ASSESSMENT–2

Submittedby

NAME.   :K.SINDHUJA

PIN NUMBE.   :20T91A0533

COLLEGE.   :GIET ENGINEERING

                    COLLEGE

E -MAIL.   :kadalisindhuja6672@gmail.com

SectionA:DataWrangling(Questions)

1. Whatistheprimaryobjectiveofdatawrangling?
   - a)Datavisualization
   - b)Datacleaningandtransformation
   - c)Statisticalanalysis
   - d)Machinelearningmodeling

   ANSWER:b)Datacleaningandtransformation

   The primary objective of data wrangling is to clean and transform raw data into a usable format for analysis. This involves tasks such as removing errors, handling missing values, restructuring data, and preparing it for further analysis.

2. **Explainthe techniqueused toconvertcategoricaldata intonumericaldata.Howdoesithelpin data analysis?**

   Explanation:
   - Onecommontechniqueusedtoconvertcategoricaldataintonumericaldataiscalled **"one-hotencoding."**
   - Inone-hotencoding,eachcategoryis representedbyabinaryvectorwhereonlyonebit is 1 (hot) and the rest are 0 (cold). Each bit corresponds to a category, and the position of the hot bit indicates the category's presence or absence.

   Forexample,ifyouhaveacategoricalvariable"color"withthreecategories:red,green,andblue, one-hot encoding would represent them as follows:

   Red:[1,0,0]
   Green:[0,1,0]
   Blue:[0,0,1]
   - One-hot encoding helps in data analysis by allowing categorical data to be used in machine learning algorithms and statistical analyses that require numerical input.
   - It prevents the algorithm from assuming a natural ordering or hierarchy among the categories, which might not exist in reality. Additionally, it avoids biasing the analysis towardsanyparticularcategorybyrepresentingeachcategoryequallyandindependently.

3. **HowdoesLabelEncodingdifferfromOneHotEncoding?**

Explanation:

Label Encoding and One-Hot Encoding are two techniques used to convert categorical data into numerical data, but they differ in their approach and application:

1. Label Encoding:

- InLabelEncoding,eachcategoryisassignedauniqueintegerlabel.
- This technique is suitable for **ordinal categorical variables**, where there is a clear ranking or order among the categories.

For example, if you have categories like "low," "medium," and "high," Label Encoding might assign them integer labels like 0, 1, and 2, respectively.

- The problem with Label Encoding arises when the categorical variable doesn't have an inherent order, as assigning numerical labels might introduce unintended relationships or biases in the data.

2. One-Hot Encoding:

- In One-Hot Encoding, each category is represented as a binary vector where only one bit is 1 (hot) and the rest are 0 (cold).
- This technique is suitable for **nominalcategoricalvariables**,where there is no inherent order among the categories.
- One-Hot Encoding creates new binary columns for each category, where a 1 indicates the presence of that category and 0 indicates its absence.

For example, if you have categories like "red," "green," and "blue," One-Hot Encoding would create three binary columns, one for each category, with a 1 indicating the presence of that color and 0s for the others.

- One-Hot Encoding avoids the issues of assuming ordinal relationships among categories and ensures that each category is treated independently in analysis.

4. **Describe a commonly used method for detecting outliers in a dataset. Why is it important to identify outliers?**

WhatAreOutliers?

- Outliersaredatapointsthatsignificantly deviatefromtherestofthedataset.Theycan be either much larger or significantly smaller than other values.
- Outliersmayresultfromnaturalvariation,measurementerrors,orotheranomalies.

Themostcommonlyusedmethodfordetectingoutliersinadatasetis **IQR(IQRstandsfor Interquartile Range.)**

- Sortdata,calculateQ1,Q3,andIQR.
- Identifydatapointsbeyondthefences.

- Removeorhandleoutliersappropriately.

WORKING:InterquartileRange(IQR)Method:

- Calculatethefirstquartile(Q1)andthirdquartile(Q3).
- Determinetheinterquartile range(IQR=Q3-Q1).
- Defineanupperfence(Q3+1.5*IQR)andalowerfence(Q1-1.5*IQR).
- Anydatapointbeyondthesefencesisconsideredanoutlier.

