Assignment 6

Machine learning

Ans1 (C)

Ans 2(B)

Ans 3(C)

Ans 4(A)

Ans 5(B)

Ans 6(A,D)

Ans 7(B,C)

Ans8(A,C)

Ans 9(A,B)

Ans 10 It is a metric to evaluate how well is our model. As we increase the number of independent variables in the equation the r2 increasse as well but that doesn't mean that the new independent variables have any correlation with output variable.In other words ,even witnh the addition with the new features in our model,it is not necessary that our model will yield better results but r2 value will increase .To rectify this problem we use adjusted r2 value which penalizes excessive use of such featureswhich donot correlate

with the output data .Adjusted r2 is a modified version of R- squared that has been adjusted for the number of predictors in the model .The adjusted R-squared increases when the new term improves bthe model more than would be expected by chance .It decreases when the predictor improves the model by less than expected.

Selecting the model with the highest value of R-squared is not a correct approach as the value of R-squared shall always increase whenever a new feature is taken for consideration even if the feature is unrelated to the response.

The alternative is to use adjusted R- squared which penalises the model complexity .

Ans 11)Lasso regression penalises the model based on the sum of magnitude of the coefficients .It will neglect the features which have no relationship with label .It ignores and acts as feature selection .Zero importance to the features having no relationship with the label.Lasso regression takes the magnitude of the coefficients .It is the modification of linear regression,where the model is penalized for the sum of

absolute values of the weights .Thus the absolute values of weight will be reduced .

Ridge regression:

It gives little importance to the negligible features which have no relationship with label .Very little importance to the features having no relationship with the label.It penalizes the model based on the sum of squares or magnitude of the coefficients.It shrinks the coefficients forb those predictors which contribute the very less in the model but have huge weights .very close to zero but never makes them exactly zero.

Ans 12)VIF or Variance Inflation Factor is the measure of the amount of multicollinearity in regression analysis.Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.This can adversely affect the regression results .VIF is used to detect the severity of multicollinearity in regression analysis.VIF score of an independent variable represents how well the variable is explained by other variables.A VIF between 5 to 10 indicates high correlation that may be problematic .

A value of 3 or below is a suitable value of a VIF in a regression modelling .As the VIF increases , the less reliable the regression results.

Ans 13)Scaling the data makes it easy for the model to learn and understand the problem. It is the preprocessing step . It is the process of normalizing the dataset.The scaling is used for making the datapoints generalised so that the distance between them will be lower .Larger differences between the datapoints increases the uncertainity in the results in the model.

Ans 14)Different metrics to check the goodness of fit in linear regression are:

MSE(Mean squared error):common metric for regression .It is te average of the squared difference between the predicted and the actual value .

MAE(Mean absolute error):

It is the average of the absolute difference between the target value and the value predicted by the model.

R-SQUARED:Measures the relationship between model and the dependent variable .

ROOT MEAN SQUARED ERROR(RMSE):square root of mean squared error.Most of the times people use RMSE as an evaluation metrics when working with deep learning techniques.

Ans 15)sensitivity=TP/TP+FN=1000/1050=0.95

SPECIFICITY=TN/TN+FP=1200/1450=0.82

PRECISION=TP/TP+FP=1000/1250=0.80

ACCURACY=TP+TN/TP+TN+FN+FP=1000+1200/2500=

0.88

RECALL=TP/(TP+FN)=0.95

WORKSHEET 6 SQL

Ans 1) (A,C,D)

Ans 2)(A,C,D)

Ans 3)(C)

Ans4)(C)

Ans 5)(B)

Ans 6)(B)

Ans 7)(A)

Ans 8)(C)

Ans 9)(A)

Ans 10)(A)

Ans 11)Denormalization :It is a process used on normalized database to increase the performance.It is a technique used by database administrators to optimize the efficiency of the database by adding data to one or more tables .Denormalization is used to combine multiple table data into one so that query can be performed quickly.The primary goal of Denormalization is to achieve faster execution of the queries .

