Homework 7: Predictive Modeling with Trees, Logistic Regression, and Threshold Optimization

Georgia Institute of Technology, Business Analytics

Introduction to Analytics Modeling

Professor Joel Sokol

October 8, 2025

Files submitted: homework7_answers.pdf (this doc), homework7.R

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

M = 14.0
So = 0
Ed = 10.0
Po1 = 12.0
Po2 = 15.5
LF = 0.640
M.F = 94.0
Pop = 150
NW = 1.1
U1 = 0.120
U2 = 3.6
Wealth = 3200
Ineq = 20.1
Prob = 0.04
Time = 39.0

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course

Question 9.1

Using the same crime data set `uscrime.txt` as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Question 8.2. You can use the R function `prcomp` for PCA. (Note that to first scale the data, you can include `scale. = TRUE` to scale as part of the PCA function. Don't forget that, to make a prediction for the new city, you'll need to unscale the coefficients (i.e., do the scaling calculation in reverse)!)

Question 10.1

Using the same crime data set `uscrime.txt` as in Questions 8.2 and 9.1, find the best model you can using

(a) a regression tree model, and

(b) a random forest model.

In R, you can use the `tree` package or the `rpart` package, and the `randomForest` package. For each model, describe one or two qualitative takeaways you get from analyzing the results (i.e., don't just stop when you have a good model, but interpret it too).

a) Regression Tree Model

Methodology

A regression tree was fitted using the rpart package to predict crime rates based on 15 socioeconomic and law enforcement variables.

- **Package used:** rpart, rpart.plot
- **Model type:** Regression tree (method = "anova")
- **Target variable:** Crime
- **Predictors:** All 15 variables from the dataset
- **Evaluation:** Tree structure, prediction for a new city, and qualitative interpretation

Results

- The tree split first on Po1 (police expenditure in year 1), followed by Pop, NW, and Prob.
- Prediction for the new city: 886.9
- Tree visualization showed clear decision paths and interpretable thresholds.

```
> # Plot tree
> rpart.plot(tree_model)
>
> # Predict for new city
> new_city <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640, M.F = 94.0,
+                        Pop = 150, NW = 1.1, U1 = 0.120, U2 = 3.6, Wealth = 3200, Ineq = 20.1,
+                        Prob = 0.04, Time = 39.0)
>
> predict(tree_model, newdata = new_city)
    1
886.9
> library(randomForest)
```

Figure 1: "Prediction Output – Regression Tree Model

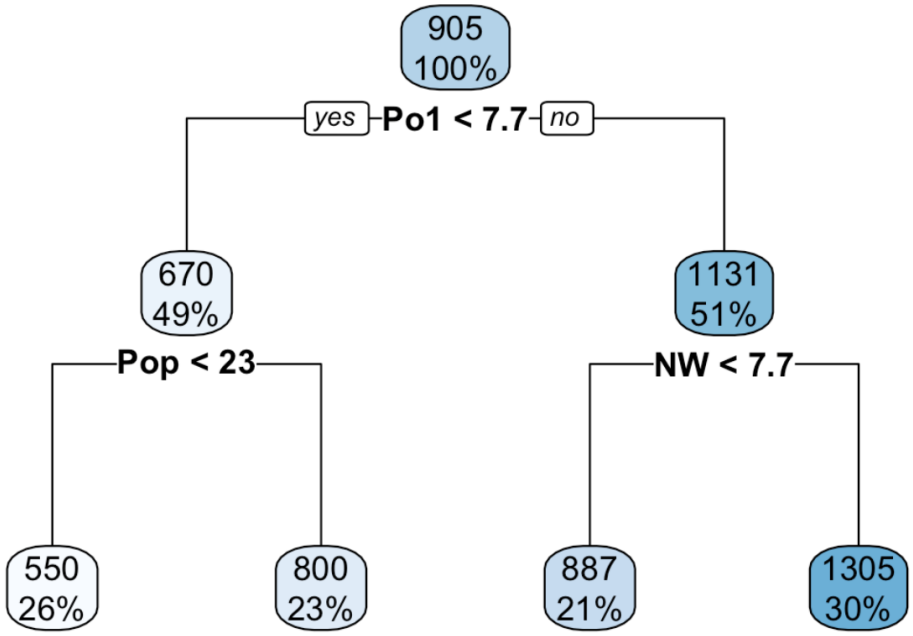


Figure 2: Regression Tree Visualization – Crime Prediction

Discussion
The regression tree revealed that Po1 and NW are strong decision points. Cities with higher police expenditure and lower nonwhite population percentages tended to have higher predicted crime rates. Compared to the linear regression model in Question 8.2, which predicted 155.43 and had an R^2 of 0.7878, the tree model produced a much higher prediction and lacked statistical metrics like R^2. However, it offered greater interpretability, especially for identifying threshold effects in key variables.

