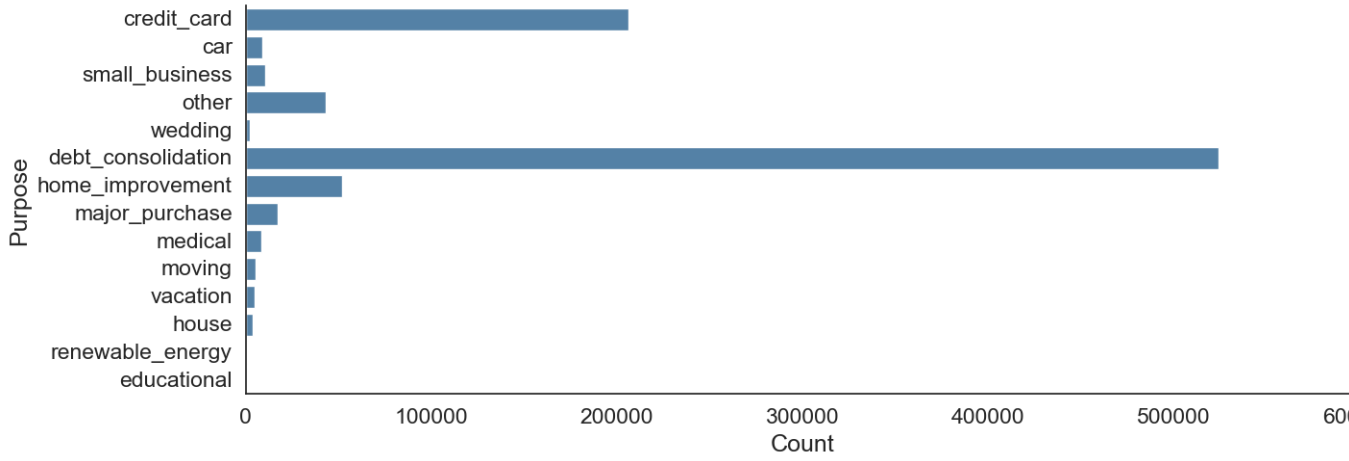

Figure 7: Histogram for loan purpose.

those with high interest payment are both close to 50%, which implies this feature may follow an uniform distribution.

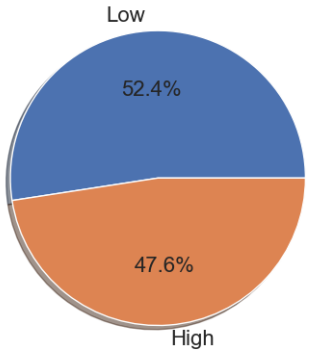

Figure 8: Pie chart for purpose.

3.2 Correlation

After analyzing the descriptive statistics of 10 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, except for the responses variable "grade". 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.

