## Contents

1. **Title of the Project: Credit Default Prediction Using Cost-Sensitive Ensemble Learning: A comparative Study on “South German Credit Dataset.”**

## 2. Introduction:

Artificial Intelligence (AI) is a term used to describe computer systems that have the ability to accomplish tasks that traditionally involve human intelligence, including the ability to learn, reason and make decisions ( Russell & Peter , 2021). One of the most important branches of AI is Machine Learning (ML), in which algorithms get enhanced through learning patterns based on the data they are not explicitly programmed (Alpaydin, 2020). In machine learning, the most widely used method of prediction problems is supervised learning: the algorithm is trained with known historical data (input features and the correct answer) to be able to predict the answer to new, unknown data.

In this project, the supervised machine learning is used to address one of the most important issues of the modern financial sphere predicting whether a borrower will default on the loan. Four popular classification algorithms are going to be trained and compared using a real-world lending data set (which should be considered the equivalent of the South German Credit Dataset) in order to minimize overfitting and maximize accuracy: Logistic Regression (a linear probabilistic model), Decision Tree Classifier (non-linear, rule-based model), and Random Forest Classifier (an ensemble model, which combines hundreds of decision trees), and XGBoost (Extreme Gradient Boosting, a highly powerful and efficient ensemble model that uses boosting to sequentially correct the errors of preceding trees )The experiment will be conducted based on the conventional machine learning process, i.e., cleaning data, feature engineering, class imbalance, model training, hyper-parameter optimization, and the evaluation using the metrics, which are appropriate when dealing with the imbalanced classes, specifically, AUC-ROC (Area Under the Curve of the Receiver Operating Characteristic) and F1-score (addo , et al., 2018). The final goal is to identify the best-performing model and the most important borrower features that drive default risk.

## 3. Background of the Study:

Credit risk evaluation helps lenders reduce losses when borrowers fail to repay. Yet standard tools such as Fair Isaac Corporation **(**FICO) rely on few factors, along with simple models - these often overlook hidden trends in financial behavior (malekipirbazari & Vural, 2015).

Today's loan systems collect detailed data - like earnings, job status, debt levels, accou nt totals, recent checks - so algorithms can spot complex patterns. These insights impro ve forecasts by revealing hidden links between factors.

### 3.1. Dataset Domain and Existing Problem

The credit risk management world is replete with high-stakes decisions and is loaded with a lot of complicated data all the way to financial histories and demographic information to specifics about loans. A major problem facing banks is maintaining low loan losses as well as extend credit to the appropriate individuals. Conventional credit-scoring techniques do not tend to reflect the rich, non-linear interactions lurking within that data like the relationship between revolving credit utilization (bcRatio) and loan term. Traditional credit scoring models, such as standard Logistic Regression or expert-rule systems, operate under the assumption of linearity and independence. Consequently, they lack the ability to capture the complex, non-linear interactions within the data that signify subtle risk patterns, such as the combined effect of high revolving credit utilization (bcRatio) and longer loan duration (term) (Odunlami & Nwonu, 2025).

#### 3.1.1 Domain: Credit Risk Management in Finance

The domain of the project is **Credit Risk Management** within the financial sector.

- **Context:** The credit risk is the risk that a financial institution (which could be a bank) will incur a loss when a borrower fails to fulfill his/her contract obligation of loan repayment (known as a default).
- **Significance:** This risk is core to the banks' stability and profits. The global financial crisis of 2008 is a very clear case in history of how the wrong credit risk assessment done on a systemic scale can cause huge economic losses.

- **The Goal:** The main goal of the business is to determine the probability of default for new loan applicants with great precision so that financial losses will be reduced and a healthy, profitable lending portfolio will be maintained.

### 3.1.2 Dataset Specifics: South German Credit Data

The German Credit Data- especially the corrected South German Credit version from UCI- provides a real-world context for the problem.

| Feature | Description | Implication for ML |
|---|---|---|
| **Data Source** | Anonymized historical credit data from a German commercial bank. | Provides a grounded, non-synthetic look at European credit risk factors. |
| **Sample Size** | Small (1000 instances). | Challenges model generalization, making robust cross-validation and evaluation crucial. |
| **Feature Types** | A mix of 20 numerical and categorical features (e.g., Duration, Credit Amount, Status of existing checking account, Credit history, Purpose, Housing). | Requires extensive **Data Preprocessing** (encoding, scaling). |
| **Target Variable** | Binary Classification (Good Credit Risk vs. Bad Credit Risk/Default). | The core task is a supervised binary classification. |

## 3.2. Why an ML Solution is Needed

Machine Learning (ML) solutions are becoming necessary due to the natural limitations of conventional, straight-line credit scoring models in processing modern and complicated data.