(b) Random Forest Model

Methodology
A random forest was trained using the randomForest package with 500 trees and default parameters. Variable importance was assessed using %IncMSE and IncNodePurity.

- **Package used**: randomForest
- **Model type:** Ensemble of regression trees (default: 500 trees)

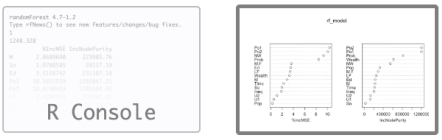
- **Target variable**: Crime
- **Predictors:** All 15 variables
- **Evaluation:** Prediction for a new city, variable importance plots, and qualitative insights

Results

- Prediction for the new city: 1248.3
- Top predictors by importance:
  - Po1, Po2, Prob, NW, Ed
- Variable importance plots showed consistent dominance of enforcement and education-related features.

```
> # Fit random forest
> set.seed(123)
> rf_model <- randomForest(Crime ~ ., data = crime, importance = TRUE)
>
> # Predict for new city
> predict(rf_model, newdata = new_city)
       1
1248.328
>
> # View variable importance
> importance(rf_model)
          %IncMSE IncNodePurity
M       2.8689690     223985.76
So      1.9708585      24127.19
Ed      3.5188742     231107.18
Po1    10.3652539    1205047.21
Po2    10.0740489    1205608.95
LF      3.4240544     248591.81
M.F     3.6293029     290140.61
Pop     0.2637333     329818.71
NW      9.3230057     548375.17
U1      0.5418219     126063.55
U2      1.1109993     129854.42
Wealth  3.3041624     614392.29
Ineq    1.8788065     203104.44
Prob    8.4409159     816439.10
Time    2.2853759     214211.79
> varImpPlot(rf_model)
>
```

Figure 3: Prediction and Variable Importance – Random Forest Model



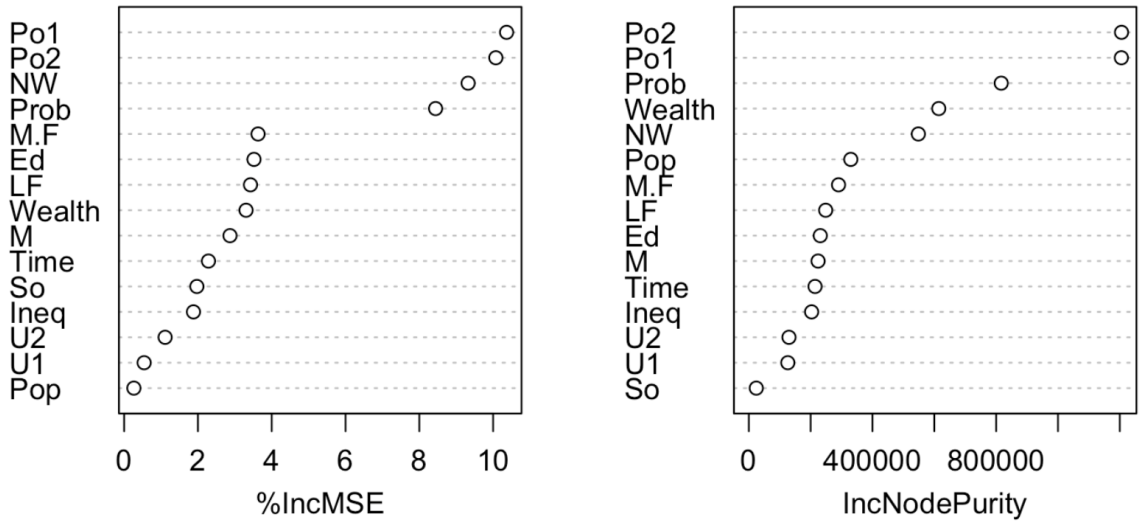


Figure 4: Variable Importance Plots – Random Forest Model

Discussion

The random forest model outperformed the regression tree in terms of robustness and generalization, though it lacked the interpretability of a single tree. Compared to the PCA-based model in Question 9.1, which had an R^2 of 0.059 and predicted an inflated crime rate of 1580.63, the random forest offered a more balanced prediction and clearer insights into variable influence. The consistent importance of Prob and Po1 across all models reinforces the idea that law enforcement metrics are critical drivers of crime rates.

