H:\Risk Measurement Tools Course\Website\ncsu_logo.gif

**BUS 443: Business Analytics**

**Data Mining Techniques, Examples, and Case**

Data Mining is a popular technique in the category of Predictive Analytics. Regardless of the method chosen by the data analyst, several steps must occur. The analyst must extract a sample of data that is relevant to the business problem under consideration and prepare (manipulate) the data to put it in a form suitable for formal modeling. Once the data is clean, the analyst may construct the model, using an appropriate data-mining technique (regression, classification trees, *k*-means) to accomplish the desired data-mining task (prediction, classification, clustering, etc.). Lastly, the analyst must assess the model by evaluating it against other models and checking its performance on the data set.

**Data Cleansing**

The data analyst must clean the data before using any data mining technique.

* Missing data must be determined or removed (by deleting the observations (records or rows))
* Categorical variables must be converted to binary (dummy) variables
* Erroneous data and outliers must be investigated

**UnSupervised Learning Methods**

In unsupervised methods of data mining there is no outcome variable to predict; rather, the goal is to use the variable values to identify relationships between observations, In short, these methods are used to discover patterns in data. When observations include continuous variables, **Euclidean distance** is the most common method to measure similarity between observations. We will cover two unsupervised learning methods: cluster analysis and association rules.

**Cluster Analysis:**

Cluster analysis is often used to divide observations into different homogeneous groups. The goal of Cluster Analysis is to discover patterns in a massive set of data! The underlying premise is that we as analysts do not know what the patterns are. Market researchers do this frequently for market segmentation to identify clear segments to target specific advertising campaigns.

Cluster Analysis is not a hypothesis-testing procedure like other analytical methods. Cluster analysis requires an asset of variables over which to build the clusters. We need a way to ***measure the similarity*** between the observations in the clusters. We use cluster observations to build a typology for types of cases (aka groups or clusters). The number of typologies may be large.

Objective 1: Find intra-cluster (within) similarities that are as alike as possible (homogenous)

Objective 2: Find inter-clusters (between) differences that are as different as possible (heterogeneous)

We have already studied how to use correlation to find similarity between variables. In a way, clustering is like this, but it discovers patterns of similarity among observations across several variables. Correlation is too limited for cluster analysis. In correlation, the bigger the number (e.g., r = .99), the more **similar** the variables are. In clustering, the bigger the number (e.g., distance between observations), the more **dissimilar** they are. Thus, we use Euclidian geometry to measure these distances. The straightest distance between two points is a straight line. **Squared Euclidian distance** has its roots in the Pythagorean triangle. To find the distance between any two points you take the difference on the vertical axis, square it, and add it to the squared difference on the horizontal axis. That gives you the squared length of the distance of two points, which turns out to the hypotenuse of a right triangle. You can use it to measure the distance between two respondents on any number of variables.

**Cluster Analysis Algorithms:**

Now, we need an algorithm to search through the numbers. There are many from which to pick.

**Hierarchical:** This approach is bottom-up and good for smaller data sets. There are many from which to pick, but a popular one is Ward’s Minimum Variance. Start with a sample size of N and gradually cluster them together until you end up with one big cluster at the end. You do not know the number of clusters at the beginning. The advantage of this method is that the solution will detect the number of clusters. The disadvantage is that once observations are placed in a cluster, they never move, even if they belong somewhere else later. Be sure to look at a tree diagram (dendrogram) to graphically view the iterations and determine the optimal number of clusters. Examples of this using XLMiner are found in Chapter 6, Essentials of Business AnalyticsProblems 2 and 4.

**Non-Hierarchical.** This approach istop-down; good for large data sets; the analyst must specify the number of clusters, k.) (most popular is **k-Means**): You pick the number of clusters first, then the algorithm (through iterations) determines how close each respondent is to each cluster. After all observations have been assigned to a cluster, the resulting cluster centroids are calculated (these cluster centroids are the “means” of k-means clustering). Advantage: Things are allowed to iterate; Disadvantage: You must pick the number of clusters first. An example of this using XLMiner is found in Chapter 6, Essentials of Business AnalyticsProblem 6.

