Assignment #3

David Caban

Bronsin Jabraeili

Scott Jue

Dr. Joe Wilck

Northwestern University

MSDS 411 – Unsupervised Learning Methods

April 30, 2023

Abstract

This paper discusses the use of unsupervised learning methods to improve the credit card industry's customer risk management. The "German Credit Card Case" dataset is utilized to show the efficacy of unsupervised learning in credit classification and customer extension. There is a review of several other studies that have used unsupervised learning techniques in addressing similar issues in the credit card industry. The methodology used for this study includes several preprocessing steps taken prior to fitting a baseline logistic regression model on credit card application data, including recoding and consolidating categories, converting variables to a factor, and performing a log transformation. The logistic regression model's performance was assessed using cross-validation and precision and recall metrics. Additionally, a separate cutoff value was used based on the company's risk tolerance and cost matrix. After further data preparation and normalization, three different autoencoder models with different activation functions and regularization methods were applied as an unsupervised pre-training technique. Model 2 was selected as the best-performing autoencoder model and was used to pretrain the credit data before refitting the logistic regression model. After comparing the results between the baseline logistic regression model and the pretrained logistic regression model, it was determined that the logistic regression model without the use of an autoencoder model outperformed the model with autoencoder pretraining in terms of precision, F1 score, and total cost, although the auto-encoder model did perform better in recall. In conclusion, the baseline logistic regression model was chosen as the optimal model, despite the potential benefits of pre-training the logistic regression model using unsupervised learning techniques. In the absence of better performance metrics, the most simplistic model should always be chosen due to the benefits of interpretability and less unnecessary code to slow the program down.

Introduction

To successfully grow their business, credit card companies must properly manage risk when extending their customer base. This is a challenging endeavor as it can be significantly more costly to extend credit to a "bad" customer versus extending credit to a "good customer". However, this assumes that a company has a methodology that allows it to bucket potential customers into either classification. Traditional methods such as logistic regression are primary options for performing this bifurcation. Optimal solutions typically incorporate unsupervised learning methods to augment supervised models, resulting in greater accuracy and time-saved.

The "German Credit Card Case" can be used to illustrate the utility of blending approaches or simply using an unsupervised method, The dataset consists of 1,000 observations across 22 variables. These variables include measures and characteristics such as age, credit history, credit amount, marital status, employment status, and purpose of credit being sought, among others. As a result, this dataset provides a useful and practical context in which to explore the benefits of using unsupervised learning in the credit classification and eventual extension process.

Literature Review

Several research papers and articles have been published centering on the use of unsupervised methods in addressing challenges in the credit card industry. For example, a study by Eric Umuhoza, Dominique Ntirushwamaboko, Jane Awuah, and Beatrice Birir (2011)

outlined a strategy that could be used to offer personalized credit products in Africa by using an unsupervised behavioral-based segmentation model. This study demonstrated the efficacy of the model and it was built using anonymous data from the Commercial International Bank of Egypt. Similarly, the Machine Learning Group at the Université Libre de Bruxelles (2010) used unsupervised techniques to identify fraudulent credit card transactions. As noted in the abstract of the paper, traditional supervised techniques are frequently used in fraud detection using existing data, however, unsupervised methods proved to be effective when applied to new data. They tended to identify anomalies and detect new patterns of nefarious activity. Further, it is noted that the two methods are complimentary. Finally, Wu and Wang (2018) used the unsupervised learning technique Self Organizing Map to partition individuals using credit information with the goal of limiting default. The paper highlights how clustering was used to separate individuals into different segments. Models were then built for each cluster and this improved prediction accuracy.

Methodology

Prior to creating the baseline logistic regression model and applying unsupervised pre-training methods the data was preprocessed. First, it was discovered that there are no observations for "female single" in the personal_status variable. Therefore to fix this issue, the personal_status variable was converted into a factor and “female single” was not included as one of the factor levels. Additionally, the purpose variable had categories with low or no frequencies. As such, these were re-coded and consolidated into existing categories or an “other” category.

Finally, since the credit_amount variable is highly skewed, a log transformation was performed to reduce the observed skewness.

The data was then used to fit a baseline cross-validation logistic regression model to classify credit card applicants as good or bad using all explanatory variables, except for the "foreign_worker" variable. The dataset is divided into 5 subsets using the cross-validation method, with one subset as the testing set and the remaining subsets being used to train the logistic regression model. This approach is iterative, as the model is fitted on 5 different folds and tested on 5 different subsets during each iteration. The model's performance is assessed based on precision, recall, F1-Score, and total cost. Two sets of performance metrics are computed, one using a standard cutoff of 0.5 and the other using a cutoff value of .086 which is based on the company's risk tolerance. The total cost is based on the costs of granting credit to a bad customer and not granting credit to a good customer.

The unsupervised pre-training technique used in this project was an autoencoder model. Prior to inputting the data into the model, the data was prepared by creating a design matrix. This converts the categorical variables into dummy variables using one-hot encoding. Using the design matrix, the data normalized using a min-max normalization method. The last step prior to building the autoencoder models was creating a train test split using a 80/20 split.

The keras_model_sequential() function was used to create three different autoencoder models. All three models consist of the same encoder dense layers with 20, 15, 10, and 5 hidden units, respectively. The decoder layers then reverse the encoder and consist of four dense layers with 10, 15, 20, and the same number of 44 input units as the training data, respectively. Then each model is compiled using “binary_crossentropy” as the loss function ,”adam” as the

optimizer, and “accuracy” as the metric. The first autoencoder model (Model 1) uses “sigmoid” as the activation function. The second autoencoder model (Model 2) that was tested uses “ReLU” as the activation function. The final model is similar to Model 2 but includes a dropout layer with a 0.2 rate as a regularization technique to help prevent overfitting. Then each autoencoder model was trained using 100 epochs and a batch size of 32. Model 2 was shown to have the lowest loss score and the highest accuracy out of the three models. As such, Model 2 was selected to pretrain the credit data prior to refitting the logistic regression model.

