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?

Regression analysis is predictive modeling technique to find the relationship between a dependent variable Y" and one or more independent variables "X"

Linear regression has a numeric value for both Y and X – you can see the relationship in a scatterplot

Logistic regression predicts the probability of a binary (yes/no) outcome.

So, we are assessing if the independent variable(s) are more or less likely to produce a dependent yes or a no, a 1 or a 0.

The independent variable(s) can be:

- **Continuous**—like price or weight
- **Discrete, ordinal**—data on a scale e.g. 1-5, very good/good/etc
- **Discrete, nominal**— like colours, or gender

The independent variables should not have a high correlation with each other.

Examples of use are deciding who to issue a credit card to, which customers are most likely to respond to an offer

Advantages

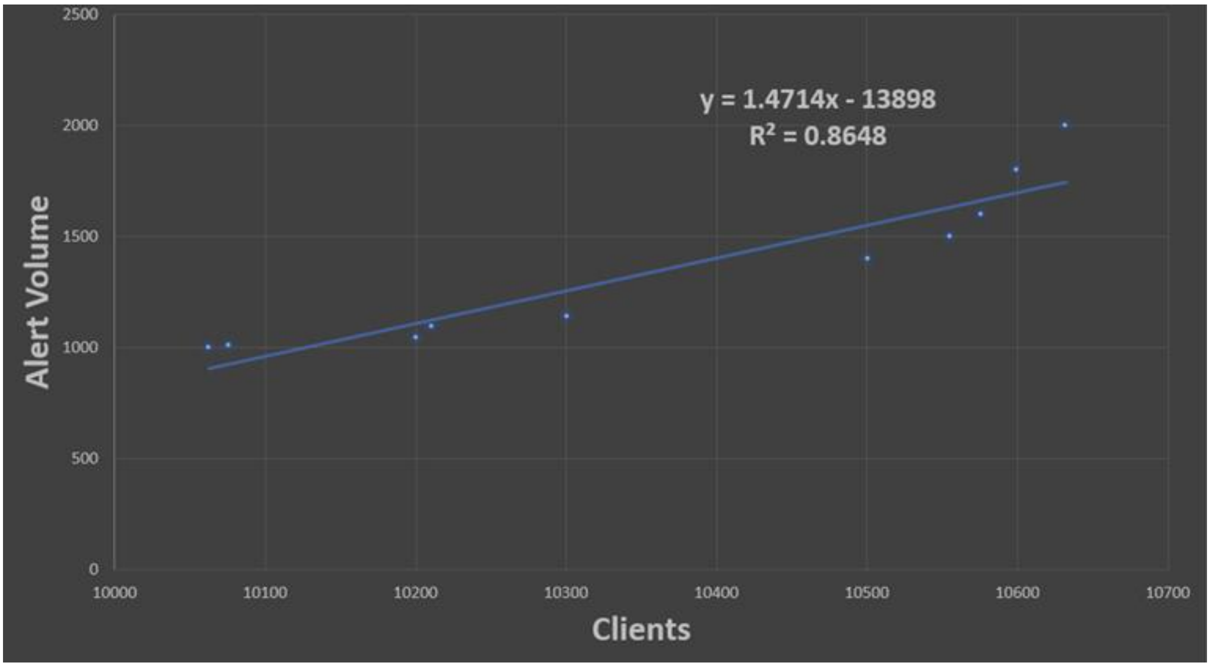
- Easier to implement and easier to add to machine learning
- Works well when the output are clearly separable
- Clear, useful insights

Disadvantages

- Don't get a continuous outcome – i.e. you don't get likelihood – you just get a yes/no
- Practically, it's unlikely that the output are clearly separable
- Needa a large data size

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?



This is a positive relationship. As customers increase, so does fraudulent activity

R^2 is close to one, so there is a strong relationship.

This is suitable model for Pig E Bank to use. They can expect that as they increase numbers of customers to see fraud alerts rise.

If we look just at the final four data points though, we can see a different pattern.

A regression line would be much steeper, suggesting a higher volume of alerts.

As customers are added and new data is gained, this model should be monitored and evaluated.

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 here, as we are looking at how one quant variable (price of oil) is affected by another quant variable (unemployment rates)

This suggests a Linear regression as the dependent variable oil price is continuous.

(If we defined the oil price as above or below a threshold, then a logistic regression could be used)

- **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 here – we are looking to classify a group of customers (as Romantic Comedy fans), based on what we already know about them

A decision tree could be used here (possibly with nodes like have they watched romantic comedies before?, age threshold, gender, in a relationship etc).

If there are many possibly ways to construct the decision tree, a random forest could be used to refine to one model.

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?

The independent variable Y is based on Alert Volume i.e. when we predict fraud, not when fraud actually occurs.

From the details given here, we do not know how the Alert Volume has been modelled – and there could be many different types of bias that are in that model.

For instance, the sample may not reflect the population of all customers

Has the model updated, as different types of fraud, or avoiding detection have developed?

In particular, as the bank grows its customer base, it’s possible it will start to attract a different demographic of customer. Does this model account for that? Or does it now start to produce more or fewer true or false positives?