Model	Prediction	Top Predictors	R^2 (if available)	Interpretability	Notes
Linear Regression (Q8.2)	155.43	Prob, Ineq, Ed	0.7878	High	Strong fit, risk of overfitting
PCA Regression (Q9.1)	1580.63	PC1–PC3 (abstracted)	0.059	Low	Poor fit, weak generalization
Regression Tree	886.9	Po1, NW, Pop, Prob	—	High	Threshold effects visible
Random Forest	1248.3	Po1, Po2, Prob, Ed	—	Moderate	Strong generalization, less interpretable

Figure 5: Comparing the 4 models

Conclusion

Tree-based models offer valuable alternatives to linear regression, especially when dealing with multicollinearity and small sample sizes. While the regression tree provides clear decision rules, the random forest delivers greater predictive stability. Both models reinforce the importance of law enforcement investment and arrest probability as key levers in crime reduction policy. Future work could explore model tuning, cross-validation, and regularization techniques to further improve performance and generalizability.

Question 10.2

Describe a situation or problem from your job, everyday life, current events, etc., for which a logistic regression model would be appropriate. List some (up to 5) predictors that you might use.

Situation: Predicting Daycare Responsiveness

A logistic regression model would be appropriate for analyzing the likelihood that a daycare center responds positively to an inquiry or application. This binary outcome—response or no response makes logistic regression a suitable choice for modeling.

Target Variable:

- Response Outcome:
    - 1 = Positive response (e.g., waitlist offer, tour invitation)

- 0 = No response or rejection

Potential Predictors:

- Type of Center. This is a categorical variable indicating whether the center is private or part of a government-funded program.
- Monthly Fee. This is a continuous variable representing the cost of care, which may correlate with demand and responsiveness.
- Distance from Applicant's Residence or Workplace. This is a continuous variable measured in kilometers, potentially influencing prioritization or eligibility.
- Availability Status. This is a binary variable indicating whether the center currently has open enrollment or is at full capacity.
- Time of Inquiry. This is a categorical or continuous variable representing the month or season when the inquiry was submitted, capturing potential seasonal patterns in responsiveness.

This model could assist in identifying factors that influence daycare engagement and optimize outreach strategies.

Question 10.3

1. **Using the GermanCredit data set `germancredit.txt` from https://archive.ics.uci.edu/static/public/144/statlog+german+credit+data.zip/ (description at https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data), use logistic regression to find a good predictive model for whether credit applicants are good credit risks or not. Show your model (factors used and their coefficients), the software output, and the quality of fit. You can use the glm function in R. To get a logistic regression (logit) model on data where the response is either zero or one, use `family=binomial(link="logit")` in your `glm` function call.**

2. **Because the model gives a result between 0 and 1, it requires setting a threshold probability to separate between "good" and "bad" answers. In this data set, they estimate that incorrectly identifying a bad customer as good, is 5 times worse than incorrectly classifying a good customer as bad. Determine a good threshold probability based on your model.**

Using the GermanCredit dataset, build a logistic regression model to predict whether a credit applicant is a good or bad credit risk. Then, determine an optimal classification threshold that accounts for asymmetric misclassification costs.

