Credit Risk Prediction Using Machine Learning – A Comprehensive Report

1. Project Overview

This project focuses on predicting credit risk using machine learning algorithms, with the objective of classifying loan applicants as either **low-risk (good)** or **high-risk (bad)**. The analysis uses the **German Credit Dataset**, a widely recognized benchmark in financial risk modeling.

2. Problem Statement

Financial institutions need to make informed decisions when assessing the likelihood that a borrower will default on a loan. Traditional methods often rely on rigid credit scoring systems that fail to capture non-linear patterns or adapt to changing trends. This project explores the use of supervised machine learning models to:

- Automate the risk classification process.
- Improve the accuracy of credit risk predictions.
- Provide interpretability for stakeholders through model explainability tools.

3. Dataset Summary

- **Dataset**: UCI German Credit Data
- **Total Samples**: 1,000
- **Features**: 20 input features including both categorical and numerical types (e.g., credit history, loan purpose, employment status, savings account).
- **Target Variable**: `Credit Risk` (Good = 1, Bad = 0)

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Credit_per_Duration	Risk
0	67	male	2	own	unknown	little	1169	6	radio/TV	194.833333	0
1	22	female	2	own	little	moderate	5951	48	radio/TV	123.979167	1
2	49	male	1	own	little	unknown	2096	12	education	174.666667	0
3	45	male	2	free	little	little	7882	42	furniture/equipment	187.666667	1
4	53	male	2	free	little	little	4870	24	car	202.916667	1

