Term Project

Saravanan Janarthanan

DSC530 Data Exploration and Analysis, Bellevue University

Professor: Fadi Alsaleem

Week 12 – Assignment due date: 3/2/2024

Data Exploration and Analysis Project Summary

Summarize the following: Statistical/Hypothetical Question

- Outcome of your EDA
- What do you feel was missed during the analysis?
- Were there any variables you felt could have helped in the analysis?
- Were there any assumptions made you felt were incorrect?
- What challenges did you face, what did you not fully understand?

Outcome of your EDA

- The Exploratory Data Analysis (EDA) results rejected the null hypothesis suggesting no relationship between the response variable, loan_status (determining loan approval), and the predictor variables.
- It was observed that out of 11 predictor variables (comprising 7 continuous and 4 categorical variables), all exhibited statistical significance concerning the response variable.
- Eliminating duplicates ensured that the model remains robust and avoids overfitting.
- Trimming outliers from three variables, namely person_age, person_income, and person_emp_length, aided in reducing distribution tail and skewness, especially due to the presence of non-factual values.
- Examination of the five-point summary, histograms, box plots, and density plots in relation to loan_status provided insights into the predictive variable effectiveness.
- Probability Mass Function (PMF) analysis conducted on two datasets—applicants of all ages and mortgage home owners—indicated a consistent probability distribution within subgroups
- Cumulative Distribution Function (CDF) analysis helped identify the age distribution of applicants, revealing that 75% were aged below 30 years and 95% below 40 years

Data Exploration and Analysis Project Summary

- Analytical distribution revealed that the selected person_age variable fits both empirical and analytical distributions when compared to the standard normal distribution
- Scatter plots and correlation coefficients (Pearson and Spearman) suggested a weak or small positive correlation between person_age and loan_amount, yet the p-value established this relationship as statistically significant.
- Permutation of data between groups reaffirmed the absence of a difference in credit-on-file data distribution between the two groups, signifying their statistical significance
- Regression analysis provided confirmation of the statistically significant relationship between multiple predictor or independent variables and the response variable, thus rejecting the null hypothesis of no relationship between dependent and independent variables

What do you feel was missed during the analysis?

- Felt that the depth of correlation among different sub-groups was overlooked, hindering a comprehensive understanding for improved predictions.
- Missed multivariate analysis entails examining relationships among multiple variables concurrently.
- Feature scaling involving standardizing numerical variables to ensure they operate within a comparable range, thereby preventing any individual variable from exerting undue influence over the analysis.

.

Were there any variables you felt could have helped in the analysis?

The following variables contributed to the analysis:

- person_age - Age of the person
- loan_amnt - Loan amount

Data Exploration and Analysis Project Summary

- person_home_ownership - Home ownership - whether person owns home or mortgage or rented
- cb_person_default_on_file - Historical default (previous default history from Bureau)
- loan_percent_income - Percent income - % of income represented by loan

Were there any assumptions made you felt were incorrect?

It was initially assumed that the dataset contained only a few factual errors. However, upon closer examination, it became apparent that there were numerous factual errors, making remediation challenging. For instance, one notable error was the exceptionally high loan-to-income ratio, which reached 83%

What challenges did you face, what did you not fully understand?

By employing permutation hypothesis testing and comparing the means across shuffled data groups, an attempt was made to ascertain whether there would be an effect or difference between the groups. However, it remained uncertain how this approach contributed to the EDA analysis if the samples were mixed