Predicting Credit Risk

Caleb Puckett

cpuckett@bellarmine.edu

1/16/2024

This project will utilize predictive analysis to determine credit risk of individuals based on many factors, some of which include gender, family size, income, education level, and house and vehicle possession. Although a credit score communicates the likelihood of defaulting on a loan, this project will attempt to analyze factors not used as determinants of this score. Credit scores are determined by an individual's history of credit, on-time payment history, number of credit lines, rates of credit use, and recent requests for additional credit. These factors have been chosen for good reason, but in this project, an attempt will be made to link them with more general identifiers of one's situation. The data was gathered online and will be primarily adjusted in Python. Data cleaning and manipulation and exploratory data analysis will be performed in Python, followed by the use of machine learning for predictive purposes.

The primary focus of this project is to consider factors affecting credit risk. More specifically, this will hopefully predict the likelihood of default on a credit payment with a high level of accuracy. Some of the variables which will be used for predictive purposes are gender, family size, income, education level, and house and vehicle possession. It is intuitively expected that higher income and higher education level would be correlated with lower risk of credit default. Possession of a house or vehicle would also seem to be correlated with lower risk of credit default, especially because in many instances, the process by which these things are obtained is responsible use of loans and credit. The other variables are not as clear in their effect, so it will be interesting to evaluate the results after the necessary analysis.

Credit scores can already accurately predict the likelihood of default on a credit payment. These scores are not only used to assist lenders in their decisions of lending and giving lines of credit, but also by employers to gauge a potential employee's level of financial and personal trustworthiness. Credit scores fall short, however, because their only predictive power is determined by previous demonstration of responsible use of borrowed funds. This creates difficulty when an individual wants to demonstrate their trustworthiness as a borrower but has no history of borrowing funds. This project will attempt to link this creditworthiness to factors that do not require previous experience borrowing funds or responsibly using lines of credit. Credit risk is the basis of multiple billion dollar industries and is thus important to understand and predict as thoroughly as possible.

Decision Tree will be one of the initial machine learning models used. Although it does not go as in-depth as some other models, it is well suited to the results of these data. Because these results will seek to predict one of two options, or whether a borrower will default on their loan or not, Decision Tree will likely draw meaningful conclusions and fit the data in this model well. This will help us identify the importance of all involved variables, showing us the degree to which they appear to affect credit risk. It is worth noting that Decision Tree will not be as accurate as some other models because of its simplicity.

Another machine learning model utilized in this analysis will be Random Forest. Because of its high level of accuracy in comparison to Decision Tree modeling, Random Forest will be very useful in its ability to draw thoroughly backed conclusions from these data. Where a Decision Tree model is a combination of decisions, a Random Forest model is a combination of a multitude of Decision Trees. Because of this, it will be much more accurate, although it will naturally also take much longer to run. Similar to Decision Tree, Random Forest will tell us the level of significance of the explanatory variables with respect to predicting credit risk. It will be interesting to see the ways in which the results of the Decision Tree model differ from those of the Random Forest model.

Kernel Support Vector Machine (SVM) will also be one of the machine learning models used in this analysis. This will be a good fit because it has the ability to show more complex relationships between variables in cases where these relationships are not linear. Kernel SVM is more complex than SVM as the dataset grows larger, so it is worth being wary of the potential difficulties that could come with this complexity. As opposed to traditional linear regression, Kernel SVM will add extra dimensions to better understand the data. Because it is not limited to only linear relationships, Kernel SVM will provide much more robust results than linear regression.

This project will likely involve some data manipulation in Excel, although the vast majority of work will be done in Python. There may also be some visualization done in Tableau or PowerBI depending on the visuals that would best fit the data. The data will need to be cleaned and adjusted based on our goals, outliers will need to be identified, and exploratory data analysis will be used to better understand the data. After this has all been done, at least three different machine learning functions will be performed. Many packages will be imported into Python to efficiently complete these goals. Some of these include Pandas, Matplotlib, NumPy, Caret, and Skicitlearn. Pandas will be used for data manipulation and analysis and will be utilized consistently throughout the project. Matploblib and NumPy are most useful for visualization of data and will be primarily used during exploratory data analysis. Scikitlearn and Caret will be used primarily in the machine learning portion of this project. All of these, and more, will be used in Python to complete robust analysis of the data.

Among its many other effects, credit risk evaluation affects the credit card industry, banking industry, and the insurance industry very heavily. The goal of this project is to identify variables that correlate with risk of credit default, as well as the degree of correlation. This will primarily be done using machine learning models, or, more specifically, the Decision Tree model, the Random Forest model, and the Kernel Support Vector Machine model. The variables considered here are not factored into credit scores, so the discovery of high degrees of correlation between them and credit default risk would be interesting to understand. It is likely that any variables which appear to be significant in this project will also be correlated with the current determinants of credit scores. This would not take away from these results, however, because many of them can exist in the absence of credit history.

References

Data obtained from:

https://www.kaggle.com/datasets/rikdifos/credit-card-approval
prediction?select=application_record.csv