Riya Mokashi

An Analysis of Loan Repayment

Executive Summary

This document is an in-depth examination of the loan repayment dataset as well as my thought process when creating a suitable algorithm to predict future values. For this experiment, we could accurately predict loan repayment rates after being given statistics on certain schools.

Data Summary

The data itself is quite expansive, reaching over a thousand variables and values. Specifically, we reached 8705 data values. For this reason, I chose not to try to extract values, but to instead examine the data in its entirety.

The data, in this case, is set up so that each row or row_id is an individual school and the variables are various statistics of the school above.

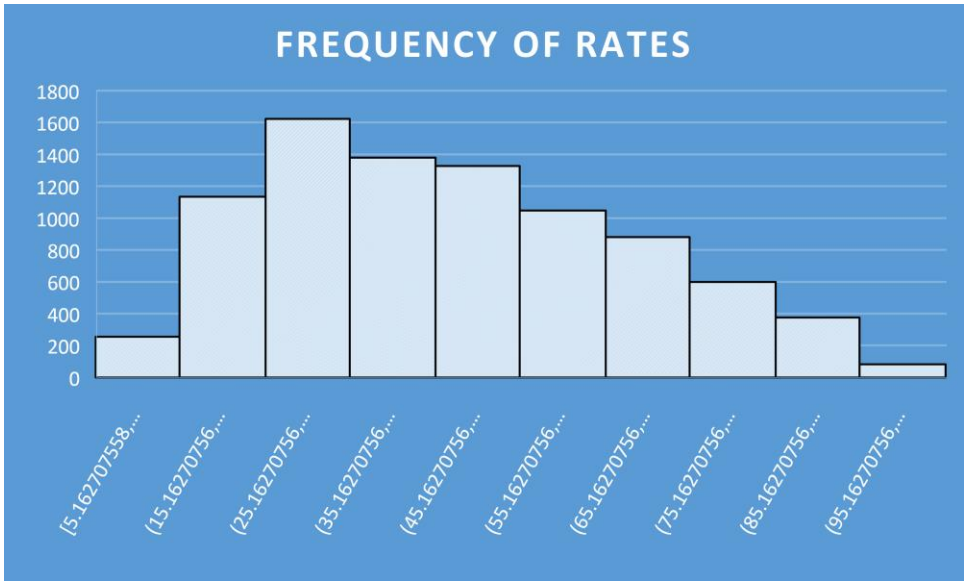
First and foremost, I examined the different categories of variables.

- ACADEMICS
    - These variables describe the various educational programs certain percentages of students are involved in. The Academics variable will be valuable in that it will be able to pinpoint individual correlations between loan repayment and academic courses and enrollments.
    - This category contains 228 separate variables. In this context, a 0 means ‘does not offer,' a one means ‘does offer,' and a two means ‘does offer over distance learning.'
- ADMISSIONS
    - This provides the admissions statistics for individual schools when first examining it can be noted that a majority of the ACT SAT values are blank. This variable contains valuable information about the baseline education levels of entrants into these schools. Such information can also correlate with demographic variables.
- AID
    - This variable is strongly correlated with a school’s ability to offer financial aid aimed towards students. Aid is important as it lowers the need for a student to look for big loans. If offered enough financial aid a student should be able to collect smaller loans which will, in turn, be easier to pay off in the long run.
- COMPLETION
    - End indicates levels of the finalization of the academic programs included above. Completion may or may not be a variable of statistical significance. This variable can point to the number of students who thought it prudent to go through their entire schooling, and low levels of completion can indicate lower repayment rates.
- COST
    - Cost is simply the overall cost of attendance. This variable can be closely tied to financial aid in that it is a fundamental cause for repayment rates. For example, a

school with a lower tuition will, in turn, have less need for large loans, and perhaps a greater loan repayment rate.

