**Decision Tree Modeling Project in R - HMEQ\_Scrubbed.csv**

**Name:** Ellen Alves  
**Program:** MS in Business Analytics  
**Dataset:** HMEQ\_Scrubbed.csv

### **Introduction**

This project aims to develop decision tree models for both classification and regression using the HMEQ\_Scrubbed.csv dataset. The steps include predicting loan default (classification), predicting financial loss (regression), and building a combined probability and severity model.

### **Data Reading and Initial Exploration**

**Executed code:**

# Load libraries  
library(rpart)  
library(rpart.plot)  
library(ROCR)  
  
# Read the data

PATH <- "/Users/ellenalves/Desktop/master"

FILE\_NAME <- "HMEQ\_Scrubbed.csv"

INFILE = paste( PATH, FILE\_NAME, sep="/" )

setwd( PATH )

df = read.csv( INFILE )  
  
# Data structure and summary  
str(df)  
summary(df)  
head(df)

**Initial observations:**

* Numeric and categorical variables examined
* Presence of TARGET\_BAD\_FLAG and TARGET\_LOSS\_AMT variables

### 

### **Classification - Predicting Loan Default**

**Target variable:** TARGET\_BAD\_FLAG  
**Excluded variable:** TARGET\_LOSS\_AMT

**Models developed:**

*Decision Tree using Gini index*

A diagram of a flowchart

AI-generated content may be incorrect.

*Decision Tree using Entropy (Information Gain)*

A diagram of a graph

AI-generated content may be incorrect.

**Top important variables:**

| Rank | Gini Variables | Importance | Entropy Variables | Importance |
| --- | --- | --- | --- | --- |
| 1 | M\_DEBTINC | 570.021010 | M\_DEBTINC | 762.591210 |
| 2 | IMP\_DEBTINC | 128.539072 | IMP\_DEBTINC | 188.922871 |
| 3 | IMP\_DELINQ | 77.371518 | IMP\_DELINQ | 68.152477 |
| 4 | M\_VALUE | 51.334486 | IMP\_CLAGE | 40.125205 |
| 5 | IMP\_CLAGE | 36.076295 | LOAN | 34.053718 |
| 6 | LOAN | 25.645675 | M\_VALUE | 30.094365 |
| 7 | IMP\_DEROG | 22.501563 | IMP\_DEROG | 12.037746 |
| 8 | M\_DEROG | 9.540586 | IMP\_VALUE | 10.263083 |
| 9 | IMP\_VALUE | 8.551021 | IMP\_YOJ | 3.436136 |
| 10 | M\_DELINQ | 7.632469 | IMP\_CLNO | 3.075170 |
| 11 | M\_NINQ | 6.311465 | IMP\_MORTDUE | 1.219274 |
| 12 | IMP\_YOJ | 4.323751 |  |  |

**ROC Curves:**

A graph with a line and a point

AI-generated content may be incorrect.

Gini AUC : 0.8433084

Entropy AUC: 0.8293732

**Summary and Analysis:**

Based on the ROC curves and the AUC values, the **Gini model outperformed the Entropy model**. The Gini tree achieved an AUC of **0.8433**, while the Entropy tree achieved an AUC of **0.8294**. This indicates that the Gini-based decision tree had a slightly better ability to differentiate between defaulters and non-defaulters. Therefore, for this dataset, the Gini model is recommended for classification tasks.

**Confusion Matrix**

*Gini Model (fG)*

|  |
| --- |
| Actual  | 0 | 1  -------|------|-----  Pred 0 | 4468 | 370 <- True negatives, False negatives  Pred 1 | 303 | 819 <- False positives, True positives |

*Entropy Model (fE)*

|  |
| --- |
| Actual  | 0 | 1  -------|------|-----  Pred 0 | 4515 | 458 <- True negatives, False negatives  Pred 1 | 256 | 731 <- False positives, True positives |

I also evaluated the models using confusion matrices. The **Gini model correctly identified 819 defaulters**, compared to **731 identified by the Entropy model**. Additionally, the Gini model had **fewer false negatives (370 vs. 458)**, which means it missed fewer defaulters. This reinforces that the Gini-based decision tree performs better for this classification task, especially when identifying high-risk individuals is crucial.

