Exam
Machine Learning and Predictive Analytics – UFCFMJ-15-M

Student ID: **21051422**

Section A

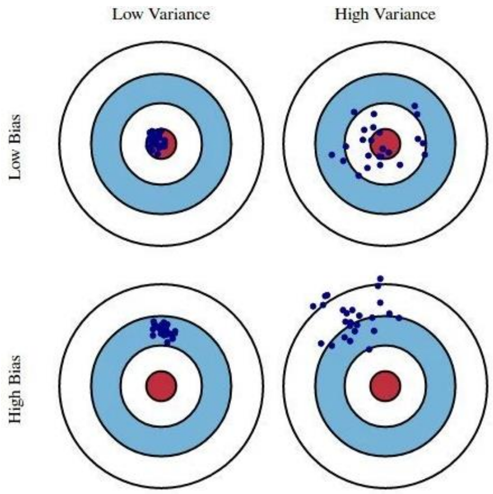
Answer 1:

Lasso and Ridge are ensembles of linear regression models. For knowing the difference between two methods, it becomes more important to understand the base behind these methods.

A brief explanation on linear regression models.

Linear regression models function on the relation between dependent variable, and one or more independent variables. Dependent variable is also known as response or target variable whereas independent variables are referred as explanatory variables or predictors (datacamp.com). The intercept and coefficient are found for that relation using least squares approach minimizing the R^2, which helps in determining the values for dependent variables for the given predictor/s. The most efficient application for these methods is predicting the future – example, prediction of prices, weather forecasting.

The predictors are critically characterized on the bases of bias and variance, where bias is the difference between true and predicted values and variance is the uncertainty (spread) in predicted values.



Source: kdnuggets.com

Difference between two ensembles – Lasso and Ridge.

Both Ensembles are mainly used to enhance the performance of linear regression models by optimizing the model complexity. This is known as regularization which aims at the reduction of variance by introducing some bias. There are some differences in the function of both methods.

Lasso:

Least Absolute Shrinkage and Selection Operator, as the name itself suggests on how Lasso regression functions, the main aim here is to remove the variable with high variance and thus making their coefficients to absolute zero (L1 penalty). It works on absolute value of magnitude of coefficients and thus, it makes the interpretability of the model simpler by reducing the number of predictors. The parameter of initialization is alpha but are smaller than Ridge regression and for the same value of alpha, Lasso fits poorly.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

When to use?
Lasso performs best in sparse model, which means when the predictors are large in number but only few of them holds importance. Example, while prediction of house price from data set containing information of houses and stars, variables like number of rooms, area of house, neighborhood matters important rather than distance of stars.

Why?
It simply removes the predictors which has less importance on dependent variable, thus by doing so it gives better model which is easy in interpretation and computationally less expensive.

Ridge:

Ridge regression decreases the model complexity by keeping all the variables in model instead of removing them by penalizing those variables with high variance, and thus making their coefficients near to zero (L2 penalty). It works on square of the magnitude of coefficients. The parameter of initialization is alpha, and when alpha increases the complexity of model reduces, which helps reducing overfitting of the model. The best alpha parameter is often selected by cross-validation technique.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

When to use?
Its application can give good results where there exists multicollinearity between predictors in the dataset. Example, for diamond price prediction, the length width and depth are highly correlated to carat, thus creates multicollinearity.
Also performs well, when the number of predictors are higher than the number of observations.

Why?
It minimizes square of coefficients by adding the penalty equivalent for squaring the magnitude of coefficients and thus introduces bias and removes multicollinearity.

Answer 2 [a.]:

Entropy using $\log_2$

$$\text{Entropy: } H(S) = -\sum_{i=1}^{C} p_i \cdot \log(p_i) = -\sum_{i=1}^{C} \frac{|S_i|}{|S|} \cdot \log(\frac{|S_i|}{|S|})$$

