ASSIGNMENT NO: - 6

STATISTICS WORKSHEET-6

Question no	Answer
1	b) The outcome of flip of a coin
2	a) Discrete
3	a) pdf (probability density function)
4	c) mean
5	a) variance
6	a) variance
7	c) 0 and 1
8	b) bootstrap
9	b) summarized.

Q10 Answer: -

Sr.No.	Boxplot	Histogram
1	Box plots, on the other hand, are more useful when comparing several data sets	To determine the underlying probability distribution of a data, histograms are preferred
2	Box plots, on the other hand, attempt to do the same thing but fail to provide a good picture of the variable’s distribution.	Histograms give you a good idea of how a variable is distributed
3	The median, interquartile range, maximum, and minimum values of the data are displayed in a box plot, which is used to compare multiple groups of data	A histogram divides the data into ranges, which then plots the frequency with which data occurs in each range.
4	A boxplot can tell you whether a data set is symmetric (where the median is in the centre), but it cannot tell you how to shape the symmetry like a histogram.	Although histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed.
5	Box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.	It takes large space than box plot to display the results.

6	Box plot may be little clearer in terms of the center and outliers in the distribution.	Histogram may be less clear in terms of the center and outliers in the distribution.
7	In practice, a sample size of at least 30 data values would be sufficient for both tools.	In practice, a sample size of at least 30 data values would be sufficient for both tools.

Q11 Answer: -

Metrics are measures of quantitative assessment commonly used for assessing, comparing, and tracking performance or production. Generally, a group of metrics will typically be used to build a dashboard that management or analysts review on a regular basis to maintain performance assessments, opinions, and business strategies. Executives use them to analyze corporate finance and operational strategies. Analysts use them to form opinions and investment recommendations. Portfolio managers use metrics to guide their investing portfolios. Furthermore, project managers also find them essential in leading and managing strategic projects of all kinds. Overall, metrics refer to a wide variety of data points generated from a multitude of methods.

Every business executive, analyst, portfolio manager, and the project manager has a range of data sources available to them for building and structuring their own metric analysis. This can potentially make it difficult to choose the best metrics needed for important assessments and evaluations. Generally, managers seek to build a dashboard of what has come to be known as key performance indicators (KPIs). In order to establish a useful metric, a manager must first assess its goals. From there, it is important to find the best outputs that measure the activities related to these goals. A final step is also setting goals and targets for KPI metrics that are integrated with business decisions. Academics and corporate researchers have defined many industry metrics and methods that can help shape the building of KPIs and other metric dashboards.

Several businesses have also popularized certain methods that have become industry standards in many sectors. DuPont began using metrics to better their own business and, in the process, came up with the popular DuPont analysis which closely isolates variables involved in the return on equity (ROE) metric. GE has also commissioned a set of metrics known as Six Sigma that are commonly used today, with metrics tracked in six key areas: critical to quality; defects; process capability; variation; stable operations; and, design for Six Sigma.

Q12 Answer: -

In quantitative research, data are analyzed through null hypothesis significance testing, or hypothesis testing. This is a formal procedure for assessing whether a relationship between variables or a difference between groups is statistically significant.

- Null and alternative hypotheses

To begin, research predictions are rephrased into two main hypotheses: the null and alternative hypothesis.

A null hypothesis (H0) always predicts no true effect, no relationship between variables, or no difference between groups.

- An alternative hypothesis (Ha or H1) states your main prediction of a true effect, a relationship between variables, or a difference between groups.

Hypothesis testing always starts with the assumption that the null hypothesis is true. Using this procedure, you can assess the likelihood (probability) of obtaining your results under this assumption. Based on the outcome of the test, you can reject or retain the null hypothesis. In quantitative research, data are analyzed through null hypothesis significance testing, or hypothesis testing. This is a formal procedure for assessing whether a relationship between variables or a difference between groups is statistically significant.

The significance level can be lowered for a more conservative test. That means an effect must be larger to be considered statistically significant. The significance level may also be set higher for significance testing in non-academic marketing or business contexts. This makes the study less rigorous and increases the probability of finding a statistically significant result. As best practice, you should set a significance level before you begin your study. Otherwise, you can easily manipulate your results to match your research predictions.

Q13 Answer: -

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well. Let us understand with few examples.

