Capstone Project

Financial Risk Forecasting: An Introductory Analysis of German Credit Data

Strahinja Nakic (500809487)

strahinja.nakic@torontomu.ca

Toronto Metropolitan University

Professor Tamer Abdou

June 28, 2025

CIND 820-DAH

Abstract

This project focuses on using classification to identify credit risk. Banks and lenders lose money when people do not repay their loans, so it is important to figure out who might default before giving out credit. The analysis will use the German Credit dataset from the UCI Machine Learning Repository, which includes 1,000 applicants and 20 attributes related to their personal and financial background, such as job status, credit history, and loan purpose. The first step will be to clean the data, explore patterns, and convert text categories into numbers. After that, the plan is to build and test three basic models: Decision Tree, Logistic Regression, and Naive Bayes. Their performance will be compared using accuracy, recall, and confusion matrices.

My main research questions are the following:

1. Which personal or financial features are most linked to credit risk?
2. How well do basic classification models predict whether someone is a good or bad credit risk?
3. What can this process teach us about larger topics in financial risk and anti-money laundering?

This topic is meaningful to me because of my background in billing and revenue at KPMG and Canadian Tire. I am now moving toward a career in risk and anti-money laundering. I believe that this project is a good step toward learning how financial behavior can be understood through data, and it helps prepare me for future work in compliance and fraud detection.

Article 1 – Bagging Supervised Autoencoder Classifier for Credit Scoring (Abdoli, Akbari, & Shahrabi, 2021)

Credit risk is both a practical and conceptual issue. It has always demanded approaches that can evolve alongside the increasing complexity of consumer behavior. In their 2021 paper, Abdoli, Akbari, and Shahrabi propose a hybrid machine learning model designed to improve the accuracy of credit classification systems. What sets their work apart from other machine learning experiments is not just the use of advanced techniques, but the conscious attempt to strike a balance between feature extraction and model performance. The authors introduce a “bagging-based” supervised autoencoder classifier, or BSAC, that blends neural embedding with ensemble modeling (Abdoli, Akbari, & Shahrabi, 2021). The way the model is constructed signals a deeper understanding of the shortcomings of traditional algorithms, especially when working with datasets as noisy and categorical as the German Credit dataset. What makes their article so important and thereby serving as my primary article, is not only its originality but also its practical relevance. It doesn’t merely present a theory, it tests it directly on the German Credit dataset, which adds a level of credibility to the findings, especially for students or researchers working on similar real-world data like in this capstone (Abdoli, Akbari, & Shahrabi, 2021).

The authors approach the problem from two angles. First, they use supervised autoencoders to distill important features from the original dataset. Rather than feeding every single variable directly into the model, an autoencoder helps compress the input data into a smaller set of features. They then apply bagging, a type of ensemble strategy that uses multiple classifiers to reduce variance and boost model stability. The end result is a hybrid model that reportedly outperforms traditional techniques like logistic regression and Naive Bayes (Abdoli, Akbari, & Shahrabi, 2021). Though the paper leans heavily into complex modeling, it doesn't lose sight of core principles like performance evaluation. The models are tested using metrics that actually matter to financial professionals, like precision, recall, and F1-score. These methods go above just merely chasing just overall accuracy. These are the same metrics planned for this capstone, which is part of what makes this article a useful guide for validating those choices (Abdoli et al., 2021). This article is not something I'd model entirely. The depth and complexity of the techniques used such as autoencoders and ensemble classifiers aren't ideal for a beginner-level project. The value here is not in replicating what Abdoli et al found, it's in the context. It confirms that the German dataset remains one of the best small scale credit datasets available for educational and experimental work. More importantly, it offers a clear example of how credit data can be reshaped, cleaned, and filtered to improve outcomes and paint a wider picture of the actual data (Abdoli, Akbari, & Shahrabi, 2021).

While I won't be building a supervised autoencoder from scratch, I do plan to incorporate some of the preprocessing logic and model evaluation ideas into the simpler models that form the backbone of this project. In many ways, this article represents the upper limit of what can be done with the dataset if one were to pursue a more complex route. Its results also serve as a benchmark for what can be achieved in terms of accuracy and model robustness. The authors argue that their model achieves better performance not just because of the algorithm itself, but because of how the features are selected and reshaped (Abdoli, Akbari, & Shahrabi, 2021). This insight strengthens my understanding of the importance of thoughtful data preparation. Lastly, the significance of this article lies in how it frames the classification problem. It doesn't just throw algorithms at the data; it considers structure, meaning, and variability in the process. That attention to nuance is what elevates the paper beyond just another machine learning comparison. For this project, which aims to apply foundational models like Decision Trees, Logistic Regression, and Naive Bayes to the same dataset, this paper serves as an essential point of reference. It offers a vision of what is possible while reinforcing the choices being made for a simpler, more transparent, and AML-conscious approach to credit risk classification (Abdoli, Akbari, & Shahrabi, 2021).

Article 2 – Credit Scoring for Good or Bad: A Comparative Study (Baesens, Van Gestel, Stepanova, Suykens, and Vanthienen, 2003a)

The question of how best to predict creditworthiness has long been central to the work of financial institutions, and Baesens et al. (2003a) offer a foundational contribution to this discussion by comparing a range of statistical and machine learning techniques in the context of credit scoring. What makes this paper especially useful is not just the number of models it compares but the scope and depth of its analysis. The authors examine models such as logistic regression, support vector machines, decision trees, neural networks, and Bayesian classifiers, evaluating their strengths and weaknesses using real financial data such as chequing account status, Credit history, etc. Their conclusions prove that no single model outperforms the others across all datasets or business goals (Baesens et al., 2003a). That insight aligns perfectly with the approach taken in this capstone, which is to compare several foundational models without assuming that one will be universally superior to the other. A helpful aspect of the study is that it includes the German Credit dataset as one of the benchmarks for testing the models. This validates the relevance of that dataset and provides a concrete baseline for evaluating my own model results. For example, if my logistic regression model achieves a similar accuracy to what is reported in the study, it gives me confidence that my data preparation and modeling decisions are on the right track (Baesens et al., 2003a).