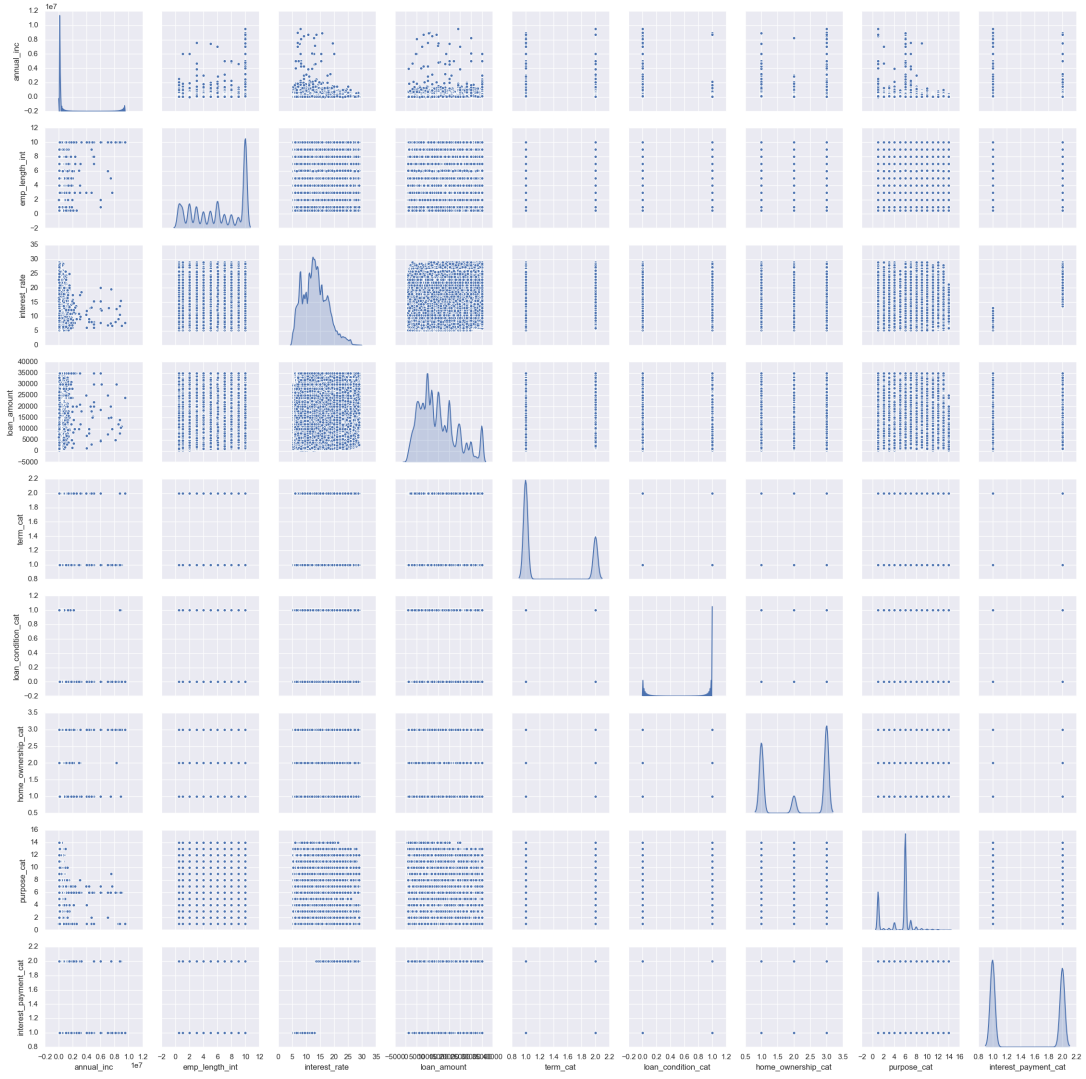

Figure 9: Scatterplot of pairwise features.

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 10, except the diagonal values, there are only 2 relevantly large values in this table, which are both 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.

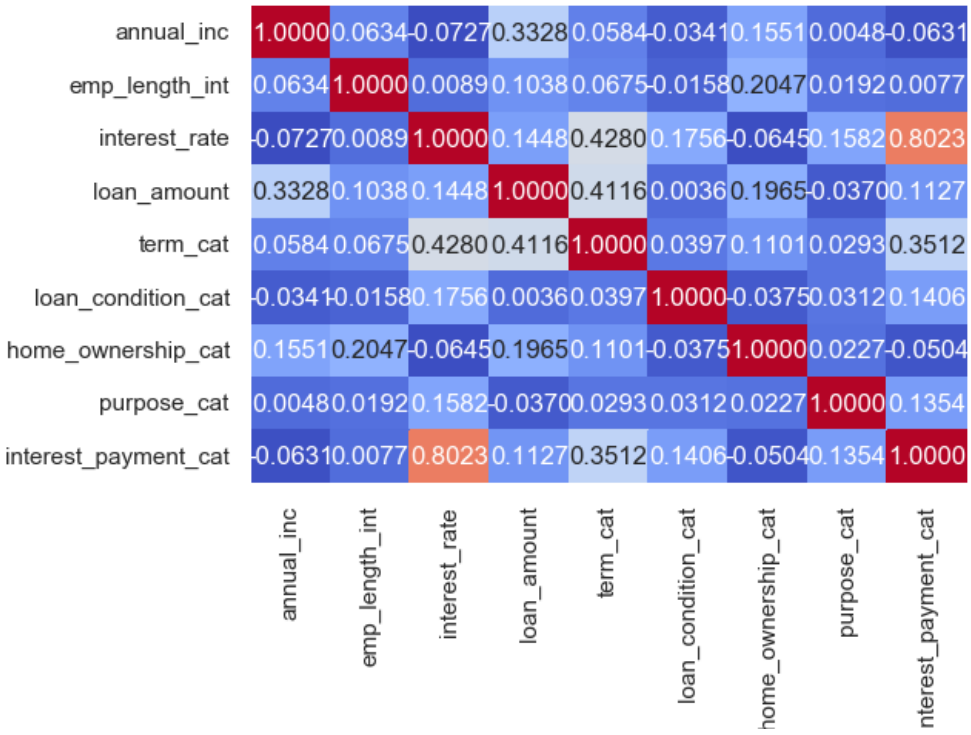

Figure 10: Correlation heat map.

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 11 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.

