Jaineel Upadhyay

Goal: To analyze data of customers who Lending Club have issued loans to and try to decipher how much loan amount people take on as debt and who are more likely to default on their loans.

Analysis:

* Summarized Information of the data:
  + The average amount of loan given by Lending Club evaluated to $12,435.17, with the highest amount being $35,000 and lowest amount being $1,000.
  + The average net annual income of a customer of Lending Club was $57,013.98, with the highest and an outlier being 7135346.00 and the minimum being -$14,540, indicating someone having more debt than income.
  + The percentage of gross income amounting to loan was at average, 0.2 with the highest being 0.45, and lowest being 0, meaning customers never got more debt than they could afford.
  + However, the average debt to gross monthly income ratio was found to be 16.90 on average, with 34.99 being the highest and 0 being the lowest. The average gross annual income = $57,013.98 + (0.2 \* $57,013.98) = $68,416.776, which is the average percentage of gross income taken as loan and the annual net income, ignoring taxes and other debts.
  + Dividing this value by 12, we get the average monthly income of $5,701.398. Again, assuming the average debt-to-income (gross income) being 16.90% . Therefore,
  + Average debt / gross monthly income = 16.90% => $963.53. Therefore, the gross income, ignoring taxes, is at around $6,664.928, or $79,979.136 per annum.
  + Around 43% of Lending Club’s members don’t have a property and live on rent.
  + Around 13% have defaulted on their loan(s).
* As seen in scatter plot, most incomes are under $60,000 and loans being pretty balanced in terms of amounts.
* Highest counts out loan amounts come from ~$10,000 to ~$12,500. Also, interestingly, there seems to be more loans of $30,000 and lesser than $35,000 than loans of amounts $25,000 and less than $30,000.
* The boxplot of the adjusted\_annual\_inc with loan\_amount seems to suggest the for the loan amounts to be around $10,500.
* Also, the correlations between the fields were calculated and stored in a table. The highest correlation appeared to be between percentage of loan vs loan amount, indicating a good and directly proportional relationship between both the fields. This also indicates, that as the loan amount increases, the percentage of gross income it affects also increases, indicating people taking more debt despite non-rising incomes.
* Second highest appeared to be between adjusted annual income and loan\_amount with the correlation coefficient being 0.349. Despite a weaker linear relation, this basically provides a considerable argument for correlation between loan amount and the income earned, mostly gross income.

Machine Learning:

* Multiple Regressions:
  + The fields used for multiple regressions are pct\_loan\_income, adjusted\_annual\_inc (fields with stronger correlation with loan\_amnt), as such loan\_amnt being the output (y value).
  + The intercept (b) coefficient comes out at -1.672e+0.3 with x1 coming at 4.987e+04 and x2 coming at 7.153e-02.
  + The r-squared or coefficient of determination value came out to be 0.572, indicating a moderate but good relationship between the regression model and the data points.
  + This provides very strong support to the loan amount-income argument.
  + Regression Equation : y = -1672.426 + (49876.66)x2 + (0.071539)x1
* Naïve Bayes Model:
  + Many loan defaults appear to happen with loan amounts around $10,000 as seen by the frequency polygons plot.
  + Similarly, majority of loan defaults tend to happen with customers having approximately 10 open credit accounts.
  + Running the Naïve Bayes model, interesting results tend to appear.
  + The accuracy of the Naïve Bayes Classifier was 0.8743, indicating an 87.43 % chance of accurately predicting when a person has defaulted on their loans or not.
  + However, the Cohen’s Kappa value appears to be -2e-04 (Minor value differences), indicating bad classification accuracy, or no level of agreement. It also means 0-4% of data is reliable, indicating faulty information. Information from ([1](https://towardsdatascience.com/interpretation-of-kappa-values-2acd1ca7b18f))([2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/)).
  + The classifier predicted 19,332 False outputs and 2770 True outputs as False. Total 19,341 referenced False outputs.
  + The classifier predicted 9 False referenced cases and 1 True case as True out of 2771 referenced True outputs.
  + Summarizing this model practice, a graph was plotted to show the importance of predictor variables in predicting the output of a person defaulting on loan or not.
  + It showed the total net household income or adjusted\_annual\_inc, with most importance in predicting the output of loan defaulted. With debt-to-income ratio and the percent of gross income equivalent to loan amount as having some significance.
  + The least significant ones were the number of public recorded bankruptcy, the number of open credit accounts and number of accounts 120 days or more past due, with the other variables having minor significance.

Therefore, forgoing Kappa value, there appears to be a situation visible with regards to loan defaults. Already established using multiple regression and correlation coefficients, that loan amounts (debt) increase along with percentage of income as loan. We can now establish that there appears to be lesser income rise (weak correlation) and increasing debt-to-income ratio. Coupling this with the Naïve Bayes model, where we see the net annual income, the debt-to-income ratio and the pct\_loan\_income being major predictors of loan defaults, we can conclude that the major contributor of factor in loan defaults is percentage of gross as income and adjusted annual income, as loan amounts (debt) tend to rise, incomes don’t rise quicker and % of income as loan increases, also increasing debt-to-income ratio.