ABHYASMITRA.IN

STUDY NOTES

BUSINESS ANALYTICS

Unit III – Predictive and Prescriptive Analytics

3.1 Trend Lines

A trend line (also called a line of best fit or regression line) is a straight or curved line drawn through a scatter plot or time-series chart that best represents the overall direction (trend) of the data. Trend lines help analysts identify whether data is increasing, decreasing, or staying flat over time.

> **Definition**
> *"A trend line is a line superimposed on a chart that reveals the general pattern or direction of the data over time."*

Types of Trend Lines

- **Linear Trend Line:** A straight line used when data increases or decreases at a roughly constant rate. Best for data with a consistent direction. Example: Steady growth in annual revenue.
- **Exponential Trend Line:** A curved line used when data increases (or decreases) at an accelerating rate. Example: Viral spread of a social media post.
- **Polynomial Trend Line:** A curved line used when data fluctuates. The degree of the polynomial determines how many bends the line has. Example: Sales with seasonal peaks and troughs.
- **Logarithmic Trend Line:** A curved line that rises quickly and then levels off. Example: Product adoption curve.
- **Moving Average Trend Line:** Smooths out short-term fluctuations to show the long-term trend. Example: 30-day moving average of stock prices.

R-Squared (R^2) Value

The R^2 value (coefficient of determination) measures how well the trend line fits the data. It ranges from 0 to 1. An R^2 of 1 means the trend line perfectly fits the data. An R^2 of 0.85 means the trend line explains 85% of the variation in the data.

3.2 Regression Analysis – Linear and Multiple

Regression analysis is a statistical method used to examine the relationship between one dependent variable (the outcome we are trying to predict) and one or more independent variables (the factors that influence the outcome). It is one of the most widely used predictive analytics techniques.

> **Definition**
> *"Regression analysis is a statistical process for estimating the relationships among variables. It helps understand how the dependent variable changes when any one of the independent variables is varied."*

Simple Linear Regression

Simple linear regression examines the relationship between ONE dependent variable (Y) and ONE independent variable (X).

Equation: $Y = a + bX + \varepsilon$

- **Y:** The dependent variable (what we want to predict). Example: Sales revenue.
- **X:** The independent variable (the predictor). Example: Advertising spend.
- **a (Intercept):** The value of Y when X = 0.
- **b (Slope):** How much Y changes for each one-unit increase in X.
- **ε (Error term):** The random error or residual not explained by the model.

Example: If b = 5 and a = 10,000, it means for every additional ₹1,000 spent on advertising, sales increase by ₹5,000.

Multiple Linear Regression

Multiple linear regression examines the relationship between ONE dependent variable (Y) and TWO or MORE independent variables (X_1, X_2, X_3...).

Equation: $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + \varepsilon$

Example: Predicting a house price (Y) using area in square feet (X_1), number of bedrooms (X_2), and location score (X_3).

Aspect	Simple Linear Regression	Multiple Linear Regression
Number of Predictors	1 (X)	2 or more (X_1, X_2, X_3...)
Equation	$Y = a + bX$	$Y = a + b_1X_1 + b_2X_2 + \ldots$
Complexity	Low	Higher
Example	Sales vs Ad Spend	Sales vs Ad Spend + Season + Region

3.3 Predictive Modeling

Predictive modelling is the process of using statistical and machine learning techniques to build a model that can predict future outcomes based on historical data. It is the core of predictive analytics.

> **Definition**
> *"Predictive modelling is the process of creating a statistical or machine learning model that forecasts the probability of future events based on patterns discovered in historical data."*

Types of Predictive Models

- **Classification Models:** Predict which category an observation belongs to. Output is a class label. Example: Will a customer churn? (Yes/No). Algorithms: Logistic Regression, Decision Trees, Random Forest, Neural Networks.

- **Regression Models:** Predict a continuous numerical value. Example: What will next month's sales be? Algorithms: Linear Regression, Multiple Regression, Ridge Regression.
- **Time-Series Models:** Predict future values based on past values in time-ordered data. Example: Forecast electricity demand for next 7 days. Algorithms: ARIMA, Exponential Smoothing, Prophet.
- **Clustering Models:** Group similar observations together. Example: Identify customer segments based on purchase behaviour. Algorithms: K-Means, Hierarchical Clustering.

Predictive Modelling Process

- **Step 1 – Define Objective:** What are you trying to predict?
- **Step 2 – Collect and Prepare Data:** Gather historical data, clean it, and engineer relevant features.
- **Step 3 – Select Algorithm:** Choose the right model type based on the problem.
- **Step 4 – Train the Model:** Fit the model on the training dataset.
- **Step 5 – Validate the Model:** Test the model on unseen test data and measure accuracy.
- **Step 6 – Deploy and Monitor:** Use the model for predictions and monitor its performance over time.

3.4 Forecasting Techniques

Forecasting is the process of making predictions about future values based on historical data and analysis. Accurate forecasting helps businesses plan inventory, budget, staffing, and strategy.

1. Time-Series Forecasting

Uses historical data collected over regular time intervals to predict future values. Key components of time-series data:

- **Trend:** Long-term increase or decrease in the data over time.
- **Seasonality:** Regular, predictable patterns that repeat at fixed intervals (daily, weekly, yearly). Example: Ice cream sales peak every summer.
- **Cyclical Patterns:** Longer-term fluctuations linked to economic cycles (not fixed intervals).
- **Irregular/Random:** Unpredictable random variations.

2. Moving Average

The Moving Average method forecasts the next value as the average of the most recent N values. It smooths out short-term fluctuations. A Simple Moving Average (SMA) treats all past values equally. A Weighted Moving Average (WMA) gives more weight to recent values.

3. Exponential Smoothing

Exponential smoothing gives exponentially decreasing weight to older observations. Recent observations have more influence on the forecast. The smoothing parameter α (alpha) controls how much weight is given to the most recent value ($0 < \alpha < 1$). Higher α = more reactive to recent changes.

4. ARIMA (AutoRegressive Integrated Moving Average)

ARIMA is a sophisticated time-series model that uses the relationship between current and past values (autoregression), differences to make the series stationary (integration), and past forecast errors (moving average). It is denoted as ARIMA(p, d, q) where p, d, q are parameters.

5. Regression-Based Forecasting

Uses regression analysis to forecast future values based on relationships with other variables (predictors). Example: Forecasting demand based on price, promotions, and season.

3.5 Data Mining – Definition and Approaches

> **Definition**
> *"Data mining is the process of discovering meaningful patterns, correlations, anomalies, and insights from large datasets using statistical, mathematical, and machine learning techniques."*

Data mining is sometimes called Knowledge Discovery in Databases (KDD). It is the process of automatically finding hidden patterns in very large amounts of data that would be impossible to find manually.

Key Approaches in Data Mining

- **Classification:** Assigns data items to predefined categories. Example: Classify email as spam or not spam. Algorithms: Decision Trees, Naive Bayes, SVM, Neural Networks.
- **Clustering:** Groups similar data items together without predefined categories. Example: Group customers into segments based on purchasing behaviour. Algorithms: K-Means, DBSCAN.
- **Association Rule Mining:** Finds relationships and co-occurrence patterns. Example: "Customers who buy bread also buy butter" (Market Basket Analysis). Algorithm: Apriori.
- **Regression:** Predicts continuous values. Example: Predict house prices based on features.
- **Anomaly (Outlier) Detection:** Identifies unusual patterns that differ significantly from the norm. Example: Detect fraudulent credit card transactions.
- **Sequence Mining:** Discovers patterns in sequences of events over time. Example: "Customers who buy a camera then buy a memory card within a week".

