TITLE: DDA FULL STACK 2021

Subtitle: A complete report of tasks, Exploratory Data Analysis, Model Building and Results

Author

- Amritha Subburayan

Due Date: 02/01/2021

School: University at Buffalo, Suny

Table of Contents

CASE STUDY 1:

Section1: Determining the issues in the Dataset

1) Describe the dataset and any issues with it.

The given dataset contains the list of all past loan applicants that is obtained from the lending club platform. The aim is to predict the interest rate based on the past loan details which can help any individual to analyze for how much interest one can lend and borrow money respectively.
Below are the issues which was figured out when analyzing the dataset:

- Few columns have missing values:

  Missing values can be found in the following columns:

  1) emp_title

  2) emp_length

  3) annual_income_joint

  4) verification_income_joint

  5) debt_to_income_joint

  6) months_since_last_delinq

  7) months_since_90d_late

  8) months_since_last_credit_inquiry

  9) num_accounts_120d_past_due

- There occurs multi-Collinearity in the dataset, where many independent variables are highly correlated with each other:

  Multicollinearity can be found between these columns:

  1) installment and loan amount - 0.94

  2) balance and loan amount - 0.93

  3) annula income joint and total credit limit - 0.54

  4) Inquires last 12 m and months since last credit inq - 0.51

  5) total credit lines and open credit lines - 0.76

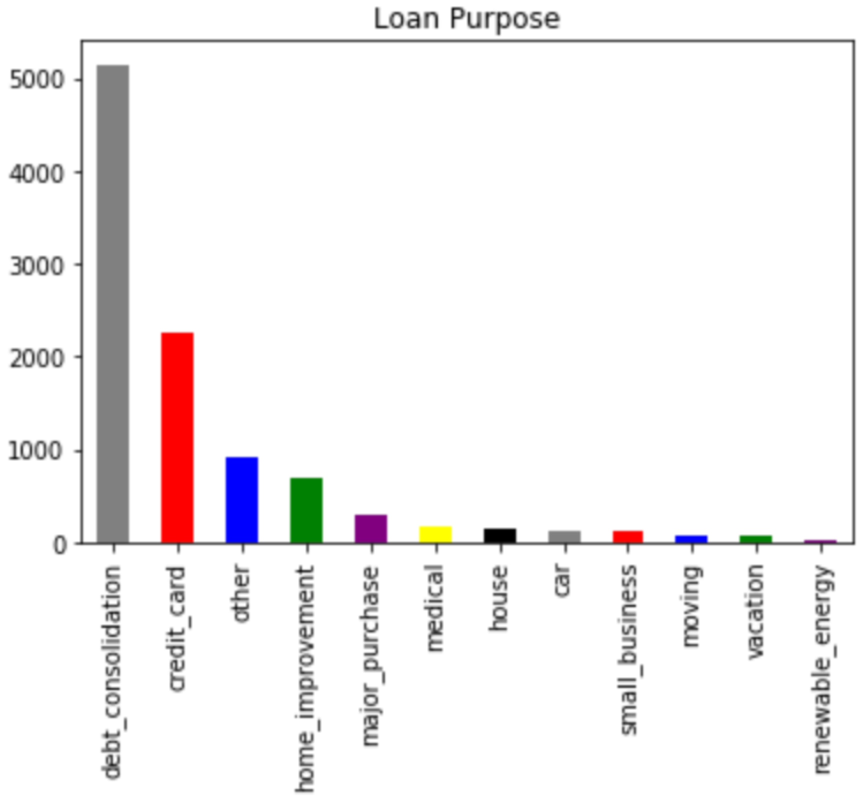
  6) num satisfactory acc and total credit lines - 0.75

