STOUT internship Assignment Case\_study1 Name – Gautami Lad

* **Describe the dataset and any issues with it.**
* This dataset is about loans already made by customers through lending club
* I developed a model based on the interest rates offered to those customers based on their income, credibility.
* This dataset have 10000 rows and 55 columns
* The independent variables and their description can be found on the below link

<https://www.openintro.org/data/index.php?data=loans_full_schema>

* The interest rate is dependent variable for this dataset which we have to predict using regression model
* The model developed will predict the interest rate for the customer if we have all the information for independent variables
* The dataset contains NA and zero values. The most of the NA values are valid e.g if for a column annual\_income\_joint the NA present in the dataset implies that in the loan application there is no joint holder income that is to be considered for the loan.
* Issue- The column debt to income is important but from the data it doesn't look like they are ratios. They are more like percentage e..g if debt to income has value 18.01 then it says 18% is debt for the total income

**Data cleaning –**

* Earlier I used to work in a bank as finance officer in the loans department
* I used my business acumen to eliminate the columns that are irrelevant to predict the interest rates and also created some new columns by using existing columns
* In my first screening of dataset i removed the following columns as these are irrelevant for the model we are trying to build emp\_title , emp\_length, state, application\_type, loan\_status, issue\_month, initial\_listing\_status, disbursement\_method.
* I added column annual income (which is primary applicant income) with column annual income joint and created new column total annual income as it will give us the total income for the household. Because at the end the total income from all sources is what matters. Then I removed the primary applicant annual income column and joint annual income column
* It is observed from the dataset that the column verification income joint has only 1495 rows (out of 10000) and >90% of them match the verification status with primary holder so I dropped the column 'verification income joint'
* I replaced the NaNs in the column debt\_to\_income\_joint with zero. I did not impute those values with other values as these are not missing values. It is possible the joint holder who do not have income would not any debt that is zero(none). It is also possible for someone to not have any debt.
* The column earliest\_credit\_line have year values. I subtracted the loan issue year(2018) with earliest credit line year so we would get how much old the earliest credit line is. E.g if earliest credit line was in 2001 then 2018-2001 will tell us that the credit line is 17 years old.
* I dropped the column total credit lines as the column open(active) credit line is more important than total credit lines.
* I calculated the percentage credit utilized and created a new column percentage\_credit\_utilized from total credit limit and available credit limit and removed columns total\_credit\_limit, total\_credit\_utilized
* I removed columns num\_accounts\_120d\_past\_due, num\_accounts\_30d\_past\_due as they had more than 95% zeros.
* Removed the column current\_accounts\_delinq as it has 9999 zero values
* I created a new column by calculating the percent number cc carrying balance accounts by dividing number of cc carrying balance and num total cc account and multiplying by 100 as percentage of number of cc carrying balance is important rather than all the three columns for total and active cc num account and removed the respective columns
* Removed the columns stating balance, principal or interest paid as these column values are not required for interest rate prediction. The columns are generated after the disbursement of loan
* Then converted the categorical data to numerical labels.
* Generate a minimum of 5 unique visualizations using the data and write a brief description of your observations.
* I used correlation matrix plot, box and whisker plot, density plot, histogram and bar plot to visualize the data.
* The detailed explanations are in the .pdf and .pynb files
* Create a feature set and create a model which predicts *interest rate* using at least 2 algorithms. Describe any data cleansing that must be performed and analysis when examining the data.
* I used standardization to transform the attributes with gaussian distribution to standard gaussian distribution with a mean of 0 and standard deviation of 1.
* The standardization worked better for my random forest regressor model
* After standardization I used PCA (principal component analysis) which is a feature selection technique. The PCA technique compresses the datasset (i.e. reduces the number of columns) but keeping the variance of dataset.
* Visualize the test results and propose enhancements to the model, what would you do if you had more time. Also describe assumptions you made and your approach.
* I used random forest regressor to predict the interest rates. The R2 value I got is 0.84 which shows a good model
* The R2 score is calculated to see the correlation between the actual Y and predicted Y (interest rate) values.

**Assumptions** –

* The dataset is for already disbursed loans
* The model does not take into consideration if there are any special offers of interest rates for different emp-title, locations, time of year

**If I would have more time-**

* I could use only one model due to the time frame constraint
* The dataset is huge and requires high computation power and could not gather resources to run other regression models
* I failed to run polynomial regression for this dataset on my laptop
* I learnt that I need to learn how to utilize the cloud computing so that I would not face this problem in the future.

URL for both the case study

https://github.com/gautami91/Stout-Case-Study