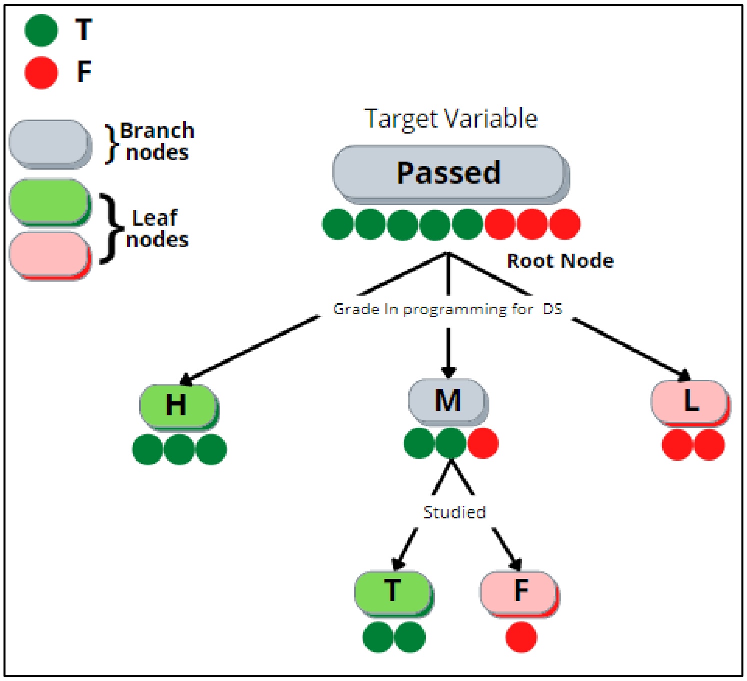
H (Passed): - 5/8*log (5/8) - 3/8*log (3/8) = **0.95**

H (Passed | Grade in Programming for DS): P(H)*H(H) + P(M)*H(M) + P(L)*H(M) = 3/8*0 + 3/8*0.91 + 2/8*0 = **0.34**

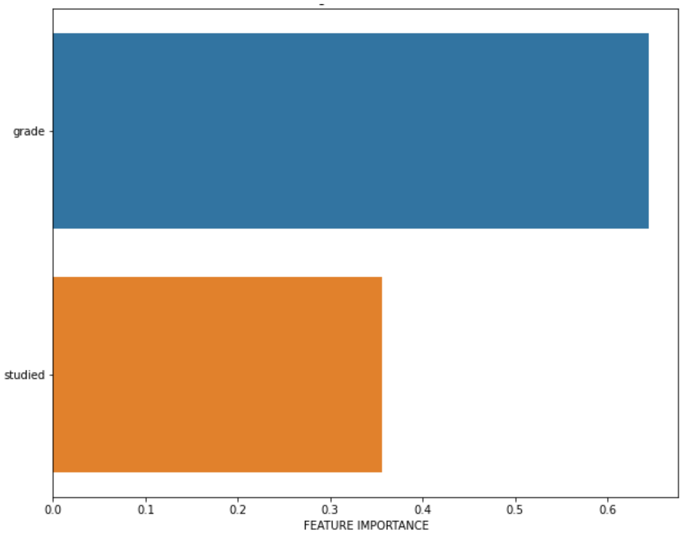
For calculating the entropy in python, I have used code provided by (Hisham) in the lecture of decision tree.

Answer 2 [b.]: Decision Tree and justification of choices with feature importance ranking.

Decision Tree



Feature Importance



The optimal decision tree has been developed using only two features as “Study hour” shows no importance, which could be justifiable as the same class for all observations that literally do not make any classification.

The variable “Grade in Programming for DS” has been selected for Root node according to ID3 classification, as it classifies the dataset in more appropriate way. It gives higher information of 0.61 in comparison with the information gain obtained from splitting the data with “Studied” variables, for which the information gain was only 0.16. Also, visually it could be seen from above decision tree diagram that the first split gives 2 clear leaf nodes while when checked with “studied” as root node it failed to give any leaf node.

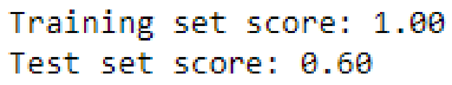
The facts mentioned justifies my choice of giving “Grade in programming for DS” as 1st rank, followed by “studied”. And no rank has been assigned to “study hours”.

Answer 2 [c.]:

- Accuracy of the learnt tree.

Accuracy: The accuracy of a machine learning algorithm in classification problem is a performance measurement to find how often a data point is classified by the algorithm correctly. It is the count of correct predictions out of all data points.

