DMWAY

A Billion Dollar Question: What Predicts Loan Defaulting?

By Dr. Ronen Meiri, Founder & CTO, DMWay Analytics

A Billion Dollar Question: What Predicts Loan Defaulting?

Introduction

In the not-so-distant past, it would have taken a data scientist a lot of time, and your company a lot of money, to analyze loan data to help predict whether you should approve a loan and, if so, under what conditions. Today, predictive analytics and machine learning make it easy to reduce risk and get paid.

The goal

The goal of this white paper is to present the analytical methodology for estimating the risk of defaulting the loan. The same analytic process with minor changes can be applied and used for different loan and credit lines.

The content of the models below is based on data available from Lending Club, which was downloaded from Kaggle. The aim here is not to get the best model, but to demonstrate the process of evaluating risk using real data. Data is never perfect. But understanding the process leads to more reliable models.

Data preparations can be done with almost any analytic tool, from a business intelligence tool to Excel. Handling the data is the most technical process.

Credit lines

The input data contains two types of loans, loans with 36-month terms and loans with a payment period of 60 days. As a rule of thumb, it is always better to develop a model for each line of credit. In cases where the payment period is a continuous number, one can use the period as a predicting variable in the mode. However, since in this example we have only two periods, we will focus on a loan model for a period of 36 months. In total, the dataset contains 621,125 loans for 36 months and 266,254 loans for 60 months.

To make all the analysis on the same scale, we look at loans that have completed the 36-month period and those that defaulted. The latest payment date (or the maximum of "last_pymnt_d") is January 2016. As such, any loans issues after January 2013 will have a payment period of less than 36 months.

Defining the target variable ("bad loans")

The target variable is a column in the dataset that defines for what the model is trained. In this example, we would like the model to be able to differentiate between "bad loans" and "good loans." We start by investigating the loan status described in the table below.

Charged Off	9071
Current	33
Default	12
Does not meet the credit policy. Status: Charged Off	649
Does not meet the credit policy. Status: Fully Paid	1,789
Fully Paid	63,357
In Grace Period	6
Issued	0
Late (16-30 days)	6
Late (31-120 days)	81
Total	75,004

We see that there are two groups that are not clear and probably need to be excluded from the analysis: "Does not meet the credit policy. Status: Charged Off" and "Does not meet the credit policy. Status: Fully Paid." It is not clear why loans are given to users that "Does not meet the credit policy" and there is not much documentation on these categories.

Now we need to define what is considered a "bad loan." For example, is a loan payment that is 16 days overdue considered a bad loan? Such a loan does possess more risk than a fully payed loan, but it is still in the process of collection. It makes sense in this case to define bad loans as "Charged Off" and the population as all the other loans that are ("fully paid"). The rest of the categories represent loans that are in the process of payment and it will take some time to evaluate if they are fully payed or not.

The target variable is defined as:

Select the population of loans with a status “fully paid” or “charged off.”

Create a new column in the database called “Default” that is TRUE for loan status “Charged Off” and FALSE otherwise.

How to create the feature:

In this specific example, Lending Club has already made the file available for us, but still we need to understand some concepts of how to make a good dataset to get a reliable model.

Causality

Predictive models are a major component of the process that is known as data-driven decisions (DDD). In this case, the model will be used as part of the underwriting of loan applications for approving loan applications and setting the interest rates based on the expected risk. The data that can be used for the analysis is data that is known at the point in time where we evaluate the loan application. In this dataset, any future data like "total_pymnt" (payments received to date for total amount funded) that indicate the total payment from the inception of the loan until the date the file is extracted because it is data we only know after that is later than the date the loan is issued). This concept is known as causality, and we need to make sure that data from the future is not leaking into the model.

Attributes

Creating features is always more art than science, however we would like to share some guidelines on how to create good features.

Start by looking at the dimensions that influence the risk, these can be the user characteristics, sociodemographic information, credit scoring or assets, for example.

Once the dimensions are identified, look at what features are relevant. For example, credit scoring can include external credit scoring data purchased from other sources, user characteristics that may include age, occupation, annual income and so on.