In addition, it provides perspective on the expected range of outcomes and allows me to interpret my own results with greater context. Their use of the German dataset also reinforces its usefulness for the scope of this academic project as finding access to real bank data is limited. What Baesens et al. (2003a) do well is shine light to the trade-offs that exist between accuracy and interpretability. For example, while neural networks may offer slightly higher predictive performance in some scenarios, they often lack the transparency needed in regulated environments (Baesens et al., 2003a). This reflects the realities faced by professionals in the compliance, fraud prevention, and anti-money laundering space, where it is important to have an understanding of how a model arrives at a decision. This alone can be just as important as the decision itself. For this reason, the study shows the value of simpler models like logistic regression and decision trees (Baesens et al., 2003a). These models are easily interpretable but also tend to be more stable and easier to deploy in production settings. That way of thinking is central to my project, which prioritizes model clarity and real-world usability over theoretical complexity. The article also stands out for its discussion of evaluation metrics needed for a favorable outcome. Instead of just relying solely on accuracy, the authors incorporate a wide range of metrics, including AUC, precision, recall, and the Gini coefficient. This perspective on evaluation confirms my own decision to look beyond accuracy and consider confusion matrices and recall as primary indicators of model success (Baesens et al., 2003a). In highly imbalanced datasets, where the number of good credit cases often outweighs the bad, focusing only on accuracy can create a false sense of model performance. Having only a few customers that are classified as a bad risk is still not an acceptable metric to present. The inclusion of these metrics also helps us understand what a good model looks like, especially in applications where the cost of misclassification can be high. Misclassification can put huge financial risk on banks and deprive clients of access to credit (Baesens et al., 2003a).

Finally, the authors touch on the importance of data preparation. Baesens et al. (2003a) make it a point to stress that the quality of input data has a significant impact on model outcomes, often moreso than the choice of the algorithm itself (Baesens et al., 2003a) This insight is relevant for my capstone because I am working with categorical and mixed-type features that need to be encoded and transformed properly to reflect quantifiable results. To conclude, Baesens et al. (2003a) provide a thorough and methodical evaluation of the models most used in credit scoring in financial institutions. Their findings support my methodological decisions, reinforce my evaluation strategy, and strengthen the rationale for using the German Credit dataset. The paper is valuable not just for what it teaches about models, but for how it encourages thoughtful and precise decision making in the modeling process. The article serves as an important reminder that predictive power alone is not the end goal in credit risk analysis. The models we build must also be understandable, repeatable, and ethically sound. For these reasons, this article plays a central role in my literature review and guides many of the design choices in this project (Baesens et al., 2003a).

Article 3 – Machine Learning in Consumer Credit Risk Analysis: A Review (Warse, 2020)

There is a growing need for machine learning applications within consumer credit scoring. To put it simply, this is not just a trend, but a reflection of how rapidly the financial services industry is evolving. Recently, TD has faced a historic U.S. Department of Justice (DOJ) fine of nearly 3 billion $USD over their negligence in cracking down on known money laundering within its own institution. In this review article, Warse (2020) dives into the performance of various classification algorithms for predicting credit risk, drawing on benchmark datasets that include the German Credit dataset, which is also the foundation of this capstone project. The usefulness of this article doesn't lie in proposing a new model, but in offering a structured comparison of multiple machine learning techniques and establishing what works best under real world data conditions. The article covers a wide range of models from logistic regression and decision trees to support vector machines, k-nearest neighbors, and neural networks. Each of these models is tested and compared using consistent evaluation metrics (Warse, 2020). This method offers a fair and objective look at each model's strengths and weaknesses. One of the most important takeaways from the paper is the confirmation that models like logistic regression and naive Bayes, despite being simpler, are far from obsolete. They perform well when compared to more sophisticated models like for example, neural networks and ensemble methods. This is useful for a beginner focused project such as this one, where the goal is not to implement the most complex system possible, but to develop a solid understanding of model building and evaluation using interpretable methods (Warse, 2020).

The author's findings support the idea that good predictive performance can be achieved even without heavy tuning or advanced algorithms. This helps build confidence in my decision to rely on logistic regression, decision trees, and naive Bayes as my primary tools for this capstone. The article places a significant emphasis on the importance of preprocessing, especially when dealing with datasets that have a mix of categorical and numerical features (Warse, 2020). Warse (2020) explains that how the data is prepared has a major effect on the outcome of the model. This includes steps such as normalization, handling missing values, and transforming categorical variables into numeric format. These are the steps planned in this capstone project. Warse (2020) devotes a large portion of the discussion to this stage of the process and this is justifiable given the time and focus that will be spent preparing the German Credit dataset before model training can begin. Another important point made in the article is that accuracy alone should not be the sole performance measure. Warse (2020) argues for an approach that includes metrics like precision, recall, F1 score, and the area under the Receiver Operating Curve (ROC). The reason these metrics are important in imbalanced datasets is because credit scoring tends to levy heavily towards one side where one class, most often good credit, tends to dominate (Warse, 2020). Only focusing on accuracy makes it possible to miss models that are better at identifying the smaller but more critical class of risky borrowers (Warse, 2020). This aligns directly with the plan for this capstone to use confusion matrices, and recall in evaluating model performance, making this article relevant and necessary in shaping the evaluation strategy (Warse, 2020).

The models tested in the article include more complex options like support vector machines and neural networks, Warse (2020) does acknowledge the interpretability trade-off. Meaning, while these models may slightly outperform simpler algorithms in some cases, they are often harder to explain and therefore less useful in environments where accountability and transparency matter (Warse, 2020). To put it simply, if these results cannot be explained easily, they are unlikely to live up to scrutiny. This is true in domains related to compliance and anti money laundering, where model decisions need to be explained to financial auditors or regulatory bodies. Warse (2020) endorses the use of simpler models, specifically logistic regression and decision trees. The techniques the author employs, specifically in using cross validation to guarantee fair performance comparisons is a sound method to achieve validation in model development (Warse, 2020). Warse also argues that applying cross validation across all models guarantees that the results are not just due to random chance or overfitting. The methodological consistency gives this article more than just academic weight. The study becomes a blueprint for how to properly test and compare machine learning models in real world applications. To conclude, Warse (2020) presents a grounded and practical overview of classification models in credit scoring (Warse, 2020). The paper supports my selected modeling techniques, informs my evaluation plan, and highlights the relevancy of data preparation before any model can be created. Therefore, this article affirms that quantifiable insights and strong predictive results can be obtained using standard tools, provided these tools are applied thoughtfully.

