Machine Learning Models Applied to German Credit Data

Michael Grandinetti

2024-03-16

Contents

Introduction

Credit risk assessment is a critical function in financial institutions, aiming to determine the likelihood that a borrower will default on their obligations. Accurate modeling of default risk enables lenders to make informed decisions, manage financial exposure, and comply with regulatory standards. In this project, we explore the German Credit Data Set, a well-known benchmark in credit scoring research, to develop predictive models that can effectively classify whether a customer is likely to default.

The dataset includes 1,000 individuals with 20 diverse attributes ranging from financial history and employment status to personal demographics and credit-related details. The binary response variable, Default, indicates whether or not a customer defaulted on their loan. The primary goal of this analysis is to build and evaluate statistical and machine learning models that can predict default status with high accuracy and interpretability.

Several modeling techniques were applied, including logistic regression, linear and quadratic discriminant analysis, decision trees, K-nearest neighbors, and ensemble methods like Random Forest and XGBoost. Performance was assessed using 5-fold cross-validation, focusing on metrics such as AUC, sensitivity, and specificity. This study ultimately seeks to identify a model that balances predictive power with practical utility in the context of credit risk classification.

German Credit Data Set

The German credit dataset was originally provided by Prof. Dr. Hans Hofmann from the University of Hamburg and distributed by the UCI Machine Learning Repository. It contains 1,000 records of loan applicants in Germany, and the goal is to classify applicants as creditworthy (Good) or not creditworthy (Bad) based on various financial and personal attributes. The dataset contains 20 attributes/variables that will be used to predict the Default status of the 1,000 individuals included in this dataset. Below is a table showing each attribute contained in the dataset, as well as the meaning and data type of each attribute.

Table 1: German credit dataset attribute summary.

Attribute	Variable	Type	Description
Response	Default	Binary	0 = No default, 1 = Default
1	Checking account status	Categorical	A11–A14: <0 DM, 0–199 DM, 200 DM, none
2	Duration in months	Numerical	-
3	Credit history	Categorical	A30–A34: from no credit issues to critical account
4	Purpose	Categorical	A40–A410: car, furniture, education, etc.
5	Credit amount	Numerical	-
6	Savings account/bonds	Categorical	A61–A65: <100 DM to unknown
7	Employment since	Categorical	A71–A75: unemployed to 7 years
8	Installment rate	Numerical	Percentage of income
9	Personal status & sex	Categorical	A91–A95: male/female, marital status
10	Other debtors/guarantors	Categorical	A101–A103: none, co-applicant, guarantor
11	Residence duration	Numerical	Years at current residence
12	Property	Categorical	A121–A124: real estate, insurance, car, unknown
13	Age	Numerical	Age in years
14	Other installment plans	Categorical	A141–A143: bank, stores, none
15	Housing	Categorical	A151–A153: rent, own, free
16	Credits at this bank	Numerical	Number of credits
17	Job	Categorical	A171–A174: unskilled to management
18	Maintenance dependents	Numerical	No. of people liable for maintenance
19	Telephone	Categorical	A191–A192: none or yes
20	Foreign worker	Categorical	A201–A202: yes or no

Below in Table 2 includes a summary of the dataset where the counts for the factor levels of each qualitative attribute and general distribution statistics for the attributes that have numerical data types included in this dataset are shown.

Table 2: The number of observations for each factor level of the qualitative attributes and distribution statistics for the numerical attributes in the German credit dataset.

```
##  Default checkingstatus1    duration    history      purpose        amount
##  0:700   A11:274         Min.   : 4.0   A30: 40   A43    :280   Min.   :  250
##  1:300   A12:269         1st Qu.:12.0   A31: 49   A40    :234   1st Qu.: 1366
##          A13: 63         Median :18.0   A32:530   A42    :181   Median : 2320
##          A14:394         Mean   :20.9   A33: 88   A41    :103   Mean   : 3271
##                          3rd Qu.:24.0   A34:293   A49    : 97   3rd Qu.: 3972
##                          Max.   :72.0             A46    : 50   Max.   :18424
##                                                   (Other): 55
##  savings   employ     installment    status     others    residence property
##  A61:603   A71: 62   Min.   :1.000   A91: 50   A101:907   1:130     A121:282
##  A62:103   A72:172   1st Qu.:2.000   A92:310   A102: 41   2:308     A122:232
##  A63: 63   A73:339   Median :3.000   A93:548   A103: 52   3:149     A123:332
##  A64: 48   A74:174   Mean   :2.973   A94: 92              4:413     A124:154
##  A65:183   A75:253   3rd Qu.:4.000
##                      Max.   :4.000
##
##       age        otherplans housing    cards     job      liable    tele
##  Min.   :19.00   A141:139   A151:179   1:633   A171: 22   1:845   A191:596
##  1st Qu.:27.00   A142: 47   A152:713   2:333   A172:200   2:155   A192:404
##  Median :33.00   A143:814   A153:108   3: 28   A173:630
##  Mean   :35.55                         4:  6   A174:148
##  3rd Qu.:42.00
##  Max.   :75.00
##
##  foreign
##  A201:963
##  A202: 37
##
##
##
##
##
```

Exploratory Data Analysis (EDA)

We will start analyzing the German credit dataset using Exploratory Data Analysis (EDA), which is used to summarize the attributes of the German credit data set using statistical plots. The analysis includes a correlation plot for the numerical attributes in the dataset shown in Figure 1. Then we will use bar plots showing the distribution of Default Status, Housing Status, and Employment Status shown in Figure 2. And lastly, we look at box plots showing the distribution of Loan Duration, Age, and Savings Amount all grouped by Default Status show in Figure 3.

The correlation plot below shows the correlation coefficient calculated between the numerical attributes in the German credit dataset. A correlation coefficient is a statistical measure used to quantify the strength of the linear relationship between two variables. The correlation coefficient can take on values from -1 to 1, where a coefficient of 1 would suggest a perfectly positive linear relation between the two variables and a coefficient of -1 would suggest a perfectly negative linear relation between the two variables. The correlation plot below shows the correlation coefficients between two attributes by both the size and color of the circles found in the plot.

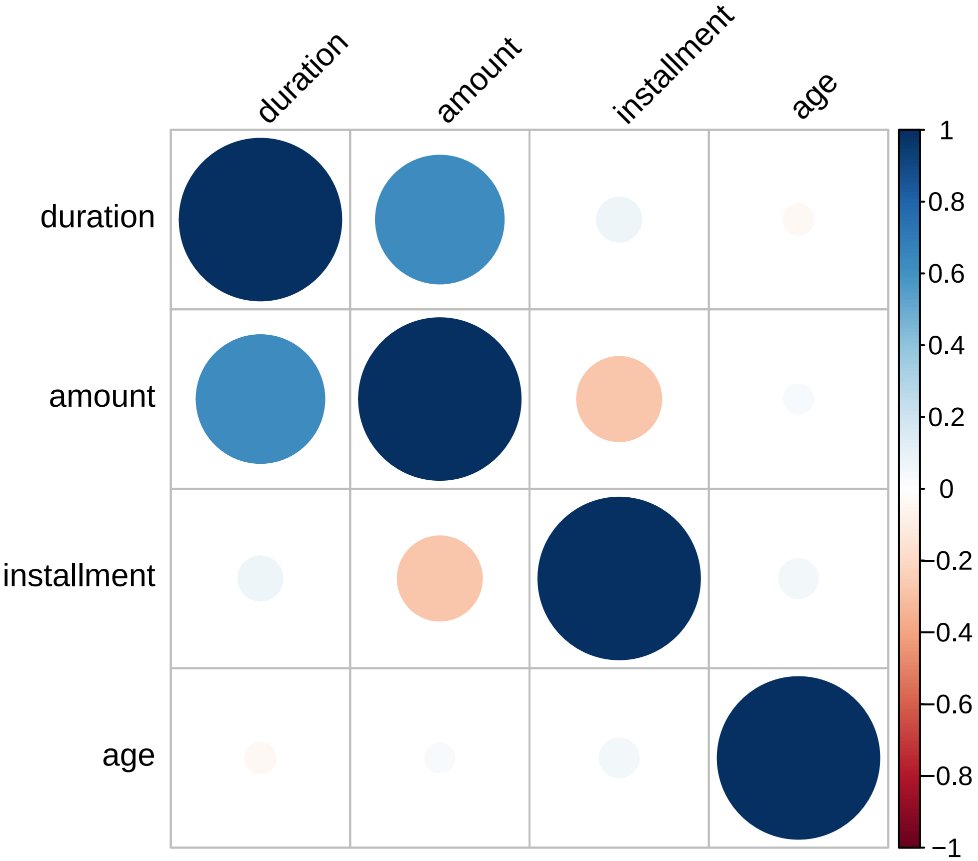

Figure 1: Correlation plot for the numeric attributes from the German credit dataset.

