Assignment Classification & Comparing

---

1. ***Do some Exploratory Data Analysis (EDA) on the data and report what you find. What is the shape of the data set? How many rows and columns are in the data?***

By using "str (data)" I conclude that in this dataset there are 20,000 observations(rows) and 21 variables (columns). Using the "summary(data)" we are able to check the average amounts of each column to get an idea of the dataset. For example in tot_balance column we have an average of 107,439 meaning most users fall around this range of total balance. We can also see that most users earn between $140k - $189. Most have card balances of around $12k which is the mean.

2. ***Which variables contain missing values? How would you go about dealing with these missing values? Pick one method for handling missing values so that in the remainder of the assignment, you are working with data that has no missing values.***

The variables containing missing values are:

1- "Percentage of open credit cards with over 50% utilization " with **1958** missing values

2- "Annual income (self reported) with **1559** missing values

3- "Education level: with **1** missing value.

For numerical columns I replaced the missing values with the median so it wouldn't be affected by outliers and follow the general trend.

For the categorical since there was only one missing value, so I chose to replace it with the mode to keep the information without losing anything from it.

3. ***Are there any duplicates in the data? Data types?***

When running the commands in R there seems to be no duplicates in this data.

Assignment Classification & Comparing

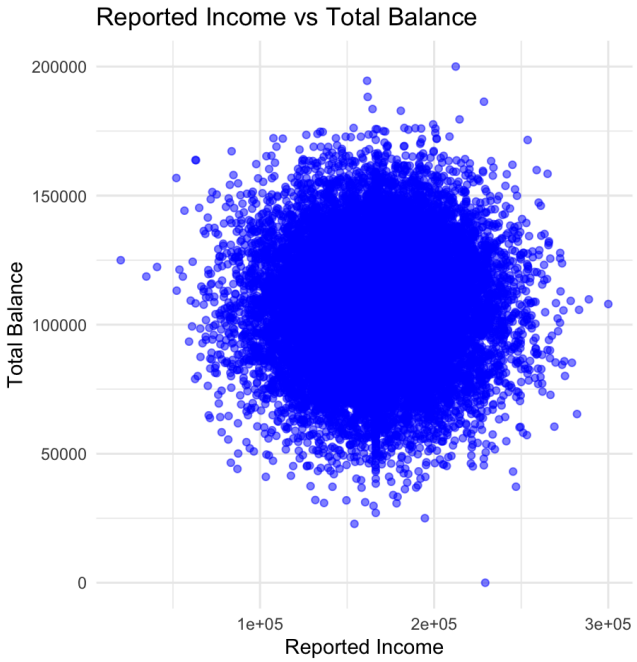
---

4. ***How would you handle duplicates in the data? How about variables that have been classified as the wrong data type (e.g., a date column classified as a text column)?***

To handle duplicates in a dataset it is most common to delete the duplicates to reduce redundancy, unless needed but that is rare.

With variables that have been classified as the wrong data first we identify the incorrect data types and convert to appropriate data types as needed.

5. ***Plot any two variables in the data and describe your graphs.***



This graph illustrates the relationship between the total debt owed by the applicant and their annual income. What I can conclude from this graph is that those with income between $100,000 - $200,000 have approximately between $50,000 - $150,000 in debt. Also that those with income between $100,000 - $200,000 have the majority of debt.

Assignment Classification & Comparing

---

6. ***Which education level is underrepresented in the data?***

The "other" category seems to be the lowest, but when talking about the ones defined, the "graduate" level is the most underrepresented. It has a count of 2406 and the proportion of 0.1203.

7. ***Are the classes in the default status variable ("Def_Ind") balanced in the data? If they are not balanced, suggest ways to correct this imbalance.***

The classes in the default statues seem to be imbalanced with the non default (0) outweighing the default (1).

0	1
Frequency 18,000	**Frequency** 2000
Proportion 0.9	**Proportion** 0.1

Since we are using the data to review credit card applications to determine which ones should be approved this would be more of a data centric approach. There are a couple different ways of correcting imbalance such as **undersampling(a lot of data)** which decreases the number of examples in the majority class and **oversampling(less data)** which increases the number of examples in the minority class. In this case I believe using the undersampling method works best to reduce the number of majority examples without losing much of the information.

8. ***How would you describe the distribution of "rep_income"? Is it skewed or approximately normal?***

The distribution of "rep_income" is approximately symmetrical. We know this because the skewness is -0.00474 this indicates it is slightly left skewed which we can also see in the histogram but OVERALL it is symmetrical.

Assignment Classification & Comparing

---

9. ***Group default status ("Def_Ind") by education level ("rep_education"). Which education level is more likely to default on loans?***

The education level most likely to default on loans is high school graduates with a rate of 11.57%

10. ***Does anything else stand out?***

Just that the other category has the lowest default rate and it would be beneficial to know what the other category consists of.

KNN MODEL:

- ***Confusion Matrix:*** From the confusion matrix we conclude that there are 3590 True Negatives(correctly predicted as non default) , 388 False Negatives(Actual defaulters (1) but predicted as non-defaulters(0)) , 10 False Positives(wrong prediction of defaulter) , and 12 True Positives(correctly predicted defaulters)
- ***Accuracy:*** The accuracy is 90.05% meaning that is how correct the model is.
- ***Precision:*** **12 / 10 + 12 = 12 / 22 = 0.545 Meaning 54.5% of predicted defaulters are actually defaulters.**
- ***Recall:*** **12 / 12+388 = 12 / 400=0.03 Meaning it only detected 3% of actual defaulters**

DECISION TREE:

- ***Confusion Matrix:*** From the confusion matrix we conclude that there are 3574(TN), 353(FP), 26(FN), and 47(TP).
- ***Accuracy:*** The accuracy is 90.52% meaning that is how correct the model is.
- ***Precision:*** **47 / 47 + 353 = 47 / 400 = 0.118 = 11.8%**
- ***Recall:*** **47 / 47 + 26 = 47 / 73 = 0.644 = 64.4%**

Assignment Classification & Comparing

---

4. Which features in the data are the most important for predicting default status ("Def_Ind")

The most important features seem to be avg_bal_cards 100.00%

num_acc_30d_past_due_12_months 69.96%

num_mortgage_currently_past_due 57.16%

num_acc_30d_past_due_6_months 55.05%

tot_amount_currently_past_due 54.29%

5. Which model performs better at classification? Why?

I believe the Decision Tree Model performs better because it has a slightly better accuracy, it is easier to interpret such as showing which features matter most.