TO 414 Group Project #2: Prosper Data Analysis

Document created by Sanjeev Kumar

Project Specification Update

Q1. What are all these variables?

As you would have noticed, the dataset has a **lot** of variables. Most of these variables are intuitive to understand - but several are not. So here is a quick explanation of the variables in the dataset.

number_of_days : How many days did the loan took to default. For loans that did not default, this shows the days until the last observation.

principal_balance : Loan Principal Balance Outstanding. It should be zero for all complete loans (i.e. loan has been paid off).

loan_status and loan_status_description : Categorical variable showing the last status of the loan. Values include Current (status = 1) - loan is being paid off on schedule, Defaulted (status = 3) - loan has already defaulted, Completed (status = 4) - loan has been paid off completely, Chargeoff (status = 2) - loan has defaulted and the balance has been sold to a collection agency for collection. For our analysis purposes, you have the option of combining Current and Completed together; and Defaulted and Chargeoff together.

loan_origination_date : The date the loan originated (was funded).

amount_funded : The amount that was borrowed for the loan. This is the starting balance.

prosper_rating : Prosper rating of the listing for the loan at the time the listing was created. Possible values: AA, A, B, C, D, E, HR, N/A. This is Propser's evaluation of the credit risk profile of the loan.

borrower_rate : Interest rate at which the loan was originated.

listing_term : Length of the loan in months. Possible values are 36 or 60.

listing_monthly_payment : Monthly payment that the borrower needs to make. Note that monthly payment is not an independent variable as it can be calculated using amount funded, borrower rate and listing term.

scorex : Credit score - provided in different ranges. Ranges possible are: < 600, 600-619, 620-639, 640-649, 650-664, 665-689, 690-701, 702-723, 724-747, 748-777, 778+.

prosper_score : A custom risk score built using historical Prosper data. The score ranges from 1 to 11, 11 having the lowest risk.

listing_category_id : Broad Category of the Listing. The data is integer coded and we no longer have direct visibility into what different categories are.

income_range, income_range_description, stated_monthly_income : Income level of the borrower. Data provides stated monthly income and then groups them into one of several ranges. Each range is then given an integer code.

income_verifiable : Logical value stating whether the income information provided by the borrower has been verified or not.

dti_wprosper_loan : Debt to income ratio including the monthly payment of the prosper loan. This value should be a fraction less than 1. You would note that some values are coded as 1000000 - this is indeed an invalid value and may have been code for something else. You should explore what this code may mean and how to handle this value.

employment_status_description, occupation, months_employed : Self explanatory. You will see a lot of ”other” as occupation - it is because borrowers were asked to choose their occupation from a list of options and the list was not broad enough leading to many borrowers choosing ”other”.

borrower_state, borrower_city : Self explanatory.

lender_indicator : Whether the borrower also has a lender Role in the marketplace. 0 = Holds borrower role only, 1 = Holds both borrower and lender roles.

monthly_debt : Monthly debt payments currently being made by the borrower. This usually include such items as auto loans, monthly minimum payments on credit cards, student loans, mortgage etc.

current_delinquencies, delinquencies_last7_years : This is count information from credit report regarding delinquencies and public records (see below). Delinquencies refer to delay in making a payment. Public records include adverse information such as bankruptcies (rare) and collection activity (more frequent).

public_records_last10_years, public_records_last12_months : See above.

first_recorded_credit_line : Date of the first credit line in the credit report of the borrower.

credit_lines_last7_years, inquiries_last6_months : Count of how many credit lines were opened in last 7 years and how many credit inquiries were made on borrower's credit report in last 6 months.

amount_delinquent : Indicates the past due amounts owed by the borrower. This includes accounts included in Chapter 13 bankruptcies and other unpaid derogatory account balances.

current_credit_lines : The number of open or closed accounts in the borrower's name that the borrower is paying on time.

open_credit_lines : The total number of open accounts.

bankcard_utilization : A percentage value determined by the sum of the balances owed on open credit cards divided by the sum of the cards' credit limits.

total_open_revolving_accounts : The total number of open revolving credit lines a borrower has. Revolving accounts include credit cards.

installment_balance, real_estate_balance, revolving_balance : Balance outstanding on installment accounts, real estate accounts and revolving accounts. Installment account include items such as auto loans or purchase loans.

real_estate_payment : Monthly payment amount on real estate credit accounts.

revolving_available_percent : Percentage of revolving credit available. Note that this is the inverse of the bankcard utilization value we saw before.

total_inquiries, total_trade_items, satisfactory_accounts : Count of total number of inquiries on the credit report, total number of accounts and total number of satisfactory accounts.

now_delinquent_derog, was_delinquent_derog : Number of accounts that are currently delinquent or derogatory and number of accounts that were delinquent or derogatory in past.

delinquencies_over30_days, delinquencies_over60_days, delinquencies_over90_days : Number of accounts where payment is now late by over 30 days, over 60 days and over 90 days respectively.

is_homeowner : Logical value indicating whether the borrower owns his/her home or not.