Analyzing the correlation plot above gives the highest correlation between attributes of Credit Amount and Loan Duration showing a correlation of about 0.625 with confidence intervals of (0.5857, 0.6614) at the 5% level, showing a statistically significant correlation. Although 0.625 is moderate in strength it does suggest a moderate positive linear relation with respect to Credit Amount and Loan Duration. The correlation coefficient between the Installment Rate as a percentage of disposable income and Credit Amount was calculated to be about -0.271 with confidence intervals of (-0.3278, -0.2129), indicating statistical significance at the 5% level. This correlation would show a slight negative linear relation with respect to the Installment Rate as a percentage of disposable income and Credit Amount. The correlation coefficient between Loan Duration and the Installment Rate as a percentage of disposable income was calculated to be about 0.0747 with confidence intervals of (0.0128, 0.1361) suggesting a statistical significance at the 5% level, however it should be noted how small the coefficient estimate is indicating a very small linear relation between the two attributes.

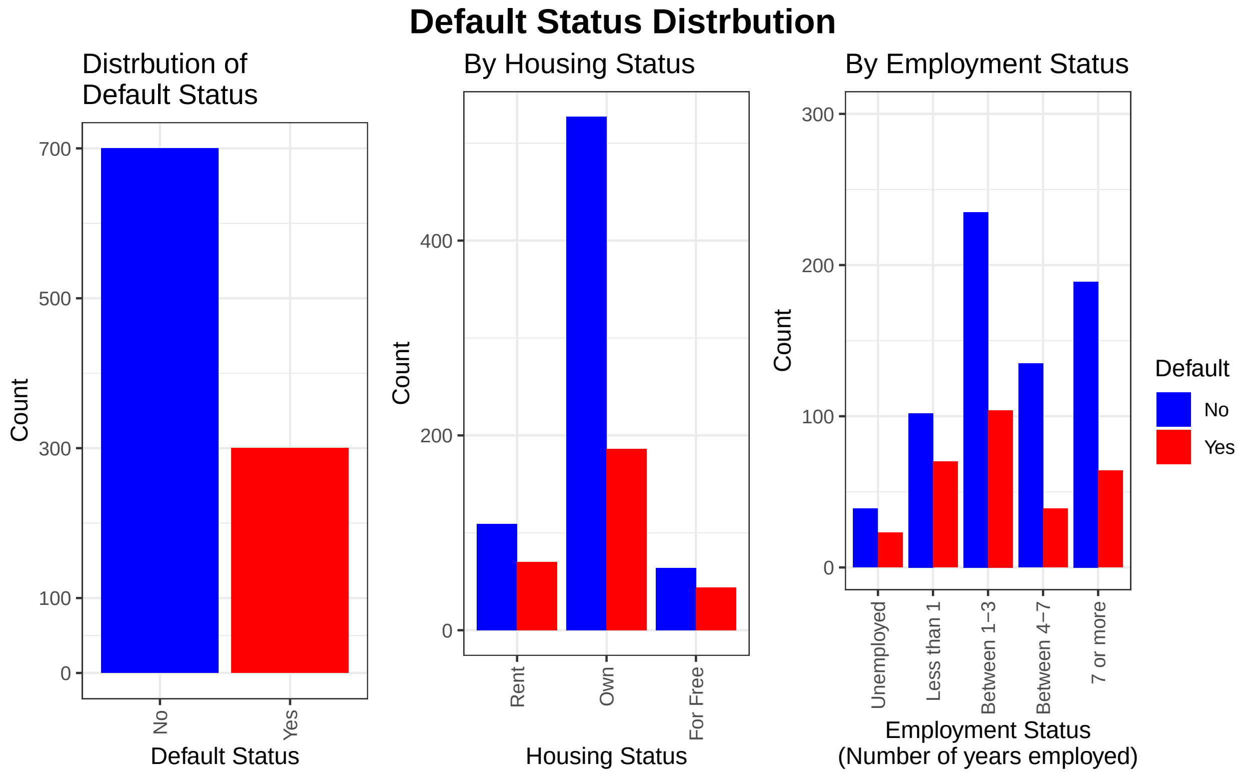

Figure 2: Bar plots showing the distribution of Default Status (left), bar plots for Housing Status (center), Employment Status (right) grouped by Default Status.

Starting with the left plot in Figure 2 showing the distribution of Default Status in the German credit data set shows a 70/30 split of the non-Default class and the Default class, as the non-Default class is represented by 700 observations and the Default class is represented by 300 observations. The bar plot on the right for the distribution of the Default and non-Default classes partitioned by employment status shows that the Unemployed group Defaulted with a rate of about 37.10%, the people working less than a year had a Default rate of about 40.70%, those who have worked between 1-3 years had a Default rate of around 30.68%, those who have worked between 4-7 years Defaulted with a rate of about 22.41%, and the people have worked for over 7 years had a Default rate of about 25.30%.

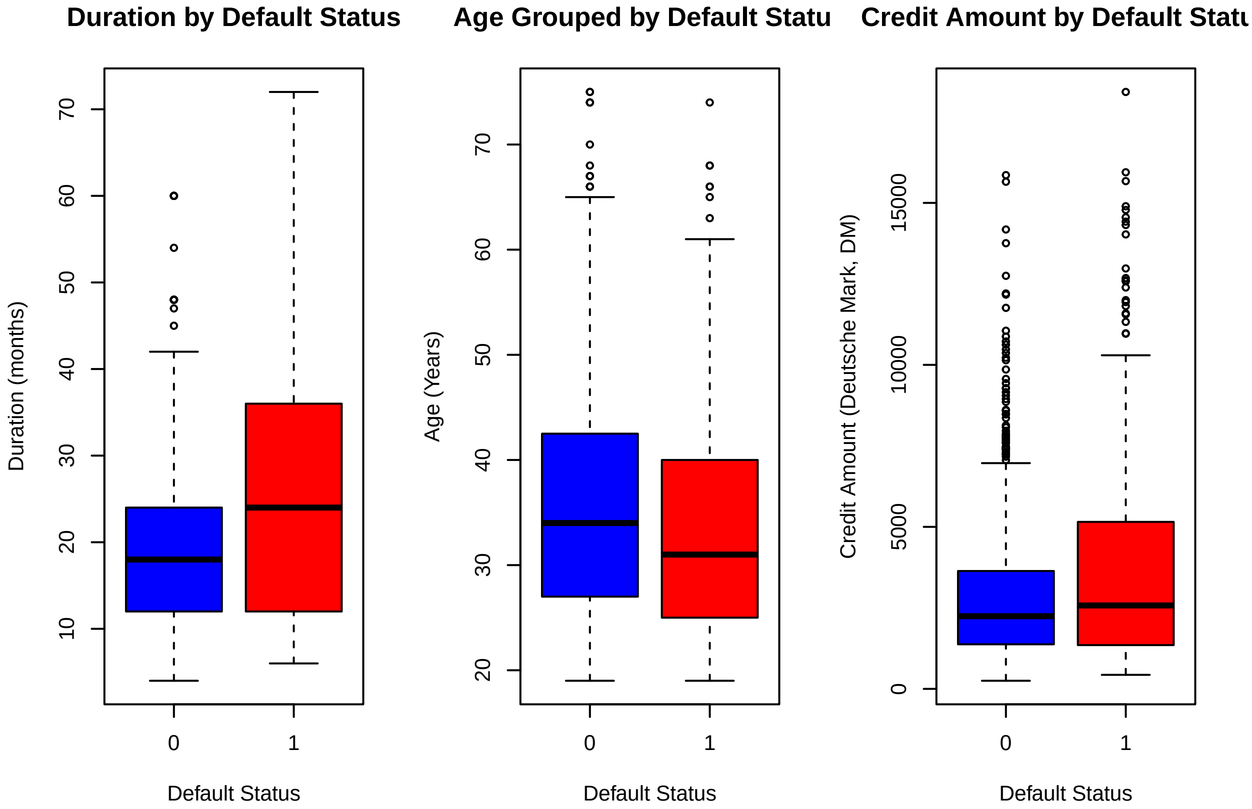

Figure 3: Box plot of the distributions of Loan Duration (left), Age (center), and Credit Amount grouped by Default Status.

