CS 7641: Predicting Credit Risk

Team (Group 4):

Ian Wood

Tejas Vermani

Hugh Weston

Christian Derolf

Introduction/Background

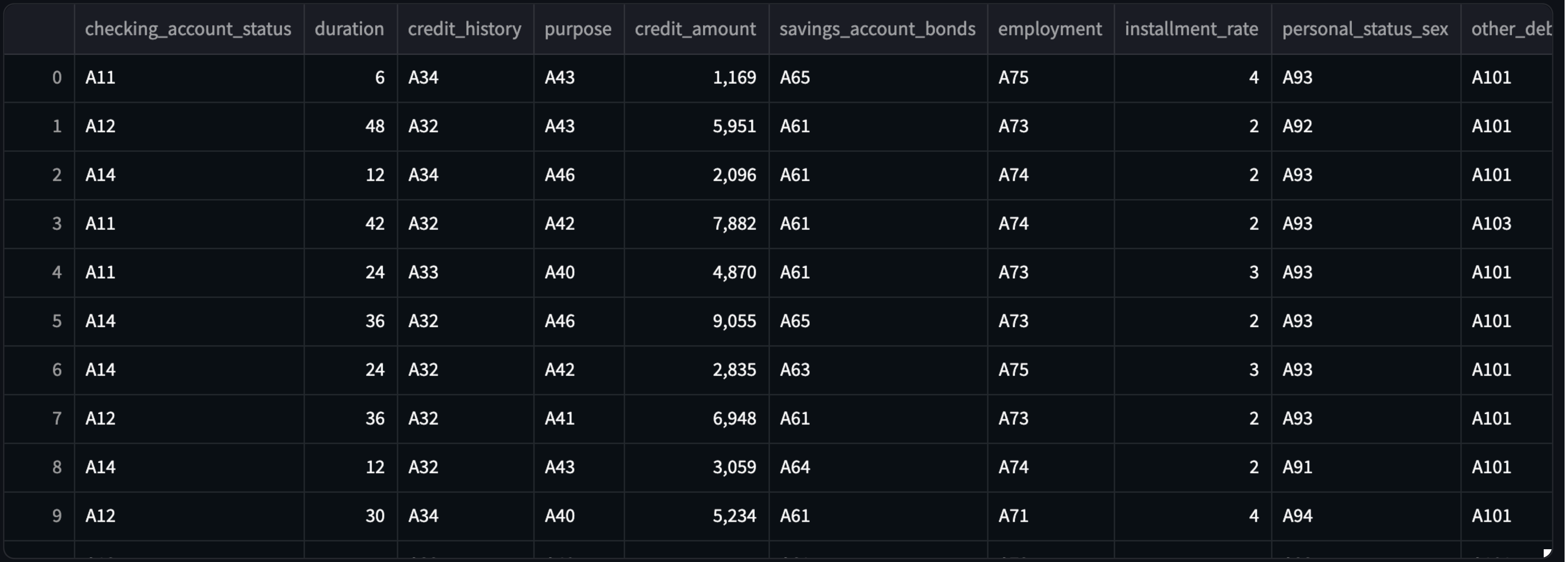
Literature Review:

Machine learning algorithms have become a crucial component in enhancing credit risk assessments. Approaches such as decision trees can identify reliable borrowers while minimizing false positives [1]. Research further suggests that combining models can mitigate incorrect classifications [2]. More recently, neural networks that learn from unstructured data have shown great potential [3]. Together, the application of these models can contribute to more accurate credit risk evaluations.

Dataset and description:

- This is a dataset from 1994, donated to UC Irvine by Dr. Hans Hofmann, detailing different characteristics of account holders at a in Germany. It contains features such as credit amount, credit history, credit duration, age, etc. It also has a binary metric for whether the specific account holder has “good” credit or “bad” credit. Each row within the dataset is a specific account holder.

