Problem Solving with Advanced Analytics

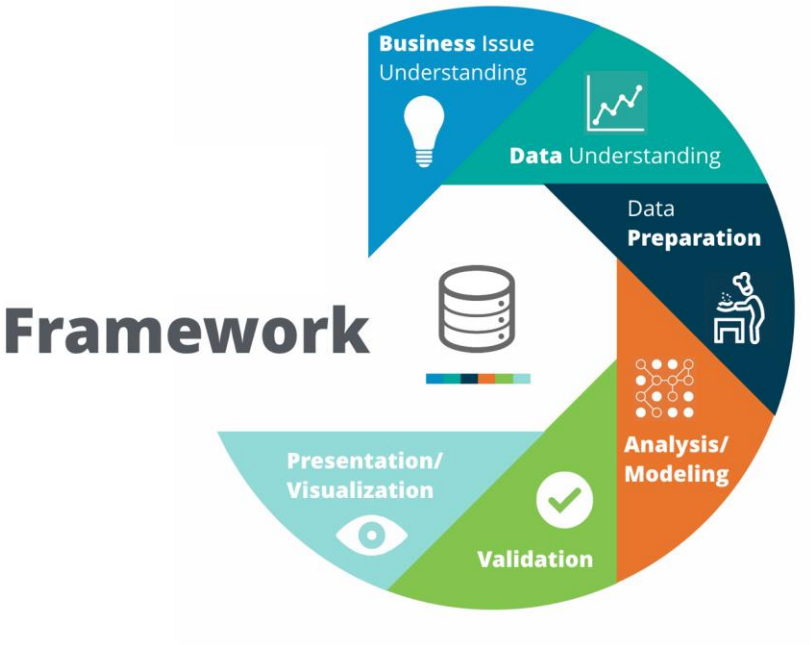
Lesson 1: The Analytical Problem

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

This framework was originally developed by data miners to generalize the common approaches to defining and analysing a problem. In this course, we will call CRISP-DM the "Problem Solving Framework".

The framework is made up of 6 steps:

1. Business Issue Understanding
2. Data Understanding
3. Data Preparation
4. Analysis/Modeling
5. Validation
6. Presentation/Visualization



Business Issue Understanding:
a) What decision needs to be made?
b) What information is needed to inform that decision?
c) What type of analysis will provide the information to inform that decision?

"This initial phase focuses on understanding the project objectives and requirements from a business perspective and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used."

Data Understanding:

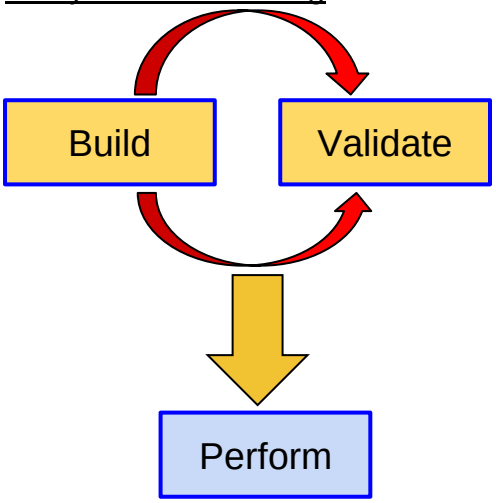
a) What data is needed?
b) What data is available?
c) What are the important characteristics of the data?

"The data understanding phase starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information."

Data Preparation:
It generally involves the following operations on data:

a) Gather: Collection of data from multiple sources in the organization
b) Cleanse: Resolving the issues with data set before analysis, in the form of incorrect/ missing data
c) Format: Formatting the data
d) Blend: Blending or combining of data with other datasets to enrich with additional variables
e) Sample: Sampling the dataset and work with a more manageable number of records

Analysis and Modeling:



"In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed."

Validation:

a) Observe the key results on the model
b) Ensure the results make sense within the context of the business problem
c) Determine whether to proceed to the next step or return to a previous phase
d) Repeat as many times as necessary

"At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to the final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached."

