Risk Prediction Model to Enhance Loan Approval Strategies

Indrajith Wasala Mudiyanselage

In this project, the primary goal was to build a credit risk assessment model using the German Credit Dataset. The dataset contains customer and loan-related features, and the objective was to predict whether an applicant is likely to be a good credit risk (repay the loan) or a bad credit risk (default). This model is designed to help financial institutions make informed decisions about loan approvals and manage their lending risks effectively. The German credit dataset has 1000 observations on 21 variables. Response variable is (1, 0) someone is likely to repay their loan (a good credit risk) or not (a bad credit risk). Predictor variables includes both categorical and numerical variables such as Status of existing checking account, Credit history, Credit amount, Present employment, Age, etc.

1 Data Cleaning and Prepossessing

1. Missing Value Handling : We begin by checking and handling missing values. Numerical features with missing values are imputed using the mean or median, depending on the distribution of the data. Categorical features with missing values are replaced with the mode (most frequent value) or a new category, like "Unknown".

2. Normalization : To scale the features to a common range, we use min-max normalization, which transforms numerical features to a scale between 0 and 1. This ensures that features with large ranges do not dominate the modeling process.

3. Log Transformation for Skewed Data: Skewed numerical features are transformed using logarithmic transformations to reduce skewness, which helps improve the performance of machine learning models like regression or tree-based algorithms.

4. Rebinning for Ordinal Features: For categorical features that have too many levels (e.g., "Checking Account Status"), we may group them into fewer, more meaningful bins. This process, called rebinning, reduces noise and improves model performance.

2 Perform Exploratory analysis and Feature Selection

1. Figure 1 show the class conditional distributions of the predictor variables. If for a predictor, the distributions across the two classes of Default differ, that indicates potential association between the predictor and response. For quantitative variables we have shown boxplots and for qualitative variables we have plotted barplots.

   From the plots, almost all the variables seem associated with response. For a more quantitative analysis, we can conduct chi-square tests of association between Default and the qualitative predictors and fit univariate logistic model between each quantitative variable and Default. From this, we can infer that *checkingstatus1*, *duration*, *history*, *purpose*, , *amount*, *savings*, *employ*, *status*, *others*, *property*, *age*, *otherplans*, *housing*, *foreign* are important in prediction of the response variable *Default*.

2. We perform backward subset selection using AIC criteria to build a reasonably good model. Lower AIC indicates a better model. First we fit the full model with all the predictor variables and its AIC is 997.4. To perform backward elimination using AIC, variables with the highest p-values (indicating they are not statistically significant) are iteratively removed from the model. After each variable removal, the AIC of the new model is recalculated. The process continues until minimum AIC value is obtained. Final model includes variables *checkingstatus1*, *duration*, *history*, *purpose*, *amount*, *savings*, *installment*, *status*, *others*, *residence*, *otherplans*, *housing*, *tele* and *foreign* with lowest AIC = 984.8.

