Unit - 5

Predictive Analytics

Linear Least squares:

→ We often want to draw a straight line that best represents the general trend of the points. This is called finding the regression line, and we use Least square method

→ We minimize the sum of the squares of the vertical distances (residuals) between each point and line.

→ In simple words: we make sure the line is as close as possible to all points on average.

Least Square Method:

→ The Least Squares Method is a statistical technique used to find the line (or curve) that best fit a set of points.

For straight line: $y = mx + b$

[y - dependent variable
x - independent variable
m - slope
b - y-intercept]

Limitations of Least Squared Method:

→ It only shows the relationship between two variables.
→ It does not explain other causes (or) effects.
→ It doesn't work well if data is not evenly distributed
→ It is very sensitive to outliers., which can distort the results.

Least Square Method Graph:

→ The line we find shows the trend between x and y.
→ A smaller residual means a better fit.
→ Two types of residuals:
  → Vertical Residuals
  → Perpendicular Residuals

given data points:

$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

We find m and b using:

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{(\Sigma y) - m(\Sigma x)}{n}$$

Example:

Sol

x	y	xy	x^2
1	2	2	1
2	4	8	4
3	5	15	9
4	8	32	16
5	7	35	25

$n = 5$

Total $\Sigma x = 15$ $\Sigma y = 26$ $\Sigma xy = 92$ $\Sigma x^2 = 55$

Find slope:

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{5(92) - (15)(26)}{5(55) - (15)^2}$$

$$= \frac{460 - 390}{275 - 255} = \frac{70}{50} = 1.4$$

Find intercept b:

$$b = \frac{(\Sigma y) - m(\Sigma x)}{n}$$

$$= \frac{26 - 1.4(15)}{5} = \frac{26 - 21}{5} = \frac{5}{5} = 1$$

Regression line,

$$y = 1.4x + 1.$$

(11). Implementation of Predictive Analytics

i) Follow a Standard Process (like CRISP-DM)

→ CRISP-DM (Cross Industry Standard Process for Data Mining) is a common framework for building predictive analysis systems.

ii) Key Actions for Successful Implementation:

a) Define your goals:

→ Be clear on what you want to predict
→ Different goals need different kinds of data.

b) Set Success Criteria:

→ Know how will you measure success.
→ Set realistic objectives.

c) Develop your Infrastructure:

→ Predictive analysis need lots of data.
→ Built strong systems for data collection and sharing.
→ Keep communication open among all teams.

d) Start Early Predictions:

→ Use historical data even it's imperfect
→ Early predictions help to your test.
→ Adjust the model as you gather data.

e) Keep your data updated:

→ Predictive model fails data is old.
→ Consumer behaviour changes quickly.
→ Stay Accurate and regular update.

Technical Skills Needed for Predictive Analytics Implementation:

i) Software Engineer's Role:
   → Needed a proactive mindset.

ii) Required Skills:
   → knowledge of machine learning and data Science
   → Strong in Regression models
   → Skilled in classification models.

iii) Advanced Skills:
   → Know model optimization and validation methods.
   Eg:
   → PCA
   → k-fold Cross validation.

iv) Big Data Handling:
   → Must work with large datasets.
   → Tasks involve
      * Data Collection
      * Data Preprocessing
      * Data Storage
      * Data analysis
      * Data visualization.

v) Tools and Platforms:
   → Tools like: Apache Spark, Tableau, Matlab.
   → Frame work for Machine Learning:
      * TensorFlow
      * Tensor Board
      * Py Torch, Keras, Sonnet, DL4J.

vi) Programming Languages:
   → Python
   → R

vii) Using IBM Tools: (Watson Studio, Watson ML)

viii) Specialized AI Capabilities: (NLP) / Tools for NLP [NLTK, Apache

(iii). Goodness of Fit:

→ Goodness of fit checks how well observed (actual) data matches expected (theoretical) data.