- Time that an Interviewer spends with a candidate

If you are applying for a vacancy and are asked to wait for the interview, you can simply use the exponential distribution to roughly estimate the timing of your interview and predict the time that for how long would it go. For this purpose, the only requirement is that the average time that the interviewer takes to finish the interview of previous candidates is well known.

- Average Time a Call Centre Employee Spends with the Customer

If the average time that a call centre executive takes to complete his/her given task to communicate with a customer is known to be twenty minutes, then the probability that the executive will handle eight customers per hour can be estimated with the help of exponential distribution. This helps the managers draft an appropriate schedule to tackle all the customers approaching the firm with their concerns. This helps improve the customer satisfaction index.

- Purchasing Flight Tickets

Most travellers tend to buy their flight or train tickets a few days prior to their actual journey to avoid any last-minute hustle. Assume that maximum customers tend to buy their tickets fifteen days before they actually execute a journey, then the probability that a person will book his/her ticket ten days prior to his actual date of commencing the journey can be calculated with the help of exponential distribution. This helps the flight managers maintain an appropriate customer to occupancy ratio in advance. It also helps the transport marketing managers draft the appropriate deals and offers to attract potential customers and enhance sales.

- Life Span of Electronic Gadgets

Exponential distribution finds its prime application in calculating the reliability of electronic gadgets such as a laptop, battery, processor, mobile phone, etc. It helps the engineers and manufacturers to know an approximate time after which the product will get ruptured. The engineers use this data to improve the quality of their products by replacing the low-quality components with those having comparatively high quality.

Q14 Answer: -

Let us say you run a customer satisfaction survey with a sample of 9 and rate their overall satisfaction scores on a scale of 1 to 10. You get an average of 5.22. You know that in general, you tend to retain customers with a score over 3, so you are satisfied, because this indicates that you are still above where you want to be. But then, suddenly, you lose 6 of those 9 customers. You go back to look at your data, and you find these scores: 1, 3, 3, 3, 3, 5, 9, 10, 10

The median of this group is a 3, indicating that at least half of your customers or more were unhappy. The scores became lopsided because of the unexpected 10's, and you missed out on an important part of your data – the midpoint that indicated that as many as half of your customers or more were dissatisfied with your company.

Median can play a major role in things like income level research as well, because a few millionaires may make it look like the socio-economic status of your sample is higher than it really is.

Whenever a graph falls on a normal distribution, using the mean is a good choice. But if your data has extreme scores (such as the difference between a millionaire and someone making 30,000 a year), you will need to look at median, because you will find a much more representative number for your sample.

Q15 Answer: -

Likelihood refers to how well a sample provides support for values of a parameter in a model.
Example: -
Suppose we have a spinner split into thirds with three colors on it: red, green, and blue. Suppose we assume that it is equally likely for the spinner to land on any of the three colors. If we spin it one time, the probability that it lands on red is 1/3. Now suppose we spin it 100 times and it lands on red 2 times, green 90 times, and blue 8 times. We would say that the likelihood that the spinner is equally likely to land on each color is very low.
When calculating the probability of the spinner landing on red, we simply assume that P(red) = 1/3 on a given spin. However, when calculating the likelihood, we are trying to determine if the model parameters (P(red) = 1/3, P(green) = 1/3, P(blue) = 1/3) are correctly specified.

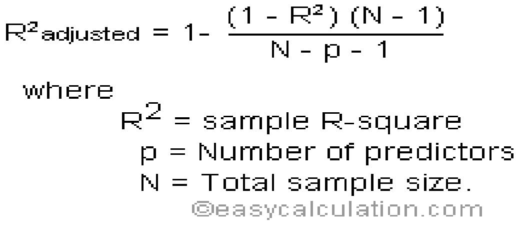
ASSIGNMENT NO: - 6

MACHINE LEARNING

Question No	Answers
1	C) High R-squared value for train-set and Low R-squared value for test-set.
2	B) Decision trees are highly prone to overfitting.
3	C) Random Forest
4	B) Sensitivity
5	B) Model B
6	A) Ridge & D) Lasso
7	B) Decision Tree & C) Random Forest
8	A) Pruning & B) L2 regularization
9	A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

Q 10 Answer: -

The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.
The adjusted R2 will penalize you for adding independent variables (K in the equation) that do not fit the model. Let us understand Why? In regression analysis, it can be tempting to add more variables to the data as you think of them. Some of those variables will be significant, but you cannot be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables.
While values are usually positive, they can be negative as well. This could happen if your R2 is zero; After the adjustment, the value can dip below zero. This usually indicates that your model is a poor fit for your data. Other problems with your model can also cause sub-zero values, such as not putting a constant term in your model.
Conclusion: - Use the adjusted R-square to compare models with different numbers of predictors