4. Data Preprocessing

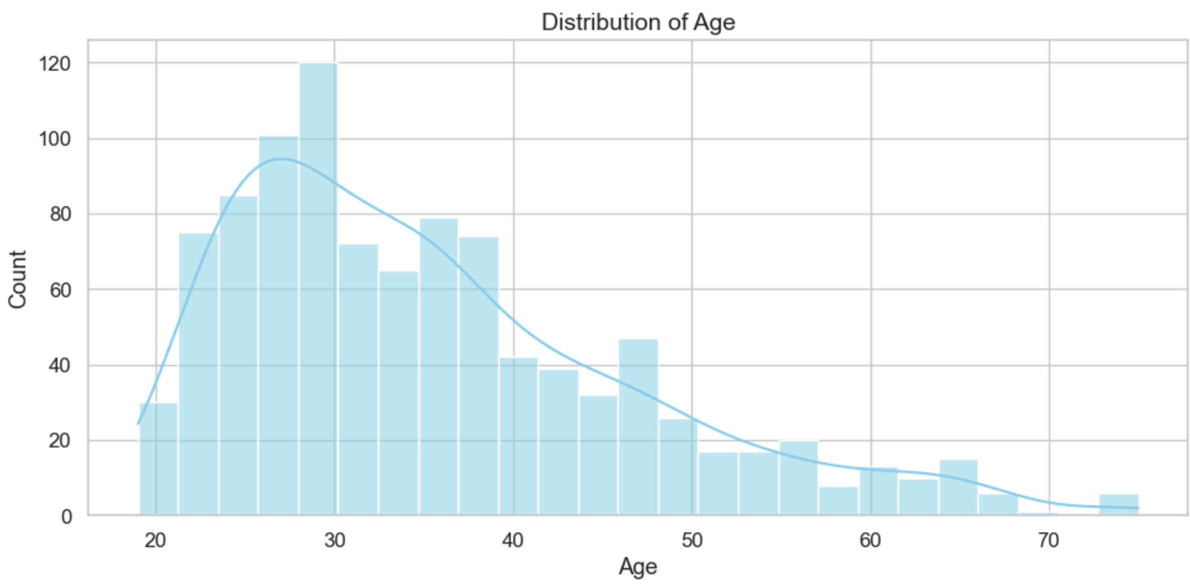
4.1 Data Cleaning & Transformation

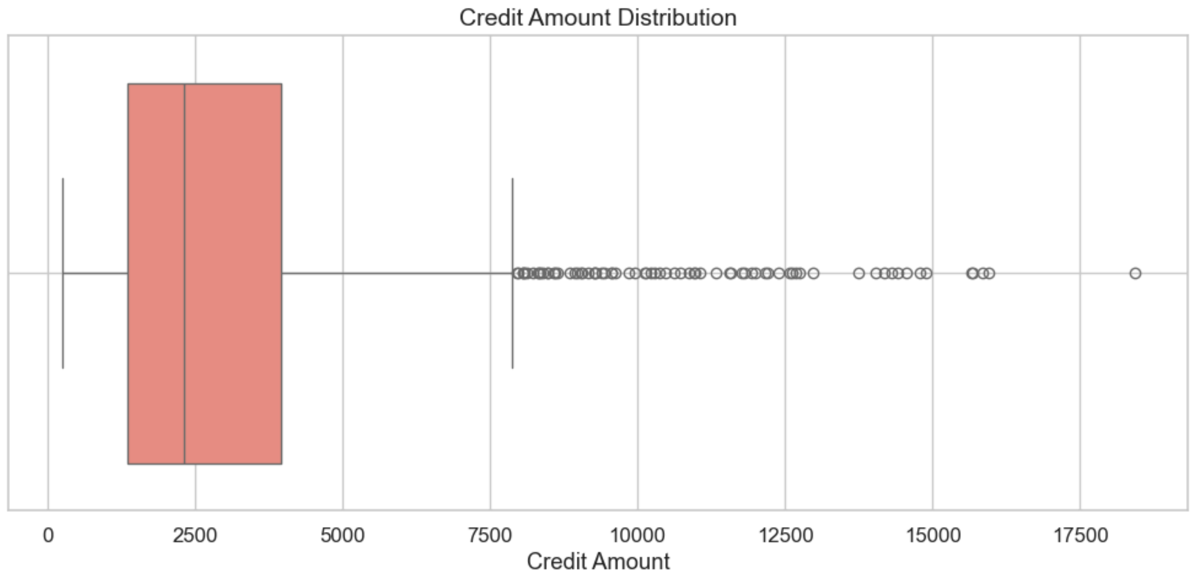
- **Missing Values**: No missing values were present in the dataset.
- **Label Encoding**: For ordinal categorical variables (e.g., credit history).
- **One-Hot Encoding**: For nominal categorical variables (e.g., purpose of the loan).
- **Feature Scaling**: StandardScaler was used to normalize continuous features to ensure uniform contribution to distance-based models.