7) total credit lines and num total cc accounts - 0.77

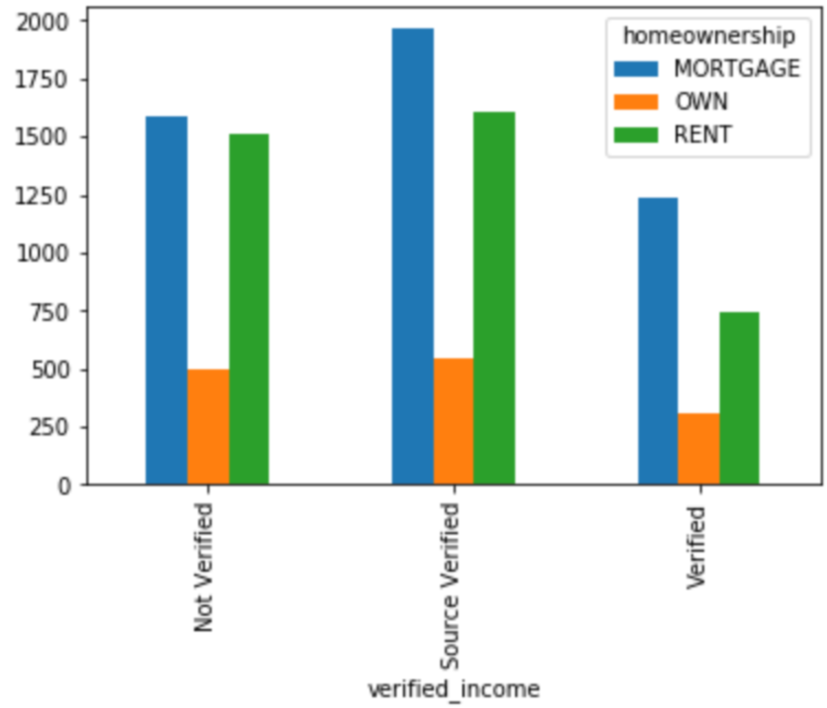
8) total credit lines and num open cc accounts - 0.62

Section2: Exploratory Data Analysis

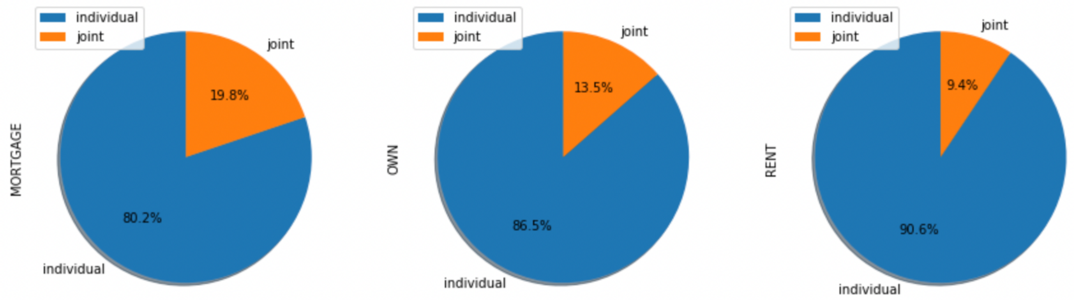
2) Generate a minimum of 5 unique visualizations using the data and write a brief description of your observations. Additionally, all attempts should be made to make the visualizations visually appealing



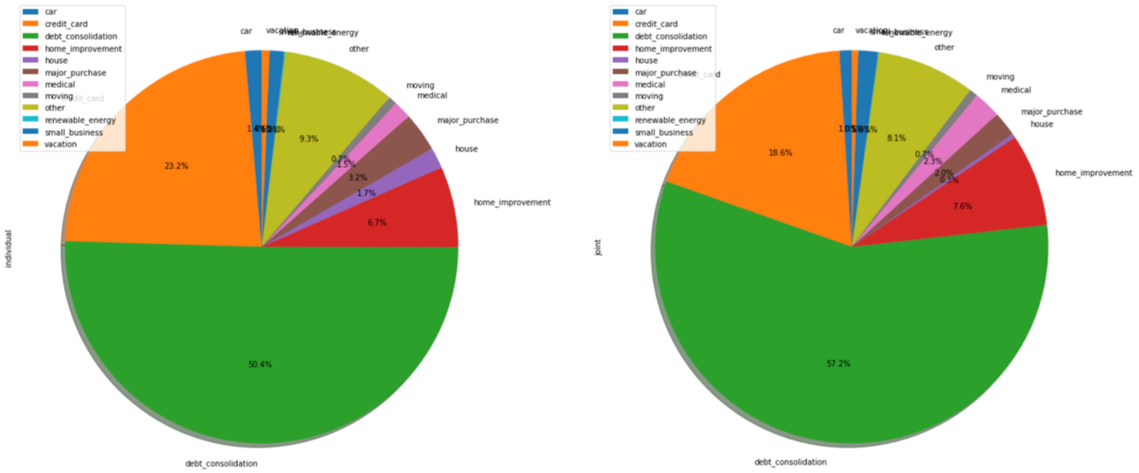
a) Above plot shows the loan purpose for which an individual borrows loan. From above plot, we can see that most of the people borrow loan to pay off their other small loan debts. Second highest is for credit card. From this we can say that people use more money on debt consolidation and credit card payments.



