5.5. Intro to Predictive Analysis
By Mary Kane

Answers 5.5

Step 1: Understanding Regression

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?

Logistic regression is a statistical modeling technique used to predict the probability of a binary event. Instead of predicting a continuous value (as in linear regression), logistic regression predicts the probability that an observation belongs to one of two classes. For example, whether the customer leaves the bank or stays.

The logistic regression estimates the probability that the dependent variable equals 1 ("Left Customer") based on one or more predictor variables.

What are the differences with Linear Regression?

Feature	**Linear Regression**	**Logistic Regression**
Target/output	Continuous (ex: balance, credit score)	Binary category (0/1) (ex: Left / Stay)
Example prediction	Predicts $7500.25	Predicts probability 0.83
Assumes	Linear relationship	Non-linear relationship between predictors and class probability
Errors	Uses least squares	Uses maximum likelihood

Why linear regression is *not* good for classification? If you use linear regression to predict a binary outcome, you might get predicted values less than 0 or greater than 1, which **don't make sense as probabilities**. Logistic regression ensures outputs are always between 0 and 1.

Mathematically, **Logistic Regression** uses the sigmoid function to constrain predicted values between 0 and 1:

$$\hat{p} = \frac{1}{1 + e^{-z}}$$

where:

- $\hat{p}$= predicted probability
- $z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$(the linear predictor)

Because output is a probability, logistic regression is **perfect for classification tasks**.

5.5. Intro to Predictive Analysis
By Mary Kane

Linear regression:

Predicts a numeric score:

“Given age, balance, and credit score — what continuous value best describes this outcome?”

Logistic regression:

Predicts a probability:

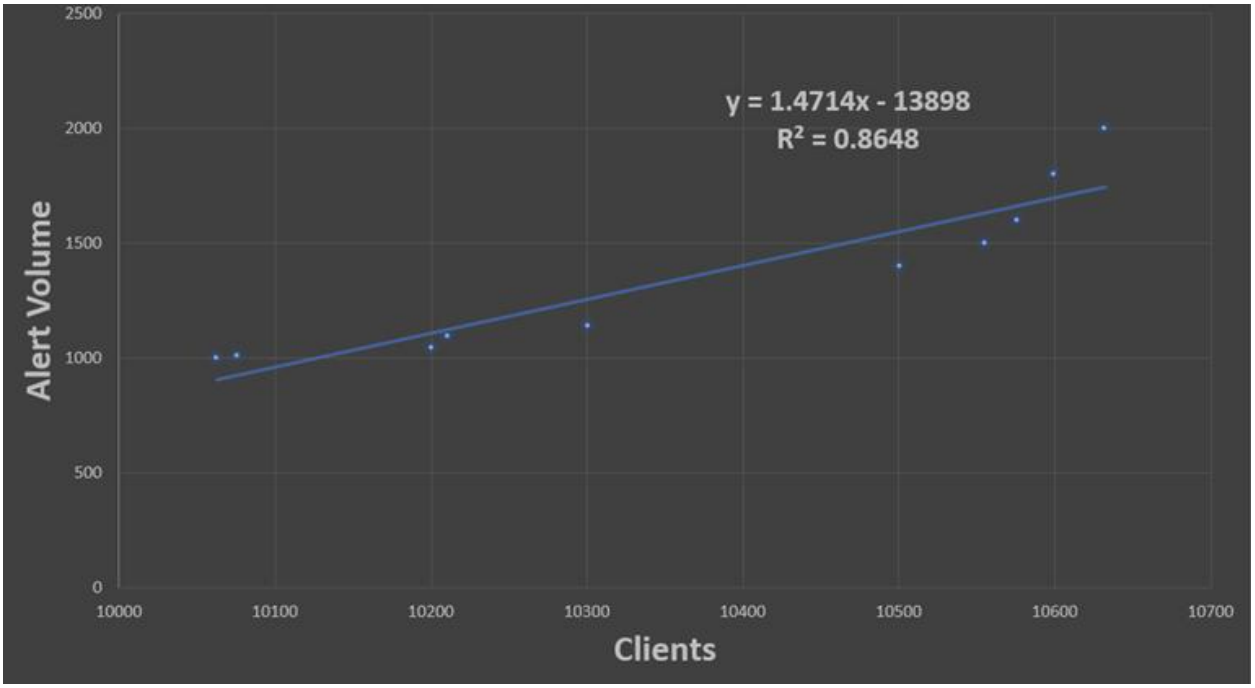
“Given age, balance, and credit score — what is the *probability* this customer will leave the bank?”

If the probability is ≥ 0.5, we might classify them as “will exit”; if < 0.5, “will stay.”

5.5. Intro to Predictive Analysis
By Mary Kane

Step 2: More on Linear Regression

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?



The plot shows a "strong positive linear relationship" between the number of clients and the number of fraud alerts.

The regression equation is:

(Y) Alert Volume = 1.4714 x **Clients** – 13898

The slope (1.4714) means that on average for each additional client, the model predicts about 1.47 more alerts.

Model fitness:

The model's **R^2 = 0.8648**, meaning about **86% of the variation in alert volume** is explained by the number of clients. This is a good fit for the prediction of a simple one-variable model, especially for forecasting within the range of the data shown.

Step 3: Differentiating between Models

5.5. Intro to Predictive Analysis
By Mary Kane

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 a classification model to validate your theory? What specific algorithm would you use for this predictive model and why?

  Answer: I would use the regression Model because the oil price is a continuous numerical value (measurable and quantitative).

  The hypothesis proposes that unemployment rates influence global oil prices; we can apply a linear regression to test the relationship between independent variables. (Unemployment rate by country and 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?

  Answer: I would use a Classification model. Will the customer watch the movie (yes/no) – it's a binary outcome. The algorithm Logistic regression can predict the probability of the customer watching the movie; it also works well with numerical and categorical predictors (ex: viewing history, genre preferences)

Step 4: Bias in Your Data

5.5. Intro to Predictive Analysis
By Mary Kane

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?

If I were involved in collecting the data for the linear regression "number of clients to fraud alert volume" the bias that can arise includes:

- Sampling Bias: If only certain regions or client segments were included in the data. The sample should represent all the regions and all the branches of the bank.
- Measurement bias: if more than one analyst is manipulating the data.
- Temporal bias: Is the data representing all year around or an specific season or economic event. The trends can vary and it will affect the model and its effectiveness.