- YEAR
    - Indicates the years at hand. This is useful for dating the schools, however, beyond this, it is likely not a measure of statistical significance.
- SCHOOL
    - Indicates specific physical statistics of the schools, not to be mixed up with the fact that every category of variables is inherently a statistic of the school. This can be variables such as facilities offered and location. This can show some of the demographic cause and effects on repayment rates. Such statistics are suitable for inferring a bit about the student population as well.
- STUDENT
    - Is indicative of the statistics of the overall student population of the school at hand. This, like the SCHOOL category, is useful for looking at demographic cause and effects towards loan repayment rates. Students coming from wealthier backgrounds may need smaller loans leading to greater repayment rates.

The graph below shows various frequencies of the loan repayments. I initially created this chart so I could get a feel for what kind of values to look out for when executing my model. It should be noted that I merely did this so I could visually see the average, median, and mode of the loan repayment rates. This is so I can get a feel for what kind of values I should expect when creating my model.



To complement this data, I further explored the data set by finding the mean, max, min, and median of the repayment rates. This was merely done through excel within the same excel.CSV file.

Category	Value
Min	5.162708
Max	100.4736
Median	44.85505
Mean	47.37086

Now that we can see the summary of the data it would be good to take a minute and look for errors. One thing we know is that one value is >100 meaning over 100% of people repaid their loans, an occurrence that is simply not possible. This is good to keep in mind for when we continue to our algorithm building as it will help us weed out errors and unwanted values.

Beyond this point, I thought it would be fruitless to continue trying to gain statistical knowledge on the data. With the overall structure in mind, I would now be able to commit this data set to a model and train and test a model that would suit this context.

Working with the Excel Files

To make the process easier, we will be slightly manipulating the Excel spreadsheets. Upon examination, we can see that the train_values sheet has all the information we need, while train_labels has what we need to find (repayment_rate). For our algorithm, we will be copying and pasting the repayment_rate column into train_values so that we can feed the data into our algorithm in one go. This may be a bit unorthodox, but the creation of a simple, yet efficient, the algorithm can stem from changing the format of your data files as well. After doing this, I was able to create an algorithm that will require only one input.

After completing this, the Excel spreadsheet looks like this:

	A	B	C	D	E	F
1	repaymen	row_id	academics	academics	academics	academics
2	19.91653	3	0	0	0	2
3	55.00939	4	0	0	0	0
4	60.20351	5	0	0	0	0
5	19.51239	6	0	0	0	1

The row_id column has been highlighted to show how the row_id values should line up when transferring costs. Here the repayment rates have been inserted on the left of the rest of the data and are going to allow us to carry out our algorithm.

Riya Mokashi

Creating our Algorithm

This test called for Azure ML as it offers a user-friendly machine learning interface. To begin our first step should be to load the data sets into the machine learning studio. Once we have done that, we can now upload them into our blank experiment. In this case, we would upload train_values(1) which would contain our repayment rates as well as all the other information. This can be done with the upload data module that can be found in Azure.

