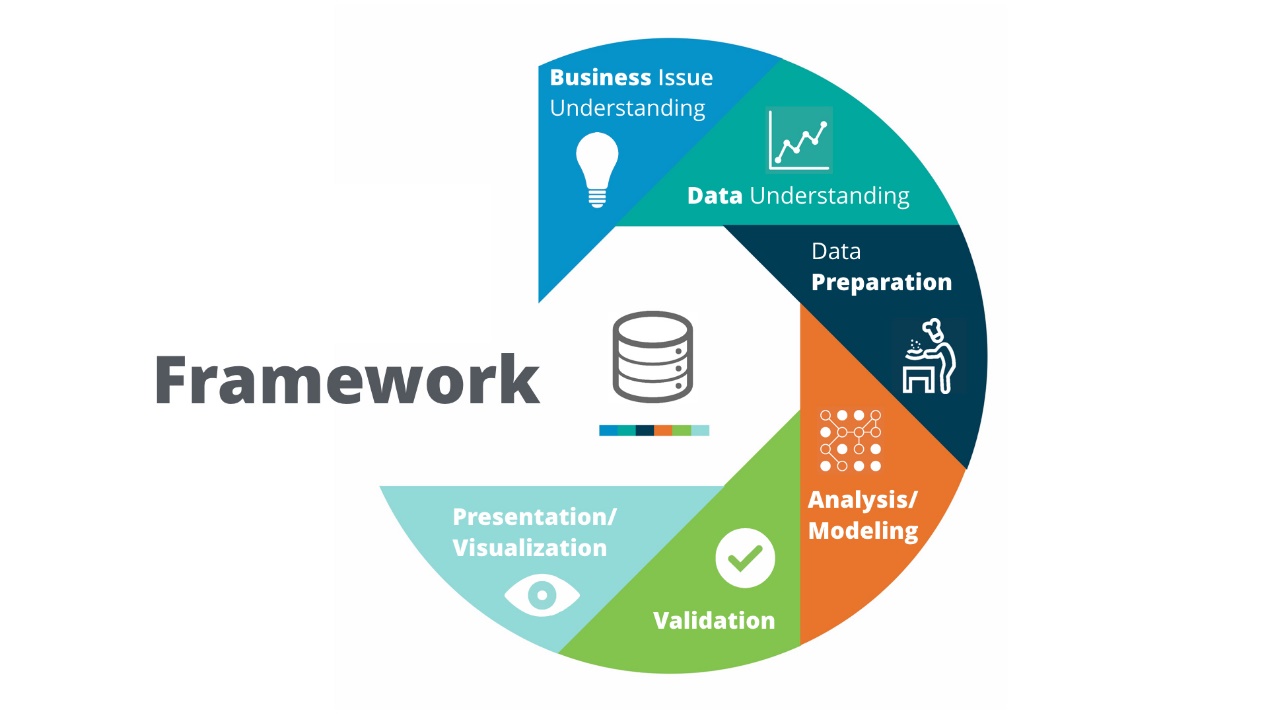
1. Cross Industry Standard Process for Data Mining (CRISP-DM)

This framework was originally developed by data miners in order to generalize the common approaches to defining and analyzing a problem.