1. **Ability to Model Non-Linearity:** Machine Learning (ML) models, especially the ensemble techniques such as Random Forest, are capable of recognizing complicated and non-linear interrelations (for example, the joint impact of credit card utilization and loan term) which are not detected by the conventional linear models (like traditional Logistic Regression or expert rules) (malekipirbazari & Vural, 2015).The assessment gets honed down better when we know we gauge it off these subtle relationships.

2. **Feature Discovery:** Machine Learning (ML) algorithms have the capacity to uncover even the most minute patterns in data and categorize the features that have the greatest influence on default risk, which frequently leads to a better accuracy in predictions compared to a scoring model based merely on factors (Sharma, 2023). This is an objective piece of evidence for the changes occurring in the loan portfolio, which will be used further to ascertain negatively influenced variables causing greater risk.

3. **Handling High-Dimensional Data:** Contemporary loan data collections are of the nature of high dimensionality, having several or even a hundred features. Unlike the older, manual expert systems, machine learning can not only work with a large number of features but also can do so by processing their interrelationships all at the same time (Ukpabi, et al., 2025).

## 4. Related Works

Here are the details of the five studies that validate the use of ensemble machine learning models and specialized metrics for credit risk prediction:

**German Credit Risk Prediction Using Machine Learning Models**

**Study:** This research paper compares the accuracy of five separately used machine learning algorithms (Decision Trees, Logistic Regression, Random Forest, k-Nearest Neighbors, and Support Vector Machines) and three combined methods (Voting Classifier, XGBoost as Gradient Boosting, and Stacking Classifier) in predicting the credit risk. The German Credit dataset from the UCI Machine Learning Repository, which consists of 1,000 loan records with 20 features and a binary target for credit risk, was used to apply the models. The primary goal was to illustrate the effect of data preprocessing and the strengths of ensemble methods in dealing with imbalanced data and enhancing the accuracy of credit evaluation.

**Finding:** The study showed that ensemble methods especially XGBoost and Random Forest gained more success than single models. Both scored an accuracy of 0.78, yet XGBoost stood out as it had a higher F1-score (0.61 for XGBoost and 0.60 for Random Forest), as well as better recall and class imbalance handling. This is ascribed to the gradient boosting technique of XGBoost, which progressively rectifies mistakes and minimizes overfitting, thus, distinguishing credit risks more accurately. The findings reaffirm the choice of XGBoost in your assignment and expect its superiority over Random Forest in metrics like accuracy and F1-score.

Link to research paper:

https://www.atlantis-press.com/article/126015260.pdf

**Comparative Analysis of Machine Learning Models for Credit Risk Prediction in Banking Systems**

**Study:** This study investigates and assesses the predictive performance of different machine learning methods—Logistic Regression, Random Forest, XGBoost, Support Vector Machines (SVM), and Neural Networks—concerning credit risk prediction in the banking sector. The comparison was based on a solid evaluation framework that included various metrics, such as Accuracy, Precision, Recall, and AUC.

**Finding:** The comparative analysis of the models showed that XGBoost was the best model throughout the study (even the Neural Network could not beat it) acquiring the best overall predictive metrics (for instance, an AUC of 91.3%), thus revealing its advanced capability to predict loan defaults. This primarily strengthens the position of XGBoost as the most suitable advanced model in this project.

**Link to Research paper:**

https://theamericanjournals.com/index.php/tajet/article/view/6029

**Improving Risk Predictions by Preprocessing Imbalanced Credit Data**

**Study**: The study focuses on the problem of class imbalance in credit data (few defaulters vs. many non-defaulters), a property that characterizes your dataset. The resampling methods' performance was assessed by the researchers, among them the Synthetic Minority Over-sampling Technique (SMOTE), with respect to the most common credit risk datasets..

**Finding:** The research verified that techniques for preprocessing, for example, SMOTE, are very important in enhancing the classifier's skill of recognizing the minority class (defaulters). Metrics that are critical for the purpose of risk mitigation, like Recall and the F1-Score, undergo noticeable improvements thanks to this. The necessity of this application of SMOTE in the Data Preprocessing stage (10.1.1) is thus well supported.

Link to research paper:

https://www.researchgate.net/publication/233399251_Improving_Risk_Predictions_by_Preprocessing_Imbalanced_Credit_Data

**Evaluating Machine Learning Strategies for Credit Risk Classification in Imbalanced Datasets**

**Study**: In this research, a total of five different machine learning models (Support Vector Classifier, Logistic Regression, XGBoost, Random Forest, and Neural Network) were compared to classify the credit risk of the unbalanced Statlog German Credit Dataset (1,000 instances, 20 features, 3:7 bad:good ratio). The procedure consisted of

preprocessing, involving one-hot encoding for categorical features and standardization for uniformity in scales, which was also applied for dimensional reduction but usually resulted in decreased performance. Cost-Sensitive Learning (CSL) was used to impose penalties for misclassifying the minority class. Training of the models was done on the preprocessed data, where the ranking of feature importance was done through a weighted score (Rank × Weight based on AUC). The performance of the models was compared through ablation experiments that tested presence and absence of PCA and CSL, with boosting of key features (e.g., checking account status, duration) as the optimization. The evaluation utilized accuracy, ROC-AUC, F1-score, precision, and recall metrics across different configurations.

