Burhan Ahmed Hanif

Professor Adam Kaplener

04/28/2019

Math 390.4

A study in creditworthiness with credit scores

What is data science? Defining this term given that we are being vague is the scientific study of data. There are different subfields of data science such as supervised learning, neural nets, unsupervised learning, deep learning etc. Reality is truth, saying that what has happened, what is happening and more importantly what will happen. Our main purpose as data scientists is to model reality or approximate it with as little error as possible. A model is a systematic description of a phenomenon that shares valid and important characteristics with the phenomena. The two reasons for modelling are; explanation and understanding how the phenomena works and prediction can the model tell us what will happen in the future given certain conditions.

The model that will be built in this paper will be a mathematical model. A model that will take numerical values in its variables as inputs and give us a numerical output as a result. This numerical output will be quantified for making decisions. A famous statistician named George Box once quoted “all models are wrong, but some are useful” by George Box. This mathematical model will be made from credit reports generated by credit bureaus such as Experian, Transunion and Equifax. This model will be perhaps useful for understanding the consumer credit market. Also Using these credit reports we will generate a model and use it to predict if a person will be able to meet their debt obligation or default. This model is not an original idea because such a model already exists, and it has been developed by the Fair Isaac Corp, called a

FICO score. This score is used by almost 90% of credit lenders in the United States and is now making a world recognition and being adapted by other countries as consumer credit markets are increasing around the world. The actual model itself is not publicly available but certain content about how it works is. Model building accepts this assumption of stationarity that given behavior of the past will continue in the future unless it changes completely.

Credit bureaus mentioned earlier generate financial reports about an individual credit information by acquiring information from creditors, debtors, public records, debt collection agencies etc. They work with creditors such as banks, mortgage lenders, credit card companies etc. these bureaus do not make the decision if anyone should have credit extended to them they just generate reports and assign a credit score or give to their clients what is called a credit report. Prior to credit scores the model of lending institutions were human underwriters whom reviewed each application and made decisions on their expert opinions. This was extremely time consuming and the error rates were high.

What is the philosophy behind having a credit score? A credit score sets a standard for the consumer credit market. It makes socio-economic equality for the consumer credit market given that there have been prior injustices towards minorities in the past. For example, now they meet compliance standards of equal housing lending. It gives a metric for a consumer standing defined by the credit markets. For our purpose of modeling we have a mathematical model that can interpreted by its features and its output can be comprehended by the consumer and its lender. The output is a number between (0-1000) 0 being no credit, 200 being the worst credit and 1000 being perfect credit. Other benefits are; If the model is implemented by the lender it can be automated to a system and output or a decision can be generated in the matter of seconds.

But if a lending institution does not have an automated system, then a human can decide just using the metric of credit score and a threshold.

In our case for this model we will be implementing the specific branch of data science called supervised learning. Supervised learning has three ingredients, the first one is historical data which we will call our training data also known as our “D” which consists of our features (p’s) and outputs (y’s) and there is an “n” amount of them in our sample size. Our x’s in our data set will consist of payment history, declared bankruptcy, dollar amount of payments made to credit cards, credit utilization, length of credit history, long-term debt and employment status, other incomes, by combining all the points generated by these features will give us a credit score. We will predict our y as probability of default on debt obligation use the credit score. Second ingredient is we will need candidate functions that will model our best guess function also known as ***H***. Third and final ingredient is our ***A*** this the space of all algorithms that we can use. We will be working and referring to this equation. ***g= A(D,H)***.

Before building any model, a data scientist must be certain that his features (x’s) of his data set are valid for what he is trying to model. For instance, if we are trying to model stock prices of IBM and if one of our features is did it rain in Toronto Canada, this feature doesn’t have any relevance for our model but if one of our features tells us about their overhead on salaries this year this feature has relevance to their expenses. Ethics is another standard that must be considered for data scientists, especially now because of what social network websites such as Facebook have been doing with their data, the point to make is that there should be legal means of how the data was acquired and how the data should be used. Ethics is not our main point of discussion in this paper but there are enough resources on the internet that you can refer to if the reader is interested.

