Data Science Capstone: Comprehensive Study Guide

---

1. Linear Regression & Metrics

Foundational concepts for predicting continuous values.

- **Ordinary Least Squares (OLS):** The standard method for linear regression that minimizes the Sum of Squared Residuals (SSR) between observed and predicted values.
- **R-Squared (R^2):** A metric (0 to 1) representing the proportion of variance in the dependent variable explained by the independent variables.
- **Adjusted R-Squared:** A modified version of R^2 that adjusts for the number of predictors in the model. It decreases if a new term adds no value, preventing the "inflation" of accuracy by just adding noise.
- **RMSE (Root Mean Squared Error):** The square root of the average squared errors. It measures the "average" error in the same units as the target variable.
- **Heteroscedasticity:** A violation of OLS assumptions where the variance of errors is not constant (e.g., errors get larger as the predicted value grows). Visible as a "funnel shape" in residual plots.
- **Multicollinearity:** A situation where independent variables are highly correlated with each other, making coefficient estimates unstable and difficult to interpret.
- **Coefficient Interpretation:** In a standard linear model **y = b0 + b1*x**, a one-unit increase in x leads to a **b1** increase in y (holding all else constant).

2. Generalized Linear Models (GLMs)

Moving beyond the "Normal Distribution" assumption.

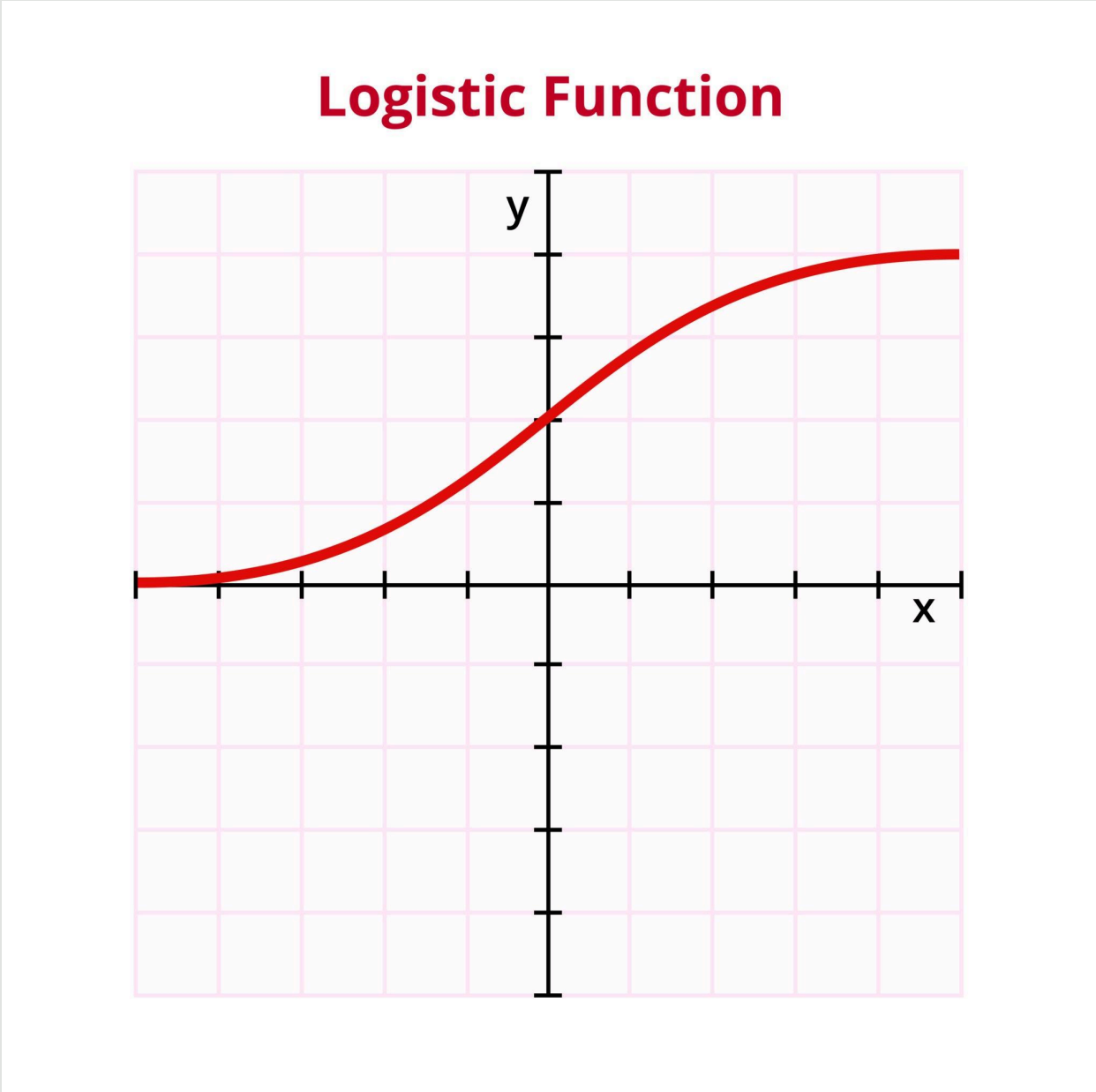
- **GLM Framework:** Extends linear regression to allow for response variables that have error distribution models other than a normal distribution.
- **Link Function:** The function that connects the linear predictor (eta) to the mean of the distribution (mu).
    - *Identity Link:* Used in Linear Regression (eta = mu).
    - *Log Link:* Used in Poisson Regression (eta = ln(mu)).
    - *Logit Link:* Used in Logistic Regression (eta = ln(mu / (1 - mu))).
- **Poisson Regression:** Used for **count data** (e.g., number of visitors per hour). Assumes the mean equals the variance.
- **Gamma Regression:** Used for continuous, positive, skewed data (e.g., insurance claim costs, wait times).
- **Tweedie Distribution:** A flexible family of distributions often used when data has a

mass at zero but is otherwise continuous (e.g., total rainfall).

3. Classification Models

Predicting categories (0 vs 1, Spam vs Ham).

- **Logistic Regression:** A classification algorithm that estimates the probability of an event occurring using the Sigmoid function to squash outputs between 0 and 1.



- Getty Images
- **Log-Odds (Logit):** The natural log of the odds ratio: **ln(p / (1-p))**. Logistic regression is linear in terms of log-odds.
- **K-Nearest Neighbors (KNN):** A non-parametric method that classifies a point based

on the majority class of its 'K' closest neighbors. Sensitive to the scale of data (requires normalization).
- **Support Vector Machine (SVM):** Finds the "hyperplane" that maximizes the **margin** (distance) between classes. Uses **Kernels** (Linear, RBF, Polynomial) to handle non-linear separation.
- **Naive Bayes:** A probabilistic classifier based on Bayes' Theorem, assuming that features are statistically independent (the "Naive" assumption).

4. Tree-Based Models & Ensembles

Handling non-linear relationships and interactions.

- **Decision Tree:** A flowchart-like structure where internal nodes represent tests on attributes (e.g., "Is Age > 30?"), and leaf nodes represent class labels. Prone to overfitting.
- **Entropy & Gini Impurity:** Metrics used to decide the best "split" in a tree. Lower values indicate purer nodes (more homogeneous classes).
- **Random Forest:** An **Ensemble** method that builds multiple decision trees on random subsets of data (Bagging) and random subsets of features. It reduces variance and overfitting compared to a single tree.
- **Grid Search:** A brute-force method for hyperparameter tuning. It works through a specified subset of the hyperparameter space (e.g., checking depths of 5, 10, and 15) to find the best performance.

5. Dimensionality Reduction

Simplifying complex data while keeping the "signal."

- **Curse of Dimensionality:** The phenomenon where data becomes sparse as the number of features (dimensions) increases, making distance-based algorithms (like KNN and Clustering) ineffective.
- **PCA (Principal Component Analysis):** An unsupervised linear technique that projects data onto orthogonal axes (Principal Components) that explain the maximum amount of variance.
- **Explained Variance Ratio:** The percentage of the dataset's information (variance) captured by each Principal Component.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding):** A non-linear technique mainly used for **visualization**. It preserves local structure (keeping similar points close) but distorts global distances.
- **LDA (Linear Discriminant Analysis):** A supervised dimensionality reduction technique that finds axes that maximize the separation between *classes*.

6. Model Evaluation & Interview Concepts

Thinking beyond "Accuracy."

- **Confusion Matrix:** A table showing True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).
- **Precision: TP / (TP + FP)**. "Of all the ones we predicted as fraud, how many were actually fraud?" (Use when False Positives are expensive).
- **Recall (Sensitivity): TP / (TP + FN)**. "Of all the actual fraud cases, how many did we catch?" (Use when False Negatives are expensive).
- **F1-Score:** The harmonic mean of Precision and Recall. Useful when you need a balance between the two.
- **ROC-AUC:** A performance measurement for classification problems at various threshold settings. AUC (Area Under Curve) represents the degree or measure of separability.
- **Train-Test Split:** Separating data into a training set (to build the model) and a testing set (to evaluate it) to prevent data leakage and measure generalization.

1.

1.