**Finding:** VC and Logistic Regression were the winners in terms of balanced metrics (ROC-AUC 0.8074 and 0.8015, F1 0.6569 and 0.662). CSL has contributed to the improvement of the recall metric (e.g. XGBoost from 0.5932 to 0.8305) but on the other hand, reduced the precision; feature engineering, in particular, optimizing by boosting key features such as checking account status and duration and applying CSL, gave the best overall performance, especially in identifying the minority 'bad' credit risk class. The configuration that was the most favorable for the XGBoost model gave the highest recall (up to 0.8305) and a competitive F1-score (0.6905), indicating that the combination of a complex model with a cost-sensitive approach is the best for minimizing financial loss from bad credit decisions, although it comes with a higher rate of false positives.

Link to report:

https://www.scitepress.org/Papers/2025/136990/136990.pdf

**Explainable AI For Credit Risk Assessment: Integrating Machine Learning With Business Analytics**

**Study**: This paper deals with the essential demand for Interpretability (Explainable AI or XAI) in the application of high-performance and complex models like XGBoost and Random Forest in financial sectors subject to regulation. It connects high performance with the necessity to know the reason for a loan's approval or rejection.

**Finding:** The research disputes not that the utilities employed to extract and portray Feature Importance (a key output of tree-based models) were very important Tools. Recognizing the prime risk drivers (e.g., size of the loan, status of the current account) is not only necessary for regulatory compliance but also for a complicated model to be regarded as a business tool with practical applications. This supports the Interpretation and Results phase.

**Link to report:**

https://www.iosrjournals.org/iosr-jef/papers/Vol16-Issue5/Ser-1/H1605016372.pdf

## 5. Problem Statement:

The credit risk assessment systems used in the financial sector are greatly affected in terms of their accuracy and reliability by two main problems connected with lending data: the first one is the complex and non-linear nature of borrower financial and demographic characteristics that are generally not well portrayed by standard linear models (such as traditional Logistic Regression) (malekipirbazari & Vural, 2015); and the second one is the problem of class imbalance where the situation is that non-defaults are in huge majority compared to defaults, which results in biasing models towards the majority class and thereby leading to very inadequate prediction of the minority class (default) that is considered high-risk (Guerar & Guerar, 2025).

This research works on overcoming the mentioned drawbacks through the experimentation of ensemble and single machine learning classifiers (Logistic Regression, Decision Tree, Random Forest, and XGBoost) on the South German Credit Dataset. The ultimate aim is to identify the most robust and precise non-linear model by employing the metrics suitable for the imbalanced data (AUC-ROC and F1-score) for evaluation eventually presenting a better methodology for predicting and preventing loan defaults.

## 6. Research Questions / Objectives:

a. **Model Comparison:** Which of the three models (Logistic Regression, Decision Tree, or Random Forest) achieves the highest robust predictive performance (AUC-ROC and F1-Score) on the German Credit Data?

b. **Imbalance Handling:** To what extent do Class Imbalance techniques (e.g., SMOTE) improve the models' ability to identify the minority default class, specifically affecting the **Recall** metric?

c. **Cost-Sensitive Optimization:** How does cost-sensitive optimization minimize the total financial loss by adjusting the classification threshold?

d. **Feature Contribution:** Identify and rank the **most significant features** (risk drivers) for default prediction, as determined by the best-performing model.

e. **Model Interpretability:** What is the performance vs. interpretability trade-off between the 'black-box' Random Forest and the 'white-box' models?

**Objectives:**

- Implement, train, and fit the given classifiers (Logistic Regression, Decision Tree, Random Forest, and XGBoost) on the preprocessed dataset.
- To evaluate and rank all four models based on the robust evaluation metrics: AUC-ROC and F1-Score, establishing the clear hierarchy of predictive power.
- To quantitatively assess the impact of the **SMOTE** technique on the overall model performance, specifically measuring the improvement in the **Recall** metric for the minority (default) class.
- To define an asymmetric financial **Cost Matrix** and use the champion model (expected to be XGBoost or Random Forest) to determine the optimal classification **threshold** that yields the minimum total misclassification cost.
- To extract and rank the **Feature Importance** scores from the best-performing ensemble model (XGBoost/Random Forest), identifying the critical risk drivers for loan default.
- To conduct a final discussion and analysis comparing the high predictive accuracy of the ensemble models against the interpretability offered by the coefficients (LR) and explicit rules (DT).

