**Lost Sales: Logistic Regression**

(Case Study adapted from “JMP Case 17 – Lost Sales: Logistic Regression, Marlene Smith, University of Colorado Denver Business School)

**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**

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

|  |  |
| --- | --- |
| **Quote** | The quoted price, in dollars, for the order |
| **Time to Delivery** | The quoted number of calendar days within which the order is to be delivered |
| **Part Type** | OE = original equipment; AM = aftermarket |
| **Status** | 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 |

**Analysis**

Begin by determining the current state of the company’s ability to win orders.

* What percentage of the quotes don’t result in subsequent orders within 30 days?

The company sells to both original equipment manufacturers (OE) and to aftermarket suppliers (AM). Sales managers expect that the “hit” rates for these two markets are vastly different.

* Using a simple contingency table or graph, does it appear that winning a sale is associated with part type (OE or AM)? That is, does there appear to be a difference in the percentage of sales won for OE vs AM suppliers? (You don’t need to do a formal statistical test)

Many of the sales managers insist that a more important factor in lost sales opportunities is related to pricing. They believe that the company’s prices are too high, and that customers are moving to lower cost providers.

To explore the relationship between Status and Quote, we can use a logistic regression model.

* Build a simple logistic regression model for Status vs Quote. Based on this model, does it appear that the likelihood that an order will be lost increases or decreases as the quoted price increases? Is quoted price a significant predictor of whether or not an order will be lost?

The remaining factor is Time to Delivery (the quoted number of days before the order will be delivered).

* Use logistic regression to explore the relationship between winning an order and delivery time (with the ultimate goal of predicting the probability of losing and order). Is time to delivery a significant predictor of whether or not the order is lost? How does time to delivery affect the likelihood of losing an order?

Finally, build a multiple logistic regression model that includes Quote, Part Type, and Time to Delivery as the explanatory variables. If any of the explanatory variables are not significant, remove them one at a time until you end up with a model where all of the explanatory variables are significant.

* Does this model appear to be useful for predicting the likelihood of a lost sale? What is the misclassification rate? What is the area under the ROC curve?
* Generate a confusion matrix. Does the model appear to be more effective at correctly predicting lost sales or won sales (or is it equally effective at predicting both)?
* Suppose that we have an order in the pipeline that is for an after market (AM) supplier with a quoted price of $1400 and an expected time to deliver of 20 days. What is the probability that this sales will be lost? Assuming that the sales person has some control over the quoted price and time to delivery, what actions could he take to decrease the probability of losing the sale (or, equivalently, to increase the probability of winning the sales)?

For further thought (please answer):

* In this analysis, you built three different models for predicting lost sales. How do the misclassification rates compare for the three models? Is this what you would expect based on the significance of the explanatory variables?
* Based on the misclassification rates, which model does the best job of predicting Status?
* How could this model be improved? For example, are there any variables missing from this data set that might improve the model’s predictive ability?