$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

©easycalculation.com

Q11 Answer: -

Sr.No.	Lasso Regression	Ridge Regression
1	Lasso model lowers the size of the coefficients and leads to some features having a coefficient of 0, essentially dropping it from the model	The Ridge Regression also aims to lower the sizes of the coefficients to avoid over-fitting, but it does not drop any of the coefficients to zero.
2	The LASSO, however, does not do well when you have a low number of features because it may drop some of them to keep to its constraint, but that feature may have a decent effect on the prediction	The Ridge Regression improves the efficiency, but the model is less interpretable due to the potentially high number of features.
3	It also does not do well with features that are highly correlated and one (or all) of them may be dropped when they do have an effect on the model when looked at together.	It performs better in cases where there may be high multi-collinearity, or high correlation between certain features
4	LASSO works better when you have more features and you need to make a simpler and more interpretable model, but is not best if your features have high correlation	Ridge Regression works better when you have less features or when you have features with high correlation
5	simpler and more interpretable	The Ridge Regression improves the efficiency, but the model is less interpretable due to the potentially high number of features.
6	It drops features that are correlated.	it does not drop features and in that case may lead to bad predictions.

Q12 Answer: -

Multicollinearity in regression analysis occurs when two or more predictor variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the regression model. The most common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model.

- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
- A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

Q13 Answer: -

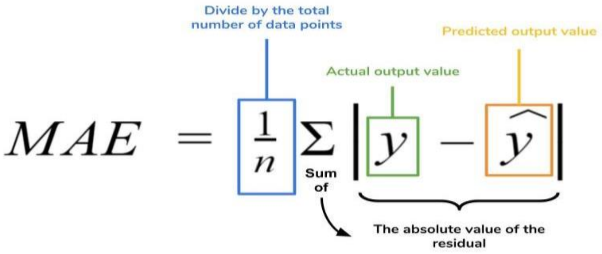
In regression modelling, scaling the goal value is a good idea; scaling the data makes it simple for a model to learn and grasp the problem. Scaling the target value is also a good idea. When we are applying machine learning algorithms to a data set, one of the procedures that falls under the umbrella of ″data pre-processing″ is scaling the data.

Q14 Answer: -

A goodness-of-fit test, in general, refers to measuring how well do the observed data correspond to the fitted (assumed) model.

To measure the performance of your regression model, some statistical metrics are used. They are-

- Mean Absolute Error (MAE): - This is the simplest of all the metrics. It is measured by taking the average of the absolute difference between actual values and the predictions.


Divide by the total number of data points
Predicted output value
Actual output value
$MAE = \frac{1}{n}\sum\left|y - \hat{y}\right|$
Sum of
The absolute value of the residual

- Root Mean Square Error (RMSE): - The Root Mean Square Error is measured by taking the square root of the average of the squared difference between the prediction and the actual value. It represents the sample standard deviation of the differences between predicted values and observed values (also called residuals).

It is calculated using the following formula,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

- Coefficient of determination or R2: - It measures how well the actual outcomes are replicated by the regression line. It helps you to understand how well the independent variable adjusted with the variance in your model. That means how good is your model for a dataset. The mathematical representation for R^2 is-

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Here, SSR = Sum Square of Residuals (the squared difference between the predicted and the average value)

SST = Sum Square of Total (the squared difference between the actual and average value)

Adjusted R2: - There is a drawback of R^2 that it improves every time when we add new variables in the model.

Think about it, whenever you add a new variable there can be two circumstances, either the new variable improves your model or not. When the new variable improves your model then it is ok. But what if it does not improve your model? Then the problem occurs. The value of R^2 keeps on increasing with the addition of more independent variables even though they may not have a significant impact on the prediction.

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - number of regressors

n - sample size

Q15 Answer: -

- Recall or Sensitivity = TP / (TP + FN) = (1000)/ (1000+1200) = 0.4545
- Specificity = TN / (TN + FP) = (50) / (50+250) = 0.1666
- Precision = TP / (TP + FP) = (1000) / (1000+250) = 0.80
- Accuracy = (TP+TN) / (TP+TN+FP+FN) = (1000+50) / (1000+50+250+1200) = 0.42