edx Username: roseenah1

Student Loan Repayment Prediction

Rojina Deuja, July 2017

1. Executive Summary

As a part of the final requirement for the completion of Microsoft Professional Capstone: Data Science, this report is prepared describing the prediction model created for the Capstone project. The datasets required to be explored were provided and the output was thus calculated using the dataset. The data set can be categorized into two different types:

a. Training set (Values available in train_values.csv and labels in train_labels.csv)

b. Test set values as test_values.csv

c. Submission Format with submission_format.csv.

The major purpose of being provided the dataset is to carry out predictions for Educational Data Mining (EDM) purposes. The dataset is representative of the US student's demographics. It helps to evaluate the student's ability to repay loans that depends on multiple factors. The dataset is obtained from the published information from US Department of Education about student loan. The dataset comprises of a total of 8705 observations or data objects. The test set consists of a total of 6391 unlabeled data that has to be predicted with optimum accuracy, by implementing the most suitable techniques.

2. Data Preprocessing

Firstly, the data was explored amongst the 444 features provided in the given data and extracted the most influential attributes for input to the model. To avoid the problems like "the curse of dimensionality" only 83 attributes were chosen as input. Then the data was fit into various machine learning models to choose the one with the best performance in terms of accuracy and low rate of error. Finally, a linear regression model was found to give the most satisfactory output.

The features selected for the model in this analysis were:

Firstly, some essential summary and descriptive characteristics of the data were calculated. Excel was used to quickly compute the following summary statistics for the repayment_rate:

Mean 47.3709

Median 44.855

Minimum value 5.1627

Maximum value 100.4736

Standard Deviation 20.9876

edx Username: roseenah1

Some graphs are shown below for insight:

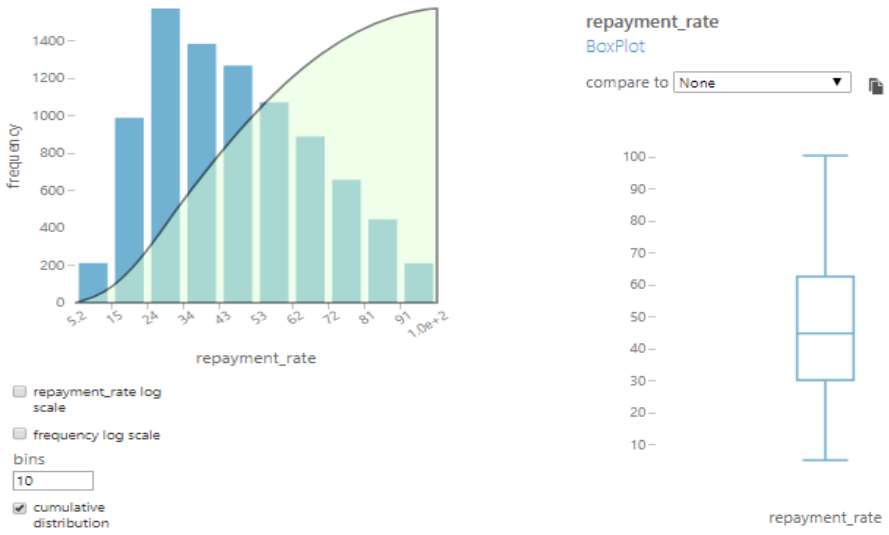

Fig. repayment_rate column histogram and box-plot

2.1 Data Cleaning

One of the major issues in the given dataset was Missing values. Thus, extensive data cleaning was incorporated. The data had two major types of missing values: a. Numeric and b. Categorical variables. The numerical data was cleaned by substituting the missing row with “0”. Similar, the categorical variables were cleaned by substituting the missing row with a value “unknown”.