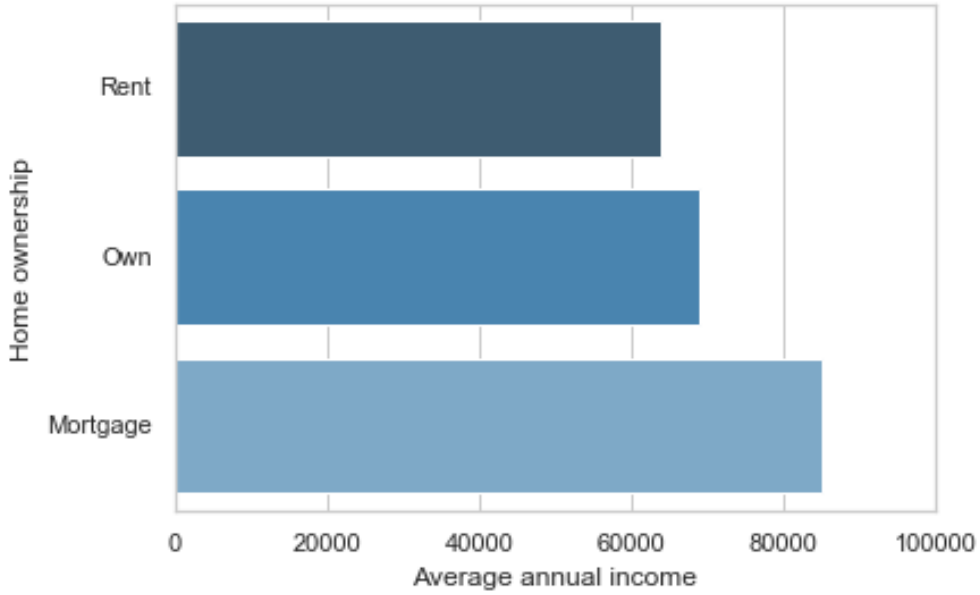

Figure 11: Income and home ownership.

4.2 Loan condition and interest rate/income

Let's move on to the relationship between loan condition and other features. From the left bar plot in Figure 12, The interest rate of "Good Loan" category is averaging about 3% larger than that of "Bad Loan". In fact, this is reasonable because the higher the interest rate, the more likely the clients fail to pay the loan. In addition, the other chart demonstrates that the average annual income of clients who belonged to "Good Loan" category is higher than those belonged to "Bad Loan" by nearly 10,000 euro, which also accords with common sense.

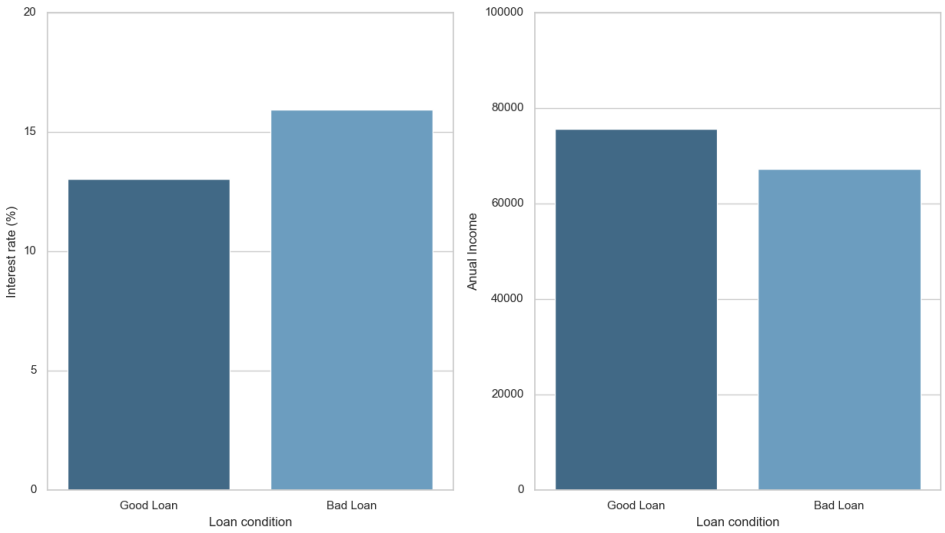

Figure 12: Loan condition and interest rate/income.

4.3 Purpose and income/loan amount

In this section, I focus on the relationship between loan purpose and other features. Based on Figure 13, the loan amount of debt consolidation and small business are averaging the highest among all the categories. This result is attributed to the fact the loan amount for commercial purposes is often higher than that for personal purposes.

Then take a look at the relationship between purpose and income. According to Figure 14, among all loan purposes, clients who apply a loan for "home improvement" and "small business" have higher income than others. This finding makes sense, because people who own a small business or require home development is more likely to be wealthy. On the contrary, clients who apply for a loan due to education have the lowest average income, which is also reasonable since educational cost is the basic expenditure for a family.

