Linear Regression (Learn about p-value, t-statistics, ANOVA)

Logistic Regression

Support Vector Machine (Kernel)

Principal Component Analysis (PCA)

Neural Networks (Backpropogation)

Convolutional Neural Network

K-Nearest Neighbor

K-Means Clustering

Agglomerative Clustering

Mathematical Part-> Statistics, Probability, Set-theory's, Matrix
Programming->python/R
Database->Mysql
Strong communication ability.
basic data structures

Interview:

use dataset provided to build a k-NN model that avoids time leakage

What is the performance of the model measured in Median Relative Absolute Error?

What would be an appropriate methodology to determine the optimal k?

Note:

dataset given has four columns [latitude, longitude, close_date, close_price]

To prevent time leakage, a home j should be considered a neighbor to home i only if the close date of j occurred prior to the close date of i. Think about making a prediction using information available to house i . You only want to use information you have available at that time. One way of doing this is to restrict yourself to neighbors that have closed prior to the close date of i .

The Median Relative Absolute Error (MRAE) is defined as median (|P_predict - P_true|/P_true)

Microsoft Data Scientist new grad interview:

I had a 5 rounds (in addition to a technical phone screen) as follows:

1. **Technical Phone Screen:** A coding question (LC similar question: last index to place a flower), and a lot of discussion around p-values, central limit theorem and a probability question.
2. **Technical Round 1**: A lot of discussion around developing different types of metrics for measuring the quality of Bing Search & its recommendations, effectiveness of Windows OS, how to evaluate these metrics to improve these products
3. **Technical Round 2**: A coding round: Q1: How can we generate random numbers from 1-10 if we have a function to generate random numbers from 1-7. Q2: Discussion around designing a distributed system for k merging sorted lists.
4. **Technical Round 3**: A brief discussion on my background and past experience, followed by a product analytics question focusing on statistical inferencing, z-tests, p-values, A/B testing.
5. **Technical Round 4**: A brief discussion regarding my background and then some discussion on statistical tests, bayesian inferencing,
6. **Technical Round 5**: A coding question (don't remember the exact question, but probably an LC Easy-Medium) and then some behavioral question on time management, conflict resolution.

Interview Question For Data Science

Introduction:

Data science is an interdisciplinary field that mines raw data, analyses it, and comes up with patterns that are used to extract valuable insights from it. Statistics, computer science, machine learning, deep learning, data analysis, data visualization, and various other technologies form the core foundation of data science.

Over the years, data science has gained widespread importance due to the importance of data. Data is considered as the new oil of the future which when analyzed and harnessed properly can prove to be very beneficial to the stakeholders. Not just this, a data scientist gets the exposure to work in diverse domains, solving real-life practical problems all by making use of trendy technologies. The most common real-time application is fast delivery of food in apps such as Uber Eats by aiding the delivery person shows the fastest possible route to reach the destination from the restaurant. Data Science is also used in item recommendation systems in e-commerce sites like Amazon, Flipkart, etc which recommends the user what item they can buy based on their search history. Not just recommendation systems, Data Science is becoming increasingly popular in fraud detection applications to detect any fraud involved in credit-based financial applications. A successful data scientist can interpret data, perform innovation and bring out creativity while solving problems that help drive business and strategic goals. This makes it the most lucrative job of the 21st century.

Data Science Interview Questions for Freshers

1. What does one understand by the term Data Science?

An interdisciplinary field that constitutes various scientific processes, algorithms, tools, and machine learning techniques working to help find common patterns and gather sensible insights from the given raw input data using statistical and mathematical analysis is called Data Science.

- It starts with gathering the business requirements and relevant data.
- Once the data is acquired, it is maintained by performing data cleaning, data warehousing, data staging, and data architecture.
- Data processing does the task of exploring the data, mining it, analyzing it which can be finally used to generate the summary of the insights extracted from the data.
- Once the exploratory steps are completed, the cleansed data is subjected to various algorithms like predictive analysis, regression, text mining, recognition patterns, etc depending on the requirements.
- In the final stage, the results are communicated to the business in a visually appealing manner. This is where the skill of data visualization, reporting, and different business intelligence tools come into the picture.

2. What is the difference between data analytics and data science?

- Data science involves the task of transforming data by using various technical analysis methods to extract meaningful insights using which a data analyst can apply to their business scenarios.
- Data analytics deals with checking the existing hypothesis and information and answers questions for a better and effective business-related decision-making process.
- Data Science drives innovation by answering questions that build connections and answers for futuristic problems. Data analytics focuses on getting present meaning from existing historical context whereas data science focuses on predictive modeling.
- Data Science can be considered as a broad subject that makes use of various mathematical and scientific tools and algorithms for solving complex problems whereas data analytics can be considered as a specific field dealing with specific concentrated problems using fewer tools of statistics and visualization.

3. What does it mean when the p-values are high and low?

A p-value is the measure of the probability of having results equal to or more than the results achieved under a specific hypothesis assuming that the null hypothesis is correct. This represents the probability that the observed difference occurred randomly by chance.

- Low p-value which means values ≤ 0.05 means that the null hypothesis can be rejected and the data is unlikely with true null.
- High p-value, i.e values ≥ 0.05 indicates the strength in favor of the null hypothesis. It means that the data is like with true null.
- p-value = 0.05 means that the hypothesis can go either way.