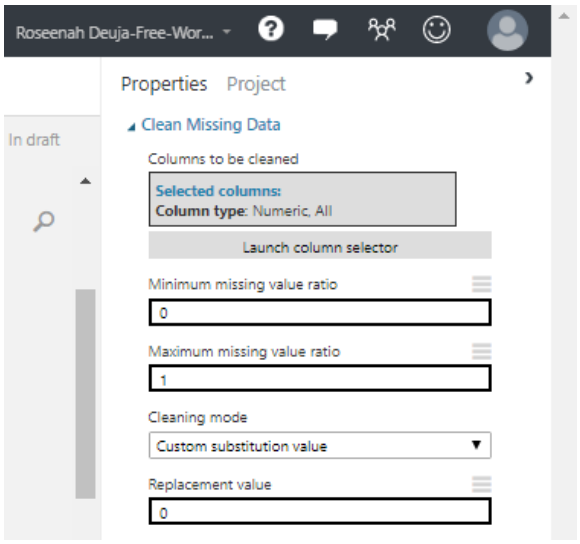

Fig. Handling of missing values (For Numeric)

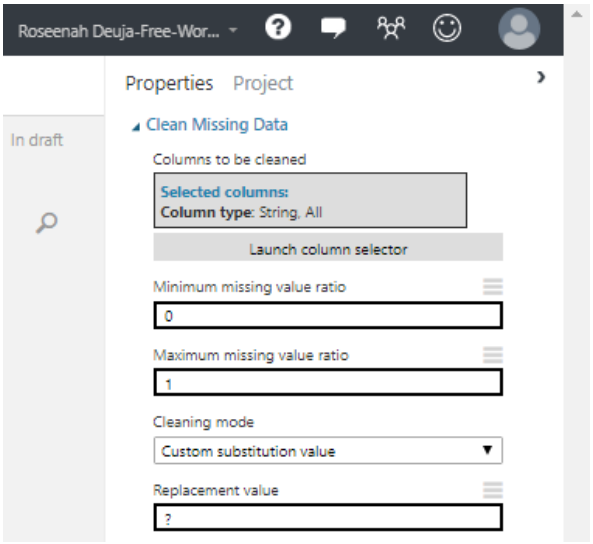

Fig. Handling of missing values (For Categorical)

edx Username: roseenah1

2.2 Data Transformation

Since the different attributes of the data fall in different range of values, it is necessary to transform them to a consistent form. For this Normalization was carried out.

For normalization, Z-Score normalization was used. It is obtained mathematically as:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

Fig. Z-score formula

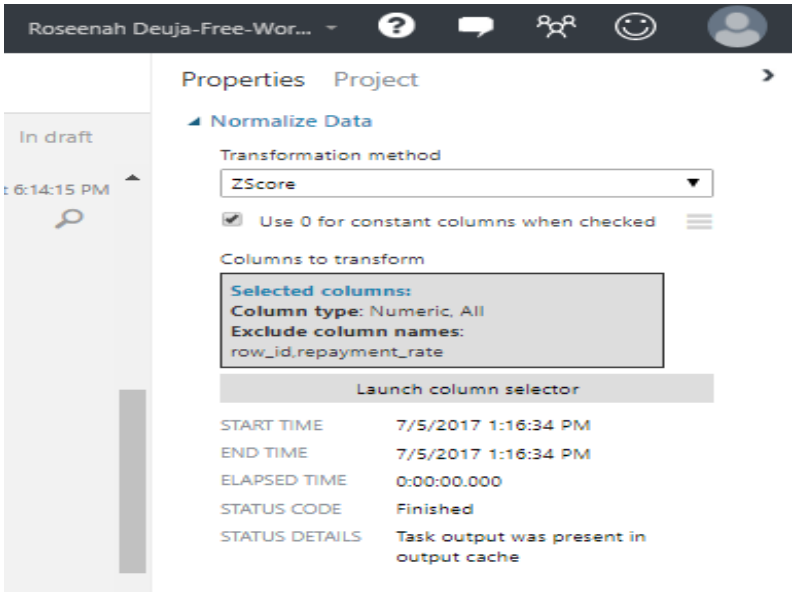

Fig. Normalization by z-score

2.3 Data Reduction

The actions preformed under this task are:

a. Removing irrelevant attributes: attribute selection, searching the attribute space.
b. Principle component analysis (numerical attributes only): After the analysis, it was found that student_demographics_avg_family_income is the most influential features amongst all features in the dataset. This step helps to determine a lower dimensional space that can best represent the data
c. Reducing the number of attribute values: Similarly, the other attributes were also chosen out of the given set. The 83 attributes chosen for analysis are:
    1. admissions__act_scores_midpoint_cumulative
    2. admissions__act_scores_midpoint_english

