5.5: Intro to Predictive Analysis

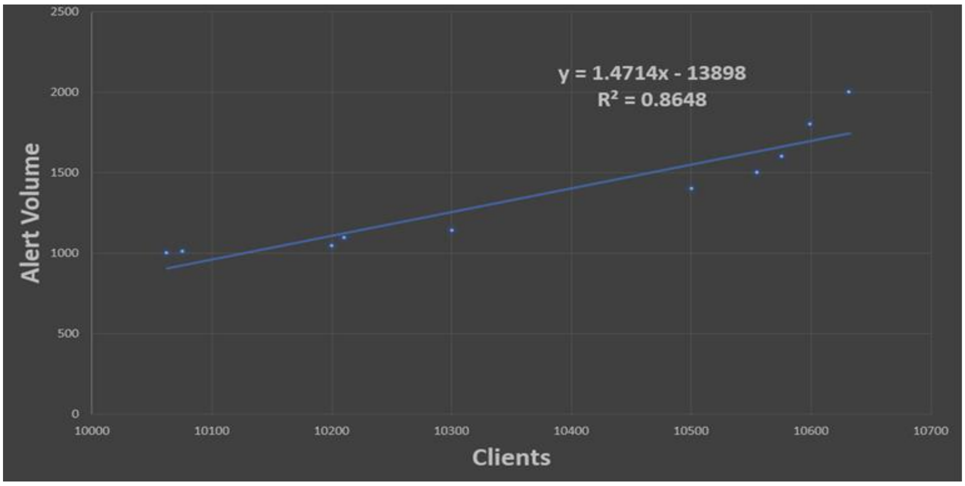
Step 1: Understanding Regression

You learned about linear regression in this Exercise, but you’d also like to know what logistic regression is. Conduct some research on logistic regression and explain how it differs from linear regression. When would you use logistic instead of linear regression and why?

Using a straight line on the cartesian plane, linear regression generally predicts continuous dependent variables, including price and age. Continuous variables, such as price and age, create a line of best fit. When predicting categorical dependent variables, logistic regression is best suited for binary targets such as "Good/Bad," "Yes/No," or "True/False." In contrast to linear regression, which can have a large range of output variables, logistic regression outputs can only be between 0 and 1. While logistic regression works with classification analysis, predicting binary outcomes based on a collection of independent variables, linear regression deals with the connection between independent and dependent variables in predictive analysis.

Step 2: More on Linear Regression

Take a look at the linear regression below. It shows a relationship between the number of clients at Pig E. Bank and the number of alerts for fraudulent activity at the bank. Describe the relationship between these two variables. Based on the results, how would you assess the fitness of this model in predicting alert volume based on the number of clients?



With an R-squared value of 0.8648, the correlation between Pig E. Bank's client base and the number of alerts for fraudulent conduct is favourable. This suggests that the volume of alerts increases with the number of customers Pig E. Bank has. The outcome variation can be described by the predictor factors in 86% of cases, according to the R2 value of 0.8648. The upward slope of the line suggests a positive correlation, but because the values on the x- and y-axes are scaled, it is challenging to pinpoint the precise positive correlation. The dots are reasonably near the line, indicating that the model is reasonably accurate in predicting alert volumes based on client numbers. A bigger sample would, however, improve these findings.

Step 3: Differentiating between Models

Read the scenarios below, then decide which predictive model you'd use in each one. Provide a short explanation for the rationale behind your decisions.

- **Scenario A:** As an analyst for a large financial institution, your job is to perform research and develop models that predict the future values of precious metals. You theorize that the global oil price can be predicted based on the unemployment rates of the top 20 countries in gross domestic product (GDP). Would you use a regression model or classification model to validate your theory? What specific algorithm would you use for this predictive model and why?

  A linear regression model is suggested to predict two continuous data points, oil prices and unemployment rates. Logistic regression is inappropriate for this type of study because it is not binary. Oil prices are the dependent variable, whereas unemployment rates are the independent variable. The programme would then focus on the top 20 countries in terms of GDP to look at the association between the two variables. Given that a binary output is not necessary, a linear regression model is better suited to this circumstance than a logistic regression model.

- **Scenario B:** You're a data analyst for an online movie provider that collects data on its customers' viewing habits. Part of your job is to support the company's efforts to display movies that customers are likely to enjoy prominently on their profile page and keep the movies they're least likely to enjoy off their profile page altogether. To this end, your company has asked you to predict which customers are most likely to watch a romantic comedy starring Adam Sandler and Drew Barrymore. Would you use a regression or classification model for this? What specific algorithm would you use and why?

  In order to forecast the binary response of viewers of a romantic comedy featuring Adam Sandler and Drew Barrymore, I would advise employing a classification model. A random forest decision tree method, which also recognises traits like gender, marital status, and previous movie likes, can be used to accomplish this. The classification model can be used to display where a consumer prefers to watch a romantic comedy, making it appropriate for dealing with categorical variables. For this objective, the random forest technique works well since it enables the use of several decision trees based on qualitative data to assess a customer's propensity to appreciate a film.

Step 4: Bias in Your Data

Imagine you were involved in collecting the data that was used in the linear regression in step 2. What types of bias could have arisen when collecting the data and why?

Given that it only represents a small portion of the Pig E. Bank customer base, the small data set may be an indication of sampling and/or collection bias. Due to the fact that the data does not reflect all customers, confirmation bias may result. Additionally, the results of the linear regression may be skewed because it only displays a tiny portion of the data. The data's original source and the random sample utilised must be noted. The limited sample size raises the possibility of sample bias, which would lead to a lack of diversity in the data that was gathered. Examining the data-collecting procedure and the "Alert algorithm" for intrinsically biased elements is crucial to preventing collection bias.