Dolores Farley

Week 6 Exercises (Include screenshots in all your responses)

**Book: Building Better Models with JMP Pro (Chapter 6)**

Use the Titanic Passengers BBM.jmp dataset posted on Blackboard for this exercise. Take some time to initially explore this new data, and then build a classification decision tree for **Survived** using the available variables (refer to Chapter 5 – Page 115 in your Building Better Models w/ JMP Pro book for information on the data and definitions of each variable). Please exclude the variables **Name** and **Home / Destination** from your analysis. **Prior to running the Partition platform for the decision tree analysis, assign a random seed using the integer value of “123”.**

Use what we learned about training and validation to **assign a holdout sample of 30% (or .30)** to help optimize the number of splits and assess your model’s performance. Leave all other default settings alone in the Partition platform, and don’t forget to assign the random seed prior to running the decision tree analysis.

1. Adjust the property setting under the red hot key to Show Split Counts. **Press Split three times.** Please describe to me in business language, as if I’m your manager, the “story” of the segment that had the highest probability of Surviving.

A screenshot of a computer

Description automatically generated

The segment that had the highest probability of surviving was women of passenger class of 1 or 2 at a 93.02 % survival rate.

1. How many splits are in your final tree (Hint: Press Go)? Please include visualization of your Split History, and short explanation on how SAS JMP arrived at this number of splits.

There are 6 splits are in my final tree. The first factor was separating the groups into male and female. Out of the female grouping the second factor was passengers of class 3 (46.05% survival rate for that group ) in one group and the other group was passenger of group 1&2 (93.33% survival rate for that group). The third split was on the male grouping. This third split was passage group of 1 at a 38.06% survival rate for class 1 and passenger class (2,3) at a 14.79% survival rate for class 2 and 3. The fourth split was for male passengers of class 2&3. In this fourth spilt the two grouping was age greater or equal to 11 at a survival rate of 12.08%, and age less than 11 at a 51.52% survival rate. The fifth spilt was male passengers of class (2,3) that was under 11 years of age that had siblings and spouses less than 3 at a 88.89% survival rate compare to 6.67% who had 3 or more siblings and spouses. The sixth and final split was for women passengers of class 3 into two groups of fare price. A fare greater or equal than 25.4667 had a survival rate of 4.35%, and a fare less than 25.4667 had a survival rate of 53.49%.

A screenshot of a computer screen

Description automatically generated

A graph with red and blue lines

Description automatically generated

A computer screen shot of a computer tree

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Which variables are the largest contributors? Expected responses are somewhat subjective, but assess the G^2 Entropy (or Information Gain) scores, and make an analyst decision on whether you view a sharp drop-off to finalize your answer.

Before we split and by looking at the G^2 Entropy (or Information Gain) and Logworth scores, we can see that the biggest factors are sex, passenger class, fare followed by port, age, parent and children, and siblings and spouses.

Before SplitA screenshot of a computer

Description automatically generated

After Split, we see the largest entropy (G^2) is sex, passenger class, and age, followed by siblings and spouses, fare, parents and children, and ports.

After Split:

A screenshot of a computer

Description automatically generated

The G^2 Entropy (or Information Gain) and Logworth accounts for factors separately. In the final answer (see answer for g) we see that the second highest survival groups (86.06%) were males in passenger class (2,3) who were less than 11 years old who had less than 3 siblings and spouses.

1. What is the validation misclassification rate for this model? Is the model better at predicting survival or non-survival?

The validation misclassification rate for this model is 17.46%. The model better at predicting non-survival 86.2% versus survival at 77.3% on the validation training set.

A screenshot of a computer

Description automatically generated

1. What is the area score under the validation ROC curve for **Survived**? Does the model do a better job of classifying **Survival** than a random 50/50 model?

The area score under the validation ROC curve for **Survived is 0.8540 or 85.40%.** The model does a better job of classifying **Survival** than a random 50/50 model because the random 50/50 model’s area is only 0.5. A perfect predictor or perfect fit would have an area of 1.

A screenshot of a computer

Description automatically generated

1. What is the lift for the model at portion = 0.3 and at portion = 0.5? Interpret these values.

The lift for the model at portion = 0.3 is roughly 2.125. This means that rows in the data table that correspond to the top 30% of the model’s predicted probabilities, the number of actual survival outcomes is 2.125 times higher than we would expect if we had chosen 30% of the rows in the data set at random.

The lift for the model at portion = 0.5 is roughly 1.75. This means that rows in the data table that correspond to the top 50% of the model’s predicted probabilities, the number of actual survival outcomes is 1.75 times higher than we would expect if we had chosen 50% of the rows in the data set at random.

A screenshot of a graph

Description automatically generated

1. Summarize the three “purest” segments with *respect to survival* as if I were your manager at work, and I wanted a business language summary of the top three groups that the Decision Tree identified. Including the overall survival rate for the total passenger population is always good to include as a frame of reference (which can be obtained using a distribution report rather than using the values listed for training and validation).

*Hint: Leaf Report can be helpful here by reinterpreting the English rules that the Decision Tree generates into business language*

A screenshot of a computer

Description automatically generated

Out of 1309 passengers only 500 people survived. That is only a 38.197% survival rate for the whole group. The top three groups that the Decision Tree identified were first the women of passenger class (1,2) at a 93.02% survival rate. The second group was male passengers of class (2, 3) under the age of 11 who had less than 3 spouses and siblings at an 86.06% survival rate. The third groups were females of Passenger class 3 who’s fare was less than 25.4667 at a 53.39 % survival rate.

A screenshot of a computer

Description automatically generated

**931 total training, 378 validations**