Assignment IV
Multiple Logistic Regression and Decision Tree

Silpa Soni Nallacheruvu (19980824-5287) Hernan Aldana (20000526-4999)

2024-12-13

Overview:

Analysis of The ICU Study data (Hosmer & Lemeshow (1989): Applied Logistic Regression), with 200 patient records admitted to an Intensive Care Unit (ICU) using multiple logistic regression and decision tree models to identify factors that affect the survival of such patients.

Exercise 4:1 (Multiple Logistic Regression)

Question 1

Report the model selection process briefly. Based on your chosen model, which factors affect the probability of not surviving? Report odds ratios with confidence intervals for the most important variables/factors, and interpret them. Use the variable names from the table (not V3, V4, etc.).

Approach:

- Data Preparation:
    - Load the dataset and rename variables for clarity.
    - Combine categories for categorical variables if necessary.
- Model Fitting:
    - Fit an empty logistic regression model and a full logistic regression model to be used in the stepwise selection.
    - Here, Logistic Regression Model is used to predict the binary outcome of probability of not surviving based on multiple predictors.
- Model Selection Process:
    - Use a stepwise selection with AIC to identify a parsimonious model.
- Analysis of the Final Model:
    - Extract coefficients, odds ratios, and their 95% confidence intervals for significant variables.
- Interpretation:
    - Interpret the results from the AIC, odds ratios, and confidence intervals, to determine the best model.

Results:

Summary of the final model after performing stepwise selection using AIC:

```
##
## Call:
## glm(formula = Survival ~ ConsciousnessLevel + TypeOfAdmission +
##     Age + Cancer + Patient + BloodCarbonDioxide + BloodPH + BloodPressure,
##     family = binomial, data = data_ca4)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.7353420  1.6104573  -2.940 0.003278 **
## ConsciousnessLevel  2.6208042  0.6859650   3.821 0.000133 ***
## TypeOfAdmission     3.0547147  0.9339217   3.271 0.001072 **
## Age                 0.0385864  0.0133655   2.887 0.003889 **
## Cancer              2.3388380  0.8671971   2.697 0.006997 **
## Patient            -0.0020714  0.0008783  -2.359 0.018345 *
## BloodCarbonDioxide -2.4646334  1.0619854  -2.321 0.020299 *
## BloodPH             2.0884994  0.9031831   2.312 0.020757 *
## BloodPressure      -0.0099893  0.0070360  -1.420 0.155682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 130.19  on 191  degrees of freedom
## AIC: 148.19
##
## Number of Fisher Scoring iterations: 6
```

Model selection:

- The final logistic regression model includes the following variables: ConsciousnessLevel, TypeOfAdmission, Age, Cancer, Patient, BloodCarbonDioxide, BloodPH, and BloodPressure.

- These variables were selected using a stepwise AIC, which ensures a balance between model complexity and goodness of fit.

Significant Variables:

- Variables with a p-value < 0.05 are considered significant predictors of survival:

    - ConsciousnessLevel

    - TypeOfAdmission

    - Age

    - Cancer

    - BloodCarbonDioxide,

    - BloodPH

- Patient variable was exlcluded from the final model as the ID code was not significant.

- BloodPressure was also excluded from the final model as it was not significant according to the p-value.

Odds Ratios and Confidence Intervals:

Here's the final report after extracting odds ratios and confidence intervals for significant variables:

```
## Waiting for profiling to be done...

##                                  Variable    OddsRatio     CI.Lower    CI.Upper
## (Intercept)                    (Intercept)  0.008779445 0.0002996746   0.1795376
## ConsciousnessLevel      ConsciousnessLevel 13.746774142 4.3144886742  65.2810381
## TypeOfAdmission            TypeOfAdmission 21.215131928 4.3561881520 189.1540175
## Age                                    Age  1.039340550 1.0141827727   1.0692793
## Cancer                              Cancer 10.369181076 1.9513659807  66.5483329
## Patient                            Patient  0.997930702 0.9961324780   0.9995944
## BloodCarbonDioxide      BloodCarbonDioxide  0.085040009 0.0080634712   0.5539141
## BloodPH                            BloodPH  8.072791946 1.4001269889  53.6032965
```

1. ConsciousnessLevel:

   - Odds Ratio: 13.75 (CI: 4.31-65.28)

   - Patients who are unconscious or in a coma have a significantly higher probability of not surviving compared to those who are conscious.