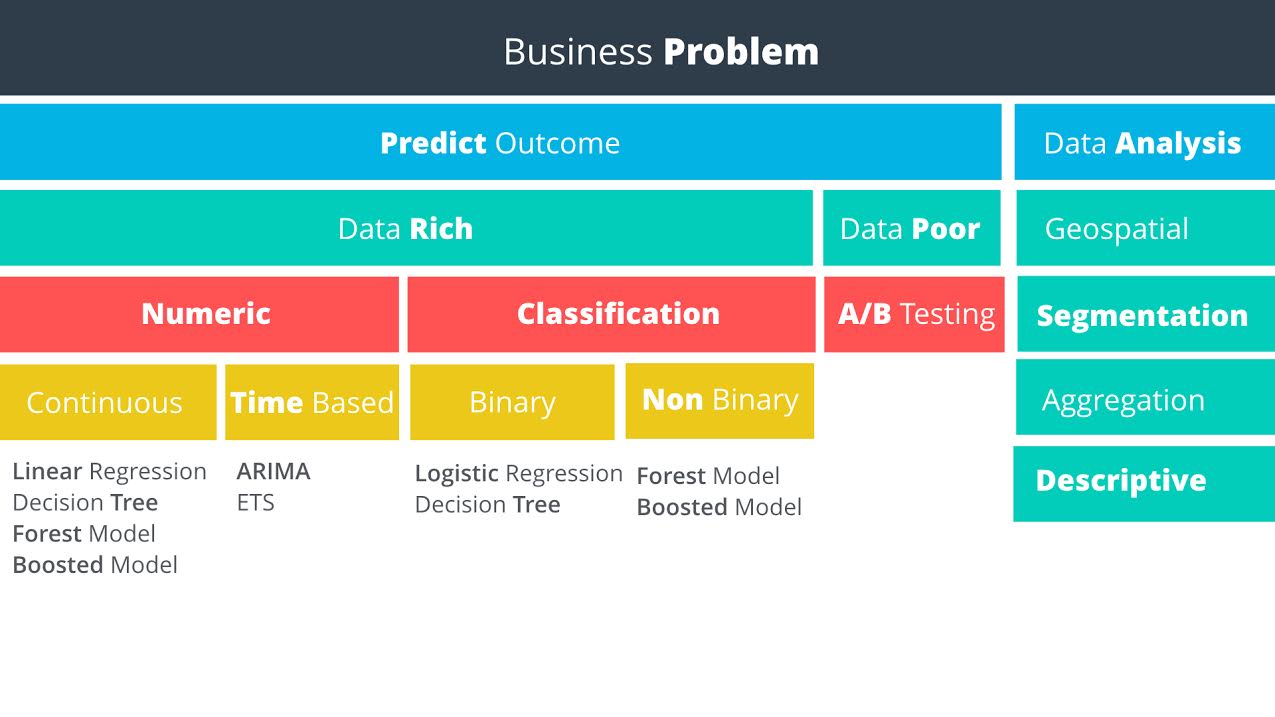


**Methodology Map**

The methodology map is a guide to determine the appropriate analytical technique(s) to solve a particular business question or problem.

The map outlines two main scenarios for a business problem:

1. Data analysis
2. Predictive analysis



#### Predictive

Predictive analytics uses existing data to predict a future outcome. For example, a company may use predictive analytics to forecast demand or whether a customer will respond to an advertising campaign.

#### Geospatial

This type of analysis uses location based data to help drive your conclusions. Some examples are:

 Identifying customers by a geographic dimension such as zip code, state, or county, or

 Calculating the distance between addresses and your stores, or

 Creating a trade area based upon your customer locations for further analysis

Some types of Geospatial analysis require the use of special software - such as software that can convert an address to Latitude & Longitude, or can calculate the drive time between two geographic points on a map.

#### Segmentation

Segmentation is the process of grouping data together. Groups can be simple, such as customers who have purchased different items, to more complex segmentation techniques where you identify stores that are similar based upon the demographics of their customers.

#### Aggregation

This methodology simply means calculating a value across a group or dimension and is commonly used in data analysis. For example, you may want to aggregate sales data for a salesperson by month - adding all of the sales closed for each month. Then, you may want to aggregate across dimensions, such as sales by month per sales territory. In this scenario, you could calculate the sales per month for each salesperson, and then add the sales per month for all salespeople in each region.

Aggregation is often done in reporting to be able to “ slice and dice” information to help managers make decisions and view performance.

#### Descriptive

Descriptive statistics provides simple summaries of a data sample. Examples could be calculating average GPA for applicants to a school, or calculating the batting average of a professional baseball player. In our electricity supply scenario, we could use descriptive statistics to calculate the average temperature per hour, per day, or per date.

Some of the commonly used descriptive statistics are Mean, Median, Mode, Standard Deviation, and Interquartile range

#### Data Rich vs. Data Poor

Do you have data on what you are trying to predict? If so, you can proceed down the data rich path, otherwise, the data poor path is your only option. See the following example that demonstrates a data poor scenario.

#### Types of Numeric Variables

The three most common types of numeric variables are continuous, time-based, and count.

#### Continuous

A continuous variable is one that can take on all values in a range. For instance your height can be measured down to many decimal places. We do not grow in even inch intervals.

#### Time-Based

A time-based numeric variable is one where you are trying to predict what will happen over time. This is often related to forecasting.

#### Count

Count variables are numbers that are [discrete](https://www.mathsisfun.com/data/data-discrete-continuous.html), positive integers. They’re called count numbers because they’re used to analyze variables that you can count. As modeling these type of variables is not common in business, we won’t be covering this topic in this course.