Presentation and Visualization:
It is the process of telling the story about the data that meets the needs of the decision-makers.
The type of presentation and visualization used should be determined based upon both the audience and the analysis itself.

a) Determine the best method of presenting insights based on the analysis and the audience both
b) Make sure the amount of information shared is not overwhelming
c) Use the results to tell a story to the audience

d) For more complex analyses, walk the audience through the analytical problem-solving process
e) Always reference the data sources used
f) Make sure the analysis supports the decisions that need to be made

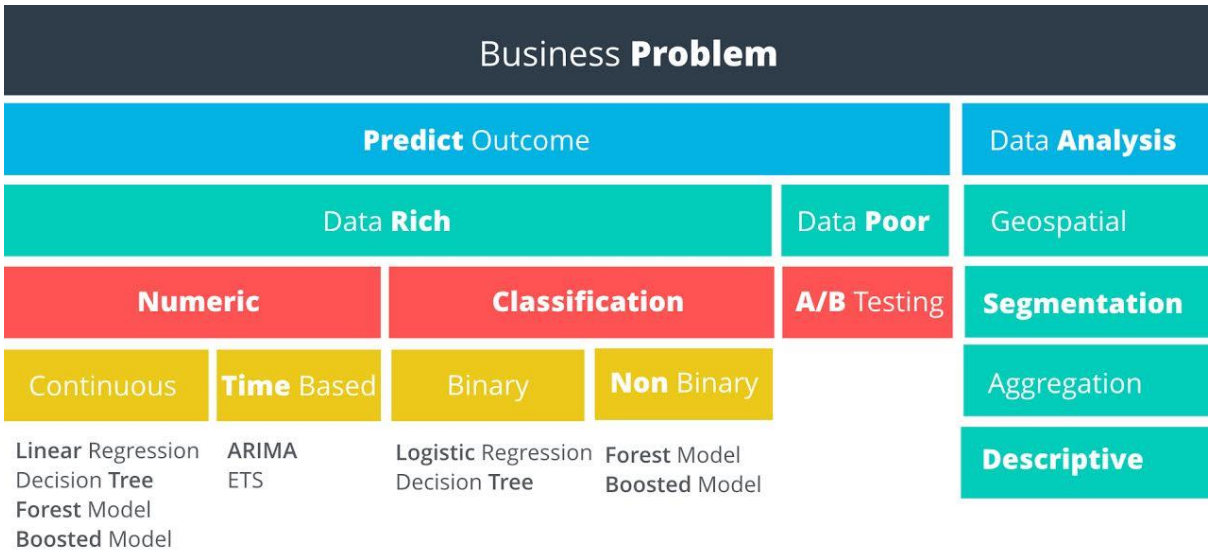
"Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process. In many cases, it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model the customer needs to understand up front the actions which will need to be carried out to make use of the created models."

Lesson 2: Selecting an Analytical Framework

Methodology Map

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

There are two scenarios involved while solving a business problem:
i) Data Analysis
ii) Predictive Analysis



Data Analysis: It refers to the more standard approach of blending together data and reporting on the trends and stats and helps to answer business questions that involve an understanding of the dataset such as “On average, how many people order coffee and a doughnut per transaction in my store in any given week?"

Non-Predictive Analysis

It includes:

a) Geospatial
b) Segmentation
c) Aggregation
d) Descriptive

Geospatial Analysis: This type of analysis uses location-based data to derive conclusions. For example, identifying customers by geographic region, calculating the distance store locations or creating a trade based upon customer locations.

Segmentation Analysis: Segmentation is the process of grouping similar data together. Groups can be simple as customers who have purchased different items, or more complex segmentation techniques like, identifying similar stores based upon the demographics of their customers

Aggregation Analysis: This methodology simply means calculating a value across a group or dimension and is commonly used in data analysis. For example, if we want to aggregate the sales data for a salesperson by month- adding all of the sales closed for each month. Then, we may aggregate across dimensions, such as sales by month per sales territory. Aggregation is often done in reporting to be able to “slice and dice” information to help managers make decisions and view performance.

Descriptive Analysis: Descriptive statistics provide simple summaries of a data sample. Examples could be the calculation of average GPA for applicants to a school, calculation of the batting average of a professional cricket player. Some of the commonly used descriptive statistics are Mean, Median, Mode, Standard Deviation and Interquartile range.