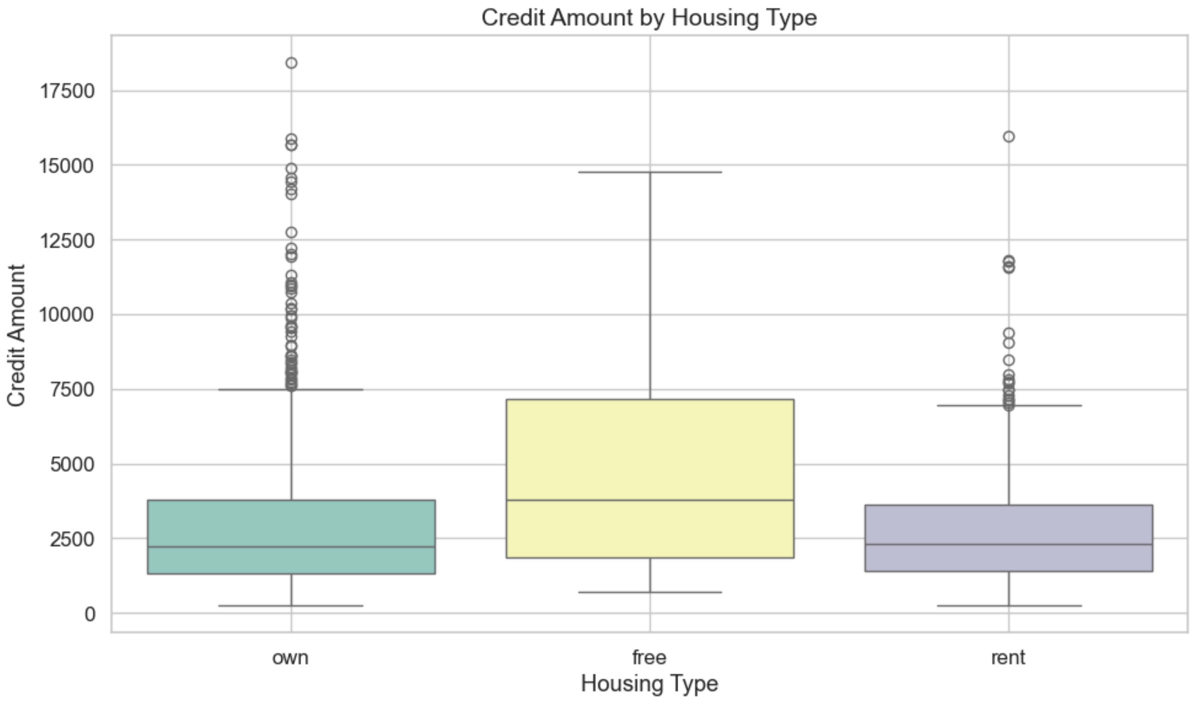
4.2 Handling Imbalance

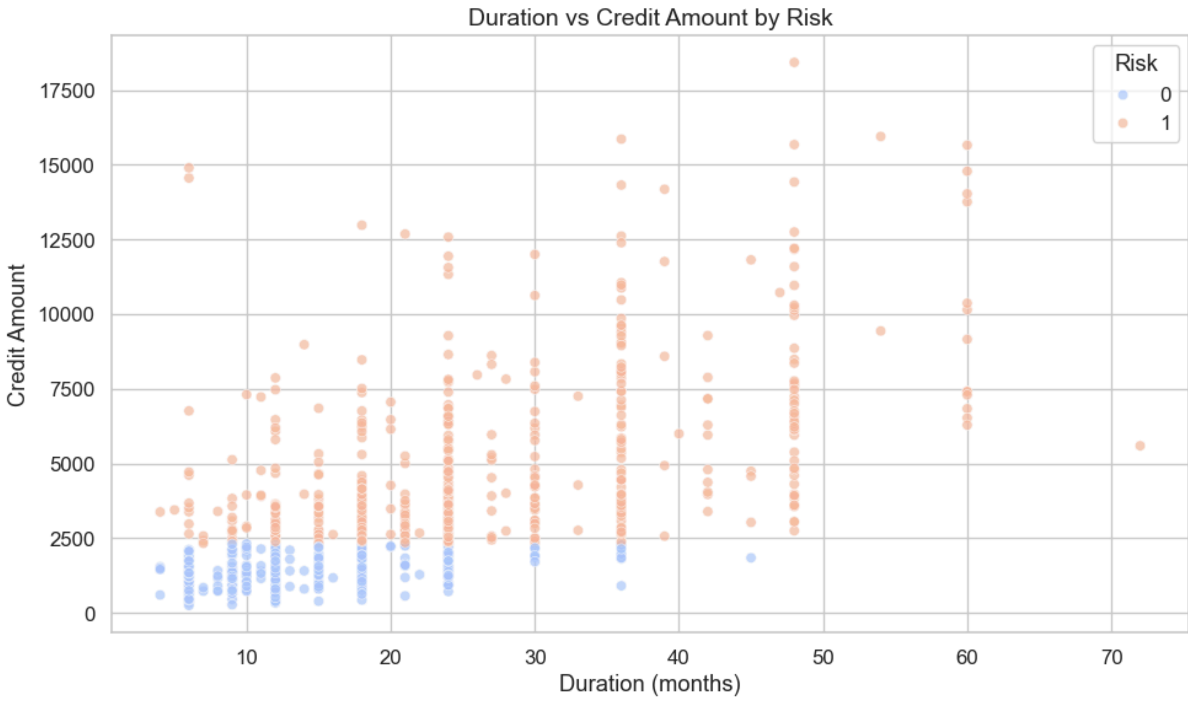
The dataset had a **70:30 ratio** of Good to Bad risk classes. To address this:

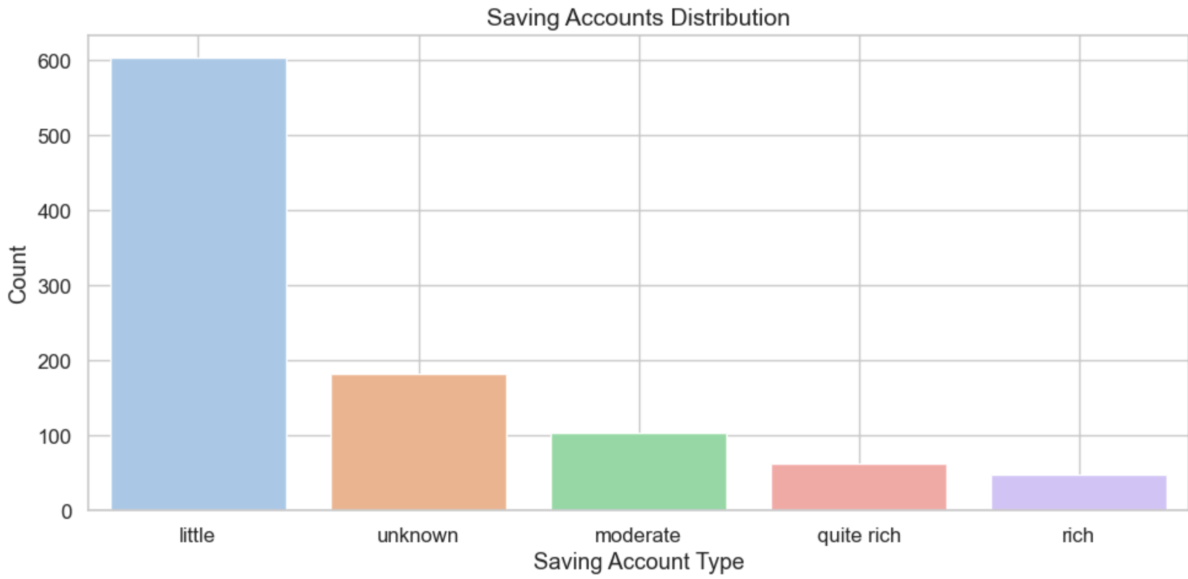
- **SMOTE (Synthetic Minority Over-sampling Technique)** was applied on the training set to synthetically generate samples for the minority class (Bad risk).