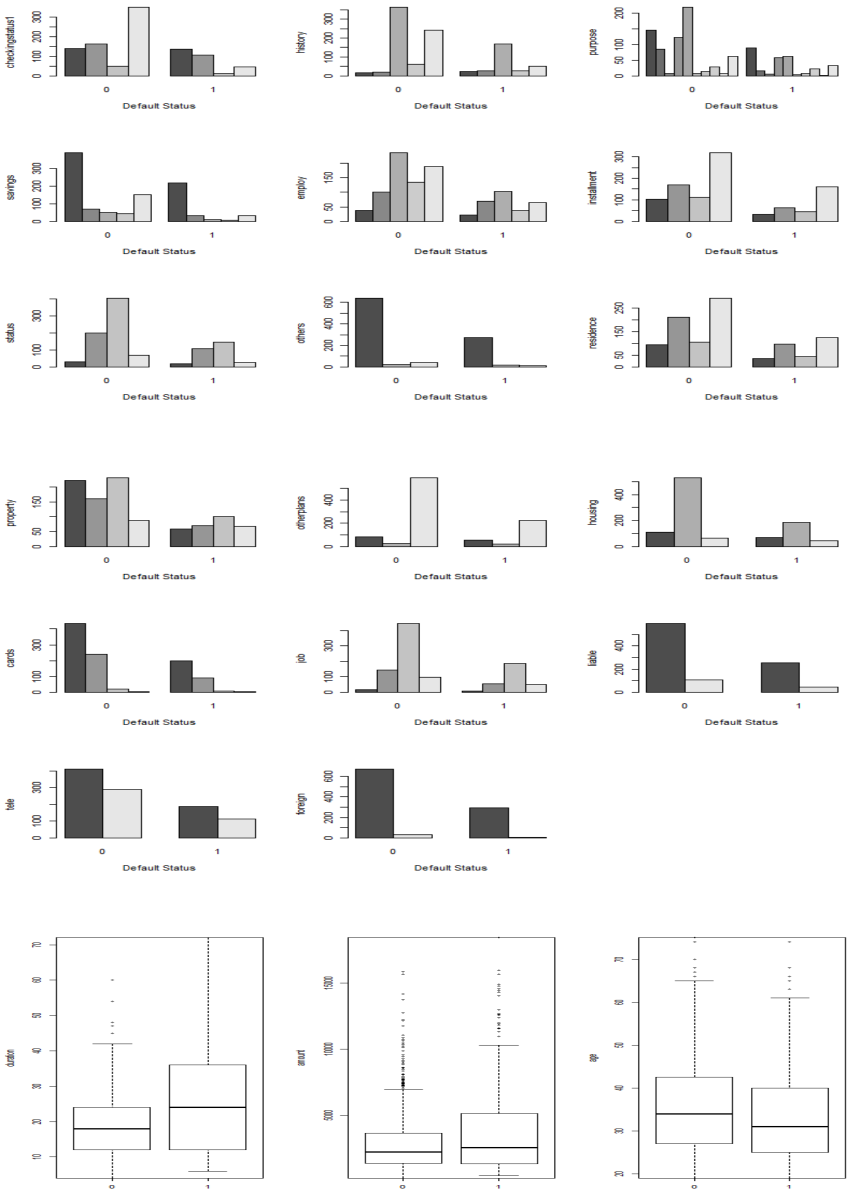

Figure 1: Bar plots and Box plots showing relationship between Default and predictors.

3 Logistic regression model for best subset

Table 1 Represent Coefficients and p values. The coefficient of the variable duration is 0.028. This means an unit increase in the value of the predictor duration, will make $e^{0.0275} = 1.0278$ times increase in the odds of one's chance of not being a defaulter. Similarly, someone having foreign A202 will have $e^{-1.4630} = 0.232$ times chance of being defaulter than someone with foreign A201. The training error obtained from the model is 0.218.

	Estimate	Pr(> $|z|$)
(Intercept)	1.3689	0.0503
checkingstatus1A12	-0.3853	0.0713
checkingstatus1A13	-1.0445	0.0041
checkingstatus1A14	-1.7776	0.0000
duration	0.0275	0.0023
historyA31	-0.1369	0.7964
historyA32	-0.8646	0.0353
historyA33	-1.0037	0.0319
historyA34	-1.5656	0.0003
purposeA41	-1.5917	0.0000
purposeA410	-1.3763	0.0759
purposeA42	-0.6695	0.0082
purposeA43	-0.8831	0.0003
purposeA44	-0.5267	0.4851
purposeA45	-0.1263	0.8153
purposeA46	0.2067	0.5951
purposeA48	-2.0474	0.0888
purposeA49	-0.7295	0.0281
amount	0.0001	0.0044
savingsA62	-0.3145	0.2620
savingsA63	-0.4431	0.2557
savingsA64	-1.4077	0.0068
savingsA65	-1.0109	0.0001
installment2	0.1647	0.5857
installment3	0.5725	0.0861
installment4	0.8885	0.0025
statusA92	-0.2195	0.5584
statusA93	-0.8258	0.0244
statusA94	-0.3557	0.4221
othersA102	0.5062	0.2070
othersA103	-1.0734	0.0105
residence2	0.7240	0.0097
residence3	0.4060	0.2124
residence4	0.2936	0.2958
otherplansA142	-0.0580	0.8866
otherplansA143	-0.6777	0.0040
housingA152	-0.5108	0.0246
housingA153	-0.2444	0.4578
teleA192	-0.3089	0.0937
foreignA202	-1.4630	0.0199