**Analysis of Cluster Analysis Results:**

We are now ready to analyze of the nature of the clusters and attempt to determine business value from them. We use statistical techniques to compare the clusters over the variables used to construct them. For example, in one cluster we may have customers that purchase high amounts of a business’ products, with another cluster having the buyers of few sales. We draw a picture to show this since this shows an internal validation of the cluster. Later, we can use external validation measures, which allow us to compare these to other variables not used in the algorithm.

*Suggested Resource: Emeritus Professor Ray Cooksey (UNE) has a YouTube channel where he discusses various types of data mining, including cluster analysis algorithms. I encourage you to check out his work* [*here*](https://www.youtube.com/watch?v=PBKxZkOzGxY)*.*

**Association Rules:**

If-Then statements which convey the likelihood of certain items being purchased together. An example of this is found in Chapter 6, Problem 8.

**Supervised Learning Methods**

The goal of a supervised learning technique is to predict! It is to develop a model that predicts a value for a continuous outcome or classifies a categorical outcome.

**Partitioning:**

Partition the data first into three sections (training set, verification set, test set)

* **Standard Partition** (use when the classification doesn’t involve a rare event
* **Partition with Oversampling** (use when the classification involves a rare event. The validation set and test set are formed to have the approximate % of rare observations in order to be representative of the overall population

**Classification Methods & Accuracy:**

* K Nearest Neighbors (kNN) (Example: Chapter 6, Problem 10)
* Classification Trees (Example: Chapter 6, Problem 14)
* Logistic Regression (Example: Chapter 6, Problem 18)
* Others you may use in industry
  1. Naïve Bayes
  2. Discriminant Analysis
  3. Neural Network

With each type of method, your model must be checked for checked for errors and usually rerun multiple times until you find the model with the least errors. Use the classification Confusion Matrix which displays a model’s correct and incorrect classifications and lift charts.

**Prediction Methods & Accuracy:**

* K Nearest Neighbors (kNN)
* Multiple Linear Regression
* Regression Tree
* Neural Network

There are several ways to measure accuracy when estimating a continuous outcome variable, but the two used in XLMiner are the average error and the root mean squared error (RMSE).

**Scoring New Data:**

After you have selected your best model, your new data can be ***scored*** (classified or predicted). In XLMiner, you may use the **Score** tool on the XLMiner toolbar or just score it inside the last run of the model.

**Data Mining Class Example**

**Clustering with KTC**

The goal of clustering is to segment observations into similar groups based on the observed variables. Cluster analysis is commonly used in marketing to divide consumers into different homogeneous groups, a process known as market segmentation.

We will use a company called KTC, a financial advising company that provides personalized financial advice to its clients. As a basis for developing this tailored advising, KTC would like to segment its customers into several groups (clusters) so that the customers within a group are similar with respect to key characteristics and are dissimilar to customers not in the group. We will look at age, gender, annual income, marital status, number of children, car loan, and home mortgage as variables.

1. Open the KTC-Small file and perform hierarchical clustering. Use the following variables: Female, Married, Car Loan, and Mortgage. Select Jaccard’s Coefficient method to measure similarity between observations and McQuitty’s method to measure similarity between clusters. (McQuitty’s method is a hierarchical clustering method.)
2. Choose 4 clusters. Copy the results from the HC\_Clusters, including the labels (B8:G38) sheet and paste into the Data sheet. Sort by the Cluster ID column.
3. Review the dendrogram. Remember that a dendrogram is a chart that depicts the set of nested clusters resulting at each step of aggregation. Review the composition of four clusters (with the horizontal line) from the dendrogram. Cluster 1 (only one observation, 1), is a single female with no car loan and no mortgage. Characterize the other three clusters.

Cluster 2: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Cluster 3: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Cluster 4: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

These clusters segment KTC’s customers into four groups that could possible indicate varying levels of responsibility, an important factor to consider when providing financial advice.

