Project Part II

Noah Keogh (njk6) S01264950

The dataset contains 150,000 entries, with 11 features for each entry [1]. These features are defined below:

1. **SeriousDlin2yrs:** Indicates whether a person has experienced 90 days past due delinquency or worse. This is a boolean parameter which is represented as 1 or 0 in the dataset. 1 being that the person has experienced serious delinquency or 0 meaning the person has not experienced serious delinquency.
2. **RevolvingUlizationOfUnsecuredLines:** Represents the total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits. This value is represented as a percentage and, therefore, ranges from 0 to 1.
3. **Age:** Age of the borrower in years. This is represented as an integer.
4. **NumberOfTime30-59DaysPastDueNotWorse:** Number of times that the borrower has been 30-59 days past due but no worse in the last 2 years. This is represented as an integer.
5. **DebtRatio:** Monthly debt payments, alimony, living costs divided by monthly gross income. This is represented as a percentage and, therefore, ranges from 0 to 1.
6. **MonthlyIncome:** This is the monthly income of the person. It is represented as a float in units of USD.
7. **NumberOfOpenCreditLinesAndLoans:** Number of open loans (installment like car lons or mortgages) and lines of credit (e.g. credit cards). This is represented as an integer.
8. **NumberOfTimes90DaysLate:** Number of times that the borrower has been 90 days or more past due. This is represented as an integer.
9. **NumberRealEstateLoansOrLines:** Number of mortgage and real estate loans including home equity lines of credit. This is represented as an integer value.
10. **NumberOfTime60-89DaysPastDueNotWorse:** Number of times borrower has been 60-89 days past due but no worse in the last 2 years. This is represented as an integer value.
11. **NumberOfDependents:** Number of dependents in family excluding themselves (spouse, children, etc.). This is represented as an integer value.

The SeriousDlin2yrs parameter is what can be used as the label for any supervised model that may be made. The reason for this is that the business objective is to predict whether a borrower is likely to have a serious default. Thus, if they are likely to default then we would not want to loan money to them, however, if they are not likely to default then we may consider loaning money to them.

After examining the data it appears that there are two features that possess missing values. The MonthlyIncome feature has 29,731 entries that are missing values. Since this is a significant number of entries ($\sim \frac{1}{5}$ of the data), I think it would be unwise to completely remove these entries. Instead, I plan to compute the median value for MonthlyIncome and fill these missing entries with that value. The other feature which has missing values is the NumberOfDependents. This attribute has 3,924 missing entries. This is a much smaller number of missing entries compared to MonthlyIncome. Though the number of missing entries vary, I will handle the missing entries in the same way. I will calculate the median value and then apply that value to the missing entries in the data. All the other features do not contain any missing values so no operations will have to be performed on those columns.

All of the features are numerically encoded values. Thus, I will not need to perform ordinal or one hot encoding transformations to the dataset. The only feature that may have required this type of transformation was the SeriousDlin2yrs feature, however, the dataset provided had already transformed the True or False statements into 1s and 0s, respectively.The only thing that may be transformed are the columns with the missing values (two mentioned previously). In order to fill those missing values I plan to use an imputer that will be utilizing the median value to fill the missing value. I do not plan to make any new features or columns, since each of the features are already well described, and I do not believe that I could add any additional features from the existing ones that may aid me.

One transformation or analysis I plan to perform regarding the features is dimensionality reduction. I plan to investigate how I may be able to use principal component analysis (PCA) in order to reduce the number of dimensions. This may aid me in making a better performing model. This is because reducing the number of dimensions may help to avoid the curse of dimensionality.

After performing some preliminary analysis of the data, it appears that the labels are imbalanced (see Figure 1). I, therefore, plan to balance the labels, as leaving the labels imbalanced may lead to the formation of biases in the final model. It appears that there are around 10,026 entries that have experienced serious delinquency and there are 139,974 entries that did not experience serious delinquency. I plan to randomly under sample the major (non-delinquency) group. to the point where the two classes are equal in number.

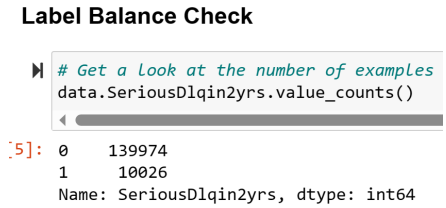

Figure 1: Depicts the number of entries for each label in the dataset.

In order to train and test the data I will make three splits on the data. I will perform a split where 80% of the data will be used for training, the other 20% of the data will then be used for testing. Instead of performing a random split, I plan to perform a stratified shuffle split so that why I will get a better representation of the overall dataset in both the training and testing datasets. This will help to ensure that there is a good distribution of each label type in the test and training data. If I performed a random split, either dataset may not capture the overall distribution of values for each of the features and could lead to issues with model performance. Furthermore, the labels with serious delinquency may not be split very well across the two datasets if a random split was performed. Thus, the stratified shuffle split should help to mitigate these issues.