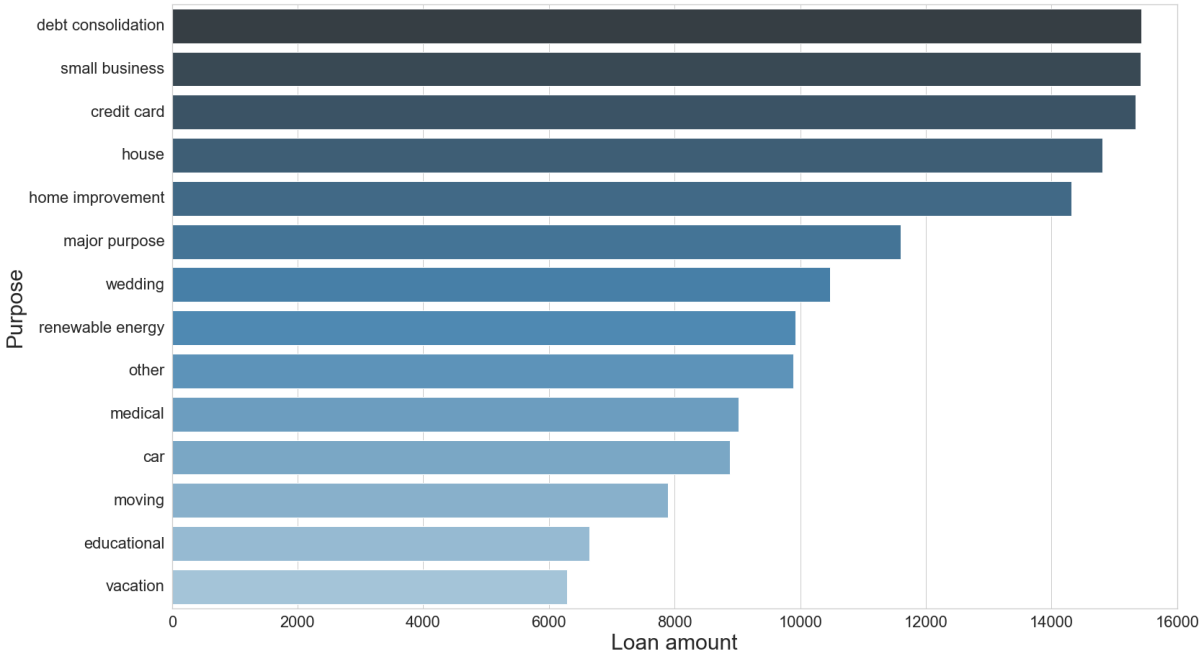

Figure 13: Loan amount and purpose.

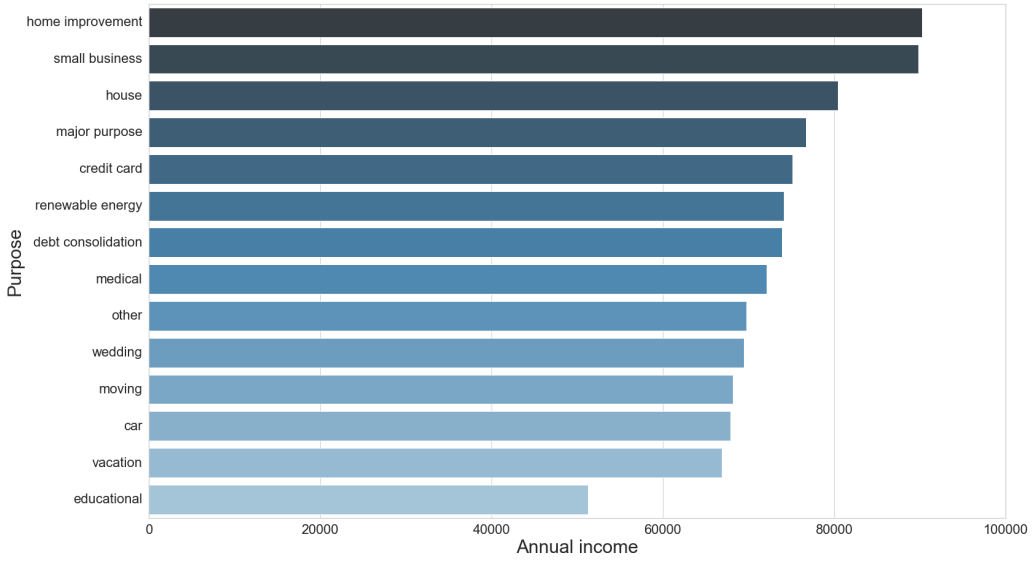

Figure 14: Income and purpose.

4.4 Employment length and loan amount

Figure 15 shows the relationship between employment length and loan amount. It is obvious that there is a positive correlation between these 2 features. The result indicates that generally this Irish bank tends to grant a loan with larger amount to those clients with longer working years. A possible reason is that a client with long employment length is generally believed to be more trustworthy than someone with short employment length.

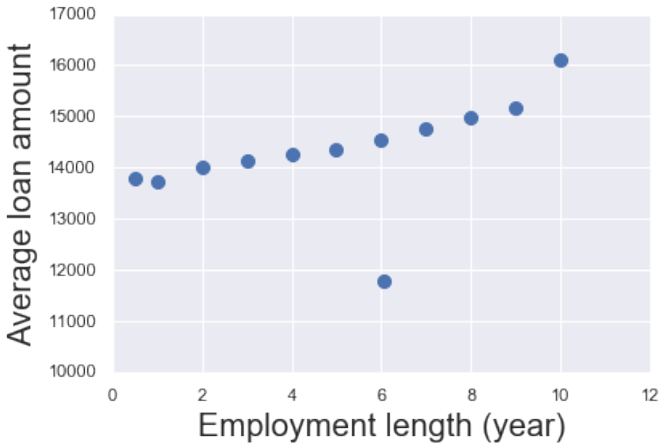

Figure 15: Employment length and loan amount.

4.5 Feature extraction by PCA

As mentioned in section 3.1, both "grade" and "loan condition" can represent the evaluation of loan application by the bank, where "grade" consists of 6 categories and there are only 2 types in "loan condition". Since a categorical variable with more types is believed to contain more information and it is easier to train with models, "grade" is a better option for the dependent variable in section 5.

