Abstract

Banks make less money when a client defaults because all future interest payments that would have been made are lost. Therefore, banks have a vested interest in providing loans to clients who will not default. Our analysis provides banks a way of reliably predicting whether a client is a good or bad risk. We defined a good risk as an investment made by the bank that is likely to be profitable. A bad risk is defined as an investment made by the bank that will likely not be repaid.

Intro

Associating risk with clients that have applied for a loan has important implications for banks. Before issuing loans to applicants, a bank might want to check how likely it is for a client to default. There are five questions we sought to answer with this data set. First, can a client's new credit balance be used as a good predictor of what will be a client's highest credit balance? Second, do clients with similar features have identical class labels? Third, are there clusters within the data set? Fourth, can a client's risk be reliably determined using a probabilistic classifier? Fifth, can decision rules be used to reliably predict whether a client is a good or bad risk?

IDA

In this phase the raw data set was cleaned in order to facilitate analysis in subsequent phases.

Some of the columns in the raw data set were irrelevant for predicting whether a client was good or bad risk to default, so these columns were discarded.

For some clients, the raw data set contained numerous rows for a single client, whereas other clients only had one row. These rows corresponded to data for different dates. For example, there would be one row containing data for April and another row containing data for May. Both of these rows would correspond to a single client. In order to simplify our analysis, only one of these entries was kept.

For dealing with missing data, the mean or median value of the column was used. Some clients had high values for bank account balance relative to the majority of other clients. Substituting missing values with the column's mean value in this case would not be appropriate, because the mean value is heavily affected by the high balances of a few clients. For columns where this was the case, the missing value was substituted with the column's median value. The median value was also used for replacing missing values in columns with categorical data, such as type of product. For columns with low variance, the missing value was replaced with the column's mean value.

To easily find data belonging to a specific client we decided it would be appropriate to index the data set by client ID.

In order to reduce the dimensionality of the feature space, we aggregated the different type of overdue payments into a single column. This column represents the number of times a client made an overdue payment.

EDA

Our focus in this phase was to find insights that either would help us develop a better understanding of the data, or reveal patterns in the data. The latter helped us select appropriate models to use in the next phase.

The data set initially contained a significant imbalance in the class labels. Out of the 1,125 samples, 900 of these had good risk labels. To not bias our analysis, we made the amount of class labels equal. This left us with 450 samples in our data set (we'll come back to this point in the recommendations section).

Another one of our discoveries in this phase was that clients with high credit amounts owed were about four times more likely to be labeled as a bad risk. We made this discovery by finding the number of clients above an agreed upon threshold of $500,000 for each class label.

We also found an interesting pattern between two of the features. We used seaborn to produce two-dimensional scatter plots of all the features. We noticed that there was a linear relationship between columns 'new_balance' and 'highest_balance.' The 'new_balance' feature represents the current balance a client owes, whereas the 'highest_balance' feature represents the highest amount a client has owed. With this information we can predict what a client's highest balance owed is given their current balance.

Out of the 17 different kinds of credit products contained in the data set, we found 62% of clients had product 10. There was 90% of clients having products 5, 6, and 10. This suggests the kind of credit products is

To get a better understanding of the data contained in each of the columns, we determined the maximum and minimum values, mean, and standard deviation for each column. The max values helped us find outliers in the amount of credit owed. The min values helped us detect negative values, which was useful for detecting anomalies that can perhaps be attributed to mistakes in documenting the data. The mean sometimes gave us representative values of the column, whereas the standard deviation gave us insight into how dispersed the values were from the mean.

Modeling

In this phase we used machine learning to develop models that capture patterns found in the data. I can't overstate how useful performing PCA was for our analysis. After plotting the transformed data set on a two-dimensional plot, we noticed amazing clusters and subclusters. This gave us valuable insight into what models we were going to select.

K-Nearest Neighbors

KNN was an obvious choice because identical class labels tended to be near one another. We used L2 norm to measure similarity between the data points. Several k values were used to see which value would produce the best results. We found that a k value of 7 produced the best results. The accuracy was 89%, the true positive rate was 65%, and the false positive rate was 10%.

K-Means

K-means was used to associate data points with one of three clusters. Three clusters were chosen because it is apparent from the plot of the transformed data that there are three clusters. This is useful because knowing which cluster the data point is in, and where, gives an indication as to its class label. We found that 100% of the tested data points were associated with the correct cluster.

Decision Tree

We chose decision tree in order to have a model that learns decision rules based on the training data. It is a popular classification algorithm, and it can be applied a wide range of problems. It is also simple when compared to other classification algorithms. An interesting aspect of decision trees is that they consider possible traces in a path to assign a class label to a data point. Decision Trees are versatile in that they can be applied to problems in various contexts. The accuracy of our model was 84%, the true positive rate was 62%, and the false positive rate was 12%.

Linear Regression

While performing EDA, I noticed a linear relationship in the scatter plot of highest balance versus new balance. This was useful because it meant we could use information about a client's new credit balance to predict what their highest credit balance will be. We can use this information to associate a client with a class label, because those with high credit balances were found to be more likely to be classified as a bad risk. Our model did a great job of capturing the patterns found in the test data. It had a coefficient of determination value of .87. We noticed that the performance of the model improved significantly as we fed the model more training data. This suggests we can obtain much better predictions of what a client's highest balance will be given their new credit balance.

Logistic Regression

We chose logistic regression because it provides an intuitive probabilistic perspective on making classification predictions. To evaluate the model, we used accuracy, recall, precision, and false positive rate. The accuracy for our model was 86.7%, the recall was 60%, the precision was 95%, and the false positive rate was 3.7%. Of the three classification models we used, this one had the second-best accuracy. One interesting observation is that the false positive rate is only 3.7% (far better than our other classification models).

Instructions for use with different data sets

The analysis performed in this report was done within the context of helping banks predict whether a client is a good or bad risk. Therefore, we would advise using the approaches found in this report for a similar data set.

Recommendations

In order to remove the imbalance in class labels, we had to reduce our sample size from 1125 to 450. Models perform better when they have a lot of data to learn from, so, in order to improve performance, acquiring more data is recommended.

As mentioned previously, we found that clients with high credit amounts owed were about four times more likely to be labeled as a bad risk. We would advise granting loans to clients that have the financial resources to pay it back.

Figures

Confusion matrix

TPR vs FPR

highest balance vs new balance

Confusion matrix

	Predicted 0	Predicted 1
Actual 0	33	8
Actual 1	18	31

TPR vs FPR

Visualization of clustered data

FPR

Confusion matrix

TPR vs FPR

Data after dimensionality reduction

Class imbalance