Kay Quiballo | MSDS 411 | 05.07.2023

Assignment 3

Abstract. An executive summary of the research.

This assignment explores the possibility of using unsupervised learning to aid in the performance of supervised learning methods. Credit card companies may take advantage of a combination of methods to identify the risks of lending credit to clientele based off their financial behaviors. In this analysis, we consider pretraining data with clustering methods such as bi-clustering and K-means to aid in the supervised method models that are traditionally used for credit risk prediction.

Introduction. Why are you conducting this research?

In this assignment, we consider what it is like to work for a credit card company. The task is to consider granting credit to customers, and to use the customer database as a means for predicting good or bad credit. Extending credit to "bad" customers incurs costs that are 5x greater than the cost of no extending credit to a "good" customer. The ideal situation is to lend credit to only "good" customers in hopes of it getting paid back in full with interest.

Using a 1994 German credit card case, we completed a traditional regression on clientele's credit data to determine indicators of being labeled a "good" or "bad" credit risks. Working for a credit card company, the analytics team performed several unsupervised learning methods to see if this type of pretraining improves the prediction accuracy of credit risk to reduce costs and raise revenue.

Literature review. Who else has conducted research like this?

In a similar study by Carcillo et al., researchers combined supervised and unsupervised methods in determining risks surrounding credit card fraud detection. By analyzing customer behavior, this study used clustering to aid in the accuracy of prediction models to identify credit card fraud (Carcillo et al. 2019).

Methods. How are you conducting the research?

The Dataset

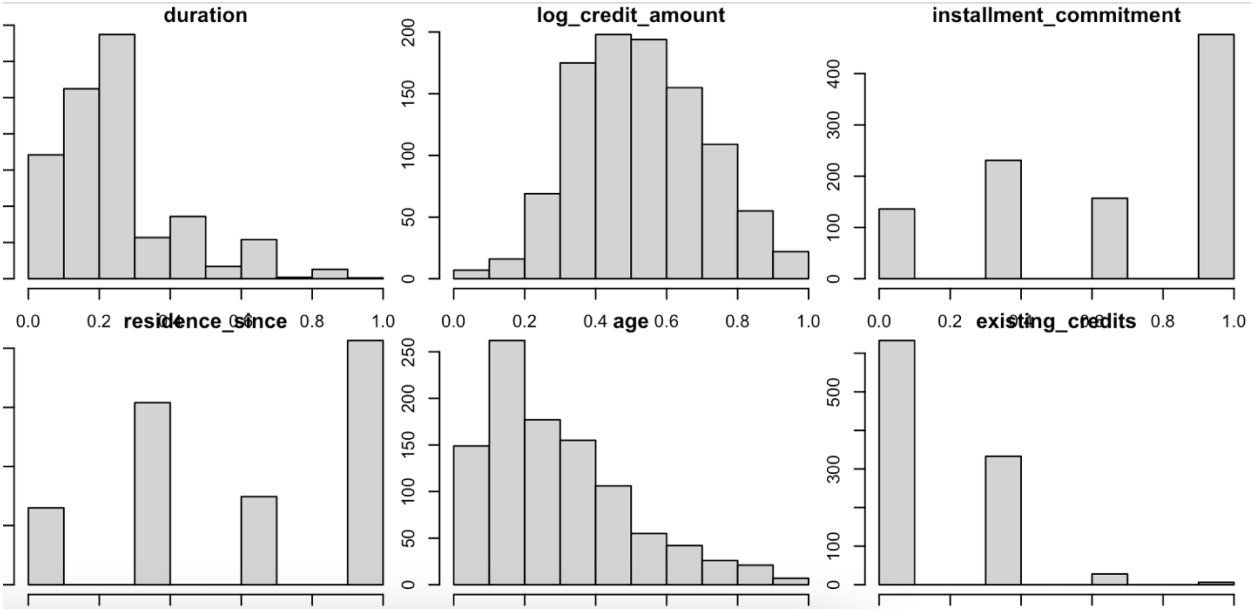
The dataset comes from the 1994 German credit card case which consists of 1000 observations and 21 variables and ranges from information such as status of existing checking account, purpose of credit, and occupation. About 30% of the records are labeled as "bad" credit risks (we focused on the "bad" credit risks in this assignment because they incur costs that are 5x greater than failing to lend credit to someone with "good" credit risks). Traditionally, credit card companies perform logistic regressions to label customers with a response of "good" or "bad" credit based on financial information.

