Explainable AI for Credit Risk Detection

Natalia Leyenda Lodares - 202108204

December 29, 2025

1 Introduction

Credit risk detection is the task of evaluating the likelihood that a loan applicant will fail to repay a loan. Financial institutions rely on this evaluation to decide whether to approve or reject credit applications and under which conditions a loan should be granted.

In this project, the problem is formulated as a binary classification task. Given information about an applicant, such as financial status, credit history, and loan characteristics, the model predicts if the applicant presents low or high risk. The goal is to support credit decision making while minimising financial losses.

1.1 Motivation

Machine learning models are commonly used in credit risk detection, they are able to learn the complex relationships in tabular data. However, these decisions have a significant impact on individuals and organisations. A rejected loan application can affect a person's financial opportunities, and incorrect approvals can lead to financial losses.

This is why we need explainability in credit risk detection. Stakeholders must be able to understand why a model produces a certain prediction, identify potential biases, and evaluate if the model is trustworthy and reasonable.

The goal of this project is to build a predictive model and apply Explainable AI techniques to analyze, evaluate, and assess the model's behaviour. Explanations are used to understand the decision process and give reasonable recommendations to make the model better.

1.2 Stakeholders

The primary stakeholders we find in credit risk detection are financial institutions. They use these models to detect risk and make decisions. Explainability helps to understand and validate the model's decisions.

Another key group to consider is the loan applicants. For them, explanations are essential to ensure transparency and fairness. It's also important to provide feedback to be able to understand why the loan was rejected, when rejected.

Finally, there are regulators and auditors as an important stakeholder group. They are the ones that inspect and evaluate the models. They look for potential bias or unintended behaviour. Explainable AI techniques help them in their task.

2 Data and Model

The experiments in this project were done using the Statlog (German Credit) dataset from the UCI Machine Learning Repository [UCI94]. This dataset contains information about credit applicants and is used in research on credit risk detection and explainable AI.

Each instance represents a loan application described by a set of financial and personal attributes.

2.1 Features and Preprocessing

The dataset consists of a mix of numerical and categorical variables that describe the financial and personal situation of loan applicants. Numerical features include variables such as the credit amount, loan duration, age, and installment rate. Categorical features cover aspects like checking account status (e.g., A14 for no checking account), savings, employment history, housing, and the purpose of the loan (e.g., A40 for a new car).

Before training the model, it was necessary to preprocess the data to ensure it was in a suitable format. Numerical variables were scaled to a standard range, while categorical variables were transformed using one hot encoding. This encoding process is what creates specific features like `cat_status_A14` or `cat_purpose_A40` seen in the later analyses.

It is important to note that while the model operates on these individual encoded features to make predictions, the explanations provided are also aggregated. This grouping of categories back into their original variables makes the results much easier to understand.

2.2 Model Choice

For this task a tree based classification model was chosen. Tree based models work well with tabular data and can capture non linear relationships. In addition, they are compatible with different explainability techniques.

The model was implemented using the `scikit learn` library.

The dataset was split into training and test sets to evaluate model performance on unseen data.

3 Model Evaluation

Model performance was evaluated with standard classification metrics. Accuracy was used to see how many predictions were right in general, while precision and recall show how the model behaves for each specific class. For this project, recall for high risk applicants is the most important metric. Failing to identify a risky applicant is a serious issue because it can lead to financial losses.

In addition, theROC AUC was also calculated. This metric measures the model's ability to rank applicants by risk and is commonly used in credit scoring applications, as it is independent of a fixed classification threshold.

3.1 Results

The trained model achieved an accuracy of 0.79 on the test set. The ROC AUC score was close to 0.80, indicating a good ability to separate low and high risk.

The recall for high risk applicants was around 0.61, which means that the model is able to identify a significant portion of risky cases, although some remain undetected. At the same time, the model showed strong performance for low risk applicants, a typical trade off in credit risk modeling.

Additional metrics support this interpretation. The F1 score was approximately 0.63, showing reasonable balance between precision and recall for the high risk class.

Finally, the mean absolute error (MAE) and root mean squared error (RMSE) values were relatively low, suggesting stable and consistent predictions.

3.2 Interpretation of Performance

Overall, the results indicate that the model performs well, but it could be better. The imperfect recall for high risk cases highlights the limitations of relying only on the model.

In credit risk detection, it is important to understand why certain predictions are made and where the model may fail.

This shows that explainability is really necessary to understand how the model decides and to see if we can trust it.

4 Explainability Techniques

While the model's predictive performance is good, metrics alone are not enough to explain a decision process to ensure the model is transparent and can be trusted.