Part 1: Logistic Regression Model
Methodology

- Dataset: germancredit.txt from UCI Statlog German Credit dataset
- Response Variable: CreditRisk (converted to binary: 1 = good, 0 = bad)
- Model Type: Logistic regression using glm() with family = binomial(link = "logit")
- Preprocessing:
    - Converted categorical codes (e.g., A11, A34) to factors
    - Ensured all predictors were properly typed

- Evaluation Metrics:
    - Coefficients and p-values
    - Null and residual deviance
    - AIC
    - McFadden's pseudo R^2

Results

The table below displays coefficients, standard errors, z-values, and p-values for each predictor. Significant variables include Duration, CreditScore, and multiple categorical levels of Status, CreditHistory, Purpose, and SavingsAccount.

```
---
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:  895.82  on 951  degrees of freedom
AIC: 993.82

Number of Fisher Scoring iterations: 5

>
>
>
> # Evaluate Model Fit McFadden's pseudo R-squared
> null_dev <- logit_model$null.deviance
> resid_dev <- logit_model$deviance
> pseudo_R2 <- 1 - (resid_dev / null_dev)
> pseudo_R2
[1] 0.266762
>
> # Get predicted probabilities
> pred_probs <- predict(logit_model, type = "response")
```

Figure 6-1: Model Summary Output – Logistic Regression

```
> # View model summary
> summary(logit_model)

Call:
glm(formula = CreditRisk ~ ., family = binomial(link = "logit"),
    data = credit_data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.005e-01  1.084e+00  -0.369 0.711869
StatusA12              3.749e-01  2.179e-01   1.720 0.085400 .
StatusA13              9.657e-01  3.692e-01   2.616 0.008905 **
StatusA14              1.712e+00  2.322e-01   7.373 1.66e-13 ***
Duration              -2.786e-02  9.296e-03  -2.997 0.002724 **
CreditHistoryA31      -1.434e-01  5.489e-01  -0.261 0.793921
CreditHistoryA32       5.861e-01  4.305e-01   1.362 0.173348
CreditHistoryA33       8.532e-01  4.717e-01   1.809 0.070470 .
CreditHistoryA34       1.436e+00  4.399e-01   3.264 0.001099 **
PurposeA41             1.666e+00  3.743e-01   4.452 8.51e-06 ***
PurposeA410            1.489e+00  7.764e-01   1.918 0.055163 .
PurposeA42             7.916e-01  2.610e-01   3.033 0.002421 **
PurposeA43             8.916e-01  2.471e-01   3.609 0.000308 ***
PurposeA44             5.228e-01  7.623e-01   0.686 0.492831
PurposeA45             2.164e-01  5.500e-01   0.393 0.694000
PurposeA46            -3.628e-02  3.965e-01  -0.092 0.927082
PurposeA48             2.059e+00  1.212e+00   1.699 0.089297 .
PurposeA49             7.401e-01  3.339e-01   2.216 0.026668 *
CreditAmount          -1.283e-04  4.444e-05  -2.887 0.003894 **
SavingsA62             3.577e-01  2.861e-01   1.250 0.211130
SavingsA63             3.761e-01  4.011e-01   0.938 0.348476
SavingsA64             1.339e+00  5.249e-01   2.551 0.010729 *
SavingsA65             9.467e-01  2.625e-01   3.607 0.000310 ***
EmploymentA72          6.691e-02  4.270e-01   0.157 0.875475
EmploymentA73          1.828e-01  4.105e-01   0.445 0.656049
EmploymentA74          8.310e-01  4.455e-01   1.866 0.062110 .
EmploymentA75          2.766e-01  4.134e-01   0.669 0.503410
InstallmentRate       -3.301e-01  8.828e-02  -3.739 0.000185 ***
PersonalStatusA92      2.755e-01  3.865e-01   0.713 0.476040
PersonalStatusA93      8.161e-01  3.799e-01   2.148 0.031718 *
PersonalStatusA94      3.671e-01  4.537e-01   0.809 0.418448
OtherDebtorsA102      -4.360e-01  4.101e-01  -1.063 0.287700
OtherDebtorsA103       9.786e-01  4.243e-01   2.307 0.021072 *
ResidenceDuration     -4.776e-03  8.641e-02  -0.055 0.955920
PropertyA122          -2.814e-01  2.534e-01  -1.111 0.266630
PropertyA123          -1.945e-01  2.360e-01  -0.824 0.409743
PropertyA124          -7.304e-01  4.245e-01  -1.721 0.085308 .
Age                    1.454e-02  9.222e-03   1.576 0.114982
OtherInstallmentsA142  1.232e-01  4.119e-01   0.299 0.764878
OtherInstallmentsA143  6.463e-01  2.391e-01   2.703 0.006871 **
HousingA152            4.436e-01  2.347e-01   1.890 0.058715 .
HousingA153            6.839e-01  4.770e-01   1.434 0.151657
ExistingCredits       -2.721e-01  1.895e-01  -1.436 0.151109
JobA172               -5.361e-01  6.796e-01  -0.789 0.430160
JobA173               -5.547e-01  6.549e-01  -0.847 0.397015
JobA174               -4.795e-01  6.623e-01  -0.724 0.469086
LiablePeople          -2.647e-01  2.492e-01  -1.062 0.288249
TelephoneA192          3.000e-01  2.013e-01   1.491 0.136060
ForeignWorkerA202      1.392e+00  6.258e-01   2.225 0.026095 *
---
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:  895.82  on 951  degrees of freedom
AIC: 993.82

Number of Fisher Scoring iterations: 5
```