The DecisionTreeClassifier model was trained on the dataset given at the beginning of the question 2. The accuracy showed 1.00 on train set, meaning all the labels were predicted correctly. However, the accuracy dropped significantly to just **0.60** on the test set introduced in question 2.c.



```
Training set score: 1.00
Test set score: 0.60
```

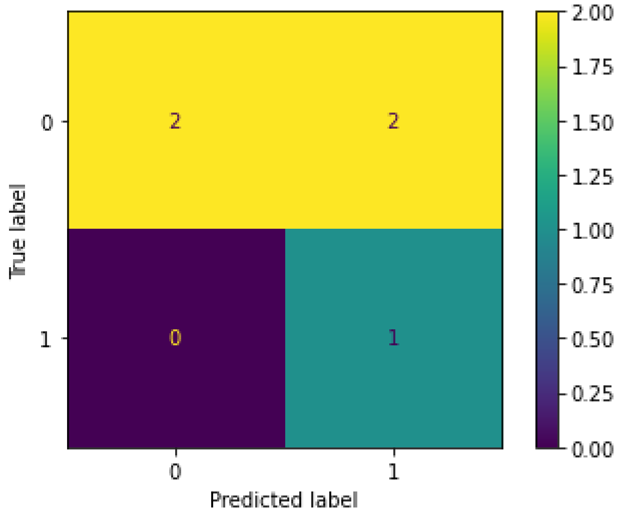
- General limitation of the Decision Tree method (Dhiraj, 2019).

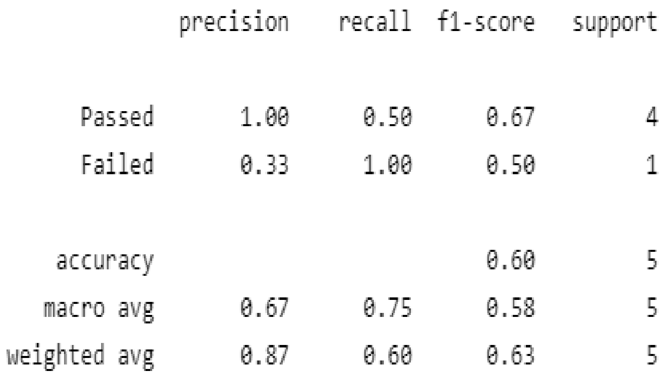
1) Decision Tree can become structurally instable largely with even a small change in the data.
2) In comparison to other algorithms, sometimes the calculations become more complex.
3) It required higher time for training the model which increases computational cost.
4) It may work well in classification, but while predicting it performs inadequately.
5) There are chances of overfitting while training as it can split so many times without stopping method.
6) It is biased towards the dominant class in dataset which has imbalance classes.

Limitations faced in the given data set.

In the training dataset, the variable "Study hour" became less important because it had only one class "E", and it did not help in making any classification. While in test dataset, this variable had 3 classes "E", "N" and "L".

- Confusion Matrix. F1 Score.





```
              precision    recall  f1-score   support

      Passed       1.00      0.50      0.67         4
      Failed       0.33      1.00      0.50         1

    accuracy                           0.60         5
   macro avg       0.67      0.75      0.58         5
weighted avg       0.87      0.60      0.63         5
```

Here, the TN (true negative) and TP (true positive) are 2 and 1 respectively, meaning that 3 predictions are correctly made from 5 test observations. And FP (false positive) shows 2 meaning that the students who failed the module, are being wrongly predicted as passed. Also, the F1 score of passed class is 0.67 because 2 out of 3 predictions were made correctly and for failed its 0.50, meaning 2 out of 4 predictions were made correctly.

Answer 2 [d.]:

Ensemble methods are the combination of multiple trees and produce much better predictions collectively rather than just one tree trained in decision tree method. A main idea behind this is the overall idea of many good professional is better than an expert (Hisham). Ensemble methods helps overcome many limitations of decision tree and thus enhances the prediction performance. It helps in handling missing values in the dataset and thus keeps the accuracy maintained. The higher dimension data is well handles using ensemble technique – Random Forests. Whereas another ensemble – Gradient Boosting improves performance by supporting many different loss functions and handles interactions well (Anuja, 2017). Ensemble techniques used many different parameters randomly and thus trains the model based on many trees trained with subset of parameters at each node. It uses voting system. Ensemble algorithms are easy to tune and train. Ensemble learning used sequential and parallel methods to exploit the dependence present in base learners. Several machine learning techniques can be used in one model that helps reduce variance and bias thus improving the predictions than decision tree.

