Statistics.

1. Measures of Central Tendency – Mean,Mode,Median
2. Measures of Spread-Variance & Standard Deviation
3. Correlation- Pearson,Chi-Square,ANOVA & Correlation does not mean Causation
4. Histogram, Normal Distribution – Symmetric around Mean, Empirical Rule
5. Central Limit Theorem
6. Hypothesis Testing- z-test,t-test

Data Preprocessing Steps.

7. Missing Values treatment- Mean/Median for continuous variables, Mode for Categorical ones
8. Outlier treatment – Boxplot,Quartiles, z-score
9. MinMaxScaling & z-score Normalization for Continuous variables
10. One Hot & Label Encoding for Categorical variables
11. Feature Enginnering based on business knowledge
12. Feature Selection using Correlation, VIF & Lasso
13. Plots – Histogram, Scatter plot, Box plot

Model Building.

<u>Supervised Models</u>

14. Linear Regression – Gradient Descent (Stochastic, Batch & Mini Batch), R2- score, Adjusted R2-score and MSE, Assumptions- VIF for Multicollinearity, Homoskedasticity, Autocorrelation of errors.
15. Logistic Regression- Odds, Logit & Sigmoid Function, Maximum Likelihood Function
16. Decision Tree-
    i. Entropy, Information Gain & Gini Impurity- Classification
    ii. Mean Squared Error – Regression.
    iii. Pruning
17. Ensemble Models:
    i. Bagging- Random Forest
        a. Boot Strapped Samples
        b. Aggregating
        c. Out Of Bag score
    ii. Boosting –
        a. Adaboost (Weight of sample, Weight of tree)
        b. Gradient Boost - MART
        c. XGBoost - DART
18. Support Vector Machines – Kernel Functions (Sigmoid, RBF & Polynomial), C and Gamma
19. K Nearest Neighbors- Eucledian , Manhattan Distance
20. Naïve Bayes Classifier- Bayes Theorem, MultinomialNB, GaussianNB & BernoulliNB

<u>Unsupervised Models:</u>

21. Principal Component Analysis – Eigen Values & Eigen Vectors
22. K Means Clustering – Elbow Curve

Loss Functions:

23. Mean Squared Error, Mean Absolute Error
24. LogLoss & Hinge Loss

Model Building Challenges:

25. Overfitting, Solution – Regularization: penalizing the coefficients (Lasso, Ridge, ElasticNet) , Cross Validation- KFold & StratifiedFold
26. Hyperparameter Tuning: GridSearchCV & RandomizedSearchCV, Bayesian
27. Imbalanced Dataset- SMOTE
28. AIC (Akaike Information Criterion) for Model Selection – Parsimonious model

Metrics to Measure Model Performance

29. R2-score, Adjusted R2-score, MSE, MAE
30. Confusion Matrix – Precision, Recall/TPR/Sensitivity, Specificity, AUC ROC, f1-score

Natural Language Processing.

31. Tokenization
32. Stopword Removal
33. Lemmatization, Stemming
34. Bag of Words, Count & Tf-Idf Vectorizers of Documents in the Corpus
35. Parts of Speech tagging
36. Regular Expressions
37. Text Classification
38. Word2Vec, Cosine Similarity

Time Series Forecasting

39. Stationarity Check – Presence of Trend, Seasonality. Check using Augmented Dickey Fuller test
40. Moving Average & Exponential Smoothing
41. Differencing to get rid of Trend, Log transformations to get rid of Variance.
42. ARIMA, Auto ARIMA
43. Autocorrelation & Partial Auto Correlation Functions

Deep Learning.

44. Neural Networks- Architecture
45. Activation Functions – Sigmoid, Tanh, ReLu, Leaky ReLu
46. Vanishing Gradient problem
47. Forward & Backpropagation
48. Dropout Regularization
49. Convolution Layers- Filters, Zero Padding & Maxpooling
50. Softmax Function
51. Model Compiling arguments: Loss, Optimizer & Metrics
52. Recurrent Neural Networks: For Sequential data

Trade-Offs:

53. Bias-Variance
54. Precision – Recall
55. Explanatory Power & Performance
56. Learning Rate & Number of Estimators

Python Concepts:

57. List, Tuple, Set, Dictionary
58. For Loops, If Conditions, List Comprehension
59. Functions, Lambda Expressions, Map,Filter, Reduce
60. Exception Handling
61. Args & Kwargs
62. OOPs
63. Libraries- Numpy, Pandas, Seaborn, Matplotlib, sklearn, imblearn, scipy, statsmodels, nltk, Spacy, Gensim, WorNet
64. Handling Date formats
65. Web Scrapping – BeautifulSoup library
66. Deploying ML Models as API
67. Microservices Architecture
68. Database Connections- MySQLdb, PySpark

Tell me something that I don't know:

69. UberEATS Estimated Time of Delivery Model.
70. Derivation of LogLoss from Bernoulli equation
71. Zomato's Weighted Ratings
72. Difference between Frequentist & Bayesian
73. SVM Kernel trick explanation – (0,3),(1,2),(2,1),(3,0), Kernels : X+Y, X-Y, XY
74. Haversine distance
75. Kubernetes & Docker concepts

Sources to Learn.

76. Andrew Ng – Coursera
77. MIT Lectures
78. StatQuest
79. Medium
80. Towards Data Science
81. Geeks4Geeks & Telusko YT Channel for Python
82. Google Developers Course
83. ML 100 Page Book
84. Kaggle's No Free Hunch – Winner Interviews