Forinstance,considermeasuringrunningtimesforcollegestudents.Trueoutliers(naturalvariations) should be retained, while other outliers (measurement errors) should be addressed.

5. **Explain how outliers are handled using the Quantile Method. EXPLANATION:**
   TheQuantileMethodforhandlingoutliersinvolvesseveralsteps:

   1. CalculateQuantiles:
   - Dividethedatasetintoquantiles,suchasquartiles(Q1,Q2,Q3)orpercentiles(e.g., 25th, 50th, 75th).
   2. IdentifyOutliers:
   - Determinethelowerandupperboundsforoutliersusingthequantiles.
   - Commonly,outliersfallbelowQ1-k*IQRoraboveQ3+k*IQR,whereIQRisthe interquartile range (Q3 - Q1), and k is a constant multiplier (e.g., 1.5 or 3).
   3. HandleOutliers:
   - Outlierscanbetreatedinvariousways:
     - Removal:Excludeoutliersfromthedataset.
     - Transformation:Applymathematicaltransformationstotheoutliers,suchaswinsorization or log transformation.
     - Capping:Capoutliersatacertainthresholdtoreducetheirimpactontheanalysis.
   4. ReevaluateAnalysis:
   - After handling outliers, reassess the analysis to ensure that the results remain valid and meaningful.

6. **DiscussthesignificanceofaBoxPlotindataanalysis.Howdoesitaidinidentifying potential outliers?**

Explanation:
A **box plot**, also known as a **box-and-whisker plot**, is a type of chart used in **explanatory dataanalysis**.
- It visually represents the distributionofnumerical data and providesinsights into the central tendency, spread, and skewness of the dataset.

1. **ComponentsofaBoxPlot**:

    - **Minimum Score**: The lowest data point (excluding outliers) represented at the end of the left whisker.
    - **LowerQuartile(Q1)**:Thevaluebelowwhich25%ofthedatafalls.
    - **Median(Q2)**:Themidpointofthedata,dividingtheboxintotwoequal parts.
    - **UpperQuartile(Q3)**:Thevaluebelowwhich75%ofthedatafalls.
    - **MaximumScore**:Thehighestdatapoint(excludingoutliers)representedatthe end of the right whisker.
    - **Whiskers**:Representscoresoutsidethemiddle50%(betweenQ1andQ3).
    - **InterquartileRange(IQR)**:The rangebetweenQ1andQ3.

2. **WhyareBoxPlotsUseful?**

    - **VisualSummary**:Boxplotsdividedataintosectionscontainingapproximately 25% of the data. They provide a quick overview of:

        - Meanvalues.
        - Dispersion(spread)ofthedataset.
        - Signsofskewness.

        OutlierDetection:Outliersareoftenvisibleasdatapointsbeyondthewhiskers.

    - **SkewnessIdentification**:
        - Symmetricdistribution:Medianinthemiddleofthebox,whiskers balanced.
        - Positively skewed (skewed right): Median closer to the bottom of the box, shorter whisker on the lower end.
        - Negativelyskewed(skewedleft):Medianclosertothetopofthebox, shorter whisker on the upper end.

Remember, box plots provide valuable insights beyond just central tendency measures (mean, median, mode.

SectionB:RegressionAnalysis

7. Whattypeofregressionisemployedwhenpredictingacontinuoustargetvariable?
Explanation:
Whenpredictinga **continuoustargetvariable**,themostcommonlyusedregressionmethod is**linear regression**. Linear regression, also known as **ordinary least squares (OLS)**, is a fundamental technique in statistics and data science. Let's explore its key aspects:
LinearRegression(OLS):

- **Purpose**: Linear regression models the relationship between a **continuous dependentvariable** (target)andoneormore**independent variables** (predictors).
- **Assumption**: It assumes a**linear relationship** between the predictors and the target.
- **Objective**: The goal is to find the best-fitting straight line (or hyperplane in multiple dimensions) that minimizes the sum of squared differences betweenthe observed data points and the predicted values.
- **Parameters**: Linear regression estimates coefficients (slopes) for each predictor variable.