3.6 Data Exploration and Reduction

Before applying mining algorithms, data must be explored and often reduced to a manageable size. With Big Data, working with the entire raw dataset can be computationally expensive and time-consuming.

Data Exploration

Data exploration involves understanding the structure, distribution, and relationships in the data using statistical summaries and visualisations before formal analysis.

- Compute descriptive statistics (mean, median, min, max, standard deviation).
- Visualise distributions using histograms and box plots.
- Check for missing values, outliers, and inconsistencies.
- Explore correlations between variables using correlation matrices.

Data Reduction Techniques

Data reduction reduces the volume of data while preserving the information needed for analysis.

- **Dimensionality Reduction:** Reducing the number of variables (features). The most common technique is Principal Component Analysis (PCA), which transforms correlated variables into a smaller set of uncorrelated components.
- **Sampling:** Using a representative sample of the data instead of the full dataset. Random sampling, stratified sampling.
- **Feature Selection:** Selecting only the most relevant variables for the analysis, removing redundant or irrelevant ones.
- **Data Aggregation:** Combining detailed data into summary-level data. Example: Aggregating daily sales into monthly totals.
- **Data Compression:** Encoding data in a more compact form without losing significant information.

3.7 Data Mining and Business Intelligence

Data Mining and Business Intelligence (BI) are closely related but distinct concepts. Both help organisations make better decisions using data, but they focus on different aspects of the data analytics process.

Aspect	Data Mining	Business Intelligence (BI)
Focus	Discovering hidden patterns	Reporting and monitoring known metrics
Output	Patterns, predictions, rules	Dashboards, reports, KPIs
Approach	Exploratory, discovery-driven	Structured, query-driven
Users	Data scientists, analysts	Business managers, executives
Tools	Python, R, Weka, RapidMiner	Power BI, Tableau, SAP BI
Time Orientation	Future (predictive)	Past and present (descriptive)

In practice, BI and Data Mining are used together: BI provides the dashboards and reports that describe what is happening, while Data Mining uncovers the hidden reasons why it is happening and predicts what will happen next.

3.8 Classification, Association, and Cause-Effect Modeling

Classification

Classification is a supervised learning technique that assigns new observations to one of a set of predefined categories (classes) based on a training dataset.

- **Binary Classification:** Two possible outcomes. Example: Loan approved (Yes/No), Email is Spam (Yes/No).
- **Multi-class Classification:** Three or more outcomes. Example: Classify a news article as Sports, Politics, or Technology.
- **Common Algorithms:** Decision Trees, Logistic Regression, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbours (kNN), Neural Networks.

Association Rule Mining

Association rules discover interesting relationships between variables in large datasets. It is best known for Market Basket Analysis – finding which products are frequently purchased together.

- **Support:** How frequently the rule occurs. Support(A→B) = Transactions containing both A and B / Total transactions.
- **Confidence:** How often the rule is correct. Confidence(A→B) = Transactions containing both A and B / Transactions containing A.
- **Lift:** How much more likely B is to be bought when A is bought, compared to B being bought alone. Lift > 1 means a positive association.
- **Example:** Apriori algorithm finds: {Bread, Butter} → {Milk} with Support=20%, Confidence=70%, Lift=2.1. This means customers who buy bread and butter are 2.1 times more likely to also buy milk.

Cause-Effect Modeling

Cause-effect (causal) modelling goes beyond correlation to identify which variables directly cause changes in others. This is critical for making effective business interventions.

- **Correlation vs Causation:** Correlation means two variables move together. Causation means one directly causes the other. Example: Ice cream sales and drowning rates are correlated (both increase in summer) but ice cream does not cause drowning.
- **A/B Testing:** The gold standard for establishing causation in business. Two versions (A and B) are tested on different user groups; the difference in outcomes reveals the causal effect of the change.
- **Regression Analysis:** When properly designed with controls, regression can establish causal relationships.

- **Structural Equation Modeling (SEM):** A statistical technique that models complex causal relationships between multiple variables simultaneously.