Data Cleaning and Manipulation

The data was binarized according to variable and response for analysis. Values such as *personal_status* "female single" that had no observations were removed from the dataset. Other categories like *purpose* that had multiple categories with very small sample sizes were grouped into "other" levels. Skewed variables like *credit_amount* were log transformed to have a more normal distribution.

Of the 44 semi-binarized variables in the cleaned dataset, 38 of them have binary responses and 6 of them are continuously distributed. Of the 6 continuous distributions (see below), *installment_commitment, residence_since*, and *existing_credits* are discrete and are split into their 4 respective, unique values. *Duration* and *age* are slightly skewed and are binarized according to above and below the median. *Log_credit_amount* has a more normal distribution and is binarized according to above and below the mean. With all variables being binarized, we are left with 1000 observations across 53 variables. All data has been scaled using min-max methods to ensure there is no bias or unequal weighting across variables.

6 non-binary, continuous variables' distributions



In the Results section, we will cover the results from the logistic regression in comparison to pretraining via clustering methods.

Results. What did you learn from the research?

Results from the logistic regression are the baseline performance of an autoencoder-based solution. Using cross-validation via training-and-testing datasets, the results are classified by accuracy measures such as precision (proportion of applicants identified as bad were actually bad) and recall (proportion of bad applicants we identify as bad). Criteria labels optimal cutoffs to weigh the cost of incorrect classification of actually bad credit to outweigh any other results (Zweig and Campbell (1993) and Gallop et al. (2003))

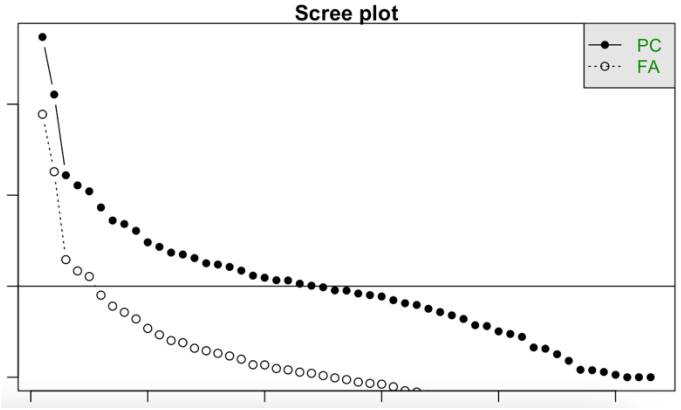
Table for Baseline Statistics

```
Cross-validation summary across folds:

  baseprecision baserecall basef1Score basecost ruleprecision rulerecall rulef1Score rulecost
1         0.538      0.475       0.505      179          0.362      0.864       0.510      130
2         0.607      0.557       0.581      157          0.418      0.967       0.584       92
3         0.592      0.509       0.547      160          0.338      0.930       0.495      124
4         0.609      0.475       0.533      173          0.372      0.932       0.531      113
5         0.652      0.469       0.545      186          0.381      0.922       0.539      121

Cross-validation baseline results under cost cutoff rules
    F1 Score:  0.532
    Average cost per fold:  116
```

As we see from the 5-folds, the base precision ranges from 0.54-0.64 and the base recall ranges from 0.47-0.56. We may also note that the "rule" columns take advantage of the cutoff criteria which penalizes more harshly for inaccurately labeling "bad" customers as "good". This results is worse precision values ranging from 0.34-0.42 but much better recall ranging from 0.86-0.97. This means that the credit card company would be much more selective at lending out credit and will often deny credit to "good" customers, but would also be much better at avoiding lending credit to "bad" customers.

The baseline results were compared to the results of unsupervised methods' classification matrices. In the following sections, we look at 2 different unsupervised clustering methods which work to group together similar customers. Such methods are typically utilized to reveal relationships between variables, but not necessarily used to answer a specific research question like model prediction.

<u>K-means from Primary Components</u>



From the Scree plot, we may note that there are 24 eigenvalues from the correlation plot that are greater than 1 indicating that the Principal Component Analysis may have 24 different components. Using these components from the binarized values, we performed a K-means analysis of 2 clusters in hopes of identifying “good” and “bad” credit risks from customers.