Q2. Should we be using all the variables in the dataset

As we just saw, we have a lot of variables here. We should not blindly use all of them in a stepwise model and reach a situation where you have a high r-square but the model makes no sense. Recall the purpose of the model - we want to uncover differences between how the market prices risk (through adjusting interest rate) and how those risk factors actually lead to loan default or not. We need models that show us the marginal effects - small items that are significant but which may get overpowered by other variables.

Issue 1: What about aggregated variables?

There are some pre-processed aggregated variables out there - like Prosper Score, Prosper Rating and Credit Score. We don't have visibility into how they were calculated. These are valid information that are available to the market - so they should have a place in the model. However, we might be better off using the constituent information that these aggregates are based on rather than the aggregates themselves.

We can resolve the issue by building two sets of models. One in which we allow aggregated variables and one in which we don't allow aggregated variables. You will see that the r-square will necessarily be lower for models that do not have aggregated variables - but these models may also allow you to explore significance of variables that would otherwise be overpowered by aggregated variables.

You may also decide to focus on only a subset of the data. For example - run a model only of High Risk borrowers (Prosper Rating HR) - that is where you are likely to find more interesting relationships. You can explore how to divide the dataset into different subsets that may behave differently.

Issue 2: Which variables to use?

Which variables to use is really a theory driven question. You should ask whether a particular variable is theoretically likely to influence your dependent variable - if yes then include in your model. However, pay attention to the facts that not all variables are independent. Some variables can be derived from others - so they do not add any additional information and should be excluded. Some variables may have many missing values - they should be treated with caution. Some variables may not have been coded well - may have values that don't make sense (like the debt to income ratio variables) - these variables may need cleaning up before they can be used.

Issue 3: What is our benchmark?

When you have these many variables, it is difficult to figure out what is the optimal model, how do we figure whether one model is better than the other, what is the benchmark or decision criteria. You should feel free to use various fit parameters we have discussed in class including AIC values and anova function for comparing between nested models. Avoid the temptation of just trying to get a high r-square value. Remember that a very high r-square value might mean that you are overfitting your data.

Parsimony has a definite value here. The smallest and simplest model that explains most of the variation often turn out to be superior.

Q3. Logistic Model or Duration Model?

For the second part of the project - figuring out what causes loan defaults - you have the option of using a logistic model or a duration model. Logistic models have easier interpretation but they do not incorporate all the information (specifically time taken to default). Duration models may be more difficult to interpret but they are more appropriate for the data we have especially for right censored data.

As always, the answer is - try both and see which works better. You may find that the additional complexity of duration model is not commensurate with the additional information (if any) provided by the duration model. However, we will not know that until we have both these models in hand.

Q4. How to compare part 1 and part 2?

The key part here is to compare the model of interest rate with the model of default. I am leaving this part open to you - you can follow any methodology/approach you want. This is a difficult task to do - both conceptually and computationally - so leave enough time to do this.

Q5. What about testing?

After trying out several different scenarios, I have taken testing out of the project. Doing testing took the project way too far from where I wanted the focus of the project to be - on model fitting, model interpretation and model comparison. We will still do testing - but I will do it as a demo in classroom rather than asking you to do it.

Q6. This PDF document looks nice. How did you make it?

Believe it or not - this document has been made in RStudio. This is a Sweave document that uses a LaTeXinstallation to build PDF documents with embedded R-code. This is a neat way to create PDFs.

Finally - We should all look to Calvin for inspiration

IF YOU DO THE JOB BADLY ENOUGH, SOMETIMES YOU DON'T GET ASKED TO DO IT AGAIN.
© 1993 Watterson/Distributed by Universal Press Syndicate
1-30

Not a bad strategy while you wait for the following:

WHATCHA DOIN' ?
I'M KILLING TIME WHILE I WAIT FOR LIFE TO SHOWER ME WITH MEANING AND HAPPINESS.