For our feature of payment history using the information given by (Finlay 4) the fico scores have created a nominal binary variable. A person earns points or loses points on the fico scale by if their accounts are up to date or how many days they are late by on each credit card. If the account is 0 days late then you earn 14 points, 1 to 30 days late 0 points earned, 31-60 days late -29 points, greater then 60 days late -41 points. The problem of creating the variable this way it doesn't tell us the what are the size of the payments a person is making relative to what they owe to the credit card company. A person can end up making the minimum payment and earn the points on the fico scale, but the information can be misleading.

Continuing the idea of payment history another variable should be added because there exists a concept of bankruptcy. A person can declare bankruptcy and take a settlement for his debt obligations. The idea behind making this a separate feature has a philosophical point that is a person is living beyond their means and the practical point for the model will be if given the chance again there is a high probability that they might repeat the same behavior. The threshold will be if a person has declared bankruptcy in the past 5 years. The way we will create this variable would be a binary variable {0,1}. If yes = 1, then -100 points on the score and if 0 no points added or subtracted.

Misleading information from the feature can be damaging for the model. Our model can take this variable and improve it by creating a factorial variable for short term debt of credit cards and each factor should be ratio of payment made and money owed. For example, if one owes 10 dollars, 2 dollars payment should worth less in points then a 7 dollars payment. Given that credit cards have different limits and interest rates, one chooses to pay them accordingly that's why the factors should be ordered and worth points accordingly. This way the information will not be misleading because we are predicting chance of default on credit scores.

Credit utilization is another important feature in the dataset. This is the ratio the amount of money borrowed on credit to the amount of your credit limit. For instance, if you have 2 credit cards one has a credit line of 10,000 with a balance of 5000 and another card with a credit line of 1000 with balance of 200, your credit utilization ratio for both cards is 47%. This feature also depends on the revolving debt that you have on your credit cards. Debt that remains unpaid on your cards, when a person doesn't pay of their account balance at the end of the billing period. For this feature we create a variable of how much is the debt/credit ratio in percent and what is the revolving debt in dollars. According to the fico corporation this accounts for 35% of your score. A variable for this would be continuous so the points allocated for the score would (1-credit utilization) * 100.

The more historical data you have the better your model will be and the same applies for your fico scores. Another feature is the length of your credit history or the length of time each account has been open for. A variable created for all the credit cards as separate sub variables to tell us length of time in months. While opening new accounts in a short period of time is bad for your credit score because it comes of as financial instability. The variable can be coded in this manner the length of time each credit card has been active then summing over the time of months. Create thresholds for the number of months and then assign points accordingly.

If we are modelling probability of defaults or paying back loans a need for an income variable is a necessity. Our most important feature for the model will tell us your employment status. The feature can be nominal variable (full-time, part-time, retired, student, unemployed). If fulltime employed (less then one year, 1-5 years, over 10 years) points granted accordingly from less to more. The others will take either zero points or negative points.

There can be other forms of incomes other then employment. Such as capital gains, annuities, asset owner, equipment leasing, consulting etc. variable creation for this will nominal binary and for what is true sum up the dollar amount and output total.

Long-term debt must be considered for more of the calculated score because it is a liability for years. Starting with student loans, house mortgage or extensive medical bills etc. We can create a variable and subset it with binary variables {0,1} if a long-term debt exists take the sum and give the output in dollars (less than 50,000, 50,000-200,000, 200,000-1 million) then subtract points from the score accordingly.

Using these features and adding the points will give us a credit score. We will train our model on the training data we already have. It is not possible that we will be able capture the exact truth or reality. Our model will be g(x) + error, in supervised learning g(x) will be the best guess for a function that models the truth plus the error that is known as (***e***). Our job as data scientists is figure out the best function and minimize the error so we can come as close to the true function as possible. The error is grouped into three sub categories; estimation error, misspecification error and error due to ignorance.

