Roll No: 67

NAME: HERAMB R. PAWAR

SUB: DAV

ASSIGNMENT

Q1 BRIEFLY Explain the phases of Data Analytics lifecycle?

→ The Data analytic lifecycle is Designed for Complex Data problems and Data Science projects. The cycle is iterative to Potray a real project. It enables, Work can return to earlier Phase as new Information is Uncovered. The phases are as follows:

Phase 1: Discovery

This phase include ~~gathering of~~:

a) gathering of Domain knowledge
b) gathering of Resources
c) gathering of Data
d) Planning & gathering People having proper skills, who can work on the given problem and can help in solution building
e) planning out Schedules and Deadlines
f) problem statement formation

Phase 2: Data preparation

This phase mainly includes dealing with data

a) ETL: Extract, Transform, Load:

It includes gathering data from various sources and transform it into a proper useable & acceptable format and loading it into a database or warehouse

b) Data Conditioning:
Making Data uniform through the process of normalization

c) Data cleaning:
Data cleaning includes removing of null values and irrelevant data from data available

d) Data Selection:
From the available data selecting some of the attributes on the basis of which data analysis would be performed

Phase 3: Model Planning:
In this phase we decide the various methods that will be required for building of Model such as clustering, association rule, regression, classification
It also includes selection of variables

Phase 4: Model building:
In this phase we build the model and execute it and setting up additional hardware and software resources if required

Phase 5: Communicating Results:
Communicating result weather the project was successful or not. Discussing the key findings and summarizing the narratives of stakeholders

Phase 6: Operationalize:

Preparing final reports, briefing, technical Documents are submitted or delivered

Q5 What are the types of text analysis?

→ There are various types of text analysis techniques

a) Sentiment analysis:

This technique involves analyzing text to determine the emotional tone of the text. It can be used to identify positive, negative or neutral sentiments

b) Text Categorization:

This involves categorizing text into predefined categories based on their content. For eg: news article can be categorized into topics such as Politics, Sports, Culture, entertainment, etc

c) Text clustering:

This technique involves grouping similar text together based on their content. Text clustering can be used to identify common themes or topics in large volume of text

d) Named Entity Recognition:

This technique involves identifying and classifying entities mentioned in text such as people, organization and locations

e) Topic Modelling:

This technique involves identifying topic's or themes that are present in a collection of Documents. Topic modelling can be used to understand the content of large volume of text, and identify common patterns and themes

Q7 What are 5 guidelines For better data visualization?

Ans)

1) Know your Audieance:
Before creating a data visualization it's important to understand who the audience is and what they are looking For. This can help to determine the type of visualization to use and level of detail to include.

2) Keep It Simple:
A good data visualization should be simple and easy to understand. Avoid cluttering the visualization with too much information or using complicated charts and graphs that can confuse the audience

3) Use appropriate Visualization:
Different types of data require different type of visualization For example bar charts are good For comparing values while line charts are better For showing trend over time. Choose the right type of visualization to effectively communicate infomation

4) Highlight key Insights:
A good data visualization should highlight the key insights or finding in the data. Use color, annotations and other visual cue to draw attention to important point in visualization.

5) Be honest:
A data visualization should be honest and accurately represent the data. Avoid using tricks such as distorting the scale or using inappropriate axis labels that can mislead the audience.

Q8) Explain ten guidelines for better table design?
→

1) Keep It Simple:
Tables should be easy to understand and use. Avoid cluttering the table with unnecessary information or formatting that can make it difficult to read.

2) Choose the right layout:
Tables can be organized in different ways such as rows and columns or a more complex structure. Choose the layout that best suits the data being presented.

3) Use clear headings:
Headings should be clear and descriptive, providing summary of the data in the table

4) Highlight key information:
Use formatting such as bold or color to highlight important information in table

5) Avoid vertical Scrolling:
Table should be fit on single screen whenever possible to avoid vertical scrolling which makes it difficult to compare data

6) Use consistent Formatting:
Consistent Formatting throughout table can make it easier to read and understand

7) Use white space effectively:
Use white space to separate different sections of table and make it easier to read

8) Keep data labels simple:
Data labels should be short and to the point avoiding long description or unnecessary details

9) Use visual cues:
Visual cues such as icons or arrows can help to draw attention to important information in the table.

