**Interview Preparation Guide**

1. Model evaluation techniques for Classification, clustering and regression?
2. R Square
3. ROC
4. Confusion Matrix
5. Chi Sqaure selection
6. GINI Index
7. Information Gain
8. Cost function
9. Regularization (L1 – Lasso(linear penalty term is added to cost function) L2- Ridge Regression( polynomial penalty term))
10. Hypothesis Testing
11. Significance level – alpha
12. P- value
13. Dimensionality Reduction Methods
14. Decision tree where to split?
15. Time Series Analysis (ARIMA etc)
16. Is L1 Regularization used for feature reduction?

**Script**

Hi my name is Prakash, I am a graduate in computer science and engineering with total 4 years of experience in Analytics field. I am keen to learn new data science technologies and currently I am learning python to upgrade my skill set.  
  
As far as my professional experience is concerned - I have total 3.5 years of exp.  
Recently I was working with Deloitte as an Associate Sol Advisor. I was a part of RISK and Financial Advisory team and we were closely working with our Internal Audit Team. We used to get Client's General Entry data(GL Data) and TRIAL BALANCE (TB) from Audit team then We help them in General Entry Testing.We help them to reconcile GL data with the given Trial Balance sheet.  
We try to find out suspected entries made into client's general entry data to mitigate the risk.We primarily use SAS, excel, Vba and some other Deloitte proprietary tools for our day to day activities.We get clients data on monthly basis, quarterly basis or yearly basis. we share Reconciliation Reports, Date Time Summary, Period Summary, Account summary, man\_sys Summary and all the discrepancies to our clients.  
  
Before this I was working with Ingenious E Brain Solutions as a Research Associate, where I worked with an E commerce client based in UK. We used to get customer's order details, payment details, Delivery Details and Return Details based on which we need to identify customer's buying behaviour, Top Selling Product category, potential fraud customers, product category hit by max no of return, City wise fraud, Suspected fraud customers, month on month sale increment etc.  
  
Why you left deloitte?  
It was a fixed term role. For last 2 years Deloitte has started hiring people on fixed term roles for their so called Busy Season (which starts from Oct and ends on Feb).Its basically the Q4 i.e Year End testing period when they do have lots of work from their existing clients. Although based on my performance they were extending my tenure by 2 months and out of 30 people they offer extension to only 5 people and luckily I was one among them. But again they were not assuring for any permanent offer. At the same time I lost my best friend in an accident in Dehradun so I was really upset and I need a break.So I did not accept the offer and decided to get another job after a break.  
  
  
Why you left your previous companies so frequently?  
My first company was Alstom. I left Alstom because I wanted to prepare for some PSU exams. I did my preparation but unfortunately, I couldn't qualify them so I decided to get back to work and I joined HGS. There I was on third party payroll. When I got an offer from Ingenious E brain solutions so I decided to join that company. Everything was going good and suddenly there came an Medical Emergency at my home, My father met an accident and he was hospitalized. Doctor advised him to have Hip joint Replacement surgery. We took our father from my hometown and we admitted him to fortis Noida and where his Surgery was done by the doctors.  
At that time I need to take care of my father. I was already on leaves for a month and it was not possible for my employer to extend my leaves further. So I decided to quit my job and I was taking care of my father. In the meantime I joined a Training Institute based in noida as a SAS Trainer. In a short span of 4 months, I trained around 100 professionals and students in SAS. It was a great experience for me.I learned too many things from that experience.Then I got an offer from Deloitte and I joined that.  
  
Again it was a contractual role. Initially I thought if I work hard there could be chances to get permanent.Later on I realized that they are not going to offer permanent role bcz of the lack of work.I got listed my name in the top performer list for the month of DEC & JAN. Although they are extending my contract for another 6 months but not giving any assurity on offering a permanent role.  
  
How soon you can join?  
I can join in a month. I do have some commitments.Right now I am working as SAS trainer and I need to wind-up my batches, hand over batches to new trainer and need to give some notice to the management as well.Second is I am living in a rented flat so I need to give a month's prior notice to my landlord.Otherwise he will deduct amount from my advance deposit and I also need to attend my cousin's wedding on 14th June. So I thing it would be better if you give me one months time.  
  
  
Do you have any offer in hand?  
Yes, currently I do have an offer from EY, they are hiring for their banking client and they are offering me 11 LPA. They are waiting for some kind of approval from their management.  
  
What is your expected salary?  
As per industry standards.It would be great If you can provide some hike over my existing offer.  
  
what is your current CTC?  
its lpa.  
  
Is there anything you want to ask?  
1. What are the metrics you will be evaluating me on.What topics will be covered.  
2.

Important Terms in ML

1.       Linear Regression – from sklearn.

2.       Logistic Regression - performs binary classification

**The sigmoid/logistic function is given by the following equation: y = 1 / 1+ e-x**

3.       **Cost Function** -

4.       **Regularization** - Overfitting happens when model learns signal as well as noise in the training data and wouldn’t perform well on new data on which model wasn’t trained on. Now, there are few ways you can avoid overfitting your model on training data like cross-validation sampling, reducing number of features, pruning, regularization etc.