Estimation error is how far is the distance of your y's and your function. Imagine you are on a x-y coordinate and there is a line y=x, you are walking on this line there are points on top of you and underneath you the distance between you and the points is the error. This can also be stated as how far are you from the truth. This distance summed up for all the points is also called the residual error. For an equation purpose this defined by h*(x) – g(x).

The next error term is our misspecification term for an equation this is f(x) – h*(x). This error can be explained by how our function is not mapping the truth function. The space between our f(x) and our best guess for the actual function. To decrease this error, we must increase our algorithm space. The final error term of error due to ignorance means simply that there is something happening in our data we are unaware of t(z) – f(x). We are ignorant on this error because there is missing information that we do not have. The way to improve our model or decrease this error we must add more relevant features to our dataset. At the end of this now we can state supervised learning as an equation $y = g(x) + h^*(x) - g(x) + f(x) - h^*(x) + t(z) - f(x)$**.**

What we must come to understand about error is that if our end goal is match the truth function by any means necessary then that would be a loss to our model. In supervised learning there is a concept of overfitting and we want to avoid this because the purpose of our model is not to fit the data we already have but to predict a future observation we get. If our in-sample error is low, then out-sample error will be high. In a nut shell this means that do not overfit the in-sample error because the out of sample prediction would be awful. A way of doing this would be to start adding more p features to our dataset known as fitting random noise to improve our model.

Supervised learning is limited only for interpolation meaning that if our new prediction is within the range of our X. supervised learning does not generally work with extrapolation where our new observation is out of range for our X. Certain models extrapolate differently an example will be polynomial functions. Prediction in extrapolation would be dire at certain times because polynomial functions can end up moving 90 degrees up and down with almost no change in x. For using prediction in polynomial functions, it is recommended not go over the power of two because prediction will suffer drastically.

Our first approach for using algorithms depends on our data, but we need a starting point and that is null model. There is a rule of thumb that no matter what fancy machine learning and highly complicated model you build but if you cannot beat the null model in prediction you are most likely fired from your job. An initial starting point is to use our model space for all linear models. For our dataset we can start with linear separation with a binary output of {-1,1} and the credit score as our X. Our algorithm will create the best straight line for linear separation and yes there will be some error. An instance with words without looking at what we have generated someone with a high credit score defaulted and some one with a low credit score did not default on their loans.

Let us increase our model space so we generate less in-sample error without overfitting. We can use the support vector machines algorithm also known as SVM. This algorithm will create the best fit linear separable line with two different wedges as close as possible to the vectors and those wedges will have some vectors on them. These vectors will be what we will use to train our model hence the name support vector machines. Another linear algorithm we can use is kth nearest neighbor also known as KNN. This algorithm takes the information about a vector, uses the information to output that their other vectors such as itself with a certain amount of error.

The algorithm of ordinary least squares known as OLS will help us a lot in this model because we will be regressing over the features individually known as linear regression or we can do multiple linear regression. This model will give us a better explanation because we can see the weights that each feature will carry so we can predict better. This model will also provide us with two different types of error metrics R squared and RMSE. The R squared will provide us a percentile of well we are fitting our model. The RMSE will provide us with a mean of how much amount of points we are off from each observation.

OLS is one of the best algorithms for fitting models without overfitting given we have enough n observations. The error metrics are easy to comprehend and explain to clients or employers. It also has many forms because of its background in math, bringing down the dimensions without the loss of information. For our model we can use its other forms such as interactions with variables that increases our r squared and decreases our RMSE without overfitting. The same can also be applied with forward stepwise and backward stepwise linear regression. This takes the best weights for the variables and runs the regression on them accordingly.

Polynomial regression as said before is unstable and opens the door to a lot of risk in the extrapolation. Polynomial regression also has high risk of overfitting because it starts modelling noise as soon as you start increasing the exponent > 2.

Bibliography

Finlay, Steven. *Predictive analytics, data mining and big data: myths, misconceptions and methods*. Palgrave Macmillan, 2014.