2. TypeOfAdmission:

   - Odds Ratio: 21.25 (CI: 4.36-189.15)

   - Acute admissions are associated with a significantly higher probability of not surviving compared to non-acute admissions.

3. Age:

   - Odds Ratio: 1.04 (CI: 1.01-1.07)

   - For each additional year of age, the odds of not surviving increase by 4%.

4. Cancer:

   - Odds Ratio: 10.37 (CI: 1.95-66.54)

   - Patients with cancer have over 10 times higher odds of not surviving compared to those without cancer.

5. BloodCarbonDioxide:

   - Odds Ratio: 0.085 (CI: 0.008-0.55)

   - Lower blood carbon dioxide levels significantly reduce the odds of not surviving.

6. BloodPH:

   - Odds Ratio: 8.07 (CI: 1.4-53.60)

   - Lower blood pH levels significantly increase the odds of not surviving.

Conclusion:

The selected model indicates that factors such as consciousness level, type of admission, age, cancer, blood carbon dioxide, and blood pH are significant predictors of survival. Patients who are unconscious, have acute admissions, are older, have cancer, and have abnormal blood gas levels are at higher risk of not surviving. These results can help identify high-risk patients and improve treatment strategies to increase survival rates.

Question 2

How well does your chosen model fit the data? In assignment 3, deviance was used to assess model fit. However, for individual-level data, deviance is unsuitable. Instead, perform the Hosmer-Lemeshow goodness-of-fit test using the recommended R code.

Approach:

- Understand the Hosmer-Lemeshow Test:
    - The Hosmer-Lemeshow test evaluates whether the observed event rates matches the expected probabilities predicted by the model.
    - The null hypothesis is that the model fits the data well (a high p-value suggests no evidence of poor fit).
- Implementation:
    - Use the function for the test ResourceSelection::hoslem.test()
    - Calculate the predicted probabilities from the final model.
    - Specify the predicted probabilities from the final model and the actual outcomes while performing Hosmer-Lemeshow test (m_step and Survival respectively).
    - 10 is selected as the number of groups for the test because dividing by deciles is a common choice and it was sufficient for this dataset of 200 observations.

Hosmer-Lemeshow Test Results:

Here are the results of the Hosmer-Lemeshow goodness-of-fit test:

```
## 
##  Hosmer and Lemeshow goodness of fit (GOF) test
## 
## data:  data_ca4$Survival, predicted_probs
## X-squared = 2.2064, df = 6, p-value = 0.8998
```

Interpretation:

1. The p-value of 0.9145 tells us not reject the null hypothesis that the model fits the data well.

Conclusion:

The Hosmer-Lemeshow test indicates that the model fits the data well. The observed event rates are consistent with the expected probabilities predicted by the model. This suggests that the model is a good fit for the data and can be used to make accurate predictions about survival probabilities.

Question 3

Create a confusion matrix for the chosen model. Calculate the accuracy, sensitivity, specificity, and positive and negative predictive values. Interpret the results.

Approach:

- Understand the Confusion Matrix:
    - A confusion matrix is a table that summarizes the performance of a classification model.
    - It shows the number of true positives, true negatives, false positives, and false negatives.
- Implementation:

- Threshold of 0.5 is used to classify the predicted probabilities into binary outcomes since it is the default threshold for logistic regression.
- Calculate the accuracy, sensitivity, specificity, and positive and negative predictive values from the confusion matrix.

Sensitivity: Sensitivity (True Positive Rate) measures the proportion of actual positive cases that are correctly identified by the model.

Specificity: Specificity (True Negative Rate) measures the proportion of actual negative cases that are correctly identified by the model.