Regularization basically adds the penalty as model complexity increases. Regularization parameter (lambda) penalizes all the parameters except intercept so that model generalizes the data and won’t overfit.

5.       R2 Score – Fitness of Good

6.       LASSO Regression - linear term (lambda) is added with the cost function also called as L1 Regularization. Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “absolute value of magnitude” of coefficient as penalty term to the loss function.

7.       RIDGE Regression – Squared term (lambda) is added with the cost function also called as L2 Regularization. Ridge regression adds “**squared magnitude**” of coefficient as penalty term to the loss function.

8.       Feature Scaling –  Rescale variables (e.g. standardizing or normalizing)

  from sklearn.preprocessing import StandardScaler

scaler = StandardScaler().fit(X)

rescaledX = scaler.transform(X)

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

scaler.fit(df.drop('TARGET CLASS',axis=1))

scaled\_features = scaler.transform(df.drop('TARGET CLASS',axis=1))

df\_feat = pd.DataFrame(scaled\_features,columns=df.columns[:-1])

df\_feat.head()

9.       **Cross Validation ( K Fold Cross Validation)** – test data is splited into k subsets and each time k-1 subsets will be used for training the model and K- th subset is used to validate the model.

from sklearn import cross\_validation

data = cross\_validation.KFold(len(train\_set), n\_folds=10, indices=False)

**10.**   **Grid search**

11.   Random Search

12.   Random Forest – How to choose next splitting node?

It is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging.

The forest that it builds is an ensemble of Decision Trees as we previously talk and most of the time it is trained with the “bagging” method. The basic concept behind the bagging method is that a combination of learning models increases the overall result.

13.   Naïve – Bias Classifier – Based on conditional probability. Best used with Categorical data.

Naive Bayes can be used in real-world applications such as:

Sentiment analysis and text classification

Recommendation systems like Netflix, Amazon

To mark an email as spam or not spam

Face recognition

P( A | B)  = P (B | A) \* P(A) / P(B)

14.   SVC - widely used in pattern recognition and classification problems — when your data has exactly two classes.

from sklearn.svm import SVC

SVM can be used in real-world applications such as:

detecting persons with common diseases such as diabetes hand-written character recognition

text categorization — news articles by topics stock market price prediction

The objective of the support vector machine algorithm is to find a hyperplane in N-dimensional space (N — the number of features) that distinctly classifies the data points.

15.   **PCA** **with Mathematic Explaination** - Reduce data redundancy and dimensionality

<https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643>

16.   Neural Network

17.   NLP

18.   Model Tuning – Hyperparameters

19.   KNN

from sklearn.neighbors import KNeighborsClassifier

20.   K-Means Clustering - needs to know in advance how many clusters there will be in your data

from sklearn.cluster import KMeans

kmeans = KMeans(n\_clusters=4)

21.   Trade-off between Bias and Variance

22.   Classification Report

23.   Accuracy

24.   Precision

25.   Recall

26.   F1 Score

27.   Confusion Matrix – Type 1 Error and Type 2 Error

28.   Bootstrap -

29.   Bagging

30.   Boosting - In short, it combines multiple weak or average predictors to a build strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix.

Boosted algorithms are used where we have plenty of data to make a prediction. And we seek exceptionally high predictive power. It is used for reducing bias and variance in supervised learning.

31.   XGBoost

32.   Gradient Decent Algo

In real time applications, it is obviously very important to have a prediction as fast as possible. For instance, in autonomous driving, it’s important that the classification of road signs be as fast as possible to avoid accidents.

