Group 6: Sean Atkinson, Sameer Khan, Parker Davis

Assignment 3: Group 6 CC Autoencoder Analysis

Abstract

In the following report we assess the utility of unsupervised pre-training to provide greater accuracy and precision with respect to the classification of potential customers. We wish to use supervised and unsupervised learning methods to find “good” customers that can be added to the credit card company’s client list. These methods will be used to not only distinguish good from bad customers, but also to reduce the frequency of mislabeling a good customer as bad and vice versa. The report will provide insight into the original logistic regression model with the addition of greater accuracy and precision by using an unsupervised variable creation model as input.

Keywords: logistic regression, unsupervised learning, autoencoder, supervised learning

Introduction

The original logistic regression model provided inference with respect to the classification of good and bad customers for granting credit card accounts. Logistic regression is a useful analysis method for solving classification problems with binary outcomes. The classification process works by using past information to calculate the probability that a given person is a good customer. An initial cutoff was set which represented a threshold for predicting bad credit. This was determined by taking the ratio of the cost of determining false negatives over the cost of false positives and multiplying that by the ratio of prevalence of positive and negative. Then a model formula was developed with cross validation in order to determine a

person's class by checking several of the existing variables in the data set. Next, the model was trained and tested on the data set to develop its decision making skills. Group6 CC believes that improvements in precision, accuracy, F1 score, and recall can be achieved with the use of autoencoders as an input into logistic regression. The management of the input variables with autoencoding will provide a new orientation of the variables and structure of the input data. The method of restructuring and pre-training will complement improvements in classification of both GOOD and BAD customers.

Literature Review

Determining good credit candidates and being able to identify fraud has long been a need of large corporations that grant credit. Dr. Lucas & Dr. Jurgovsky looked into credit card fraud detection and attempted to develop a model to best identify fraudulent charges. Two important notes that they made about the datasets is that first there are skewed class distributions. A vast majority of credit card purchases are correct and only a small percentage are fraudulent. The model developed needed to properly address this, since most unsupervised methods only work well with balanced data. Next the dataset itself has shifts in fraudulent behavior. As more fraud defense gets implemented, fraud behaviors adapt to overcome these changes (2020). In addition to this work Zou, Zhang, & Jiang used autoencoding to supply a neural network in credit card fraud detection. They used oversampling to account for the imbalance in fraud and a denoising autoencoder to reshape the data properly (2019). These models yielded 90% or higher accuracy in determining fraud, proving that autoencoding is effective in fraud detection.

Methods

The choice of unsupervised learning method is an autoencoder. The application of this method will provide dimensionality reduction and anomaly detection in the resulting input of the normalized dataset. Application of this methodology will improve the classification accuracy of Group6 CC customers. Figure 1 provides the original data results from the logistic model through 5 iterations of Precision, Recall, F1 score, and costing.

Iteration	Base Precision	Base recall	Base F1 score	Base Cost	Rule Precision	Rule recall	Rule F1 score	Rule cost
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	0.388	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

Figure 1 Baseline metrics and rule based calculations

The results for precision, recall, F1 score, and cost are broken down into two categories. Base is the standard metric measurement and rule is the metric after taking our predefined cutoff into account. Precision is the ratio between true positives and all positives; in this case it would be the percentage of actual good customers in the total population of customers that were labeled as good by the model. Recall measures the proportion of actual positives that were identified correctly, thus capturing any good customers which were labeled as bad. The F1 score combines precision and recall into a single metric by taking their harmonic mean. F1 scores are a useful way to compare the performance of different classifiers as a measure of accuracy. Group 6

predicts that the F1 score will significantly increase once an unsupervised autoencoder model is applied to the data.

Auto encoders are neural networks that compress inputs into a latent space representation so they appear closer together. The autoencoder consists of two main parts: an encoder and decoder. The encoder compresses the data into that latent space representation and the decoder reconstructs the data from the latent space representation. Auto encoders with the right constraints can provide valuable dimensionality reduction and identify more interesting relations than those found in PCA. Autoencoders are much better at reducing high dimensional non-linear data than other techniques like PCA and t-SNE.

Figures 2 and 3 display the Receiver Operating Characteristics (ROC) curve and Area Under the Curve (AUC). These curves are used as performance measurement for machine learning models. The ROC itself is a probability curve and the AUC measures the degree of separability. The greater area under the curve shows the models ability to predict true positive and tru negative values better.