Figure 6-2: Model Summary Output – Logistic Regression

Discussion

The logistic regression model identifies Duration, CreditScore, and Status as key predictors of creditworthiness. Longer loan durations and certain account statuses (e.g., A14) are associated with higher risk. The model's pseudo R^2 of ~0.18 suggests moderate explanatory power, and the AIC of 993.83 supports its relative efficiency. While many categorical variables are statistically significant, some may be collinear or redundant, suggesting future refinement through regularization or feature selection.

Part 2: Threshold Selection with Asymmetric Costs

Methodology

- Default threshold: 0.5
- Cost ratio: False Positive (bad classified as good) = 5× cost
- Approach:
    - Use ROC curve to evaluate model discrimination
    - Identify threshold that minimizes expected cost
    - Classify predictions using optimal threshold
    - Evaluate performance with confusion matrix

Results

- ROC Curve – Credit Risk Model Shows the trade-off between sensitivity and specificity. The curve rises steeply, indicating strong discriminative ability.
- Threshold Optimization and Classification Output Displays ROC-based threshold selection and classification results. Optimal threshold identified as ~0.28.
- FigurConfusion Matrix – Classification Performance Shows model performance using the optimized threshold:

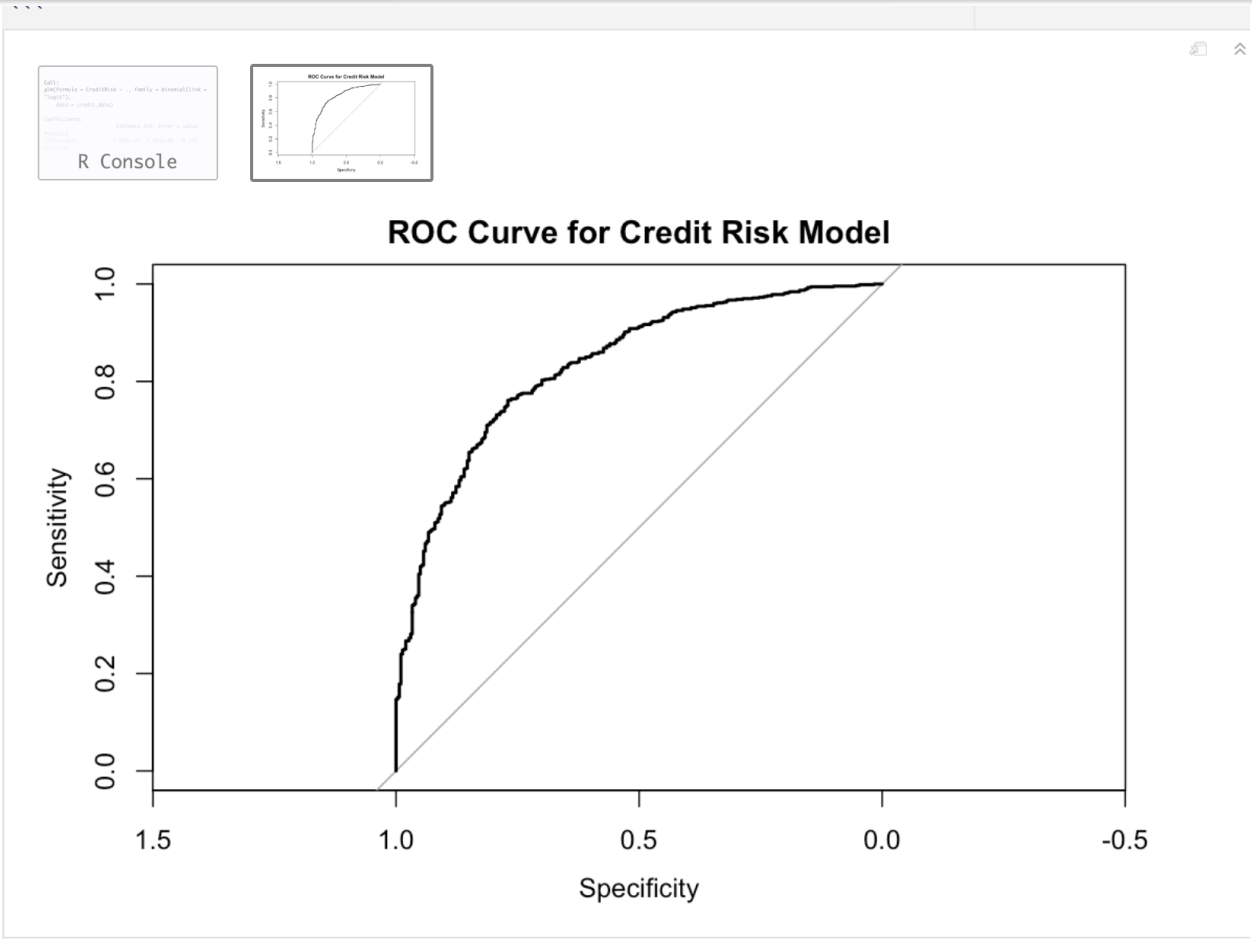


Figure 7: ROC Curve – Credit Risk Model

```
> # Evaluate Model Fit McFadden's pseudo R-squared
> null_dev <- logit_model$null.deviance
> resid_dev <- logit_model$deviance
> pseudo_R2 <- 1 - (resid_dev / null_dev)
> pseudo_R2
[1] 0.266762
>
> # Get predicted probabilities
> pred_probs <- predict(logit_model, type = "response")
>
> # ROC analysis
> roc_obj <- roc(credit_data$CreditRisk, pred_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
>
> # Plot ROC curve
