UGA Data Science Competition 2021

Ayush Kumar, Chloe Phelps, Faisal Hossain, Nicholas Sung

April 15, 2021

Contents

1 Exploratory Data Analysis and Preprocessing

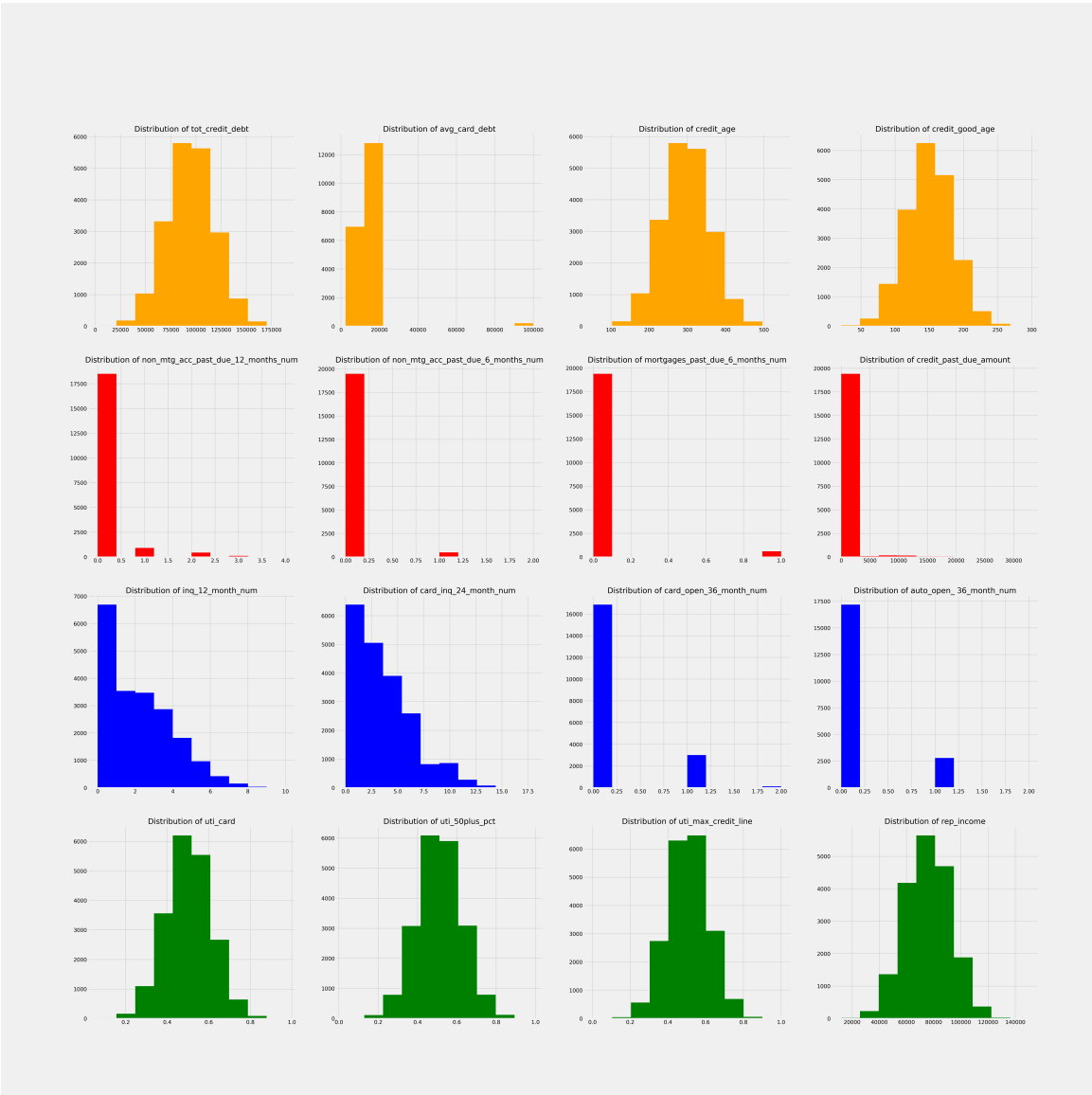
1.1 Summary Statistics for Continuous Variables

We begin our exploratory data analysis by looking into the data, and determining the type of the data for each column. For continuous numerical data, we want to visualize the distribution using histograms as well as taking a look at the following summary statistics: mean, standard deviation, the minimum and the maximum. For the categorical data we want to take a look at the possible categories, and the frequency of each category.

