Worksheet 8-MACHINE LEARNING

Ans1) B) In hierarchical clustering you don't need to assign number of clusters in beginning

Ans2) A) max_depth

Ans3) C) RandomUnderSampler

Ans4) C) 1 and 3

Ans5) D) 1-3-2

Ans6) C) K-Nearest Neighbors

Ans7) 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)

Ans8) 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.

Ans9) B) remove only one of the features

C) Use ridge regularization

Ans10) A) Overfitting

D) Outliers

Ans11) 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.

Ans12) Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. example of imbalanced data is –

· Disease diagnosis

· Customer churn prediction

· Fraud detection

· Natural disaster

The techniques that can be used to balance the data sets are as follows:

1. Resampling (Oversampling and Undersampling):

Undersampling is the process where you randomly delete some of the observations from the majority class in order to match the numbers with the minority class.

The second resampling technique is called, **Oversampling**. This process is a little more complicated than undersampling. It is the process of generating synthetic data that tries to randomly generate a sample of the attributes from observations in the minority class. There are a number of methods used to oversample a dataset for a typical classification problem. The most common technique is called **SMOTE (**Synthetic Minority Over-sampling Technique). In simple terms, it looks at the feature space for the minority class data points and considers its *k* nearest neighbours.

2. **Ensembling Methods (Ensemble of Sampler):**

In Machine Learning, ensemble methods use multiple learning algorithms and techniques to obtain better performance than what could be obtained from any of the constituent learning algorithms alone. (Yes, just like a democratic voting system). When using ensemble classifiers, bagging methods become popular and it works by building multiple estimators on a different randomly selected subset of data. In the scikit-learn library, there is an ensemble classifier namedBaggingClassifier. However, this classifier does not allow to balance each subset of data. Therefore, when training on imbalanced data set, this classifier will favour the majority classes and create a biased model.

Ans13) 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.

Ans14) GridSearchCV is a technique for **finding the optimal parameter values from a given set of parameters in a grid**. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.

The grid Search Cross-Validation technique is computationally expensive. The complexity of Grid Search CV increases with an increase in the number of parameters in the param grid. Thus Grid Search CV technique is not recommended for large-size datasets or param grids with a large number of components.