edx Username: roseenah1

3. admissions__act_scores_midpoint_math
4. admissions__act_scores_midpoint_writing
5. admissions__admission_rate_overall
6. admissions__sat_scores_average_overall
7. admissions__sat_scores_midpoint_critical_reading
8. admissions__sat_scores_midpoint_math
9. admissions__sat_scores_midpoint_writing
10. aid__cumulative_debt_number
11. aid__federal_loan_rate
12. aid__loan_principal
13. aid__median_debt_completers_overall
14. aid__median_debt_dependent_students
15. aid__median_debt_income_0_30000
16. aid__median_debt_income_30001_75000
17. aid__median_debt_income_greater_than_75000
18. aid__median_debt_independent_students
19. aid__median_debt_no_pell_grant
20. aid__median_debt_noncompleters
21. aid__median_debt_number_completers
22. aid__median_debt_number_dependent_students
23. aid__median_debt_number_income_0_30000
24. aid__median_debt_number_income_30001_75000
25. aid__median_debt_number_income_greater_than_75000
26. aid__median_debt_number_independent_students
27. aid__median_debt_number_overall
28. aid__students_with_any_loan
29. completion__completion_cohort_less_than_4yr_100nt
30. completion__completion_cohort_less_than_4yr_150nt
31. completion__completion_cohort_less_than_4yr_150nt_pooled
32. completion__completion_rate_4yr_100nt
33. completion__completion_rate_less_than_4yr_100nt
34. completion__completion_rate_less_than_4yr_150nt
35. completion__completion_rate_less_than_4yr_150nt_pooled
36. completion__transfer_rate_4yr_full_time
37. completion__transfer_rate_less_than_4yr_full_time
38. cost__avg_net_price_private
39. cost__avg_net_price_public
40. cost__title_iv_private_all
41. cost__title_iv_public_all
42. cost__tuition_in_state
43. cost__tuition_out_of_state
44. school__carnegie_size_setting
45. school__carnegie_undergrad
46. school__degrees_awarded_highest
47. school__degrees_awarded_predominant
48. school__faculty_salary
49. school__ft_faculty_rate
50. school__instructional_expenditure_per_fte
51. school__locale
52. school__main_campus
53. school__men_only
54. school__minority_serving_aanipi
55. school__minority_serving_annh
56. school__minority_serving_hispanic
57. school__minority_serving_historically_black
58. school__minority_serving_nant
59. school__minority_serving_predominantly_black
60. school__minority_serving_tribal
61. school__online_only
62. school__ownership
63. school__religious_affiliation

64. school__state
65. school__tuition_revenue_per_fte
66. school__women_only
67. student__demographics_age_entry
68. student__demographics_avg_family_income
69. student__demographics_avg_family_income_independents
70. student__demographics_median_family_income
71. student__family_income_dependent_students
72. student__family_income_independent_students
73. student__family_income_overall
74. student__parents_education_level
75. student__part_time_share
76. student__retention_rate_four_year_full_time
77. student__retention_rate_four_year_part_time
78. student__retention_rate_lt_four_year_full_time
79. student__retention_rate_lt_four_year_part_time
80. student__size
81. student__students_with_pell_grant
82. student__valid_dependency_status
83. repayment_rate

3. Linear Regression Model

Since our data mining problem is a predictive one, a prediction model was created for the data. The data to be predicted i.e. repayment_rate is a continuous attribute hence, regression was found to be the best technique to handle continuous data. Linear regression was chosen as the technique. The output of the model is the repayment_rate for the test values given.

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.

Given the representation is a linear equation, making predictions is as simple as solving the equation for a specific set of inputs.

Let's make this concrete with an example. Imagine we are predicting weight (y) from height (x). Our linear regression model representation for this problem would be:

$$y = B0 + B1 * x1$$

or

$$weight = B0 + B1 * height$$

Where B0 is the bias coefficient and B1 is the coefficient for the height column. We use a learning technique to find a good set of coefficient values. Once found, we can plug in different height values to predict the weight.