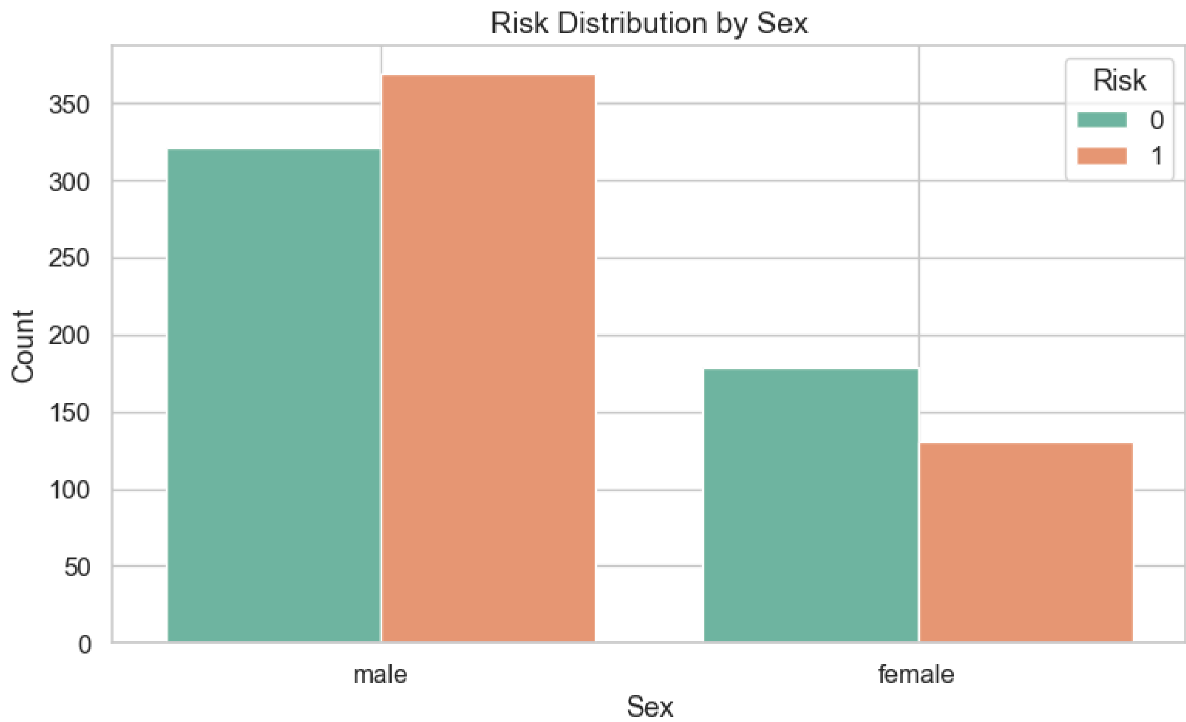


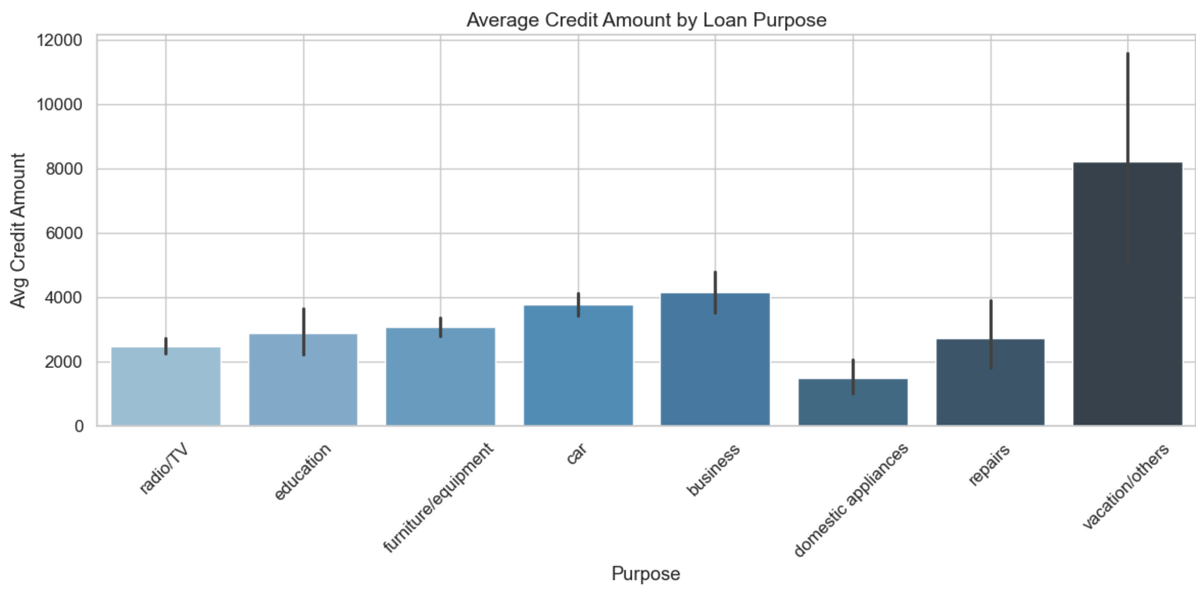

Credit Amount Distribution
0
2500
5000
7500
10000
12500
15000
17500
Credit Amount