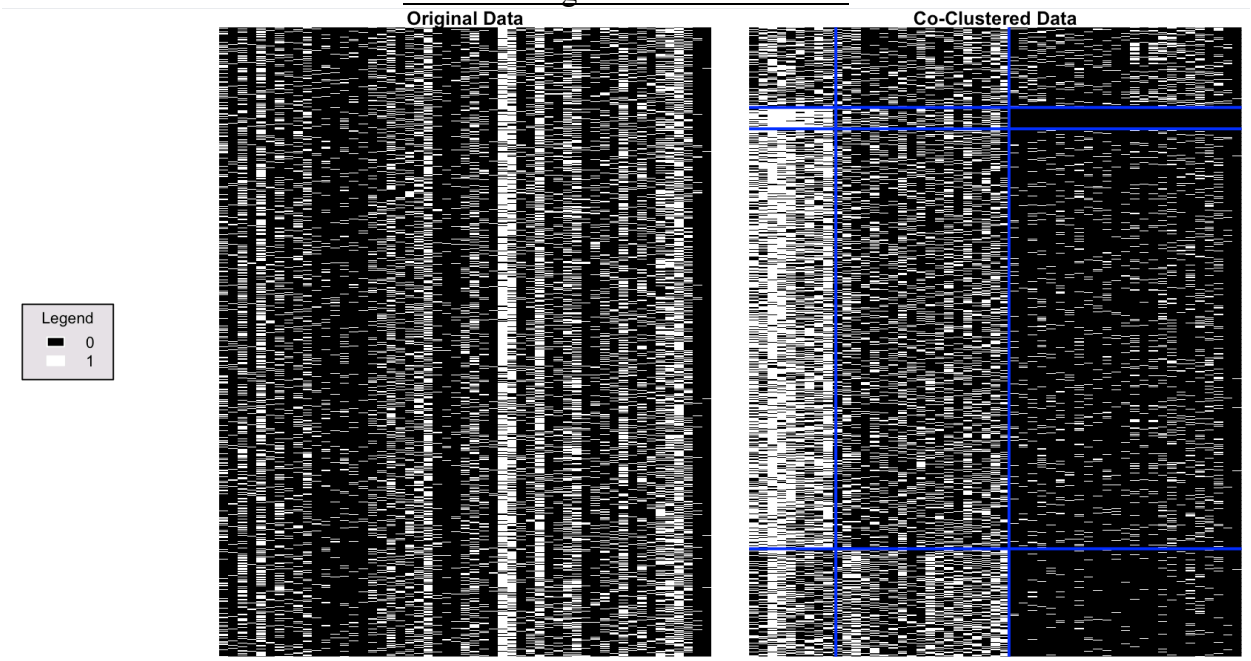
Confusion Matrix for KMeans-Clustered Data (*Default Cutoff*)

	Predicted Bad	Predicted Bad
Actual Bad	167	133
Actual Good	351	349

Using K-Means clustering with the default cutoff, we get a precision of 0.5566667 and recall of 0.3223938.

Unfortunately, K-Means clustering does not provide a probability distribution of falling within a cluster, so we cannot use the optimal cutoff value of 0.086 to create an additional confusion matrix.

Co-clustering from Binarized Data



I decided to use 4 row clusters which are respective of the 4 options of historical credit as well as 3 column clusters which are respective of the highest 3 eigenvalues (seen in the scree above). The blue borders separate the good and bad credit risks (black and white) on the right plot.

This plot from co-clustering tells us several things. Between the row clusters, it is difficult to distinguish drastic differences in credit. However, between the column clusters, it is easy to see that these components can be easily indicative of good or bad credit shown by the gradient of white to black on the right plot. Although the clustering visualization provides some insight, we may only use the row clusters in the classification matrix which is not very useful.

Confusion Matrix for Co-Clustered Data (*Default Cutoff*)

	Predicted Bad	Predicted Bad
Actual Bad	180	120
Actual Good	367	333

Using Biclustering with the default cutoff, we get a precision of 0.6 and recall of 0.3290676.