8. Identifyandexplainthetwomaintypesofregression.
Explanation:
Thetwomaintypesofregressionare:
       (a). LinearRegression
       (b). LogisticRegression

(a). LinearRegression:
- Linearregressionmodelsthe relationshipbetweenone ormore independentvariables (features) and a continuous dependent variable (target) by fitting a linear equation to the observed data.
- The relationship is assumed to be linear, meaning the change in the target variable is proportional to changes in the independent variables.
- Linear regression is used when the relationship between the variables can be approximated by a straight line and is widely applied in various fields for prediction and inference tasks.

(b). LogisticRegression:
- Logistic regression models the probability that a binary outcome occurs based on one or more independent variables.
- Unlikelinearregression,whichpredictscontinuousoutcomes,logisticregression predicts the probability of a categorical outcome **(e.g., success/failure, yes/no, 0/1).**
- Thelogisticfunction(sigmoidfunction)isusedtomapthelinearcombinationof independent variables to a probability between 0 and 1.

9. WhenwouldyouuseSimpleLinearRegression?Provideanexamplescenario.
Explanation:
- Simple linear regression is typically used when there is **a linear relationship** between a single independent variable and a continuous dependent variable.
- It's suitable for scenarios where you want to understand or predict the behavior of the dependent variable based on changes in one predictor variable.

ExampleScenario:
Let's consider a scenario in which a real estate agent wants to predict house prices based on theirsize(in square feet).Theagent believes thatthereis a linear relationshipbetweenthesize of a house and its price.

Inthisscenario:

DependentVariable(Y): HousePrice

IndependentVariable(X): HouseSize(insquarefeet)

- The agent collects data on various houses, recording their sizes and corresponding prices.Byapplying simple linearregressiontothisdata,theagentcanbuildamodelthat predicts the price of a house based on its size.

10. InMultiLinearRegression,howmanyindependentvariablesaretypicallyinvolved?

Explanation:

- In**MultipleLinearRegression**,**twoormoreindependentvariables** aretypicallyinvolved.
- Thistechniqueallowsustoestimatetherelationshipbetweentheseindependentvariables and a **single dependent variable**.
- Forinstance,youmightusemultiplelinearregressiontoanalyzehow **rainfall,temperature, and fertilizer usage** affect **crop growth**.

11. When should Polynomial Regression be utilized? Providea scenario where Polynomial Regression would be preferable over Simple Linear Regression.

Explanation:

Polynomialregressionisapowerfultechniqueusedinsituationswheretherelationshipbetween the predictor variable(s) and the response variable is **nonlinear**.

Unlike**simplelinearregression**,whichmodelstherelationshipasastraightline,polynomial regression allows for more flexibility by fitting a **polynomial equation** to the data.

Herearescenarioswherepolynomialregressionwouldbepreferableoversimplelinear regression:

1. **CurvedRelationships**:

   - Whenthescatterplotofthepredictorvariableand theresponsevariableexhibitsa **curvedpattern**,polynomialregressionisabetterchoice.
   - For instance, consider predicting a student's exam score basedon the number of hours studied.
   - Ifthescatterplotshowsanonlinearrelationship(likeaU-shape),polynomialregression can capture this curvature more effectively.

2. **FittingNonlinearPatterns**:

- Suppose you have data that follows a quadratic, cubic, or higher-degree nonlinear relationship.
- Insuchcases,simple linearregressionwouldunderperform. Polynomial regressioncan handle these situations by allowing for more complex curves in the model.

3. **AdjustedR-SquaredComparison**:

- Calculate the **adjusted R-squared** for both linear and polynomial regressionmodels.
- The adjusted R-squared represents the proportion of variance in the response variable explained by the predictor variables, adjusted for the number of predictors. The model with the **higher adjusted R-squared** provides a better fit.
- If the adjusted R-squared is significantly better for the polynomial model, itindicates that polynomial regression is more appropriate.