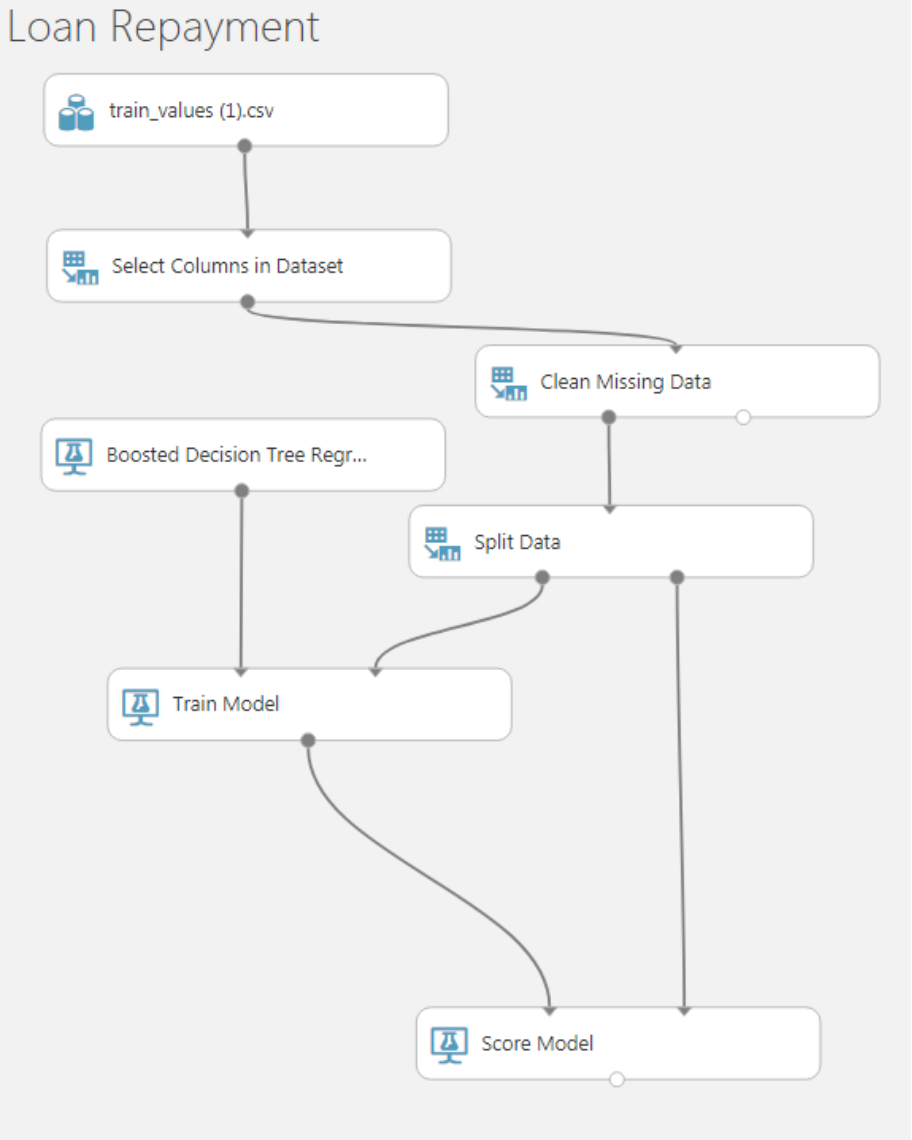
At this point, we should consider filtering out some of the columns. After close examination, I decided the only column that provided misleading data was the row_id column. After all, this column is merely an identifier and not any actual data about the loan rates or the schools. This is because the row_id column simply indicates a school's given a number, it has nothing to do with its statistics but is only an identifier.

Once this column had been removed, I was then able to implement the clean missing data module. This removes rows with missing data and ensures that the algorithm will not be swayed by n/values. I at first considered filling missing information in, but that is statistically improper as we have no real way of knowing exactly what those numbers should have.

After doing this, I split the data 75/25. This allows 75% of the data to train the model and for 25% to test the model. This makes it possible to check our accuracy but still, affords enough to train the model accurately. To a certain extent, this can be preferential. I have seen many algorithms where people tend to find it more acceptable to do an 80/20 split for their data as it provides the algorithm with more leeway to be trained.

For this experiment, I decided on a Boosted Decision Tree Regression model. This model will most accurately suit the predictive style we are going for. Once this was done, I trained and scored my model. While training I was sure to select the repayment_rate column to be the one trained for. We were already given the information that the resulting algorithm would have to be a regression model, and after careful examination of the forums, I saw that many had indicated that Bayesian models did not work well in this situation. For this model, I was prepared to experiment with various algorithms, but the Boosted Decision Tree Regression model worked accurately.

In the end, the flow chart looked like the one below.



At this point, the algorithm was trained, but now we had to use it on the Test data. To do this, I deployed a web service that allowed me to enter the Test data as a batch. This allowed me to create an application in which I could either upload a CSV or hand come in information to predict the loan repayment rates at hand.

I received an RMSE (Root Mean Square Equation) score of about 7.1 which indicated that we would be able to accurately measure the likelihood of loan repayment after considering a school's statistics.

At this point, I was able to conclude that, with the given data we were, in fact, able to create a regression model that accurately predicted the loan repayment rates for any typical school.