Before fitting the dataset with some machine learning models, it is necessary to analyze the relationship with "grade" and other 9 independent features. A visualization may be the most straightforward approach, from which I may find some insights. After that, specific algorithms can be chosen to analyse the dataset based on the pattern from the plot.

However, in addition to 'grade', there are still 9 other features. The goal is to summarize 9 features to 2 features, so that they can be shown in a 2 dimension intuitive graph. Therefore, principal component analysis (PCA) is utilized. Technical explanation will not be covered here, basically 2 new variables can explain a high percentage of variation of the original 9 features. Furthermore, by assigning 7 types of grade with 7 various colors, the following plot is generated.

According to Figure 16, although there are overlaps, it is noteworthy that the color is becoming lighter from the left to the right in the plot. Therefore, it is feasible to classify the dataset with typical classification models, which will be demonstrated in more details in the next section. Moreover, I experimented to clean the dataset by excluding some outliers discussed in the preliminary analysis section. However, the visualization is not as satisfactory as Figure 16. A possible reason is that, the so called outliers in section 3 indeed carry out a significant amount of variation explainability of the raw datasets. In other words, it is more suitable to apply classification models on the original dataset than the clean dataset.

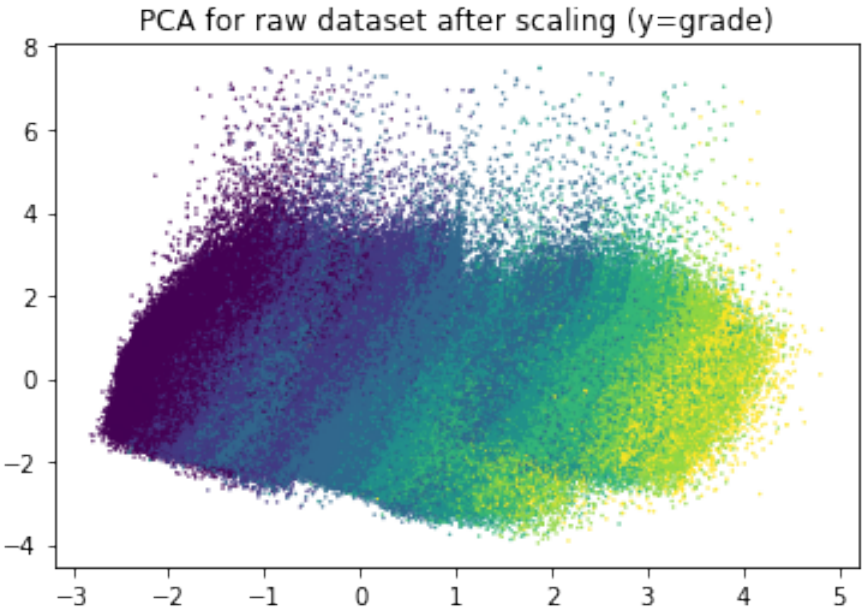

Figure 16: PCA for 9 features.

5 Model

According to the colorful Figure 16 in last section, the "grade" feature of the loan dataset is feasible to be classified to several categories. Since the total categories of grade is known in advance, which is "A", "B",...,"G", supervised learning algorithms should be applied to analyse the dataset. Before in-depth analysis with various models, it is necessary to demonstrate the experiment settings.

First of all, similar to the former section, the raw dataset should be scaled so that the values of each feature are between the same range. The reason is that if we fit the dataset to the model directly, a feature like "income" with a maximum of 200,000 euros, will have a more significant influence on the response feature "grade", compared to a feature like "interest rate". Therefore, I scaled the dataset by 3 different approaches, Min-Max Scaling, Standard Scaling and Robust Scaling. Simply speaking, most of the updated data by these 3 approaches will be between -1 and 1.

In addition, to evaluate the performance of the model, I fit the model that I generalize to the new data. This is usually done by splitting the dataset into 2 parts, training set and test set, while the former one is utilized to build the machine learning model and the latter one is used to assess how well the model performs. Therefore, the metric in this experiment is the prediction accuracy of test set.

Last but not least, to attain the highest prediction accuracy, I conducted multiple experiments in terms of different parameters of the specific algorithm. By comparing the performance of all the scenarios, the optimal model with the highest accuracy will be selected.

5.1 k-NN

Among all the classification methods in supervised learning, k-Nearest Neighbors (k-NN), is arguably the simplest machine learning algorithm. In k-NN classification, a record is classified by a plurality vote of its neighbors, with the record being assigned to the class most common among its k nearest neighbors. For example, there are 2 known categories in Figure 17, blue square and red triangle,

while the green circle is a unknown record. If k=3 is assigned in k-nn, the green circle will be classified as red triangle category, since there are 2 red triangles out of 3 nearest neighbors. However, if k=5 is assigned in k-nn, the green circle will then be classified as blue square category, because there are 3 blue squares in 5 nearest neighbors.