Article 4 – A Hybrid Model of Self-Organizing Maps and Support Vector Machines for Credit Risk Classification (Ala'raj and Abbod, 2016)

When it comes to financial classification, especially in credit risk analysis, there is always a pressing need for models that are not just accurate, but intelligent in how they process and segment information. Ala'raj and Abbod (2016) respond to this challenge by proposing a hybrid model that combines Self Organizing Maps with Support Vector Machines. This article stands out because it does not rely solely on a single algorithm to do the heavy lifting. The article separates the classification process into two stages. First, the "Self Organizing Maps" are used to reduce the complexity of the data by grouping cases together. Then, these grouped features are sent into a "Support Vector Machine" classifier for the final classification. This type of layered approach is both thoughtful and strategic **(Ala'raj and Abbod, 2016)**. It suggests that classification is not just about prediction accuracy, but also about understanding structure and eliminating noise from the data before trying to reach a conclusion **(Ala'raj and Abbod, 2016)**. What makes this article particularly useful is that the German Credit dataset is one of the benchmarks tested in the study. Since this capstone also uses that dataset, the findings become a lot more applicable. I can see how the hybrid model performs on the same data which means I can better understand the trade-offs between complex ensemble models and more straightforward algorithms **(Ala'raj and Abbod, 2016).**

Although my own project will not replicate the full complexity of this approach, the article still offers important context. It shows how breaking the modeling process into logical stages can improve the quality of predictions. It also validates the idea that careful data transformation and segmentation before modeling can be just as important as the algorithm itself.

One of the most useful lessons this article teaches is that good classification is not just about feeding raw and unprocessed data into a model. The authors also demonstrate how removing redundancies and clustering similar features improves both efficiency and accuracy **(Ala'raj and Abbod, 2016)**. I believe this further supports my decision to invest heavily in preprocessing the German Credit dataset before applying any predictive models. I will not use a "Self Organizing Map", but the idea of grouping similar values, reducing dimensionality, and treating data patterns seriously is completely transferable. Even with a Decision Tree or Logistic Regression model, cleaning and restructuring the data before modeling is going to be a core part of this capstone's methodology. Another valuable contribution of the paper is how it emphasizes model evaluation. The authors do not just stop at overall accuracy **(Ala'raj and Abbod, 2016)**.

This article is also relevant because it touches on model transparency. The authors acknowledge that models used in the financial space need to be explainable. While SVMs are known for being accurate, they can be difficult to interpret. That tension between accuracy and transparency is at the heart of this capstone, especially considering my long term goal of working in compliance and financial crime prevention **(Ala'raj and Abbod, 2016)**. In regulated environments, a model that cannot be explained clearly will always be a liability, no matter how well it performs. To conclude, Ala'raj and Abbod (2016) offer a different approach to credit risk classification that combines and merges the strengths of unsupervised and supervised learning **(Ala'raj and Abbod, 2016)**. Their hybrid model outperforms many of the standard techniques by structuring the learning process more thoughtfully. While the complexity of their solution is beyond the scope of this capstone, the ideas presented here reinforce everything that is important about this project. That includes careful data preprocessing, balanced model evaluation, and an ongoing awareness of how model decisions translate to real world outcomes. This article aids in my literature review by showing how even a small dataset can be used to build an advanced model when thoughtful methods are applied.

Article 5 – Benchmarking Classification Models for Credit Scoring (Baesens, Setiono, Mues, and Vanthienen, 2003b)

This article by Baesens et al. (2003b) stood out to me because it helped me understand that sometimes sticking to simpler models is completely fine. The authors do not try to introduce anything new. Instead, the authors test a variety of different credit scoring models using real financial data, including the German Credit dataset that I am using for this project. They go through models like decision trees, support vector machines, logistic regression, neural networks, and Bayesian networks. What I found fascinating is that they don't just focus on showing which model is best at evaluating credit risk. They look at things like accuracy, recall, and other metrics that help explain how each model performs overall. One thing that was clear from the article is that there is no perfect model **(Baesens et al., 2003b).** While neural networks did slightly better in some cases, models like logistic regression and decision trees still performed really well and were easier to work with. That gave me some reassurance because my project is not using the most advanced techniques, but instead focusing on something that I can understand and explain. Baesens et al. basically say that if a model is too hard to understand, it is less useful in real financial environments. This is especially true in areas like compliance and fraud where someone might have to explain how a decision was made **(Baesens et al., 2003b).**

The authors also discuss the importance of preparing data the importance of preparing your data. That really stuck with me because I’ve already started cleaning and encoding the dataset for this project, and I realized it’s not something that can be rushed. They mention how poor data preparation can make even the best model perform badly, and I’ve seen how messy data can not give you the results you need to conduct your work **(Baesens et al., 2003b)**. It made me feel more confident about spending time on cleaning and transforming the dataset before delving into modeling. Another point they make is that models need to be stable and easy to update over time **(Baesens et al., 2003b)**. In real life, credit scoring systems don’t just get built once and left alone. They are used for years and often adjusted. Simpler models like decision trees or logistic regression can be updated more easily and are more transparent. This kind of practical advice helped me feel better about choosing basic models for my capstone instead of trying to force in something I do not fully understand **(Baesens et al., 2003b)**. Overall, this article helped confirm that the approach I am taking with this project makes sense. I want to build models that are easy to explain, that work well enough, and that are built on clean data. Baesens et al. (2003b) back all of that up with actual results, which is why I think this paper plays a key part in supporting my research.

Article 6 – The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients (Yeh & Lien, 2009)

This article by Yeh and Lien (2009) is insightful because it compared a bunch of different models for predicting credit card defaults, which is similar to what I will be doing in this project with credit risk classification. Regardless of the fact that their focus was on credit card customers and not loans like in the German dataset, the logic and modeling process are the same. They use a dataset from a Taiwanese bank with data on 30,000 clients and run models like decision trees, logistic regression, neural networks, k-nearest neighbors, and support vector machines **(Yeh & Lien, 2009).** What I found interesting is that they did not just say one model is superior to another. They tested all of them properly using cross validation and looked at different metrics like precision, recall, F1 score, and accuracy. In the end, the neural network had the best performance overall, but logistic regression and decision trees were not far behind **(Yeh & Lien, 2009)**. For my project, that helps justify why I am using those simpler models. I don’t need a neural network to build something decent, especially when I’m still learning and want to actually understand how the model works.

