Jessica Duong
Data Immersion Achievement
Data Ethics & Applied Analytics

Intro to Predictive Analysis

Step 1: Understanding Regression

Linear regression predicts a continuous numeric outcome, such as revenue, temperature, and alert volume. It fits a straight-line relationship between preditors (inputs) and a numeric target (output).

Logistic regression is used when the outcome is categorical, most commonly binary, such as yes/no, 0/1, or churn/no churn. Instead of predicting a raw number, logistic regression predicts the probability that an observation belongs to a class (between 0 and 1). It uses the logit (log-odds) transformation so probabilities stay within a realistic range.

When to use logistic instead of linear (and why):

Use a logistic regression when the business question is a classification decision (or probability of an event), like:

- Will a customer churn (Yes/No)
- Will a loan default? (Default/No default)
- Will a transaction be fraudulent? (Fraud/ Not fraud)

Use linear regression when you need a quantity forecast, like:

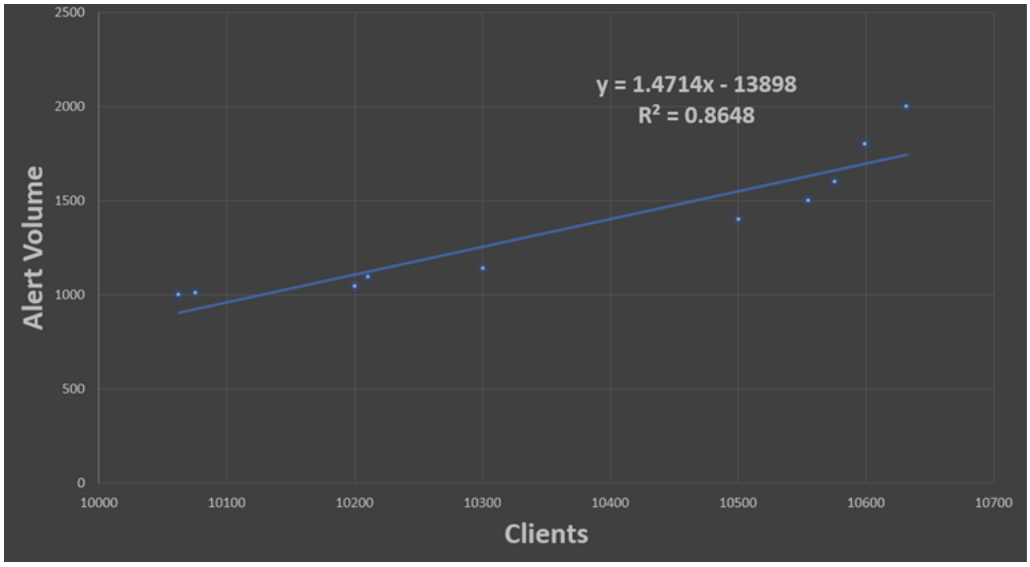
- How many alerts next month?
- How much revenue next quarter?
- What will the average handle time be?

Logistics = probability of a category

Linear = predicted numeric value

Jessica Duong
Data Immersion Achievement
Data Ethics & Applied Analytics

Step 2: More on Linear Regression



Relationship between variables:

The scatterplot and regression line show a positive linear relationship between number of clients and fraud alert volume: as the number of clients increases, the number of alerts generally increases as well.

The regression equation shown is:

Y = 1.4714x – 13898

This means that, on average, each additional client is associated with about +1.47 alerts (within the observed range of the data).

Model fitness (how good is it?):

The model is $R^2 = 0.8646$, which indicates a strong fit: about 86% of the variation in alert volume is explained by the number of clients in this dataset. That suggests this model is quite effective at predicting alert volume from client count.

Limitation:

R^2 does not prove causation, and this model may not generalize well outside the observed range. For example, if client growth changes the fraud max, alert rules, or reporting. So it's a

strong model for this dataset, but it should be validated before being used for forecasting in a live environment.

Step 3: Differentiating between Models

Scenario A:

Model type: Regression

In this scenario, I am predictive a continuous numeric value of oil price, and I want to test how unemployment rates, the predictors, relate to oil price changes (target).

Specific algorithm: Multiple Linear Regression

I would have multiple predictors (unemployment rates across top GDP countries) and one continuous outcome (oil price). Multiple linear regression is the simplest, most interpretable starting point for estimating and testing that relationship.

Scenario B:

Model type: Classification

In this scenario, I am predicting whether a customer will watch a specific movie (watch/not watch), which is a binary outcome.

Specific algorithm: Logistic Regression or Random Forest for a more advanced option

I would use logistic regression as it predicts a probability a customer will watch the movie and is easy to interpret and explain. If viewing behavior is complex or nonlinear and there are many features, random forest can improve accuracy by combining multiple trees.

Jessica Duong
Data Immersion Achievement
Data Ethics & Applied Analytics

Step 4: Bias in Your Data

If I helped collect the data in Step 2, several biases could show up:

- Sampling bias: If the data only came from certain regions, certain branches, or a short time window, the relationship may not represent the full bank population.
- Measurement bias: Fraud alerts can be inconsistent depending on rule changes, system upgrades, analyst behavior, or reporting errors. Reporting errors such as missing alerts, duplicated alerts, or different thresholds over time.
- Labeling/process bias: "Alert volume" is influenced by how aggressively the bank flags transaction, not only by true fraud. If the bank tightened alert rules while client count grew, it could inflate the relationship.
- Confounding/omitted variable bias: Other factors could drive alerts, such as economic conditions, new fraud patterns, staffing levels, or new detection tools. If those aren't included, client count may look more predictive than it really is.

Bias in collection or recording could distort the relationship and make the model look more accurate than it truly is in practice.

Sources:

https://pmc.ncbi.nlm.nih.gov/articles/PMC9747134/

https://statisticsbyjim.com/regression/interpret-r-squared-regression/

https://aws.amazon.com/compare/the-difference-between-linear-regression-and-logistic-regression/