The box plot of the distribution of Loan Duration grouped by Default Status indicates on average those who have defaulted have higher Loan Duration, as the mean value for the Default class was 24.9 months, while the mean for the non-default class was 19.2 months. We can also see from the box plot that the distribution of the Default class in terms of Loan Duration is widely spread out in comparison to the Default class which has a much more narrow interval for the distribution. The right box plot of Figure 3 showing the distribution of Credit Amount by Default Status shows similar behavior. The Default class in terms of Credit Amount has a higher mean value than the non-Default class, with a mean value of DM3938 while the non-Default class had a mean value of DM2985. The spread of the Default class is also wide in comparison to the non-Default class distribution in terms of Credit Amount. The distribution of Age with respect to Default Status shown in the center plot of Figure 3 indicates that on average those in the non-Default class are older than those who are in the Default class. The mean Age for the non-Default class was about 36 years old, while the mean Age for the Default class was 34 years old. Here the spread for the distribution of Age with respect to Default Status is very similar in both classes.

Methodology

Types of Models

Logistic Regression

Logistic regression is a generalized linear model used for binary classification. It models the log-odds of the probability of the default class as a linear combination of the input variables. The model estimates the probability that a given input belongs to class 1 (Default), using the logistic function to constrain the output between 0 and 1. Classification is typically based on whether this predicted probability exceeds a chosen threshold.

Linear Discriminant Analysis (LDA)

LDA is a probabilistic classification method that assumes each class follows a multivariate normal distribution with a shared covariance matrix. It estimates the class-conditional densities and uses Bayes' theorem to compute posterior probabilities for each class. The prediction is made by assigning the observation to the class with the highest posterior probability. LDA is optimal under the assumption of equal covariance and normally distributed predictors.

Quadratic Discriminant Analysis (QDA)

QDA extends LDA by relaxing the assumption of equal covariance matrices between classes. Each class is still modeled as a multivariate normal distribution, but with its own covariance matrix, allowing more flexible and nonlinear decision boundaries. QDA is more adaptable than LDA in situations where class covariances differ significantly but may overfit with small sample sizes.

K-Nearest Neighbors

KNN is a non-parametric classification algorithm that assigns a class label based on the majority vote of the k closest training observations (measured typically using Euclidean distance). The choice of k is a tuning parameter that balances bias and variance: lower values capture local structure, while higher values smooth out noise. No explicit training is done; all computation occurs at prediction time.

Decision Tree

Decision trees recursively partition the feature space based on values of the predictors to build a tree structure that classifies observations. At each node, the algorithm selects the variable and split that best separates the data (e.g., using Gini impurity or entropy). The tree continues growing until a stopping criterion is met (e.g., minimum node size or maximum depth). Trees are interpretable but prone to overfitting unless pruned or regularized.

Random Forest

Random forest is an ensemble learning method that builds a large number of decision trees on bootstrapped subsets of the training data, introducing randomness in feature selection at each split. The final prediction is made by aggregating (majority vote) the predictions from all individual trees. This process reduces variance and improves generalization compared to a single decision tree.

XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that builds decision trees sequentially to minimize classification error. Each new tree is trained on the residuals of the previous trees, gradually improving the model. XGBoost incorporates regularization, tree pruning, and parallel computation, making it both efficient and robust to overfitting. It is particularly effective on structured/tabular data.

5-fold Cross Validation Setup

For estimating the test error rate of each method a 5-fold cross validation resampling method was implemented where 5 validation sets of 200 observations each were used. This means we will have 5 estimates for the misclassification/test error rate that corresponds to the 5 different validation sets and training sets that will be used to model the data from the German credit dataset. So, our estimated test error rate will be an average of the 5 test error estimates based on the 5-fold cross validation. This is shown in the equation below where the ith test error rate represents the estimated test error of the ith validation set and the index j represents the jth observation.

$$
\begin{aligned}
CV_5 &= \frac{1}{5}\sum_{i=1}^{5}\left[\frac{1}{200}\sum_{j=1}^{200} I(y_{ij} \neq \hat{y}_{ij})\right] \\
&= \frac{1}{1000}\sum_{i=1}^{5}\sum_{j=1}^{200} I(y_{ij} \neq \hat{y}_{ij})
\end{aligned}
$$

Where $I(y_{ij} \neq \hat{y}_{ij})$ returns the value of 1 when the prediction for the response variable does not equal the observed value in the validation set and will return a 0 when the prediction matches the observed value.

It is important to note that the 5 folds created for this cross validation were stratified by using the createFolds() function in R. So, the distribution of the Non-Default class and the Default class in each validation set and training set used are the same for each set with 70% of the observations being observed to be the Non-Default class, while the remaining 30% of the observations observed to be the Default class.

Optimal Threshold Criterion

For comparison later, the Area Under the Curve (AUC), sensitivity, specificity, and misclassification rate were calculated. To determine which threshold value should be used the Youden Index will be calculated based on the ROC curve for each model. The ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates the performance of a binary classification model at various threshold values. The Youden Index (J) is a single statistic that summarizes the performance of a binary classifier across different thresholds using equally weighted sensitivity and specificity values from the ROC curve. For this data the sensitivity of the model will refer to the probability of predicting the Default class (Yes) correctly, while the specificity of the model will refer to the probability of predicting the Non-Default class (No) correctly. The Youden Index (J) is defined as follows:

$$J = \text{sensitivity} + \text{specificity} - 1$$

The Youden Index ranges from 0 to 1, where 0 indicates that the classifier performs no better than random chance, and 1 indicates perfect classification. The Youden Index provides a balance between sensitivity and specificity, making it especially useful when both types of classification errors (false positives and false negatives) are considered equally undesirable.

By computing the Youden Index across all possible threshold values on the ROC curve, the threshold that maximizes J is selected as the optimal threshold. This threshold is where the classifier achieves the best tradeoff between correctly identifying defaults (sensitivity) and correctly identifying non-defaults (specificity).

Another important note regarding the German credit dataset used in this report is that the numerical predictors in the data were scaled using the scale() function in R to reduce the effect that large input parameter values would have on the models in terms of adding bias to the model and to prevent overfitting.

Results

Logistic Regression

For the logistic regression method we will first fit a model using all 20 of the attributes included in the German credit dataset. We will then test for the significance of each attribute using a significance level of 5% to ultimately make a reduced model including only the attributes that have a significance level of under 5%. A summary of the fitting of the Logistic Regression model is shown below.

Table 3: Reduced Logistic Regression Model using 11 of the 20 attributes in the German credit dataset.

```
##
## Call:
## glm(formula = Default ~ checkingstatus1 + duration + history +
##     purpose + amount + savings + installment + status + others +
##     otherplans + foreign, family = binomial, data = germancredit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2211  -0.7062  -0.3938   0.7637   2.6952
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.71963    0.58766   4.628 3.69e-06 ***
## checkingstatus1A12 -0.41105    0.20971  -1.960 0.049986 *
## checkingstatus1A13 -1.06856    0.35830  -2.982 0.002861 **
## checkingstatus1A14 -1.77894    0.22637  -7.858 3.89e-15 ***
## duration            0.33830    0.10656   3.175 0.001499 **
## historyA31         -0.13614    0.51979  -0.262 0.793395
## historyA32         -0.85881    0.40460  -2.123 0.033784 *
## historyA33         -0.98000    0.46075  -2.127 0.033421 *
## historyA34         -1.58722    0.42668  -3.720 0.000199 ***
## purposeA41         -1.55580    0.35986  -4.323 1.54e-05 ***
## purposeA410        -1.57525    0.75400  -2.089 0.036690 *
## purposeA42         -0.66119    0.24866  -2.659 0.007837 **
## purposeA43         -0.89680    0.23933  -3.747 0.000179 ***
## purposeA44         -0.56350    0.74281  -0.759 0.448091
## purposeA45         -0.17819    0.53826  -0.331 0.740600
## purposeA46          0.17995    0.38464   0.468 0.639890
## purposeA48         -2.13264    1.22434  -1.742 0.081531 .
## purposeA49         -0.80928    0.32256  -2.509 0.012110 *
## amount              0.31492    0.11552   2.726 0.006409 **
## savingsA62         -0.26709    0.27390  -0.975 0.329488
## savingsA63         -0.42714    0.39239  -1.089 0.276349
## savingsA64         -1.33072    0.50378  -2.641 0.008255 **
## savingsA65         -0.96766    0.25434  -3.805 0.000142 ***
## installment         0.33990    0.09429   3.605 0.000312 ***
## statusA92          -0.12430    0.36606  -0.340 0.734175
## statusA93          -0.77049    0.35999  -2.140 0.032331 *
## statusA94          -0.27865    0.43452  -0.641 0.521342
## othersA102          0.53228    0.40010   1.330 0.183402
## othersA103         -1.02216    0.41279  -2.476 0.013278 *
## otherplansA142     -0.14165    0.40069  -0.354 0.723700
## otherplansA143     -0.65360    0.23329  -2.802 0.005083 **
## foreignA202        -1.29016    0.62201  -2.074 0.038064 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.73  on 999  degrees of freedom
## Residual deviance:  920.95  on 968  degrees of freedom
## AIC: 984.95
```

```
##
## Number of Fisher Scoring iterations: 5
```