In addition to a test and train split, I plan to split the training dataset into a training and validation set. I will perform different methods of performing such splits, such as K-fold. When deciding which model shows the most promise and should be further pursued, I will use nested cross validation in order to give a quantitative measure to the performance of each model. This quantitative measure will allow me to select the model that performed the best. Thus, I can spend more time on this model, optimizing its hyperparameters.

I have included a histogram showing the distribution of each feature (shown in Figure 2). It appears that in many features the distribution is not Gaussian. The age distribution appears to be Gaussian, and the number of open credit lines and loans appear to be close to a Gaussian distribution. The other distributions do not appear to be close to Gaussian. Instead the vast majority of people appear to have close to the same value for a given feature, with a very few number of people having much larger values, such as for income or number of real estate loans or lines. I will perform dimensionality reduction to see if I can identify outliers or identify trends in people's behavior. For example, maybe those with very high income also have a large number of real estate loans or unsecured lines of credit. PCA dimensionality reduction may be used to map and group the samples to determine if this is the case.

I have also included a correlation plot to get a sense of how correlated each of the features are to one another (shown in Figure 3). The correlation plot may be useful if feature selection is considered. Feature selection will most likely only be considered if PCA dimensionality reduction is unable to reduce the number of dimensions while maintaining a high explained variance ratio. If this is the case, feature selection will be considered. I would then investigate and decide that if two of the features are extremely correlated with one another I could omit one of features in the training of the model. This is because the two features would likely contain similar information since they are highly correlated. This would, therefore, reduce the number of dimensions in the dataset.

Based on the analysis performed above it is clear that I will need to balance the labels before model training is performed. I will also perform dimensionality reduction to avoid the curse of dimensionality and hopefully be able to train a model with better performance. I will cluster the data to see if I can identify certain financial behaviors or outliers in the dataset. These are the next steps required in order to obtain clean data that will be useful in training a model.

References

[1] Show me some credit. https://www.kaggle.com/competitions/GiveMeSomeCredit/overview.