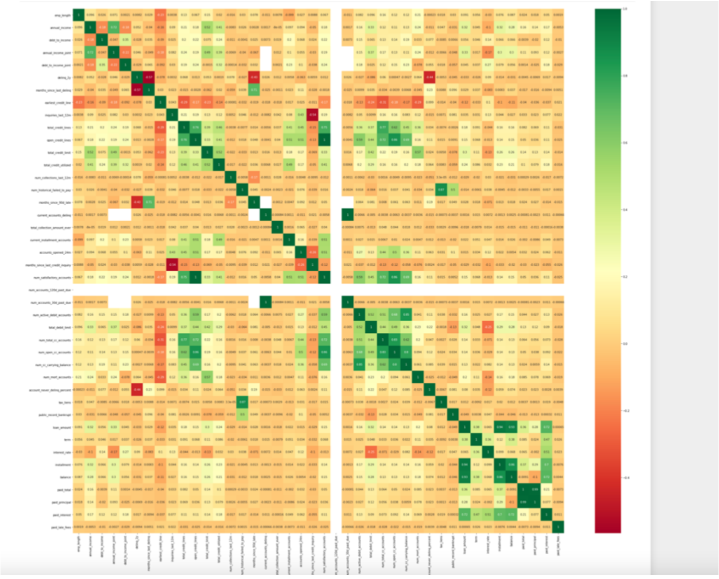
b) The above plot shows the details of accommodation of an individual. We could see that only a smaller number of people own a house which is verified, whereas many other lives in rental house and mortgage. From this we can be able to understand who can pay of the loan on time, so that we adjusted interest rate accordingly.



c) Above plot describes the relation between applicant type and homeownership. From this plot, we can interpret that an individual applicant who rents a house is more compared to others. Hence it is highly likely that individual who doesn’t have a house can borrow more money and ends up borrowing a loan to pay off the house and other personal debts.



Above plot shows the relation between application type and loan purpose. From this we can infer that joint applicant borrows more money for debt consolidation than an individual applicant. Here, we can see the loan purpose for vacation is greater for an individual.



The above heatmap shows the correlation between all the variables, using this diagram we can identify multicollinearity between the variables which can be used in data pre-processing to build a model with good accuracy.

Section3: Model Building and Evaluation

3) 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.

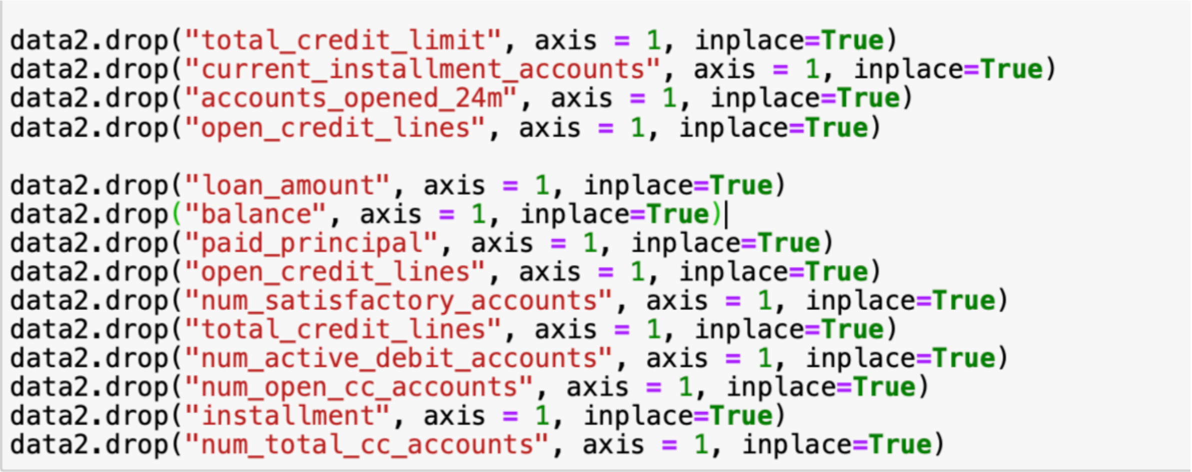
Data Cleansing:

As we could see that the given dataset has many numbers of missing values and outliers, we will follow few below techniques to pre-process the data before modelling.

1) Firstly, the summary for all the data columns is checked and the column which has 0 as the minimum value, first Quartile, second Quartile and also in 3rd Quartile are removed since those columns won't have great impact on providing useful information to predict the unseen data.
   1) delinq_2y
   2) num_collections_last_12m
   3) num_historical_failed_to_pay
   4) current_accounts_delinq
   5) total_collection_amount_ever
   6) num_accounts_120d_past_due
   7) num_accounts_30d_past_due
   8) tax_liens
   9) public_record_bankrupt
   10) paid_late_fees

2) Secondly, the correlation is checked between each independent variable using heatmap. Each one of the columns is removed by checking their correlation with target variable.