4. When is resampling done?

Resampling is a methodology used to sample data for improving accuracy and quantify the uncertainty of population parameters. It is done to ensure the model is good enough by training the model on different patterns of a dataset to ensure variations are handled. It is also done in the cases where models need to be validated using random subsets or when substituting labels on data points while performing tests.

5. What do you understand by Imbalanced Data?

Data is said to be highly imbalanced if it is distributed unequally across different categories. These datasets result in an error in model performance and result in inaccuracy.

6. Are there any differences between the expected value and mean value?

There are not many differences between these two, but it is to be noted that these are used in different contexts. The mean value generally refers to the probability distribution whereas the expected value is referred to in the contexts involving random variables.

7. What do you understand by Survivorship Bias?

This bias refers to the logical error while focusing on aspects that survived some process and overlooking those that did not work due to lack of prominence. This bias can lead to deriving wrong conclusions.

8. Define the terms KPI, lift, model fitting, robustness and DOE.

KPI: KPI stands for Key Performance Indicator that measures how well the business achieves its objectives.

- Lift: This is a performance measure of the target model measured against a random choice model. Lift indicates how good the model is at prediction versus if there was no model.
- Model fitting: This indicates how well the model under consideration fits given observations.

- Robustness: This represents the system's capability to handle differences and variances effectively.
- DOE: stands for the design of experiments, which represents the task design aiming to describe and explain information variation under hypothesized conditions to reflect variables.

9. Define confounding variables.

Confounding variables are also known as confounders. These variables are a type of extraneous variables that influence both independent and dependent variables causing spurious association and mathematical relationships between those variables that are associated but are not casually related to each other.

10. Define and explain selection bias?

The selection bias occurs in the case when the researcher has to make a decision on which participant to study. The selection bias is associated with those researches when the participant selection is not random. The selection bias is also called the selection effect. The selection bias is caused by as a result of the method of sample collection.

Four types of selection bias are explained below:

- Sampling Bias: As a result of a population that is not random at all, some members of a population have fewer chances of getting included than others, resulting in a biased sample. This causes a systematic error known as sampling bias.
- Time interval: Trials may be stopped early if we reach any extreme value but if all variables are similar invariance, the variables with the highest variance have a higher chance of achieving the extreme value.
- Data: It is when specific data is selected arbitrarily and the generally agreed criteria are not followed.
- Attrition: Attrition in this context means the loss of the participants. It is the discounting of those subjects that did not complete the trial.

11. Mention some techniques used for sampling. What is the main advantage of sampling?

Sampling is defined as the process of selecting a sample from a group of people or from any particular kind for research purposes. It is one of the most important factors which decides the accuracy of a research/survey result.

Mainly, there are two types of sampling techniques:

Probability sampling: It involves random selection which makes every element get a chance to be selected. Probability sampling has various subtypes in it, as mentioned below:

- Simple Random Sampling
- Stratified sampling

- Systematic sampling
- Cluster Sampling
- Multi-stage Sampling

Non- Probability Sampling: Non-probability sampling follows non-random selection which means the selection is done based on your ease or any other required criteria. This helps to collect the data easily. The following are various types of sampling in it:

- Convenience Sampling
- Purposive Sampling
- Quota Sampling
- Referral /Snowball Sampling

12. What is bias in Data Science?

Bias is a type of error that occurs in a Data Science model because of using an algorithm that is not strong enough to capture the underlying patterns or trends that exist in the data. In other words, this error occurs when the data is too complicated for the algorithm to understand, so it ends up building a model that makes simple assumptions. This leads to lower accuracy because of underfitting. Algorithms that can lead to high bias are linear regression, logistic regression, etc.==

13. What is dimensionality reduction?

Dimensionality reduction is the process of converting a dataset with a high number of dimensions (fields) to a dataset with a lower number of dimensions. This is done by dropping some fields or columns from the dataset. However, this is not done haphazardly. In this process, the dimensions or fields are dropped only after making sure that the remaining information will still be enough to succinctly describe similar information.

14. Why is Python used for Data Cleaning in DS?

Data Scientists have to clean and transform the huge data sets in a form that they can work with. It is important to deal with the redundant data for better results by removing nonsensical outliers, malformed records, missing values, inconsistent formatting, etc.

Python libraries such as Matplotlib, Pandas, Numpy, Keras, and SciPy are extensively used for Data cleaning and analysis. These libraries are used to load and clean the data and do effective analysis. For example, a CSV file named "Student" has information about the students of an institute like their names, standard, address, phone number, grades, marks, etc.

15. How is Data Science different from traditional application programming?

Data Science takes a fundamentally different approach in building systems that provide value than traditional application development.

In traditional programming paradigms, we used to analyze the input, figure out the expected output, and write code, which contains rules and statements needed to transform the provided input into the expected output. As we can imagine, these rules were not easy to write, especially, for data that even computers had a hard time understanding, e.g., images, videos, etc.

Data Science shifts this process a little bit. In it, we need access to large volumes of data that contain the necessary inputs and their mappings to the expected outputs. Then, we use Data Science algorithms, which use mathematical analysis to generate rules to map the given inputs to outputs.

This process of rule generation is called training. After training, we use some data that was set aside before the training phase to test and check the system's accuracy. The generated rules are a kind of a black box, and we cannot understand how the inputs are being transformed into outputs.

However, If the accuracy is good enough, then we can use the system (also called a model).

As described above, in traditional programming, we had to write the rules to map the input to the output, but in Data Science, the rules are automatically generated or learned from the given data. This helped solve some really difficult challenges that were being faced by several companies.