Assignment 3

Download the Homesite Quote Conversion Challenge data from Kaggle or from my dropbox.

The data is available at these locations:

Kaggle: https://www.kaggle.com/c/homesite-quote-conversion/data

(You can read more about this data on the Kaggle competition data page linked above.)

Please turn in a Word document with brief answers to Blackboard.

Decompress the contest data and read it into Python, R, or SAS/SAS Enterprise Miner. Your choice.

1.) **(2 pts.)** Engineer features for the day, month, and year of the original quote date.

It plausible that the day of the week, the month, or the year in which someone is first offered an insurance quote will have an effect on whether they accept the offered quote. Use the Original_Quote_Date variable to engineer three new features: quote_day, quote_month, and quote_year.

- quote_day: numeric, range from 1-7, with Sunday=1, Monday=2, and so on to Saturday=7
- quote_month: numeric, range from 1-12, with January=1, February=2, and so on to December=12
- quote_year: numeric, the four-digit year of original_quote_date

What are the values of quote_day, quote_month, and quote_year for QuoteNumber 269253 in the training set?

What are the values of quote_day, quote_month, and quote_year for QuoteNumber 213517 in the test set?

2.) **(2 pts.)** Partition the data, impute missing values, and create missing marker features.

- Randomly split the data set into 70%/30% training/validation partitions.
- Create one new numeric variable for each numeric variable with missing values, in this new variable impute missing numeric values with the mean value of the original variable.
- Create one new categorical variable for each categorical variable with missing values, in this new variable impute missing categorical values with the mode (most common) value of the original variable.
- Create one new numeric variable for each variable in the original data with missing values. These variables should be 0 when the analogous original variable is not missing and 1 when the analogous original variable is missing.
- Remove the original variables with missing values from the analysis.

What is the mode of the new imputed variable for GeographicField63 in the training set?

What is the mean of the new imputed variable for PersonalField84 in the validation set?

3.) **(4 pts.)** Build a benchmark main effects logistic regression model.

- Treat QuoteConversion_Flag as a binary target variable.
- Treat QuoteNumber as an ID variable – don't include it in the model.
- Treat all character variables as nominal or factor variables.
- Treat the new binary missing marker variables as nominal or factor variables.
- Treat all remaining numeric variables as numeric or interval variables.
- Reject Original_Quote_Date from the analysis.
- Use a 70-30 train-validation split of the original training set.
- Do not use any model selection. (If you are using a penalized regression technique, you must set all penalty tuning parameters so that no variables are excluded from the model.)

Explain the parameter estimate of the new imputed variable for PersonalField84 in terms of its odds ratio.

Explain the parameter estimate of the missing marker variable for PersonalField84 in terms of its odds ratio.

What is the lift for the top 5% of customers in the validation set – and what does this mean?

What is the posterior probability (or probability predicted by the model) for quote number 213517 in the test data – and what does this mean?

4.) **(2 pts.)** Beat the benchmark.

Use any combination of tools or techniques we have discussed so for logistic regression (significance-based model selection, penalized model selection, cross-validation, SAS, R, Python, or H2O) to beat the benchmark of 0.0835 misclassification on a 30% validation sample.

Turn in a screenshot of the model output that clearly shows the size of the validation set and the misclassification rate. Briefly explain your approach.