1)What is caret?

Caret stands for **classification and regression training** and is arguably the biggest project in R. This package is sufficient to solve almost any classification or regression machine learning problem. Caret contains tools for pre-processing, feature selection, data splitting, training and testing functions, model comparison.

2)What is Boruta?

Boruta is a wrapper approach of feature selection.

Wrapper approach is where subset of features is used to train a model using them. Based on the inferences we draw from the previous model we decide to add or remove features from our subset.

Boruta is used for feature selection: It is based on RF. For each attribute it creates a shadow attribute and this shadow attribute has all the values shuffled across. And then we create models which include shadow attribute + original and assess the importance of each attribute. Idea is if a variable is not doing better than shadow attribute in terms of importance then we should not have those variables in model. Maximum z scores across all attributes are calculated.

3)Why used this three algo only?

NB- It doesn't require as much training data, it is highly scalable with the number of predictors and data points. It is fast compared to many other algorithms. It also performs well in multi-class prediction.

--What is NB algo

It is based on Bayes theorem. (p(y/x1x2x3…))=p(x1/y).(x2/y).p(x3/y)….p(y)/p(x1).p(x2).p(x3)

Fundamental naïve bayes approach is that each feature makes an independent and equal contribution to the outcome. And the probability of all the classes for given condition is compared and the instance is classified to the class having highest probability.

DT- Outcome can be easily understood. Structure follows the natural flow of human thought. Very good with both numerical as well as categorical data.

--What is DT algo

It’s a binary tree that recursively splits the dataset until we are left with pure leaf nodes, that is data with only one type of class. There are 2 types of nodes, one is a decision node (containing branches) and other is leaf node. Initially all dataset is given to root node and depending upon the decision the dataset gets split further.

RF- It is very good when it comes to categorical data.

Decision trees are highly sensitive to training data and which could result to high variance so our model might fail to generalize.

RF is collection of multiple random decision trees and its much less sensitive to the training data. In rf first bootstrapping is performed and then aggregation and together it is called bagging.

Procedure – First we have original dataset and then we have bootstrapped dataset. Bootstrapped dataset is generated randomly and can contain duplicate values. Then decision trees are built from these bootstrapped datasets using random variables. And a test point is passed through each decision tree and aggregation is performed.

LR can be used only for 2 class outcomes. SVM and KNN are distance based which means it will be good for numeric data but not for categorical data.

4)Why used feature extraction?  
It was used to see if we get better accuracy when we take only the most important features. For DT with all features, we get accuracy of 98.66 percentage. Then we used Boruta function to remove the least important features (got 14 most imp features) and then after applying dt we got accuracy of 98.66 percent. For NB with all features, we get accuracy of 87.92 percent. Then we used Boruta function to remove the least important features (got 14 most imp features) and then after applying NB, we got accuracy of 93.62 percent.

5)Where is dataset taken from?  
Data world

6)Hyperparameters of all algo?

Hyper-parameter for DT is cp (complexity parameter) and that for NB is fL and adjust. For RF it is mtry

Cp – The complexity parameter (cp) used to control the size of decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue.

fL (Laplace correction) –

mtry – No. of variables randomly collected to be sampled at each split time. Default value is floor division of square root of number of columns in original dataset.

7)What is as.factor?

as.factor is used **when you want to convert the data type of a variable to a factor/categorical variable**.

8)What is set.seed?

 By using the set. Seed () function, you guarantee that the same random values are produced each time you run the code.

9)Why we use setseed?  
Need is for reproducible results.

10)How to interpret 3\*3 confusion matrix

Horizontal is FN

Vertical is FP