Proposal
Credit Predictor

CS 375 - H01
Potri Abhisri Barama, pb558; Nandika Karnik, nk676; Ananya Tyagi, at899; Vibha Venkataraman, vv66

I. OBJECTIVE

The primary objective of this project is to identify and discover patterns and dive deeper into the finances of a sample of the German population. We hope to identify trends in terms of their spending and saving patterns, and if any of them stand out to us as being in the high-risk category. Through our analysis, we hope to accurately discover these patterns, because oftentimes they are not obvious at first glance. We also hope to categorize people based on how they finance their lives in terms of how often they take out loans, and their rate of paying it back (as it relates to their demographics as well).

II. BACKGROUND AND MOTIVATION

We began our search for a dataset that could have a positive impact once run through a model. We were drawn to the Brain Cancer dataset but after realizing another group had chosen it, we continued our search. We then came across the *Statlog (German Credit Data)* dataset. After researching it more and realizing this was an entire country's data we realized that our outcomes could have an impact on an entire country. We hope that we will be able to discover patterns that can help the German people improve their credit.

III. DATA SOURCE

We selected the Statlog (German Credit Data) dataset, which classifies people based on a set of attributes into good or bad credit risks. The dataset includes various features such as credit history, savings account/bonds status, employment, credit amount, and more. These attributes are used to predict the creditworthiness of individuals. The dataset is available in two formats, including an all-numeric version, and is accompanied by a cost matrix to help optimize classification performance.

IV. METHODOLOGY

A. Data Preprocessing & Feature Engineering: Reduce dimensionality of data to restrict to only meaningful data and to make processing and learning from it more efficient. For example, there are 20 features in this dataset. After taking a look at the dataset, perhaps features such as 'Property' and 'Housing' can be combined. Take out NaN and null values. This is to ensure that we are not letting those values skew our results. However, upon closer inspection, there doesn't seem to be any of these values so this step might not be necessary. We also plan to make every feature have a uniform number of data points to make working with it easier. We also will consider outliers and transform the data into a normal form to reduce variance. This will happen only if necessary so we don't lose information about the data that way. Create new features from patterns in the data to help the model recognize further patterns in the data. Our last step before moving on to the next step will be to split the data into training and testing sets.

B. Model: This is a nonlinear multi-classification dataset and as such we will have to use a model that is able to conduct multivariate analysis. After taking a look at the models with this capability in the 'Baseline Model Performance' section on the dataset website, we believe *Random Forest Classification* will be a good model to work with.

C. Evaluation Metrics: We plan to evaluate the model based on metrics such as Area Under the Curve (AUC), confusion matrices to measure accuracy, and calculate values such as Precision and Recall.

V. EXPECTED OUTCOMES

We expect to find trends relating to the types of customers, the reasons behind loans, and more such features. Using the model, we will be able to group individuals based on certain characteristics. This will reveal what kinds of customers take out certain kinds of loans or how they pay them back. The age and sex features will reveal whether there is a difference between why different age groups need loans, how much they borrow, and whether this differs by sex. By the end of this project, the model will yield various insights into the patterns of borrowers.

VI. TIMELINE

WEEK	TASKS	ALLOCATION
Week 6	Data collection and reprocessing	Ananya, Vibha
Week 7	EDA	Nandika, Abhi
Week 7	Related works	Vibha, Ananya
Week 8	Model implementation	Everyone
Week 10	Model evaluation and refinement	Everyone
Week 12	Final model section and report preparation	Everyone

Week 13	Project presentation and submission	Everyone

REFERENCES

[1] GeeksforGeeks, “Interpreting Random Forest Classification Results,” GeeksforGeeks, May 27, 2024. https://www.geeksforgeeks.org/interpreting-random-forest-classification-results/

[2] D. Jain, “Data Preprocessing in Data Mining,” GeeksforGeeks, Mar. 12, 2019. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[3] “What is Feature Engineering? - Feature Engineering Explained - AWS,” Amazon Web Services, Inc. https://aws.amazon.com/what-is/feature-engineering

[4] A. Hazra and N. Gogtay, “Biostatistics series module 10: Brief overview of multivariate methods,” Indian Journal of Dermatology, vol. 62, no. 4, p. 358, 2017, doi: https://doi.org/10.4103/ijd.ijd_296_17.