Table 1: coefficient of logistic regression

4 Logistic regression model for full model by calculating error rate using LOOCV

1. The German credit dataset has 1000 observations on 21 variables. Default is the response and the rest are predictors. We fit a logistic regression model to the data with all predictors. The error rate, sensitivity, specificity and AUC presented in table 2. Moreover, figure 2 represents ROC curve.

2. Own code to estimate the test error rate. Please refer to part (b) in the R-code to find the written coed. The code produces a test error rate of 0.247.

3. For this part codes from *caret* package were used. The function produces exactly same test error rate of 0.247.

5 Linear Discriminant Analysis model

The error rate, sensitivity, specificity and AUC presented in table 2. Moreover, figure 2 represents ROC curve. The fitted model under LOOCV produces a test error rate of 0.2470.

6 Quadratic Discriminant Analysis model

The error rate, sensitivity, specificity and AUC presented in table 2. Moreover, figure 2 represents ROC curve. The fitted model under LOOCV produces a test error rate of 0.2823.

7 KNN model

Optimal k chosen via LOOCV is 77. The error rate, sensitivity, specificity and AUC presented in table 2. Moreover, figure 2 represents ROC curve. The fitted model under LOOCV produces a test error rate of 0.2880.

8 Logistic regression model for the best subset by calculating error rate using LOOCV

The error rate, sensitivity, specificity and AUC presented in table 2. Moreover, figure 2 represents ROC curve. The fitted model under LOOCV produces a test error rate of 0.2460.

9 Model Comparison: Logistic, LDA, QDA, KNN and logistic with best subset

Comparing all the methods we would like to choose the model with low LOOCV error rate. Based on the results in 2 we can see that the logistic regression with reduced variables created in problem 1 yields the best results. LDA and full logistic model is also closely follows the best method.

Model	LOOCV error rate	Error rate	Sensitivity	Specificity	AUC
Logistic (Full)	0.2470	0.2110	0.5500	0.8943	0.8366
LDA	0.2470	0.2170	0.5467	0.8843	0.8349
QDA	0.2823	0.1600	0.7800	0.8657	0.9115
KNN	0.2880	0.2880	0.7143	0.7119	0.6319
Logistic (Reduce)	0.2460	0.2180	0.5367	0.8871	0.8314

Table 2: Performance of Logistic(Full), LDA, QDA , KNN , Logistic(Reduced) on test set

10 Ridge regression, Lasso regression

1. The estimated test rate for the logistic regression fit is 0.247.
2. The penalty parameter chosen is 0.02582 and the test error rate for ridge regression is 0.245.
3. The penalty parameter chosen is 0.006249 and lasso yields a test for lasso regression is 0.241.
4. Table 3 presents the coefficients and estimated test error rates for all models. The lowest test error corresponds to the model obtained using Ridge regression and can be recommended as it is most parsimonious among the tree models.