This project uses different explainability techniques to analyze the model. Global explanations are used to show overall behavior, while local explanations focus on specific individual predictions. Additionally, sanity checks and counterfactuals help verify if the explanations are actually useful and realistic.

4.1 Global Explanation: Feature Importance

The first step in the global analysis was to look at the feature importance scores. These scores give a general idea of which inputs have the most influence on the results across the whole dataset.

Originally, this analysis was done on one hot encoded features 1. This view shows that `cat_status_A14` is by far the most significant feature. While it might seem unusual for one specific category to be so dominant, in this dataset, A14 represents applicants with no checking account. For the model, the absence of a checking account appears to be the strongest single indicator for predicting credit risk, as it likely serves as a proxy for financial stability. Other important continuous variables include `num_credit_per_month` and `num_credit_amount`. However, this representation can be difficult to read because it treats every category of a single variable as a separate entity.

To make it more interpretable, the importance scores were aggregated back into the original variables 2. All categories belonging to the same variable were grouped together, while numerical variables were kept as they were.

When grouped, `status` remains the most critical factor, representing over 20% of the total importance. This suggests that the history and existence of a bank account are more predictive than the actual amount of credit requested. Other financial factors, such as monthly burden and property, also rank highly, while demographic variables like `residence_since` have the least impact.

This aggregated view offers a much clearer and more logical summary of the model's global behavior.

4.2 Global Explanation: SHAP Values

SHAP values were used to get a more reliable and grounded explanation. This method gives a value to each feature to show how much it affects a specific prediction.

The SHAP summary plot was first checked using the one hot encoded features 3. This plot shows how high or low values of a feature move the prediction. For example, the red dots for `cat_status_A14` (no checking account) being on the left shows that this status strongly reduces the predicted risk.

Just like with feature importance, SHAP values were aggregated by their original variable groups to make them easier to understand. The grouped analysis 4 confirms that financial indicators (like account status, savings, and credit amount) are the main drivers of the model. Demographic variables, such as age or gender, appear to have much less impact.

4.3 Local Explanation: Individual Predictions

While global explanations show how the model works on average, they do not explain the logic behind a single decision. Therefore, local explanations were also used to look at specific cases.

SHAP waterfall plots, like the one in Figure 5, show how the model gets from the average base value to the final prediction for one applicant. In this example, a high credit amount (+0.96) is the main reason for an increased risk, even though the applicant's account status (-0.71) tries to pull the risk down.

The analysis covered both low and high risk cases. In low risk examples, positive factors like a stable account status usually outweighs the risks. In high risk cases, there are often multiple factors working together, making the final decision very clear and hard to change with just one small adjustment.

4.4 Sanity Check: Faithfulness of Explanations

A sanity check was done to make sure the explanations actually represent how the model works. The most important variable identified, `status`, was removed and the model was retrained.

After removing this variable, the ROC AUC dropped from 0.80 to 0.74. This change in performance proves that the variable flagged as important by SHAP and Feature Importance is

truly essential for the model's accuracy. This gives more confidence that the explanations are not just showing random correlations.

Metric	Full Model	Model without Status
Accuracy	0.79	0.74
Recall	0.61	0.54
ROC AUC	0.80	0.74
MAE	0.21	0.26
RMSE	0.46	0.51
F1 Score	0.63	0.56

Table 1: Performance comparison between the full model and the model without account status

4.5 Counterfactual Explanations

Counterfactual explanations were used to see how a prediction might change if the input data were different. This helps answer what an applicant would need to change to get a different result.

A counterfactual analysis was performed on an applicant (the same as for the local explanation) with an initial predicted probability of default of approximately 0.198. According to the previous SHAP analysis, the high credit amount and the specific loan purpose were the primary factors increasing the risk for this individual.

The following scenarios were tested to see how the model would react:

- **Reducing Credit Amount**: Lowering the requested credit from the original value to 6,000 reduced the probability of default to approximately 0.104.
- **Changing Loan Purpose**: Changing the purpose from "A42" (furniture/equipment) to "A44" (domestic appliances) also resulted in a lower probability of approximately 0.107.
- **Combined Changes**: When both the credit amount was reduced and the purpose was changed simultaneously, the predicted risk dropped significantly to 0.053.

These results show that the model responds in a logical and coherent way to actionable changes. Also, these counterfactuals give practical guidance for decision support, as they show exactly how modifying loan conditions can lead to a more favorable outcome. This analysis complements the global and local views provided by the SHAP values, offering a more complete picture of the model's behavior.

5 Actionable Insights and Limitations

5.1 Actionable Insights

The explainability analysis provides several practical insights for credit risk detection. Both the global and local explanations show that the model is mainly driven by financial stability and credit sustainability, specifically account status, credit amount, duration, and monthly burden.

