Applying Data Mining Techniques In German Credit Applications

By: Kylee Fisher (kfisher11@ycp.edu)
Faculty Mentor: Dr. Wei Chen
York College of Pennsylvania | Graham School of Business

Abstract

Money-lending is the world's second oldest profession but it wasn't until the beginning of the 20th century that credit companies were founded to share information about credit. Credit agencies often collect large amounts of data to predict when defaults or similar events can occur. This research investigates applying known data mining techniques to a German Credit dataset to investigate if those techniques can achieve equivalent, or better, results in determining whether applicants have "good credit" or "bad credit" as the results attained before predictive modeling.

Introduction

In the early stage of the historical transition to predictive modeling, humans were employed to label credit records as good or bad based on 30 variables and 1,000 records of prior credit applicants. The variables provided include information on the applicants financial and personal background, if they have any existing credit, and the reason for applying. In this study I investigate the use of several data mining techniques to predict if an applicant will be categorized as having good or bad credit. By implementing these techniques, I was able to identify the most important set of attributes when determining credit status. I was also able to identify the net profit associated with each technique and the accuracy of that technique as well.

Methodology

In order to analyze the GermanCredit.csv I used Python via the PyCharm editor with the NumPy, Pandas, and Sklearn libraries. I cleaned up the data by removing missing values, if any, and converting categorical variables into dummy variables. Once the variables were transformed, I partitioned the data into 60% training sets, 25% validation sets, and 15% test sets. I used the training and validation models to analyze three data mining techniques and then the final test set to analyze the final selected model.

Three Data Mining Techniques:

- **Neural Networks** : a high performance model for classification and prediction that mimics the biological activity in a human's brain.
- **Classification and Regression Trees (CART):** used for the systematic placement of group membership data as it maps the data into predefined groups/classes and searches for new patterns.
- **K-Nearest Neighbors (KNN):** one of the most commonly used data mining techniques for classifications in supervised learning. This model finds records in a data that have similar numerical values of a set of predictor variables and does not assume any underlying distribution in the data

Comparison Performance Metrics:

- **Confusion Matrix:** It is a summary of prediction results of a classification problem that breaks down the number of correct and incorrect predictions for each class.
- **Accuracy Score:** The percentage of predictions that the model correctly predicted.

Results

From the CART model I was able to determine that the most important variables in determining whether an applicant will have bad/good credit is the applicant's age, duration of credit in months, whether they have a checking account or not, and the amount of the credit.

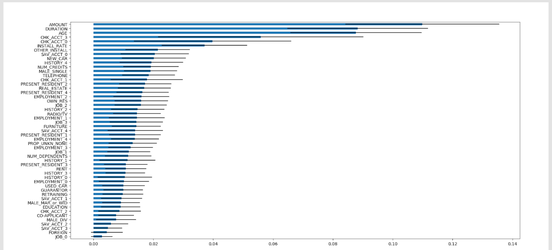

Figure 1: Importance Bar Chart

The accuracy scores of each model were very close. The Neural Network had the highest model accuracy of 70.4%, followed by KNN with a 70% model accuracy, and then the CART model with a 68.4% accuracy.

	Neural Networks	CART	K-Nearest Neighbors
Accuracy	0.704	0.684	0.7

Figure 2: Accuracy Score Table

The confusion matrices of each model can be seen in the image below. The top left square describes the number of applicants who were correctly classified as having good credit. The top right square describes the number of applicants who were falsely classified as having good credit. The bottom left square describes the applicants who were falsely classified as having bad credit. The bottom right square describes the applicants who were correctly classified as having bad credit.

NN CM	
0	0
74	176

CART CM	
28	33
46	143

KNN CM	
74	176
0	0

Figure 3: Confusion Matrices

Conclusion

The best data mining technique to predict whether an applicant will have good/bad credit is the KNN model. This model predicts that 30.7% of applicants will have bad credit while 69.2% of applicants will have good credit. The model accuracy was high (70%) and the confusion matrix showed the KNN model to have the highest net profit.

Being that this is the selected model, I ran the model again using the partitioned 15% test data set. This improved the model accuracy to 70.8% but it decreased the amount of applicants classified as having good credit to 46, instead of the original 74. I believe that KNN is the best model for this data as it takes into consideration the applicants closest to the predictor applicant and classifies the applicant based on that. .

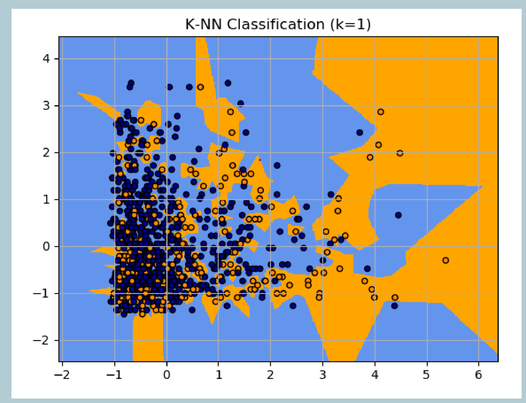

Figure 4: KNN Distribution Graph

Acknowledgements

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for business analytics: concepts, techniques and applications in Python.* Hoboken, NJ: John Wiley & Sons, Inc.