Notes

Logistic Regression

* Logistic Regression is a classification algorithm used when the target variable is binary (e.g., Yes/No, Buy/Sell, Fraud/Not Fraud), it predicts the probability that an event belongs to a particular category.

Example:

* Predict whether a stock will go up or down tomorrow based on historical data.

What is Logistic Regression?

* Unlike Linear Regression, which predicts continuous values, Logistic Regression predicts probabilities between 0 and 1.

When do we use Logistic Regression?

* Binary Classification
* Predicting Market Trends
* Spam Detection
* Credit Risk Analysis

How does Logistic Regression work?

* It calculates the probability score for an event happening.
* If p >= 0.5, we classify it as 1 (True, Yes, Up)
* If p < 0.5, we classify it as 0 (False, No, Down)

Types of Logistic Regression

* Binary Logistic Regression: Classifies into two categories (0 or 1).
* Multinomial Logistic Regression: Classifies into three or more categories (e.g., Buy, Hold, Sell).
* Ordinal Logistic Regression: Used when categories have a natural order (e.g., Low, Medium, High).

Performance Metrics for Logistic Regression

* Accuracy: % of correct predictions.
* Precision & Recall: Measures how well we classify the positive class.
* F1 Score: Balances Precision & Recall.
* ROC-AUC Score: Measures how well the model distinguishes between classes.

Example:

* A financial analyst wants to predict whether a stock’s closing price will be higher or lower than its opening price using historical data.
* The goal is to build a Logistic Regression Model that classifies stock movement as Up (1) or Down (0) based on selected market indicators.
* Objective:
  + Develop a binary classification model using Logistic Regression to predict whether a stock’s closing price will be higher (1) or lower (0) than its opening price.
* Solution approach:
  + Data Collection & Approach
    - We use historical stock data containing: Opening Price – Stock’s price at market open, Closing Price – Stock’s price at market close, Volume – Number of shares traded, Daily High & Low – Intraday price range, Technical Indicators – Moving, Averages, RSI, etc.
    - We create a binary target variable:
      * 1 (Up) 🡪 If closing price > Opening Price
      * 0 (Down) 🡪 If closing price <= Opening Price
  + Train a Logistic Regression Model
    - Split the data into training (80%) and testing (20%) sets.
    - Train the model using stock market features as input.
    - Predict whether the stock will go Up or Down.
    - Evaluate performance using Accuracy, Precision, Recall, and AUC-ROC.
  + Interpretation of Results
    - Accuracy Score: How well the model classifies Up vs. Down.
    - Precision & Recall: Measures prediction predictability.
    - Confusion Matrix: Shows correct and incorrect predications.
    - ROC-AUC Score: Tells us how well the model separates the two classes.
  + Expected Business Impact
    - Traders can make informed buy/sell decisions based on model predictions.
    - Portfolio managers can assess market trends before making large investments.
    - Algorithmic traders can automate trades using AI – driven stock movement predictions.

Evaluation Metrics for Logistic Regression

* Confusion Matrix
* Accuracy Score
* Precision & Recall
* F1-Score
* ROC-AUC

Confusion Matrix

* A table that shows the number of correct and incorrect predictions.
* Interpretations:
  + True Positives (TP) – Model correctly predicted stock went up.
  + True Negative (TN) – Model correctly predicted stock went down.
  + False Positives (FP) – Model predicted stock went up, but it actually went down (Type 1 Error).
  + False Negatives (FN) – Model predicted stock went down, but it actually went up (Type 2 Error).

Accuracy Score

* Measures how many predictions were correct overall.
* Good when data is balanced.
* Misleading for imbalanced data.

Precision & Recall

* Useful when false positives or false negatives matter more.
* Precision (Positive Prediction Value) 🡪 “When the model says a stock will go Up, how often is it correct?”
* Recall (Sensitivity, True Positive Rate) 🡪 “Of all the stocks that actually went Up, how many did the model correctly predict?”
* High Precision: Few false positives (better for stock trading to avoid buying wrongly).
* High Recall: Few false negatives (better for risk analysis to avoid missing important signals).

F-1 Score (Balances both Precision & Recall)

* A good balance when both false positives and false negative matter.
* Best for cases where false positives and false negatives have different impacts.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve)

* Measures how well the model separates the two classes (Up & Down).
* ROC Curve 🡪 Plots True Positive Rate (Recall) vs. False Positive Rate.
* AUC (Area Under Curve) 🡪 1.0 = perfect model, 0.5 = random guessing.
* Higher AUC = Better model at distinguishing Up vs. Down Stocks.
* Best for imbalanced datasets.