Big Data Analytics Final Project

Due Friday, December 7, 2018 at 11:59PM

1 Case Description

Chess Bank recently received 250 loan applications. As Chess's chief data analyst, your job is to propose a list of applicants that the bank should give loans to, with the goal of maximizing total profit from these loans.

The "application2018.csv" file contains information of these applications, including loan amount and personal characteristics such as income, credit score, age, and education level. From the bank's administrative records, you have pulled similar data from 250 historical loan applications and their payment information, contained in the "application2017.csv" file. The next step is to learn the relationship between individual characteristics and loan payment behavior from the 2017 applicants, and predict payments for the 2018 applicants.

The following variables are available for the applications:

- ~~ID -~~ A unique identification number that the bank assigned to each applicant
- ~~NAME - L~~oan applicant's first and last name
- ~~SSN - (format: X~~XX-XXX-XXX) Loan applicant's Social Security Number
- ~~DATE - (f~~ormat: MMDDYYYY) Date when the ~~l~~oan application was submitted
- LOAN_AMT - Loan amount applied for (in dollars)
- ~~AMT_DUE -~~ Loan amount plus interest to be paid (in dollars)
- ~~STATECODE -~~ Loan applicant's state of residence
- ~~AGE~~ - Loan applicant's age in years when the application was submitted
- MARRIED - Loan applicant's marital status when the application was submitted (1 if married; 0 otherwise)
- ~~EDUC~~ - Loan applicant's self reported education attainment (1 if less than high school, 2 if high school graduates, 3 if college graduates)
- W2INC_M1 - Loan applicant's wage income 1 year ago obtained from W-2 Form (in dollars)

might highly correlated

- W2INC_M2 - Loan applicant's wage income 2 years ago obtained from W-2 Form (in dollars)
- TAXDEPENDENT - Loan applicant's number of tax dependent in the previous year
- CREDITSCORE - Loan applicant's credit score
- ASSET - Loan applicant's self reported total asset (in dollars)

- DEBT - Loan applicant's self reported total debt (in dollars)
- ~~UNEMPRATE~~ - Unemployment rate in the loan applicant's state of residence when the application was submitted (in %)
- AVG_HOMEPRICE - Average home sale price in the loan applicant's state of residence

For the sample of historical applicants, you also observe

- AMT_PAID - The amount of money paid back by the application (in dollars)

There are several simplifications we will make in this exercise

1. Loan amount has already incorporated costs of loan so that the difference between amount paid and the loan amount is the bank's profit. In other words, the bank does not lose money as long as the applicant pays back the loan amount. Profit is positive if the applicant pays back the loan amount plus some interest. Profit is maximum if the applicant pays back the full "AMOUNT_DUE"
2. The bank is not constrained by how much loans it can issue in total. In other words, loans should be given to all applicants whose predicted payment is larger than the loan amount
3. Payment amount (i.e. the AMT_PAID variable) is calculated as present value of all period installments. In other words, you don't have to consider value of early payments.

2 Project Deliverables

You should deliver the following three files as part of your final project:

1. **Executive Summary**. This file, which should be in Word format, should list whether this is an individual or group assignment (up to 3 members per group, with names listed). Because I will be reporting final project results publicly to the class, give your project an appropriate Team Name to preserve your identity. The document should then have three sections: a concise overview of the case and objectives, a description of your methodology, and an actionable conclusion which should include a summary of your loan recommendations and an estimate of profits to be earned. Your executive summary should not be more than 3 pages in length. You should use the template posted to Compass.

2. **R script**. Along with your executive summary, turn in the R script used to perform your analysis. It should be created such that if the data files are in the same folder as the script, I can run your script on my computer and generate exactly the same results. (Depending on the analysis you use, you may need to set the random seed at the beginning for this to occur.)

3. **Loan File**. You should report your loan decisions in a csv file with the following format:

ID	NAME	APPROVE
⋮	⋮	⋮
522	Jerry Warner	1
523	Herbert George	0
524	Rick Franklin	1
⋮	⋮	⋮

   where the APPROVE variable indicates whether each applicant should be given the loan according to your model prediction, 1 if approved and 0 otherwise. The csv file should have 250 rows (one for each application) and 3 columns (ID, NAME and APPROVE variables).

Your final product should be a csv file in the following format:

On the last day of class, I will report the predictions and actual profit realizations for each team, and we will see which team realizes the highest profit for Chess!

3 Project Grading

1. **Executive Summary**. Your grade on this component will be based primarily on your ability to clearly communicate your objectives, methodologies, and results. Avoiding jargon, correct grammar, and proper sentence and paragraph structure will all be relevant.

2. **R script**. Your grade on this component will be based on two factors. First, how easy is it to follow your script? Clearly commenting your code and using informative naming conventions will help. Second, are your results replicable? If the raw data files are in the same folder as your script, I should be able to run your code without any modification and replicate the results in your final project.

3. **Loan File**. Your grade on the Loan file will be based on formating and the total amount of profit your prediction generates. If the file is properly formatted, the lowest grade you can receive on this component is a B+. As a special hat-tip, the project achieving the highest profit in the class will get an A+ on this component.