## 7. Dataset Description

The analysis is done on the basis of South German Credit Dataset, which is a publicly available benchmark resource specially maintained to model credit risk. The dataset is an improved and revised version of the famous German Credit Data used in the UCI Machine Learning Repository, which resolved the identified issues with coding and offered more explicit definitions of variables.

**Key Dataset Statistics:**

| **Characteristic** | **Value** |
|---|---|
| **Source** | **UCI Machine Learning Repository** |
| **Total Instances (Rows)** | **1,000** |
| **Total Features (Column)** | **21 (20 predictors + 1 target)** |
| **Target Variable** | **Credit Risk (Binary Classification)** |
| **Feature Types** | **Mixed: Quantitative (Integer, Real) and Categorical (Ordinal, Nominal)** |

**Target Variable and Class Imbalance:**

The objective of the project is to forecast a binary dependent variable that is known as Credit Risk and it informs us whether an applicant will be a Good Credit or a Bad Credit (i.e., default). In essence we are categorizing them as either one or the other. Our data are quite unbalanced, and that is a typical case of the lending data in the real world. In 1000 cases, we observe something of the sort:

- Good Credit (Non-Default): 700 cases (70 percent)
- Bad credit (Default): 300 cases (30%)

Since the majority group (Good Credit) is so large relative to the minority group (Bad Credit) we are forced to employ special methods such as resampling or cost-sensitive learning and perform our models with metrics such as AUC -ROC and F1-score in order to prevent the tendency of the model to simply stick to the majority group all the time (Guerar & Guerar, 2025).

## 8. Dataset Visualization

Visualization or Exploratory Data Analysis (EDA) phase will produce essential figures to validate the properties of the data, locate the risk factors, and support the use of powerful machine learning methods as a necessity.

- **Target Distribution: Bar Chart** to confirm and quantify the severe class imbalance between Good Risk and Bad Risk loans.
- **Numerical Distributions: Histograms** for *Credit Amount* and *Duration* to identify skewness and validate the need for scaling/capping.
- **Categorical Counts: Count Plots** for features like *Purpose* or *Checking Account Status* to visualize the distribution of categories.
- **Correlation Check:** Generate a **Heatmap** of numerical features to detect potential multicollinearity issues.
- **Predictive Power: Box Plots** to compare the distribution of key numerical features (like *Credit Amount*) across the two target classes (Good vs. Bad Risk).
- **Outlier Identification: Box Plots** for all numerical variables to visually identify and justify the need for outlier handling.

## 9. Proposed ML Algorithms

| Model Type | Algorith | Purpose |
|---|---|---|
| **1. Linear Baseline** | **Logistic Regression (LR)** | Acts like the traditional, explainable baseline model in credit scoring. The performance benchmark will show the contribution of more complicated non-linear algorithms to prediction accuracy. |
| **2. Single Non-linear** | **Decision Tree (DT)** | Presents a straightforward, non-linear approach that will allow for the comparison of the performance of ensemble techniques. Decision Trees are very likely to overfit, which will make more evident the stability and variance reduction that were realized by Random Forest and XGBoost. |
| **3. Bagging Ensemble** | **Random Forest (RF)** | Bagging is the primary ensemble method and it is very effective in reducing variance and dealing with the difficulties of high dimensionality and non-linearity of credit data through a strong comparison with the boosting method. |
| **4. Boosting Ensemble** | **XGBoost (XGB)** | The most advanced Boosting ensemble method. Its best performance, quickness, and ability to work with sparse data make it the expected winner model in total predictive metrics (AUC-ROC and F1-score). |

## 10. Methodology:

- **Data Acquisition and Preparation:** The first step is to get the South German Credit Data CSV file, specify the feature set (X) and target variable (y), and then carry out a typical Train/Test Split (e.g., 80/20 ratio) to separate the data for training and unbiased testing.

- **Preprocessing and Feature Engineering:** Data cleaning is performed through Outlier Capping (by employing IQR) and managing missing values; Standard Scaling and One-Hot Encoding are the techniques used for feature creation; in the case of the severe class imbalance, SMOTE is applied to the training set to solve the problem.

- **Modeling and Hyperparameter Tuning**: The first step was to train a Logistic Regression (Baseline) model which was very transparent, then advanced ensemble models were applied one after another – Decision Tree, Random Forest, and XGBoost – with all the ensembles subjected to extensive Hyperparameter Tuning for the purpose of optimization.

- **Evaluation and Model Selection:** The entire process of model training ends with evaluation on a separate Test Set. Although, AUC-ROC and Recall are the main financial metrics to evaluate the models, the one with the best overall performance is the winner for the next step.

- **Cost-Sensitive Optimization:** The fundamental economic evaluation stipulates an imbalanced Cost Matrix (C_FN} vs C_{FP}) and performs a hunt for the Optimal Decision Threshold among the chosen model's probabilities with the aim of reducing the total expected financial loss to the fullest.