A summary of the reduced logistic regression model is shown above, where we are left with a 11 attribute/parameter model. The attributes used in the reduced logistic regression model are the Status of existing checking account, loan duration, credit history, purpose of the loan, credit amount, savings account, installment rate in percentage of disposable income, personal status and sex, other debtors/guarantors, other installment plans, and whether the individual is a foreign worker or not. Note that even though some of the specific factor levels of the attributes included may not be statistically significant at the 5% level, however at least one factor level of each attribute is statistically significant. Trying to remove the other insignificant factor levels from the attributes would result in removing or changing the attribute entirely, hence why the statistically insignificant factor levels were included in the reduced logistic regression model.

Reduced Logistic Regression model with 5-fold CV

We can now test this reduced logistic regression model using the 5-fold CV explained previously in the methodology section and calculate the average AUC for the model trained and tested based on the 5 validation and training sets determined by the stratified folds. And we will evaluate sensitivity, specificity, and misclassification rate of the model using the optimal p-threshold determined by the Youden Index.

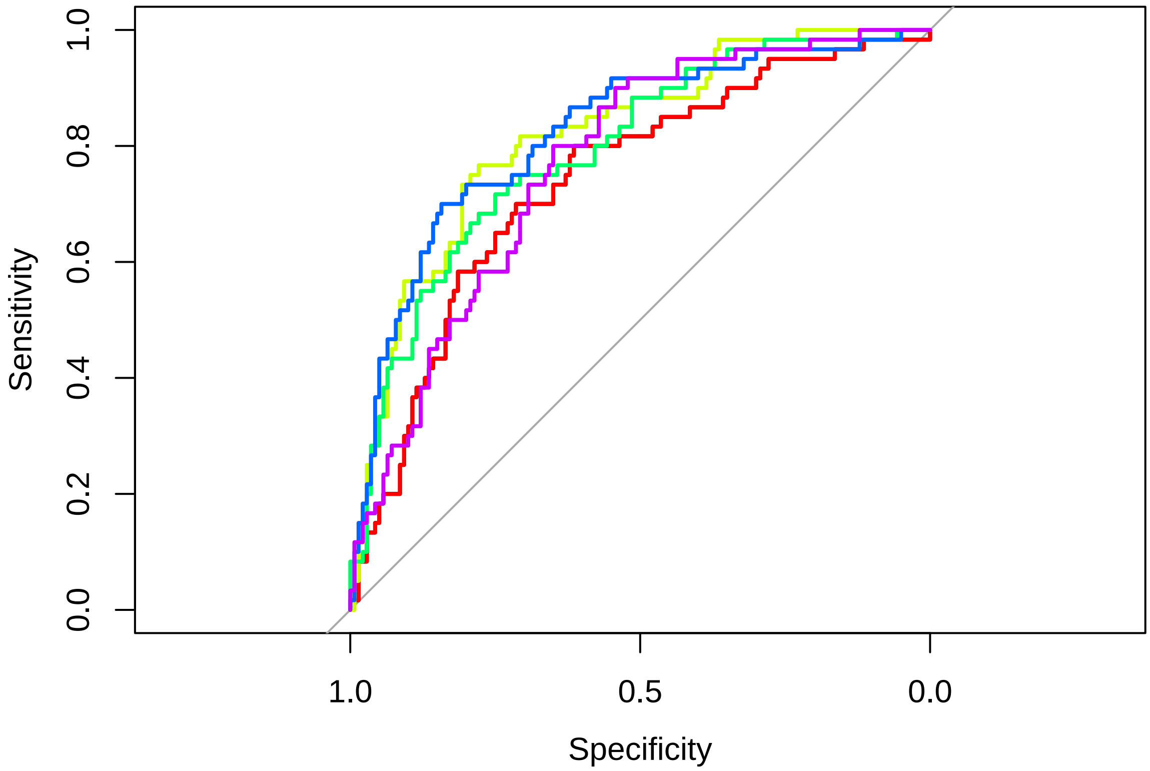

Figure 4: ROC Curves for 5-Fold Cross-Validated Logistic Regression Model.

Table 4: Average Confusion Matrix for Logistic Regression Model with 5-Fold Cross-Validation.

	No	Yes
No	101.8	14.6
Yes	38.2	45.4

```
## Average Results for Logistic Regression Model with 5-Fold Cross-Validation
## Average AUC: 0.7885,  sd = 0.03478601
## Average Sensitivity: 0.7566667,  sd = 0.04654747
## Average Specificity: 0.7271429,  sd = 0.09386812
## Average Misclassification Rate: 0.264,  sd = 0.05401389
```

The results of the logistic regression model showed an AUC of about 0.7885 with an estimated test error rate of 0.264. The average Specificity using the optimal threshold value for each fold of the 5-fold CV was estimated to be about 0.7271, while the Sensitivity using the same thresholds was estimated to be about 0.7567. This means that on average using the optimal threshold values per fold, the model correctly classified the Non-Default class around 72.71% of the time, while the model correctly classified the Default class around 75.67% of the time. The estimated test error rates of 0.264 and an AUC of 0.7885 indicates that the model has a reasonably good ability to discriminate between defaulters and non-defaulters, though there remains room for improvement. The balanced sensitivity and specificity also suggest the model performs well on both classes, its overall performance is acceptable but not outstanding, falling within the commonly accepted AUC range of 0.7–0.8 for credit risk modeling.

Linear Discriminant Analysis (LDA)

In LDA a linear combination of the predictors is found that best separates the two classes by projecting the data onto a common axis that distinguishes the classes more effectively. This is done through maximizing the distance between the means of different classes while minimizing the variance within each class. In LDA, it is assumed that each class shares the same covariance matrix resulting in linear decision boundaries when distinguishing between the two classes. The results of the LDA model for the German credit dataset is shown below.