3.9 Overview of Linear Optimization

Linear optimisation (also called Linear Programming, LP) is a mathematical technique for finding the best outcome (maximum profit or minimum cost) in a mathematical model whose requirements are represented by linear relationships.

> **Definition**
> *"Linear Programming is a method to achieve the best outcome (such as maximum profit or minimum cost) in a mathematical model whose requirements are represented by linear relationships."*

Components of a Linear Programming Problem

- **Objective Function:** The mathematical expression of what we want to maximise (profit, revenue) or minimise (cost, time). Example: Maximise $Z = 5X_1 + 4X_2$
- **Decision Variables:** The unknowns we need to find. Example: X_1 = units of Product A, X_2 = units of Product B.
- **Constraints:** The limitations or restrictions. Example: $2X_1 + X_2 \leq 100$ (labour hours), $X_1, X_2 \geq 0$ (non-negativity).
- **Feasible Region:** The set of all possible solutions that satisfy all constraints. The optimal solution lies at a corner (vertex) of the feasible region.

Methods for Solving LP Problems

- **Graphical Method:** Used for problems with two decision variables. Plot constraints on a graph, identify the feasible region, and find the optimal corner point.
- **Simplex Method:** An iterative algebraic algorithm that moves from one corner of the feasible region to the next, improving the objective function until no further improvement is possible.
- **Big M Method / Two-Phase Method:** Extensions of the Simplex Method for problems with equality or greater-than constraints.

Business Applications of LP

- Production planning: How many units of each product to produce to maximise profit.
- Resource allocation: How to assign limited resources (labour, machines, budget) to maximise output.
- Transportation problems: How to ship goods from multiple warehouses to multiple destinations at minimum cost.
- Diet/Nutrition problems: Minimum-cost diet that meets all nutritional requirements.

3.10 Non-Linear Programming and Integer Optimization

Non-Linear Programming (NLP)

Non-Linear Programming deals with optimisation problems where the objective function or at least one constraint is non-linear. Real-world business problems are often non-linear.

- **Example:** Profit = Revenue – Cost. If Revenue = Price × Quantity and Price depends on Quantity (demand curve), the relationship becomes non-linear.
- **Challenges:** NLP is much harder to solve than LP because the feasible region may not be convex, meaning there can be multiple local optima.
- **Methods:** Gradient Descent, Newton's Method, Lagrange Multipliers, Quadratic Programming.
- **Applications:** Portfolio optimisation in finance, engineering design, pricing optimisation.

Integer Programming (IP)

Integer Programming (also called Integer Linear Programming, ILP) is a special case of LP where some or all decision variables must take integer (whole number) values. This is important in real-world problems where fractional answers are not meaningful.

- **Pure Integer Programming:** All decision variables must be integers. Example: How many machines to buy.
- **Mixed Integer Programming (MIP):** Some variables are integers, others are continuous. Example: Whether to open a store (0 or 1) combined with how much stock to order (continuous).
- **Binary Integer Programming:** Variables can only be 0 or 1. Example: Should we include project A in the portfolio? (Yes=1, No=0).

Cutting Plane Method

The Cutting Plane Algorithm is a method to solve integer programming problems. It works by:

- **Step 1:** Solve the LP relaxation (ignore the integer requirement and allow fractional solutions).
- **Step 2:** If the solution is integer, stop – it is the optimal integer solution.
- **Step 3:** If not, add a "cutting plane" – a new constraint that cuts off the fractional solution from the feasible region but does not exclude any integer solutions.
- **Step 4:** Repeat until an integer solution is found.

Other methods for solving IP: Branch and Bound, Branch and Cut, Dynamic Programming.

3.11 Decision Analysis – Risk and Uncertainty

Decision Analysis provides a structured, quantitative approach for making decisions under conditions of risk (known probabilities) and uncertainty (unknown probabilities).

