**Insurance Medical Costs – Classification Lab**

A US medical insurance company is considering re-structuring their medical insurance offerings, and is wanting to decide how to design their insurance plans. The main goal is decide whether to continue to offer insurance to all applicants, or to become a boutique insurance company which carefully chooses who to insure, and only offers insurance to a select few. The company’s COO believes that there should be a correlation between specific factors and whether the medical costs are high or low, and is wanting to use Classification to see if it is possible to predict which types of factors would likely contribute to a high insurance cost.

The data provided in Classification Lab - MediCareCosts.jmp is used to draw your conclusions. The variables in the data set are as followed:

**Age** - age of applicant

**Gender** - 1 for female and 0 for male

**BMI** - Body mass index. The ideal BMI is 18.5 to 25

**Children** - Number of children covered by health insurance

**Non-Smoker- 1 if the applicant is a smoker, 0 if they are a non-smoker**

**Region** – the residence of the person within the country, East Coast (1), West Coast (2), Southern US(3), Mid-West(4).

**Cost** - 1 if the medical costs billed by health insurance is high, 0 otherwise

In using the above variables, please find the answers to the following questions. **Please use Cost as Y variable and all others as X variables**.

For a cut-off of .85 using Linear Classification (using Stepwise regression):

1. What is the rsquare of the model? What does this number suggest?

0.5681

1. What is the p-Value of the model? What does this suggest?

<0.001, the full model is significantly better than the reduced model.

1. What is the logOdds equation?

logOdds(Cost) = log(cost/(1-cost)) = 13.4065449 + age \* (-0.0511209) + BMI \* (-0.2158657) + Non-smoker \* (-5.3879649)

1. What is the level of accuracy you see in your model?

86.83%

1. In just evaluating the testing data, if you decided to only offer insurance to those who you have predicted to have a low cost medical bill, how many people would you have offered insurance to?

64+68=132

1. What is the significance of the threshold? Does increasing the threshold to .95 give us a more accurate prediction?

Threshold represents the tradeoff between false positives and false negatives. The choice of a threshold depends on the importance of TPR and FPR classification problem. No, the threshold of .95 gives us a less accuracy of 79%.

For a cut-off of .85 using Decision Tree Classification

1. Which variables are most important in making a decision?

Non-Smoker, age, and BMI are most important variables in making a decision.

1. Which is the most important variable according to your model?

Non-Smoker is the most important variable according to my model because it is the top node in the decision tree.

1. According to your model, which are the features of people who have a high medical cost?

Non-smoker < 1, non-smoker ≥ 1 & BMI < 30.02 & Age < 57

1. What is the probability a person who is a smoker and who has a BMI of less than 30.02 will have a high insurance cost? What percentage of the training data population is accounted for by this group?

0.778, 81/738=10.98%

1. ~~What is the r square value for the training and testing data set?~~
2. How many splits were made before the R square of the testing data started to drop?

4 splits, the R-squred is 0.697 greater the 5 splits of 0.691.

1. What is the level of accuracy you see in your model?

97.17%

1. Assume the following:
2. If the company **correctly** predicts someone is a high insurance cost, they do not take on the client, and do not make or lose any money.
3. If the company **correctly** predicts someone is a low insurance cost, they take on the client and make $3000 per client.
4. If the company **incorrectly** predicts someone is a high insurance cost, they turn away the client, and lose $3000.
5. If the company **incorrectly** predicts someone is going to be a low insurance cost, they accept the client and lose $5000.

If the company switches to a boutique model, they have to make at least $120,000 in profit to convince the board. Using the testing data, and based on these assumptions at a .85 cut-off rate, do you suggest the company offers boutique insurance services or continue to offer insurance to anyone who applies?

They make $148,000 in profit, so they should offer a boutique insurance services.

General Questions

1. Recall from our previous class, the ‘Baseline’ calculation (if we assign a 1 or 0 to all rows, instead of using Linear or Decision Tree Classification).
2. If we were to do this and assign a ‘0’ to all rows- essentially assuming that all applicants were low-cost applicants – what would our accuracy level be?

13.83%

1. What is the practical value of using a baseline evaluation?

A baseline is the result of a very basic model. It provides a comparison with the more complex model we build, which tells us how good our model is compared to the baseline that is simple and sometimes just like random guessing. The exact accuracy rate is not something we need to really focus on. Instead, what is important is the difference of accuracy between the prediction model and the baseline model.

1. Assume the following…. The company is wanting to be overly cautious and assumes everyone is a high cost applicant. They do not use any sort of classification model, and instead simply assigns 1’s to all applicants. The following decisions are made accordingly after applying 1’s to all rows:
2. If the company **correctly** predicts someone is a high insurance cost, they lose $500.
3. If the company **incorrectly** predicts someone is a high insurance cost, they earn $1500.
4. What is the accuracy level of the model?

86.17%

1. What net profitability will the company see?

-$134,000