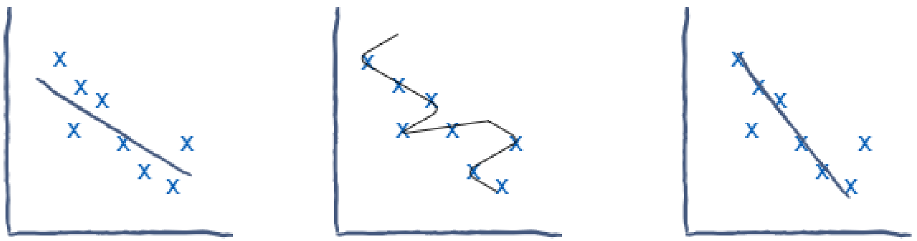
Section B

Answer 3 [a.]:

Model fitting is a way of measurement used to know how well a trained machine learning model generalizes to a dataset similar to the one it is trained on. Accurate approximating the output when provided with unseen inputs is a sign of good model fit. A properly fitted model is equipped with hyperparameters which enables the model to understand the complex relations between the predictors and target variable, on succession of which it gives insights from data and make accurate predictions. Accuracy of a good fitted model is known by comparing the difference between true and predicted values.

Fitting works automatically but it can be tuned with the help of parameters that varies in different machine learning models to make It work efficiently and reduce the errors it could make.

Model doesn't fit accurately on the test set every time which tells about the problems of overfitting and underfitting. An overfitted model performs well in training by learning the noise in the training dataset on test set but fails to fit in the test dataset or data of similar kind. On the other hand, an underfitted model performs very poor on training dataset and thus stands no chance of going further into production. Thus, the fitting of model accurately is considered an important step in machine learning.



Underfitted model

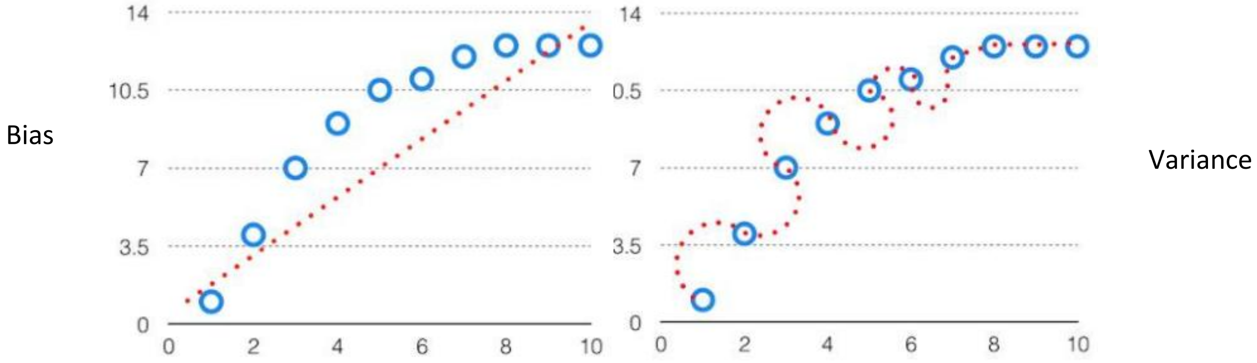
Overfitted model

Fitted model

Source: educative.io

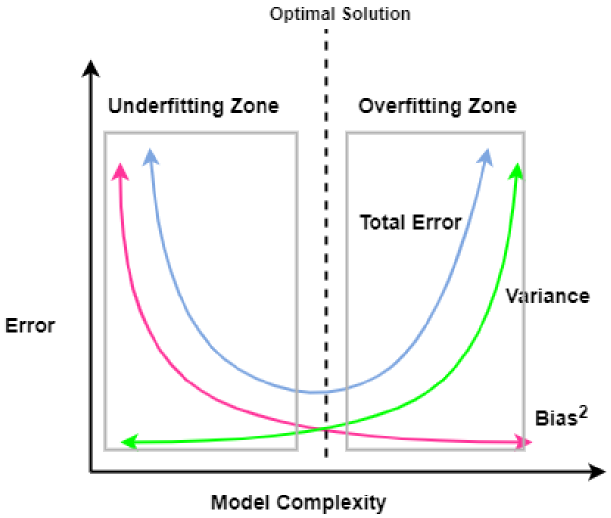
Answer 3 [b.]:

Bias is nothing but the difference of true value and the predicted value by machine learning model. High bias leads to huge errors in both training and test data sets and creates a problem of underfitting. Variance, on the other hand is the variability of predictions which explains the spread if the data set. High variance in model, if not taken care can lead the model towards overfitting.



Source: geeksforgeeks

A model's ability to balance between variance and bias is known as Bias-Variance tradeoff. It refers to the best solution that can determine a value with optimized regularization constant. Establishing a proper tradeoff would save the model from underfitting or overfitting the dataset. An algorithm is error-prone if it is too simple which may be low-variance and high-bias and inversely, a too complex algorithm with high degree may be on low-bias and high-variance.



Bias-Variance Graph

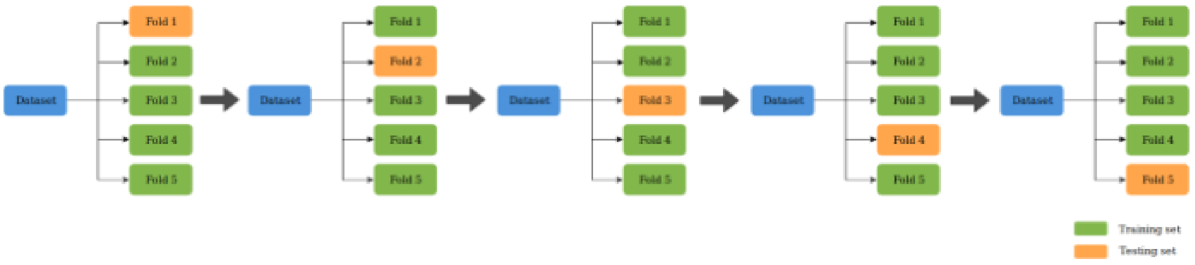
Source: geeksforgeeks

In the above image, the dashed line showing the optimal solution acts as the bias-variance tradeoff. This is very difficult to achieve in real word, but the algorithms can be kept improving with the goal to find iterative process to make the predictions accurate.

Answer 3 [c.]:

Cross Validation is one of the several techniques to evaluate model whose intention is finding the generalizability of the trained model. In other words, to check whether the model is predicting accurately on unseen data. Before taking the model into production, thorough check is necessary to ascertain the model's performance. For which, one of the traditional techniques is to split the data into train and test sets in 70/30 or 80/20 proportion, by doing so we train the model on train set and test its accuracy on test set. One big limitation of this method is one-time random split can lead into imbalance partition which leads to estimating biased generalization estimations.