### 

### **Regression - Predicting Financial Loss**

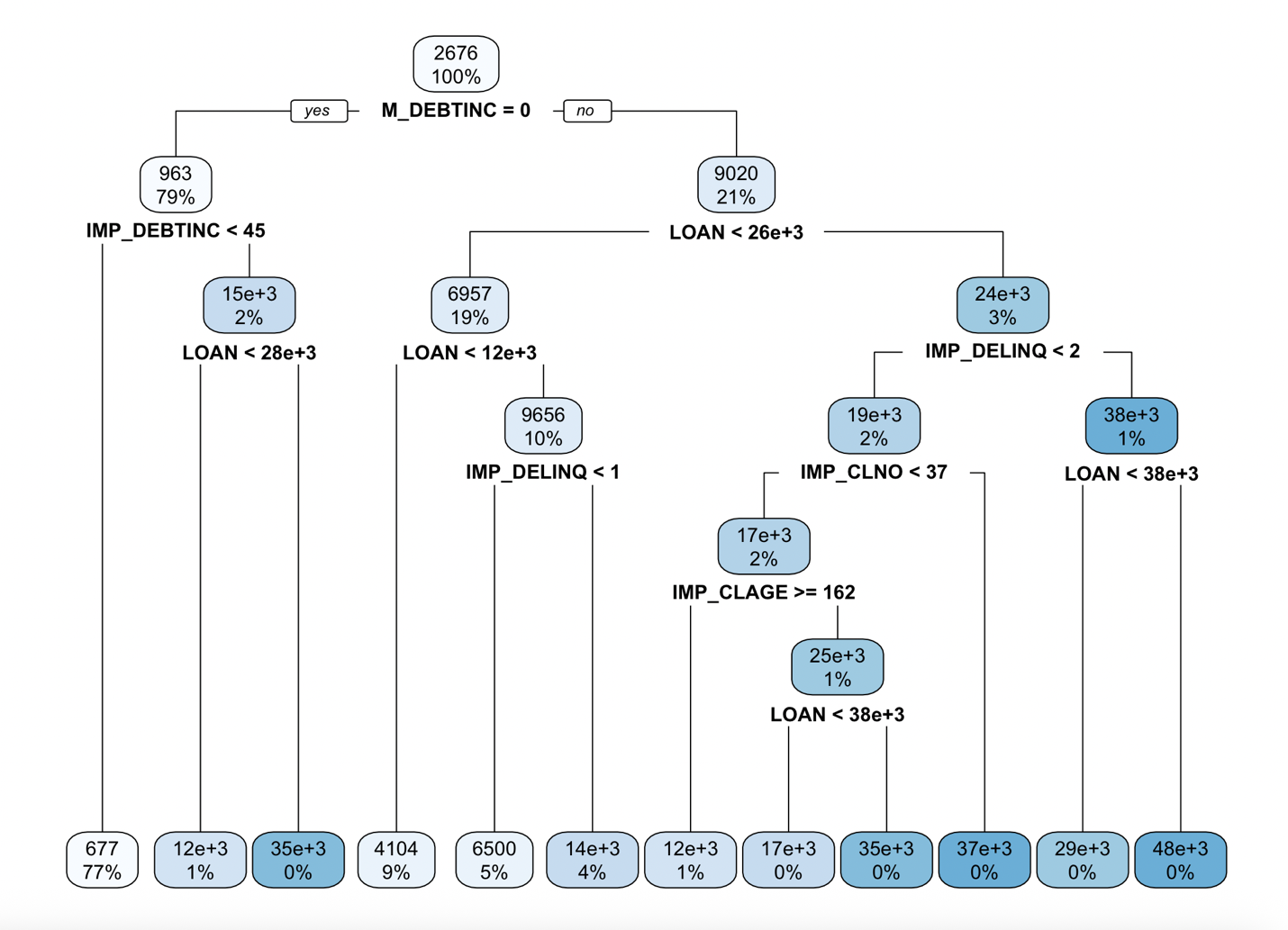
**Target variable:** TARGET\_LOSS\_AMT  
**Excluded variable:** TARGET\_BAD\_FLAG

**Executed code:**

# Prepare data  
df\_reg = df  
df\_reg$TARGET\_BAD\_FLAG = NULL  
  
# ANOVA model  
tree\_anova = rpart(TARGET\_LOSS\_AMT ~ ., data = df\_reg, method = "anova")  
rpart.plot(tree\_anova)  
tree\_anova$variable.importance  
pred\_anova = predict(tree\_anova, df\_reg)  
RMSE\_anova = sqrt(mean((df$TARGET\_LOSS\_AMT - pred\_anova)^2))  
  
# POISSON model  
tree\_poisson = rpart(TARGET\_LOSS\_AMT ~ ., data = df\_reg, method = "poisson")  
rpart.plot(tree\_poisson)  
tree\_poisson$variable.importance  
pred\_poisson = predict(tree\_poisson, df\_reg)  
RMSE\_poisson = sqrt(mean((df$TARGET\_LOSS\_AMT - pred\_poisson)^2))