Another reason this paper was useful is because it talks about how important preprocessing is before you even train a model. They explain how models that need numerical inputs, like SVMs or k-nearest neighbors, require proper normalization **(Yeh & Lien, 2009)**. That reminded me that I need to pay extra attention when converting all the categorical data in the German dataset into something the model can understand. Even though I’m using models that are less sensitive to scaling, I can’t skip that part. Yeh and Lien make it clear that messy or poorly processed data will hurt your model no matter how advanced it is. The paper also makes a strong point about balancing performance and interpretability. They bring up the idea that even if a model like a neural network performs better, a simpler model like logistic regression might still be more useful in a real business or financial setting **(Yeh & Lien, 2009)**. That’s something I’ve been thinking about a lot, especially since I want to move into risk or AML after this program. If I build something I can’t explain, I’m not going to be able to use it in the type of job I want. So seeing that this paper supports that idea was a good sign. Lastly, this article gave me a good baseline for how well my models should perform. If Yeh and Lien's decision tree model performs with a high level of accuracy, I can compare my results with theirs and see if I am on track **(Yeh & Lien, 2009)**. Even though their dataset is different, it still helps me to know what kind of performance range I should expect. That gives me something to work toward when I start testing. This article further backed up my choices and gave me more confidence to focus on clean data and simple, explainable models. It also helped me think more carefully about which evaluation metrics to use, especially when dealing with imbalanced datasets like the one in this project.

Research Questions

The primary focus of my project is to understand how I can use machine learning models to predict credit risk using the German Credit dataset. The models I picked are Logistic Regression, Decision Tree, and Naive Bayes. These models are simple, and backed by previous research that confirms they can perform well in financial classification tasks such as this. I've also had the opportunity to work with these models in previous assignments I had done for this certificate program. I developed three research questions that will guide my analysis. The first question I am exploring is "Which personal or financial features are most indicative of credit risk in this dataset?" Since the dataset contains several variables like employment status, credit history, purpose of the loan, housing situation, and age, I want to understand which of these features are the most important when predicting if someone will be a good or bad credit risk. This will be done by looking at feature importance in decision trees and how coefficients behave in logistic regression. By answering this question, I hope to make the predictions more meaningful and maybe even connect them to broader financial behavior patterns.

My second question is "How well do simple classification models perform in predicting creditworthiness?" The goal here is not to create the most complex or high performing model. That is well beyond the scope of this project. Instead, I would like to see how these simple models stack up in terms of accuracy, recall, and other evaluation metrics. This is important because in real world setting like banking or compliance, models need to be understood and explained to auditor and regulators. If a simpler model can perform well enough, it might actually be more useful than a complicated one. I want to see if that idea holds up when tested.

Lastly, my third question is: "Can insights from this project be connected to financial risk analysis or anti money laundering?" As I stated previously, project is focused on credit scoring, but the wider scope of this assignment for me is to understand financial behavior and how data can help flag risks before they happen. I want to explore whether the patterns I find here can also be relevant in other areas like fraud detection or customer due diligence (CDD). The dataset itself may be limited to credit risk, the techniques and insights might help build a foundation for future work in risk and AML. These three questions will shape how I explore the data, train the models, and interpret the results. It also connects back to both the six articles I reviewed and the career path I am currently aiming for.

Data Description

The dataset used in this project is the German Credit dataset from the UCI Machine Learning Repository. It contains 1000 records, each representing a loan applicant, and includes 20 features that describe personal and financial characteristics such as employment status, credit history, age, savings, housing, and more. The target variable is labeled as either one for good credit or two for bad credit. I plan on using the version of the dataset with categorical values meaning the non-numeric one. I chose this one because I felt that this one gives a more real world use case, as bank data is categorial by nature. The first thing I had to look into was how to prepare the data for machine learning models that require numerical input. Every column is stored as text or category values, so the data will have to be converted into a numerical format before any modeling can be done. I will do this by encoding the categorical variables into a numerical format. I have confirmed that there are no missing values in the dataset, so I will not need to worry about filling in or removing any incomplete rows, I will however include this step as I believe it is best practice to confirm this through the model and code. As this is compliance work, you can never be too careful when handling sensitive financial data such as this. I did review the dataset and it was clear that some features might be more useful than others. Features like “credit history” and “chequing account status” are directly related to credit risk. Categories like “foreign worker” or “telephone ownership”, might not carry much predictive power on their own.

The goal in this project is to test these assumptions and see what the models actually learn from the data. I will be looking at the distribution of the target variable, which is slightly skewed in the “good credit direction. Approximately 700 applicants are considered good credit risks, while about 300 are labeled as bad. This imbalance matters because it can affect how well the models predict the minority class. I will have to consider this when evaluating performance and possibly use methods like class weights or stratified sampling during training which I delved into in previous courses. I plan to calculate some basic summary statistics like average age, median loan amount, and the most common values in each categorical column. This gives me better understanding of the dataset before delving into modeling portion of this project. Any unusual values or outliers will be documented and reviewed to see if they need to be handled differently. Overall, the dataset is small but extremely detailed for my use case and evidently has been used in studies that are far more advanced than the scope of my own. I believe it will be more than enough to train and compare models, and it provides a realistic case for studying credit scoring in a way that is expected for a capstone level project.