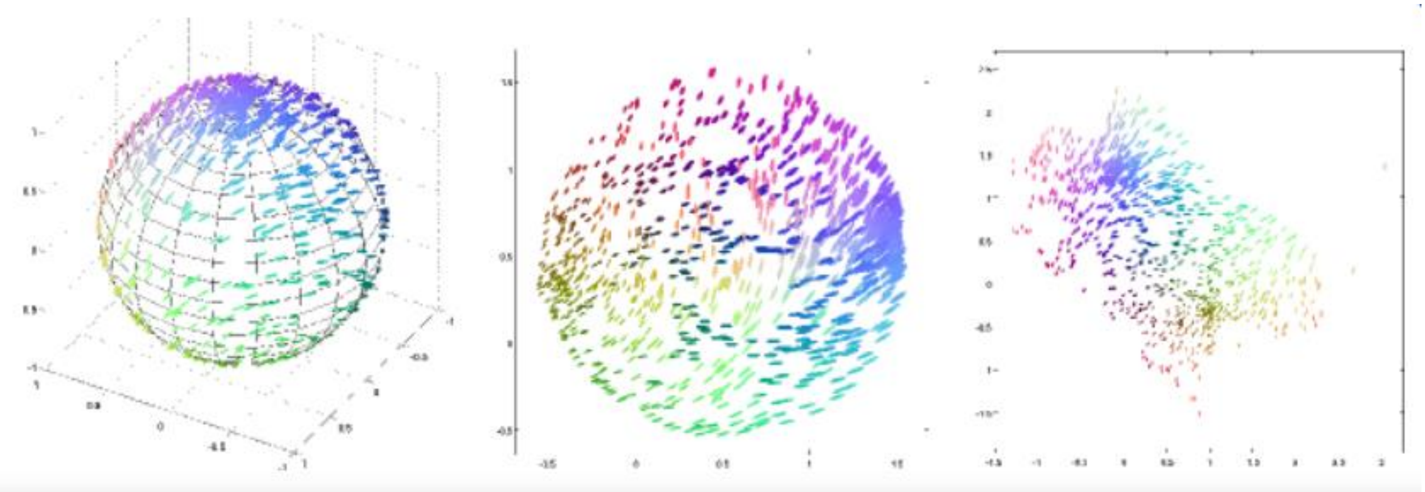
This drawback is handled by using cross validation which helps evaluate model in a better way. Here, the split is executed among n number of parts generally known as k-folds. Each fold is then used as a test set, while remaining part is used as the training set. This operation is repeated k number of times by which every observation is used for training as well as testing. Finally, the error metric is calculated taking the mean of k-folds. Commonly 5 and 10 are more commonly used as k.



As the Cross validation uses every observation for testing it enables to perform models accuracy in the most efficient way and without the need for another data, it used the present data for train and test optimizing the models accuracy.

Answer 3 [d.]:

Real life business problems require the machine learning algorithms to be trained on the dataset with too many factors/variables/dimensions. With the increase in these dimensions, it becomes difficult to visualize the training set, the solution to this is dimensionality reduction. As some of the features are redundant which adds noise It reduces the quantity of features and thus transforms data to low dimension feature space from high dimension feature space.



The advantage of dimensionality includes the reduction in storage to a greater extent thus allows to reduce the computational cost. For example, in facial recognition, High quality images are transformed into lower quality image. It makes algorithms more accurate by reducing noise (Nilesh, 2021). It also solves the curse of dimensionality.

Source: semanticscholar

Curse of dimensionality

The various phenomena arising while the analysis and organization of data in high-dimensional spaces which is not occurring in low-dimensional spaces are referred as the curse of dimensionality. The problem associates with the volume of space which increases very fast with the increase in dimensionality that makes available data sparse. At high dimensions many data organization strategies become inefficient as all object appears to be sparse. Decision trees, neural networks, k-means, SVM are some of the algorithms that suffers from curse of dimensionality.

Answer 3 [e.]:

Coefficient of Determination (R Squared: R^2)

The measurement to determine strength of linear relationship between two variables which determines the proportion of variance in the dependent variable with the change in independent variable. In other words, it can be put as how well a model fits the data. It ranges from 0 to 1, that suggests the dependency of a dependent variable between 0% to 100% on predictor/s. Thus, a high value indicated the model fits good on data and lower value defines the model does not fit well. When the model is trained only one variable, the coefficient of determination relates to Pearson’s correlation. It is very important to note that it only shows the relation and not causation.

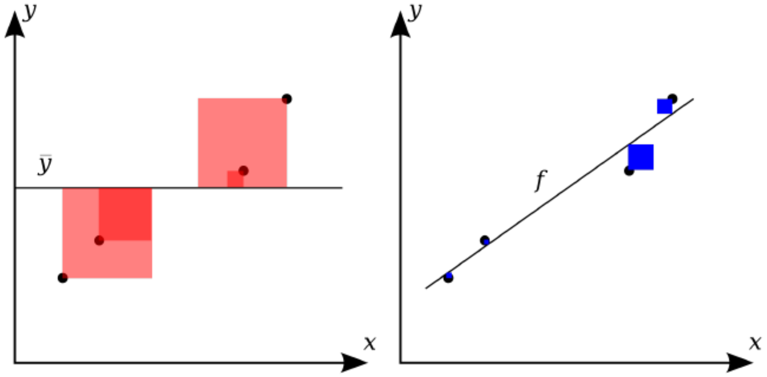
$$R^2 = \left(\frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - \left(\sum x\right)^2} \sqrt{n \sum y^2 - \left(\sum y\right)^2}} \right)^2$$