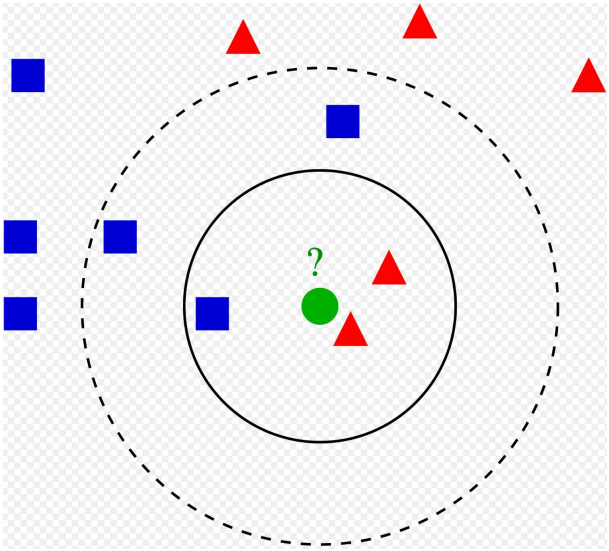

Figure 17: k-NN example.

5.1.1 Scaling methods

Before applying k-NN to the original loan dataset, I scaled the dataset by 3 statistical methods, Min-Max Scaling, Standard Scaling and Robust Scaling. I remained the default parameters for k-NN, implemented the algorithm and collected the result in Table 3. It is straightforward that the Min-Max scaling is the best methods to scale this loan dataset before applying k-NN for classification.

Table 3: k-NN prediction accuracy by various scaling methods.

Scaling methods	**Min-Max**	**Standard**	**Robust**
Accuracy(%)	85.0	81.1	80.3

5.1.2 PCA

It is ideal to utilize the raw dataset for all parts of analysis, however, since this is a 880,000 $\times$ 9 dataset, the computation cost is heavy. In fact, it takes about 20 minutes to compute the results in Table 3. To explore how the parameters of k-NN affects its prediction accuracy efficiently, modification is necessary. Similar to section 4.5, PCA will be utilized to decrease the number of features in the dataset. According to following table, it is concluded that there is a trade-off between accuracy and efficiency. When N=9, which means it is the raw dataset, the prediction accruacy is the highest, but the process is computationally heavy. On the contrary, when N=2, the speed is fast, while the accuracy is unsatisfactory, because much information is lost by the dimension reduction. Choosing N=5 as PCA components makes most sense, since it increases the efficiency by 50 times at the cost of only 1.2% of prediction accuracy, compared to when N=9.

Table 4: Accuracy and time of k-NN by various PCA components.

Number of PCA components	**N=9**	**N=5**	**N=2**
Accuracy(%)	85.0	83.8	74.5
Time(seconds)	399.2	7.5	4.7

5.1.3 Test size

The test size is the percentage of test data in the whole dataset. Since normally the training set is larger than the test set, so the range of test size is between 0 and 0.5. Intuitively, more data is trained by the k-NN, the robuster the model will be, which leads to a high prediction accuracy for the test set. Therefore, the raw dataset will be split by 90% and 10% as training set and test set in the following sections.

5.1.4 Number of k-NN neighbors

Now let's move on to the parameters of k-NN algorithm. Firstly, I focus on the number of k-NN neighbors(k), which is regarded as the most significant parameter

in this machine learning algorithm. Using few neighbors would corresponds to a complex model, while using many neighbors leads to a simple model. In an extreme case, if you consider the number of neighbors are exactly the number of all data points in the training set, the prediction accuracy for every test point will be the same.

To find out the optimal k for this situation, I created a loop function to choose from 1 to 20, as well as calculated the prediction accuracy for each number of neighbors. From the following plot, it is obvious that the prediction accuracy reaches the peak when k is 14.

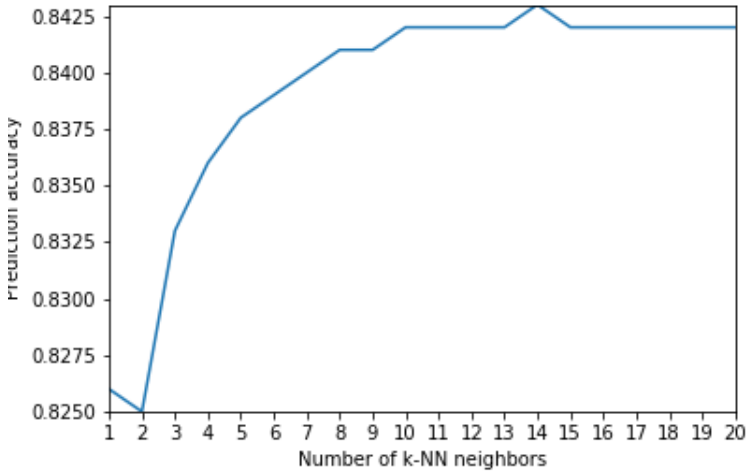

Figure 18: Prediction accuracy for various number of neighbors.

5.1.5 Weight of k-NN neighbors

As for the weight of k-NN neighbors, it contains 2 main options, "uniform" and "distance". In short, the former one means all neighbors are equally important, while the latter one implies that a closer neighbor plays a more important role. It seems like the latter option makes more sense, and the results from Table 5 accords to the intuitive judgment.