From a practical point of view, this suggests that risk can often be reduced by adjusting the loan structure. For example, lowering the credit amount or changing the loan purpose can lead to a lower risk prediction for some applicants. The counterfactual analysis confirms this, showing that realistic changes to loan conditions can significantly alter the outcome. These findings can help loan applicants by suggesting alternative loan offers.

5.2 Limitations

Even though these techniques are very useful, there are some limitations to consider. First, explanations depend entirely on the model and the quality of the data. If the training data is biased or incomplete, the explanations will reflect those flaws.

Second, SHAP values can sometimes be sensitive to features that are highly correlated with each other. Because of this, explanations should be used as supportive evidence rather than absolute facts.

Finally, the counterfactuals were created through manual changes. While they are very informative, they might not represent the optimal change.

6 Conclusion

This report has presented a complete Explainable AI case study for credit risk detection. A machine learning model was developed and tested, showing strong performance while also highlighting the trade offs between accuracy and risk detection.

The main focus was on the explainability analysis. Global views identified the most important variables, while local explanations showed the logic behind individual decisions. Sanity checks proved that these explanations are faithful to the model's true behavior, and counterfactuals demonstrated how specific changes can influence the score.

Overall, these results show that explainability is essential for building trustworthy systems in high risk fields like finance. Instead of replacing human judgment, this approach supports it by providing insights that make the model's decisions easier to understand and trust.

References

[UCI94] UCI Machine Learning Repository. *Statlog (German Credit Data)*. `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data`. Accessed: 2025. 1994.

