Overview of Final Project: Don Vetal

Introduction to Problem

I will be competing in a kaggle competition for All State Insurance. The premise of the competition is to predict what an individual customer will eventually purchase for insurance given the history of their scenario browsing. Each insurance purchase consists of seven different option choices. Each option has a set of discrete options a person can choose. An individual customer chooses only one of the options for each category.

The dataset consists of historical browsing snapshots from customers throughout their shopping experience. So the analyst can see how a customer is changing their choices before they make a purchase.

Approach

I will be conducting a tailored hidden markov chain model construction to predict what each individual customer will decide as their final insurance purchase options.

After some basic research on the data I have found that over the more than 550K rows of training data there are only a total of 735 unique combinations out of a possible ~2300 purchase option combinations are explored or purchased.

A hidden markov chain model will be able to effectively use the entirety of the information as it depends on analyzing the different states of an object as it moves through time. I also also believe I will be more successful if I use the entire dataset rather than only those records that are designated as the final shopping point. The time dimension embedded in this data drives the analysis.

My project will consist of two levels of prediction. One level will be to split the data into two distinct sets as I believe Cars that have a low age are probably more likely to choose different insurance options than higher age cars. This split will be conducted via a tree classifier.

LEVEL 2: Hidden Markov Chain Models

Step 1: From the training set I will generate an array of starting probabilities for each unique combination of insurance options for A through G option categories.

StartingMatrix = ( P(1), P(2), … , P(n) )

Step 2: Markov Chain prediction will also require a clear understanding of the probability of moving from one state to another. I will generate a 2 dimensional array that specifies the probability of transitioning from one state to another. The array will follow the pattern described below.

TransitionMatrix = ( P(1 ->1), P(1 ->2), … ,P(1 ->n) ) ,

( P(2 ->1), P(2 ->2), … ,P(2 ->n) ) ,

( P(m ->1), P(2 ->2), … ,P(m ->n) )

Step 3: I will use the scikit learn package; particularly the hidden markov model (hmm) part of the package to generate a model based upon the historical training data.

Step 4: I will predict the final purchase decision for each customer in the test set based upon the model.

KEY CHALLENGE:

One of the critical things I will need to address as this project proceeds is how I will manage the over 700 states. I may need to bin the states into groups to deal with the large number of states.