PYTHON

Ans.1)- C) %

Ans.2)- B) 0

Ans.3)-C) 24

Ans.4)- A) 2

Ans.5)-D) 6

Ans.6)- C) The finally block will be executed no matter if the try block raises an error or not.

Ans.7)- A) It is used to raise an exception.

Ans.8)- C) in defining a generator

Ans.9)- A) _abc C) abc2

Ans.10)- A) yield B) raise

Q.11 – Q.15) Link- https://github.com/sumeshyadav29/fliprobo/blob/main/Worksheet/Python.ipynb

MACHINE LEARNING

Ans.1)- B) In hierarchical clustering you don't need to assign number of clusters in beginning

Ans.2)- A) max_depth

Ans.3)- C) RandomUnderSampler

Ans.4)- C) 1 and 3

Ans.5)- D) 1-3-2

Ans.6)- B) Support Vector Machines

Ans.7)- C)) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)

Ans.8)- A, B, D

Ans.9)- B,C,D

Ans.10)- A, B, D

Ans.11)- When the categorical features present in the dataset are ordinal i.e. for the data being like Junior, Senior, Executive, Owner. When the number of categories in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption.

To fight the curse of dimensionality, binary encoding might be a good alternative to one-hot encoding because it creates fewer columns when encoding categorical variables. Ordinal encoding is a good choice if the order of the categorical variables matters.

Ans.12)- There are 7 types of techniques that can be used:

1. Use the right evaluation metrics-
   Applying inappropriate evaluation metrics for model generated using imbalanced data can be dangerous. Imagine our training data is the one illustrated in graph above. If accuracy is used to measure the goodness of a model, a model which classifies all testing samples into “0” will have an excellent accuracy (99.8%), but obviously, this model won’t provide any valuable information for us.

2. Resample the training set-
   Apart from using different evaluation criteria, one can also work on getting different dataset. Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.

3. Use K-fold Cross-Validation in the Right Way-

   It is noteworthy that cross-validation should be applied properly while using over-sampling method to address imbalance problems. Keep in mind that over-sampling takes observed rare samples and applies bootstrapping to generate new random data based on a distribution function.

4. Ensemble Different Resampled Datasets-
   The easiest way to successfully generalize a model is by using more data. The problem is that out-of-the-box classifiers like logistic regression or random forest tend to generalize by discarding the rare class.

5.Resample with Different Ratios-

The previous approach can be fine-tuned by playing with the ratio between the rare and the abundant class. The best ratio heavily depends on the data and the models that are used. But instead of training all models with the same ratio in the ensemble, it is worth trying to ensemble different ratios.

5. Cluster the abundant class-

An elegant approach was proposed by Sergey on Quora [2]. Instead of relying on random samples to cover the variety of the training samples, he suggests clustering the abundant class in r groups, with r being the number of cases in r. For each group, only the medoid (centre of cluster) is kept. The model is then trained with the rare class and the medoids only.

6. Design Your Models-

All the previous methods focus on the data and keep the models as a fixed component. But in fact, there is no need to resample the data if the model is suited for imbalanced data. The famous XGBoost is already a good starting point if the classes are not skewed too much, because it internally takes care that the bags it trains on are not imbalanced. But then again, the data is resampled, it is just happening secretly.

Ans.13)- GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters. GridSearchCV is a function that comes in Scikit-learns (or SK-learn) model selection package. So an important point here to note is that we need to have the Scikit learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit

your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

It's good to have large data sets because the larger the data set, the more we can extract insights that we trust from that data set.

Ans.14)- GridSearchCV is a method used in machine learning to find the best combination of hyperparameters for a given model. It exhaustively searches over all the possible hyperparameter combinations and evaluates them using cross-validation, allowing the user to select the best model configuration for their specific problem.

GridSearchCV can be useful when dealing with smaller datasets, as it provides a way to optimize model performance by systematically tuning hyperparameters. However, it can be computationally expensive and time-consuming when working with large datasets, as the number of hyperparameters and their combinations can become very large. In such cases, it may be more practical to use alternative techniques such as random search or Bayesian optimization, which can be more efficient and less computationally intensive.

In summary, GridSearchCV is a powerful tool for optimizing model performance, but its use should be considered in the context of the specific problem and dataset size.

Ans.15)- There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)

STATISTICS

Ans.1)- A) The probability of rejecting H0 when H1 is true.

Ans.2)- A) Correct hypothesis.

Ans.3)- A) Level of significance.

Ans.4)- B) The t distribution with n - 1 degrees of freedom.

Ans.5)- C) Rejecting Ho when it is false.

Ans.6)- D) A two-tailed test.

Ans.7)- B) The probability of committing a Type I error.

Ans.8)- A) The probability of committing a Type II error.

Ans.9)- A) $z > z\alpha$

Ans.10)- C) The level of significance.

Ans.11)-A) Level of significance.

Ans.12)- D)All of the Above.

Ans.13)- Analysis of Variance, i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables.

Ans.14)- There are three primary assumptions in ANOVA:

1. The responses for each factor level have a **normal population distribution**.
2. These distributions have the **same variance**.
3. The data are **independent**.

Ans.15)- The only difference between one-way and two-way ANOVA is **the number of independent variables**. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.