Predictive Analysis

Predictive Analysis will help businesses to predict future behaviour based on existing data, such as "Given the average coffee order, how much coffee can I expect to sell next week if I were to add a new brand of coffee?"

It can be classified into two types: a) Data Rich
b) Data Poor

Data-Rich: It can be simply described as the scenario when we have enough sample data to predict the outcome.

Data Poor: It is the scenario when we don't have enough data to predict the outcome of the analysis.

An example problem can be a prediction of incremental sales for the new product of a company while having no data related to the new product. This is considered to be a data-poor scenario. This is because we don't have any data related to the sales of the product to help us predict an outcome.

Data-Poor Business Problem: If there is not sufficient usable data to solve the problem, then we need to set up an experiment to help us get the data we need. An experiment in a business context is usually referred to as an A/B Test.

Data-Rich Business Problem: Assuming we have enough data to proceed with the analysis, our next decision is to determine whether it is a
i) numeric outcome, or a
ii) non-numeric outcome

i) Numeric Outcomes are those where an outcome is simply a number. For example, predicting the demand for electricity or the hourly temperature. Models predicting numeric data are called regression models.
ii) Non-numeric Outcomes are those trying to predict the category into which a case falls, such as whether a customer pays on time, pay late, or default on a payment. Models predicting non-numeric data are called classification models.

Examples of Numeric & Non-Numeric Outcomes:
Tricycle Manufacturer's Production Department:
A manufacturer wants to use historical production data to know how many tricycles they'll need to produce over the next six months to meet expected demand. Since the outcome the manufacturer wants to predict is a number, then the target variable is numeric. Therefore, they would use a numeric or regression model to solve this problem.

Hot & Fresh Pizza's Marketing Department:
Hot & Fresh Pizza wants to use sales data from their existing stores and respective demographic data around those stores to predict how many pizzas they'll sell at their new store location. Since the outcome that Hot & Fresh Pizza is trying to predict is the number of pizzas, then the target variable is numeric and they would use a numeric or regression model to solve this problem.

Risk Management Department at a Bank:
A bank wants to use historical data of their clients to predict whether a new customer will default on a loan, always pay on time, or sometimes pay. Since the outcome the bank is trying to predict is a category that the new customer will fall into, they would use a non-numeric or classification model to solve this problem.

Introduction to Numeric Models

Target Variables:
Target variables represent the outcome we are trying to predict. To select the right predictive model, we first determine whether the target variable is numeric or non-numeric. The type of numeric or non-numeric target variables will then help us select which model is appropriate.

Types of Numeric Variables:

a) Continuous: A continuous variable is one that can take on all values in a range. For example, our height can be measured down to many decimal places.
b) Time-based: It is to predict what will happen over time. This is often related to forecasting.
c) Count: These are discrete positive integers. They are called count because they can be counted. These variables are not common in business models.

If the target variables are continuous, we can build Continuous Models to solve the business problem, for Time-based variables we can build Time series analysis models to solve the business problem.

Introduction to Non-Numeric Models

A non-numeric variable is often called categorical, because the values of the variable take on a discrete number of possible values or categories. Examples include whether an electronic device will fail before 1000 hours or not; whether a customer will pay on-time, pay late, or default on a payment, or whether a store is classified as large, medium or small.

Types of Non-Numeric Variables:

a) Binary: If there are only two possible categorical outcomes such as Yes/No or True/False then the variables can be classified as Binary.
b) Non-Binary: If there are more than two possible categorical outcomes, such as small/medium/large or pay on-time/ pay late/ default on the payment, then the variables can be classified as non-binary.

Lesson 3: Linear Regression

We will now create linear regression models to help predict numerical data such as sales. The following concepts will be covered:

1) Linear relationship of data
2) Multiple R-squared and p-values
3) Significant coefficients
4) Modelling categorical variables

A scatter plot is plotted to describe the relationship between the number of employees and the number of tickets. It shows they have a linear relationship since we can draw a straight line through the points in the plot.

By the equation of the straight line, we can predict the values for tickets given a certain number of employees,

$$y = mx + b$$

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 known as the dependent variable.

Predictor Variable: Predictor variables are used to predict the target variable and are also known as independent variables.

