MACHINE LEARNING-8

Answer of Following Questions-

Ans 1-D) None of these

Ans 2- A) max_depth

Ans 3- A) SMOTE

Ans 4- D) 2 and 3

Ans 5-A) 3-1-2

Ans 6-D) Logistic Regression

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) Ridge will lead to some of the coefficients to be very close to 0

D) Lasso will cause some of the coefficients to become 0

Ans 9-C) Use ridge regularisation

D) use Lasso regularisation

Ans 10-A) Overfitting

C) Underfitting

Ans 11-One-Hot-Encoding has the advantage that the result is binary rather than ordinal and that everything sits in an orthogonal vector space.
The disadvantage is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.
Also Where For categorical variables where ordinal relationship exists, the one hot encoding is not enough. We have to use Label Encoder for ordinal data

.

Ans 12-An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed.
Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the
assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class.
Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.

1) Under-sampling
Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient.
By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling.

2) Over-sampling
On the contrary, oversampling is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples.
Rather than getting rid of abundant samples, new rare samples are generated by using e.g. repetition, bootstrapping or SMOTE (Synthetic Minority Over-Sampling Technique).

3) Cluster-Based Over Sampling
In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset.
Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.

4) Modified synthetic minority oversampling technique (MSMOTE) for imbalanced data
It is a modified version of SMOTE. SMOTE does not consider the underlying distribution of the minority class and latent noises in the dataset. To improve the
performance of SMOTE a modified method MSMOTE is used.

Ans 13-SMOTE: Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors,
joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors.
The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

ADASYN: Adaptive Synthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor.
The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data.
The algorithm uses Euclidean distance for KNN Algorithm. The key difference between ADASYN and SMOTE is that the former uses a density distribution,
as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the
different minority samples to compensate for the skewed distributions. The latter generates the same number of synthetic samples for each original minority sample

Ans 14-Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model.
There are libraries that have been implemented, such as GridSearchCV of the sklearn library, in order to automate this process.
Grid Search can be thought of as an exhaustive search for selecting a model. In Grid Search, the data scientist sets up a grid of hyperparameter values
and for each combination, trains a model and scores on the testing data. In this approach, every combination of hyperparameter values is tried
and when running it on larger dataset can be very inefficient.

For example, searching 20 different parameter values for each of 4 parameters will require 160,000 trials of cross-validation.
This equates to 1,600,000 model fits and 1,600,000 predictions if 10-fold cross validation is used.
While Scikit Learn offers the GridSearchCV function to simplify the process, it would be an extremely costly execution both in computing power and time.

Ans 15-There are three main errors (metrics) used to evaluate models, Mean absolute error, Mean Squared error and R2 score.

Mean Absolute Error (MAE): Lets take an example where we have some points. We have a line that fits those points.When we do a summation of the absolute value distance from
the points to the line, we get Mean absolute error.The problem with this metric is that it is not differentiable.

Mean Squared Error (MSE): Mean Squared Error solves differentiability problem of the MAE. Consider the same diagram above. We have a line that fits those points. When we do a summation of
the square of distances from the points to the line, we get Mean squared error.

R2 Score: R2 score answers the question that if this simple model has a larger error than the linear regression model. However, it terms of metrics the answer we need is how much larger.
The R2 score answers this question. R2 score is 1 — (Error from Linear Regression Model/Simple average model).

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y,
disregarding the input features, would get a R^2 score of 0.0.