Ans 12)Database cursor : say connector.It is used to connect to table cursor acts as a pointer .In database we call it a cursor .Without declaring the cursor we cannot do anything .A cursor allows row by row processing of the result sets.When we use the cursor , we can iterate or step through the results of a query and perform certain operations on each row.

Ans 13)Different types of the queries:

SELECT:extract data from database.

UPDATE:updates data in a database.

DELETE:deletes data from a database.

INSERT INTO: inserts new data into a database .

CREATE DATABASE: creates a new database.

ALTER DATABASE: modifies a database.

CREATE TABLE: creates a new table.

ORDER BY: change the order of records (increasing or decreasing)

UPDATE TABLE:update the column value .

Ans 14)Constraints specifies the rules for data in a table .It limits the type of data that can go into the table.This ensures the accuracy and reliability of data in a table.If there is any violation between constraint and data action, the action is aborted.Constraints are the rules that we can apply on the type of data in a table.ie ,we can limit on the type of data that can be stored in a particular column.

Not Null,unique ,Primary key ,Foreign key ,Default are some of the sql constraints.

Ans 15)The auto increment in sql is a feature that is applied to a field so that it can automatically generate and provide a unique value to each and every record that we enter in a table.It is a function that operates on numeric datatype .It automatically generates sequential numeric values every time when a record is inserted in a table.

STATISTICS WORKSHEET 6

Ans 1)(D)

Ans 2(C)

Ans 3)(A)

Ans 4)(C)

Ans 5)(A)

Ans 6)(D)

Ans 7)(C)

Ans 8)(B)

Ans 9)(B)

Ans 10)Both boxplot and histogram help to visualize numeric data.

Boxplot allows to compare multiple datasets better than histogram because they are less detailed and takes less space.

Boxplot provide some information tha histogram does not ie boxplot provides 25^{th},75^{th} percentile min/max and seperates the points that are considered as outliers.

Ans 11)Metrics describe the exact numbers that make up the data or the raw ingredients that make up the analytics possible .Metrics are measures of quantitative assessment used for comparing and tracking performance.Metrics are used to monitor the performance of a model during training and testing .

Choosing the good metrics are very important for the growth of company .

Good metrics can be improved, measure progress

Good metrics inspire action .

How can you choose the metrics matters the most when you want to evaluate the companys performance.

Prioritize objectives ,examine which metric is responsible for the achievement and which activities influence predictors .

To select the metrics ,Good metrics have following characteristics:

1) good metrics are important for the growth of the company.The key point is to choose metrics that clearly indicates where you are now in relation to your goals.
2) Good metrics can be improved that means there should be space for improvement.
3) Good metrics inspire action.When your metrics are important and can be improved.

   There are two main types of metrics leading indicators (measures the activities necessary to achieve goals),lagging indicators (measure the actual reports).

Ans 12)To access the statistical significance we would use hypothesis testing .Stating the null hypothesis

which is usually opposite of what we wish to test .The null hypothesis and alternate hypothesis is first.second we will calculate the p-value which is likelihood of getting test observation if the null hypothesis is true . At last we will select the threshold ie alpha value and reject the null hypothesis if the p-value is smaller than alpha .Also we will choose the critical region for the statistics to lie in ,which is enough for the null hypothesis to be rejected.

Common tests:

- One sample Z test
- Two-sample Z test
- One-sample t-test
- Paired t-test
- Chi-squared test
- Anova

Ans 13) any type of categorical data won’t have a Gaussian distribution as well as log normal amount of time mobile battery lasts or amount of time earthquack occurs.

Ans 14) Mean is the strategy used for normal distribution and median for skewed distribution .Amount spend on healthcare each year by individuals .

Real Estate calculate the median price of houses to get the idea of the home price.

Ans 15)It is the probability that an outcome is observed .The term likelihood refers to the process of determining best data distribution for a specific situation .It is proportional to the probability.