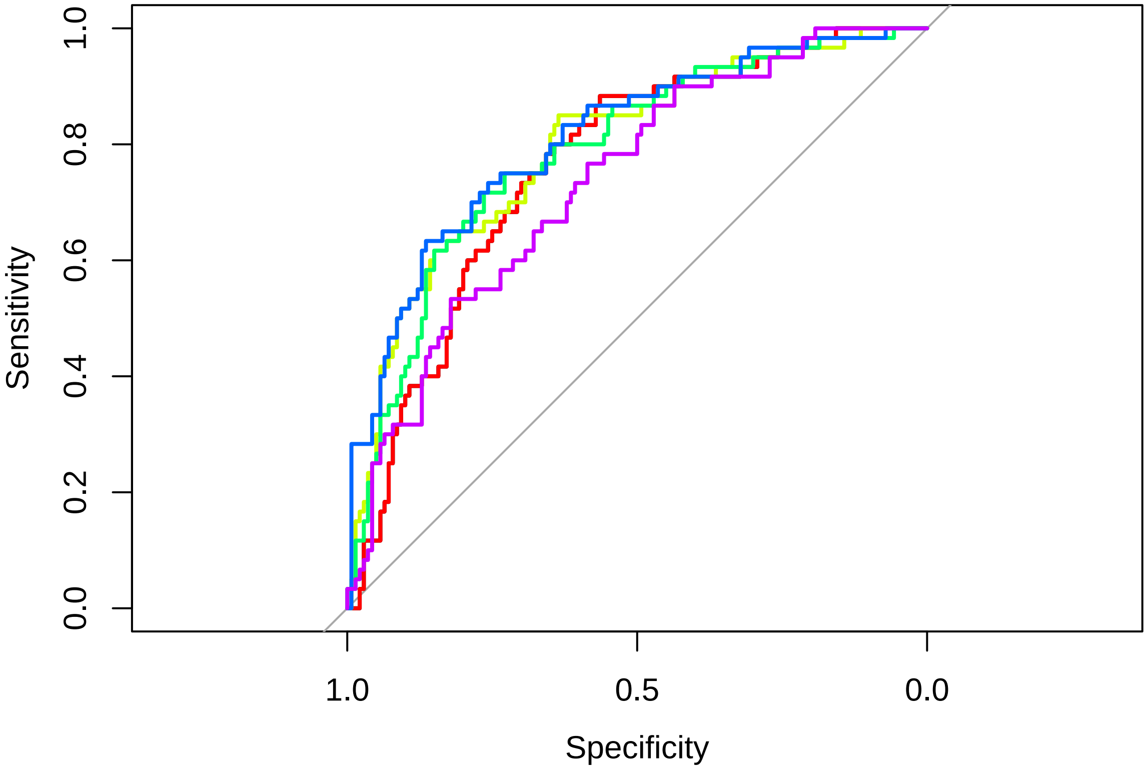

Figure 5: ROC Curves for 5-Fold Cross-Validated LDA Model.

Table 5: Average Confusion Matrix for LDA Model with 5-Fold Cross-Validation.

	No	Yes
No	102.2	16.6
Yes	37.8	43.4

```
## Average Results for LDA Model with 5-Fold Cross-Validation
## Average AUC: 0.7778333,  sd = 0.02959146
## Average Sensitivity: 0.7233333,  sd = 0.146534
## Average Specificity: 0.73,  sd = 0.1264104
## Average Misclassification Rate: 0.272,  sd = 0.05106369
```

The LDA model yielded an AUC of approximately 0.7778 and an average misclassification rate of 0.272. Based on the optimal threshold values identified by the Youden Index across each fold of the 5-fold cross-validation, the model achieved an average Specificity of about 73.00% and an average Sensitivity of roughly 72.30%. This indicates that on average, the model was able to correctly identify Non-Default cases 73.00% of the time and Default cases 72.30% of the time. With a misclassification rate of 0.272 and an AUC close to 0.78, the model demonstrates a solid capacity to differentiate between the two classes, though it does not reach exceptional performance. The relatively even sensitivity and specificity values further suggest balanced predictive ability for both defaulters and non-defaulters. Overall, the model performs reliably and falls within the generally accepted AUC range of 0.7 to 0.8 for credit risk classification tasks.

Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) fits quadratic decision boundaries by allowing for each class to have its own covariance matrix. This leads to the decision boundaries in QDA to be quadratic in nature when distinguishing between the two classes. So, QDA is ideal if one expects the variability in the different classes to come from different distributions, then QDA could be better suited.

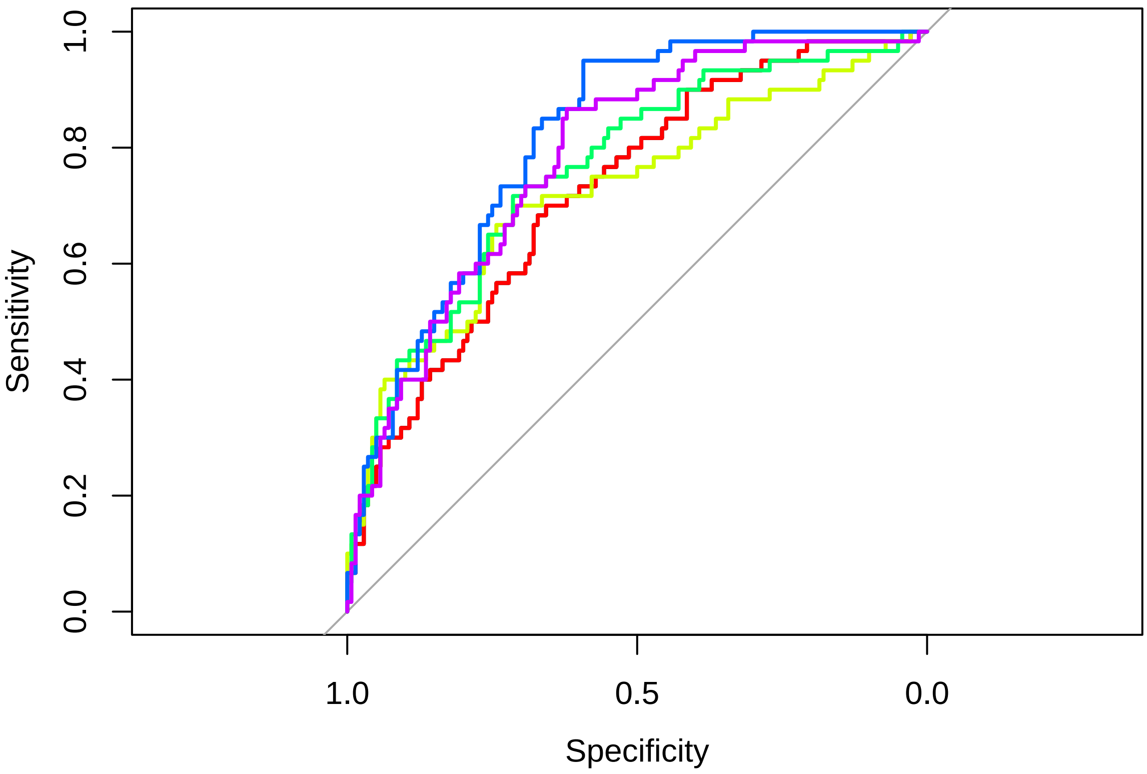

Figure 6: ROC Curves for 5-Fold Cross-Validated QDA Model.

Table 6: Average Confusion Matrix for QDA Model with 5-Fold Cross-Validation.

	No	Yes
No	92.4	12.8
Yes	47.6	47.2

```
## Average Results for QDA Model with 5-Fold Cross-Validation
## Average AUC: 0.762881,  sd = 0.03864167
## Average Sensitivity: 0.7866667,  sd = 0.1151086
## Average Specificity: 0.66,  sd = 0.05453888
## Average Misclassification Rate: 0.302,  sd = 0.01753568
```

The QDA model produced an AUC of approximately 0.7629 and an average misclassification rate of 0.302. Using the optimal probability thresholds determined via the Youden Index for each fold, the model achieved an average Sensitivity of about 78.70% and a Specificity of around 66.00%. This implies that on average, the model was more effective at correctly identifying Default cases than Non-Default cases. While the AUC of 0.7629 places the model within the acceptable range for credit risk classification, the relatively higher misclassification rate and imbalanced performance between sensitivity and specificity suggest that the model may over-prioritize detecting defaulters at the expense of correctly classifying non-defaulters. Overall, the QDA model shows reasonable discriminative ability, but its performance is somewhat less balanced and slightly weaker compared to the LDA and logistic regression models.

K-nearest neighbors (KNN)

KNN is a non-parametric classifier that identifies the k-nearest points in the training set that are closest to the test observation and classifies that observation to the class with the largest probability. To determine the optimal K for the KNN model we will estimate the test error rate for K-values ranging from 1 to 40 and choose the K that minimizes the test error rate. Cross validation was used to determine the optimal K, where the K value that produces the highest accuracy will be determined as the optimal K for the 5-fold CV KNN model. Then, the 5-fold CV KNN model used the p-threshold value determined by the Youden Index for each fold and the results for the average of the 5-folds is shown below.