Handling dates

Dates cannot be used directly in the analysis and need to be converted to duration or age or some other type of features. In this file, the date field includes only the month and the year and therefore can be converted only to duration since the relevant date of the loan application or the “issue_d” (the month that the loan was funded) is not

available. The data contains four dates: {earliest_cr_line, last_pymnt_d, next_pymnt_d, last_credit_pull_d} and only the first “earliest_cr_line” has all its values prior to the issue date of the loan. This attribute will be converted to a new attribute indicating the days since “earliest_cr_line” to the issue date of the loan.

Running the analysis

As mentioned, the goal is to be able to predict the default probability of a loan (defined as "Charged Off" in the dataset) for a user.

In predictive modeling, the statistical measures of the accuracy of the mode are not reliable to indicate how good the mode will perform on new data. Models should be evaluated on new data. In this example, we are using the default splitting parameters of the tool and take 2/3 of the data for the training process and 1/3 of the data to validate the model.

Setting the metadata (column definitions)

The list of variables that are excluded from the analysis process model are:

- Int_rate – interest rate is the outcome of the model and not known when the risk of the loan application is calculated
- All dates {issue_d, earliest_cr_line, last_pymnt_d, next_pymnt_d, last_credit_pull_d}
- url – every application has a different URL address.
- Payment columns {last_pymnt_d, last_pymnt_amnt, next_pymnt_d, last_credit_pull_d} are all unknown at evaluation of the risk of the loan application
- Total payment columns {total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, tot_coll_amt, tot_cur_bal }

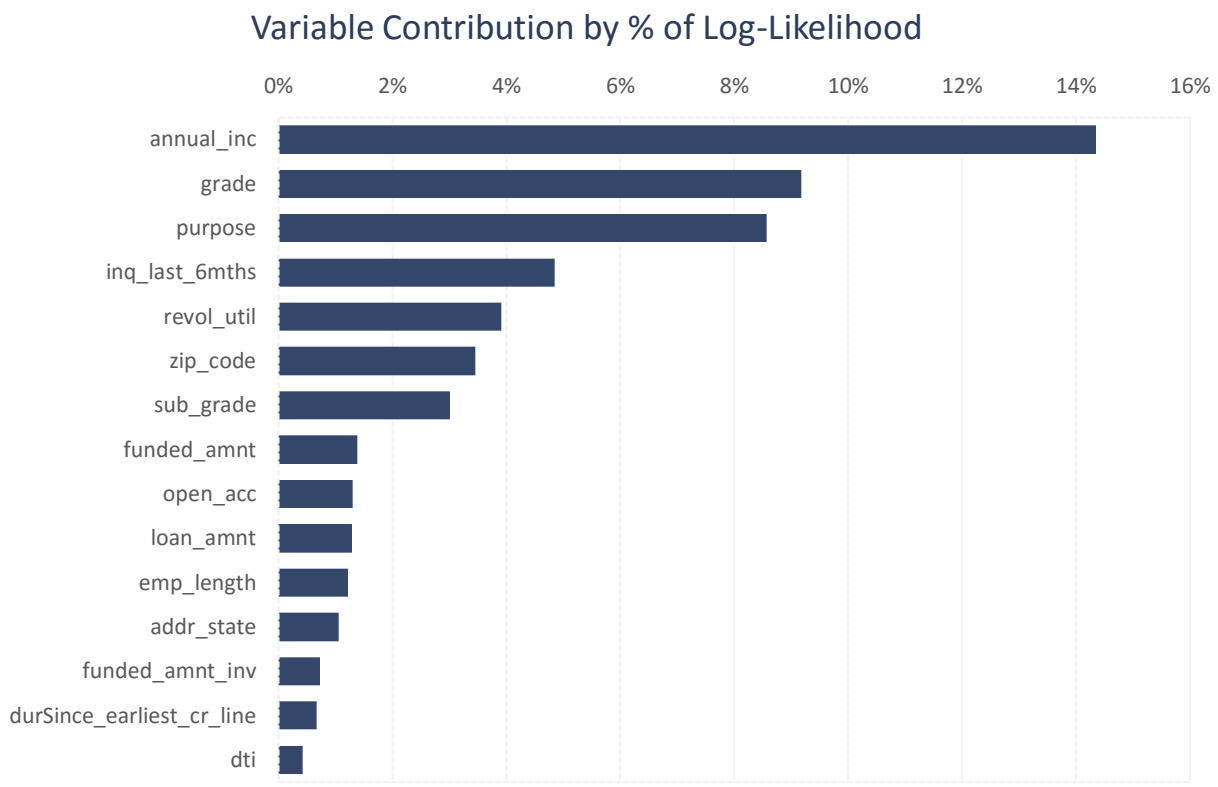
The identifier key is set to:

- Id
- Member_id

Model results

Here are the results of the analysis using DMWay Analytics model-building solution.

