DSE200x Final Project German Credit Data

Andy Tse

Abstract

Summarize your questions and findings in 1 brief paragraph (4-6 sentences max). Your abstract needs to include: what dataset, what question, what method was used, and findings.

I am going to use the German Credit data from the UCI Machine Learning Repository. For this dataset, I will want to do further understanding on consumer credit. The approaches that I am using are Random Forest, Decision Tree, k-Nearest Neighbors, and Naive Bayes. These are all supervised learning methods. k-Means Clustering will be used for unsupervised learning approach as a test for accuracy comparison. In addition, there are eight characteristics out of 21 that are correlated.

Motivation

Describe the problem you want to solve with the data. It may relate closely with your research question, but your goal here is to make your audience care about the project/problem you are trying to solve. You need to articulate the problem you are exploring and why (and for whom) insight would be valuable.

In this project, I will be finding variables to determine the differences between the amount of debt that the customers have, and how long they have it. I am planning to find the correlation with its variables for its factors. The visualizations will be shown in the correlation plot and pairplot using the seaborn package. The detail is to highlight the most correlated variables involved.

Dataset(s)

Dataset Link: https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/

I am using the German Credit Dataset from the UCI Machine Learning Repository. For the dataset, the columns are labeled from numbers 1 to 21. There are 1000 rows and 21 columns in the dataset.

The following names are translated below:

- Attribute 1: Status of existing checking account
- Attribute 2: Duration in month
- Attribute 3: Credit history
- Attribute 4: Purpose
- Attribute 5: Credit amount
- Attribute 6: Savings account/bonds
- Attribute 7: Present employment since
- Attribute 8: Installment rate in percentage of disposable income
- Attribute 9: Personal status and sex
- Attribute 10: Other debtors / guarantors
- Attribute 11: Present residence since
- Attribute 12: Property
- Attribute 13: Age in years
- Attribute 14: Other installment plans
- Attribute 15: Housing
- Attribute 16: Number of existing credits at this bank
- Attribute 17: Job
- Attribute 18: Number of people being liable to provide maintenance for
- Attribute 19: Telephone
- Attribute 20: Foreign worker

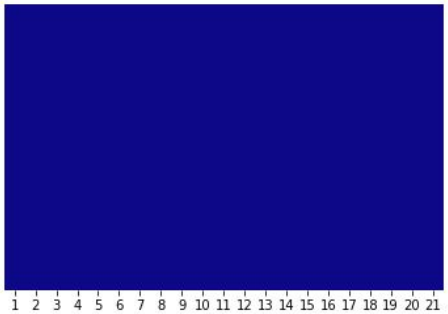
Data Preparation and Cleaning

At a high-level, what did you need to do to prepare the data for analysis? Describe what problems, if any, did you encounter with the dataset?

For this dataset, there are no missing values and all the values turned out to be false. I have use the seaborn package command of heatmap to show the missing data. Also, I have use the ‘is.null().any()’ command to detect missing data. The data is good to go and ready for analysis!

```
In [10]: gcredit.isnull().any()
Out[10]: 1     False
         2     False
         3     False
         4     False
         5     False
         6     False
         7     False
         8     False
         9     False
         10    False
         11    False
         12    False
         13    False
         14    False
         15    False
         16    False
         17    False
         18    False
         19    False
         20    False
         21    False
         dtype: bool
```

```
In [11]: sns.heatmap(gcredit.isnull(),yticklabels=False,cbar=False,cmap='plasma')
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x218c62542b0>
```



Research Question(s)

Which features have the highest correlation between the credit amount and credit history?

Out of the four supervised learning approaches: k-Nearest Neighbors, Random Forest, Decision Tree, and Naive Bayes, which method yields a higher accuracy?