- Dataset: https://www.openml.org/search?type=data&sort=runs&status=active&id=31



	checking_account_status	duration	credit_history	purpose	credit_amount	savings_account_bonds	employment	installment_rate	personal_status_sex	other_deb
0	A11	6	A34	A43	1,169	A65	A75	4	A93	A101
1	A12	48	A32	A43	5,951	A61	A73	2	A92	A101
2	A14	12	A34	A46	2,096	A61	A74	2	A93	A101
3	A11	42	A32	A42	7,882	A61	A74	2	A93	A103
4	A11	24	A33	A40	4,870	A61	A73	3	A93	A101
5	A14	36	A32	A46	9,055	A65	A73	2	A93	A101
6	A14	24	A32	A42	2,835	A63	A75	3	A93	A101
7	A12	36	A32	A41	6,948	A61	A73	2	A93	A101
8	A14	12	A32	A43	3,059	A64	A74	2	A91	A101
9	A12	30	A34	A40	5,234	A61	A71	4	A94	A101

Problem Definition

Problem:

- Lending is the backbone of the global economy. In order for financial institutions to lend money, they need proportionate reward for the risk they are taking. Taking on too much risk as a financial institution can lead to mass-defaults and recessionary periods.

Motivation:

- We desire to help financial institutions evaluate the risk associated with lending, to ensure a more stable economy. With our supervised learning methods, we aim to determine whether an individual is a “good” or “bad” credit risk, comparing it to a 21st target feature within the dataset (target). With unsupervised learning, we will seek to gain insights into the patterns or groupings of customers based on their creditworthiness, reveal distinct groups of customers, and reveal which features most contribute to variance in the data.

Methods

Data Prepocessing

To make sure the data is handled correctly as we pass it into our ML models and algorithms, there are certain preprocessing methods we need to employ to make sure the format of the data is correct and efficient. The exact methods are detailed below:

Missing Data:

- To handle potentially missing or invalid data points within our dataset we decided we may need to employ a method to replace these missing values.
- Specifically, something like the SimpleImputer class from the scikit learn library would solve this problem by replacing blank values with averages.
- However, upon researching online, the dataset we are using is high quality and therefore does not have any blank or missing values, so we did not need to employ these strategies.

Categorical Data:

- Our dataset has a large amount of categorical data for many of the entries, which is obviously not ideal for passing into ML models.
- We decided we needed some type of categorical data encoder to make these fields quantitative to pass into our model, either using OneHotEncoder from scikit learn or get dummies from pandas.
- For our dataset, we used get dummies from pandas to encode categorical data which works by creating a new column for each category and assigning a binary value for that column.

ML Models/Algorithms

We plan to use both supervised and unsupervised ML models to predicts different aspects using our credit risk dataset. Using supervised learning we plan to predict whether a user's credit risk should be GOOD or BAD using the labels for GOOD and BAD within the dataset itself. Using unsupervised learning, we plan to predict more fine-grained categories for credit risk beyond just GOOD and BAD. Here are the following models:

Supervised Learning:

- LogisticRegression model from scikit learn is a supervised learning model that will be used to help predict whether a user’s credit risk is GOOD or BAD based on features.
- RandomForestClassifier will also be used to help predict if a user’s credit risk is GOOD or BAD as it handles cases where certain features are more indicative of credit risk.
- Support Vector Machine (SVM) will also be used to predict credit risk off of existing labels.

- Quantitative metrics such as a confusion matrix will be used to compare the effectiveness of these models

Unsupervised Learning:

- GaussianMixture will also be used from scikit learn as an unsupervised learning to help identify a more nuanced form of credit score beyond just a binary value of GOOD or BAD.
- Kmeans will also do soemthing similar in terms of clustering but using a different method
- The effectivenesss of these techniques will be measured using quantitative metrics using PCA plots and silhouette scores.