Linear Regression in Google Sheets/ Microsoft Excel :

a) Using Google Sheets, we can determine the slope of the X and Y variables by using =SLOPE(data_y, data_x) formula.
b) We then calculate the intercept of the regression line by =INTERCEPT(data_y, data_x) formula.
c) Now that we have the required data, we can predict the Y through the straight-line formula $y = mx + b$.

Validation

Now that we have performed the analysis, we need to validate the results of the Linear Regression model. Or in other words to determine how good our model is.

a) Using =CORREL(data_y, data_x), we can determine the correlation between the target and the predictor variable. It is also denoted as r. The range of r is from +1 to -1. The lesser the value of r closer to 0, the better is the correlation between the variables.
b) Now to check how good the data fit the regression line, we will calculate the Coefficient of determination or r-squared. It is a coefficient between 0 and 1. It

is described as the percent variance in observations explained by the model. An R-squared value greater than 0.7 is considered to be a strong model, a value of 0.5 is good, while 0.3 cannot be taken as useful. The formula is =RSQ(data_y, data_x).

Multiple Linear Regression

For Linear Regression we used the formula, $y = mx + c$. But what if we have more than one predictor variables. For that, we use Multiple Linear Regression which is given by the formula,

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$;

Y = Target variable

x_1, x_2, x_3 = Predictor variables

b_0 = Y-intercept

Dataset: Linear Example Data

MLR in Sheets/ Excel:

a) We choose the predictor variable as “Number of Employees" and target variable as “Average Number of Tickets” and create a scatter-plot. We can observe that both the variables are linearly dependent and it's a good practice to analyse the individual variables first before you run your variables through the linear regression model.
b) We then create a scatter-plot between the “Value of Contract” and “Average Number of Tickets”. Even though the contract values are made up of discrete numbers, we can still see a linear relationship between the contract value and the number of tickets.
c) Since both the predictor variables show good linear relation with the target variable, we will use them in the MLR model.
d) Use Regression from Data Analysis in Excel and input X-values as “Number of Employees” and “Value of Contract” and input Y-values as “Average Number of Tickets”.
e) For MLR, the R-squared value will be spiked because of the multiple number of predictor variables. For this reason, we will be using the adjusted R-squared value. For more information, check out this link.

Linear Regression using non-numeric predictor variables

We have discussed having more than one predictor variables in regression by using MLR. Now if we want to add a categorical predictor variable to our equation then it won’t work as we can’t do math using string variables. Therefore, we need to transform the variable into a numeric variable.

But we just cannot transform by assigning numbers to each category. We always have to keep in mind that we need to make a linear relationship between the target and predictor variables.

For example, if we have the following dataset

State	Region	Avg_income	Pct_under_18	Expenditures
AL	Southeast	3724	0.332	208
TN	Southeast	3946	0.315	212
ME	Northeast	3944	0.325	235
DY	Southeast	3967	0.325	216
HI	West	5613	0.386	546
OH	Midwest	5012	0.324	221
AR	Southeast	3680	0.32	221
SD	Midwest	4296	0.33	230
NH	Northeast	4578	0.323	231
MO	Midwest	4672	0.309	231

And we want to predict the per pupil student expenditures in a state. For this we will be using three predictor variables: Avg_income, Pct_under_18, and Region. But the 'Region' is a categorical variable with four values: West, Midwest, Northeast, and Southeast. So, we can assign 1, 2, 3, 4 respectively and run the model. The results will give the following equation:

Expenditures = -530 + 0.073 Avg_Income + 1406.36 Pct_Under_18 + 6.53 region

Analysing the equation,

- The coefficient on average income implies that for every one dollar of additional average income, the state spends 7.3 more cents on each pupil.
- The coefficient on percent under 18 implies that for every one additional percentage of the population under 18, the state spends about $1406.36 more per pupil.
- The coefficient on region implies that for every increase in region, the state spends about $6.53 more per pupil. This statement does not make any sense, since the numbers in the region variable are basically categories and not exact values. Therefore, we cannot use this format.

The solution to the above problem is to use what is known as Dummy variables. A dummy variable can only take on two values, generally zero or one. We would add one dummy variable less than the number of unique values in the categorical variable. So, if the variable is binary, then we will add one dummy. If there are four categories then we will add three dummy variables.