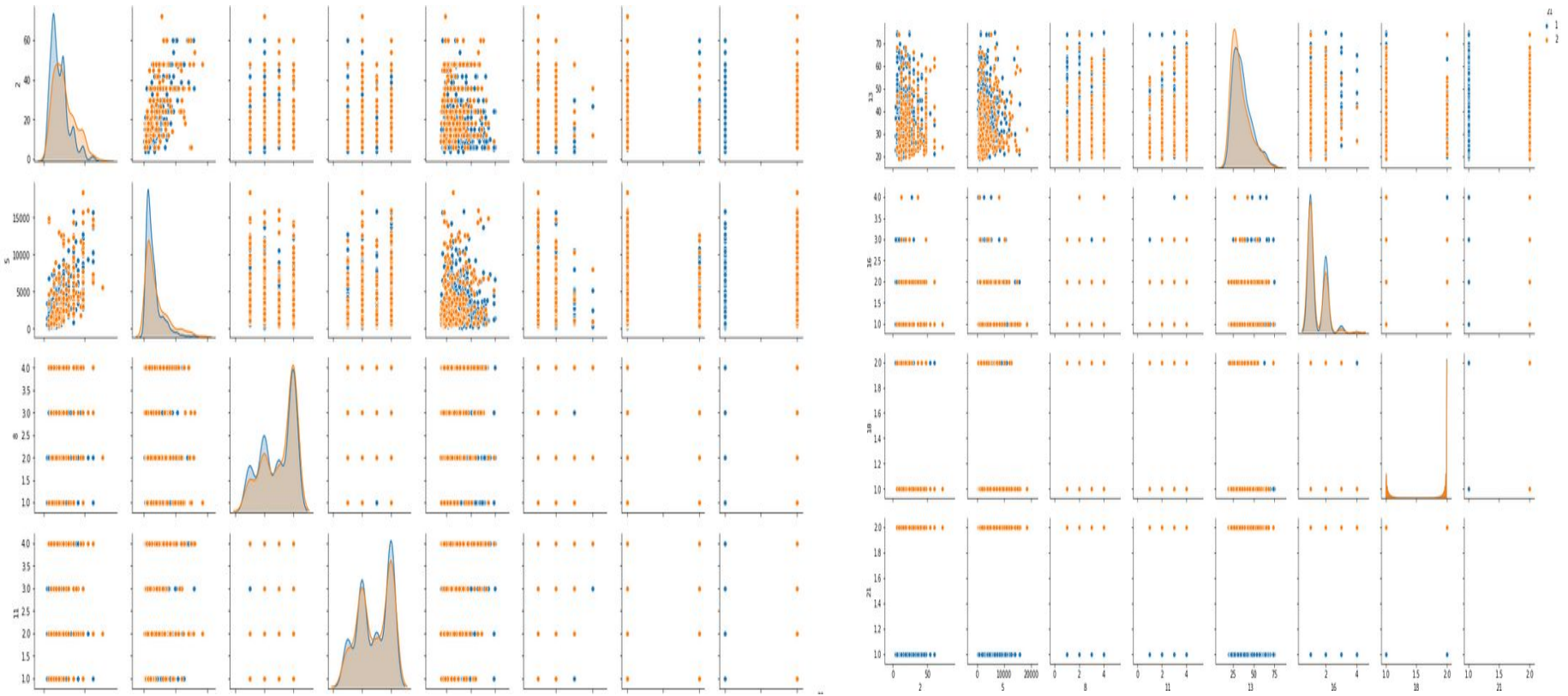
Will unsupervised learning method k-Means Clustering be better or worse than the supervised learning?

