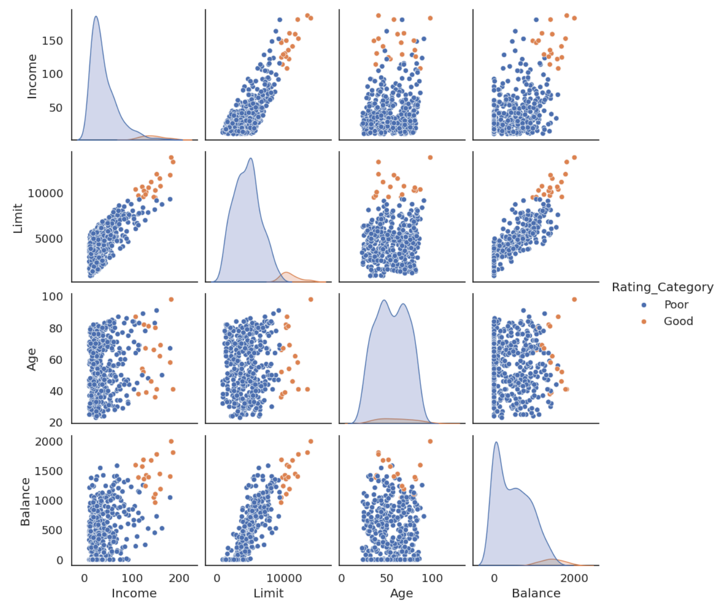
CSci 4521 Homework 1 Writeup

Data Analysis



The graphs reveal several key insights and trends. Individuals with higher income and limits tend to have good credit ratings as seen by the clusters of orange points in those graphs being higher up. Income and limit show a clear and strong correlation with credit rating. Though the correlation is not as strong as income and limit, balance also shows a similar trend with individuals who have a higher balance often having a good credit rating. Age does not show a strong correlation with credit rating as indicated by the clusters being more randomized with no clear pattern. Overall, there is more of an indication of financial factors being an influence towards credit rating rather than age. Based on the results, individuals looking to improve their credit rating should focus on raising their limit, as the correlation between a higher limit and a good credit rating is strong. Maintaining a healthy income will also improve your credit rating since those with a higher income manage their debt accordingly. Also, managing credit card balances will prevent any debt accumulation that will cause the credit rating to decrease. Since age does not play a huge factor in credit rating, younger individuals can still achieve good credit if they focus and ensure secure financial habits.

Model Evaluation

```
KNN Classifier Performance:
Training Accuracy: 0.9966
Testing Accuracy: 0.9896
Training Precision: 0.9965
Testing Precision: 0.9937
Training Recall: 0.9963
Testing Recall: 0.9924
Training F1-score: 0.9963
Testing F1-score: 0.9928
```

```
All Good Credit Classifier Performance:
Training Accuracy: 0.0377
Testing Accuracy: 0.0391
Training Precision: 0.9641
Testing Precision: 0.9431
Training Recall: 0.0373
Testing Recall: 0.0606
Training F1-score: 0.0027
Testing F1-score: 0.0069
```

**Note: all metrics output are based on a singular run so they may change if you run it again.*

To evaluate the KNN Classifier and the All Good Classifier, we normalized the input data and split it into training and testing datasets where we calculated the accuracy, precision, recall, and F1-score on both datasets. The KNN Classifier was given the k-value 3.

On the training data, the KNN classifier achieved near-perfect scores, with a precision of 0.9965, recall of 0.9963, accuracy of 0.9966, and an F1-score of 0.9963. In contrast, the All Good classifier struggled, exhibiting a training precision of 0.9641 but very low recall (0.0373), accuracy (0.0377), and F1-score (0.0027).