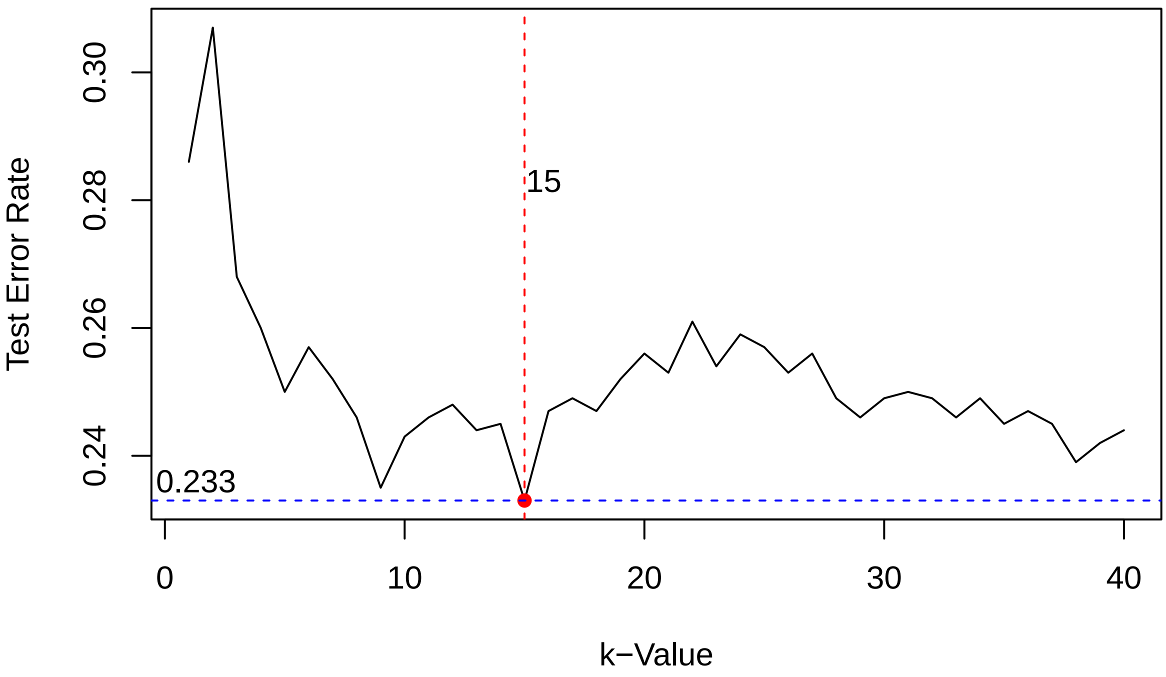

Figure 7: Test Error/Misclassification Rate for k-Values Ranging from 1 to 40 for the KNN Model.

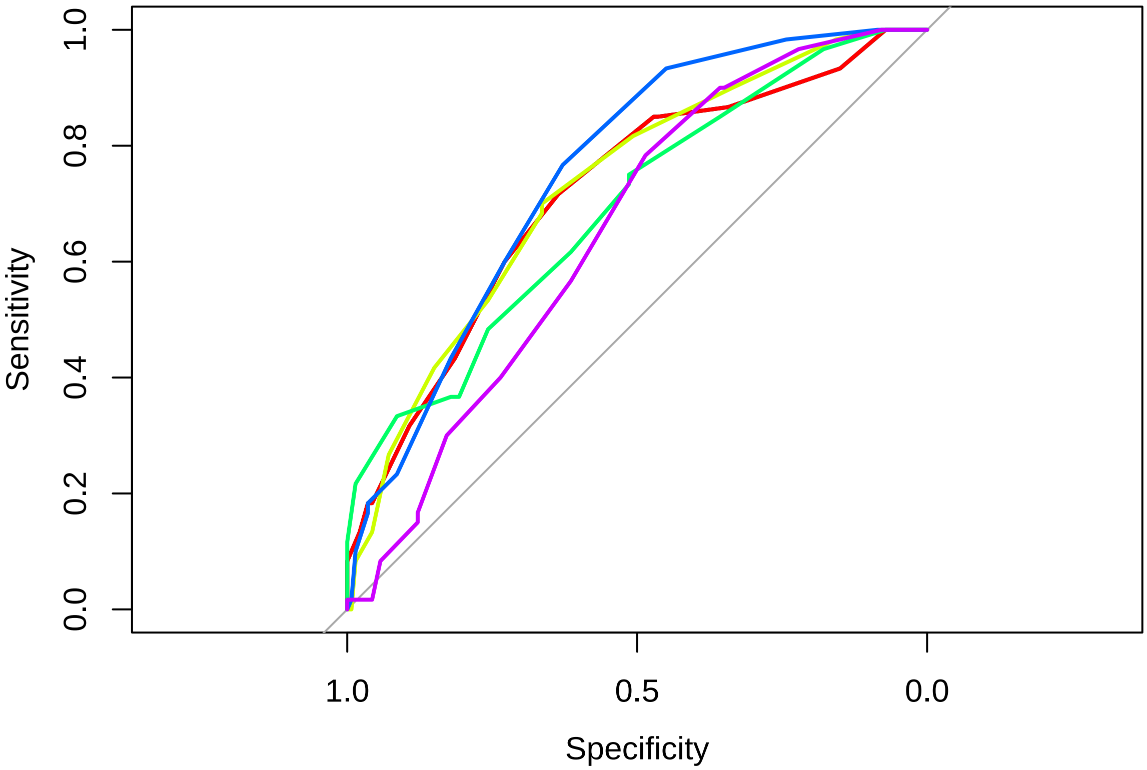

Figure 8: ROC Curves for 5-Fold Cross-Validated KNN Model with optimal k-Value of 15.

Table 8: Average Confusion Matrix for the k = 15 KNN Model with 5-Fold Cross-Validation.

	No	Yes
No	82	15.4
Yes	58	44.6

```
## Average Results for KNN Model with k = 15 and 5-Fold Cross-Validation
##  Average AUC:  0.7082381 ,  sd =  0.03820653
##  Average Sensitivity:  0.7433333 ,  sd =  0.03456074
##  Average Specificity:  0.5857143 ,  sd =  0.08001913
##  Average Misclassification Rate:  0.367 ,  sd =  0.04881086
```

Looking at the results of the KNN model the AUC of was determined to be about 0.7082 using an a K value of 15. The average misclassification rate for the KNN model using K = 15 was estimated to be around 0.367. The probability of predicting the Non-Default class is given by the specificity of about 0.5857 and the probability of predicting the Default class is given by the sensitivity of about 0.7433. Even with the AUC of over 0.7 the model has a high misclassification rate, very poor specificity and the sensitivity is within the range of the other models fitted so far.

Decision Tree

For the Decision Tree model, performance was evaluated using 5-fold cross-validation to estimate test error rates and assess classification performance. Prior to cross-validation, the tree was fit to the full training data and pruned using cross-validation to identify the optimal tree size — determined by the complexity parameter that minimized the number of misclassifications. This optimal tree size was then used consistently across all five folds. For each validation set, class predictions were made using the probability estimates from the pruned tree, and the optimal probability threshold for classification was selected using the Youden Index. This approach ensures that the threshold used to convert predicted probabilities into class labels maximizes the model's combined sensitivity and specificity.