Methods

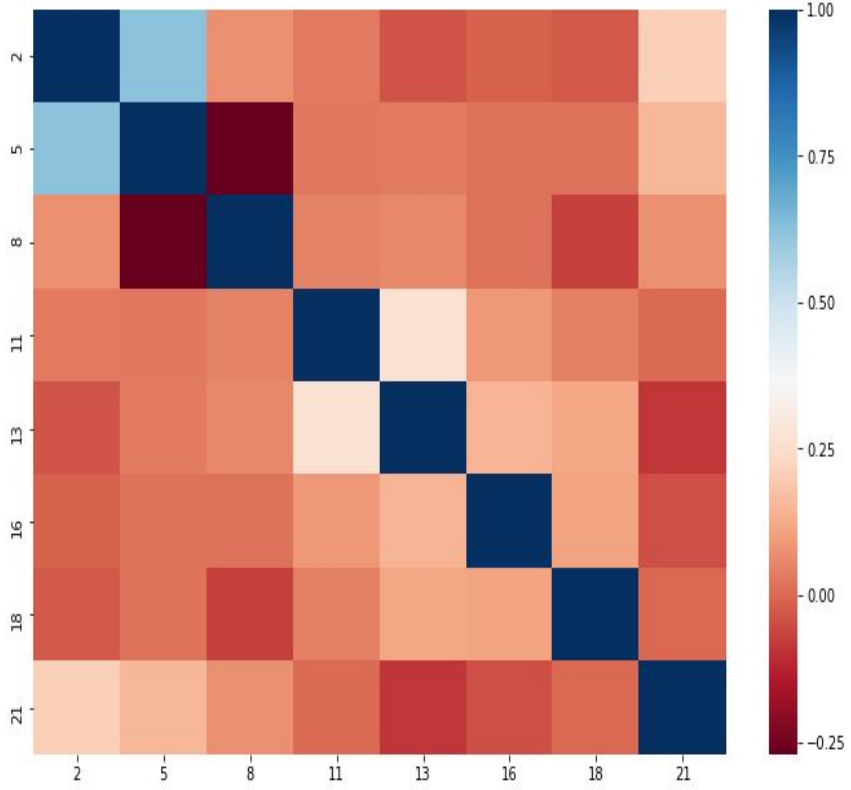
What methods did you use to analyze the data and why are they appropriate? Be sure to adequately, but briefly, describe your methods.

For this project, I have used k-Nearest Neighbors, Decision Tree, and Random Forest Methods. They are appropriate to use because this dataset contains all the labeled targets. Also, I will use k-Means Clustering to determine what happens if the targets are unlabeled.