Confusion Matrix for Co-Clustered Data (*Optimal Cutoff = 0.086*)

	Predicted Bad	Predicted Bad
Actual Bad	158	142
Actual Good	337	363

Using Biclustering with the optimal cutoff, we get a precision of 0.5266667 and recall of 0.3191919.

Conclusions. So, what does it all mean?

Comparison of Precision and Recall by Methods

Method	**Cutoff**	**Precision**	**Recall**
Logistic Regression	Default	0.54-0.64	0.47-0.56
Logistic Regression	Optimum	0.34-0.42	0.86-0.97
K-Means Clustering	Default	0.56	0.32
Bi-Cluster	Default	0.60	0.33
Bi-Cluster	Optimum	0.53	0.32

When looking at precision, it looks like the clustering methods perform on par: the default bi-cluster performing the best of the clustering methods with an accuracy of 0.60 which is comparable to the overall best precision: the default logistic regression's precision of 0.54-0.64.

However, in the context of working for a credit card company that penalizes harsher on lending credit to those who are labeled as "bad" customers, we rely on recall as the ultimate decision maker. Even when the optimal cutoff is used, all of the unsupervised clustering methods performed much more poorly in terms of recall (0.32-0.33) in comparison to the best logistic regression model's recall (0.86-0.97).

Overall, it is important to note that unsupervised learning methods can provide invaluable information in an exploratory data analysis. They may reveal patterns and relationships between predictors. It is also of note to point out that unsupervised learning methods are useful for detecting anomolies such as fraud (which was also pointed out in the lit review section) (Lucas and Jurgovsky (2020) and Pang et al. (2020)).

But in this situation, especially when it comes to predicative accuracy, optimal cutoff values, and precision and recall for a specific research question, there are too many criteria that unsupervised learning methods alone cannot answer. Overall, I advise that the credit card company continue to use logistic regression to create predictive models and use them to classify good and bad credit risk among customers.

Appendix:

Data Dictionary for German Credit Case

```
Information available from OpenML: https://www.openml.org/d/31

Author: Dr. Hans Hofmann
Source: [UCI](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)) - 1994
Please cite: [UCI](https://archive.ics.uci.edu/ml/citation_policy.html)

German Credit dataset
This dataset classifies people described by a set of attributes as good or bad credit risks.

This dataset comes with a cost matrix:
```
      Good Bad (predicted)
Good   0   1   (actual)
Bad   5   0
```

It is worse to class a customer as good when they are bad (5), than it is to class a customer as
bad when they are good (1).

### Attribute description

1. checking_status Status of existing checking account, in Deutsche Mark.
2. duration Duration in months
3. credit_history Credit history (credits taken, paid back duly, delays, critical accounts)
4. purpose Purpose of the credit (car, television,...)
5. credit_amount Credit amount
6. savings_status Status of savings account/bonds, in Deutsche Mark.
7. employment Present employment, in number of years.
8. installment_commitment Installment rate in percentage of disposable income
9. personal_status Personal status (married, single,...) and sex
10. other_parties Other debtors / guarantors
11. residence_since Present residence since X years
12. property_magnitude Property (e.g. real estate)
13. age Age in years
14. other_payment_plans Other installment plans (banks, stores)
15. housing Housing (rent, own,...)
16. existing_credits Number of existing credits at this bank
17. job Job
18. num_dependents Number of people being liable to provide maintenance for
19. own_telephone Telephone (yes,no)
20. foreign_worker Foreign worker (yes,no)
21. class Credit classification (good or bad) response to be predicted
```

Resources:

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, Gianluca Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Information Sciences,Volume 557,2021,Pages 317-331,ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.05.042.
(https://www.sciencedirect.com/science/article/pii/S0020025519304451)

```
Cross-validation summary across folds:
  baseprecision baserecall basef1Score basecost ruleprecision rulerecall rulef1Score rulecost
1         0.538      0.475       0.505      179         0.362      0.864       0.510      130
2         0.607      0.557       0.581      157         0.418      0.967       0.584       92
3         0.592      0.509       0.547      160         0.338      0.930       0.495      124
4         0.609      0.475       0.533      173         0.372      0.932       0.531      113
5         0.652      0.469       0.545      186         0.381      0.922       0.539      121
Cross-validation baseline results under cost cutoff rules
    F1 Score:  0.532
    Average cost per fold:  116
```