

Data Scientist technical Interview

Data Scientist technical interview questions & tasks

February 17, 2022

PaSHA HOLDING

# Notes

1. Please avoid sharing the technical interview file and questions with third parties.
2. Using open sources to understand concepts better while solving the tasks is totally allowed and recommended. Nevertheless, it is forbidden to replicate answers from these sources. To ensure the originality of the solutions Plagiarism Check will be performed. Likewise, you should refrain from getting direct help from another person.
3. You are given open tasks meaning that there is not an explicit answer to the questions rather it is expected to put a Data Scientific approach and solution to the questions.
4. It is recommended to elaborate solutions as much as you can while keeping them succinct.
5. You should not only give ultimate answers to the questions but also explain your way of solving them and why do you prefer such an approach.

# R2 vs MAPE (30 points)

Imagine your regression model has low R2 and MAPE (Mean Absolute Percentage Error). At first glance, low R2 is a sign of bad predictions by the model, however, low MAPE might be considered as a good result. In what conditions this (low R2 and low MAPE simultaneously) may happen? Would it be due to some relation between actual and predicted values? How would you interpret such a model? (Consider that the model has R2 = 20% and MAPE = 10%).

# Cancer Detection Model (20 points)

You are building a cancer detection model from a given data. What would be your main focus on the predictions of the model? What performance metric would you use to evaluate the model?

# Imputing missing values (30 points)

Below you are given a sample of a dataset from Company X. Dataset contains employees’ personal details and there is a single missing value in the Salary column. Fill in the missing value.

|  |  |  |  |
| --- | --- | --- | --- |
| EmployeeID | Experience (in years) | Age | Salary |
| 1 | 5 | 31 | 2100 |
| 2 | 10 | 40 | 3700 |
| 3 | 1 | 24 | 4900 |
| 4 | 1 | 24 | 700 |
| 5 | 2 | 26 | NA |
| 6 | 3 | 27 | 1100 |
| 7 | 2 | 25 | 850 |
| 8 | 8 | 30 | 3000 |
| 9 | 4 | 28 | 1900 |
| 10 | 7 | 29 | 2400 |

# Service Duration (20 points)

Consider that there is a business process that customers are given loans. For each customer, the service time of the business process may differ depending on various reasons. Compute representative usual service time and also identify abnormal services. Explain your work.

|  |  |
| --- | --- |
| Customer | Service Duration (in minutes) |
| 1 | 15 |
| 2 | 18 |
| 3 | 23 |
| 4 | 14 |
| 5 | 19 |
| 6 | 95 |
| 7 | 14 |
| 8 | 28 |
| 9 | 26 |
| 10 | 20 |