1. Now, perform k-Means clustering on this data set using a k of 3 over variables age, income, and children. Select the checkbox for Normalize input data. Enter 3 in the # Clusters box and 50 in the Iterations box. In the Options area choose Random Starts. Increase the No. of Starts to 10. Click Finish.
2. Analyze the results. View the Cluster Centers on the KM\_Output sheet. Notice that Cluster 1 consists of the youngest customers with the largest families and low incomes. How would you characterize the other two clusters?

Cluster 2 consists of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_;

Cluster 3 consists of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

If KTC decides these clusters are appropriate, they can use them as a basis for creating financial advising plans based on the characteristics of each cluster.

The second table under **Cluster centers** displays the between-cluster distances between the three cluster centers. The **Data summary** table displays the within-cluster distances and reveals that Cluster 1 is more homogeneous. By comparing the between-cluster distance in the second table under **Cluster centers** to the within-cluster distance in **Data summary** table, we observe that the observations within clusters are much more similar than the observations between clusters.

**Data Mining Case**

**A Collection of Data Mining Problems – Due 10/29/15**

**PART 1: DATA MINING TECHNIQUES TO FIND “PATTERNS” – UNSUPERVISED LEARNING**

**Problem 1: Hierarchical Cluster Analysis with the Football Bowl Subdivision (FBS)**

We started this example in class and will now do some further analysis. The Football Bowl Subdivision (FBS) of the National Collegiate Athletic Association (NCAA) consists of over 100 schools. Most of these schools belong to one of several conferences, or collections of schools, that compete with each other on a regular basis in collegiate sports. Suppose the NCAA has commissioned a study that will propose the formation of conferences based on the similarities of the constituent schools.

1. Open the FBS file (found in the Chapter 6 textbook files) that contains rows of information on constituent FBS schools. Apply hierarchical clustering with **10** clusters using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Use **Ward’s method** as the clustering algorithm. Be sure to **normalize** the data. Copy the assigned cluster column to the data sheet.
2. Use a Pivot Table on the data in the HC\_Clusters sheet to identify the cluster with the largest ***average*** football stadium capacity. Which cluster and school have the highest?

Stanford is the outlier (cluster 10)

Cluster 3 and Tennessee Volunteer has the highest AVG football stadium capacity

1. How would you characterize the universities in this cluster?

Universities in this cluster has very high stadium capacity. 12 out of 15 of them have capacity over 80,000. Tennessee Volunteers, Alabama Crimson, and Texas Longhorns are the top 3 highest capacity in this cluster. This cluster is also ranked the 4th in highest average endowment compare to the other clusters.

Most schools in cluster 3 are on 30-35° latitude and -118°- -82° longitude. Their endowment and enrollment vary widely.

1. What is the smallest cluster (the one with the fewest observations) and what makes it unique?

Cluster 10 is the smallest cluster with only 1 observation (Stanford Cardinal). This cluster is unique because it is the outliner and because its average endowment is double or even treble higher than other clusters, which is $16,502,606, raking the 1st highest in AVG endowment, while the 2nd rank is only 5,651,058. Also because it is not geographically located near any of the other schools.

1. Examine the dendrogram on the HC\_Dendrogram worksheet (as well as the sequence of clustering stages in the HC\_Output sheet). What number of clusters seems to be the most natural fit based on the distance?

According to the dendrogram, eight clusters seems to be the most natural fit based on the distance (not too dense, not too broad). This would be dropping cluster 10 (the one with only 1 school by itself) and cluster 7 (the second smallest cluster, with only 4 schools), and merging them into the larger clusters.

1. Create another pivot table and count the number of schools per cluster. Analyze the results. Why aren’t these cluster results appropriate, or (restated) why should we rerun the cluster analysis using different variables or a different number of clusters?

Overall, the number of schools are distributed quite evenly across the clusters. Except cluster 5 is quite bigger and cluster 10 is way too small with only 1 school. Especially cluster 10, I assume it’s out by itself because it’s not geographically located near any of the other schools. This location is in south Wyoming, and the closest location (according to the dendrogram). I think we should rerun because the abnormal number of school in cluster 5 and 10. It would be better, I think, if we have Athletic Revenue variables

1. Apply hierarchical clustering again with **10** clusters using just latitude and longitude as the variables. Be sure to **normalize** the data and specify **single linkage** as the clustering method. Use a Pivot Table on the data in HC\_Clusters. You can also visualize the clusters with a scatter plot with longitude as the x-variable and latitude as the y-variable. Compare the clusters to the previous method. Which is the better method?