10) Test the table:
Before finalizing the design, test the table with potential user to ensure that it is easy to understand

Q2 What is importance of regression model?
Ans Regression models are a type of statistical analysis used to model relationship between a dependent variable and one or more independent variables. The importance of regression models are as follows:

1) Predictive analysis:
Regression models are widely used in predictive analysis where they can be used to forecast values of the dependent variables based on the values of independent variables

2) Relationship Identification:
Regression model can help to identify relationship between the dependent and the independent variable. This helps to understand underlying factors that affects the dependent variable

3) It helps in improved decision making and problem solving

3) Control for Confounding Variables:
Regression model allows to control confounding variables, which can help to isolate the effect of independent variable on dependent variable. This is important because it can help to eliminate bias and increase accuracy of the analysis

4) Model Selection:
Regression models can help to identify the important variables and their relationship with the dependent variable. This can help to optimize the model and improve its predictive power.

5) Hypthesis testing:
Regression models can be used to test hypotheses about the relationship between the dependent variable and independent variable. this can help to identify the statistically significant relationship and provide insights to underlying mechanism that drive the relationship.

Q3 Explain logistic, simple and multiple linear regressions?
Ans → Regression is a statistical method used to model relationship between one or more independent variables and a dependent variable. The goal of regression is to find a mathematical equation that can predict the value of dependent variable based on value of independent variable. It is often used in situations where we want to understand how changes in one variable affect another. There are several types of regression they are as follows:

Simple linear regression:
In this type of regression there is only one independent variable and one dependent variable

Multiple linear Regression:
In this type of regression there are 2 or more Independent variable and one dependent variable

Both simple and multiple employ a regression line known as best fit line. The linear connection is defined as $y = c + mx + e$ where 'c' denotes intercept, 'm' denotes slope of line and 'e' is the error term

#Logistic Regression:
When the dependent variable is discrete, the logistic regression technique is applicable. In other word this technique is used to compute the probability of mutually exclusive occurences such as pass/fail, true/false, 0/1 and so forth. Thus the target variable can take only one of 2 values and a Sigmoid curve represents its connection to independent variable and the probability has a value between 0 and 1

(B4) a) What is time Series?
b) Explain ARIMA MODEL?
c) Explain Box Jenkins Methadology?

Ans A) Time Series Analysis is a statistical technique used to analyze data overtime. In simpler terms it involves analyzing data that is collected at regular intervals over a period of time to identify patterns, trends and relationship

eg: time Series analysis could be used to analyze monthly sales data of a company over the

past few years to identify seasonal trends or forecast future sales. Time series analysis is a powerful tool for understanding how data changes over time and how it can be used to make informed decisions based on past trends and future predictions. It includes components (like trend, season, cycle, irregularity & level.

Ans B) ARIMA Model:

#) ARIMA stands for Autoregressive Integrated Moving Average. It is a time series model that combines the Autoregressive (AR), Integrated (I) and Moving average (MA) model. It is a more powerful version of ARMA model which can be used to model time series data that exhibit trend and seasonality.

#) The notation used to describe ARIMA model is ARIMA (P, d, Q) where "p" is order of auto-regressive component, "d" is the degree of differencing and "q" is the order of moving average component

#) The AR component models the influence of past value on current value, the "I" component models the trend and "MA" component models the influence of past error terms on current value

The formula for ARIMA (P, d, Q) model is:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Ans e)

The Box Jenkin's methadology is widely used approach for time Series analysis and forecasting. It involves a Series of Step that are followed to identify, model and forecast a time series.

Here are steps involved in box Jenkin's methadology:

1) Identify the time Series:
The First Step is to identify the time Series data that needs to be analyzed. This involves collecting historical data on variable of interest such as sales or stock price over a period of time.

2) Visualize the data:
The next step is to visualize the data using plot such as time series plots, Scatter plots or histogram This helps to identify any patterns, trends or seasonality in data.

3) Check for Stationarity:
Stationarity is key assumption in time Series which means that the statistical properties of data remain Constant over time. The next step is to check weather time Series is stationary or not. These can be done by checking for presence of trends, seasonality or periodicity in the data

4) Make data Stationary:
If the data is not stationary the next step is to make it stationary using techniques such as differencing, logarithmic transformation or detrending

5) Identify order of differencing:
If differencing is required the next step is to identify appropriate order of differencing. This can be done by plotting autocorrelation function (ACF) and partial auto correlation function (PACF) of the differenced data and selecting lag order that minimizes the AIC or BIC criterion.