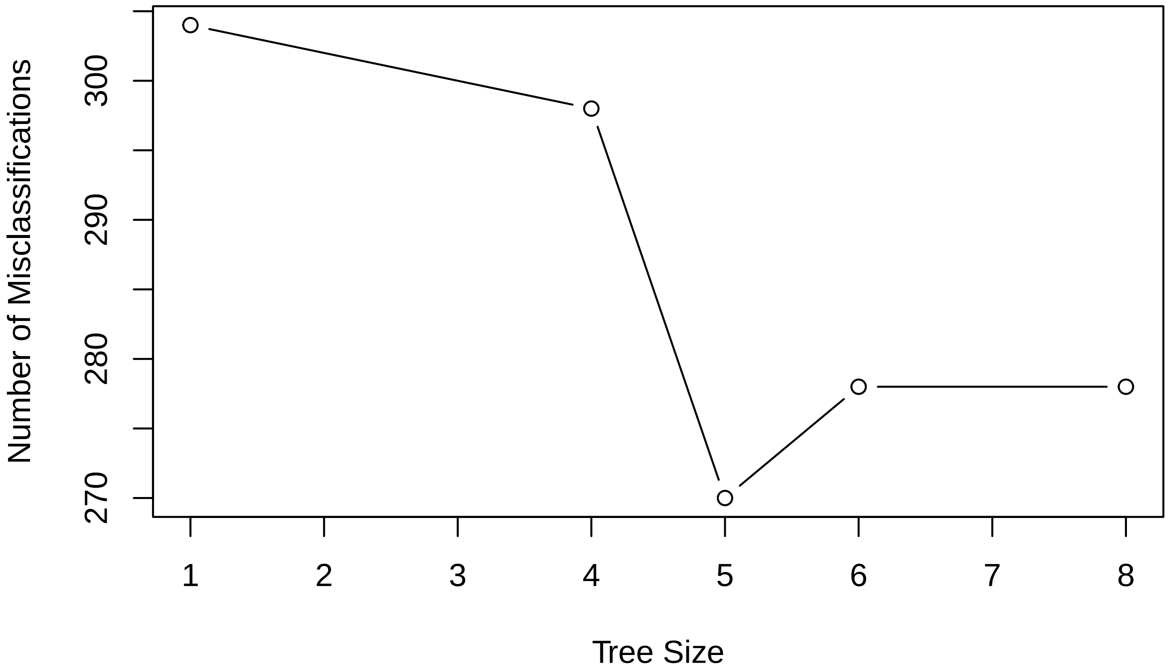

Figure 9: The Number of Misclassifications for a Decision Tree Model for Tree Sizes ranging from 1 to 8.

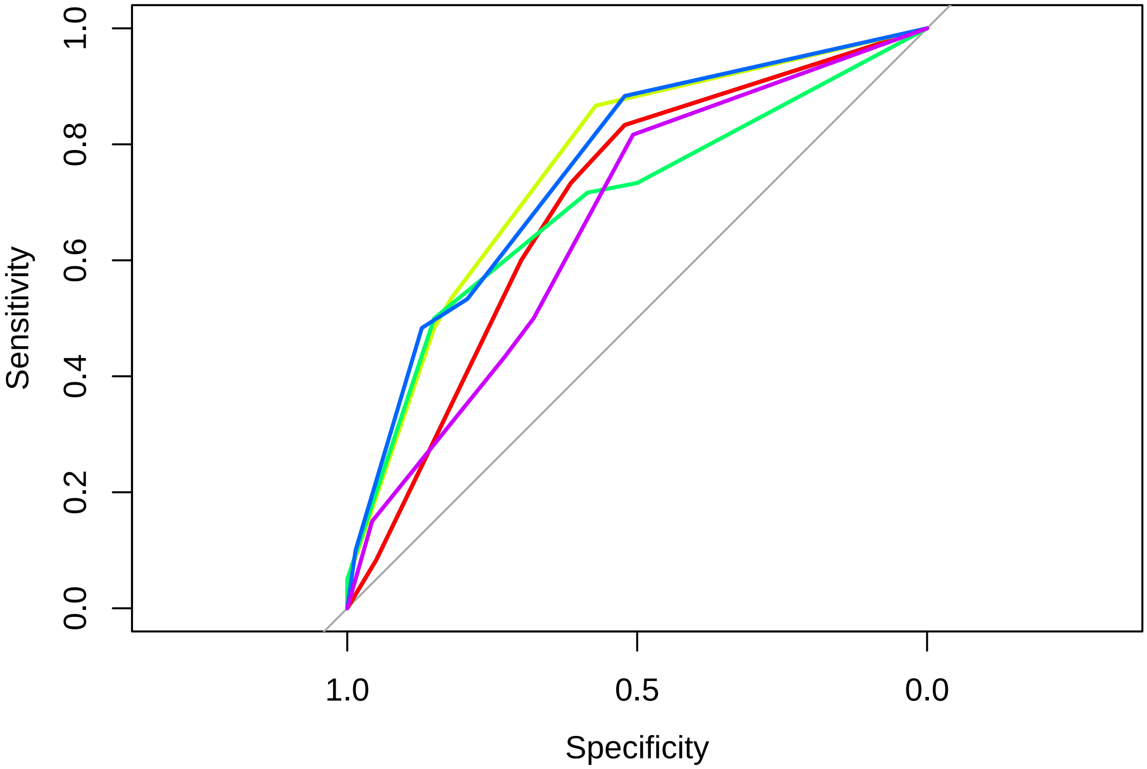

Figure 10: ROC Curves for 5-Fold Cross-Validated Decision Tree with a Pruned Tree Size of 5.

Table 9: Average Confusion Matrix for Pruned Decision Tree Model of Tree Size 5 with 5-Fold Cross-Validation.

	No	Yes
No	83.2	13.2
Yes	56.8	46.8

```
## Average Results for Decision Tree Model with size of 5 and 5-Fold Cross-Validation
##  Average AUC:  0.7158452 ,  sd =  0.04198185
##  Average Sensitivity:  0.78 ,  sd =  0.1587276
##  Average Specificity:  0.5942857 ,  sd =  0.1450194
##  Average Misclassification Rate:  0.35 ,  sd =  0.05755432
```

From the decision tree model a tree size of 5 was determined to be the tree size that minimizes the number of misclassifications for the German credit data set. Using a tree size of 5 for each fold in the 5-fold CV, the decision tree model produced an AUC of about 0.7158 with an average misclassification rate of 0.35. The decision tree model produced an average sensitivity of 0.81 and an average specificity of about 0.6914 using the optimal p-thresholds determined by the Youden Index. The performance of the model suggests that while it has a strong ability to correctly identify defaulters (as indicated by the high sensitivity), its overall discriminative power is somewhat limited compared to other models, as reflected in the relatively lower AUC and higher misclassification rate.

Random Forest

For the Random Forest model, classification performance was evaluated using 5-fold cross-validation. Within each fold, a forest of 500 trees was trained on the corresponding training set. Class probability estimates were generated for the validation set using the trained forest, and the optimal classification threshold was determined by maximizing the Youden Index on the ROC curve. This threshold was then used to convert probabilities into binary class predictions for that fold. Performance metrics, including sensitivity, specificity, AUC, and misclassification rate, were computed at the optimal threshold. Average confusion matrices and summary statistics were aggregated across all folds to assess model stability and generalization. This approach ensures that threshold selection is tailored to the model's calibration in each fold, and that performance is fairly estimated using out-of-sample predictions.

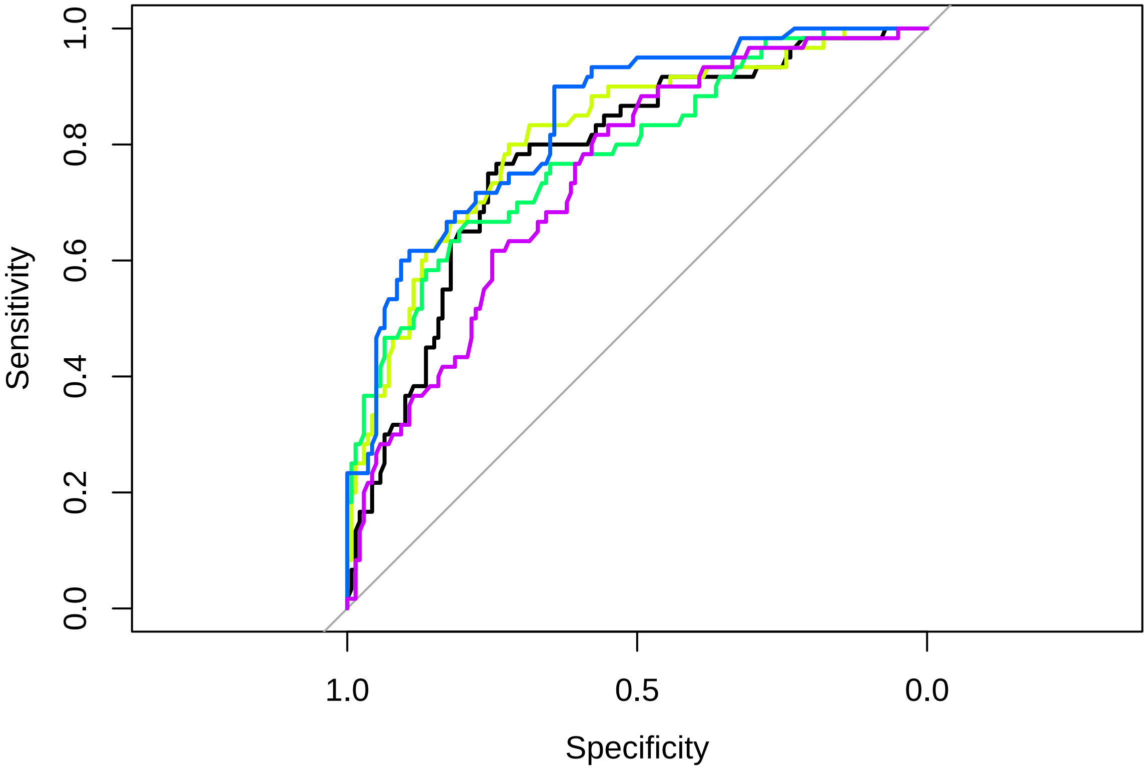

Figure 11: ROC Curves for 5-Fold Cross-Validated Random Forest Model.