There are a lot more single item clusters when using this method as comparted to the Ward’s Method. Also, cluster 2 is abnormally large (with 98 observations) compared to the other clusters, and there are even 4 cluster have only 1 observation by itself (cluster 4,6,9,10). The other method was more even all around, so I would go with the Ward’s Method for this case.

**Problem 2: k-Means Cluster Analysis with the Football Bowl Subdivision (FBS)**

1. Open the FBS file used in Problem 1 and copy the data to a new workbook. Delete the cluster column from the hierarchical clustering in Problem 1.
2. Apply k-Means clustering with k=10 using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Specify 50 iterations and 10 random starts and normalize the data.
3. Analyze the resultant clusters. What is the smallest cluster (the one with the fewest observations)?

**Clusters analysis:**

**Distance between centers**:

* Cluster 5 and 10 are the most similar (close) to each other with the smallest distance among other clusters: 32,461 (highlighted in pink in the excel file). These two clusters also have pretty close average longitudes.
* Cluster 6 and 1: highest distance number 7571512, which mean they are the most distinct from each other. (highlighted in red)

**Stadium Capacity**: Cluster 9 has the highest stadium capacity: 80072, while cluster 4 is the lowest rank with only 28846 in stadium capacity.

**Smallest cluster**: cluster 8. There is only one observation in the cluster which makes it unique from others

**Largest cluster:** cluster 5 and 2 with twenty-two observations.

1. What is the least dense (aka most diverse) cluster, as measured by the largest average distance in the cluster? What makes the least dense cluster so diverse?

According to the data summary:

Cluster 6 has the largest average distance between observations within the cluster, therefore, it is the least dense cluster (or the most diverse cluster). Diversity in a cluster caused by the lack of similarity between the observations. Least dense means that this cluster has the widest avg. distance within the cluster itself.

The lack of similarity (difference) can be easily seen through the variables analysis:

* Stadium Capacity: the range of capacity is wide. Smallest number is 30964, biggest number: 56000. Different values in this variable contributes to the diverse in cluster.
* Latitude: the values in this variable are quite diverse but not as much as other variables. Most schools are on 37◦-39◦ latitude. The range of latitudes in this cluster seem a bit larger than other clusters.
* Longitude: the range of longitudes in this cluster is larger than other clusters as well.
* The wider the range of latitude or longitude is, the further the schools are located. (Distance from A school to B school in this cluster will be further than in other clusters). If these schools are spotted on the map, we can see that they stay a bit far away from each other than in other clusters. That’s why we call it less dense.
* Enrollment: enrollment also has a contribution in cluster’s density. The enrollment value in this cluster is diverse in each observation.

Cluster 5 and 2 have the most observations, but that doesn’t make them most diverse.

If I were the administrator, I would want the average distance within a cluster to be small (more similar between observations). Larger number mean the cluster is more broad/diverse.

The Summary of cluster 10 is copied below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster ID** | **Stadium Capacity** | **Latitude** | **Longitude** | **Endowment ($000)** | **Enrollment** |
| 10 | 52480 | 38.8632 | -104.759899 | 56600 | 4413 |
| 10 | 40000 | 41.362343 | -74.027317 | 73190 | 4624 |
| 10 | 38016 | 38.41295 | -82.433767 | 83810 | 13966 |
| 10 | 35542 | 36.12775 | -95.916407 | 800925 | 4092 |
| 10 | 34000 | 38.97165 | -76.503033 | 89780 | 4576 |
| 10 | 31500 | 36.1021 | -80.26291 | 1058250 | 7351 |
| 10 | 30964 | 35.821828 | -90.685768 | 39479 | 13900 |
| 10 | 56000 | 38.22475 | -85.741156 | 772157 | 21153 |
| 10 | 50000 | 39.1901 | -96.589981 | 337460 | 23863 |
| 10 | 39790 | 36.17155 | -86.784829 | 3414514 | 12836 |

1. What problems do you see with the plan of defining the school membership of the 10 conferences directly with these 10 clusters?