Using pandas and the information sheet here are the numerical variables:

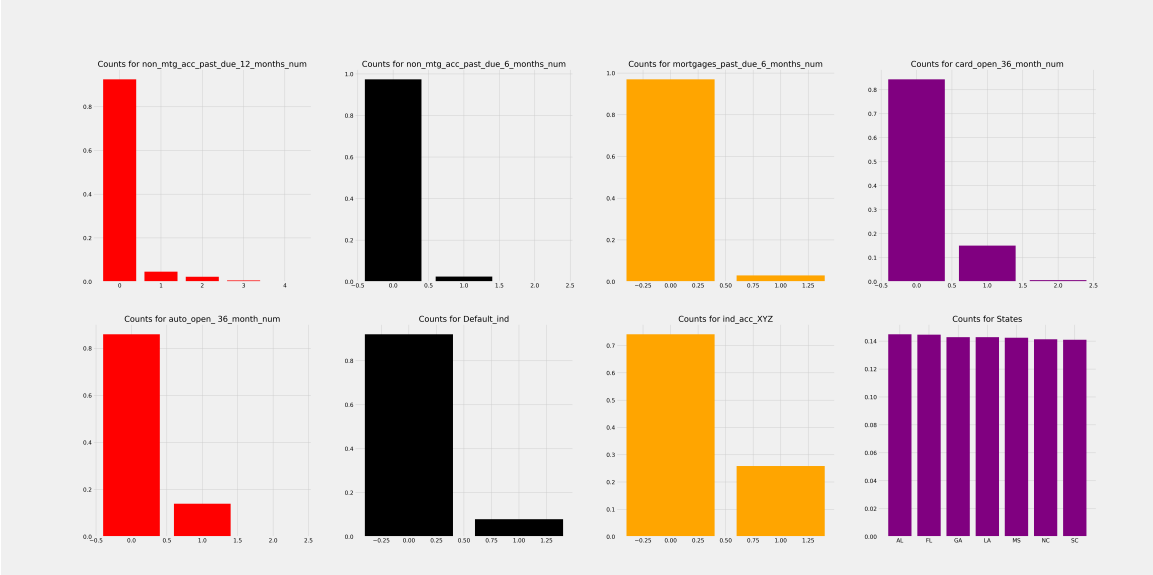
	mean	std	min	max
tot_credit_debt	94563.702530	23546.443862	2367.430000	188890.960000
avg_card_debt	14088.235475	9314.495936	2363.120000	99999.000000
credit_age	296.697000	61.711702	54.000000	545.000000
credit_good_age	149.771750	34.016476	21.000000	296.000000
non_mtg_acc_past_due_12_months_num	0.111350	0.433890	0.000000	4.000000
non_mtg_acc_past_due_6_months_num	0.027400	0.171903	0.000000	2.000000
mortgages_past_due_6_months_num	0.030200	0.171142	0.000000	1.000000
credit_past_due_amount	329.287867	2073.899357	0.000000	32662.980000
inq_12_month_num	1.762700	1.740816	0.000000	10.000000
card_inq_24_month_num	3.409600	2.926697	0.000000	18.000000
card_open_36_month_num	0.163050	0.386099	0.000000	2.000000
auto_open_ 36_month_num	0.141000	0.349607	0.000000	2.000000
uti_card	0.503157	0.109354	0.065120	0.969289
uti_50plus_pct	0.511007	0.113456	0.033749	0.988964
uti_max_credit_line	0.507629	0.108624	0.005174	1.000000
rep_income	75499.511666	16361.955146	12000.000000	150000.000000

By looking at the descriptive statistics we can see if there are any major outliers, and determine how we can use these variables in our data. We see that minimum and maximum values for many of the continuous variable are very extreme for Total Credit Debt, Average Credit Debt, and Credit past due amount. We also see that many variables seem to have very similar distributions, which may be a problem in regards to multi-collinearity. To investigate this we can create a correlation matrix and re-scale certain variables to try to preserve information, while combating multi-collinearity.



Looking at the histogram gives us a better picture of the distributions of the variables because it allows us to visualize the shapes of the variables. We can see that the outliers for Average Credit Debt are seriously impacting the distribution in comparison to Total Credit Debt. This visualization also allows us to look at some variables which by their description seem to be numerical, but display behavior that is more characteristic of categorical variables. Variables like the number of mortgages past due or non-mortgages past due would be better suited to being dummy variables rather than continuous ones.

1.2 Summary Statistics for Categorical Variables



2 Logistic Regression Model

3 Feed Forward Neural Network Model

4 Model Comparison & Evaluation

5 Future Decision Making

5.1 Previous Customers & Bias

5.2 Explaining Model Decisions