→ It helps measure how well a statistical model fits the real-world data.

→ The methods are usually applied to univariate data.

(ii) Goodness-of-Fit Tests:

a) Chi-Square Test:

→ Check if observed data matches a specific expected distribution.

→ Process: compared observation vs expected values to accept or reject null hypothesis.

b) Kolmogorov-Smirnov (K-S) Test:

→ Checks if two samples come from same distribution.

→ Process: Measures the maximum difference between the two empirical distribution functions.

→ Also used to test normality assumption in ANOVA.

c) Anderson-Darling (A-D) Test:

→ Checks goodness of fit with more focus on the tails (ends) of distribution.

→ To find known critical values.

→ More Sensitive to differences in tails to K-S Test.

d) Shapiro-Wilk (S-W) Test:

→ Tests if a sample follows a normal distribution.

→ Used when you have one continous variable and want to check for normality.

5. Weighted Reasoning.

Imbalanced class Distribution : In ML, sometimes one class (category) has many more (or) much fewer data points than other.

Example:

→ In fraud detection : 99% normal transaction(vs). 1% fraudulent ones.

→ Detecting Fraud

ii) Resampling Method:

→ Randomly selecting data points again from existing dataset.

→ Why Resampling?

→ To create a larger dataset.

→ To estimate properties of a dataset when you have limited data.

→ Useful for * Training Models more effectively

* Making Predictions more reliable.

iii) Common Resampling Methods:

a) Bootstrapping with Cross-validation:

Bootstrapping:

→ Repeatedly sample from the dataset to create many new datasets.

→ Each new dataset has duplicate points.

Cross-validation:

→ Test the model on different subsets of the data to evaluate its performance.

b) Cross-Checking (Cross-validation):

→ Methods where you split and test the data in different ways to validate the model.

iv) Types of Cross-validation:

a) Validation set Approach:

→ Randomly split the dataset into two:

* Training Set: Used to built the model
* Validation Set: Used to test the Model

b) Leave-One-out-Cross-Validation (LOOCV):

→ More accurate than validation set approach.

→ For every data point:

* Leave one observation out as valid set
* Use all other data points as traing set.
* Train the model, then test it on left-out point.

E-6

Multiple Regression:

→ Multiple Regression is a statistical method used to predict the values of one dependent variable (Y) based on the values of two (or) more independent variables ($X_1, X_2, \ldots, X_p$).

→ It is an extension of simple Linear Regression, which only uses one independent variable.

→ In multiple regression, many factors together influence the outcome.

Formula:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p + e$$

Where:

Y = Dependent variable (the one we predicting)

→ $X_1, X_2, \ldots, X_p$ (Independent variables)

→ b_0 = Intercept

→ $b_1, b_2, \ldots b_p$ = Regression coefficients

→ e = Error Term

When is Multiple Regression Used:

→ To find relationships between several factors in outcome.

→ To predict outcomes based on several inputs.

→ Analyze cause-and-effect models.

→ Used in business, finance, medicine, technology.

Example:

Predicting Fuel Company Stock Price

Step 1: Identify the Predictive Variables (Xs):

The analyst wants to predict the stock price (Y) of fuel company.

The independent variables chosen:

X_1 = Interest Rate = 5% = 0.05

X_2 = Crude oil Price = $50 per barrel

X_3 = Transport Cost = $25 per 100 barrels.

Step 2: Formula Setup:

Formula:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Given:

$b_0 = 50$ (Starting stock Price)

Co-efficients:

$b_1 = 0.015$

$b_2 = 0.33$

$b_3 = 0.8$

TIME TABLE

Step 3 Plug in values

Substitute:

$$Y = 50 + (0.015 \times 0.05) + (0.33 \times 50) + (0.8 \times 25)$$

Add:

$$Y = 50 + 0.00075 + 16.5 + 20$$

$$Y = 86.50075$$

$$Y \approx 86.5$$

Result:

The predicted stock price is $86.5

→ Stock Prices moves from $50 to $86.5.