The number of schools in each cluster varies from each other. For instance, cluster 8 only has 1 school, while cluster 5 and 2 have 22 schools. This will cause a time and management issue. Because it would be much easier and less time consuming to set up membership for just 1 school; on the other hand, it would take more time and more work to deal with 22 schools, people will have to get in line and wait for it..

The number of schools in each cluster needs to be even so the schools can play the same number of teams.

**Problem 3: Both Types of Cluster Analysis with the Football Bowl Subdivision (FBS)**

The NCAA has a preference for conferences consisting of similar schools with respect to their endowment, enrollment, and football stadium size, but these conferences must be in the same geographic region to reduce traveling costs. Take the following steps to address this desire.

1. Apply k-means clustering again (in a new worksheet) using latitude and longitude as variables with k=3. Be sure to normalize and specific 50 iterations and 10 random starts. Then create one distinct data set (one spreadsheet) for each of the three regional clusters (east, west, and south).
2. For the west cluster, apply hierarchical clustering with Ward’s method and use normalized data to form two sub-clusters using football stadium capacity, endowment, and enrollment as variables. Use a PivotTable on the data in HC\_Clusters to report the characteristics of each cluster.
3. Do the same for the east cluster, using three sub-clusters.
4. Do the same for the south cluster, using four sub-clusters.
5. What problems do you see with this plan? How could this approach be tweaked to solve the problem?

**Problem 4: Market Basket Analysis on Cookie Monster, Inc. (Problem 8 in our Textbook)**

Cookie Monster Inc. is a company that specializes in the development of software that tracks Web browsing history of individuals.

1. Open the CookieMonster file and review the binary matrix format. The entry in row and column indicates whether the column website was visited by the row user. Using a minimum support of 800 transactions and a minimum confidence of 50%, use XLMiner to generate a list of association rules.
2. Review the top 14 rules. What information does this analysis provide Cookie Monster regarding the online behavior of individuals? Be sure to address the lift ratios (and the meaning of the lift ratios) in common terms that a business user would immediately understand.

Use lift ratio. Then translate into business language: use the work likelihood, percentage, suggestions /recommendation?

**PART 2: DATA MINING TECHNIQUES TO “PREDICT” – SUPERVISED LEARNING**

**Problem 5: k-Nearest Neighbors Data Mining for Finding Undecided Voters for Campaign Organizers** Read the description of this problem (# 10) of our textbook. Complete this problem and check your answers against the solution provided on our textbook website.

**Problem 6: Logistic Regression to Predict the Oscars**

Read the description of this problem on p. 316 (#18) of our textbook. A description of logistic regression is found on page 299. Use the Oscars data to create a logistic regression equation to predict whether a movie will win the Best Picture Oscar Award using information on the total number of Oscar nominations that a movie receives and whether the movie won the previous Golden Globe award (1 = movie won; 0 = movie lost).

1. Partition all of the data using the ChronoPartition variable.
2. Construct the model using winner as the output variable and Oscar Nominations, Golden Globe Wins, and Comedy as the input variables. What is the resulting logistic regression calculation?
3. What is the overall error rate on the validation data?
4. Use the model to score the new data (2011). Which movie did the model select as the most likely to win the 2011 Best Picture Award?

**Problem 7: Logistic Regression to Predict the Organic Customers using SAS Visual Statistics**

Access Visual Analytics from the Teradata University Network (TUN) site. Open the ORGANICS\_VISTAT data set and use this data to create a logistic regression equation to predict which customers will buy organic foods.