**Machine Learning Algorithms and Python Packages**

1.       Linear Regression – **from sklearn.model\_selection import LinearRegression**

2.       Model Validation – **from** **sklearn.model\_selection import LinearRegression**

3.       Logistic Regression

4.       Train – Test Split - **from sklearn.model\_selection import train\_test\_split**

5.       SVM - **from sklearn.svm import SVC**

**clf = SVC(kernel='linear')**

6.       PCA -

7.       Decision Trees -

8.       KNN

9.       K Means Clustering

10.   Random Forest

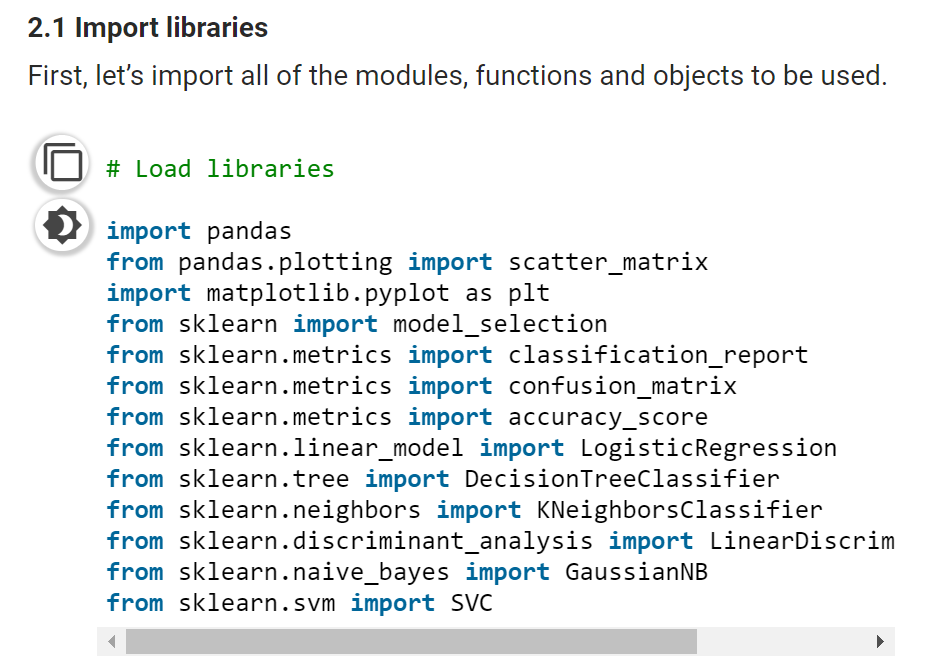
11.   NLP

12.   Model Validation

13.   Confusion Matrix

14.   Classification Report

15.   K- Fold Cross Validation - help us keep overfitting under control.



import pandas

from pandas.plotting import scatter\_matrix

import matplotlib.pyplot as plt

from sklearn import model\_selection

from sklearn.metrics import classification\_report

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import accuracy\_score

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.discriminant\_analysis import LinearDiscriminantAnalysis

from sklearn.naive\_bayes import GaussianNB

from sklearn.svm import SVC

**Python**

1.       List comprehension

2.       Lambda expression

3.       Fibonaci series

4.       Generator

5.       Class

6.       Vector

7.       Array

8.       Pandas library

9.       Functions

10.   Converting categorial variables to numeric

11.   Dataframe manipulation using Pandas

12.

**Statistical Terms**

13.   **P Value** - P is a statistical measure that helps researchers to determine whether their hypothesis is correct. It helps determine the significance of results.

[P value](https://www.wallstreetmojo.com/p-value-in-excel/) is a number between 0 and 1. There are tables, spreadsheet programs and statistical software to help calculate the p-value. Level of significance (α) is a pre-defined threshold set by the researcher. It is generally 0.05. A very small p-value, which is lesser than the level of significance indicates that you reject the null hypothesis. P value which is greater than the level of significance indicates that we fail to reject the null hypothesis.

2. Significance Level –(Alpha) - Probability of rejecting null hypothesis when it is True.

3. Confidence interval

4. Central Limit Theorem

5. Normal Distribution

6. **Type-1 Error –** When we reject Null Hypothesis when it is True. A type I error is the false rejection of the null hypothesis.

7**. Type-2 Error –**Fail to reject Null Hypothesis when it is False.

**8.** T -test

**9.** Z – Test

**Mathematical Terms**

1.       Probability Theory

2.       Calculus

3.       Matrix

4.       Vectors

5.       Transpose

6.       Matrix and Vector operations

7.       Eigenvalues

8.       Eigen Vectors

9.       Derivatives

10.   Partial Derivatives

11.   Log

12.

Machine learning Questions

1. What are assumptions of Linear Regression? What are the most common estimators for linear regression?
2. What is the formula for logistic regression? How it is used for binary classification?
3. How the decision tree decides on its split?
4. What advantages does a Decision Tree model have?
5. What is the difference between a Random Forest and versus Boosting Tree Algorithms?
6. Given a dataset with features X and Labels Y, what assumptions are made when using Naive Bayes Theorem?
7. What is Overfitting and what causes it? What ways can you use to avoid Overfitting?
8. Describe the difference between Accuracy, Precision and Recall?
9. What metrics can be used to evaluate a Regression task?
10. Describe how the Support Vector Machine (SVM) algorithm works?

|  |  |
| --- | --- |
|  | |
| - s |  |

**NEW**

**Important Terms in ML**

1.     Linear Regression – from sklearn.

2.   Logistic Regression - performs binary classification

**The sigmoid/logistic function is given by the following equation: y = 1 / 1+ e-x**

3.   **Cost Function** –

4.   **Regularization** - Overfitting happens when model learns signal as well as noise in the training data and wouldn’t perform well on new data on which model wasn’t trained on. Now, there are few ways you can avoid overfitting your model on training data like cross-validation sampling, reducing number of features, pruning, regularization etc.

Regularization basically adds the penalty as model complexity increases. Regularization parameter (lambda) penalizes all the parameters except intercept so that model generalizes the data and won’t overfit.

5.    **R2 Score** – Fitness of Good- Other metric to evaluate the performance of linear regression is R-square and most common metric to judge the performance of regression models. R² measures, “How much the change in output variable (y) is explained by the change in input variable(x).

          R-squared is always between 0 and 1:

0 indicates that the model explains NIL variability in the response data around its mean.

1 indicates that the model explains full variability in the response data around its mean.

**In general, higher the R², more robust will be the model.**

6.   **LASSO Regression** - linear term (lambda) is added with the cost function also called as L1 Regularization. Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “absolute value of magnitude” of coefficient as penalty term to the loss function.

**Asked Interview Questions**