> **Definition**
> *"Decision analysis is a systematic, quantitative, and visual approach to addressing important choices in any area of business. It involves evaluating options under conditions of uncertainty to determine the course of action most likely to achieve the decision-maker's objectives."*

Risk vs Uncertainty

- **Risk:** The probabilities of different outcomes are known or can be estimated. Example: A 60% chance of high demand and 40% chance of low demand.
- **Uncertainty:** The probabilities are unknown. Example: Launching a product in a completely new market with no historical data.

Decision-Making Environments

- **Decision under Certainty:** The outcome of each decision is known with complete certainty. The decision-maker simply chooses the option with the best known outcome.
- **Decision under Risk:** Multiple possible outcomes exist, each with a known probability. Use Expected Value (EV) to evaluate decisions. EV = Σ(Probability × Payoff)
- **Decision under Uncertainty:** Probabilities are unknown. Various criteria are used:
- Maximax (Optimistic) criterion: Choose the option with the maximum possible payoff (best case).
- Maximin (Pessimistic) criterion: Choose the option with the maximum minimum payoff (best worst case).
- Minimax Regret criterion: Minimise the maximum regret (opportunity cost) of a decision.
- Laplace criterion (Equal Probability): Assume all outcomes are equally likely and choose the highest average payoff.

Decision Trees

A Decision Tree is a visual and analytical tool for structuring decisions under risk. It maps out decisions, chance events, and their outcomes in a tree structure.

- **Decision Node (Square □):** A point where the decision-maker makes a choice.
- **Chance Node (Circle ○):** A point where a random event occurs (with associated probabilities).
- **Terminal Node (Triangle):** The final outcome with its payoff value.
- **Expected Monetary Value (EMV):** Roll back from the terminal nodes, calculating the weighted average payoff at each chance node, then select the decision that maximises EMV.

3.12 Text Analytics and Web Analytics

Text Analytics (Text Mining)

> **Definition**
> *"Text analytics (or text mining) is the process of transforming unstructured text data into meaningful, structured information using natural language processing (NLP), machine learning, and statistical methods."*

Most data in the world is unstructured text – social media posts, customer reviews, emails, news articles, chat logs. Text analytics makes this data usable for business insights.

- **Sentiment Analysis:** Determining whether text expresses a positive, negative, or neutral opinion. Example: Analysing customer reviews to understand overall product satisfaction.
- **Topic Modelling:** Discovering the main themes or topics in a collection of documents. Example: Finding the most discussed topics in customer support tickets.
- **Named Entity Recognition (NER):** Identifying and classifying entities in text – people, organisations, locations, dates, monetary values.
- **Text Classification:** Categorising text into predefined categories. Example: Classifying news articles as Sports, Politics, or Finance.
- **Keyword Extraction:** Finding the most important words or phrases in a document.
- **Document Summarisation:** Automatically generating a brief summary of a long document.

Web Analytics

> **Definition**
> *"Web analytics is the measurement, collection, analysis, and reporting of website data to understand and optimise web usage."*

Web analytics helps businesses understand how users interact with their websites and digital properties, enabling data-driven improvements to user experience, content, and marketing.

- **Key Web Analytics Metrics:**
- Sessions / Visits: Total number of visits to the website in a period.
- Page Views: Total number of pages viewed. Repeated views of a single page are counted.
- Unique Visitors: Number of distinct individuals who visited the website.
- Bounce Rate: Percentage of visitors who leave after viewing only one page.
- Average Session Duration: Average time spent on the website per visit.
- Conversion Rate: Percentage of visitors who complete a desired action (purchase, sign-up).
- Traffic Sources: Where visitors came from (organic search, paid ads, social media, direct).
- **Google Analytics:** The world's most widely used web analytics platform. It tracks user behaviour, traffic sources, conversions, and much more for free.

- **A/B Testing in Web Analytics:** Testing two versions of a webpage (A and B) to determine which performs better on a key metric such as click-through rate or conversion rate.