**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.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 1: "Prediction Output – Regression Tree Model**

A screenshot of a computer screen

AI-generated content may be incorrect.

**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 R² of 0.7878, the tree model produced a much higher prediction and lacked statistical metrics like R². 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.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 3: Prediction and Variable Importance – Random Forest Model**

**A screenshot of a graph

AI-generated content may be incorrect.**

**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 R² 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.

A screenshot of a computer screen

AI-generated content may be incorrect.

**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**

# 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](https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data%20) ), 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.

1. **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 R²

**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.

A screenshot of a computer program

AI-generated content may be incorrect. **Figure 6-1: Model Summary Output – Logistic Regression**

**A screenshot of a computer code

AI-generated content may be incorrect.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 R² 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:

A graph of a credit risk model

AI-generated content may be incorrect.**Figure 7: ROC Curve – Credit Risk Model**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**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.

A screenshot of a computer

AI-generated content may be incorrect.

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