**CSC 435 LAB 05 Points: 20**

**Logistic Regression and Decision Tree Analysis**

Your Team Member names:

1. Stryder R. Coleman

2. Padraic Darby

**Background**

In many industries throughout the world, suppliers compete for business by submitting quotes for work, services or products. A key criterion used to determine the winning quote is the dollar amount of the quote, but other factors include expected quality, estimated delivery time of the product, or quoted completion time of the work.

The focus of this case is a supplier of equipment to the automotive industry. The products of interest in this case are various precision metal components used in a range of automotive applications, such as braking systems, drive trains, and engines. Some of the products will be used in the manufacture or assembly of new automobiles (i.e. original equipment), while others will be used as replacement parts in automobiles already on the road (i.e. aftermarket).

**The Task**

The supplier wants to increase sales and expand its market position. Many of the quotes provided to prospective customers in the past haven’t resulted in orders. Do the data provide any indication why? Are there certain situations that make it more or less likely that a customer will place an order?

**The Data:** Lost Sales.csv

The data set contains 550 records for quotes provided over a six-month period. The variables in the data set are listed in Table 1:

Table 1 - Details of the variables in the data file (Meta Data).

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Details** | **Value/(Units)** | **Data Type** |
| A | Quote | Dollars – Quoted Price | Continuous |
| B | Time to Delivery | Days - The quoted number of calendar days within which the order is to be delivered | Continuous |
| C | Part Type | OE/AM - OE = original equipment; AM = aftermarket | Nominal |
| D | Status | Won/Lost - Whether the quote resulted in a subsequent order within 30 days of receiving the quote: Lost = the order was not placed; Won = the order was placed. | Nominal |

**Purpose:**

This lab introduces students to Logistic Regression analyses. After completing tasks in this lab, you should be able to:

* Partition data into Training and Validation data
* Perform Simple and multiple logistic regression models using R
* Compare the results of training data and validation data by interpreting confusion matrix
* Interpret the Odds ratios

**Lab Instructions:**

1. Many of the sales managers insist that a more important factor in lost sales opportunities is related to pricing (Quote). They believe that the company’s prices are too high, and that customers are moving to lower cost providers. Test whether this is true. Provide rationale for your answer. Run the LR and interpret the results. (Hint: yourModel: Status vs. Quote).

***(For each step provide the code and the result)***

1. What is the P- Value for the model? What is the conclusion?

0.397. Quote is not a good predictor variable for Status and should not be used.

1. Compute and Interpret the Odds Ratio:

0.99997 is the odds ratio; and the confidence interval is 95%

A Win is going to occur slightly less than a lost

1. Compute and Interpret the confusion matrix:
   1. Overall accuracy rate:

0.5504587

* 1. Overall misclassification rate:

0.4495413

1. Do you think that the important factor in lost sales opportunities is related to pricing (Quote)? Why or why not?

No, the p-value is high and the misclassification rate is high.

1. Test whether Time to Delivery (the quoted number of days before the order will be delivered) is an important factor in determining lost sales. Run the LR and interpret the results. (Hint: yourModel: Status vs. Time to Delivery).

***(For each step provide the code and the result)***

1. What is the P- Value for the model and what is the conclusion?

6.26e-06. That Time to Delivery is a good predictor variable for predicting Status

1. Compute and interpret the Odds Ratio:

0.98292, so a win is going to occur slightly less than a lost

1. Compute and interpret the confusion matrix:
   1. Overall accuracy rate:

0.6146789

* 1. Overall misclassification rate:

0.3853211

1. Do you think that the important factor in lost sales opportunities is related to delivery time? Why or why not?

Yes, the p-value is low and accuracy rate is fairly high.

1. Partition the data into Training and Validation (70/30) data sets (use stratified random on Status). Build a multiple Logistic Regression model with Status as dependent variable. Quote, Time to Delivery, and Part Type as independent variables.
2. Identify & interpret the p-values for the model & the independent variables.
3. Quote: 0.347794
4. Time to delivery: 3.22e-05
5. Part Type: 0.031650

Time to delivery and Part Type are good predictor variables.

1. Use Prediction to explore the effects of these three predictors.
   1. What is the probability of winning if the Part Type is AM, Quote = $5,000, and time to delivery is 20 days?
   2. What is the probability of losing if the Part Type is OE, Quote = $10,000, and time to delivery is 50 days?
2. Compute and interpret the Odds Ratio for all three independent variables.

Quote 0.99997

`Time to Delivery` 0.98392

`Part Type`OE 0.60716

All the odds ratios are low.

1. What are the managerial implications - summarize what you have learned from this analysis?

Time to Delivery and Part Type are the best predictor variables for predicting Status because they have low p-values.

1. Partition the data into Training and Validation (70/30) data sets. Build a Decision Tree model with “Status” as dependent variable and all other variables as independent variables***. (provide the code and the result)***
   1. list the R-square and the overall accuracy for both Training and Validation sets.

Training: Test:

Accuracy: 0.6709845 Accuracy: 0.5853659

Error: 0.3290155 Error: 0.4146341

* 1. List the top three variables in the order of importance in identifying the Status.

Time to Delivery , Part Type, and Quote

* 1. Calculate the Sensitivity and specificity rates and identify the number of False-Positive and False-Negative cases only for the Validation data set.

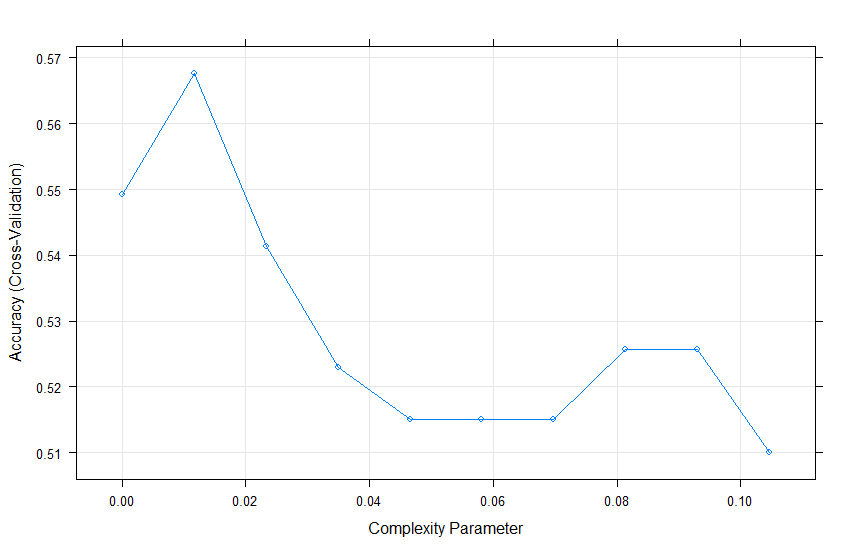
predicted. Lost Won False Negatives: 54/81

Lost 27 14 False Positives: 14/83

Won 54 69 Sensitivity: 69/83

Specificity: 27/81

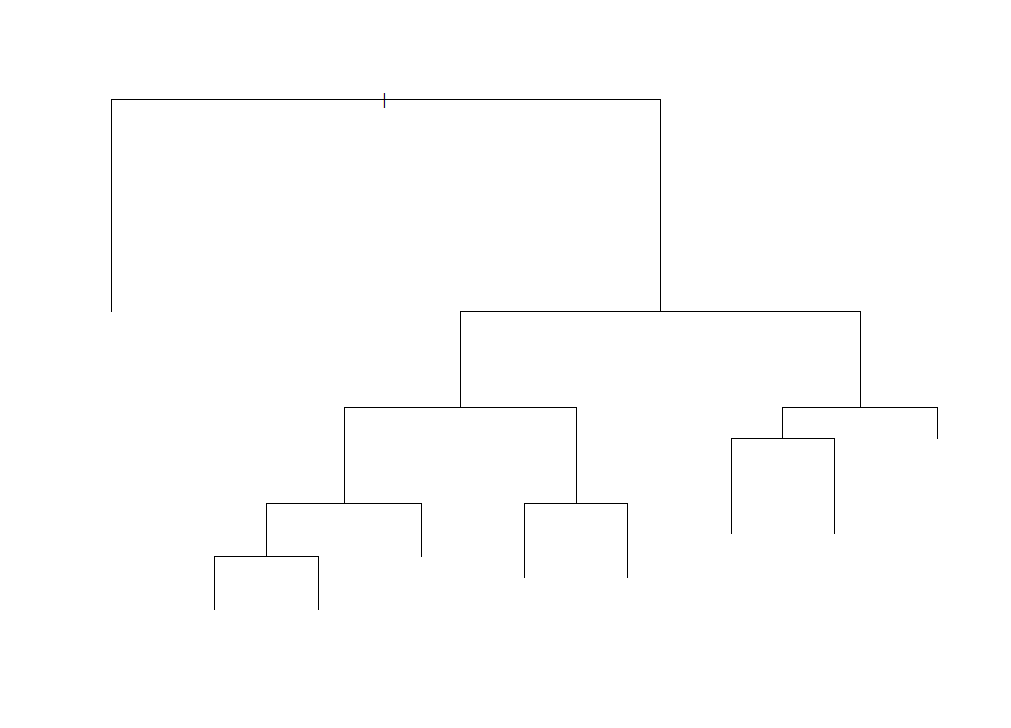
* 1. Build a pruning Decision Tree model with “Status” as dependent variable and all other variables as independent variables. Test with 10 different values of **cp** to choose the optimal **cp** that maximize the cross-validation (“cv”) accuracy.
  2. Plot the model accuracy vs different values of cp.



* 1. Print the best tuning parameter cp that maximizes the model accuracy.

2 0.01163467

* 1. Plot the final tree model.



* 1. List the model accuracy rate on the test data

Accuracy: 0.5853659

Error: 0.4146341