Data Analysis Report

Bank Loan Default Analysis & Prediction

Evening Group J

November 28th, 2016

CIS 417 Introduction to Business Analytics

Ella Xiaoyu Wan | Jason Enderton | Florens de Meyer

Jordan van den Beuken | Axel Vandevelde

Table of Contents

1. Project Outline

The context of our dataset revolves around an American loan provider for small loans, e.g. for credit card financing, home improvements, car financing, other major purchases, etc. Mortgages are not provided by this company. This company is already using a staff of analysts to help determine whether loans are going to be paid back or not by assigning a grade. An analyst will reject a loan right away in case the risk is deemed too high (i.e. the grade falls below a threshold). The other loans are accepted (after term and interest rate are agreed upon), in which case the loan can be fully paid back, or defaulted upon, both with or without potential late payments occurring before the final outcome is reached.

The dataset at our disposal contains about 1300 historical loans which are either completely paid back, or defaulted upon. That is, still running loans or loans which were rejected outright are not included. This dataset can be used to predict whether someone is going to default on their loan or not, but also for descriptive analytics (e.g. In which states do most defaults occur? What are the correlations between the variables? etc.). The columns of the dataset are explained below. Note that the data is anonymized: only the first 3 digits of the zip code are shown.

variable name	**explanation**
default	Indication whether this customer defaulted (yes) on their loan or paid it back in full (no).
issue_date	The month in which the loan was funded.
annual_income	The reported annual income provided by the borrower during registration.
employee_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
employee_title	The job title supplied by the borrower when applying for the loan.
grade	Backoffice-analyst assigned loan grade.
home_ownership	The home ownership status provided by the borrower during registration. Values are: RENT, OWN, MORTGAGE, OTHER.
monthly_installment	The monthly payment owed by the borrower.
interest_rate	Interest rate on the loan.
loan_amount	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, it will be reflected in this value.
nr_mortgages	Number of mortgage accounts.
months_since_last_delinquency	The number of months since the borrower's last delinquency.
nr_active_bank_accounts	Number of currently active bankcard accounts.
remaining_principal	Remaining outstanding principal for total amount funded.
nr_derogatory_public_records	Number of derogatory public records.

nr_bankruptcies	Number of public record bankruptcies.
purpose	A category provided by the borrower for the loan request.
term_months	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_balance	Total current balance of all accounts.
total_rec_late_fee	Late fees received to date.
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

2. Descriptive Analysis (Tableau)

2.1 Data Pre-Processing

2.1.1 Date

In order for Tableau to read the date column as an actual date, we had to create a calculated field that translated the string to a date format. We used the following expression to create this "corrected date" variable in Tableau: **DATEPARSE ("MMM/yy" , [Issue Date])**, where "Issue Date" is the name of the variable with the date as a string.

2.1.2 Zip Code

Zip codes from the data we received were anonymized though replacing the last two digits with X's. Zip code thus read, as an example, 146XX. We tried to use these first three digits of the zip code in Tableau since they indicate a region within a state, but it seemed that Tableau is not capable of dealing with this 3-digit zip code of the US. To work around this, we substituted the anonymized terminal digits with 01, so our example then would read as 14601. While this might cause some zip codes with different terminal digits but the same first three digits to group, it allows us to map a rough approximation of the loan data we have to indicate areas that may not be as detailed as an exact zip code but are far more specific than the whole state.

3. Predictive Analysis (R)

3.1 Loading the Dataset

When loading the dataset into R, use the "Import Dataset". Load the *clean_all_c.csv* file and make sure to rename it to "default" before importing.

3.2 Data Pre-Processing

3.2.1 Date

Just like in Tableau, we had to convert the date variable to a date format, that can be read by R. Originally, the date variable was of the type "factor", because the date doesn't contain a day and is thus not recognized as a date by R.

3.2.2 Employee Title

Employee_title was automatically created as a factor variable, however we intended to use it in text mining. Therefore, we had to convert this variable from a factor to a string.

3.3 Feature Selection

First, we tried to build a predictive model on all variables. However, this is not the right way to do it, since some variables considered are irrelevant for the prediction model. When we built the model, the accuracy on the training set was almost 100%, with a much lower accuracy on the test set, which indicates overfitting. This should be avoided at all costs. Therefore, we did some analysis on what variables are relevant for our predictive model. We used common sense as well as a correlation matrix, which can be seen in our Tableau workbook.

So instead we did the following:

To start, we left out the following variables: *issue_date, remaining_principal* and *total_received_late_fees*. The reasoning behind this being that, whenever we want to predict whether a new client is going to default on their loan or not, there is no data available for these variables (e.g. there is no issue date, remaining principal, nor fees because the loan has yet to be issued).

Next, we excluded highly correlated variables which we had discovered through visualization in Tableau. There is a high correlation between *grade* and *interest_rate* (higher grade = higher interest rate), so we left out the latter. The same goes for *monthly_installment* and *loan_amount*; obviously, the higher the loan amount, the higher the monthly installment, so, again, we left out the latter variable. Another one is *nr_bankruptcies* and *nr_derogatory_public_records* since bankruptcies appear on your public record, so we left out the *nr_bankruptcies* variable as well. Lastly, we left out *zip_code* and *employee_title*. This is because these variables contain hundreds of unique values, which results in a huge importance assigned to these variables, which is somewhat misleading.

3.4 Results

3.4.1 Accuracy

With our initial training set (e.g. including all variables to build the model) we were able to achieve a very high level of accuracy (almost 100%). However, much of this accuracy did not translate over in our test set (about 68%). Variables in our initial set were resulting in overfitting and so these variables, including *employee_title* and *zip_code* as mentioned above, were removed from our model for better accuracy in our test set.

Once we set our predictive criteria and ran our test set again, we were able to come up with a training set accuracy of just under 78% and a test set accuracy of about 73%.

3.4.2 Precision/Recall

With our test set confusion matrix, we ended up with 83 predicted positive, or defaulted loan, results of which 19 were false positives and 64 were true positives, giving us a precision of just over 77%. Meanwhile, our test set generated 824 total predicted negative results, that is non-defaulted loans. Of these negative predictions, 641 ended up being true negative with only 183 false negative. This gives us a total of 247 results either true positive or false negative for a recall rate of just under 26%.