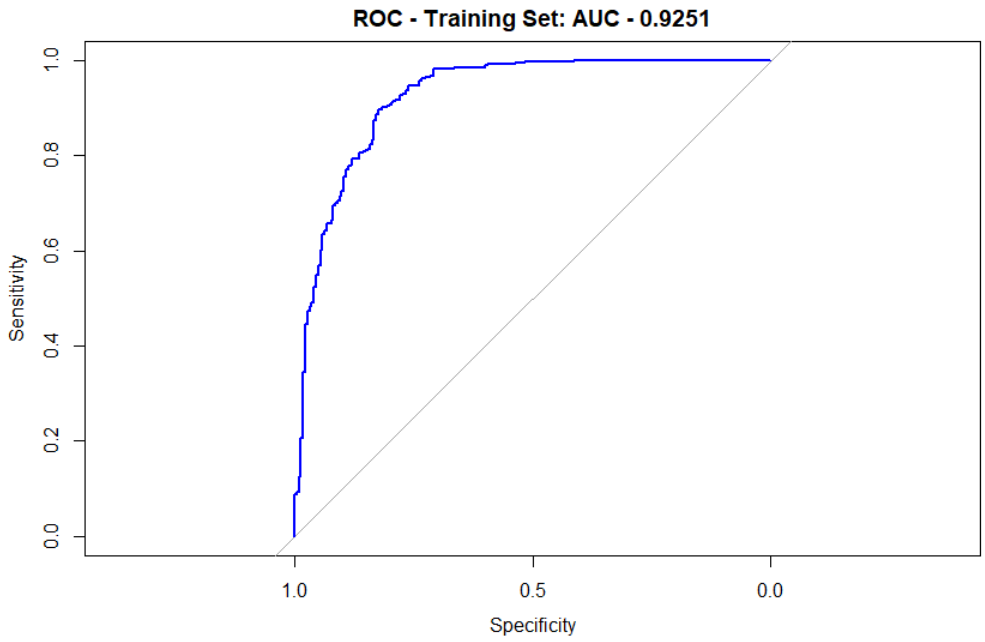

Figure 2 Training set - Area Under the Curve results

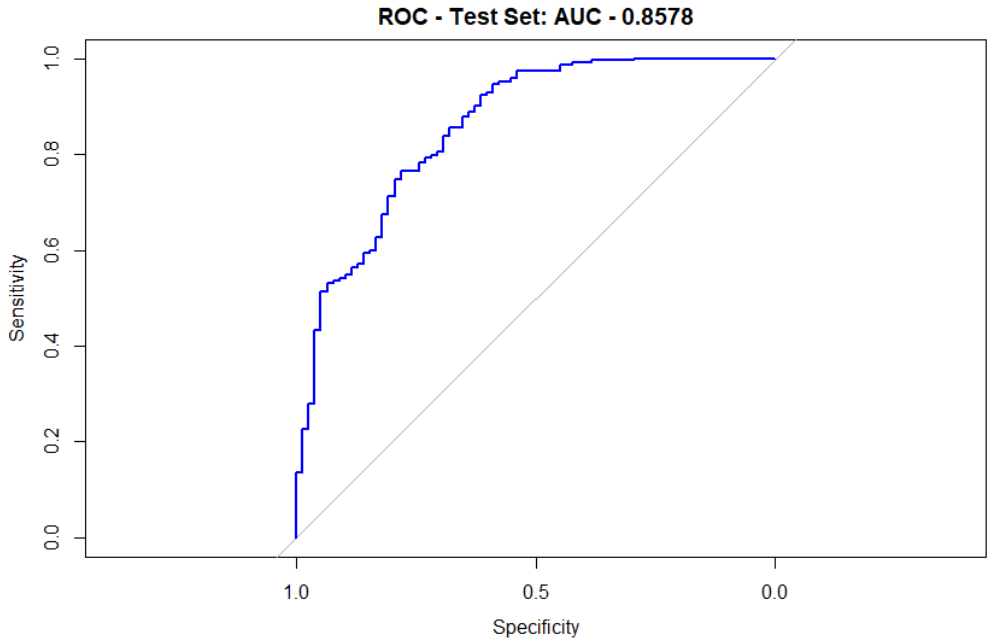

Figure 3 Test set - Area Under the Curve results

It can be seen that in the training data the AUC is around 0.92. In a perfect world with an AUC of 1 this would indicate perfect class separation and therefore no false positives or negatives would ever be predicted. In the train case the AUC of 0.92 indicates very good class separation. Next the test case shows an AUC of 0.857. Although this is lower than the training data it is still an acceptable level. This means on test data the model has an 85.7% chance of being able to distinguish a good credit candidate from a bad one.

