Case Study

Lending Club Data EDA

Sarun Natarajan

Data Information

- Shape of the dataset (39717, 111)
- Many columns irrelevant for EDA
- There were many columns with null values
- Needed cleanup of values to make it useful for insight generation
- Need to remove everything on customer behaviour - as this information will not be available when a new applicant approach for loan
- Only applicant demographics and loan attributes to be considered
- All current loan status to be removed as they are in progress loan accounts and can default in future

Data Cleanup activities done

- Removed current loan status records
- Dropped columns which were all empty
- Transformed interest rate column
- Numerical representation of grades column
- emp_length column cleaned
- Removed customer behaviour attribute columns
- Removed outliers from annual income column

Analysis Done

- Univariate Analysis
- Bivariate Analysis
- Segmented univariate analysis
- Correlation

Final Conclusion from EDA

- - Large number of loan applicants has less than 10 years of experience
- - Annual income of the applicants is bit right skewed as majority of them earn below 80k.
- - DTI provides a perfect bell curve representing normal distribution
- - Every year the number of loan application has increased
- - 14.8% of the loan applicant has defaulted
- - Loan defaulters who listed the purpose as debt consolidation has defaulted the most
- - Applicants with less experience do tend to default more, because if the count of less experiences are stacked together, it will be clearly taller compared to 10+ years
- - Grade B, C and D applicants tend to default the most
- - Lower the grade, higher the interest rate