With the increase in in predictor variable, even if that has least or no correlation, we can see the increase in R^2. This becomes a problem in some cases, because even if the new variables are not useful, it will increase the R^2. This gives rise to a new approach called Adjusted R^2 also known as fitting by weighted least square where the residuals are generalized.

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

Coefficient of Determination use for Linear Regression models

Linear regression models helps in prediction of the target variable assuming that they are linearly correlated with the independent variables. The coefficient of determination is used to measure how accurately those predictions are made. For this, firstly the residuals are found by summing the squares or absolute difference of original and predicted values and is found.



Source: ML regression session (Hisham)

The left image shows absolute difference and right shows the squared distance. The distance are measured by the Euclidean, cosine and Chebyshev methods. It is also known as MSE (mean squared error). MSE itself is a performance measuring method for regression models and acts as a base for other models like RMSE (Root mean squared error), MAE (Mean Absolute Error). When the values are high MSE can be biased towards them thus that is solved by square rooting it. R2 or MSE shows only the association magnitude and does not states whether it is statistically significant or not.

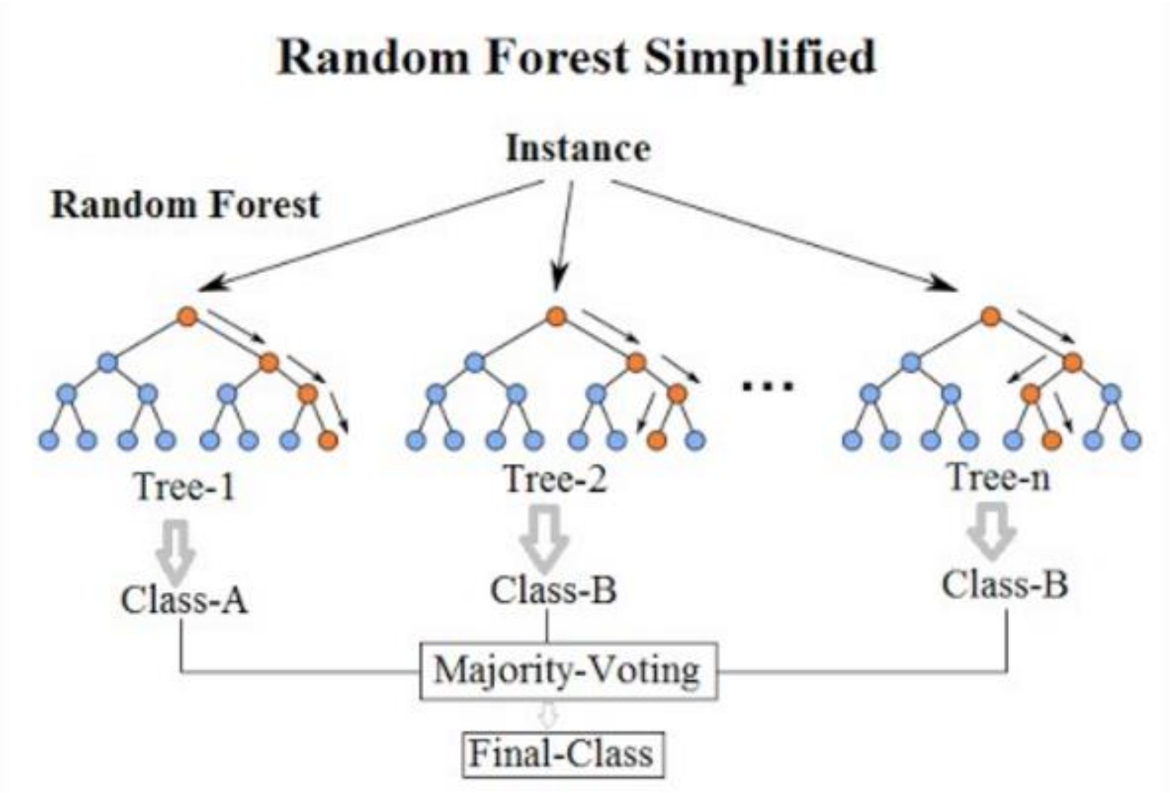
Answer 5 [a.]:

Random Forest is one of the most popular machine learning algorithms in machine learning, it handles both regression and classification problems. It was first created by Ho in 1995, later in 2006 it was registered as trademark by Leo Breiman and Adele Cutler (Wikipedia), as of 2019 it was owned by Minitab, Inc. Business models use RF as blackbox.

RF is an ensemble method of decision tree method and works on the base concept of decision tree but eliminates its drawbacks. Decision tree works on mostly if and else statement arguments on each node, From the Root node a series of questions are asked and based on how important they are, the rank is being assigned to split the node. The quality of node is mainly checked by using the metrics such as Gini impurity, information gain (entropy) for classification or MSE for regression. Then branch nodes are created and are developed until leaf nodes are formed for accurate classification.

There are several ensemble methods of decision tree like stacking, boosting and the most popular – Bagging. RF algorithm is a further extension of bagging method. The feature randomness feature of RF creates a random sample of features and ensures that the correlation between decision tree is less. The key difference is RF selects subsets of features while all possible features splits are considered by Decision tree.

RF are tuned by 3 main hyperparameters. These are the number of trees, node size and number of features sampled. IT takes random samples from the data sets through bootstrap but keeping one third data for testing known as out of bag. Several trees with are fully grown without any pruning and the result of each tree is collected. Now according to regression problem, the average of all trees' result is taken and for classification the most frequent variable is chosen. Finally, to finalize the prediction, cross validation is done on out of bog set. This is how the whole Random Forest method works. They are easy in application, robust and gives high performance after little feature engineering and tuning parameters.



Source: Wikipedia

Answer 5 [b.]:

Feature importance is a technique through which the score for each independent variable or input features for a given model, can be calculated. The importance of each feature is represented by this score. Not every tree is trained will all features and thus the trees internally are not corelated. Feature importance is a crucial step in training random forests algorithm. It becomes very important to decide from what variable's condition we can determine the root node. As the first split is plays a significant role in the model's accuracy. And each dataset is into 3 buckets thus it becomes crucial to know the importance of each feature derived from how pure the bucket is.

To measure the relative importance of each feature on prediction is very easy and a great feature of random forests. To give a better intuition, features that are selected at the top of the trees are in general more important than features that are selected at the end nodes of the trees, as generally the top splits lead to bigger information gains.

Some key features of RF are

- Diversity
- Immune to curse of dimensionality
- Parallelization
- Train-test-split
- Stability

There are 2 simple methods for determining the feature importance in random forests.

- Mean Decrease impurity
  Firstly, RF is combination of several DT. Each node in tree is a result of a condition of a variable. Impurity is measured for choosing optimal conditions. For classification it is Gini impurity or information gain. And for regression tree it is variance. According to this measure, the decreased impurity the each feature can be averaged.

- Mean Decrease Accuracy
  This is another method where the impact of each feature is measured on model accuracy. According to this method is permutation of the value of features and measure the change in accuracy of model.

Answer 5 [c.]:

Although RF is a high performing and widely used algorithm, it cannot be over trusted for its limitations which are as follows:

1) Undoubtedly RF has more accuracy than decision tree, but they lose intrinsic interpretable as DT.
2) RF required more resources and time for execution of its algorithm; thus, it becomes computationally very expensive.
3) Trust and confidence cannot be built as similar to DT because it doesn't allow to confirm that model has learned realistic information.
4) In a situation where there is a linear correlation between target and predictive variables, RF may fail to enhance base learner's accuracy.
5) It fails to improvise accuracy when the data has many categorial variables.
6) RF becomes too slow and ineffective when it comes for real time predictions.
7) There is a high chance of overfitting if feature engineering and hyperparameter tuning is not done efficiently.
8) It is not a descriptive tool which will just make predictions without giving any descriptions.
9) It fails to determine the significance of each variable
10) It is high on the risk of overfitting the data set.