1. Create a boxplot that shows affluence grade and age by organics purchase indicator, just as you did for one of the mini-cases on the midterm exam.
2. Click the Logistic Regression tool to begin using Visual Statistics.
3. Construct the model using **Organics Purchase** as the output variable and age, gender, and recent purchase variables as the input variables.
4. Remove any variables that are not statistically significant. What is the resulting logistic regression calculation?
5. What is the overall r square for the model?
6. Use the assessment plots to determine the effectiveness of the model.
   1. Look at the **Lift Chart**, which measures the *model’s* *effectiveness*. A lift chart is a graphical representation of the advantage (or lift) of using a predictive model to improve on the target response vs. not using a model. It is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained *with* and *without* the predictive model. When the lift in the lower percentiles of the chart is higher, the model is better.

The chart shows two lines: one that represents the model you built; and one that represents the best, achievable model, or a perfect classifier. When the Model line is closer to the Best line, especially in the lower percentiles, the model is better.

Restated, lift is the ratio of the percentage of captured events within each percentile bin to the average percentage of responses for the model. Cumulative lift is calculated by using all the data up to and including the current percentile bin. In this example, what is cumulative lift at the 20th percentile? Is this value low? If so, a low value indicates additional variables or interaction effects should be considered to improve the model. This lift value means that if the supermarket chain sent coupons to the top 20 percent of customers selected by this model, it could expect to see \_\_\_\_\_\_\_\_\_\_ times more customers purchasing organic products than if the same number of customers were randomly selected.

* 1. Assess the **ROC (Receiver Operating Characteristic) Chart**, whichmeasures *classification* or predictive *accuracy of a logistic model*. The classification accuracy of a model is demonstrated by the degree that the ROC curve pushes upward and to the left. This degree can be quantified by the area under the curve. The area ranges from 50 (for a worthless model) to 100 (for a perfect classifier).

Restated, a ROC chart displays the ability of a model to avoid false positive and false negative classifications. A false positive classification means that an observation was identified as an event when it was actually a nonevent (aka a Type I error). A false negative classification means that an observation was identified as a nonevent when it is actually an event (aka a Type II error). The specificity of a model is the true negative rate. To derive the false positive rate, subtract the specificity from 1. The false positive rate, labeled 1-Specificity is the X-axis of the ROC chart. The sensitivity of a model is the true positive rate (Y-axis) A good ROC chart has a steep initial slope and levels off quickly.

What does the ROC chart suggest?

* 1. Assess the **Misclassification Chart**, which displays how many observations were correctly and incorrectly classified for each value of the response variable. In this case, the misclassification plot displays how many observations were correctly and incorrectly classified as *bought* (positive) or *did not buy* (negative) organic products. How many customers were classified as false positives? Should this model be refined more in light of this?

1. Click on the **Parameter Estimates** tab in the summary table. Click on the **z Value** column two times to sort descending. Which variables have a high influence on predicting whether a customer will buy organic food?

**Problem 8: Logistic Regression to Predict the PVA Donors using SAS Visual Statistics**

Access Visual Analytics from the Teradata University Network (TUN) site. Open the PVA\_DATA data set and use this data to create a logistic regression equation to predict who is likely to donate to the PVA.

1. Create a boxplot that shows last gift amount and age by donor indicator.
2. Click the Logistic Regression tool to begin using Visual Statistics.
3. Construct the logistic regression model using **Donation** as the response variable. Select the Advanced link and make sure that the event level is set to Donated.
4. Change pep\_star from a measure to a category.
5. Select file\_avg\_gift, file\_card\_gift, home\_value, house\_income, last\_gift\_amount, lifetime\_avg\_gift\_amt, lifetime\_gift\_count, lifetime\_gift\_range, lifetime\_max\_gift\_amt, lifetime\_min\_gift\_amt, lifetime\_prom, months\_since\_first\_gift, months\_since\_last\_gift, number\_prom\_12, age, card\_prom\_12 as continuous effects.

Select frequency\_status\_97nk, gender, home\_owner, in\_house, income\_group, overlay\_source, recency\_status\_96nk, and pep\_star as classification effects.

1. On the Properties tab, select **Information missingness** and **Use variable selection**.
2. Remove any variables that are not statistically significant.
3. Remove some of the outliers via the Residual plot.
4. View the response profile tab on the Summary table.
5. What is the resulting logistic regression calculation?
6. Check the various assessment charts and comment on the usefulness of the model.