   Below are the column details, that are removed based on the correlation heatmap.



```
data2.drop("total_credit_limit", axis = 1, inplace=True)
data2.drop("current_installment_accounts", axis = 1, inplace=True)
data2.drop("accounts_opened_24m", axis = 1, inplace=True)
data2.drop("open_credit_lines", axis = 1, inplace=True)

data2.drop("loan_amount", axis = 1, inplace=True)
data2.drop("balance", axis = 1, inplace=True)
data2.drop("paid_principal", axis = 1, inplace=True)
data2.drop("open_credit_lines", axis = 1, inplace=True)
data2.drop("num_satisfactory_accounts", axis = 1, inplace=True)
data2.drop("total_credit_lines", axis = 1, inplace=True)
data2.drop("num_active_debit_accounts", axis = 1, inplace=True)
data2.drop("num_open_cc_accounts", axis = 1, inplace=True)
data2.drop("installment", axis = 1, inplace=True)
data2.drop("num_total_cc_accounts", axis = 1, inplace=True)
```

3) The dataset had more outliers, which can cause bias in predicting the test data. Hence the values which are far away are removed with the help of boxplot analyzation.
   Below are the columns that has more outliers compared to other columns,

   1) inquiries_last_12m
   2) total_credit_utilized
   3) current_installment_accounts
   4) accounts_opened_24m
   5) months_since_last_credit_inquiry
   6) total_debit_limit
   7) num_cc_carrying_balance
   8) num_mort_accounts
   9) paid_total
   10) paid_interest

4) Few unrelated columns are removed based on my intuition which will not be impactful in predicting the interest rate.
   1) State
   2) Initial_listing_status
   3) Disbursement_method

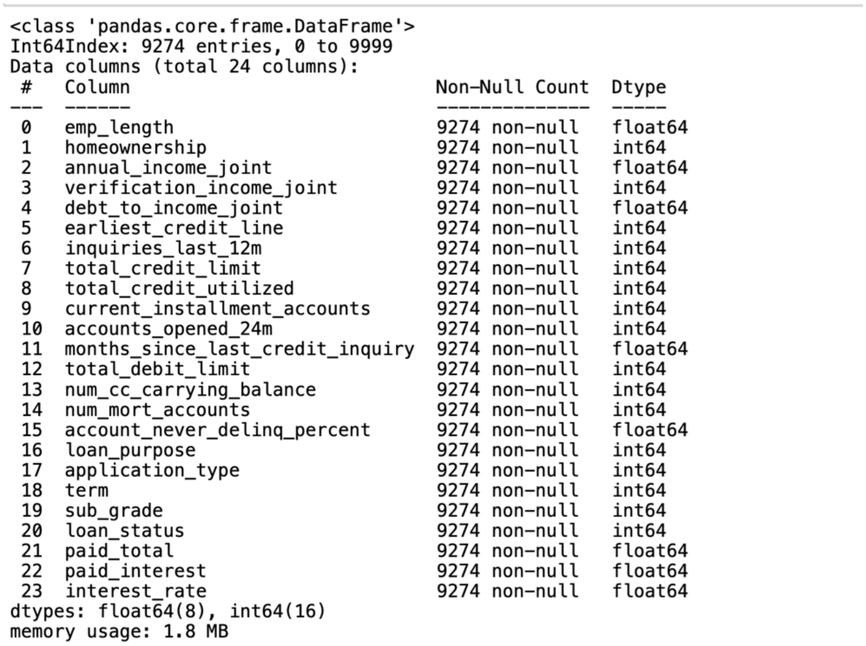
5) Below columns are merged together to a single column since it has same kind of information.
   1) combined debt income with debt income joint
   2) combined annual income with annual income joint
   3) combined verification income with verification income joint

6) For below columns, missing values are replaced with the mean.
   1) months_since_last_credit_inquiry
   2) emp_length

Feature Set and Model Building:

Feature Set:

Below are the final independent variables and dependent variables details which are used to build the regression model.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9274 entries, 0 to 9999
Data columns (total 24 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   emp_length                       9274 non-null   float64
 1   homeownership                    9274 non-null   int64
 2   annual_income_joint              9274 non-null   float64
 3   verification_income_joint        9274 non-null   int64
 4   debt_to_income_joint             9274 non-null   float64
 5   earliest_credit_line             9274 non-null   int64
 6   inquiries_last_12m               9274 non-null   int64
 7   total_credit_limit               9274 non-null   int64
 8   total_credit_utilized            9274 non-null   int64
 9   current_installment_accounts     9274 non-null   int64
 10  accounts_opened_24m              9274 non-null   int64
 11  months_since_last_credit_inquiry 9274 non-null   float64
 12  total_debit_limit                9274 non-null   int64
 13  num_cc_carrying_balance          9274 non-null   int64
 14  num_mort_accounts                9274 non-null   int64
 15  account_never_delinq_percent     9274 non-null   float64
 16  loan_purpose                     9274 non-null   int64
 17  application_type                 9274 non-null   int64
 18  term                             9274 non-null   int64
 19  sub_grade                        9274 non-null   int64
 20  loan_status                      9274 non-null   int64
 21  paid_total                       9274 non-null   float64
 22  paid_interest                    9274 non-null   float64
 23  interest_rate                    9274 non-null   float64
dtypes: float64(8), int64(16)
memory usage: 1.8 MB
```

Regression Model:

Support Vector Regression and Random Forest model are built for the given dataset and the accuracy is determined.

The performance of both the model is evaluated with help of R Square Value, MSE and RMSE values. The given dataset is split into training set (75% data) and test set (25% data). Below is the performance metrics details for both the model.

Model	Training Accuracy (R square)	Testing Accuracy (R Square)	Testing MSE	Testing MAE
Support Vector Regression	99.17%	98.25%	0.02	0.09
Random Forest	99.97%	99.95%	0.01	0.0

From above table, we can interpret that Random Forest performed well in both seen and unseen dataset with accuracy of 99.95%. Hence, we can say that Random Forest model is more generalizable to unseen data. Also, MSE value is less compared to Support Vector Regression.

Model	Training MSE	Testing MSE	Training MAE	Testing MAE
Support Vector Regression	0.01	0.02	0.06	0.09
Random Forest	0.0	0.0	0.0	0.01

It can be inferred that the Random Forest model didn't overfit, since it performed same way as like in training set. Since the error rate has remained same in both seen and unseen dataset, we can consider Random Forest as best model to predict the interest rate whereas in Support Vector Regression, error rate has increased in testing set.

Section4: Results Interpretation

4) 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.

The Random Forest model predicts the unseen data with great accuracy; hence we could consider this model has best fit in predicting interest rate. If I had more time, I would probably check for more outliers and negate them to understand how the model works fine for lesser amount of data. Also, analyzing and data cleaning play's crucial major role in building a model, so with more time, I can analyze the important features and its impact with target variables and understand how interest rate is calculated based on various factors.

Section6: Source Code and Description

Source Code/ File Name	Description
Stout Case Study 1.py Stout Case Study 1.ipynb	This program builds models, evaluates different models, visualizes and displays the result
Index.html WEB LINK URL: https://amritha29.github.io/Stout_DDA_FULL_STACK_21.github.io/ GIT HUB URL : https://github.com/Amritha29/Stout_DDA_FULL_STACK_21.github.io	This file holds all the visualization plots

Case Study #2

There is 1 dataset(csv) with 3 years' worth of customer orders. There are 4 columns in the csv dataset: index, CUSTOMER_EMAIL (unique identifier as hash), Net Revenue, and Year.

For each year we need the following information:

- Total revenue for the current year
  Ans = 31417495.030000016

- New Customer Revenue **e.g., new customers not present in previous year only**
  2017 – Net_Revenue = 31417495.030000016
  2016 - Net_Revenue = 25730943.59
  2015 - Net_Revenue = 29036749.189999994

- Existing Customer Growth. To calculate this, use the Revenue of existing customers for current year –(minus) Revenue of existing customers from the previous year
  Subtracting 2017 customer revenue with 2016 customer revenue
  Ans = 5686551.440000016

- Revenue lost from attrition

- Existing Customer Revenue Current Year
  2017 Customer Revenue = 31417495.030000016

- Existing Customer Revenue Prior Year
  2016 Customer Revenue = 25730943.59

- Total Customers Current Year
  - Year 2017 = 249987

- Total Customers Previous Year
  - Year 2016 = 204646

- New Customers - 399972
- Lost Customers – 202365

Source Code/ File Name	Description
Case Study 2.py Case Study 2.ipynb	This program contains code for all the evaluations
Index.html GIT HUB URL : https://github.com/Amritha29/Stout_DDA_FULL_STACK_21.github.io	Link for github repository