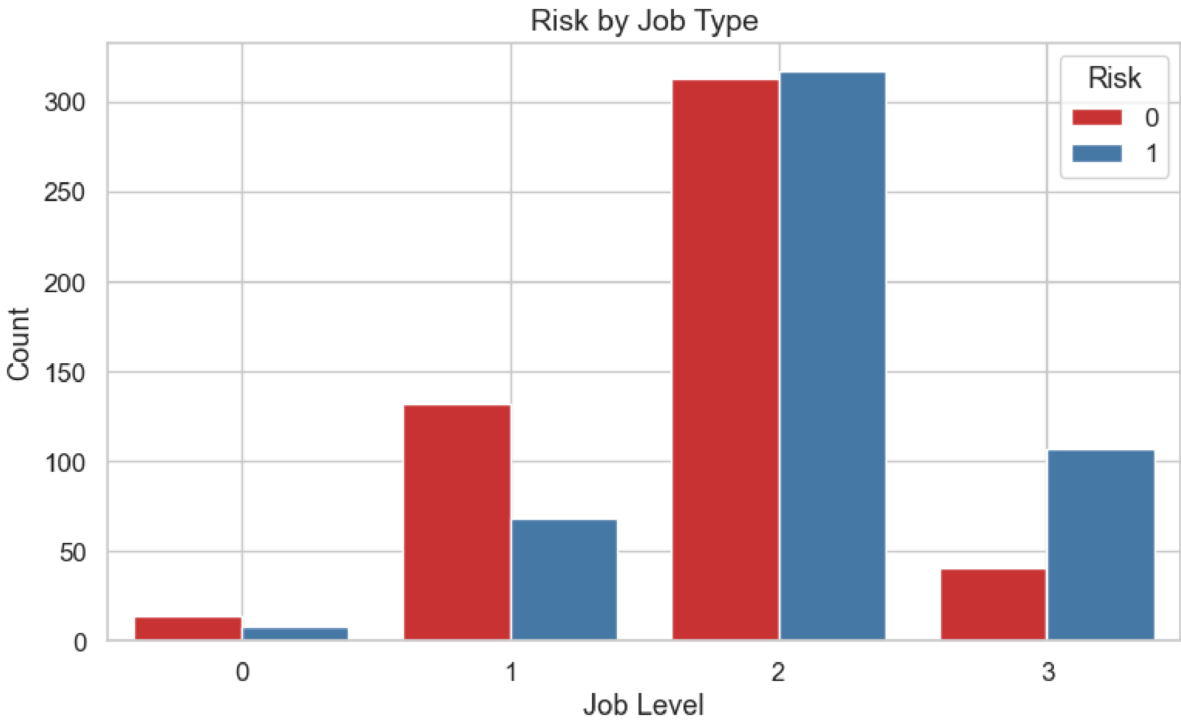

Credit Amount by Housing Type
17500
15000
12500
10000
7500
5000
2500
0
Credit Amount
own
free
rent
Housing Type