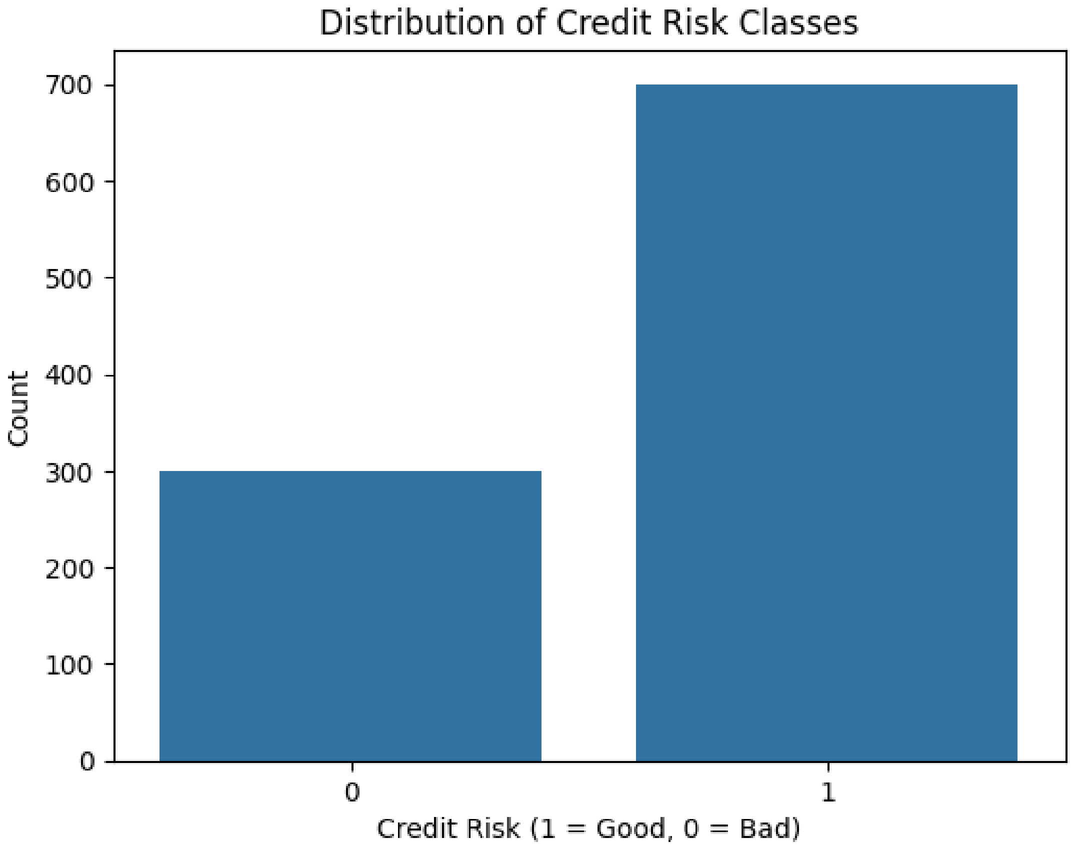

Figure 1: Distribution of the target variable shows that the dataset is moderately imbalanced, with a higher number of good credit cases (1) compared to bad credit cases (0).

Exploratory Analysis

I spent some time looking into the dataset to understand what I was working with. Most of the variables are categorical so I began to look at the frequency of values in each column. As an example, in the chequing account status column, most applicants had either no chequing account or a very low balance. This already hinted that many people applying for credit might not have strong financial backing, which could be useful when trying to predict risk. I also noticed that loan purposes like furniture and new car were the most common, while things like education or repairs were less frequent. When looking at age, I found that most applicants were between their mid twenties and early forties. There were a few older individuals, but they were not the majority. The average loan amount also varied a lot, but there were definitely some people asking for large loans compared to the rest. These higher values could be outliers, or they could reflect real financial need. I will keep an eye on those rows during modeling.

Another pattern I saw was that most people were renters, with fewer owning property or living with family. It will be interesting to see if that plays a role in credit decisions. I also checked the balance of the target variable. As I mentioned earlier, out of 1000 total cases, around 700 were labeled as good credit and about 300 as bad. This means the data is imbalanced, and it could affect how well the models identify bad credit risks. If I just go by accuracy, the model might lean toward always predicting good credit since that is the majority class. This is why I plan to rely more on metrics like recall and confusion matrix to get a better understanding of how well the models are really performing. Overall, this early exploration helped me make sense of the data and gave me a better idea of which features might be important. It also gave me a few ideas on how I can clean and transform the data before moving on to the next step.

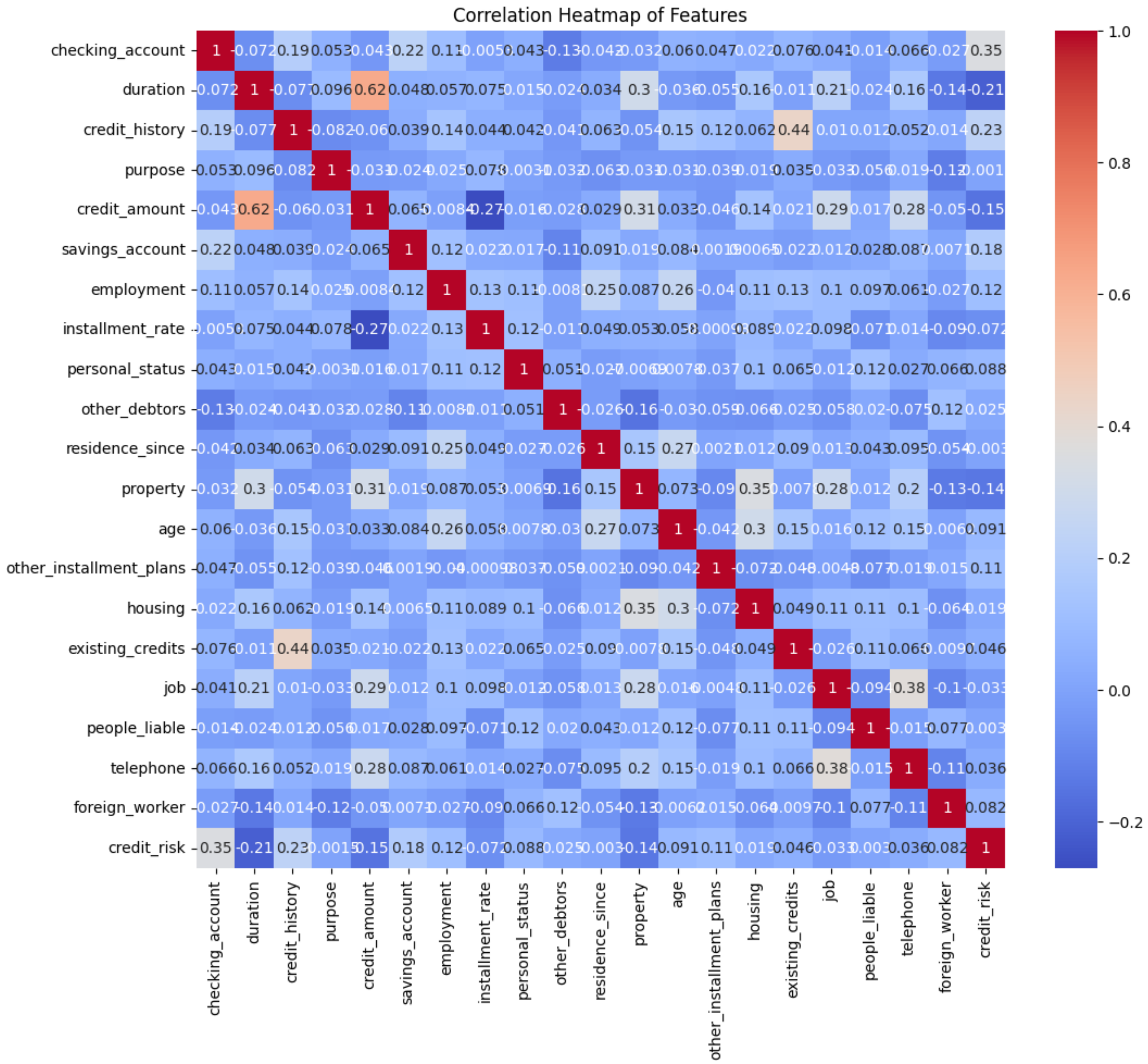

Figure 2: Heatmap showing pairwise correlations between features. Most variables show weak correlations, indicating minimal multicollinearity.