After converting the region variable to dummy variables we get the following dataset

State	Avg_income	Pct_under_18	Expenditures	Southeast	Northeast	Midwest	West
AL	3724	0.332	208	1	0	0	0
TN	3946	0.315	212	1	0	0	0
ME	3944	0.325	235	0	1	0	0
DY	3967	0.325	216	1	0	0	0
HI	5613	0.386	546	0	0	0	1
OH	5012	0.324	221	0	0	1	0
AR	3680	0.32	221	1	0	0	0
SD	4296	0.33	230	0	0	1	0
NH	4578	0.323	231	0	1	0	0
MO	4672	0.309	231	0	0	1	0

And the MLR equation is as follows,

Expenditures = β 0 + β1 Avg_Income + β2 Pct_Under_18 + β3 midwest + β4 southeast + β5 west

Each of the dummy variables take values 0 or 1. Note that we haven't taken any variable for northeast. This is because the equation needs a baseline value that cannot be coded into a dummy variable. If a state is in the northeast, then the value for the other regions will be zero.

Evaluating an Equation

School exp = -468 + (0.067 x avg income) + (1349 x pct under 18) - (14.4 x midwest) - (9.3 x southeast) + (16.5 x west)

Now suppose we have a state with average income of 4011, % under 18 of 32.5%, and was in the northeast region. We will be calculating the School expenditure as follows:
⇒ School exp = -486 + (0.067 x avg income) + (1349 x pct under 18) - (14.4 x midwest) - (9.3 x southeast) + (16.5 x west)
⇒ School exp = -486 + (0.067 x 4011) + (1349 x 0.325) - (14.4 x 0) - (9.3 x 0) + (16.5 x 0)
⇒ School exp = -486 + 268.737 + 438.425 - 0 - 0 + 0
⇒ School exp = 230.162
We are just plugging the values for each variable given, and for the dummy variables we are plugging in 0 since none of them is northeast.

Coefficient of the Dummy Variable

The coefficient of each dummy variable is the average difference between the state expenditures in one region, compared to states in the northeast region, assuming all other variables are held constant.

Alteryx

Alteryx provides analysts with the ability to easily prepare, blend, and analyze all of their data using a repeatable workflow, then deploy and share analytics at scale for deeper insights. Analysts can connect to and cleanse data from data warehouses, cloud applications, spreadsheets, and other sources; easily join this data together; then perform analytics – predictive, statistical and spatial – using the same intuitive user interface, without writing any code.

Downloading Alteryx

a) Download the AlteryxDownloadManagerNon_Admin from the Product Support Page https://www.alteryx.com/designer-trial/alteryx-free-trial
b) After downloading and executing the installer you will be given options for the version of Alteryx Designer to download. Choose the Alteryx Designer 2019.x with R-based Predictive Tools (advanced) for 64-bit.
c) It will take a while to install the package since the whole application is about 3.5GB large!

Building a Linear Regression Model in Alteryx Designer

Dataset link: Linear-Example-Data

a) Drag an “Input Data” tool into the canvas and click on the drop-down option in Connect a File or Database field in the Configuration Pane in the left side. Choose the location where you saved the dataset.
b) From the Predictive ribbon in the tools pane, drag “Linear Regression” to the canvas and connect it to the “Input Data”. Click on the “Linear_Regression” and in the Configuration Pane change the Target Variable to Average Number of Tickets and check all the predictor variables except Client ID.
c) Right click on the “Linear Regression” and select Add Browse After (Reports). This will create a visual report of the Linear Regression model.

The following is the report of the dataset given above:

Record Report

1 **Report for Linear Model Linear_Regression**

2 *Basic Summary*

3 Call:
lm(formula = Average.Number.of.Tickets ~ Number.of.Employees + Value.of.Contract + Industry, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-88.942	-15.643	0.809	16.065	103.038

6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.586e+00	4.225e+00	0.3753	0.70767
Number.of.Employees	9.849e-02	1.317e-02	7.4759	8.79e-13 ***
Value.of.Contract	9.989e-05	4.489e-05	2.2252	0.02683 *
IndustryRetail	-2.124e+01	3.601e+00	-5.8963	1.01e-08 ***
IndustryServices	-1.421e+01	4.549e+00	-3.1234	0.00197 **

