German Credit Worthiness

Vidit Kalani, Shoun Abraham

<https://github.com/ViditKalani/MSCS-5610-Data-Mining-Final-Project>

**Abstract**

In this paper, we analyzed the German Credit Dataset from the Caret Library. By analyzing this dataset, were able to create a model that allows us to predict Credit worthiness of a citizen based on a few factors related to credit. In order to create this model, we used techniques such as PCA analysis and the KNN recommendation system. We tested our model by using a training/testing set from the dataset. This form of analysis is integral to the finance industry. Further research into these models will benefit credit rating agencies greatly, as they will be able to rate credit worthiness quickly and accurately. Ultimately, by using techniques such as the KNN recommendation system, we were able to create a model with an estimated accuracy of (TO BE ADDED).

**Introduction**

The German Credit Dataset is a dataset that examines a German's credit worthiness and other factors associated with this worthiness. This dataset is interesting because it contains a plethora of information regarding credit worthiness. It is interesting to see the factors that are needed (and collected) to try to predict credit status. There are many entities that could benefit from analysis of this dataset. Credit Rating agencies could benefit from this dataset as they are primarily interested with evaluating a person's credit. With better prediction models, credit rating agencies could predict credit status more accurately and faster. Consumers could also stand to benefit from analysis of this dataset. With a good prediction model, a consumer could be able to predict the status of their own credit. Consumers would also learn more about what factors contribute to evaluation of credit.  With this project, we plan to obtain more information about credit rating/worthiness using outside research. Using our research and dataset, we plan to create a model to predict credit worthiness. This model will be able to predict a person's credit worthiness given important attributes of credit worthiness to a decent degree of certainty.

**Related Work**

**Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry:**

This article describes the importance of credit rating among credit issuers and investors. This paper also describes the problems associated with many credit rating models. In order to overcome these problems, a hybrid model was created from two existing models.

**Mining multiple private databases using a kNN classifier**

This article gives insight into using the kNN classifier. This helps us to learn more about the model we plan to create.

**Using data mining to improve assessment of credit worthiness via credit scoring models**

This article examines the history and importance of credit scoring models to the financial industry. It also examines the role of data mining to improve the assessment of credit worthiness using credit scoring models. Finally, this paper compared the accuracy of different credit rating models.

**Data Set**

This dataset is available from the UCI Machine learning website, or in the Caret library through R. This dataset contains 1000 samples. This dataset has two values for credit eligibility: good or bad. Predictors of this dataset include checking account status, duration, credit history, purpose of the loan, amount of the loan, savings accounts or bonds, and others. Upon initial examination of this dataset, it appears that it is relatively clean. We don’t believe we need to spend much time cleaning the dataset. The dataset we used also does not contain missing values. This will be very useful when we create a proper model. In order to predict the effectiveness of our data, we plan on splitting our dataset into training and testing sets. We have not yet determined the exact split to use. [MORE INFORMATION ABOUT DATASET WILL BE ADDED IN THE FUTURE]

**Methods/Models**

1. PCA Analysis
   1. [discussion and results of PCA Analysis]
2. KNN Recommendation System
   1. [discussion and results of kNN]
3. Logistic Regression
   1. [discussion and results of Regression Analysis]

**Results and Discussion**

[TO BE ADDED]

**Conclusions**

[TO BE ADDED]

**Contributions**

Vidit Kalani: KNN recommendation system, project set-up

Shoun Abraham: PCA Analysis, project write-ups

**References**

You-Shyang Chen, Ching-Hsue Cheng, Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry, Knowledge-Based Systems, Volume 39, 2013, Pages 224-239, ISSN 0950-7051,

<https://doi.org/10.1016/j.knosys.2012.11.004>.

(<http://www.sciencedirect.com/science/article/pii/S0950705112003139>)

Bee Wah Yap, Seng Huat Ong, Nor Huselina Mohamed Husain, Using data mining to improve assessment of credit worthiness via credit scoring models, Expert Systems with Applications, Volume 38, Issue 10, 2011, Pages 13274-13283, ISSN 0957-4174,

<https://doi.org/10.1016/j.eswa.2011.04.147>.

(<http://www.sciencedirect.com/science/article/pii/S0957417411006749>)

Xion, L., Chitti, S., & Liu, L. (n.d.). Mining multiple private databases using a kNN classifier. Retrieved March 20, 2018, from <https://dl.acm.org/citation.cfm?id=1244102>

Cai, Y., Ji, D., & Cai, D. (2010, June). A KNN Research Paper Classification Method Based on Shared ... Retrieved March 20, 2018, from <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/07-NTCIR8-PATMN-CaiY.pdf>

Wang, B., Liao, Q., & Zhang, C. (2013, August). Weight Based KNN Recommender System. Retrieved March 20, 2018, from <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6642782>