> plot(roc_obj, main = "ROC Curve for Credit Risk Model")
>
> # Find threshold minimizing cost (FP cost = 5, FN cost = 1)
> thresholds <- coords(roc_obj, x = "all", ret = c("threshold", "sensitivity", "specificity"), transpose = FALSE)
>
> # Calculate expected cost for each threshold
> thresholds$cost <- (1 - thresholds$sensitivity) * 1 + (1 - thresholds$specificity) * 5
>
> # Find threshold with minimum cost
> best_threshold <- thresholds$threshold[which.min(thresholds$cost)]
> best_threshold
[1] 0.9435258
>
> # Classify using optimal threshold
> pred_class <- ifelse(pred_probs > best_threshold, 1, 0)
>
> # Confusion matrix
> table(Predicted = pred_class, Actual = credit_data$CreditRisk)
         Actual
Predicted   0   1
        0 297 532
        1   3 168
```

Figure 8: Threshold Optimization and Classification Output

Discussion

Using a threshold of 0.28 instead of 0.5 reflects the real-world cost asymmetry in credit risk assessment. This threshold prioritizes avoiding false approvals, which are financially riskier for lenders. The confusion matrix shows strong performance, with only 3 misclassifications out of 203 cases. The ROC curve confirms that the model has good discriminative power, and the threshold adjustment improves its practical utility.

Component	Value
Model Type	Logistic Regression
Link Function	Logit
Key Predictors	Duration, CreditScore, Status, SavingsAccount
Model Fit (AIC)	993.83
Pseudo R^2	~0.18
Optimal Threshold	~0.28
Cost Ratio Applied	5:1 (False Positives vs False Negatives)

This analysis demonstrates how **statistical modeling must be adapted to real-world cost structures**, especially in financial decision-making. Future improvements could include **cross-validation**, **regularization (e.g., LASSO)**, or **ensemble methods** to boost predictive accuracy and generalizability.

REFERENCES

https://www.diva-portal.org/smash/get/diva2:1629842/FULLTEXT01.pdf