Pairplot



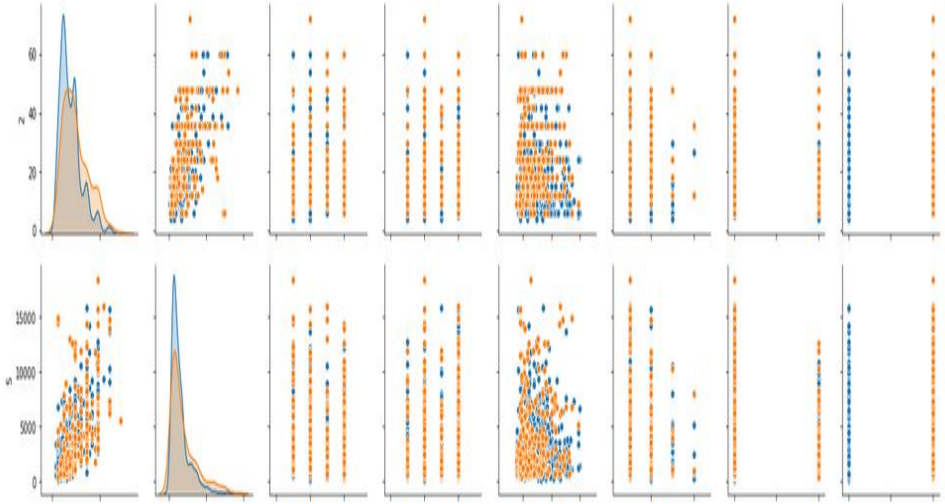
Correlation Plot

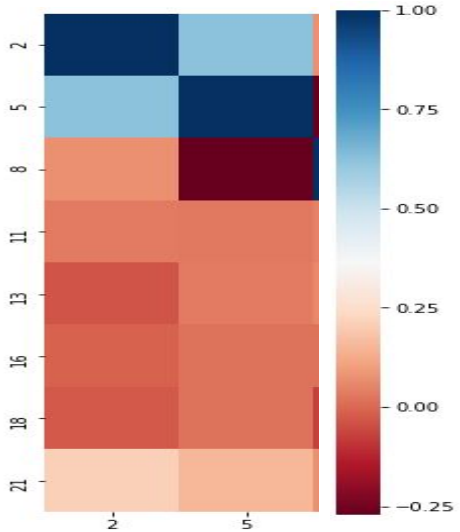


	2	5	8	11	13	16	18	21
2	1.000000	0.624984	0.074749	0.034067	-0.036136	-0.011284	-0.023834	0.214927
5	0.624984	1.000000	-0.271316	0.028926	0.032716	0.020795	0.017142	0.154739
8	0.074749	-0.271316	1.000000	0.049302	0.058266	0.021669	-0.071207	0.072404
11	0.034067	0.028926	0.049302	1.000000	0.266419	0.089625	0.042643	0.002967
13	-0.036136	0.032716	0.058266	0.266419	1.000000	0.149254	0.118201	-0.091127
16	-0.011284	0.020795	0.021669	0.089625	0.149254	1.000000	0.109667	-0.045732
18	-0.023834	0.017142	-0.071207	0.042643	0.118201	0.109667	1.000000	-0.003015
21	0.214927	0.154739	0.072404	0.002967	-0.091127	-0.045732	-0.003015	1.000000

Findings

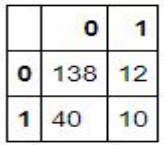
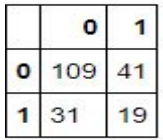
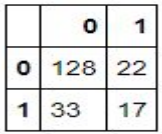
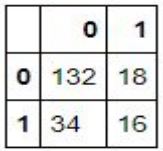
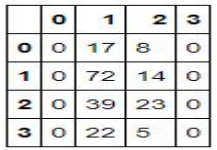
By looking at the pairplot and correlation plots from the slides, it has indicated that attributes 2, 5, 8, 11, 13, 16, 18, and 21 are the ones that are the most correlated factors out of 21 that are in the dataset. However, the 2nd and the 5th elements of the dataset are the most correlated, which equates to the duration of month and the amount of credit that the person have. As shown, the 2nd and 5th factors are shown in the light blue color as they are at the higher scale. The pairplot and correlation plot on this slide refers to the most correlated factors as shown below.





	2	5
2	1.000000	0.624984
5	0.624984	1.000000
8	0.074749	-0.271316
11	0.034067	0.028926
13	-0.036136	0.032716
16	-0.011284	0.020795
18	-0.023834	0.017142
21	0.214927	0.154739

Findings/Conclusions

	kNN	Decision Tree	Random Forest	Naive Bayes	k-Means Clustering
Accuracy Score	74%	64%	72.5%	74%	47.5%
Confusion Matrix					

	0	1
0	138	12
1	40	10

	0	1
0	109	41
1	31	19

	0	1
0	128	22
1	33	17

	0	1
0	132	18
1	34	16

	0	1	2	3
0	0	17	8	0
1	0	72	14	0
2	0	39	23	0
3	0	22	5	0

By comparing the results with these four methods, Naive Bayes and kNN are the most accurate results to perform the analysis, while Decision Tree has the worst as it is a significant difference for supervised learning methods. Looking at the whole picture, unsupervised learning using k-Means Clustering is far more inaccurate than supervised learning.

Limitations

With the accuracy in the lower to mid 70% range for supervised learning methods and 47.5% using k-Means Clustering, there could be different features that will be used to determine if the model could be improved. I will want to see if there are any differences with the non-correlated features to re-evaluate the model, as it is done with the most correlated features. This dataset will be explored more in depth with different methods particularly with Convolutional Neural Network and Recurrent Neural Network.

Acknowledgements

Where did you get your data? Did you use other informal analysis to inform your work? Did you get feedback on your work by friends or colleagues? Etc. If you had no one give you feedback and you collected the data yourself, say so.

No one has given feedback.

References

I have done the work on my own. Also used the references from the UCI Machine Learning Repository along with the scikit-learn documentation as well.