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

8 Residual standard error: 27.616 on 295 degrees of freedom
Multiple R-squared: 0.585, Adjusted R-Squared: 0.5794
F-statistic: 104 on 4 and 295 degrees of freedom (DF), p-value < 2.2e-16

9 *Type II ANOVA Analysis*

10 Response: Average.Number.of.Tickets

	Sum Sq	DF	F value	Pr(>F)
Number.of.Employees	42624.98	1	55.89	8.79e-13 ***
Value.of.Contract	3776.22	1	4.95	0.02683 *
Industry	26700.37	2	17.5	6.54e-08 ***
Residuals	224987.16	295		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficient Estimates

In the equation of MLR, we have seen the b_0, b_1, b_2, b_3. These are the estimates of b's. They represent the magnitude of the relationship between each predictor variable and the target variable. For example, the coefficient on the Number.of.Employees means that each additional employee will lead to roughly an additional 0.1 tickets, holding all other variables constant. In the Estimate column of the report above, we can find the different coefficient estimates of predictor variables.

P-Value

The p-value is the probability of having no relationship between the target and predictor variable. In other words, the lower the p-value, the higher the probability that there is a relationship between the variables or there exists a correlation between the target and predictor variables. When a predictor variable has a p-value

below 0.05, the relationship between it and the target variable is considered to be statistically significant.

FYI, *Statistically Significant means that a result is not likely to occur randomly rather there is a cause which is attributable to it*.
In the Pr(>|t|) column, we can see the p-value of the predictor variables, as well as the stars next to it indicating how significant the variable is. Generally, we'll want to remove variables from the model that are not statistically significant predictors of the target variable.

R-Squared

Here in the above report, we can see the R-squared value is 0.585 while the adjusted R-squared value is 0.5794.
Remember, R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. The higher the r-squared, the higher the explanatory power of the model.

Scoring the Dataset

Previously we applied the results of a Linear Regression manually by using the linear regression equation. As we add more predictor variables, manual calculations can become more complicated. The Score tool makes this easy by automatically applying the linear regression equation from a Linear Regression tool.

To use the Score tool, we connect the O (object) node of the Linear Regression tool to one of the input nodes of the Score tool. On the other node, we connect the data we are using to make a prediction. We must make sure that all the predictor variables used in the Linear Regression model are present in the input dataset used in the Score tool.

The resulting dataset will be identical to the input dataset but with one additional field called the Score or X. This field represents the predicted values for the target variable in the linear regression model.

Formula Tool

A formula tool is used to formalize the data in a particular way we want. To use a Formula, we first attach an input dataset. Then we select from the configurations what record we'd like to use. Lastly, we build the formula by writing the formula manually in the Expressions section or selecting the variables and expressions from the middle section.

Predictive Model Playbook

Summary: Linear Regression is a statistical method used to predict numeric 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 continuous numeric variable, such as price, revenue, customers, 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


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).


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.


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.


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.


Score

Lesson 4: Practice Project

Predicting Diamond Prices

Problem Statement

A jewellery company wants to put in a bid to purchase a large set of diamonds, but is unsure how much it should bid. In this project, we will use the results from a predictive model to make a recommendation on how much the jewellery company should bid for the diamonds.

Project Details

A diamond distributor has recently decided to exit the market and has put up a set of 3,000 diamonds up for auction. Seeing this as a great opportunity to expand its inventory, a jewellery company has shown interest in making a bid. To decide how much to bid, we will use a large database of diamond prices to build a model to predict the price of a diamond based on its attributes. Then we will use the results of that model to make a recommendation for how much the company should bid.
Note: The diamond price that the model predicts represents the final retail price the consumer will pay. The company generally purchases diamonds from distributors at 70% of that price, so your recommended bid price should represent that.