**Models developed:**

*Decision Tree using ANOVA*



**Top important variables**

*ANOVA Model*

| **Rank** | **Variable** | **Importance** |
| --- | --- | --- |
| 1 | M\_DEBTINC | 64,758,513,590 |
| 2 | LOAN | 64,443,856,477 |
| 3 | IMP\_DEBTINC | 19,307,937,442 |
| 4 | IMP\_DELINQ | 18,468,415,581 |
| 5 | IMP\_VALUE | 9,985,413,830 |
| 6 | IMP\_CLNO | 8,640,006,256 |
| 7 | IMP\_MORTDUE | 7,345,104,792 |
| 8 | IMP\_CLAGE | 5,561,821,234 |
| 9 | M\_VALUE | 3,812,596,217 |
| 10 | IMP\_DEROG | 3,423,606,021 |
| 11 | FLAG.Reason.HomeImp | 2,487,025,698 |
| 12 | FLAG.Reason.DebtCon | 2,376,139,202 |
| 13 | M\_DEROG | 1,695,086,247 |
| 14 | M\_DELINQ | 1,384,320,435 |
| 15 | M\_NINQ | 1,101,806,061 |
| 16 | IMP\_YOJ | 803,802,835 |
| 17 | M\_YOJ | 727,900,700 |
| 18 | FLAG.Job.Other | 569,633,461 |
| 19 | M\_MORTDUE | 363,950,350 |
| 20 | FLAG.Job.Self | 269,034,105 |

The most influential variable in the ANOVA regression model was M\_DEBTINC, followed closely by LOAN, IMP\_DEBTINC, and IMP\_DELINQ. This suggests that both actual and imputed debt-to-income ratios, as well as the loan amount and delinquency history, play a significant role in predicting the loss amount. Job status and reason for the loan (debt consolidation or home improvement) also contributed moderately.

**RMSE**

ANOVA: 4848.417

Decision Tree using POISSON

A diagram of a number system

AI-generated content may be incorrect.

**Top important variables:**

*POISON Model*

| **Rank** | **Variable** | **Importance** |
| --- | --- | --- |
| 1 | M\_DEBTINC | 18,534,649.01 |
| 2 | IMP\_DEBTINC | 6,636,788.15 |
| 3 | LOAN | 5,093,017.45 |
| 4 | IMP\_DELINQ | 1,989,199.88 |
| 5 | IMP\_VALUE | 765,775.84 |
| 6 | M\_VALUE | 731,438.40 |
| 7 | IMP\_MORTDUE | 390,250.40 |
| 8 | IMP\_DEROG | 292,575.36 |
| 9 | FLAG.Reason.HomeImp | 214,334.43 |
| 10 | FLAG.Reason.DebtCon | 197,111.13 |
| 11 | IMP\_CLNO | 82,289.11 |
| 12 | IMP\_YOJ | 24,796.57 |
| 13 | FLAG.Job.Self | 12,398.29 |

The Poisson regression model identified M\_DEBTINC and IMP\_DEBTINC as the most important predictors of loss amount, followed by LOAN and IMP\_DELINQ. These variables represent the borrower’s debt-to-income ratios and delinquency history, which align with financial intuition. Other variables such as home improvement and debt consolidation loan reasons also contributed to the model, though to a lesser extent.

**RMSE**

POISSON: 5558.973

**Summary and Analysis:**

The ANOVA model achieved a lower RMSE (**4848.42**) compared to the Poisson model (**5558.97**), indicating that it predicted the loss amount with higher accuracy. Since RMSE measures the average magnitude of prediction error, the lower value confirms that the ANOVA regression tree is more suitable for this regression task.

### 

### **Probability x Severity Model**

**Executed code:**

# Classification model  
tree\_prob = rpart(TARGET\_BAD\_FLAG ~ ., data = df\_flag, method = "class")  
prob\_default = predict(tree\_prob, df\_flag, type = "prob")[,2]  
  
# Severity model  
df\_severity = subset(df, TARGET\_BAD\_FLAG == 1)  
tree\_severity = rpart(TARGET\_LOSS\_AMT ~ ., data = df\_severity, method = "anova")  
  
# Improved visualization  
rpart.plot(tree\_severity,  
 type = 2,  
 extra = 101,  
 fallen.leaves = TRUE,  
 compress = TRUE,  
 space = 0,  
 cex = 0.6)  
  
# Prediction and RMSE  
loss\_pred = predict(tree\_severity, df)  
expected\_loss = prob\_default \* loss\_pred  
RMSE\_combined = sqrt(mean((df$TARGET\_LOSS\_AMT - expected\_loss)^2))  
print(RMSE\_combined)

Classification model

In the Probability × Severity model, I first built a classification tree to predict the probability of default (TARGET\_BAD\_FLAG) using rpart. I then used predict() with type = "prob" to extract the probability that each record would default.