Results

As highlighted in the methodology section above, there were two approaches taken in this study: logistic regression and utilizing an autoencoder to pre-train the data prior to running the logistic regression. The respective recall, F1-Score, and total cost will serve as a variety of key metrics to holistically assess which method is optimal for deployment. It is critical to assess the performance of the models based on this blend of metrics to tell a complete story.

Beginning with the logistic regression utilizing cross validation sans autoencoding, the output resulted in an average precision of 0.374, recall of 0.923, F1 score of 0.532, and Cost of 116 (See Exhibit A). The precision score is quite low, meaning that the logistic regression suffered in properly measuring true positives. As mentioned previously, this is just one portion of the story and when considering the complexity of credit customer classification the output is more comprehensible. The recall score conversely is reasonable by being greater than 90%, meaning the logistic regression correctly identified true positives out of all the actual positive instances in the dataset. This relationship of a low precision score but a high recall score means the algorithm is unable to distinguish between true positives versus false positives. Scrutinizing

the dataset can reveal if the data collection process is causing this phenomenon. A measure that describes the overall performance of the model, considers both precision and recall, and balances between the two is the F1 score. The F1 score of this model was 0.532 which is slightly higher than the reasonable threshold of 0.500.This reveals that despite the imbalance between precision and recall, overall the model can be trusted to classify whether a candidate is “good” or “bad.” Finally, the cost function of the model was 116. This is difficult to establish a threshold for assessment; however, the objective is to minimize the cost as much as possible. Therefore, this will be compared with the second approach.

The logistic regression utilizing the autoencoders established similar results and patterns. The precision score was 0.303 and the recall score was 0.986. This follows the same trend as the previous model which excluded the auto-encoders. The F1 score of 0.464 is also around the 0.500 mark however, if the threshold of reasonability is established to be 0.500 then this does fail to meet expectations. This determination is lenient as recommender systems such as this one are more subjective than other classification algorithms such as health screenings. Finally the cost score was 140 (See Exhibit B).

When comparing the two models, the logistic regression without the auto-encoders outperformed the one without auto-encoders in precision, F1 score, and cost (See Exhibit C). The only metric that the auto-encoder model outperformed in was recall. Overall, the consensus view is that the optimal model to deploy is the first model that does not include auto-encoding. The results are fairly similar, but one must always choose the better performing model. Another benefit of the logistic regression model excluding autoencoders is that it is more simplistic. Although unsupervised learning models typically improve supervised learning methods, in this

case logistic regression, it is not guaranteed that it will be the case. The final verdict is that the more simplistic model with better performance will be selected and utilized.

Conclusion

An objective approach was taken to test whether the initial hypothesis that unsupervised learning methods, in this case auto-encoding, will improve logistic regression for customer segmentation. The final output revealed that the unsupervised learning method resulted in similar but slightly unperformed versus not utilizing auto-encoding techniques. In data science, there are endless possibilities when tackling a problem. Researchers must consider multiple options and approach the problem with an open mind. In this instance, the best approach to classify credit customers is to keep it simple by using logistic regression without unsupervised learning methods to manipulate the data.

References

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 557* (2021): 317-331.

E. Umuhoza, D. Ntirushwamaboko, J. Awuah and B. Birir, "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa." *SAIEE Africa Research Journal 111, no. 3* (2020): 95-101.

Paldino, G.M., Lebichot, B., Le Borgne, YA. et al. The role of diversity and ensemble learning in credit card fraud detection. Adv Data Anal Classif (2022)

Wu, H. and Wang, C.-C. (2018) Customer Segmentation of Credit Card Default by Self Organizing Map. American Journal of Computational Mathematics, 8, 197-202. https://doi.org/10.4236/ajcm.2018.83015

Appendix

Exhibit A - Outputs for Logistic Regression Sans Autoencoder

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


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

Exhibit B - Outputs for Logistic Regression with Autoencoder

```
Cross-validation summary across folds:
  baseprecision baserecall basef1Score basecost ruleprecision rulerecall
1         0.000      0.000       0.000      311         0.297      0.966
2         0.538      0.115       0.189      276         0.310      1.000
3         0.273      0.105       0.152      271         0.287      0.982
4         0.294      0.085       0.132      282         0.297      0.983
5         0.529      0.141       0.222      283         0.325      1.000
  rulef1Score rulecost
1       0.454      145
2       0.473      136
3       0.444      144
4       0.457      142
5       0.490      133


Cross-validation pre-training results under cost cutoff rules:
    F1 Score:  0.464
    Average cost per fold:  140
```

Exhibit C - Comparison Between Methods

Logistic Regression without Auto-Encoders

K-Folds	Precision	Recall	F1 Score	Cost
1	0.362	0.864	0.510	130
2	0.418	0.967	0.584	92
3	0.338	0.930	0.495	124
4	0.372	0.932	0.531	113
5	0.381	0.922	0.539	121
	0.374	0.923	0.532	116

Legend

	Out Perfomed
	Under Performed

Logistic Regression with Auto-Encoders

K-Folds	Precision	Recall	F1 Score	Cost
1	0.297	0.966	0.454	145
2	0.310	1.000	0.473	136
3	0.287	0.982	0.444	144
4	0.297	0.983	0.457	142
5	0.325	1.0000	0.490	133
	0.303	0.986	0.464	140