Results and Discussion

Expected Results

- Supervised Learning Metrics: Accuracy: measures the percentage of correct predictions out of total predictions. *F1-Score: the harmonic mean of precision and recall and is useful for imbalanced data to strike a balance between false positives and false negatives.
- Unsupervised Learning Metrics: Silhouette Score: This is a common metric for clustering algorithms, measuring how well data points fit within their cluster.
- Goals: We want to maximize accuracy and F1 scores in our classification models, ensure the models can generalize well to new data, and use clustering algorithms to find distinct groups of credit profiles.
- Expected Results: Supervised Learning: Models like Random Forest will perform well, given the structured nature of the data. We anticipate that precision and recall metrics will be reasonably high due to the clarity of creditworthiness indicators. Unsupervised Learning: We expect clustering algorithms to group clients based on their credit profiles into categories like high-risk, medium-risk, and low-risk.

Supervised Models

Train supervised models

Supervised Analysis

Implementation

In this project, we implemented supervised learning models to predict credit risk using a dataset containing various features related to credit profiles. The primary goal was to classify credit risk into categories such as "Good" and "Bad" based on the provided features. We utilized two popular machine learning algorithms: Logistic Regression and Random Forest.

The process began with data preprocessing, where we converted categorical variables into a numerical format using one-hot encoding. This step ensured that the machine learning models could effectively process the data. We then split the dataset into training and testing sets, with 80% of the data used for training and 20% reserved for testing. This split allowed us to evaluate the models' performance on unseen data, ensuring their ability to generalize.

We trained Logistic Regression, Random Forest, and SVM models on the training data and evaluated their performance using various metrics. Specifically, we calculated training and testing accuracies to assess how well the models fit the training data and how well they generalized to the testing data. Additionally, we generated classification reports and confusion matrices to provide a detailed breakdown of the models' performance, including precision, recall, and F1 scores.

When we trained the logistic regression model, we would get a "ConvergenceWarning" error. This error occurs when the model fails to converge to a solution within the maximum number of iterations. To address this issue, we increased the maximum number of iterations to 1000, allowing the model to converge successfully, and we used StandardScaler to scale the features before training the model. This got rid of the ConvergenceError, and improved both training and testing accuracy by ~2%.

To visualize the results, we plotted confusion matrices and ROC curves for each model. The confusion matrices helped us understand the distribution of true positives, true negatives, false positives, and false negatives, while the ROC curves provided insights into the models' ability to distinguish between the "Good" and "Bad" credit risk categories. By comparing these metrics and visualizations, we could identify potential issues such as overfitting or underfitting and make informed decisions about model improvements.

Results

There were a few results of note that we observed from our supervised learning models:

- The testing accuracy of our logistic regression model is slightly higher than that of our random forest model, implying a better balance. This may be because we used StandardScalar from the scikit-learn package, which ensures that all features contribute equally to the model, not allowing features with larger ranges to dominate the learning process.
- The relatively low testing accuracy of the Random forest model can be explained, however, by its tendency to overfit to the training dataset, which we did not address (yet).
- The accuracy of logisitic regression was higher on the testing dataset than the training dataset. By default, the scikit-learn LogisticRegression function applies L2 regularization, which is meant to intentionally underfit a model to the training dataset to avoid overfitting, which is why the model performs as well, even better on the testing dataset.
- SVM had similar metrics for test accuracy as logistic regression, but a slightly higher train accuracy most likely due to the fact it is less underfit.
- The Random Forest model had a higher accuracy on the training dataset than the testing dataset (100% vs ~78%). This is likely due to the model overfitting the training data, which is a common issue with Random Forest models.
- Random Forest implementation had higher precision for the "Bad" classification, but lower recall, and vice versa. This indicates that the model is better at identifying "Bad" credit risks but may miss some "Bad" credit risks in the process. We can attribute this to the model's tendency to overfit the training data, as mentioned earlier.