⇒ This also shows sensitivity of stock Price.

5.7

Non-Linear Regression.

→ Non-Linear Regression is a method used when the relationship between the independent variable X and the dependent variable Y is not a straight line (not the form of $Y = mx + b$).

→ Relationship is curved - like exponential, logarithmic, trigonometric

Goal of Non-Linear Regression:

-> Main goal is to minimize the sum of squares:

* For each data point, find the difference between the actual Y and the predicted Y from the model.

* Square that difference (so negative don't cancel out)

* Add up all these squared differences.

	Linear	Non-Linear
Shape	Straight Line	Curved Line
Formula	Linear in Parameter	Parameter appear in non-linear ways
Solving	Direct formula	Need iterative methods.

Example: Population growth

→ Predicting population growth over times is often non-linear. (S-shaped curve).

→ Early growth is slow → then fast → then slow again (logistic curve).

5.8

Logistic Regression

→ Logistic Regression is a statistical method used to predict binary outcomes - two possible outcomes,

→ Yes/No, Pass/Fail.

Purpose of Logistic Regression:

→ Predict the probability that something belongs to one of the two classes.

→ Analyze how input variables (age, gender, income) affect an outcome (like "buy" (or) "not buy")

How Logistic Regression works:

1. Takes input variables (SAT score, GPA).
2. Calculates a score (called "log-odds")

$$\text{log-odds} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

3. Converts the log-odds into a probability using logistic function.

$$P = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n)}}$$

Logistic vs Logistics

→ Logistics (Supply chain management (moving goods, delivering))
→ Logistic = (Math term related to probability curves (S-curves).

Applications in Logistic Regression

→ Marketing
→ Banking
→ Healthcare
→ Weather Apps
→ Political Polling

Why is Logistic Important:

→ Simplifies probability calculations into easy-to-use Models.
→ Makes ML possible for binary classification problems.

Example:

Predict if a loan application approves (1) or denied (0).

Key Assumptions of Logistic Regression:

1. Independent variables should not be too similar
2. Observations must be unique
3. Linear relationship between input variables and log-odds.
4. Large sample size is better - more data = More accurate predictions
5. Binary Outcome. (Yes/No, Pass/Fail).

Advantage	Disadvantage
→ Easy to implement	→ Struggles with complex relationships
→ Works well with simple, linearly seperable data	→ Needs large sample size for accurate results.
→ Interpretable co-efficients	→ Only predicts two classes without modification

Tools used in Logistic Regression

→ SPSS, SAS
→ Python
→ R
→ Excel
→ Data Warehouse

5.9

Parameter Estimation:

→ When we build a mathematical model of something real - like a battery - we need to put numbers into the model.

Example of these numbers (parameters) in batteries:

→ Internal resistance
→ Diffusion rates
→ Reaction rates

→ Using real measurements (like battery voltage, current, temperature) to calculate the right values for these parameters, so that the model behaves like the real battery.

Parameter Estimation Important:

i) Accuracy - If the model has wrong numbers, it will give wrong predictions (eg. how long battery lasts). Correct parameters make the model accurate.

ii) Adapt to Changes - Battery changes over time (due to aging, temperature, etc). Estimation allows to model to update as battery changes.

iii) Optimization and Safety:

→ In BMS (Battery Management System) we use the model to control charging and discharging.

→ If the parameters are wrong, the system could overcharge (or) damage battery.

→ Correct Parameter help optimize and protect the battery.