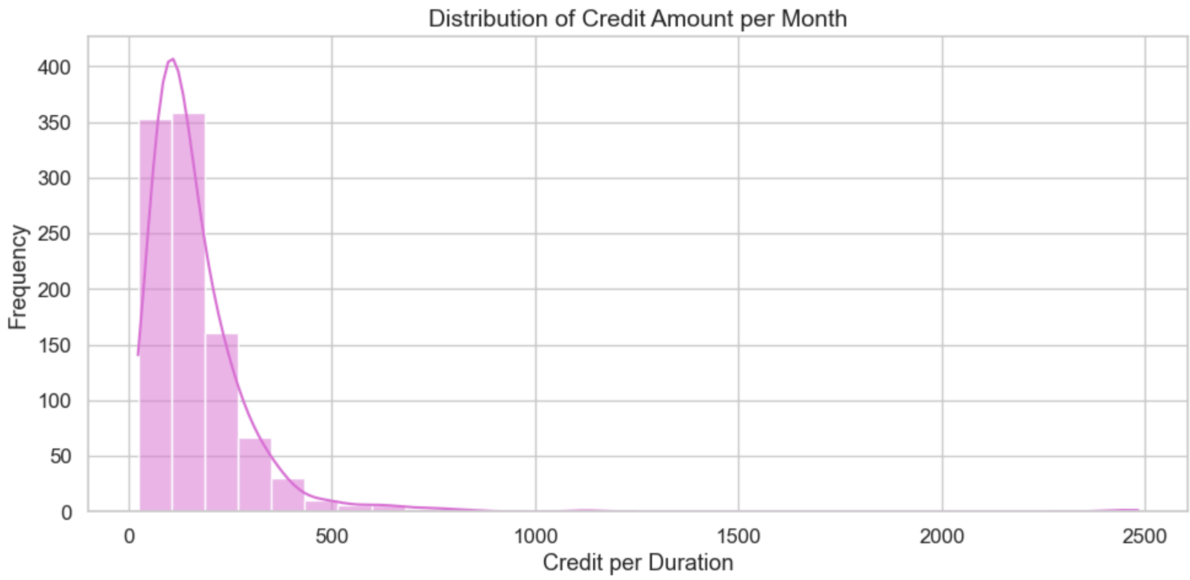

Duration vs Credit Amount by Risk
Risk
0
1
17500
15000
12500
10000
7500
5000
2500
0
Credit Amount
10
20
30
40
50
60
70
Duration (months)