A diagram of a number of numbers

AI-generated content may be incorrect.

**Top important variables**

Classification model

| **Rank** | **Variable** | **Importance** |
| --- | --- | --- |
| 1 | M\_DEBTINC | 570.021010 |
| 2 | IMP\_DEBTINC | 128.539072 |
| 3 | IMP\_DELINQ | 77.371518 |
| 4 | M\_VALUE | 51.334486 |
| 5 | IMP\_CLAGE | 36.076295 |
| 6 | LOAN | 25.645675 |
| 7 | IMP\_DEROG | 22.501563 |
| 8 | M\_DEROG | 9.540586 |
| 9 | IMP\_VALUE | 8.551021 |
| 10 | M\_DELINQ | 7.632469 |
| 11 | M\_NINQ | 6.311465 |
| 12 | IMP\_YOJ | 4.323751 |
| 13 | M\_CLN\_ | 4.256569 |
|  |  |  |

The classification model identified M\_DEBTINC and IMP\_DEBTINC as the top predictors of default probability, followed by delinquency and value-related variables. This confirms that the borrower’s debt-to-income ratio and credit history are strong indicators of default risk.

Severity model

For the severity component, I filtered the dataset to include only records with TARGET\_BAD\_FLAG = 1, since loss severity is only relevant for accounts that actually defaulted. I then built a regression tree using the rpart function with the anova method to predict TARGET\_LOSS\_AMT.

To improve the visualization of the decision tree and avoid overplotting issues, I customized the rpart.plot() function using the arguments compress = TRUE and space = 0. This helped reduce spacing between branches and made the tree structure easier to read.

This adjustment was especially helpful because the original tree was too large to fit properly, even when reducing the font size with cex = 0.15. The final plot became much clearer for presentation and documentation purposes.

A diagram of a company

AI-generated content may be incorrect.

**Top important variables**

Severity model

| **Rank** | **Variable** | **Importance** |
| --- | --- | --- |
| 1 | LOAN | 95,490,401,040 |
| 2 | IMP\_VALUE | 12,784,086,193 |
| 3 | IMP\_MORTDUE | 9,235,762,508 |
| 4 | IMP\_DEBTINC | 8,190,794,036 |
| 5 | IMP\_CLNO | 5,197,172,526 |
| 6 | FLAG.Reason.HomeImp | 4,166,142,650 |
| 7 | FLAG.Reason.DebtCon | 3,945,210,843 |
| 8 | IMP\_NINQ | 1,284,908,173 |
| 9 | IMP\_YOJ | 742,530,790 |
| 10 | M\_MORTDUE | 620,243,894 |
| 11 | IMP\_CLAGE | 444,203,863 |
| 12 | IMP\_DELINQ | 243,986,582 |
| 13 | IMP\_DEROG | 88,840,773 |

The severity model found LOAN to be the strongest predictor of loss amount, followed by variables like IMP\_VALUE and IMP\_MORTDUE. This suggests that the original loan amount and the imputed values for mortgage and debt-to-income ratio have a significant impact on how much is lost when a borrower defaults.

**Combined model RMSE:** … 1] 4997.231

**Final comparison:**

The Probability × Severity model produced an RMSE of **4997.23**, which is higher than the ANOVA model (**4848.42**) but lower than the Poisson model (**5558.97**).  
Although the direct ANOVA regression remains the most accurate for predicting loss amount in this dataset, the combined Probability × Severity model provides additional business insight by separating the probability of default from the loss given default. This can help risk managers design targeted strategies for high-risk, high-loss customers.

**Final Conclusion:**

 **Best classification model:** Gini — it had the higher AUC (**0.8433**) and better confusion matrix results (fewer false negatives, more true positives) compared to Entropy.

 **Best regression model:** ANOVA — it achieved the lowest RMSE (**4848.42**), indicating higher prediction accuracy for loss amounts.

 **Combined model:** The Probability × Severity model **was not more effective** than the direct ANOVA model in terms of RMSE (**4997.23** vs **4848.42**), but it added valuable insight by separating default risk and loss severity, which can support more granular risk strategies.

 **Key takeaway:** This project reinforced the value of using both classification and regression trees to analyze financial risk, and showed that modeling probability and impact separately can improve business understanding even if it does not always outperform a direct regression.