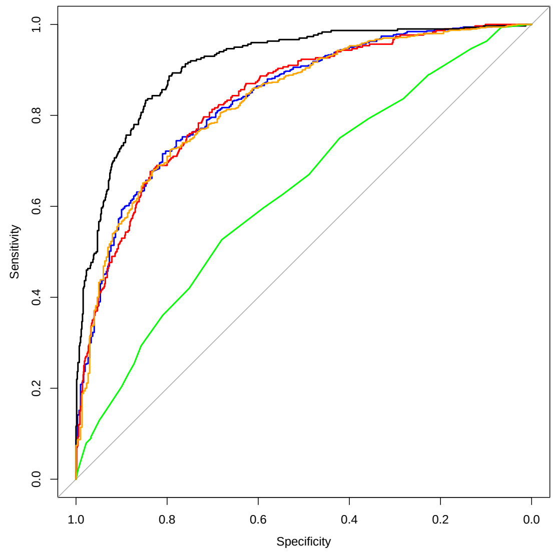

Figure 2: ROC curves for Logistic(Full)(Blue), LDA(Red), QDA(Black) , KNN(Green) , Logistic(Reduced)(Orange) on test set.

Coefficients	Logistic	Ridge	Lasso	Coefficients	Logistic	Ridge	Lasso
(Intercept)	0.8683	0.1588	0.2799	jobA172	0.4416	0.0234	0.0000
age	-0.0128	-0.0096	-0.0074	jobA173	0.4694	0.0729	0.0000
amount	0.0001	0.0001	0.0001	jobA174	0.3691	0.0582	0.0000
cards2	0.4050	0.2780	0.2642	liable2	0.2628	0.1740	0.0776
cards3	0.2741	0.0907	0.0000	otherplansA142	-0.0888	-0.0068	0.0000
cards4	0.4550	0.2866	0.0000	otherplansA143	-0.6475	-0.4871	-0.4690
checkingstatus1A12	-0.3834	-0.1954	-0.2727	othersA102	0.4329	0.3743	0.2388
checkingstatus1A13	-0.9739	-0.7198	-0.7805	othersA103	-0.9828	-0.7718	-0.7514
checkingstatus1A14	-1.7800	-1.3590	-1.5935	propertyA122	0.2698	0.1884	0.0000
duration	0.0280	0.0250	0.0262	propertyA123	0.1607	0.1360	0.0000
employA72	0.0666	0.2393	0.1943	propertyA124	0.7367	0.4292	0.1456
employA73	-0.2293	0.0185	0.0000	purposeA41	-1.6605	-1.0899	-1.1110
employA74	-0.7634	-0.4315	-0.3947	purposeA410	-1.4854	-0.8824	-0.6603
employA75	-0.2213	-0.0407	0.0000	purposeA42	-0.7481	-0.3738	-0.2781
foreignA202	-1.4614	-0.9971	-0.9062	purposeA43	-0.8743	-0.5579	-0.5427
historyA31	0.1690	0.5902	0.5081	purposeA44	-0.5109	-0.2026	0.0000
historyA32	-0.5672	-0.0523	0.0000	purposeA45	-0.1603	0.0922	0.0000
historyA33	-0.9496	-0.3289	-0.1992	purposeA46	0.1130	0.2420	0.1991
historyA34	-1.4957	-0.7464	-0.6900	purposeA48	-1.9309	-1.1937	-0.8421
housingA152	-0.4573	-0.3630	-0.3267	purposeA49	-0.6888	-0.2902	-0.2117
housingA153	-0.6303	-0.2172	0.0000	residence2	0.7613	0.4060	0.2862
installment2	0.2641	-0.0146	0.0000	residence3	0.5246	0.2418	0.0376
installment3	0.6260	0.2429	0.1467	residence4	0.3885	0.1322	0.0000
installment4	0.9369	0.5010	0.4813	savingsA62	-0.3638	-0.2201	-0.0517
statusA92	-0.2616	0.0467	0.0309	savingsA63	-0.3664	-0.4059	-0.2139
statusA93	-0.8427	-0.3886	-0.3490	savingsA64	-1.4604	-1.0705	-0.9735
statusA94	-0.3764	-0.0984	0.0000	savingsA65	-0.9732	-0.7542	-0.7286
teleA192	-0.2848	-0.2217	-0.1460				

Table 3: Summary of coefficients