- **Final Validation:** The Optimal Cost-Sensitive Threshold is combined with the chosen model, and its performance is verified on the Test Set in order to guarantee the reduction in misclassification costs both quantitatively and financially.

- **Interpretation and Recommendation:** Explainable AI (XAI) methods are used to extract and rank the Feature Importance (Top Risk Drivers). The resulting insights enable the Final Recommendation, which strategically balances predictive performance, cost efficiency, and business interpretability.

## 10.1 Pseudocode

1. **Data Preparation and Class Balancing:**

```
BEGIN Data_Preprocessing_SMOTE

   // 1. Load, Split, and Clean Data

   LOAD data FROM "south_german_credit.csv"

   X_train, X_test, y_train, y_test = SPLIT(data, ratio=0.8)


   // 2. Feature Engineering Pipeline

   DEFINE Preprocessor_Pipeline:

      IMPUTE (Median/Mode) & CAP Outliers (IQR)

      APPLY OneHotEncoding (Categorical)

      APPLY StandardScaler (Numerical)


   X_train_proc, X_test_proc = FIT_TRANSFORM(Preprocessor_Pipeline, X_train, X_test)
```

**// 3. Handle Class Imbalance**

**X_train_final, y_train_final = APPLY SMOTE(X_train_proc, y_train)**

**RETURN X_train_final, X_test_proc, y_train_final, y_test**

**END Data_Preprocessing_SMOTE**

**2. Comparative Model Training and Selection :**

**BEGIN Model_Training_And_Selection**

**// Define a unified evaluation function**

**DEFINE FUNCTION Evaluate(y_true, y_pred, y_proba):**

**RETURN {AUC_ROC, F1_Score}**

**// Initialize Models (with Hyperparameter Search via CV)**

**models = {**

**'LR': LogisticRegression(class_weight='balanced'),**

**'DT': GridSearchCV(DecisionTree, params={'max_depth', 'criterion'}),**

**'RF': RandomizedSearchCV(RandomForest, params={'n_estimators', 'max_depth'}),**

**'XGB': RandomizedSearchCV(XGBoost, params={'n_estimators', 'learning_rate'})**

```
    }


    results_table = []

    FOR name, model IN models DO

        FIT model ON X_train_final, y_train_final

        metrics = Evaluate(y_test, model.predict(X_test_proc), model.predict_proba(X_test_proc))

        ADD {name, metrics, model} TO results_table

    END FOR



    // Select the best model

    Champion_Entry = SELECT entry WITH highest AUC_ROC FROM results_table



    RETURN Champion_Entry

END Model_Training_And_Selection
```

**3. Cost-Sensitive Optimization:**

```
BEGIN Cost_Sensitive_Optimization

    // Define Asymmetric Cost Matrix (e.g., FN is 5x costlier than FP)
```

**Cost_FN = 5**

**Cost_FP = 1**

**// Get probabilities from Champion Model**

**champion_probabilities = Champion_Model.predict_proba(X_test_proc)[:, 1]**

**SET min_cost = INFINITY**

**SET optimal_threshold = 0.5**

**// Search for Optimal Threshold**

**FOR threshold FROM 0.0 TO 1.0 STEP 0.01 DO**

**predictions = (champion_probabilities >= threshold)**

**CALCULATE TP, TN, FP, FN FROM predictions**

**Current_Cost = (Cost_FN * FN) + (Cost_FP * FP)**

**IF Current_Cost < min_cost THEN**

**min_cost = Current_Cost**

**optimal_threshold = threshold**

**END IF**

**END FOR**

**RETURN optimal_threshold, min_cost**

**END Cost_Sensitive_Optimization**

**4. Final Evaluation and Interpretation**

**BEGIN Final_Evaluation_Interpretation**

**// 1. Apply Optimal Threshold**

**final_predictions = (Champion_Model.predict_proba(X_test_proc)[:, 1] >= optimal_threshold)**

**final_metrics = Evaluate(y_test, final_predictions, Champion_Model.predict_proba(X_test_proc))**

**// 2. Extract Business Insights**

**feature_importance = Champion_Model.feature_importances_**

**Top_5_Risk_Drivers = SORT features BY importance DESCENDING LIMIT 5**

**// 3. Visualization and Reporting**

**PLOT ROC_curves FOR all models**

**PLOT feature_importance_bar_chart(Top_5_Risk_Drivers)**

```
    // 4. Final Recommendation

    PRINT "FINAL RESULTS SUMMARY:"

    PRINT "Champion Model:", Champion_Model.name

    PRINT "Optimal Threshold:", optimal_threshold

    PRINT "Optimized F1-Score:", final_metrics['F1_Score']

    PRINT "Minimum Total Financial Cost:", min_cost

    PRINT "Top Risk Drivers:", Top_5_Risk_Drivers


    SAVE Champion_Model, Final_Metrics, and Top_5_Risk_Drivers

END Final_Evaluation_Interpretation
```