Table 5: k-NN prediction accuracy by various weight.

Weight	**Distance**	**Uniform**
Accuracy(%)	83.8	81.6

5.1.6 Type of distance

The last parameter I researched is the type of distance. There are 2 major measurement for distance, euclidean distance and manhattan distance. According to Table 6, k-NN with the manhattan distance outperforms the one with euclidean distance by a small margin.

Table 6: k-NN prediction accuracy by various type of distance.

Type of distance	**manhattan**	**euclidean**
Accuracy(%)	84.6	84.4

5.1.7 Summary

In conclusion, the prediction accuracy of k-NN attains the highest in this experiment, when the scaling method is Min-Max, the test size is 0.1, the weight is "distance", the number of neighbors is 14 and the type of distance is manhattan. It is noteworthy that although I conducted PCA(n=5) in some of the above sections, the main concern is the prediction accuracy, therefore, in the last step I will apply k-NN to the raw dataset without conducting PCA. As a result, the optimal prediction accuracy of the feature "grade" by k-NN is 85.8%, which is satisfactory for a real-life case.

5.2 Decision Tree

Another popular classification model is the Decision Tree model. It is a flowchart-like structure in which each internal node represents a test on a feature. Each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to the bottom leaves represent overall classification rules. For instancebased on Figure 20, a person who is under 30 years of age and does not eat many pizzas, or who is more than 30 and do exercise in the morning, will be classified as a fit person.

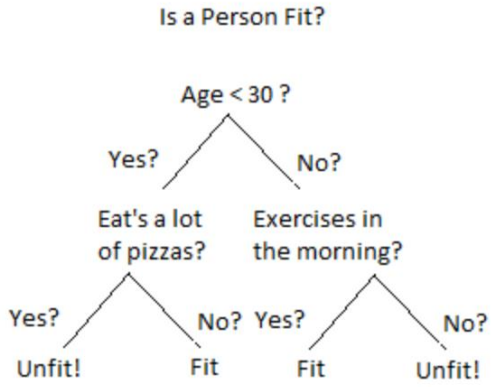

Figure 19: Decision tree example.

5.2.1 Maximum depth

Unlike k-NN, Decision Tree algorithm is invariant to the scaling of the data. Since each feature is processed separately, preprocessing steps like Min-Max scaling are not necessary. In addition, implementing Decision Tree on the raw dataset takes only few seconds, therefore, there is no need to conduct PCA.

After splitting the original loan dataset to training set and test set, I applied the Decision Tree model with default parameters to the training set, fitted the test set with the model, as well as computed the prediction rate. Surprisingly, the accuracy is already 94.4%. Although the performance is satisfactory, I will tune some parameters to test if higher accuracy can be obtained.

The first and foremost parameter to be focused on is the maximum depth. When applying default setting, the nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples, which may lead to overfitting. Intuitively, maximum depth means the total layers of the Decision Tree, for example, maximum depth is 2 in Figure 20. By plotting the prediction accuracy over various values of maximum depth in the following figure, it is straightforward that when the maximum depth is larger than 13, the prediction rate achieves the highest and become stable, which is about 95.7%.

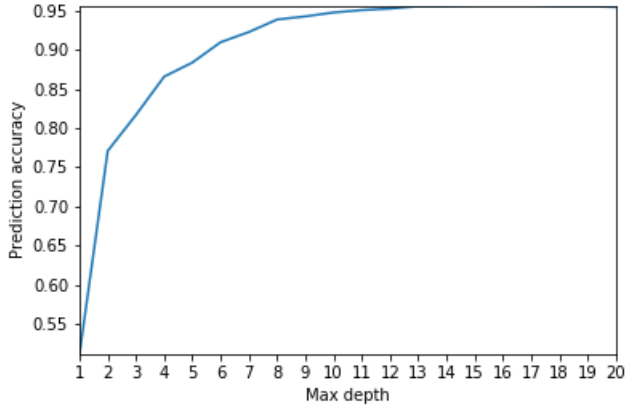

Figure 20: Prediction accuracy for various maximum depth.

5.2.2 Criterion

The next parameter to be analyzed is the criterion. There are 2 main criterion, “gini” and "entropy", which stand for Gini impurity and information gain respectively. From Table 7, Decision Tree with criterion "gini" outperforms that with "entropy" by a small margin.

Table 7: Decision Tree prediction accuracy by various criterion.

Criterion	**gini**	**entropy**
Accuracy(%)	95.7	95.4

5.2.3 Minimum samples split

Then we move on to the "minimum samples split" parameter, which is the minimum number of samples required to split an internal node. This parameter is related to the maximum depth. Since if the minimum number of samples required to split a node is small, more decision paths will be created, which leads to a large maximum depth. As shown in the below graph, when the minimum samples split is too small, the prediction accuracy is relevantly low, due to the overfitting issue. But in general, the accracy is stable, ranging from 94.4% to 95% when this parameter is under 20.