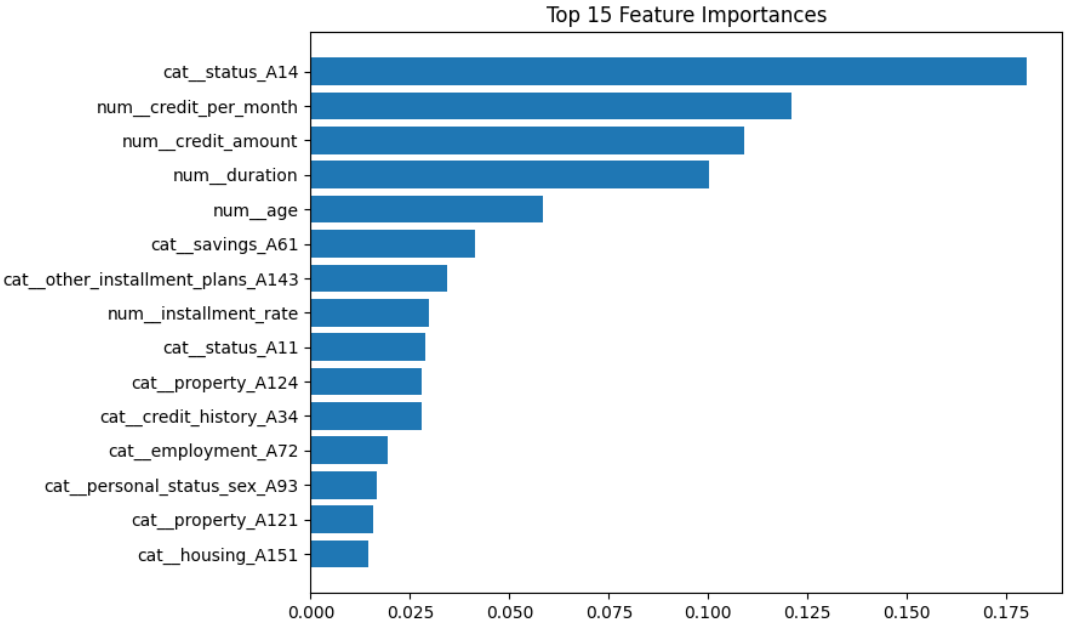

Figure 1: Feature Importance

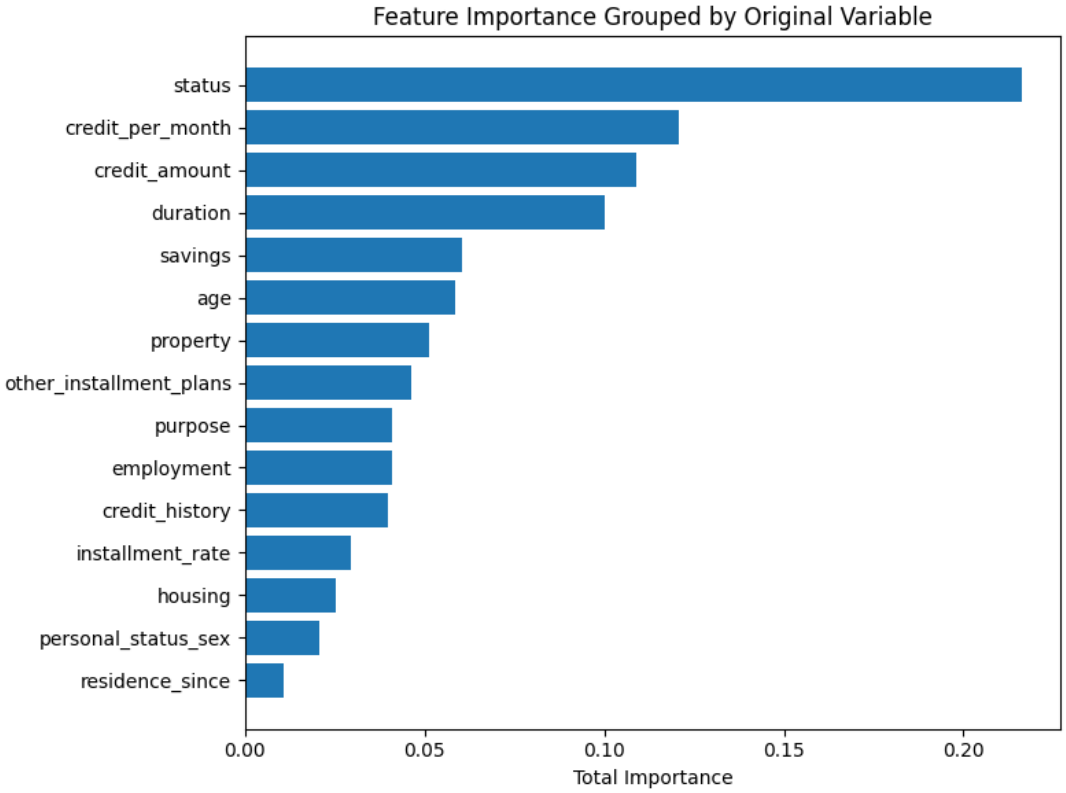

Figure 2: Feature Importance Aggregated

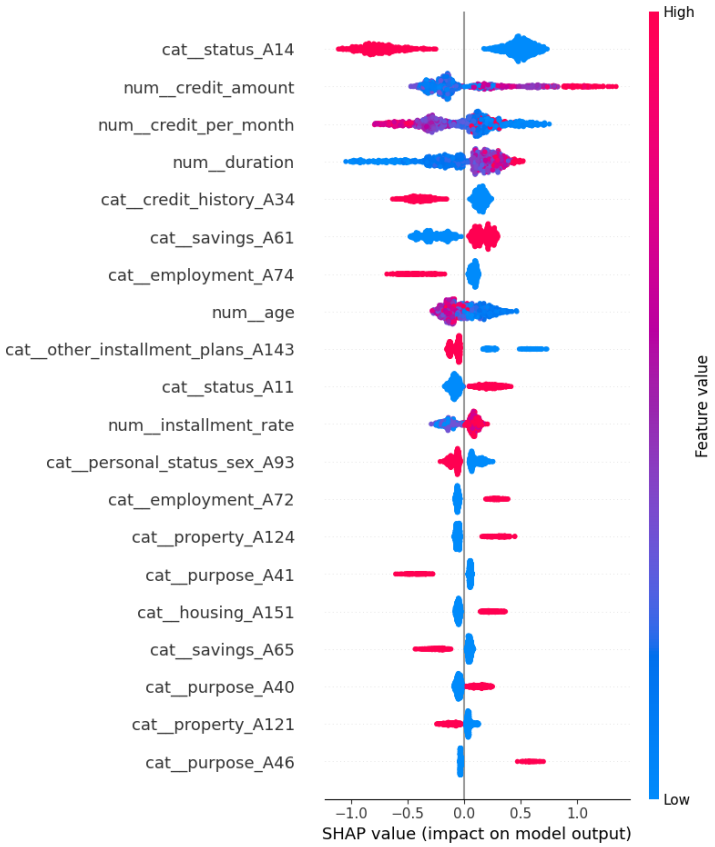

Figure 3: SHAP

group	shap_importance
status	0.743920
savings	0.321487
credit_amount	0.301652
employment	0.282891
purpose	0.268841
credit_per_month	0.267048
credit	0.261617
duration	0.255720
other	0.231696
property	0.175776
housing	0.136741
age	0.132955
personal	0.124533
installment_rate	0.106279
existing_credits	0.052031

Figure 4: SHAP Aggregated

Figure 5: SHAP for local explanations