How is Parameter Estimation Done: (Techniques)

i) Least Square Estimation:

→ Minimize the square of the difference:
  * What the model predicts
  * What we actually measured
→ Simple and very popular, when data has noise.

ii) Maximum Likelihood Estimation (MLE):

→ Find Parameters that make the observed data most likely under a probability model. (Works when you know how noisy your measurement is)

iii) Bayesian Estimation:

→ Start with a guess (belief) about parameters, and improve it as new data comes. (Mixes old knowledge with new).

iv) Gradient Descent:

→ Start somewhere and slide downhill (reduce error step-by-step) to find the best (parameters).

v) Genetic Algorithms:

→ Try lots of parameter sets, and evolve the best ones, like survival of the fittest.

vi) System identification: Plays with the battery (like applying a signal) and watch how it responds. Then guess the model parameters.

Special Case: Batteries:

Two Major Types of Battery:

① Electrochemical Models:

→ These models describe real physical reactions inside the battery.

→ Parameter includes: diffusion co-efficients, reaction rates, etc.

Techniques:

i) Curve Fitting: Fit model curves to experimental curves.

ii) Optimization Algorithms: Find the best parameters.

iii) Parameter Sweeping: Try a range of values systematically.

Also, we do Sensitivity Analysis,

→ Find out which parameters matter the most.

→ Helps simplify the model and focus on important factors.

② Equivalent Circuit Models (ECMs):

→ Simplified models with resistors and capacitors.

→ Parameters: Resistance values, Capacitance values.

Techniques used:

→ Least Square Methods: (For both linear and Non-linear).

→ System identification

→ Frequency Respose Analysis (Nyquist Plot)

→ Time Domain Methods

→ Black-Box Modeling (Using ML like neural networks. when internal structure is unknown).

5.10

Time Series Analysis:

→ Studying how data changes over time, using data collected at regular time intervals (like every hour, day, month, etc).

Instead of randomly collecting data, in time series we:

→ Collect data points regularly (eg. every hour (or) every day)

→ Analyze how things change across time.

→ Time series is not just about data, it's how data changes over time.

Why do we need a lot of data points:

→ If you only have a few data points, you might miss important trends (or) get fooled by random noise.

→ More data points = More reliable analysis

→ You can also catch seasonal effects (Eg. Pongal sales always increase).

What can we do with Time Series:

i) Understand patterns

ii) Detect Trends

iii) Predicts the future (Forecasting)

TIME TABLE

Why Organization Use Time Series Analysis:

1. Understanding the Past and Present:
   * See long-term trends (or) repeating patterns
   * Find out why something is happening
   * Use visualizations to spot patterns easily.
2. Predicting the Future:
   * Guess what might happen next based on past patterns.
   * This helps in better planning.

Example:

→ A School collects 5 years of students achievements data.

→ By analyzing the data over time,
   * Spot which students were at risk.
   * See if students were improving or falling behind.
   * Make better decisions about where to focus help.

5.11

Moving Averages:

→ A Moving Average (MA) is a technique to smooth out price data over time.

→ It reduces the impact of short term, random price fluctuations.

→ It gives the clear view of the trend.

→ Commonly used in Finance and stock Marketing.

Why we use Moving Averages:

→ Smooths the data
→ Shows trend direction clearly
→ Helps find support/resistance levels.
→ Confirms trend reversals using crossovers.

Longer MA = More lag

Eg: 200-day MA lags more than 20-day MA.

Types of Moving Averages

1. Simple Moving Average (SMA):

→ Simple average of stock prices over a specific number of days.

Formula:

$$SMA = \frac{\text{Sum of closing prices over } n \text{ days}}{n}$$

Example:

If closing prices over 5 days are: 100, 102, 104, 106, 108

$$SMA = \frac{(100+102+104+106+108)}{5} = 104$$

2. Exponential Moving Average (EMA):

→ Gives more weight to recent prices (reacts faster to price changes).

Two steps to calculate,

1. First, find the SMA (starting value)
2. Then apply smoothing factor:

$$\text{Smoothing factor} = \frac{2}{(\text{no. of days} + 1)}$$

Eg: 20 days, $\frac{2}{20+1} = 0.0952$

3. EMA Formula:

EMA today = (Price today × s) + (EMA yesterday × (1-s)).

where, s - smoothing factor.

