**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?**

Linear regression is used to determine an output with a continuous dependent and one or more independent variable. Which will result in a set of data points between these two relationships on a scatter plot. For example: You might use regression logistics to understand the pricing of a house with its relationship to size and location.

While there can be many variables in linear regression, there is only a binary outcome in logistic regression such as “yes or no” or “true or false”. An example of logistic regression is if the outcome meets a certain criteria, and is more focused on probability. For example a statement can be “If the house is bigger than 1000 sqr feet then it is 75% more likely to be more expensive than others in a better neighborhood.”

**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 relationship between the number of clients and the number of fraudulent activity alerts. The linear regression model shows that as the number of clients increases, the volume of alerts also tends to increase. The R² value of 0.8648 indicates that approximately 86.48% of the variation in alert volume can be explained by the number of clients. This strong R² value suggests that the number of clients is a good predictor of alert volume. Therefore, the model demonstrates that an increase in clients is likely to result in a corresponding increase in alerts.

**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. Research tells you that rising oil prices will increase the cost of producing precious metals, impacting their value. 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?**

I would use a regression model to predict the relationship between oil prices and unemployment rates because regression is specifically designed for forecasting continuous outcomes, such as oil prices. My hypothesis is that rising oil prices may correlate with changes in unemployment rates. To test this hypothesis, I would start with a linear regression algorithm, which is effective in capturing straightforward, linear relationships between variables. If the data shows a linear relationship, this model would provide a clear understanding of how changes in unemployment rates might predict future oil prices.

**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?**

I would use a classification model for this scenario because the goal of the project is to answer an “no” or “yes” question for viewers to predict whether a customer is likely to watch a romantic comedy starring Adam Sandler and Drew Barrymore, which is a binary outcome. I would use the logistic regression since it is a probability algorithm that will measure the viewers past interests, age, gender, past program view to measure the probability how likely the person will watch a romantic comedy with these two actors.

**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?**

Like in our past projects, data collection is rarely perfect, and several types of biases could arise when collecting bank data. One example is **selection bias**, which could occur if the data is collected from only one bank region. This may result in a dataset that doesn't accurately represent the broader client base, as different regions might have varying client demographics, income levels, and banking behaviors. These differences could skew the relationship between the number of clients and alert volume.