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?

Answer: Logistic regression is a statistical method used for binary classification problems, where the outcome variable is categorical and has only two possible values (e.g., yes/no, true/false, 0/1). Unlike linear regression, which predicts continuous numeric values, logistic regression predicts the probability of an event occurring. It uses a logistic (sigmoid) function to transform the predicted values into probabilities, typically ranging from 0 to 1.

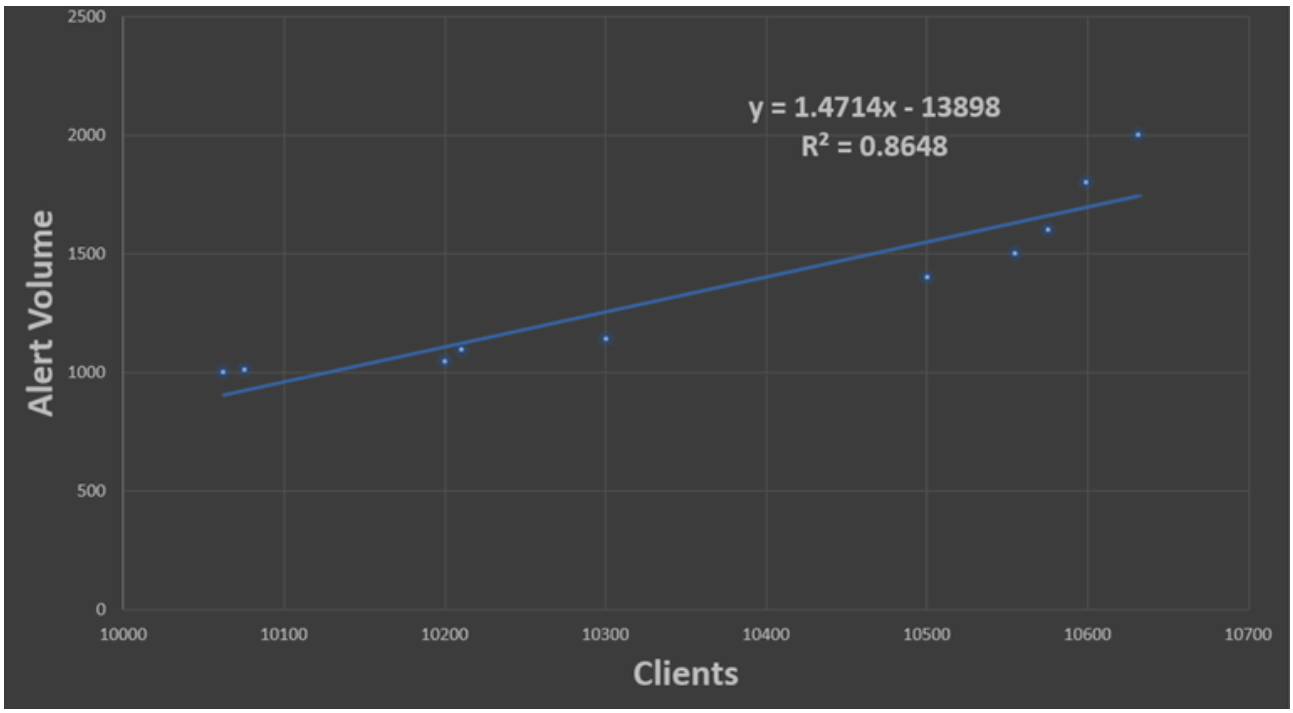
Logistic regression is ideal when the dependent variable is binary or when dealing with classification tasks like predicting whether a customer will churn or not, whether an email is spam or not, etc. It's especially useful when you want to understand the probability of an event's occurrence based on various predictor variables.

Linear regression, on the other hand, is suited for predicting continuous outcomes and establishing relationships between variables.

In summary, use logistic regression for binary classification problems when the outcome variable is categorical, and linear regression for predicting continuous numeric values or when exploring relationships between variables.

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?



There is a positive correlation between Clients and Alert Volume. The more Clients Pig E bank has, the more Alert is there. But its not completely 1:1 relationship. As we can see in this chart, R score is not exactly 1, it is 0.86 and dots aren't close to the line which means there is moderate correlation between them. If the dots were closer to the line, the there would be strong relationship.

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 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?

  Answer: For Scenario A, where the goal is to predict the global oil price based on the unemployment rates of the top 20 countries in GDP, a regression model would be more appropriate. Regression models are used for predicting continuous numeric values, such as oil prices, based on independent variables like unemployment rates.
  Since the target variable (global oil price) is a continuous value, we can employ a linear regression algorithm for this predictive model. Linear regression is a straightforward and interpretable model that can establish a linear relationship between the independent variables (unemployment rates) and the dependent variable (global oil price). It will allow us to analyze the direction and strength of the relationship between the two variables and provide a numeric prediction for the oil price based on the given unemployment rates.

- **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?

Answer: For Scenario B, where the goal is to predict which customers are most likely to watch a romantic comedy starring Adam Sandler and Drew Barrymore, a classification model would be appropriate. Classification models are used for predicting categorical outcomes, such as whether a customer is likely to watch a specific type of movie or not (in this case, a romantic comedy starring Adam Sandler and Drew Barrymore).

A suitable algorithm for this predictive model is the logistic regression algorithm. Logistic regression is commonly used for binary classification tasks, where the outcome is either "yes" or "no," in this case, whether a customer will watch the specific movie or not. The algorithm uses the logistic function to calculate probabilities, allowing us to predict the likelihood of a customer watching the romantic comedy based on relevant features like their past viewing habits, movie preferences, and other user-specific data collected by the online movie provider.

By using a logistic regression model, the company can identify the customers who are most likely to enjoy the romantic comedy and prominently display the movie on their profile page, thereby enhancing user experience and engagement on the platform.

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?

1. Sampling Bias: The data may have been collected from a specific segment of clients or alerts, leading to an underrepresentation or overrepresentation of certain groups, which might not accurately reflect the entire population.
2. Selection Bias: The selection of clients or alerts for inclusion in the dataset could be biased, skewing the results in favor of particular patterns or characteristics.
3. Reporting Bias: The reporting of fraudulent activity or the number of clients could be influenced by various factors, leading to inaccurate or incomplete data.
4. Time Period Bias: Data collected during a specific time period might not capture long-term trends or seasonal variations, affecting the generalisability of the results.

It is essential to recognize and address these biases to ensure the linear regression model's results are representative and applicable in predicting alert volume based on the number of clients at Pig E. Bank.