Interpretation of Moving Averages:

Rising MA - Upward Trend

Falling MA - Downward Trend

Short MA crosses above Long MA - Bullish signal (Buy)

Short MA crosses below Long MA - Bearish signal (Sell).

5.12

Missing Values:

→ Missing values are ~~groups~~ gaps in the dataset where information is missing for a variable.

→ They looks like:

→ Blank cells

→ "Null" entries

→ "NA" (Not Available)

→ "Unknown"

→ Special numbers like -999 to missing.

Why are ~~possible~~ Missing values a problem:

→ Smaller sample size

→ Introduce Bias

→ Block certain Analyses.

Why do dataset have missing values:

→ Human errors
→ Technical issues
→ Privacy concerns
→ Data Processing errors
→ Nature of variable

Types of Missing values.

MCAR (Missing Completely At Random) - Missingness is Random.

MAR (Missing At Random) - Missingness depends on other variables.

MNAR (Missing Not At Random) - Missingness depends on the missing value itself.

How are missing values shown:

→ Blank Cells - Empty set in Excel sheet
→ Special codes - "NA", "NULL", "Unknown"
→ Special numbers - (-999), (-1) used.

Importance to Handle Missing values correctly:

→ Maintain data quality
→ Avoid bias in your results.
→ Make statistical models work correctly.

5.13

Serial Correlation:

→ Serial correlation happens when a variable is correlated with its own past values over time.

Example: If today's stock price is highly related to yesterday's price, then the prices show serial correlation.

→ If serial correlation exists:
The Data follows the pattern.

→ If no serial correlation:
Observation is independent (no memory of past)

Serial Correlation in Time Series Models:

→ Errors (the difference between real and predicted values) can also show serial correlation.

→ If a model overpredicts one year, it may keep overpredicting in the next years → this is error serial correlation.

Serial correlation in Technical Analysis (Finance):

→ Technical analysts (especially in stock markets) use serial correlation to:
* Find patterns in stock Prices.
* Predict future price movements based on past behavior.
* Identify profitable opportunities (Buy/Sell)

Where did Serial Correlation Come From:

Engineering - To study how a signal changes over time

Economics & Finance - To study how economic/market data is connected over time.

5.14 Auto-Correlation:

→ Autocorrelations measures how strongly a variable relates to itself over different time intervals.

→ It checks the relationship between a variable and its past (lagged) values.

→ Also called Serial Correlation.

How Autocorrelation Works:

→ A value at Time T is compared with its value at T_1, T_2, etc.

→ Helps to detect patterns and trends across a time series.

Example:

→ Temperature on one day often correlates with the temperature on the next day (positive auto correlation).

Positive (vs) Negative AutoCorrelation

Positive – Rise today - Rise Tomorrow [Smooth upward]

Eg: Temperature rising steadily.

Negative – Rise today – Fall tomorrow [zig-zag pattern (up and down)]

Eg: See-saw (or) alternating pattern.

Lag in Autocorrelation:

→ Lag = Time gap between two observations

→ Lag 1 autocorrelation:

Today's values (vs) Yesterday's value.

→ Lag 30 autocorrelation:

Today's value (vs) same date next month.

TIME TABLE

Test for Assumption:

→ Durbin - Watson (DW) Statistic is commonly
→ DW value Ranges.

DW Value	Meaning
2	No autocorrelation.
Close to 0	Strong Positive autocorrelation
Close to 4	Strong negative autocorrelation

→ Use statistical software (like R, Python, SPSS) to DW.

Auto correlation in Technical Analysis:

→ Technical analysts use autocorrelation:
* Find short-term trends.
* Predict future price movement.

Example:

→ If a stock has been rising and shows positive autocorrelation, it is likely to rise futher for a short time.

→ Trader may buy and hold for short-term profit

→ Autocorrelation focuses on past patterns, not company fundamentals.

→ Best for short-term trading decisions, not long-term investment.