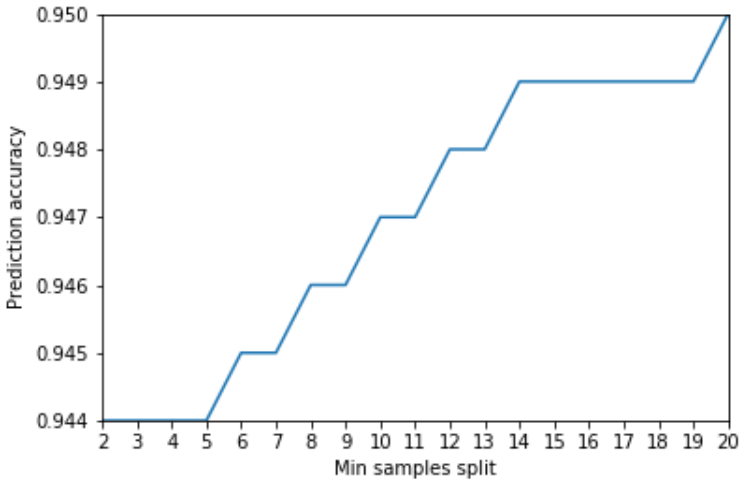

Figure 21: Prediction accuracy for various number of neighbors.

In addition, I have experimented Decision Tree model by tuning other parameters, such as "splitter", "min samples leaf", "max features" and so on. However, these scenarios are outperformed by the default setting.

5.2.4 Ensemble (Random Forest)

Ensembles are methods that combine multiple machine learning models to create more powerful models. One typical example is the Random Forest. A random forest is basically a collection of decision trees, where each tree is slightly different from the others. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction. I set parameters of both models as default values, then computed the prediction accuracy in the following table. In this case, the prediction accuracy by these 2 models have little difference.

Table 8: Prediction accuracy by different models.

Model	**Decision Tree**	**Random Forest**
Accuracy(%)	94.9	94.3

5.2.5 Visualization

One of the advantages of the Decision Tree model is that it is easy to interpret and visualize. For this loan dataset, the prediction accuracy achieves 95.5% when the max depth is 15. The visualization of a decision tree with 15 layers of nodes is a little bit messy. To clearly demonstrate how each node is split by a specific feature, I visualized the process of a decision mdoel when max depth is 4. Based on the following decision tree graph, the first selection criteria is whether the interest rate of a record is less than 12.2. If an application satisfies this condition, it will be classified to left node in the second layer, otherwise it belongs to the right node. The 8 nodes at the bottom are called the leaves or the decisions, which show the ultimate classification. For example, the left bottom leaf implies that if a records satisfies the condition that the interest rate is less than 12.2, 8.1 and 6.0, it probably belongs to grade A, because 11510 of 11617 records which satisfy these conditions turn out to be a grade A application.

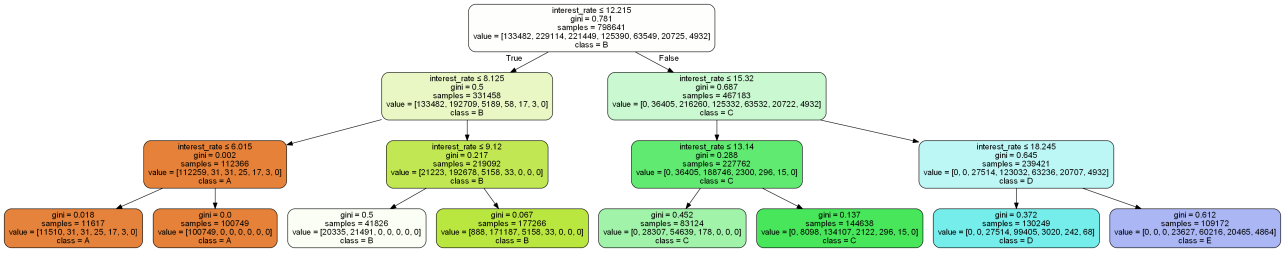

Figure 22: A decision tree when max depth is 4

5.2.6 Summary

In this section, another machine learning algorithm is utilized to predict the classification of feature "grade" in the loan dataset. Since Decision Tree is invariant of scaling, some cleansing steps can be skipped. Setting the max depth as 15 and others as default values, the prediction accuracy on the test set is around 95.7%, which is generally regarded as an ideal outcome.

6 Conclusions

Loan application is almost inevitable in the group-up world. How the management team of a bank classify the loan applications and decide whether to grant a loan or not, is preoccupying the minds of vast amount of applicants around the world. In this report, we first conducted exploratory data analysis on the loan dataset of an Irish bank, including descriptive statistics of each feature and the correlation between all the features. Moreover, further visualization about specific pairs is generated to drive more business insights. After deciding to set the feature "grade" as the responses variable, 2 machine learning classification models, k-NN and Decision Tree are applied to the training set of loan dataset to predict the class of grade in the test set. By tuning different parameters, the prediction accuracy increased from 81.8% to 95.7%, which is satisfactory in a real-world project.

In the future investigation, we will attempt to study how the feature "loan condition" (good or bad) is influenced by other variables. Additionally, more machine learning models with different conditions would be implemented.