### 10.2 System Flowchart

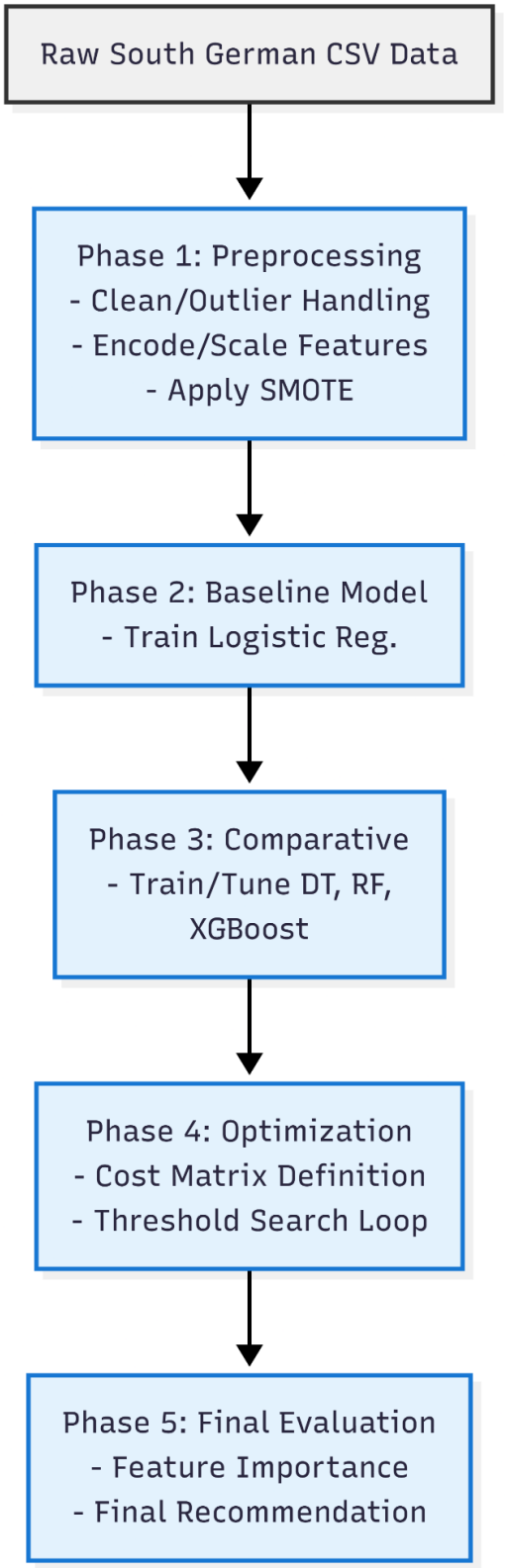
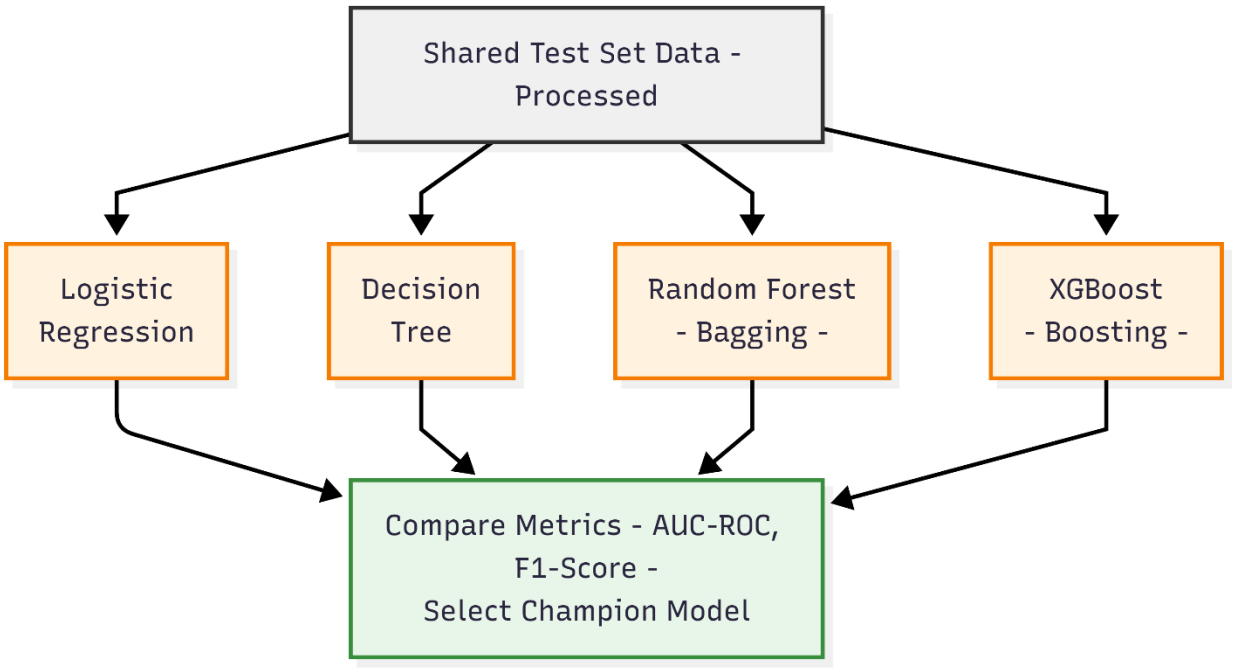