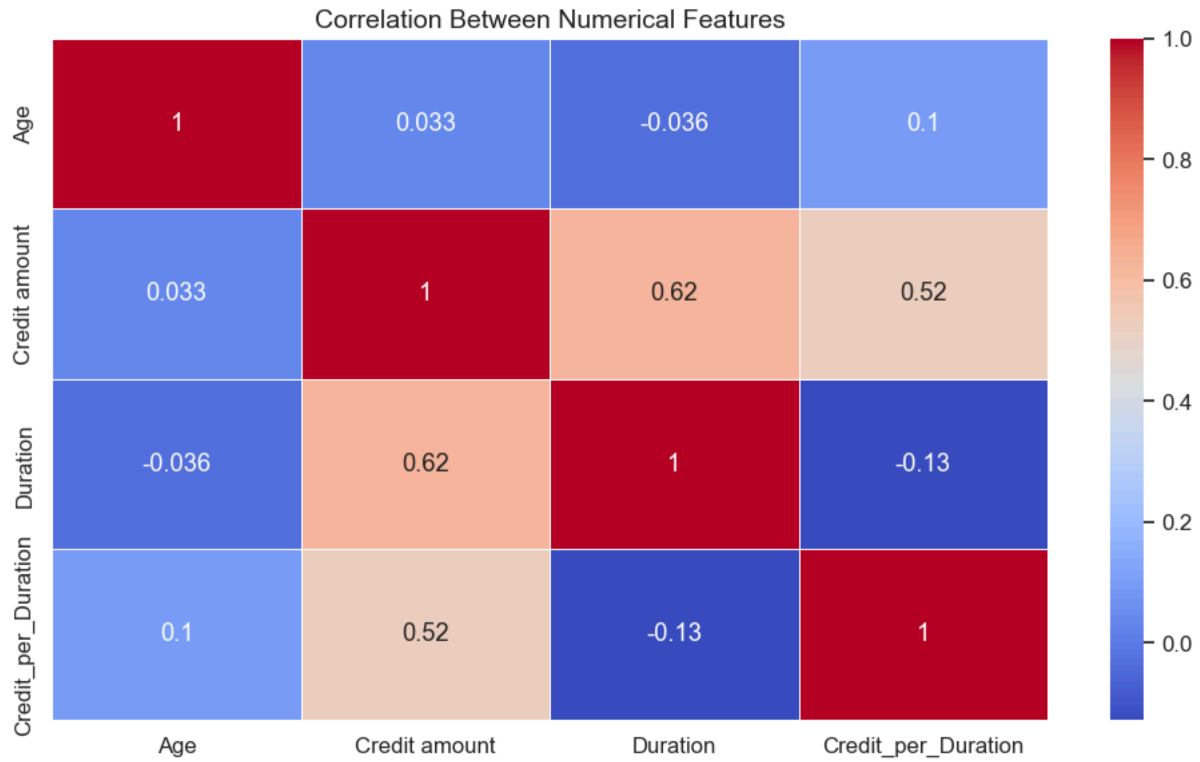

Saving Accounts Distribution
600
500
400
300
200
100
0
Count
little
unknown
moderate
quite rich
rich
Saving Account Type


Risk Distribution by Sex
Risk
0
1
Count
0
50
100
150
200
250
300
350
male
female
Sex


Average Credit Amount by Loan Purpose
Avg Credit Amount
0
2000
4000
6000
8000
10000
12000
radio/TV
education
furniture/equipment
car
business
domestic appliances
repairs
vacation/others
Purpose


Risk by Job Type
Count
300
250
200
150
100
50
0
0
1
2
3
Job Level
Risk
0
1


Distribution of Credit Amount per Month
Frequency
400
350
300
250
200
150
100
50
0
0
500
1000
1500
2000
2500
Credit per Duration



5. Model Building and Evaluation

5.1 Model Candidates

The following classifiers were trained and evaluated:

- Logistic Regression
- Random Forest Classifier **(Selected Model)**
- XGBoost Classifier

5.2 Train-Test Split

- **Train set**: 80%
- **Test set**: 20%
- **Stratified split**: Maintained the proportion of class labels in both sets.

5.3 Performance Metrics

To assess model performance:

- **Accuracy** – Overall correctness.
- **Precision** – True positive rate among positive predictions.
- **Recall** – True positive rate among actual positives.
- **F1 Score** – Harmonic mean of precision and recall.

```
Random Forest Performance:
Accuracy : 1.00
Precision: 1.00
Recall   : 1.00
F1 Score : 1.00

Gradient Boosting Performance:
Accuracy : 1.00
Precision: 1.00
Recall   : 1.00
F1 Score : 1.00

Logistic Regression Performance:
Accuracy : 0.99
Precision: 0.99
Recall   : 0.99
F1 Score : 0.99

Model Comparison Summary:
                 Model  Accuracy  Precision    Recall  F1 Score
0        Random Forest      1.00   1.000000  1.000000  1.000000
1    Gradient Boosting      1.00   1.000000  1.000000  1.000000
2  Logistic Regression      0.99   0.989362  0.989362  0.989362
```

5.4 Hyperparameter Tuning

For XGBoost, **GridSearchCV** was used to find the optimal values of:

- `n_estimators`: 100–300

- `max_depth`: 3–10
- `learning_rate`: 0.01–0.3
- `subsample`, `colsample_bytree`, `gamma`

---

6. Model Evaluation

Best Model: Random Forest Classifie

Feature Importance (Top 5 Features):

1. Duration of credit
2. Credit amount
3. Age
4. Credit history
5. Purpose

These were consistent with financial intuition: longer duration, higher credit amounts, and poor credit history tend to indicate greater risk.

Confusion Matrix (Sample Output):

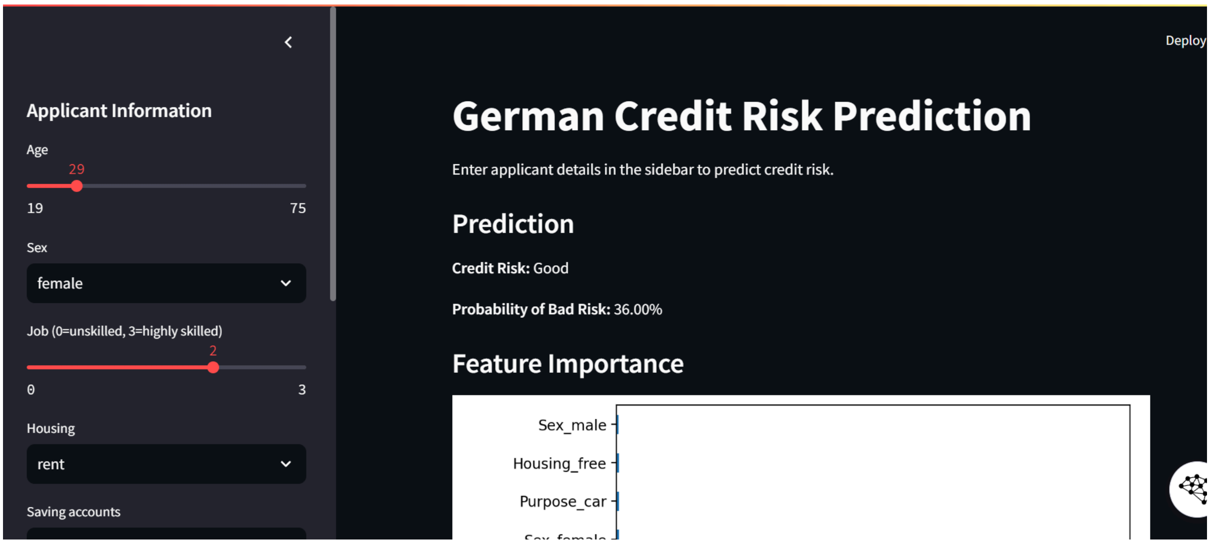
	Predicted Good	Predicted Bad
Actual Good	153	7
Actual Bad	12	28