Techniques and Tools

For this project, I plan on using three classification models. These models were mentioned in almost every research paper I reviewed. These are Logistic Regression, Decision Tree, and Naive Bayes. I chose these models because they are easy to understand, they do not require advanced tuning to work, and most importantly, they give results that can be explained. Based on what I have read and the size of the German Credit dataset, I believe these models are enough to build something useful and within the scope of this assignment. Logistic Regression is often used for binary classification problems and works by estimating the probability that a given input belongs to a certain class. In this case, it will help me predict whether a person is a good or bad credit risk. The output of logistic regression is easy to interpret and the coefficients show how each feature affects the final prediction. That is helpful when trying to explain why the model gave someone a certain risk score. This is the simplest and straight to the point model so it serves as a good starting point for me.

The second model I will use is the Decision Tree. I like this model because it builds a tree like structure where each node represents a decision based on one feature. This makes it easy to follow the logic and see how the model splits the data. Decision Trees are also good at handling both categorical and numerical features without needing too much preparation. Since most of the features in the dataset are categorical, this model fits well with the structure of the data. Lastly, I will use the Naive Bayes model. This is based on probability and works well with categorical data, especially when the dataset is not large. It also makes the assumption that the features are independent from each other. Although this might not be completely true, in my use case, I believe it will still work well. Naive Bayes is fast to train and useful as a benchmark to compare against the other two models. For tools, I will be working in Python using libraries like pandas and NumPy for data preparation, scikit learn for modeling, and matplotlib or seaborn for any visualizations. These tools were used in earlier assignments and are appropriate for my current working experience with Python. Nevertheless, these are powerful enough to complete this project within my scope. I am also using Jupyter Notebook as my main workspace because it allows me to test and document my steps as I go.

Methodology and Project Plan

This project will follow an approach that mirrors the process I have learned throughout the program. Like all previous project, it will start with data preparation. Since the dataset I choose to use is mostly made up of categorical features (I am using the non-numeric version), the first step will be encoding these values into numbers so the models can read them. I will also confirm that there are no duplicate rows and confirm that there are no missing values. While I am nearly certain there are none of these, I do believe it is good practice to build a habit of doing this, especially when in a real-world environment, where I would be working with client's financials. I would also look at the class balance between good and bad credit risk labels. This step is important because it sets up everything else that follows. After the data is prepared, I will move into exploratory analysis. This means validating how the features are distributed, spotting any odd or unexpected values, and looking at how different columns relate to the target variable. I already did a quick version of this earlier, but now I will go deeper. I will use things like value counts, group by comparisons, and maybe a few visualizations to better understand the data. This step will also help me decide which features to focus on when I start modeling.

The third phase is model training. Each of the three models, Logistic Regression, Decision Tree, and Naive Bayes, will be trained on the same training data. I will make sure to use the same split for each model so the results are comparable. After training, I will test each model using the test set and evaluate how well they perform using accuracy, recall, and a confusion matrix. As mentioned earlier, the dataset is imbalanced, and to remedy that, I will pay close attention to how well each model detects bad credit risk cases. A model that just predicts everything as good credit might have high accuracy but would be useless in practice. After evaluating the models, I will compare their results and look at which one performed best overall. I will also look at feature importance or model coefficients to see which variables had the most influence on the predictions. The final step will be summarizing everything I learned. I will look at how the models performed, what features mattered most, and whether the results give any insight into financial behavior that could apply to broader risk or compliance work. As I mentioned previously, I am fully aware that my abilities are not at the level to build a "perfect model". I do believe that after completing the previous courses and additional studying I have done over the last year and a half I have the capability to build a model that works well enough to give useful results and that I can explain clearly and competently. That is the goal for my capstone, and I believe this plan gives me a straight forward and doable path to get there.

Final Results

After I completed the modeling phase, I compared the results of all three algorithms: Logistic Regression, Decision Tree, and Naive Bayes. All the models were trained on the same version of the cleaned German Credit dataset. I used accuracy, recall, confusion matrix, and 5-fold cross-validation to evaluate performance. I choose these because they give a more complete picture of how well the models are actually performing, especially given that the dataset had some imbalance between good and bad credit risk classes.

Logistic Regression performed the best overall with it reaching an accuracy of 76 percent and had a recall score of 0.864, meaning it was able to correctly identify good credit risks most of the time. This matched what I found in the literature review, where Logistic Regression was frequently cited for its strong performance in financial related classification tasks. This model is also highly interpretable, which I believe would be useful in a compliance focused settings like risk management or AML work where model decisions need to be explained clearly.

Naive Bayes was the second-best model, with an accuracy of 76 percent and a recall of 0.807. This is also a strong result, especially considering that Naive Bayes is a very fast and simple model. Even though it makes assumptions about feature independence that might not always hold, the model still well even though it was much simpler. That gives it some value in cases where resources or time are limited and a quick decision is better than none at all.

The Decision Tree model had the lowest performance out of the three, with 71.5 percent accuracy and 0.764 recall. However, it still worked well enough to be considered a decent baseline model. In future iterations, it could potentially be improved using ensemble methods like Random Forest or boosting techniques, which were mentioned in a few of the academic papers I reviewed. For this phase, I focused only on simpler, more explainable models to stay aligned with the original project goals.

Across all three models, I used 5-fold cross-validation during training to make sure that the results were not just a fluke from one train-test split. I found the cross-validation scores to be consistent with the test set accuracy, which helped verify that the models were performing reliably. As part of the model evaluation, confusion matrices were included to better understand which types of errors each model was making. For example, Logistic Regression was the strongest at minimizing false negatives, meaning it was better at not misclassifying good borrowers as risky. That’s important in a financial setting where denying credit to reliable applicants could have major consequences.