- Based on the confusion matrix, logistic regression results in the least false positives, which is highly important in the credit assessment industry as the most economic damage/risk is a result of false positives.
- SVM usually results in the most false positives, conversely

Unsupervised Models

Set Parameters for Unsupervised Models

Train unsupervised models

Unsupervised Learning

Implementation and Analysis

We implemented unsupervised learning models to predict credit risk using a dataset containing various features related to credit profiles. We wanted to extend beyond the binary classifications of good and bad credit risk with our unsupervised learning models. Instead we hoped to find groupings more similar to good, medium, and bad. We utilized two popular unsuperised learning algorithms: K-Means and Gaussian Mixture Model (GMM). We again used StandardScalar and one-hot encoding for preprocessing, but this time disregarded the final column altogether as that contained the labels for supervised learning. We attempted to cluster datapoints with KMeans and GMM into a user-specified number of clusters. Both methods yielded relatively low silhouette scores, indicating poor clustering performance. This suggests that the data may not naturally form distinct groups suitable for these clustering techniques. We also attempted to implement DBSCAN to detect clusters of arbitrary shapes and identify noise points. However, due to the high dimensionality and density of the dataset, DBSCAN struggled to find meaningful clusters without labeling most points as noise. Overall, the unsupervised methods were not particularly effective for this dataset, due to challenges such as overlapping features and lack of clear cluster separation, which make unsupervised clustering less suitable compared to supervised approaches for predicting credit risk in this context. Furthermore, industry standards for credit risk assessment are typically based on supervised learning models, which have the advantage of being able to learn from labeled data and make predictions based on known outcomes. Those in industry would also prefer the flexibility of defining what constitutes a good or bad credit risk, rather than relying on unsupervised clustering to determine these categories.

Results

There were a few results of note that we observed from our supervised learning models:

- Kmeans visually performs better on 3 cluster but has a worse silhouette score, we find that the outliers present in the data obfuscate this metric slightly
- The spherical nature of the data and well chosen centroid can be reasons for K-means' improved performance over GMM

- When we ran the unsupervised models with 2 clusters, we did notice something of note; data points were clustered by the column "purpose", where the value A410 indicated "other" as the purpose, as opposed to the defined categories (car (new/used), furniture, radio/tv, domestic appliance, repairs, education, business, vacation).
- This indicates that the purpose of the loan is a significant factor in determining credit risk, which is consistent with industry standards.
- Aside from a group of outliers, it is overall quite challenging to assign non-binary classifications to credit risk due to the overall density of the dataset
- The silhouette scores are relatively low. When identifying 3 clusters using GMM and K-means, correlating to Good, Medium, and Bad credit risk, we received a silhouette score of 0.1 and 0.05 respectively. This indicates that the data was not clustered very effectively using our unsupervised methods.
- The computational load of performing the unsupervised analysis was significantly higher than the supervised analysis and also did not provide many particularly impactful insights. If the project were to be continued, an emphasis on supervised modelling would be prudent as the time, complexity, and results are all more favorable.

References:

[1] A. G. Roy and S. Urolagin, “Credit Risk Assessment Using Decision Tree and Support Vector Machine Based Data Analytics,” Advances in Science, Technology & Innovation, pp. 79–84, 2019, https://doi.org/10.1007/978-3-030-01662-3_10.

[2] S. Agrawal, P. Ahirao, S. Kumar, and P. Dere, “Credit Score Evaluation of Customer Using Machine Learning Algorithms,” SSRN Electronic Journal, 2021, https://doi.org/10.2139/ssrn.3867420.

[3] M. Megdad and S. Abu-Naser, “Credit Score Classification Using Machine Learning,” International Journal of Academic Information Systems Research (IJAISR), vol. 8, no. 5, pp. 1–10, 2024. https://philarchive.org/archive/MEGCSC-2.

	Member	Contribution
0	Tejas	Gantt Chart, Introduction, Potential Results and Discussion, Quantitative Metrics, Analysis of Supervised Models
1	Ian	Methods, Visualizations, Analysis of Supervised Models, Next Steps
2	Christian	Literature Review, Problem Definition, Dataset description, Data Preprocessing, Implementation of Supervised Models
3	Hugh	Streamlit deployment, Dataset description, Data Preprocessing, Implementation of Supervised Models

Download Gantt Chart

Download Gantt Chart