**German Credit Worthiness**

Shoun Abraham, Vidit Kalani

Team Credit

https://viditkalani.github.io/GermanCredit.hithub.io/

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. This includes, but is not limited to: age, duration of credit, credit purpose, and job information. 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. With outside research and careful examination of the dataset, we hope to learn more about how credit worthiness is assigned and the various factors that are used for determining credit. We also hope to create a model to predict credit worthiness. Using the information from our dataset, our model will be able to predict a person's credit worthiness given all the attributes of credit worthiness. We can evaluate our model by splitting our dataset into a testing and training set and testing our model against the training set. Our dataset is large enough to support this method.

This data has two classes for credit eligibility: good or bad. There are predictions related to attributes, such as: checking account status, duration, credit history, purpose of the loan, amount of the loan, savings accounts or bonds, employment duration, Installment rate in percentage of disposable income, personal information, other debtors/guarantors, residence duration, property, age, other installment plans, housing, number of existing credits, job information, Number of people being liable to provide maintenance for, telephone, and foreign worker status.

This dataset is available from the UCI Machine Learning website or in R studio, via the Caret library.

<http://topepo.github.io/caret/data-sets.htmlin>

In this Caret Document, chapter 23, we found the German Credit Dataset without missing values.

<https://www.kaggle.com/uciml/german-credit/data>

Through Kaggle, we found the German Credit Dataset with missing values. We plan to use the Caret library German Credit Dataset because this dataset does not contain missing values.

We are using R-Studio to explore a machine learning related problem. The tidyverse and caret libraries are two libraries we plan to use.  We may use libraries for data visualization, such as the ggplot2 library. We may also need to implement the KNN and KNN-Recommendation Models to our data. Finally, if we switch to using the dataset with missing values, we may need to use the MICE library to replace the missing values using multiple imputations. Other tools and packages may be needed as we progress with this project.

Literature:

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>