Table 10: Average Confusion Matrix for Random Forest Model with 5-Fold Cross-Validation.

	No	Yes
No	97.2	12.6
Yes	42.8	47.4

```
## Average Results for Random Forest Model with 5-Fold Cross-Validation
## Average AUC: 0.7935238,  sd = 0.03646234
## Average Sensitivity: 0.79,  sd = 0.08465617
## Average Specificity: 0.6942857,  sd = 0.08739425
## Average Misclassification Rate: 0.277,  sd = 0.04563442
```

The results of the Random Forest model showed an average AUC of approximately 0.7935 with an average misclassification rate of 0.277. Using the optimal p-thresholds determined by the Youden Index across the 5-fold cross-validation, the model achieved an average sensitivity of 0.79 and a specificity of about 0.6943. These results indicate that the model performs well in identifying defaulters, while maintaining a reasonable ability to correctly classify non-defaulters. With an AUC nearing 0.8 and balanced performance across classes, the Random Forest model demonstrates strong overall predictive power and robustness, outperforming several of the simpler models in terms of discrimination ability.

XGBoost

For the XGBoost model, classification performance was assessed using 5-fold cross-validation. In each fold, the training set was used to fit a gradient-boosted tree model with 100 boosting rounds and a logistic objective. Factor variables were converted to numeric codes for compatibility with the XGBoost framework. After generating predicted probabilities on the validation set, the optimal classification threshold was determined by maximizing the Youden Index from the ROC curve. Binary class predictions were then made using this threshold, and evaluation metrics including AUC, sensitivity, specificity, and misclassification rate — were calculated for each fold. Averaging these results provided a stable estimate of model performance across all validation folds.

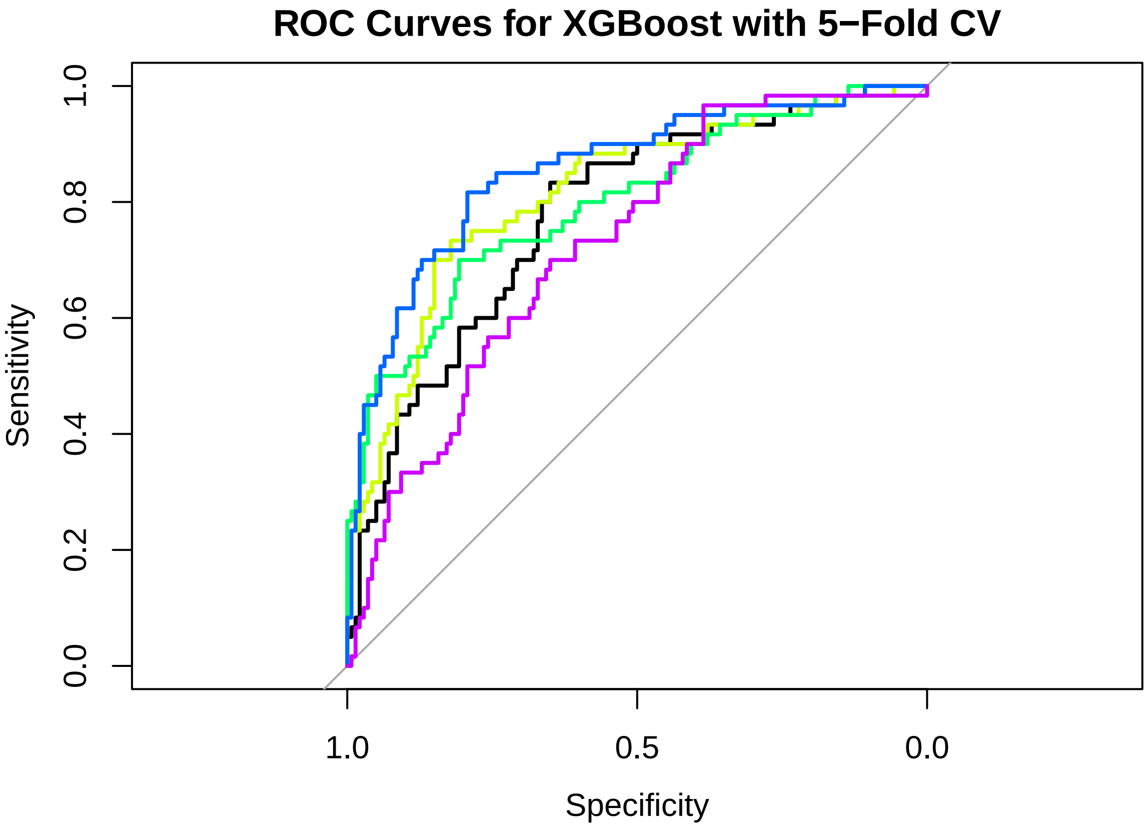

Figure 12: ROC Curves for XGBoost Model using 5-Fold Cross-Validation.

Table 11: Average Confusion Matrix for XGBoost Model with 5-Fold Cross-Validation.

	No	Yes
No	96.8	11.4
Yes	43.2	48.6

```
## Average Results for XGBoost Model with 5-Fold Cross-Validation
## Average AUC: 0.7952619,  sd = 0.04793366
## Average Sensitivity: 0.81,  sd = 0.1038161
## Average Specificity: 0.6914286,  sd = 0.1842248
## Average Misclassification Rate: 0.273,  sd = 0.1007844
```

The results of the XGBoost model indicate strong predictive performance, with an average AUC of approximately 0.7953 and an average misclassification rate of 0.273. Across the 5-fold cross-validation, the model achieved an average sensitivity of 0.81 and an average specificity of about 0.6914, using the optimal p-thresholds determined by the Youden Index. These results suggest that the XGBoost model effectively distinguishes between classes, with a particularly strong ability to correctly identify positive cases. The combination of high sensitivity and competitive AUC highlights the model's robustness, making it one of the best-performing models evaluated in this analysis.

Comparison of Models/Conclusion

Among the models evaluated, the best-performing classifiers include logistic regression, LDA, QDA, Random Forest, and XGBoost. While each model demonstrated acceptable predictive ability for credit default classification—falling within the commonly accepted AUC range of 0.7–0.8—they differ in performance characteristics that make them more suitable for different business or banking contexts.

XGBoost and Random Forest models offered the strongest overall predictive performance, with AUCs of approximately 0.7953 and 0.7935, respectively. XGBoost slightly outperformed other models in terms of both sensitivity (0.81) and test error rate (0.273), making it a particularly attractive option when the primary goal is to minimize missed defaults—a common priority in high-risk lending environments or when identifying potentially costly defaults is critical. Random Forest, with comparable metrics, provides a balance between model complexity and interpretability and may be preferred when slightly more transparency is needed in decision-making.

Logistic Regression and LDA performed similarly, with AUCs of approximately 0.7885 and 0.7778, respectively. Both models achieved balanced sensitivity and specificity, making them suitable for low-to-moderate risk portfolios or contexts where model simplicity, interpretability, and regulatory transparency are important. These models are also easier to implement and explain to stakeholders, which may be necessary in consumer credit settings or smaller financial institutions with compliance constraints. QDA, while still within the acceptable performance range (AUC 0.7629), showed a higher test error rate (0.302) and imbalance between sensitivity and specificity, suggesting it leans more toward detecting defaulters at the cost of misclassifying non-defaulters. This behavior might be desirable in risk-averse contexts where catching as many defaulters as possible is prioritized, but it may also result in more false positives, which could unnecessarily restrict credit access for reliable customers.

Ultimately, model selection should align with the operational/business priorities and risk tolerance of business/banks using these type of machine learning models on credit data.