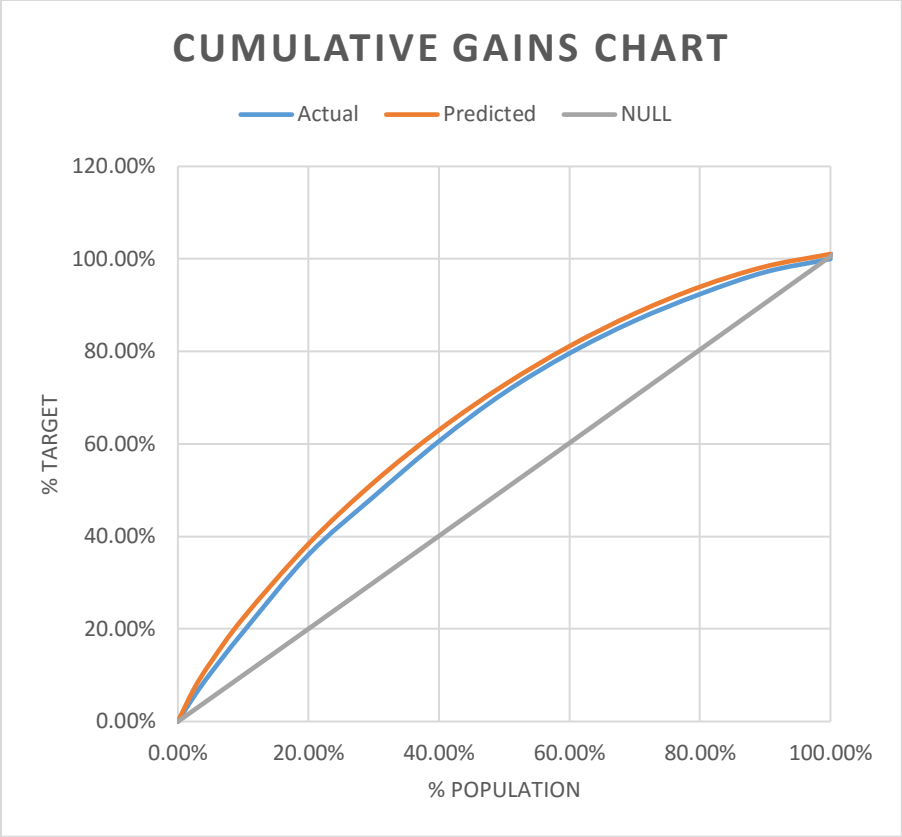
The list of variables and their contribution to the model:



The model is using the current Lending Club “Grade” of risk, but that it is not considered the most important variable in the model. We see that annual income has the highest contribution to the model. The third most influential parameter is the purposes of the loan.

The stability of the model is measured as the accuracy of the model to new data. The Cumulative Gains Chart is a representation of the quality of the model. The *x* axis represent the proportion of the data points (number of rows) from 0 to 100 and the *y* axis represent the proportion of the total data points that are classified as True (loans that have defaulted). The gray line represents the expected relation between the *x* and *y* axis without a model. Taking a sample of 10 percent of the entire population will result in average 10% of the total users who had defaulted on their loans.

Using the model, we can do better by taking the first decile in the validation dataset with the highest probability to default we can capture 20% from the total defaults, or twice the expected average. This is a good model since these are loans that were issued using a risk model. We see clearly that using DMWay Analytics we can perform better than the current model. We can also see that the model predictions are close to the actual values.



About DMWay Analytics

DMWay Analytics is disrupting the data science world by providing an autonomous predictive analytics solution. The AI-based and ML-powered solution enables every subject matter expert (non-scientist) to build his or her own predictive models within hours to days, versus the traditional development time of months to produce the models. The DMWay Analytics platform is versatile and adaptive to all industries.

For more information visit our site: www.dmway.com