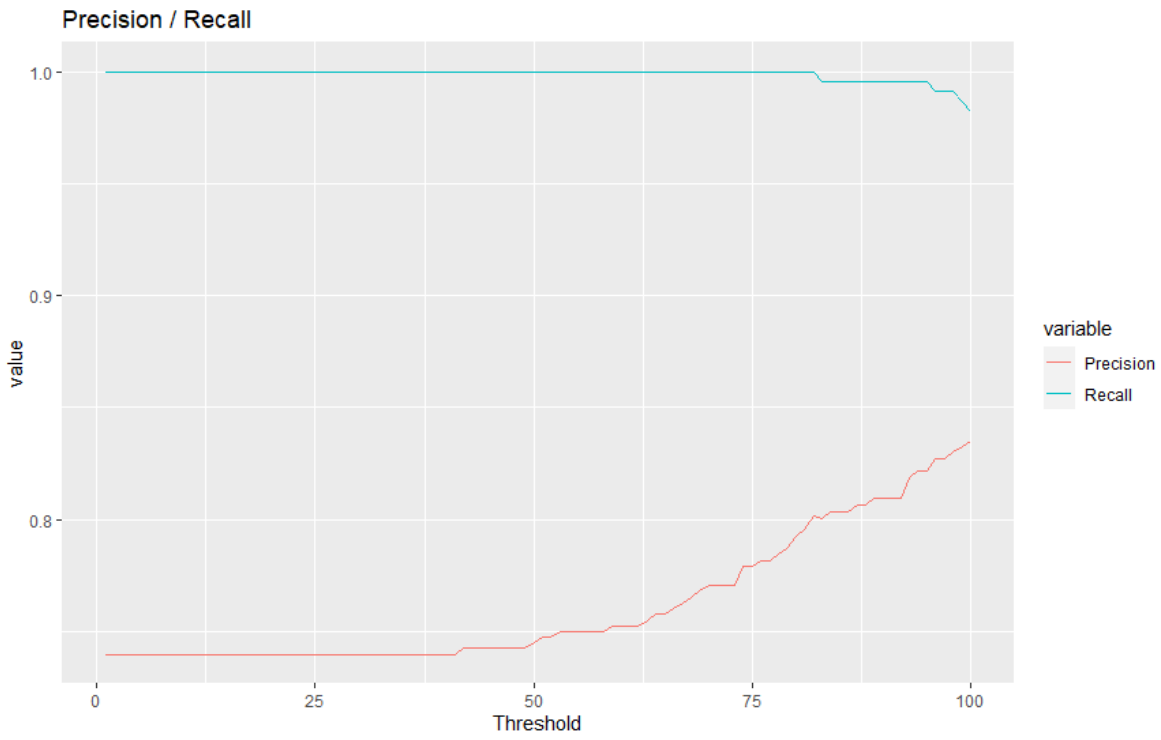

Figure 4 Precision and Recall results

Precision is high at around 75% and improves over time to 85%. The recall maintains a steady line which also shows how the precision of the model can improve without sacrificing the recall ability. In viewing the resulting metrics (see Figure 5) from the assessment above the following confusion matrix is found identifying the accuracy of the model. Using a cutoff value of 3 the resulting matrix was found.

Prediction	**0**	**1**
0	115	2
1	107	476

Figure 5 Confusion matrix for the supervised model

The confusion matrix shows that 109/700 predictions were incorrect. The 107 incorrect predictions do exist in the false negative category. This is more ideal because the false negative would deny someone good credit. False positives are a 5X more expensive mistake to make and are therefore desired less. It is better to be more conservative on giving credit out.