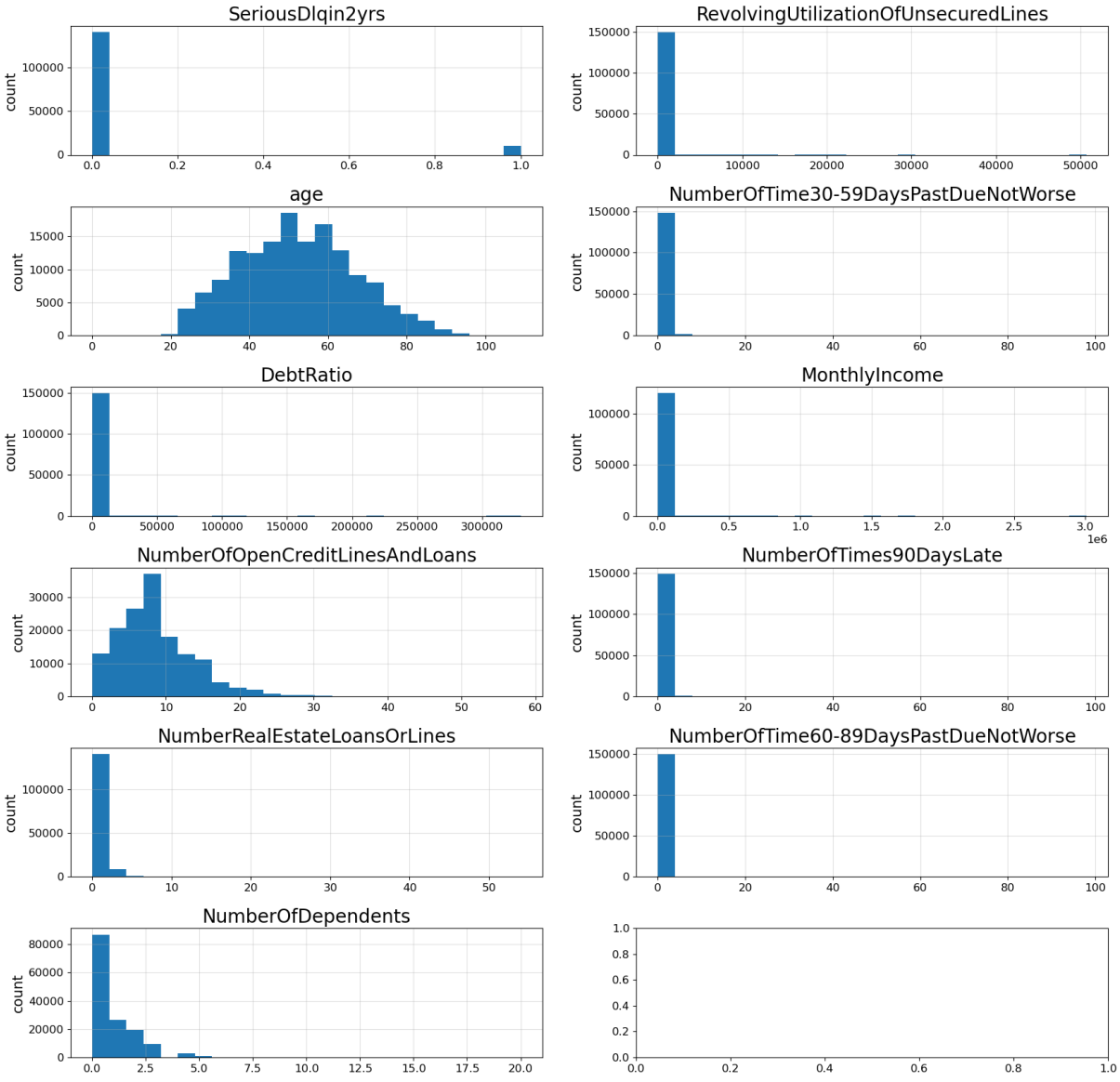

Figure 2: Depicts the distribution of values for each feature in the dataset.

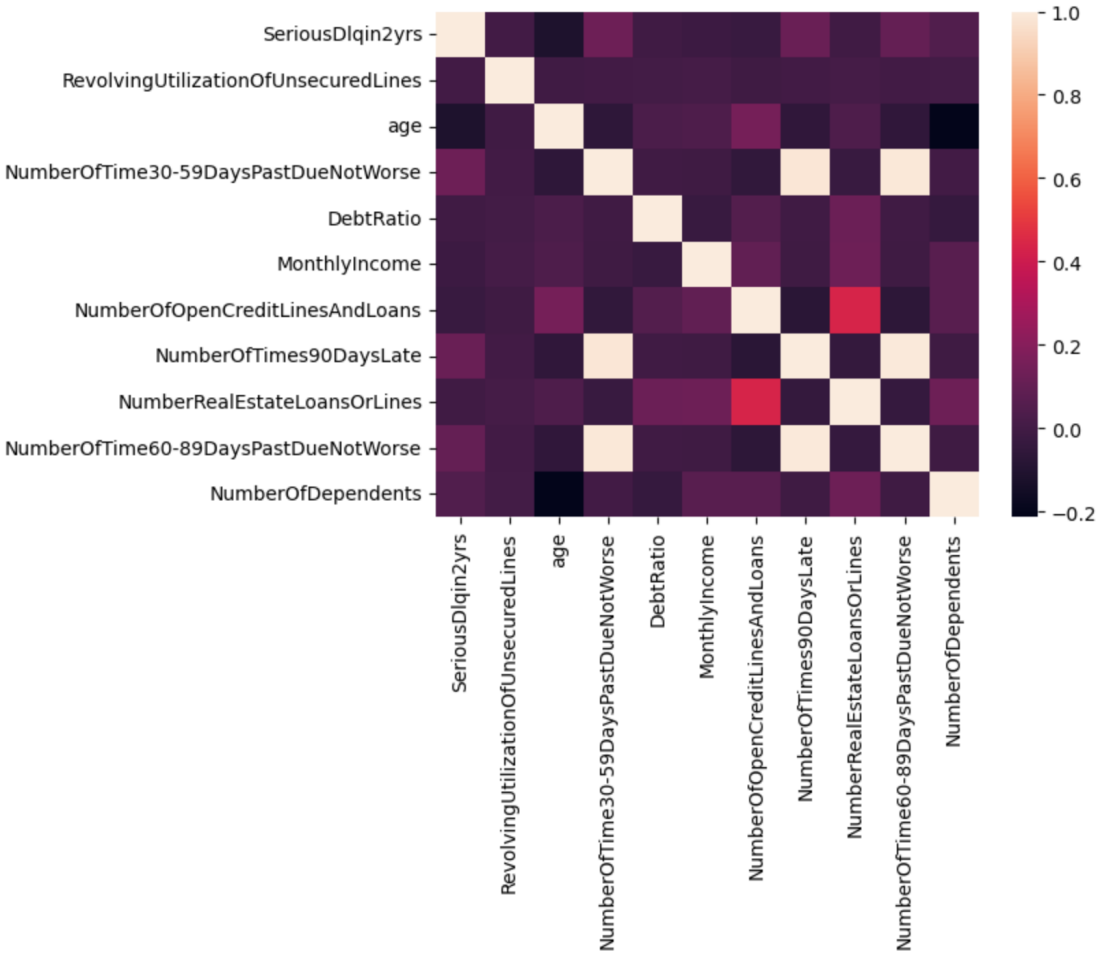

Figure 3: Depicts the correlation of each feature to every other feature in the dataset.