#### Linear Regression

Imagine we have the data displayed in the scatter plot. It appears that we have a linear relationship between the number of employees and the number of tickets. The relationship appears to be linear since it seems like we can draw a straight line through the data.

If we know the equation of the line, we can predict values for tickets given a certain number of employees. The simple equation for a line is:

#### y = mx + b

Y = Target Variable

X = Predictor Variable

m = Slope of the line

b = Y-intercept

##### Target Variable

The target variable is the variable we are trying to understand and predict. It is also referred to as the dependent variable. In our example, we are trying to predict Y, or the average number of tickets.

##### Predictor Variable

Predictor variables are used to try to predict the target variable and are also known as independent variables. In the example there is just one predictor variable, X, or the number of employees. It is used to predict the number of tickets based.

#### Validation

Now that we’ve performed the analysis and run the Linear Regression Model, we need to validate the results of the model. In other words, is there a way to measure how good the model is? Or in this case, is the linear expression we calculated a good fit of our data?

##### Step 1: Correlation

Using the correlation function CORREL(data\_y, data\_x), we can calculate the correlation between the target and predictor variable. This value is often referred to as r. The range of r is from -1 to +1. The closer r is to plus or minus 1, the higher the correlation between x and y. In our example, the value of r is 0.987, indicating a strong correlation.

##### Step 2: Calculate r-squared

While a strong correlation is good, we really want to know how well the data fits our line. Fortunately, we can get a sense of how good the formula is at approximating the data by calculating the coefficient of determination, or r-squared. R-squared is a coefficient between 0 and 1. R-squared is interpreted as the percent of variance in observations that is explained by the model, or the explanatory power of the model. An R-squared value close to 1 would mean that nearly all variance in the target variable is explained by the model. An R-squared value close to 0 would mean that nearly none of the variance in the target variable is explained by the model.

##### Caution about interpreting R-squared

How you interpret R-squared depends heavily on the problem you're trying to model and the data you use. For tough problems, a very low R-squared may be acceptable. Also, a high R-squared may result from a poor model. However, in general, the higher the R-squared the better, especially as you add and remove predictor variables to determine the strongest predictive model

#### Transforming Categorical Variables - Good Example

A much better way of using categorical variables in regression is to use what are called dummy variables. A dummy variable can only take on two values, generally zero or one. You would add one dummy variable for one less than the number of unique values in the categorical variable. So if the variable is binary, you'd add one dummy. If there are four categories, you'd add three dummy variables.

Going back to our example, let's now use dummy variables to represent the categorical variable, region. To represent the four categories west, midwest, northeast, and southeast, you’ll need to add three dummy variables. Let’s create a dummy variable for midwest, southeast, and west.

**Expenditures = β 0 + β1 Avg\_Income + β2 Pct\_Under\_18 + β3 midwest + β4 southeast + β5 west**

Each of the variables takes on a value of either 1 or 0. If the state is in the southeast, then the value for the southeast variable would be 1 while the other two variables would be zero.

Now we didn't create a variable for northeast. That’s because the equation needs a baseline value that is not coded into a dummy variable. If a state is in the northeast, then the value for all three of the dummy variables would be zero. *You always create one less dummy variable than the number of categories* to make sure that one category is represented by zero values for the dummy variables. That one category, in this case the northeast region, becomes the category that others are compared to.

Note: Many software tools, like Alteryx, will transform categorical variables into dummy variables automatically.

#### In Altyrex Linear Regression Model, you will see two options in model configuration:

#### Additional Options in Linear Regression Tool

##### Omit a model constant

The first “Omit a model constant” will produce a model that does not have a B0 value in the linear regression equation. If the situation that is being modeled should not have a b0 value then you would select this option. In this example b0 could be interpreted as the Average number of baseline tickets that any client would have, regardless of other variables. Since this is reasonable, we likely would include a model constant, and would not check the box. It can also be used if the model is producing a b0 value that does not make sense. In the case of this example, if we were getting a model constant that was a negative value, then it would not make sense for the situation, and we could choose to ignore it (essentially set it to 0).

##### Use a weight variable for weighted least squares

The second option, “Use a weight variable for weighted least squares” allows the user to set a weighting value to each row of data. An example of when you might want to use this would be if you wanted to weight clients who were well established more than other clients who were relatively new because you felt that the Average number of tickets would be more accurate for the established clients.