In case of the model, the repayment_rate is given as the linear combination of the chosen 83 attributes as:

repayment_rate =-0.7232 * admissions__act_scores_midpoint_english + 1.1902 * admissions__act_scores_midpoint_math + 0.5365 * admissions__admission_rate_overall + 0.719 * admissions__sat_scores_average_overall -0.4737 * admissions__sat_scores_midpoint_math +…and so on.

edx Username: roseenah1

In context of Azure ML Studio, the model is a simple linear regression. There are a few types of linear regression modules, the one named “Linear Regression”, found under “Regression” was used.

The "Learning Rate" parameter allows us to set how much difference we see from tree to tree. This describes quite well as "the learning rate determines how fast or slow the learner converges on the optimal solution. If the step size is too big, it might overshoot the optimal solution. If the step size is too small, training takes longer to converge on the best solution." The "Random Number Seed" parameter allows us to create reproducible results for presentation/demonstration purposes. The model was tested before selecting it as final Model. For testing, Model was trained with 60% of the combination of train_values and train_labels, and tested with the remaining 40% yields following results after the split:

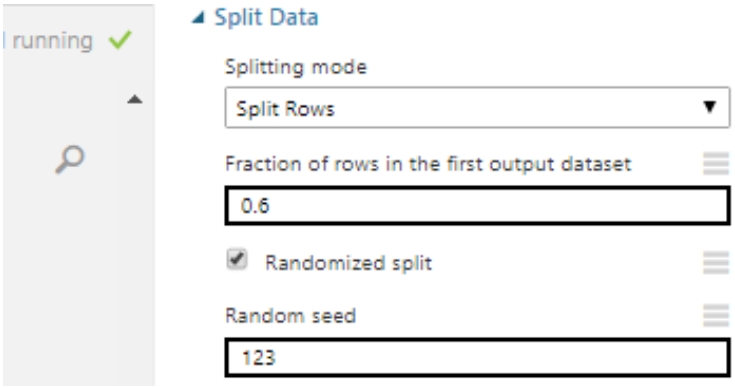

Fig. Splitting of data

The data after Score Label step is:

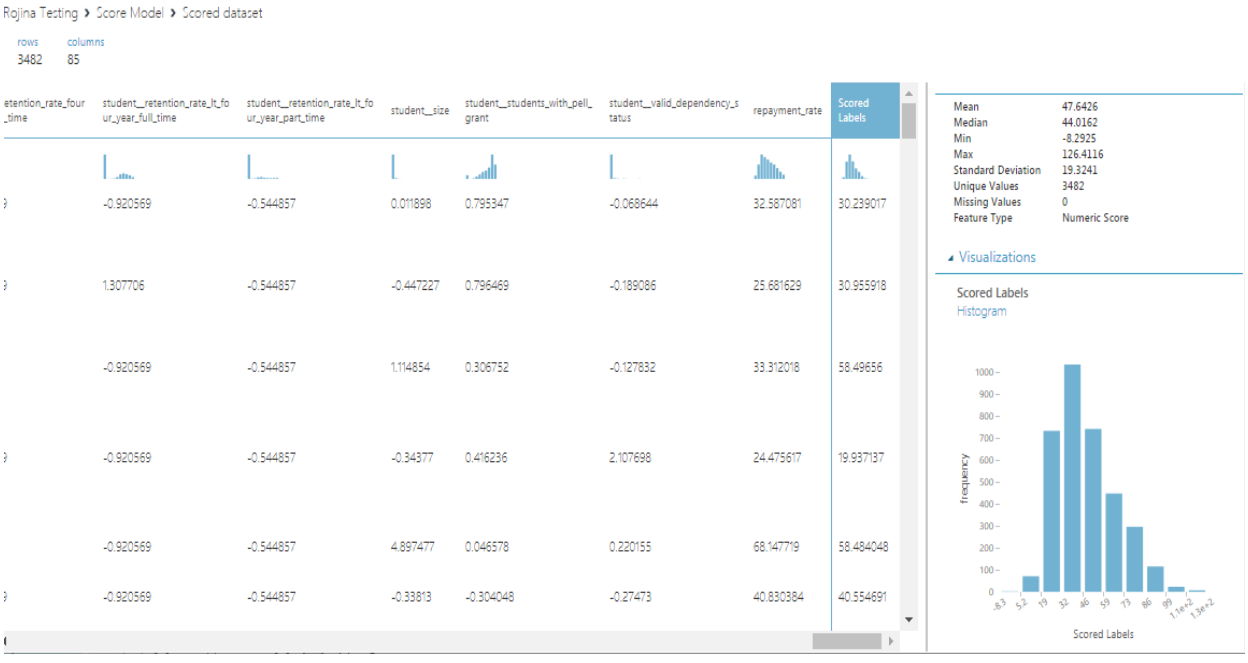

Fig. Visualization of scored dataset

edx Username: roseenah1

Mean Absolute Error: 5.528142

Root Mean Squared Error: 7.421821

Relative Absolute Error: 0.352459

Relative Squared Error: 0.120887

Coefficient of Determination: 0.903207

The module with the highest value for Coefficient of Determination is taken as best. And, here the Coefficient of Determination is near to 0.1 so it is taken as significant. Hence, the model was chosen as final regression model.

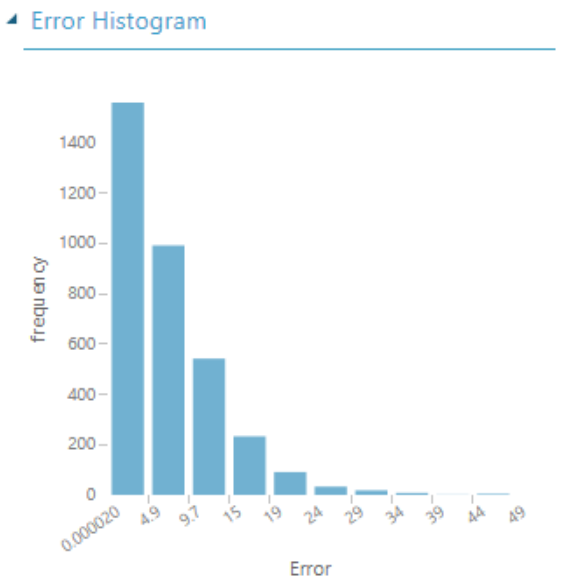

Fig. Error histogram of the linear regression model

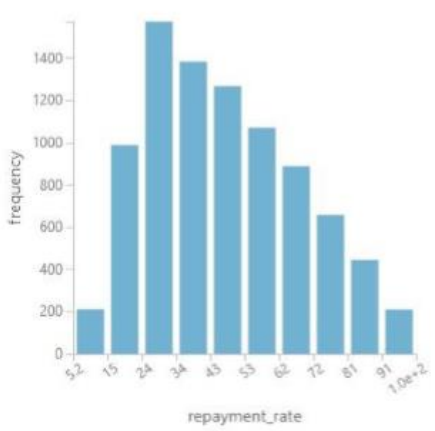

Fig.Repayment_rate acutal data histogram

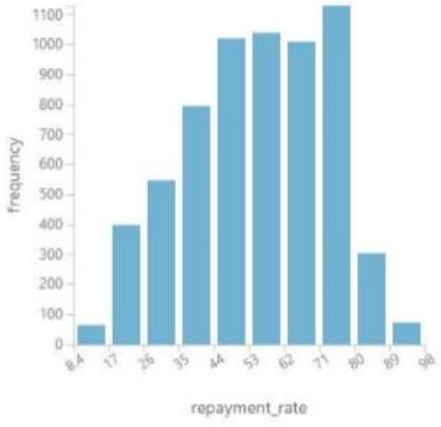

Fig. Repayment_rate predicted histogram

edx Username: roseenah1

4. Conclusion

Thus, going through the critical data mining techniques, the required value with Root Mean Squared Error (RMSE) of 9.492911 was predicted. The linear regression model was best suited for the data, upon individual evaluation and the error obtained is acceptable. Hence, the model was concluded with the above procedures.