Training Dataset: diamonds.csv
Test Dataset: new_diamonds_new.csv

Step 1: **Understanding the Data**

There are 2 datasets, *diamonds.csv* contains the data on which we will build the Linear Regression Model and we will apply the model on the *new_diamonds_new.csv* dataset. The description of the different predictor variables are as follows:

a) Carat: It represents the weight of the diamond (numerical variable).
b) Cut: It represents the quality of the cut of the diamond, and falls into the 5 categories- fair, good, very good, ideal, and premium (Categorical variable).
c) Clarity: It represents the internal purity of the diamond, and falls into 8 categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF (in order from least to most pure) (Categorical Variable).
d) Color: It represents the color of the diamond, and is rated D through J, with D being the most colorless (and valuable) and J being the most yellow.

Step-2: **Input the Data**

First we input the *diamonds.csv* by using the Input Tool and connecting it to the Select Tool to deselect the features not needed i.e. Column1, cut_ord, clarity_ord as Alteryx automatically convert categorical variables to dummy variables.

Step-3: **Building the Model**

We then connect the Select Tool to the Linear Regression Tool from the Predictive Tab in the Toolset bar. To get the report of the model performance, we connect the I (Interactive Reports) node to a Browse Tool *(Interactive Report node is available only on the latest version of Alteryx, which gives a more graphic enriched report, use it only if you have a good GPU, otherwise use the R (Report) node)*. Run the workflow using Ctrl+R or click on the Run Workflow button above the canvas.

This is the report that I've generated:

R SQUARED
0.916

ADJUSTED R SQUARED
0.916

MEAN ABSOLUTE ERROR
804.543

MEAN ABSOLUTE PERCENT ERROR
44.763

MEAN SQUARED ERROR
1338368.448

ROOT MEAN SQUARED ERROR
1156.879

F-STATISTIC
30379.29 on 18 and 49981 degrees of freedom

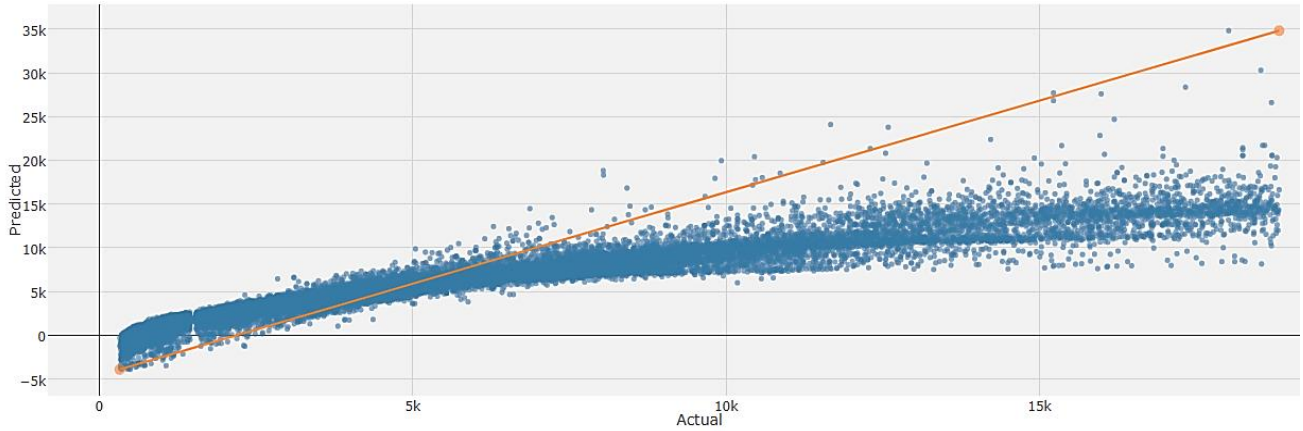
RESIDUAL STANDARD ERROR
1156.58 on 49996 degrees of freedom

Display Advanced Statistics

Search:

Variable	Estimate	Impact	Confidence
(Intercept)	-7382	7382.29058605259	***
carat	8887	8887.41193964138	***
cutGood	682	682.167003029958	***
cutIdeal	1017	1017.09019854827	***
cutPremium	889	889.256661250102	***
cutVery Good	867	867.075288458184	***
colorE	-205	205.242761001335	***
colorF	-299	298.671194903821	***
colorG	-499	498.560001118475	***
colorH	-966	966.199474874904	***

This the model performance graph below. As we can see from the scatter plot, the values tend to curve from 5K in the Actual axis, the linear regression line does not quite fit all the data points. But since the lesson was based on Linear Regression, I won't introduce polynomial regression over here. Also, the R-squared value is enough for the model to be accurate.



