Interview Questions – Set 1

1. How do You deal with a dataset having more features than data points? For example, I have 500 variables and 200 data points. How do I deal with it?
2. How do You try to reduce features even before applying any ML Model?
3. What is VIF and its threshold. Are You applying it on independent variables or dependent variables?
4. How do You deal with imbalanced dataset?
5. About cross validation, What exactly it does?
6. How do You reduce overfitting? What are the ways?
7. What is AUCROC curve and how do You chose the best threshold value? Why do You go for AUCROC instead of other metrics
8. Various feature selection techniques
9. Stratified k fold, K fold cross validation
10. What is Gradient Descent and what exactly happens in it?
11. What are the assumptions of Simple Linear Regression ?
12. What is Heteroscadasticity?
13. What is Recall and where do we use it?
14. What is Multi collinearity and How do You deal with it?
15. What VIF exactly says ?
16. What is Variance, Correlation and why collinearity is important?
17. Any clustering techniques that You had worked on?
18. Why p-value is 0.05 and why not other values ?
19. Where do we use ANOVA ?
20. Difference between RandomForest and XGBoost ?
21. In Randomforest, how do You take samples? Based on what criteria the samples are selected ?
22. RandomForest criteria and look for full parameters of RF.
23. Architecture of logistic Regression
24. What is Sigmoid function? What exactly it is ?
25. What is activation function and what are other activation functions for logistic?
26. If You have large dataset, how do You deal with it? Suppose if You have 10MB of data, the model execution time is more, then how do You deal with this scenario
27. What are the different approaches to reduce model execution time?
28. Though infrastructure handles the huge volume, still taking more execution time . Within the model what can You do to reduce execution time?
29. How do You select features?
30. Before implementing RF, how do You select features?
31. Have You ever tried to optimize the model ? Different optimization techniques available?
32. How Gradient descent help to optimize Linear Regression?
33. After building Your model, Your model is having more bias and less variance. How do You deal with this scenario.
34. What are the different approaches You follow in the above scenario.
35. Suppose looking at the prediction , You can say the model is biased, then what You will do.
36. How do You handle Underfitting ? What are the techniques of Underfitting?
37. Optimization vs Regularization techniques.
38. When will You go for Regularization techniques.

39. Explain L1 and L2 regularization.
40. What is coefficient in Linear regression analysis? What it says ?
41. How do You describe p-values for each coefficient, how are they related and what p-values do ?
42. Is p-value siginificant in regression and how do we calculate it?
43. R^2 vs adjusted R^2
44. Difference between collinear and correlation.
45. How do You interpret categorical variable in a model ? You have one dataset having many categorical variables in the model , how do You deal with them.
46. OneHotEncoding vs LabelEncoding
47. PIN is a categorical variable ?
48. A column named PIN has 20 labels, how many features would be created?
49. In Your dataset there are outliers, but they are important and cannot remove them. How do we deal this scenario.
50. Different Hypothesis tests used so far.
51. How will You use oversampling and explain with an example. What it does exactly?
52. Have You ever implemented the above scenario.
53. How entropy is different from GINI index.
54. How GINI and entropy helps to classify the data.
55. Suppose TOI had different newspapers and it requires different volumes of newspapers on daily basis. How many volumes they will print on daily basis and what are the ML techniques can be helpful in this scenario. Which algorithm is suitable for daily prediction
56. Signal vs Noise
57. Different metrics for classification problem.
58. Difference between F1-score and accuracy.
59. In layman terms , how do You explain F1-score and accuracy
60. In logistic regression, what is classification accuracy?