Then you could add a column of data, setting the number 2 for each established client and a 1 for each new client. This would weight the established clients twice as much in determining the equation for the linear regression. For the sake of this example, we’re not going to use weighting.

How to interpret Regression Results are found on separate PDF

# Data Preparation:

Garbage in = garbage out

Data types:

* Structured: rows and columns
* Unstructured: no structure at all
* Semi-structured: Log files, XML, …

Data Issues:

Dirty Data:

* Parsing: delimiter, csv: ‘,’ separation. 🡪 (Altryx)Text to column
* extra characters: 🡪( Alteryx): Formula, string functions
* duplicate records
* typos
* not updated (accuracy)
* incorrect data

Missing Data:

How to find NULL data?

Use summarize tool, all fields, count Nulls

Field summary tool

If too much, remove field (select), else, you can remove records (isNull)

* **Deleting** missing data is often the default method because of it's simplicity.
  + Alteryx 🡪 field summary
* **Imputation, (**mean, median, mode)
* **Multiple imputation (Advanced)**
* **Full Information Maximum Likelihood**

Outliers:

* Incorrect data vs abnormal but correct data
* To calculate the upper fence and the lower fence, here are the exact steps:

1. Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. You can use the Excel function *QUARTILE.INC or QUARTILE.EXC*

2. Calculate the Interquartile Range: *IQR = Q3 - Q1*

3. Add 1.5 \* IQR to Q3 to get the upper fence: *Upper Fence = Q3 + 1.5 \* IQR*

4. Subtract 1.5 \* IQR to Q1 to get the lower fence: *Lower Fence = Q1 - 1.5 \* IQR*

* There are additional methodologies that people use such as z-scores or standard deviations. See [here](http://www.real-statistics.com/sampling-distributions/identifying-outliers-missing-data/) to learn more about using z-scores. You can read more about violin plots [here](http://help.alteryx.com/9.5/Violin_Plot.htm).
* If abnormal, make two models, someone should decide

To know categorical data distribution, frequency table tool

Data Formatting:

Transposing (Pivot) 🡪 Alteryx: Transpose

Aggregating 🡪 Alteryx: Summarze

Crosstabulation 🡪 Alteryx: Cross Tab

Data Blending:

**Unioning**:

**Joining:**

**Fuzzy Matching**: use algorithms to develop normal joins, like compare (strings) with uncertainty into account

**Spatial Blending:**

**Types of Spatial Data**

All of these location data examples are represented by points, lines, or polygons.

**Points**

A point, also referred to as a centroid, is in the form of a latitude and longitude which we use to pinpoint its exact location.

**Lines**

A line is a string of latitudes and longitude locations.

**Polygons**

Polygons are made up of a series of longitude and latitude coordinates defining all of the vertices of a region. The drivetime boundary shown below represents how far you can drive out from a center location in 10 minutes and is made up of 2782 vertices.

Alteryx 🡪 long, lat data 🡪 create point tool

* Spatial tools

# How to choose predictors:

### Correlation Coefficients

* [Pearson Correlation Coefficient](http://wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)
* [Spearman's Rank Correlation Coefficient](http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient)
* [Hoeffiding's Independence Test](http://en.wikipedia.org/wiki/Hoeffding's_independence_test)

Alteryx 🡪 association analysis

Stepwise algorithm, connect data entered to linear regression and output of linear regression

Note that the Stepwise Regression tool is a tool to help you **reduce** and figure out which predictor variables have a good chance of being in the model, but it is not a tool that can automatically find all of the appropriate predictor variables in one run.

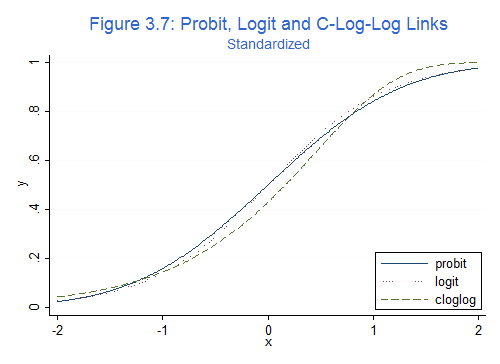
There is still a lot of work that needs to get done to explore the predictor variables that the Stepwise Regression tool gives you. The Stepwise Regression tool will speed up the process for you in choosing predictor variables.

For validation:

Model Comparison Tool

# Introduction to Logistic Regression

Logistic regression is one of the most basic forms of regression modeling. It’s part of a family of “generalized linear models” or GLM for short. This basically means that the formula is very similar to that of a linear regression. However since the target variable is binary, instead of a continuous numeric variable, the target variable has to be modified to fit this GLM formula. See the video below for more on the structure of Logistic Regression.