Metric	Training dataset result	Test dataset result
Accuracy	0.8443	0.8433
95% CI	(0.8153, 0.8704)	(0.7972, 0.8826)
No Information Rate	0.6829	0.74
P-Value [Acc > NIR]	< 2.2e-16	1.235e-05
Kappa	0.5884	0.5141
Mcnemar's Test P-Value	< 2.2e-16	2.976e-08
Sensitivity	0.5180	0.4487
Specificity	0.9958	0.9820
Pos Pred Value	0.9829	0.8974
Neg Pred Value	0.8165	0.8352
Prevalence	0.3171	0.2600
Detection Rate	0.1643	0.1167
Detection Prevalence	0.1671	0.1300
Balanced Accuracy	0.7569	0.7153

Figure 6 Autoencoded Training and Test dataset metrics

Figure 6 shows a table of prevalent result metrics to evaluate the autoencoders performance. Most notable it can be seen that the accuracy of the model from both the test and train data sets lies around 84%. Additional confidence interval information and detection rates

can be seen as well. With this information the model was determined to be valid for use and the confusion matrix below which shows the number of true positive, true negative, false positive and false negative determination. True answers are the goal but not all false answers are created equally. It is much cheaper for the company to make a false negative prediction and deny a good person credit when compared to giving a bad person credit. Using the Test dataset we find the following results from the training:

Prediction	**0**	**1**
0	35	4
1	43	218

Figure 7 Confusion matrix for the supervised test dataset model

Reduced the features to 10 and we are still able to identify 218 out of 222 bad potential credit customers - look to see how many bad customer were in the test data set - breaks this down to good versus bad customers - as bad customer cost x5 more than good customer, having better performance to detect the bad credit customer is much more beneficial. Overall, the total of 300 test elements were correctly predicted in 253 cases and 47 in the incorrectly identified. Here of note is the utility of 43 incorrectly identified as bad but actually good. These are a lower cost in comparison to those being bad and incorrectly identified as good.

Results

The results of the autoencoder unsupervised model is a drastically improved version compared to the logistic regression originally applied. Due to the nature of credit data often being skewed due to a vast majority of people making good purchases. Due to their ability to generate their own relationships between variables, unsupervised methods can identify important

relations that supervised methods can miss. Due to the skew, supervised methods would need specific weights to be applied to find the types of relationships unsupervised can find naturally.

Iteration	New Precision	New Recall	New F1 Score	New Cost
1	0.719	0.621	0.836	57
2	0.729	0.522	0.843	68
3	0.712	0.706	0.832	65
4	0.694	0.667	0.819	55
5	0.729	0.600	0.843	59

Figure 8 Pretrained metrics

Based on the resulting calculations we have an improved F1 score of 0.836 (84%) compared to the base model with no feature engineering of 0.532 (53%). The average cost per iteration also was significantly lower than the previous baseline and rules based cost model We see for the revised model we now have a cost of $60.80 whereas the average cost previously was $116.00.

Conclusion

Supervised learning methodologies are often strongest with clean and well represented data. In a situation for determining a good credit candidate this can be very useful. On the other hand supervised learning struggles when it comes to underrepresented data. Underrepresented data is often the case in anomaly detection which makes it difficult for supervised methods to account for these. Unsupervised learning is strong at anomaly detection due to it defining its own weights for relationships in the data which is not easily accomplishable in supervised methods

(Anderson, 2021). Pang et al. also mentions the difficulty in detecting anomalies due to their seemingly random occurrences and behavior. They also detail 3 types of anomalies known as: point, conditional, and group anomalies. Point anomalies are single point occurrences such as abnormal health indicators. Conditional anomalies are repeatable given a certain set of specific conditions. Finally, group anomalies are a subset of data out of the norm (Lucas & Jurgovsky, 2020). In the credit card application existence of point anomalies could eventually form a grouping of anomalies in the unsupervised model. From the data and models developed in this model it can be seen that the unsupervised methods edge out over the supervised methods provided through logistic regression. This is likely due to the natural skew in the credit data; unsupervised models tend to perform better due to their own relationship creation which is more powerful with data that has imbalance representation.

References

Anderson, M. (2021). Supervised vs unsupervised machine learning approaches. Supervised vs Unsupervised Machine Learning Approaches. Retrieved February 26, 2022, from *https://www.itransition.com/blog/supervised-vs-unsupervised-learning*

Jiang, X., Zhang, Y., Zhang, W., & Xiao, X. (2013, October). A novel sparse auto-encoder for deep unsupervised learning. In 2013 *Sixth international conference on advanced computational intelligence (ICACI)* (pp. 256-261). IEEE.

Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. *arXiv preprint arXiv:2010.06479.*

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR), 54(2), 1-38.*

Zou, J., Zhang, J., & Jiang, P. (2019, August 30). Credit card fraud detection using Autoencoder neural network. arXiv.org. Retrieved February 26, 2022, from *https://arxiv.org/abs/1908.11553*