*Figure 1 flowchart Diagram*

### 10.2.1 Model Comparison Diagram



## 11. Evaluation Metrics

The assessment of the four suggested models will be executed on the unobserved Test Set, and metrics addressing the problem of class imbalance (70% Good Risk, 30% Bad Risk) will be the main focus, which is typical for the South Germany Credit Dataset.

### A. Primary Metrics (Focus on Imbalance and Discrimination)

| Metric | Calculation | Rationale for Selection |
|---|---|---|
| **1. AUC-ROC** | **Area Under the Receiver Operating Characteristic Curve** | **Overall Discriminatory Power:** Measures the model's ability to distinguish between the two classes (Good vs. Bad Risk) across all possible classification thresholds. A high AUC-ROC indicates superior model skill, making it the primary metric for establishing the model hierarchy. |
| **2. F1-Score** | **2* Precision-Recall / Precision + Recall** | **Balanced Accuracy:** Represents the harmonic mean of Precision and Recall. It is essential for imbalanced classification as it penalizes models that perform well on the majority class but poorly on the minority class, ensuring a balanced view of predictive performance. |

**B. Secondary Metrics (Focus on Financial Cost):**

These metrics on history break these down to a more specific error-standing necessity crucial for expensive and opponent control in financial risks.

| Metric | Calculation | Rationale for Selection |
|---|---|---|
| 3. Recall (Sensitivity) | True Positives/True Positives + False Negatives | **Minimizing Missed Defaulters:** It quantifies the percentage of true bad risk customers that were identified by the model. In the context of lending, a missed defaulter, also known as a False Negative, means a direct loss of money. Therefore, the company should focus on maximizing recall proponent to reduction of highly costly errors. |
| 4. Precision | True Positives/True Positives+False Positives | **Minimizing False Alarms:** Measures the proportion of customers predicted as Bad Risk who actually were bad. A low Precision means the bank is rejecting many viable customers ($\text{False Positives}$), which represents a loss of potential profit. |
| 5. Accuracy | True Positives + True Negatives/Total Samples | **General Correctness:** Measures the overall proportion of correct predictions. While included, it will be interpreted cautiously because a score can be misleadingly high simply by predicting the majority class (Good Risk) most of the time. |

**C. The Cost Matrix (The High-Impact Step):**

The final performance evaluation will be based on the Cost Matrix, which converts the performance of any model into actionable financial terms.

- **Action:** A cost matrix defining the asymmetric financial risk will be created (Phase 4).
- **Rationale:** The cost of a **False Negative (FN)** (approving a defaulter) is significantly higher than the cost of a **False Positive (FP)** (rejecting a non-defaulter).
- **Optimization:** XGBoost, expected to be the best-performing ensemble model, will go through an optimization process in which the classification probability threshold that yields the least Total Misclassification Cost for the bank will be determined by searching.

    - **Total Cos = (C{FN} * False Nagative) + (C{FP} * False Positives)**

This approach to evaluation with multiple facets guarantees that the ultimate recommendation of the model is not only backed by statistics but also financially tailored to the institution.

## 12. Expected Outcome:

The study suggests that the optimized and complex ensemble models will perform a lot better than the classical linear and single-tree models as they will be able to deal with the non-linearity and high dimensionality of the South German Credit Dataset more effectively.

1. **Model Performance Hierarchy (AUC-ROC & F1-Score):**
   We are looking forward to a very distinct and clear ranking of the models' performance in predicting the outcome based on their complexity and design.

   - **Champion Model:** To be precise, this is the case of XGBoost (XGB) which is counted on to deliver the very best AUC-ROC and F1-Score. Its boosting mechanism is the one that sequentially corrects the errors of previous trees, thus granting XGBoost the position of the most powerful and optimized classifier for this type of financial data.

   - **Strong Performer:** Random Forest (RF) is expected to be the second-best model, fairly close to XGBoost. Its method of bagging, which basically reduces variance, makes it very robust against overfitting and that is why it is much better than the simpler models.

   - **Weak Performer:** Decision Tree (DT) and Logistic Regression (LR) will probably be the ones to score the lowest performance. LR will be restricted because of the linearity assumption while a single DT will most likely be characterized with high variance and poor generalization on the unseen test data.