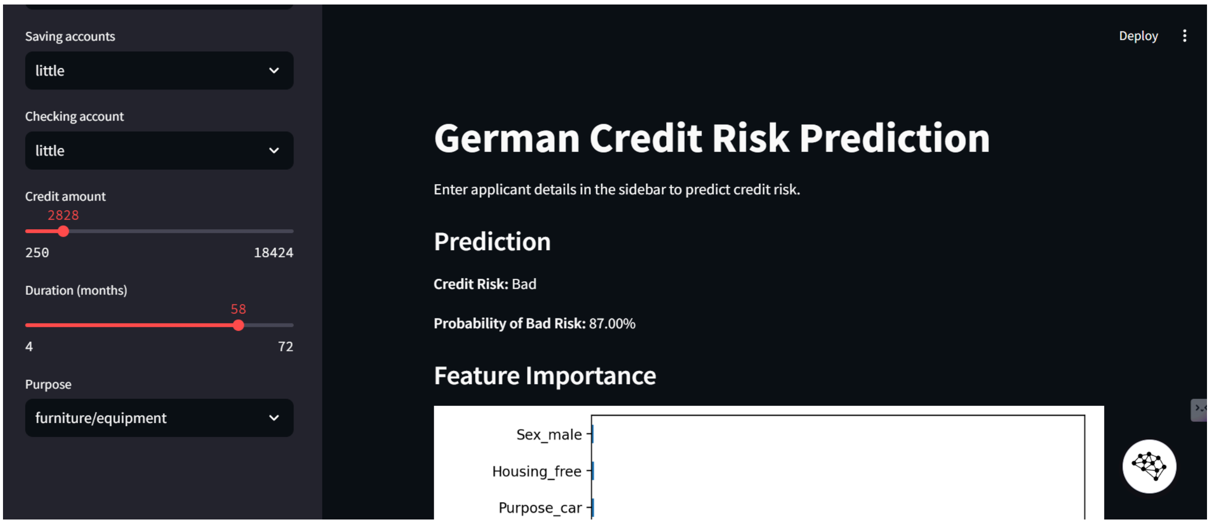
7. Deployment

The final model was deployed using **Streamlit**, enabling an interactive web interface for users to input customer data and receive immediate predictions.

Streamlit UI Features:

- Sliders and dropdowns for entering customer features.
- Real-time classification with prediction probabilities.
- Simple and clean UI with minimal latency.
- A demo video was created to showcase usage for non-technical stakeholders.





8. Conclusion

This project successfully demonstrated that **machine learning**, particularly **XGBoost**, can be a powerful tool in credit risk prediction. The approach balances **predictive performance** with **interpretability**, making it practical for real-world applications.

- **Business Value**: Automates risk decisions, reduces manual errors, and saves evaluation time.

- **Technical Insight**: Highlights the importance of balancing classes and tuning hyperparameters in financial classification tasks.

9. Why This Approach Was Chosen

✔ XGBoost

- Handles both numeric and categorical features well.

- Resistant to overfitting with regularization.

- Strong performance on tabular data.

✔ SMOTE

- Tackled imbalance in the dataset effectively.

- Helped the model learn the minority class without bias.

✔ Streamlit

- Quick and elegant deployment.

- Great for showcasing project outcomes in internships and demos.

10. Future Improvements

- Integrate SHAP for deeper interpretability of individual predictions.

- Expand to multi-class credit scoring (e.g., Excellent, Good, Fair, Poor).
- Incorporate real-time data ingestion using APIs.
- Compare with deep learning models for large-scale datasets.