Step-4: **Scoring the model**

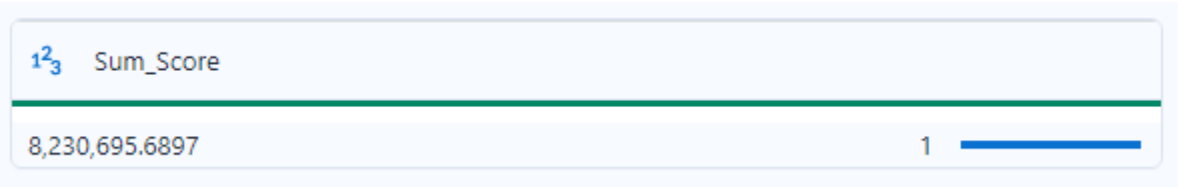
Now that we know our model performs perfectly, we will use the model to score the *new_diamonds_new.csv* dataset. We will input the new dataset similar to Step-2 and then use Step-3 to select the features we will be needing. Then we will connect the output of the Select Tool to the Data (D) node of the Scoring Tool and connect the output node (O) of the Linear Regression Model to the model node (M) of the Scoring Tool.

In order to visualize the scoring of the model on the new dataset, we will connect the output of the Scoring Tool to a Browse Tool, and to visualize the different predictor variables versus the scoring, we will connect a Scatterplot Tool with the output of the Scoring Tool, and browse the visualization by connecting it's output with a Browse Tool.

Step-5: **Summarizing the model**

Next, we connect the output of the Scoring Tool to a Summarize Tool to add all the predicted prices together. This gives us an estimate of the total retail value of the batch of 3,000 diamonds. In order to do this, we will select the Score Field from the Fields box in the Configuration Pane, and then select Sum from the Add drop-down, which will create a new field Sum_Score.

To come up with a bid price, we multiply this bid by 70%, since the company targets purchases of diamonds at 70% of retail value, by connecting the Summarize Tool with a Formula Tool and in the Configuration Pane, we will select the Output Column as Sum_Score and write the formula in the Expression Box: [Sum_Score]*0.7. We will connect a Browse Tool to show the recommended bidding price.



This is my recommended bidding price that is $8,230,695.69

Lesson 5: Project

Predicting Catalog Demand

Business Problem

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Details

- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price-cost) on all products sold through the catalog is 50%.
- Make sure to multiply the revenue by the gross margin first before you subtract out the $6.50 cost when calculating the profit.

Also, a short report is to be written with the recommendations outlining the reasons why the company should go with the recommendations to the manager.

Steps to solve the problem

Step-1: **Business and Data Understanding**

A description of the key business decisions that need to be made.

Note: Clean data is provided for the project, so we can skip the data preparation step of the Problem Solving Framework.

Step-2: **Analysis, Modeling, and Validation**

Build a Linear Regression model, then use it to predict sales for the 250 customers.

Step-3: **Writeup**

Once we have our predicted or expected profit, a brief report is to be made with the recommendation to whether the company should send the catalog or not. **Hint:** We want to calculate the expected revenue from these 250 people in order to get expected profit. This means we need to multiply the probability that a person will buy our catalog as well. For example, if a customer were to buy from us, we predict this customer will buy $450 worth of products. At a 30% chance that this person will actually buy from us, we can expect revenue to be $450 x 30% = $135.

Data

p1-customers.xlsx - This dataset includes the following information on about 2,300 customers. **Important**: You should build your model on this dataset and not p1-mailinglist.xlsx.

p1-mailinglist.xlsx - This dataset is the 250 customers that you need to predict sales. This is the list of customers that the company would send a catalog to. Use this dataset to estimate how much revenue the company can expect if they send out the catalog. It includes all of the fields from *P1_Customers.xlsx* except for Responded_to_Last_Catalog so this variable cannot be used in the linear regression model since it could not be applied to the mailing list data set. It also includes two additional variables.

- Score_No: The probability that the customer WILL NOT respond to the catalog and not make a purchase.
- Score_Yes: The probability that the customer WILL respond to the catalog and make a purchase.