In terms of acceptance criteria, I primarily looked at recall and cross-validation accuracy. Since the goal is to identify good credit risks, I prioritized recall over precision. The threshold for acceptance was roughly 75 percent recall or higher. Both Logistic Regression and Naive Bayes met this, but Logistic Regression stood out for its simplicity, speed, and reliability. There were some limitations to this phase of the project. I did choose to use Google Colab which had limited customization options compared to working in a full desktop environment. However, for the scope of this capstone, it worked well enough. In terms of techniques, I intentionally went into this project with the idea, that I wanted to use basic classification models to stay within what I believed my skill level to be. If I was to continue this project beyond it’s current scope, I would consider testing more complex methods like Random Forest, Support Vector Machines, or Neural Networks.

To conclude, this phase successfully proved that even basic machine learning models, when properly trained and validated, can deliver solid insights into credit risk. The results confirm what was expected from both the abstract and the literature review. The next steps would be to either fine-tune these models or try out more advanced techniques, depending on the resources available and the goals of the organization using the models.

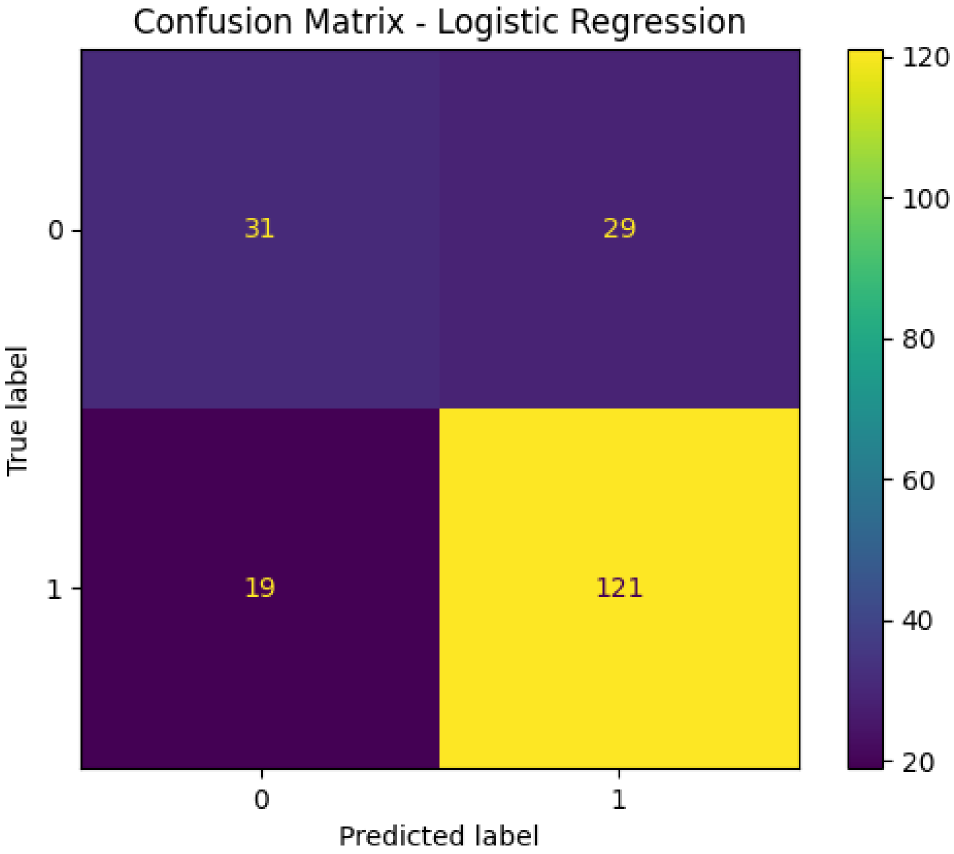

Figure 3: Confusion matrix for Logistic Regression, which achieved the best recall. “Worst” credit cases (label 0) were correctly identified, demonstrating the model's reliability in credit risk assessment.

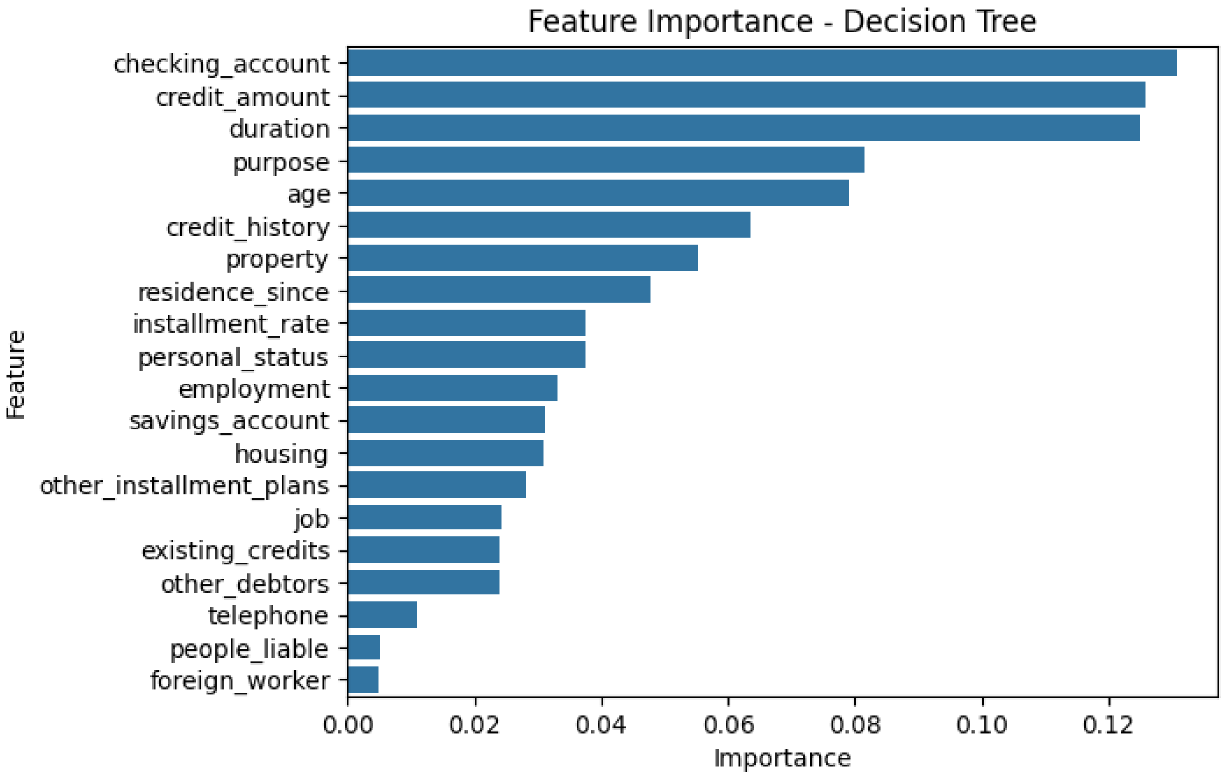

Figure 4: Feature importance as calculated by the Decision Tree model. Variables like checking account status and credit history contributed most to the classification decision.

