Task -1

1. Random Forest

A. What is the random forest? How does it work? Explain with clarity

Random forest is one of ensemble technique. To understand about random forest we will first understand what is ensemble technique. Ensemble learning is a technique in machine learning which takes the help of several base models and combines their output to produce an optimized model. Two main type of ensemble techniques are bagging and boosting. The “bagging” method is a type of ensemble machine learning algorithm called Bootstrap Aggregation. Bootstrap Aggregation can be used to reduce the variance of high variance algorithms such as decision trees.

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems in R and Python. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps:

- Step-1: Select random K data points from the training set.
- Step-2: Build the decision trees associated with the selected data points (Subsets).
- Step-3: Choose the number N for decision trees that you want to build.
- Step-4: Repeat Step 1 & 2.

ESTIMATORS

- n_estimators= The required number of trees in the Random Forest. The default value is 10. We can choose any number but need to take care of the overfitting issue.
- criterion= It is a function to analyse the accuracy of the split. Here we have taken "entropy" for the information gain.

B. How is it better than Decision Tree?

When using a regular decision tree, you would input a training dataset with features and labels and it will formulate some set of rules which it will use to make predictions. If you entered that same information into a Random Forest algorithm, it will randomly select observations and features to build several decision trees and then average the results.

Sometimes, because this is a decision tree-based method and decision trees often suffer from overfitting, this problem can affect the overall forest. This problem is usually prevented by Random Forest by default because it uses random subsets of the features and builds smaller trees with those subsets.

Random Forest is set up in a way that allows for quick development with minimal hyper-parameters (high-level architectural guidelines) which makes for less set up time.

C. How is it different from Xgboost?

Random forest builds trees in parallel, while in xgboost, trees are built sequentially. How trees are built: random forests builds each tree independently while gradient boosting builds one tree at a time. Combining results: random forests combine results at the end of the process (by averaging),

XGBoost delivers high performance, Its training is very fast and can be parallelized / distributed across clusters Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

2. How does Support Vector Machine work? When should you use it?

Support vector machine algorithm uses the concept of hyperplane which is used to distinguish between two classes. It is preferred over other classification algorithms because it uses less computation and gives notable accuracy. It is good because it gives reliable results even if there is less data. There can be many hyperplanes that can do this task but the objective is to find that hyperplane that has the highest margin that means maximum distances between the two classes, so that in future if a new data point comes that is to be classified then it can be classified easily. The concept of transformation of non-linearly separable data into linearly separable is called Cover’s theorem **– “**given a set of training data that is not linearly separable, with high probability it can be transformed into a linearly separable training set by projecting it into a higher-dimensional space via some non-linear transformation**”.** Kernel tricks help in projecting data points to the higher dimensional space by which they became relatively more easily separable in higher-

- Polynomial Kernel- The process of generating new features by using a polynomial combination of all the existing features.
- Radial Basis Function (RBF) Kernel- The process of generating new features calculating the distance between all other dots to a specific dot. One of the rbf kernels that is used widely is the Gaussian Radial Basis function.
- Linear Kernel - It is more generalized form of linear kernel and distinguish curved or nonlinear input space. Following is the formula for polynomial kernel.

In support vector machine, it finds lines or boundaries that correctly classify the training dataset. Then, from those lines or boundaries, it picks the one that has the maximum distance from the closest data points.

We use SVC when classes capable of performing binary and multi-class classification on a dataset. Numeric predictions problem can be dealt with SVM, It is effective on datasets that have multiple features

- Face Detection

Classifies the images of people’s faces in an environment from non-face by creating a square box around it.

- Bioinformatics

The Support vector machines are used for gene classification that allows researchers to differentiate between various proteins and identify biological problems and cancer cells.

- Text Categorization

Used in training models that are used to classify the documents into different categories based on the score, types, and other threshold values.

- Generalized Predictive Control(GPC)

Provides you control over different industrial processes with multivariable version and interactor matrix. GPC is used in various industries like cement mills, robotics, spraying, etc.

- Handwriting Recognization

SVM is widely used to recognize handwritten characters and test them against pre-existing data.

3. Given that your data has outliers, which regression algorithms you will try to use and why? Select at least one of them and explain its advantages and disadvantages

Regression algorithms are most commonly used for continuous datatypes cases. There are almost 16 regression methods which can be used depending on the situations. Like polynomial, simple linear, random forest, lasso, ridge, Elastic Net, Principal component regression, Ordinal, poisson etc. It has been said that if dependent variable is continuous and model is suffering from collinearity or there are a lot of independent variables, you can try PCR, PLS, ridge, lasso and elastic net regressions. We should select the final model based on Adjusted r-square, RMSE, MSE. If we are working on count data, you should try poisson, quasi-poisson and negative binomial regression. Ridge, lasso and elastic net regressions techniques are used to correct overfitting issue. Quantile regression is the extension of linear regression and can be used when outliers, high skeweness and heteroscedasticity exist in the data. Quantile regression is robust to outliers.

So lets discuss about most frequently used regression algorithm i.e Linear Regression model: Linear regression is a statistical method that enables us to summarise and study relationships between two continuous variables. Linear regression is a linear model wherein a model that assumes a linear relationship between the input variables and the single output variable . Here the output variable can be calculated from a linear combination of the input variables. When there is a single input variable, the method is called a simple linear regression. When there are multiple input variables, the procedure is referred as multiple linear regression.

Linear regression algorithm are mostly used in financial portfolio prediction, salary forecasting, real estate predictions and in traffic in arriving at ETAs.

Advantages of Linear Regression are it is a very simple algorithm that can be implemented very easily to give satisfactory results. Furthermore, these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms. Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms. The mathematical equations of Linear regression are also fairly easy to understand and interpret. Hence Linear regression is very easy to master.

Disadvantage of linear regression is it is prone to underfitting, Outliers have a very big impact on linear regression's performance. Even we need to check the multicollinearity in the data set before using this model