We use logit

# Decision Tree:

#### Important Concepts to Review

**Root Node Error**: A percentage of how many of the data points went to the incorrect terminal node (predicted incorrectly) when all of the data points are validated against themselves within the entire training set (the Estimation dataset).

**Pruning Table**: Lists out the levels in the decision tree with their related error terms with cross-validation samples.

**Confusion Matrix**: A matrix (or table) that lists out all of the possible prediction results when we validate our model against our validation set. This confusion matrix is one of the best methods to review the accuracy and precision of your model as well as to understand any model bias in classifying your data points.

After we run this Logistic Model we can see that the categorical variable Status has some classes that are not statistically significant. There a couple of things to keep in mind when something like this occurs and why in this case I kept Status in my model.

1. It is not possible to keep only part of a categorical variable
2. If NO class in a categorical variable is significant, then you may consider not using the variable at all.
3. If ANY class in a categorical variable is significant, then you should keep it.
4. A combination of classes in a categorical variable is possible with testing, but it should make logical sense though.
   * In this case Status.Gold is significant but Status.Silver & Status.Platinum are not. To create a Gold vs. Other does not make logical sense since Gold is between Silver and Platinum.

Contributing Author: Shaun Lippy

Logistic Regression

Summary: Logistic Regression is a statistical method used to predict binary outcomes by analyzing the outcome’s relationship with one or more predictor variables.

**STEP 1: SELECT TARGET AND PREDICTOR VARIABLES**

Target Variable:  The target variable is the variable we are trying to predict with the model. This should be a binary variable: yes/no, true/false, 0/1, etc.

Predictor variables: The predictor variables are used to help predict the target variable. Predictor variables should be: (1) Relevant to the target variable, (2) not highly correlated to other predictor variables, and (3), do not have a high number of missing values

Useful Alteryx tool: Association Analysis

**STEP 2: PREPARE DATA**

Preparing the data includes dealing with issues such as missing, dirty, or duplicate data; removing outliers; blending and formatting data, etc. Your final dataset should include one row for each outcome and set of predictor variables.

Estimation and validation samples: Next, split the data set into two parts:  one part for Estimation (for training the model) and one part for Validation (to help us verify that we are creating a useful model).

Useful Alteryx tool: Create Samples

**STEP 3: BUILD AND RUN THE MODEL**

Run the model with the target and predictor variables. Observe the statistical significance of each of the predictor variables by looking at the p-value in the output. If it’s below 0.05, then the relationship between the target and predictor variable is statistically significant. If not, it is not significant and can be excluded from the model. R-squared is an estimate between 0 and 1 of the explanatory power of them model, and can be used to compare models and select the best one.

Using a technique called “stepwise regression” can automatically identify the best combination of predictor variables.

Useful Alteryx tools: Linear Regression, Stepwise

**STEP 4: MODEL VALIDATION**

Apply the model to the validation sample and observe how accurately the model predicts the outcomes. This step helps avoid overfitting and helps you understand how accurate your predictions will be on new data.

Useful Alteryx tool: Model Comparison

**STEP 5: APPLY THE MODEL TO MAKE PREDICTIONS**

Apply the model to a new dataset to make predictions. This dataset should have all the predictor variable values, which are passed through the model to predict the unknown target variable value. The prediction will be a number between 0 and 1, representing the likelihood of positive outcome.

Useful Alteryx tool: Score

#### Overfitting

When a model is overfitted, it means that the model has focused too much on creating a highly accurate model on the Estimation data at the expense of being able to predict new data well. This means that the model is only a good model just for the Estimation dataset and is not a good model for any new dataset.

Decision Trees are prone to an error called over fitting, where the model fits the sample data too well, and as a result, does not predict future results as well as it should.

A technique that helps to eliminate this issues is the Random Forest Model.

**Random Forest Model**

* A Forest Model creates hundreds of trees, called an ensemble of decision trees
* Each tree is created by different randomly generated chunks of the original data.
* It looks at the results as a whole to make a prediction.

Each individual tree created still has overfitting issues, but when you look at the results as a whole, the overfitting gets averaged out by all of the other trees.

#### Important Definitions

**Out of the Bag Error Rate**

Explains how well the model performed with the cross-validation set in the estimation data. This gives a good understanding of how solid the model performs with just the estimation data.