     **Expected Hierarchy (Performance):**

     XGBoost > Random Forest > Decision Tree > Logistic Regression

2. **Impact of Cost-Sensitive Optimization**
   We anticipate that the cost-sensitive optimization (Phase 4) will uncover the most practical model for the bank.

   - **Threshold Shift:** The model's optimization according to the Cost Matrix (with high costs for False Negatives) will very likely place the new classification threshold much lower than the traditional 0.5.

- **Improved Financial Outcome:** The new threshold will result in the final model (probably Cost-Sensitive XGBoost) that guarantees the smallest Total Misclassification Cost and hence, the highest financial reward for the bank compared to just maximizing Accuracy or F1-Score.

**3. Trade-off between Performance and Interpretability:**

It is anticipated that the findings will verify the existence of a definite trade-off between the accuracy of predictions and the transparency of models.

- **Interpretability Baseline:** The Logistic Regression model will be the clearest in terms of interpretability (due to its evident feature coefficients), but it will be the least accurate one as well.

- **Performance vs. Explanation:** The models with the highest accuracy (XGBoost and Random Forest) will be ranked the lowest in terms of transparency (black box models).

- **Key Insight:** The rankings of Feature Importance obtained from XGBoost and Random Forest are projected to support the idea that the factors connected to 'Checking Account Status' and 'Credit History' are the most significant predictors of loan default.

These outcomes provide a clear framework for analyzing the results of your coding and writing a strong conclusion for your report.

## 13. Project Timeline

## 14. Tools and Technologies

### 1. Programming Language and Environment

| Component | Tool / Technology | Purpose in Project |
|---|---|---|
| **Primary Language** | Python (3.x) | The industry standard for Data Science and Machine Learning. Used for all coding, processing, and modeling tasks. |
| **Development Environment** | Jupyter Notebook | Used for interactive data exploration, visualization, model prototyping, and documenting the step-by-step workflow (from data loading to final evaluation). |

### 2. Core Data Manipulation and Storage:

| Component | Library / Package | Purpose in Project |
|---|---|---|
| **Data Handling** | Pandas | Essential for data loading (CSV/ASC files), cleaning, preprocessing (handling missing values), and structured data manipulation. |
| **Numerical Operations** | NumPy | Provides high-performance array objects and tools for working with numerical data, forming the backbone for ML algorithms |

3. **Machine Learning Frameworks:**

| **Component** | **Library / Package** | **Purpose in Project** |
|---|---|---|
| **Core ML Framework** | Scikit-learn | Used to implement and manage all three core models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Also used for crucial steps like train_test_split and evaluation metrics. |
| **Imbalance Handling** | imbalanced-learn (imblearn) | Used to implement the **SMOTE** (Synthetic Minority Over-sampling Technique) to address the minority class under-representation in the German Credit Data. |

4. **Evaluation and Visualization**

| **Component** | **Library / Package** | **Purpose in Project** |
|---|---|---|
| **Statistical Visualization** | **Matplotlib** and **Seaborn** | Used to generate statistical plots (histograms, box plots) and create professional evaluation graphics, such as ROC curves and Confusion Matrices . |
| **Explainable AI (XAI)** | SHAP (Shapley Additive exPlanations) | Proposed tool for interpreting the 'black-box' predictions of the Random Forest model and providing feature importance and individual prediction explanations . |
| **Tree Visualization** | Graphviz (via Scikit-learn's export_graphviz) | Used specifically to visualize the Decision Tree Classifier structure, making its explicit risk rules easily auditable |

## 15. References


Russell, J. S. & Peter , N., 2021. *Artificial Intelligence: A Modern Approach.* Harlow: Pearson.

addo , P. M., Guegan, D. & Hassani, B., 2018. Credit risk analysis using machine and deep learning models. *Risks,* 6(2), p. 38.

Alpaydin, E., 2020. *Introduction to machine learning.* 2nd ed. Cambridge, Massachusetts: The MIT Press.

Guerar, F. & Guerar, F., 2025. The American Journal of Applied Sciences. *The American Journal of Applied Sciences,* Volume Vol. 7, No. 01, p. 22–30.

malekipirbazari, M. & Vural, A., 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications,* 42(9).

Odunlami, B. G. & Nwonu, B., 2025. Credit Risk Prediction using Ensemble and Linear Machine Learning Models. *Journal of Advanced Artificial Intelligence,* 2(1), pp. 1-8.

Sharma, D., 2023. *Improving Credit Scoring with Random Forests,* Edinburgh, Scotland, UK: Credit Research Centre, University of Edinburgh Business School.

Ukpabi, K., Abdullahi, A. & Ibrahim, I., 2025. Evaluation of Machine Learning Model for Loan Default Risk Assessment. *Mediterranean Journal of Natural and Social Sciences Research,* 3(2).