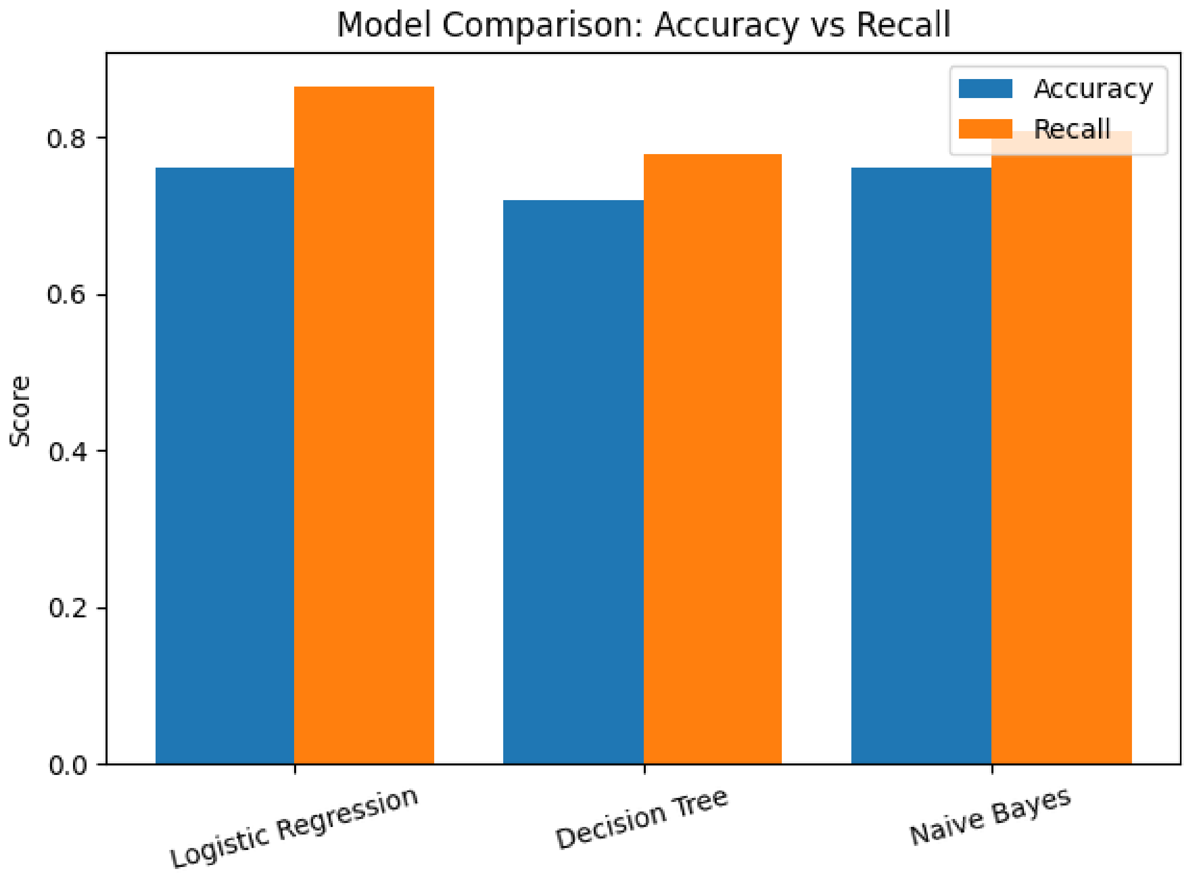

Figure 5: Comparative performance of three models across accuracy and recall. Logistic Regression had the highest recall, making it the most reliable option for identifying good credit risks.

Conclusion and Perspective

This project set out to explore how basic machine learning models could be used to classify individuals as good or bad credit risks using the German Credit dataset. After going through a complete data pipeline, I was able to test and compare three different classification models: Logistic Regression, Decision Tree, and Naive Bayes.

The results showed that Logistic Regression was the most effective model, providing a good balance of accuracy and recall while also being interpretable and fast. This was in line with what I expected based on my literature review, where most often the Logistic Regression was often highlighted for its success in credit and financial risk-related tasks. The model consistently performed well across both test sets and cross-validation, which confirmed its reliability. Naive Bayes also showed promising results. Despite being a very simple model with strong assumptions, it still managed to reach similar accuracy levels. This initially surprised me because it suggested that even lightweight models can work well in situations where computation or explainability is a concern. The Decision Tree model, while not as strong in terms of accuracy, still served as a solid benchmark and could easily be improved with ensemble methods in future work.

From a learning perspective, this project gave me hands-on experience with the complete machine learning workflow. I was able to practice converting categorical data into numerical form, applied scaling, and saw how these preprocessing steps directly impacted model performance. I also learned the importance of choosing evaluation metrics that reflect the project goals. In this case, using recall to focus on identifying reliable borrowers. One key insight was realizing that a model's performance is not just about accuracy, but also about how well it handles different types of prediction errors. There is certainly room for improvement in a few areas. The primary hinderance is that I only used three basic models. If this were a Masters/Production-level project, I would experiment with more advanced algorithms such as Random Forest or Gradient Boosting to try and boost performance.

References

Abdoli, M., Akbari, E., & Shahrabi, J. (2021). Bagging supervised autoencoder classifier for credit scoring. *Applied Soft Computing, 108*, 107442. https://doi.org/10.1016/j.asoc.2021.107442

Ala'raj, M., & Abbod, M. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications, 64*, 36–55. https://doi.org/10.1016/j.eswa.2016.07.005

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003a). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*(3), 312–329. https://doi.org/10.1287/mnsc.49.3.312.12740

Baesens, B., Van Gestel, T., Stepanova, M., Suykens, J., & Vanthienen, J. (2003b). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Hofmann, H. (1994). *Statlog (German Credit Data)* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77

Warse, K. (2020). Machine learning in consumer credit risk analysis: A review. *International Journal of Engineering Research and Technology, 9*(5), 1203–1209. https://www.ijert.org/machine-learning-in-consumer-credit-risk-analysis-a-review

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020