You can think of it in the same terms as an R-squared.

**Confusion Matrix**

Shows again how well the model performed on the original, estimation data.

Compared to the "Out Of The Bag Error Rate", the confusion matrix does a better job at representing where errors occurred in classifying the data.

**The Percentage Error for Different Number of Trees graph**

Helps us see what the correct number of trees is to use, so we can avoid over computing.

What we are looking for is the number of trees it takes to minimize the error of each of the items, so basically, where does it flatline?

After we determine the ideal number of trees, we can change subsequent Forest Models and run our data with the smaller number of Decision Trees.

**Predictor Variables**

Which predictor variables matter the most in relation to this model? This is very helpful in determining which variables are most associated with our data on and we can focus on for future analysis.

Contributing author: Matthew David

Decision Tree and Forest Models

Summary: These are two classification models. These models help identify what group a data point belongs to. Decision Tree and Forest models can help predict classification of categorical or continuous variables.

**STEP 1: Create sample**

In any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data.

Useful Alteryx tool: Create Sample

**STEP 2: Model Settings**

Select a target variable and predictor variables, you can include as many predictor variables as you would like because the model will only use variables that work best. Specify the number of records needed to allow for a split, the smaller the number the more splits you will get. In the Forest Model you can choose the number of trees to use.

Useful Alteryx tool: Forest, Decision Tree

**STEP 3: Interpreting the Report**

Root Node Error in the Decision Tree model is the percentage of how many of the data points went to the incorrect terminal node (predicted incorrectly) when all of the data points are validated against themselves within the entire training set (the Estimation dataset). The Pruning Plot lists out the levels in the decision tree with their related error terms with cross-validation samples.

The Variable Importance Plot is a bar graph that’s length indicates the importance of the predictor variables. The Confusion Matrix is a matrix (or table) that lists out all of the possible prediction results when we validate our model against itself.

The Out of the Bag Error Rate for the Forest Model explains how well the model performed with the cross-validation set in the estimation data. Similar to R-squared. The Percentage Error for Different Number of Trees graph helps us see what the correct number of trees is to use, so we can avoid over computing in the future. What we are looking for where does the graph flatline?

Useful Alteryx tools: Forest, Decision Tree

**STEP 4: Model Comparison**

Use the fit and error measures, Accuracy which represents the overall accuracy, the number of correct predictions of all classes divided by total sample number. The F1 score is calculated the following way, precision \* recall / (precision + recall) You can read more about [precision and recall](https://www.google.com/url?q=https://en.wikipedia.org/wiki/Precision_and_recall&sa=D&ust=1577710165178000). There will also be a confusion matrix in this reprot to show how the models compared to the validation set. This confusion matrix is one of the best methods to review the accuracy and precision of your model as well as to understand any model bias in classifying your data points.

Useful Alteryx tool: Model Comparison

**STEP 5: Score Data**

Apply the model by attaching a score tool to the data set you are trying to classify and the model object.

Useful Alteryx tool: Score

Boosted Models

Summary:

**STEP 1: Create sample**

In any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data.

Useful Alteryx tool: Create Sample

**STEP 2: Model Settings**

Select a target variable and predictor variables, you can include as many predictor variables as you would like because the model will only use variables that work best. For a Boosted model it is best to set the target type in the model customization tab. Your options are Continuous, Count, Binary Categorical or Multinomial Categorical.

Useful Alteryx tool: Boosted Model

**STEP 3: Interpreting the Report**

The Variable Importance Plot is a bar graph that’s length indicates the importance of the predictor variables. The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specified assessment criteria

Useful Alteryx tools: Boosted Model

**STEP 4: Model Comparison**

Use the fit and error measures, Accuracy which represents the overall accuracy, the number of correct predictions of all classes divided by total sample number. The F1 score is calculated the following way, precision \* recall / (precision + recall) You can read more about [precision and recall](https://www.google.com/url?q=https://en.wikipedia.org/wiki/Precision_and_recall&sa=D&ust=1577792182707000).

The Confusion Matrix is a matrix (or table) that lists out all of the possible prediction results when we validate our model against our validation set. This confusion matrix is one of the best methods to review the accuracy and precision of your model as well as to understand any model bias in classifying your data points.

Useful Alteryx tool: Model Comparison

**STEP 5: Score Data**

Apply the model by attaching a score tool to the data set you are trying to classify and the model object.

Useful Alteryx tool: Score