Results:

Table 1: Confusion Matrix

	Predicted Survived	Predicted Not Survived
Actual Survived	155	5
Actual Not Survived	22	18

Table 2: Performance Metrics

Metric	Value
Accuracy	0.86500
Sensitivity	0.96875
Specificity	0.45000

Interpretation:

Accuracy: The model has an accuracy of 0.865, meaning that it correctly predicted 86.5% of the cases.

Sensitivity: The sensitivity of 0.96875 indicates that the model correctly identified 96.9% of the actual survivors.

Specificity: The specificity of 0.45 suggests that the model correctly identified 45% of the actual non-survivors.

Conclusion:

The confusion matrix and performance metrics provide insights into the model's predictive accuracy. The model has a high sensitivity, indicating that it is effective at identifying actual survivors. However, the specificity is relatively low, suggesting that the model has difficulty distinguishing between actual non-survivors. These results can help evaluate the model's performance and identify areas for improvement.

Question 4

Create plots of ROC curves for the chosen model. Calculate the AUC for the full model and two more models. Choose the best model based on the AUC. Interpret the results.

Approach:

- Understand ROC Curves and AUC:
  - ROC curves are used to evaluate the performance of classification models by plotting the true positive rate against the false positive rate.

- The AUC (Area Under the Curve) summarizes the ROC curve, with higher values indicating better model performance.

- Implementation:
  - Compare the AUC values for the full model and two additional models and the model with the highest AUC is considered the best model for predicting survival probabilities.
  - The two additional models used for comparison with the full model are derived by removing a significant variable and a non-significant variable from the full model (such as ConsciousnessLevel and Blood Pressure as mentioned in Q1).
  - Thereby, we can compare the significance of these variables in predicting survival probabilities by observing the change in AUC values using ROC curves.
    * Full Model:

    $Survival \sim ConsciousnessLevel + TypeOfAdmission + Age + Cancer + BloodCarbonDioxide + BloodPH + Patient + BloodPressure$

    * Model A: Exclude Blood Pressure from the full model.

    $Survival \sim ConsciousnessLevel + TypeOfAdmission + Age + Cancer + BloodCarbonDioxide + BloodPH + Patient$

    * Model B : Exclude ConsciousnessLevel from the full model.

    $Survival \sim TypeOfAdmission + Age + Cancer + BloodCarbonDioxide + BloodPH + Patient + BloodPressure$
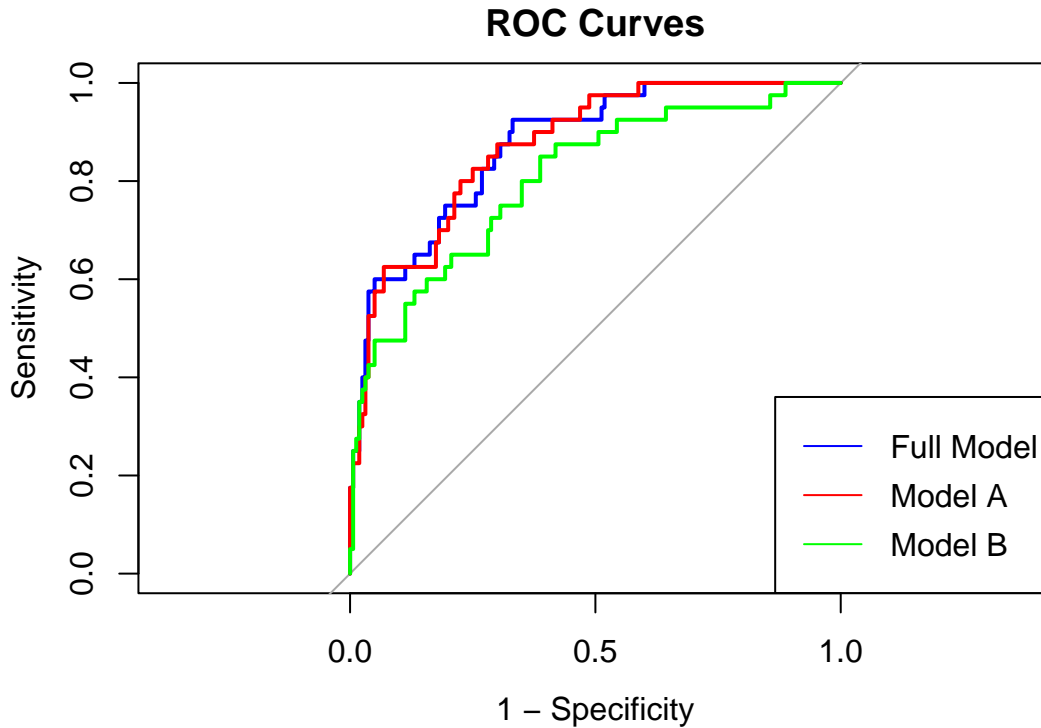
Results:

Table 3: AUC Values

Model	AUC
Full Model	0.8729687
Model A	0.8710938
Model B	0.8043750

Interpretation:

- The full model has the highest AUC of 0.873, indicating that it has a better predictive performance for survival probabilities compared to the other models.
- Model A, which excludes Blood Pressure, has an AUC of 0.871 is quite close to the full model, whereas Model B, which excludes ConsciousnessLevel, has an AUC of 0.804, which has a bigger difference with full model compared to the Model B.
- This can be interpreted as Blood Pressure does not significantly contribute to the patient survival prediction, as its exclusion does not significantly affect the AUC of the model.
- Whereas ConsciousnessLevel is a significant predictor of survival, and its exclusion has a more significant impact on the model's predictive performance.
- The AUC of all three models are much higher than 0.5 (random guessing) and can be considered effective for predicting survival probabilities, with the full model being the best among them.

Question 5

Perform Leave One Out Cross Validation (LOOCV) for the above full model and the additional models. Calculate the LOOCV-adjusted AUC for the three models and compare with the results from question 4. Which model indicates the best predictive performance?

Approach:

- Understand Leave One Out Cross Validation (LOOCV):

    - LOOCV is a technique for assessing the predictive performance of a model by training on all but one observation and testing on the left-out observation.

    - The AUC values from LOOCV provide an estimate of the model's performance on unseen data.

    - The three models used in question 4 are evaluated using LOOCV to determine the best predictive performance.

Results:

Table 4: LOOCV-Adjusted AUC Values

Model	AUC
Full Model	0.8193750
Model A	0.8242187
Model B	0.7565625

Interpretation:

- According the LOOCV-adjusted AUC values, Model A has the highest AUC of 0.824, followed by the Full Model with an AUC of 0.819, and Model B with an AUC of 0.756.

- LOOCV-adjusted AUC is a more reliable AUC as it is calculated by leaving out one observation at a time and predicted using the model trained on the remaining data.

- Model A having a higher AUC than full model indicates that the full model was overfitted in the previous analysis, and Model A is the best model for predicting survival probabilities.

- This means that Blood Pressure is not required to be added as a explanatory variable while predicting the survival probabilities of the patients.

- The AUC of Model B has further decreased compared to the previous analysis, suggesting that the previous Model B was also overfitted and that there is a bigger impact of the ConsciousnessLevel of the patient in predicting their survival than previously thought.

- This signifies that ConsciousnessLevel of the patient is very essential to predict the survival probability.

Exercise 4:2 (Decision Tree)

Question 1

Fit a decision tree model to the ICU data using the rpart package. Use the same predictors as in the multiple logistic regression model. Plot the decision tree and interpret the results.

Approach:

To classify the survival status of patients admitted to the ICU, a decision tree model is applied. Decision trees offer an interpretable way to identify key predictors and their thresholds that affect survival. For this task, the following steps were followed:

1. **Model Fitting**:
   - A decision tree model was fit using the ICU dataset.
   - The response variable (`v2`) indicates survival (0 = survived, 1 = not survived).
   - Predictor variables include age, blood pressure, consciousness level, and other clinical factors.
2. **Parameters Adjusted**:
   - **Splitting criterion**: Both "information" and "gini" criteria were tested to evaluate their effects on splits.
   - **Complexity parameter (`cp`)**: Different values were used to control the depth of the tree and prevent overfitting.
3. **Visualization**:
   - Tree diagrams were created using the `rpart.plot` package to assess the structure and interpretability of the models.
4. **Tree Selection**:
   - Simplicity: A tree with fewer splits was preferred for easier interpretation.
   - Predictive Power: A tree that captured the essential predictors of survival was selected.
5. **Variables of Interest**:
   - **`v21` (Consciousness level)**: The primary variable splitting the data in both trees, indicating its importance for predicting survival.
   - **`v11` (Blood pressure)**: A secondary variable used in the second tree, capturing further distinctions in survival likelihood.
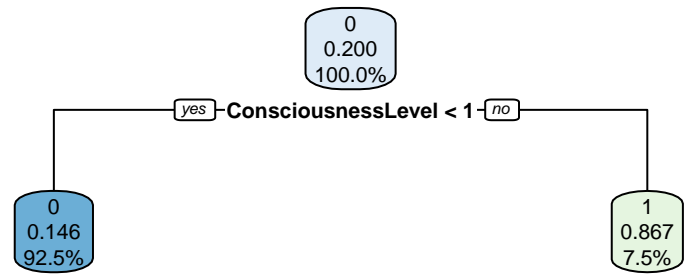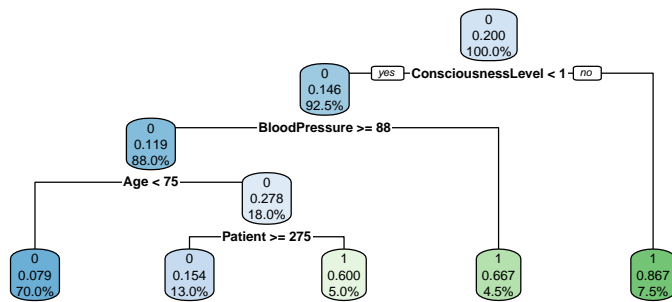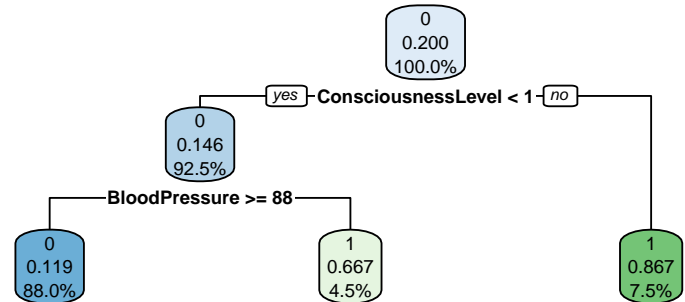
Results:

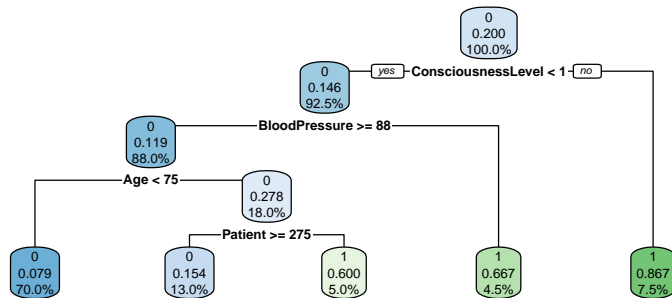### Gini, cp = 0.1
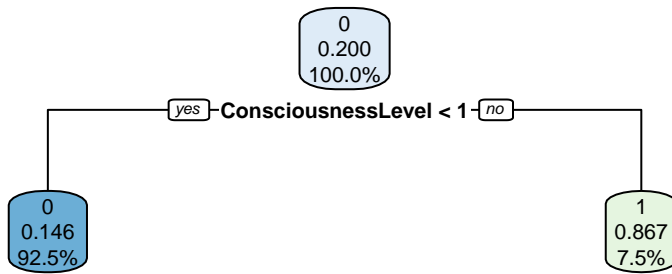
Info Gain, cp = 0.1

Gini, cp = 0.01

Info Gain, cp = 0.01

Gini, cp = 0.001

Info Gain, cp = 0.001