On the testing data, the KNN classifier maintained its strong performance, achieving a precision of 0.9937, recall of 0.9924, accuracy of 0.9896, and an F1-score of 0.9928. While slightly lower than the training scores, these results indicate good generalization. The All Good Classifier, however, continued to perform poorly on the testing data, with a precision of 0.9431, recall of 0.0606, accuracy of 0.0391, and an F1-score of 0.0069. Again, the All Good Classifier shows its inability to accurately classify instances, likely due to its oversimplified approach. However, one thing to note is that its precision for both its training and testing was very high, being 0.9641 and 0.9431 respectively. This is because precision focuses on the correctness of positive predictions and the imbalance in the dataset for good vs. poor credit ratings (more poor credit ratings than good credit ratings) likely added to the high number.

Additionally, in this classification task, only three features (Income, Balance, and Limit) were used instead of four, as they showed a clear correlation with credit rating, unlike age, which did not contribute meaningfully to the classification process. This feature selection ensures that the model doesn't overfit while maintaining high performance.

For real-world applications, the KNN classifier is a clear choice. Its consistently high precision, recall, accuracy, and F1-score on both training and testing data demonstrate its effectiveness and ability to generalize. The All Good classifier's extremely low recall and F1 score indicate that it is not capable of effectively identifying instances of "good credit," making it practically useless for any real-world application.

Based on the output metrics, recall, and F1 score emerge as the most important evaluation metrics for determining the effectiveness of the classifiers. While the accuracy and precision of the KNN Classifier and the precision of the All Good Classifier are high, the recall and F1 score for the All Good Classifier are extremely low indicating its effectiveness or lack thereof. The All Good Classifier fails to correctly identify poor credit ratings which is in its design as it is used as a baseline for comparison. This makes it clear that the KNN Classifier is the better choice overall as not only are its accuracy and precision high (almost close to one) but so are its recall and F1 score. Since recall measures the ability to correctly identify poor credit ratings and F1 score balances both precision and recall, these metrics provide a reliable assessment of a model's effectiveness, in this case, the KNN Classifier.

Applying the Model

The region expected the applicant resides in is the East. This is determined by running the KNN Classifier with an input of region instead of credit rating which these strings are first converted into numbers that map to 0 being East, 1 being South, and 2 being West (similar to how 0 is good credit and 1 is poor credit for credit rating). Once this classifier is run with k = 3, the algorithm finds that the region for those 3 nearest neighbors is classified as 0 or East. This means that this applicant likely also resides in the East. Additionally, the East has many metropolitan areas where young professionals often rent apartments near their workplaces. The applicant's income of $60,000 aligns well with salary expectations in these regions, where the cost of living tends to be higher. Additionally, their credit behavior, with a combined credit limit of $7,500 and an outstanding balance of $1,500, suggests responsible credit usage, which may be more in line with financial patterns observed in the East compared to other regions.

Using the KNN Classifier, the model predicts that this person has a good credit score (above 670). The KNN Classifier was used with $k = 3$, meaning that the model considers the 3 nearest neighbors when making a classification. A smaller k ensures that the patterns in the data are preserved, there is a more complex decision boundary with a low bias and is suitable for imbalanced data like our given dataset. Additionally, with $k = 3$, the KNN Classifier performed very high across all metrics (see: Model Evaluation). The features used in the application are the same as before (income, limit, and balance) as those factors showed the most significant correlation in credit rating. Income affects an individual's ability to make payments, while credit limit and outstanding balance determine the credit utilization ratio. Before the KNN Classifier was applied, the input data was normalized to ensure all the features were on the same scale, and the data was split into training and testing datasets to evaluate the model's performance. The model predicts a good credit score primarily because of the applicant's low credit utilization ratio as well as the individual having two credit cards which implies some financial history. The individual's stable income further reinforces their ability to manage debt responsibly. Given these factors, the KNN analysis supports the classification of this individual as having a good credit score.

The smallest change to the consumer's stats that could flip the prediction would likely be an increase in the credit utilization ratio. If the individual's balance increased from $1,500 to $4,000 with the limit staying the same ($7,500), their credit utilization would be greater than 50%, likely negatively impacting their credit rating. Similarly, if there were a decrease in the credit limit (which would also increase the credit utilization ratio), their credit rating would likely be lowered. Since this plays a huge factor in credit rating, this small change would likely switch the variables from good to poor.