12. **What does a higher degree polynomial represent in Polynomial Regression? How does it affect the model's complexity?**

Explanation:

In polynomial regression, the degree of the polynomial represents the highest power of the independent variable (predictor) in the regression equation. For example, in a polynomial regression with a degree of 2, the equation might look like this:

$$y=\beta_0+\beta_1x_1+\beta_2x_1{}^2+\beta_2x_1{}^3+. \ldots\ldots\beta_nx_1{}^n$$

Here X^2 represents the squared term of the independent variable (X). Similarly, in a polynomial regression with a degree of 3, the equation might include terms up to (X^3), and so on.

- However,increasingthedegreeofthepolynomialalsoincreasesthemodel'scomplexity.
- This increasedcomplexity canleadto overfitting, wherethemodelcaptures noiseinthe training data rather than the underlying true relationship between the variables.
- Asaresult,themodelmaynotgeneralizewelltounseendata,leadingtopoor performance on new data.

13. **HighlightthekeydifferencebetweenMultiLinearRegressionandPolynomialRegression.**

Explanation:

The key difference between multiple linear regression and polynomial regression lies in the nature of the relationship they model between **the independent variable(s)** and the **dependent variable**:

Multiple Linear Regression: In multiple linear regression, the relationship between the dependent variable and the independent variables is assumed to **be linear**.

- ➤ The model includes **multiple independent variables**, but each of these variables is raised to the power of 1, and the model equation is a linear combination of these variables.
- ➤ The equation takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Here, $X_1, X_2, \ldots, X_n$ represent the independent variables, and

$\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients associated with each independent variable.

Polynomial Regression: In polynomial regression, the relationship between the dependent variable and the independent variable is **not constrained to be linear.** Instead, the model allows for higher-order terms of the independent variable to capture nonlinear relationships.
The model equation takes the form:
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_d X^d + \varepsilon$$
Here, X represents the independent variable, and $X^2, X^3, \ldots, X^d$ represent the higher-order polynomial terms up to degree d. $\beta_0, \beta_1, \beta_2, \ldots, \beta_d$ are the coefficients associated with each term.

14. Explain the scenario in which Multi Linear Regression is the most appropriate regression technique.

Explanation:

Multiple linear regression is most appropriate when you have a single dependent variable and two or more independent variables (the variables used to predict the dependent variable), and the relationship between the dependent variable and the independent variables is assumed to **be linear.**

scenario:

Suppose you are working for a real estate agency, and your task is to predict the selling price of houses based on various factors. You have data on houses that were sold in a particular area,

and for each house, you have information such as the size of the house (in square feet), the number of bedrooms, the number of bathrooms, the age of the house, and the neighborhood.

1. **Dependentvariable:**Thesellingpriceofthehouses.

2. **Independentvariables:**Thesizeofthehouse,thenumberofbedrooms,thenumberof bathrooms, the age of the house, and potentially other factors like the neighborhood.

Youcan use multiple linearregressionto builda modelthat predicts the selling price ofa house based on these independent variables. The assumption is that there is a linear relationship between each independent variable and the selling price, holding other variables constant.

15. Whatistheprimarygoalofregressionanalysis?

<u>Explanation:</u>

The primary goal ofregressionanalysis is tounderstandand **quantify therelationship between one or more independent variables (predictors) and a dependent variable (outcome).**

Theprimarygoalsofregressionanalysisare:

1. **Estimate Parameters**: Determine the coefficients or parameters that best describe the relationship between the independent variables and the dependent variable.
2. **PredictOutcome**:Use the estimated regressionequation tomake predictions about the dependent variable for new or unseen data based on the values of the independent variables.
3. **Inferential Purposes**: Draw statistical inferences about the relationship between variables, including testing hypotheses about the significance of individual predictors and overall model fit.
4. **Model Interpretation**: Provide insights into the strength, direction, and nature of the relationship between variables, helping to understand the underlying mechanisms driving the outcome variable.