ACSG-555: Data Mining

Project 1

The data set for this assignment is called *Project 1 Data Set*, and is available in Canvas. It is used to investigate churn rates for a phone company.

Include an Executive Summary (at the beginning) of your most salient findings.

Explain all steps and results clearly and cogently, in a MS Word document, so that a reasonably intelligent though statistically naïve manager could understand it. *You need to include all graphics in your report.* Your narrative should be clear and concise, accompanied by supporting evidence in the form of graphics and tables. Please repeat the instructions in your report.

You will need to submit your work in Canvas.

A telephone company is interested in determining which customer characteristics are useful for predicting churn, customers who will leave their service.

Your task is to uncover patterns in the customer data that will help the company identify which types of customers are most (least) likely to churn.

You will not be able to complete this assignment in a weekend. I recommend that you "move in" with this assignment, and get really into it.
I encourage you to go beyond the letter of the assignment, but only after you have covered all of the required stuff.

Unfortunately, there is no data documentation available. The fields are as follows:

State	discrete.
account length	continuous.
area code	continuous.
phone number	discrete.
international plan	discrete.
voice mail plan	discrete.
number vmail messages	continuous.
total day minutes	continuous.
total day calls	continuous.
total day charge	continuous.
total eve minutes	continuous.
total eve calls	continuous.
total eve charge	continuous.
total night minutes	continuous.
total night calls	continuous.
total night charge	continuous.
total intl minutes	continuous.
total intl calls	continuous.
total intl charge	continuous.
number customer service calls	continuous.
Churn	Discrete

1. Perform data preparation on the data set, if needed. Give evidence that there are no problems with data quality or missing data.
   - **a.** Which variables show anomalous behavior? How shall we deal with this?
   - **b.** Which field will yield no usable statistical or graphical information, as a surrogate for the ID field?

2. Examine the variables graphically.
   - **a.** For two interesting categorical variables, construct a distribution of the variable. Comment on each.
   - **b.** Examine the distribution of all numeric variables, using histograms. (Need not include in report.)
   - **c.** Make a little table listing (in alphabetic order) the variables which are not normally distributed, along with the transformation function needed to induce normality (e.g. log).
   - **d.** For each variable, perform the transformation to induce approximate normality. Provide before / after histograms for all such variables.

3. Examine the variables statistically.
   - **a.** For all the numeric variables, find the mean, median, standard deviation, min and max. Put the results in a table, with the variables in alphabetical order.
   - **b.** Normalize all the numeric variables, using either (i) z-scores, or, (ii) min-max normalization.

4. Relationships between variables.
   - **a.** Plot Day Mins vs. Day Charge. Comment. How shall we deal with this?
   - **b.** Construct a scatter plot between any two numeric variables that you find interesting (not those in (a)). Comment.
   - **c.** Report any high correlations between any two variables. What would be the effect of keeping two highly correlated variables in the model? What should be done? Do so.

5. Data Manipulation.
   - **a.** Derive two interesting new variables to assist in the analysis. Use functions of the other variables. One derivation may be trivial (e.g., a sum), but the second should be more sophisticated (using at least two operations). Why do you find these new variables interesting and useful? Should they replace other variables or not? Provide graphics and descriptive statistics that describe the behavior of your new variables. Discuss.
   - **b.** Apart from churners/non-churners, are there any interesting subsets of records to be selected for special attention? Selecting and analyzing them may increase the precision of your analysis for important subsets of customers. Why do you find this subset of the data interesting and useful? Provide graphics and descriptive statistics that describe the behavior of your subset. Discuss.
   - **c.** Discretize (make categorical) a relevant numeric variable which you think will be explicatory of churn. This can be done using histograms.

6. With a view to uncovering customer churn patterns, investigate how each relevant variable is associated with Churn.
   - **a.** For each relevant categorical variable, construct a distribution of the variable with a churn overlay. You may wish to normalize to increase contrast. Comment on each.
   - **b.** For each relevant numerical variable, construct a histogram of the variable with a churn overlay. You may wish to normalize to increase contrast. Comment on each.
   - **c.** For each relevant categorical variable, construct a table of each relevant variable against churn.
   - **d.** For the two variables in 5a above, construct a distribution or histogram of the variable with a churn overlay. Comment on the usefulness of your new variable in the light of this evidence.
   - **e.** For the subset of records in 5b, compare the proportion of churners in this subset with the proportion of churners not in the subset (Note: Don't compare against the entire data set, which includes your subset). Discuss.
   - **f.** Discuss your results clearly, cogently, and completely.

7. Find a pair of numeric variables which are interesting with respect to churn. That is, for a pair of variables, construct a scatter plot with a churn overlay. If things look uniform, then this is not particularly interesting. We are looking for differences within the scatter plot (churn vs. non-churn), which can help us understand the relationship between the two variables with churn. Now, if there seems to be a horizontal or vertical differentiation, then this is not bivariately interesting, as the churn behavior is altering only along one of the axes. We want to find churn behavior changing simultaneously along both axes.