6) Identify the ARMA model:
Once the data is stationary the next step is to identify appropriate autoregressive (AR) and moving average (MA) orders for ARMA model. This can be done by analyzing ACF and PACF plots of differenced data and selecting the orders that best fit the data.

7) Estimate model parameters:
Once the order are selected the next step is to estimate the model parameters using maximum likelihood estimation or another suitable method.

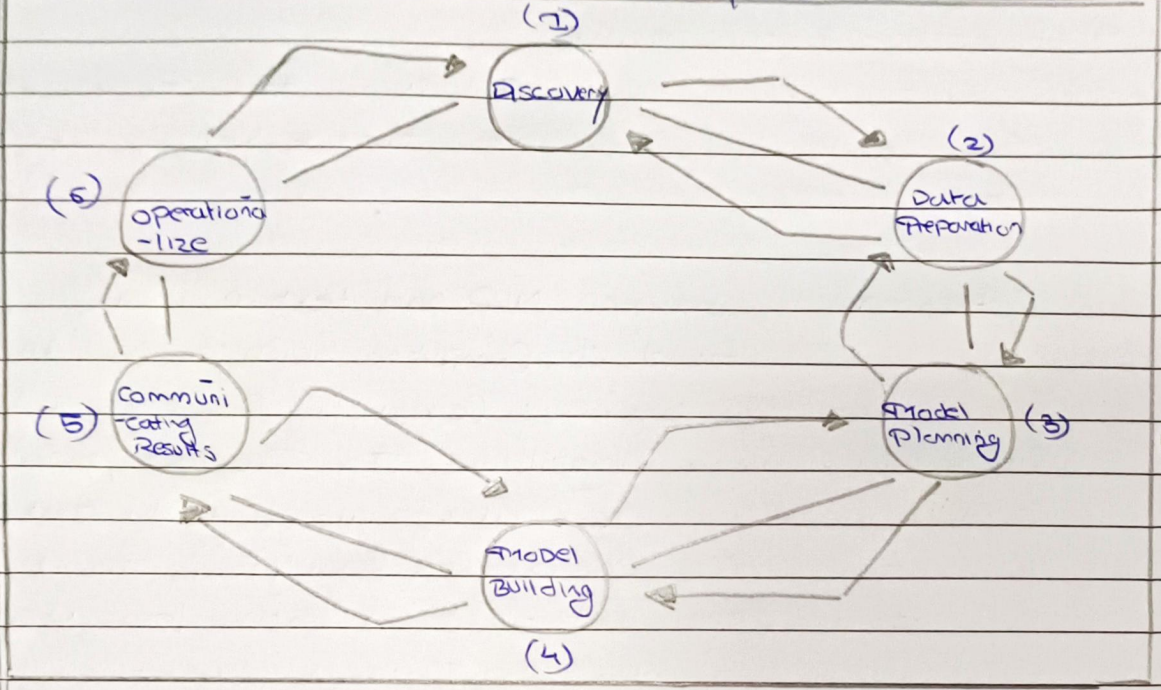
8) Check the model fit:
The next step is to check the fit of model by comparing model predictions to the actual data. This can be done using diagnostic plots such as residual plots, normal probability plots or QQ plots.

9) forecast future value:
Once the model is validated, the final step is to use it to forecast future value of the ~~serie~~ series.

Q6 Differentiate between NLP and text analysis?
Ans NLP is a subfield of Computer Science and AI concerned with interaction between computers and human language. It involves using algorithms to process, analyze and understand natural language text data. NLP aims to create computer programs that can understand and interpret human language in same way human do. This involve task such as speech recognition, language translation and sentiment analysis

Text analysis on the other hand is a broader field that involves using statistical and ML techniques to analyze and extract insights from unstructured text data. It includes various methods of processing text, such as pre-processing feature extraction and text mining. Text analysis can be used for a variety of application including sentiment analysis, topic modelling and text classification

Q1 PHASES OF DATA ANALYTICS LIFE CYCLE



Q4 (c) THE BOX JENKINS APPROACH: