Data Scientist Interview Master Guide - Vivek Choudhari

Must-Have Technical Questions & Answers

Q: Walk me through a machine learning project you led.

A: Sure. One major project I led was chest X-ray pathology detection. We built a CNN-based model to detect 5 diseases, including Cardiomegaly and Pneumothorax. I handled data curation, preprocessing (normalization, augmentation), model training with transfer learning, and achieved 93%+ validation accuracy. We deployed the model on a Linux server and integrated it into the hospital's clinical viewer, reducing diagnosis time by 80%.

Q: How do you handle missing values in time series data?

A: I begin by checking the pattern of missing data - random or systematic. For time series, I use methods like forward fill, interpolation, or rolling average. If it impacts model accuracy, I create a missing-flag feature or use model-based imputation. For critical gaps, I may drop that segment. My goal is to preserve the trend and seasonality.

Q: What is LSTM and how is it used in time series prediction?

A: LSTM stands for Long Short-Term Memory. It's a type of RNN that avoids vanishing gradient problems by using memory cells and gates. In time series, LSTM captures long-term dependencies - like in sensor data or energy usage patterns - where past values influence future predictions.

Q: When would you use SVM vs Decision Trees?

A: SVM works well for small to medium datasets with clear margins between classes. It's effective with high-dimensional data, like text. Decision Trees are better when interpretability is required and the data has clear hierarchical splits.

Q: Explain SHAP and LIME - when would you use each?

A: Both are model explainability tools. SHAP uses game theory to explain global and local predictions. LIME explains individual predictions by fitting local interpretable models. I use SHAP for dataset-level insights and LIME for specific predictions.

Q: How do you deploy a model into production?

A: I package the model using Flask or FastAPI. Then deploy on a Linux server or Docker container, exposing an API endpoint. I monitor model input drift and performance metrics.

Q: How do you ensure a model is production-ready?

A: It must generalize well on unseen data. I validate performance metrics, test edge cases, ensure logging, version control, and monitor live inputs post-deployment.

Time Series Focused Questions

Q: What challenges arise in industrial time series modeling?

A: Missing or irregular data points, sensor drift, high variable correlation, multivariate dependencies, and cold-start

issues. I address these using LSTM, feature engineering, and anomaly detection.

Q: How do you preprocess sensor data for ML modeling?

A: Convert timestamp to datetime, create lag features and rolling stats, normalize, handle missing values, and resample if needed.

Q: What metrics do you use to evaluate time series models?

A: MAE, RMSE, MAPE, and R^2 depending on the context and business requirement.

Cloud, Tools & Infra

Q: Have you deployed ML models on cloud?

A: I've deployed on Linux servers using Flask. While I haven't done full cloud deployment, I'm Azure AI-900 certified and understand blob storage, ML endpoints, and CI/CD basics.

Q: What visualization tools do you use?

A: Matplotlib, Seaborn, Power BI, Tableau. Used them for EDA, customer segmentation, and reporting results to stakeholders.

Behavioral / HR Questions

Q: Tell me about yourself.

A: I'm a Data Scientist with 4+ years of experience in ML and AI, with projects in healthcare and utilities. I specialize in deep learning, NLP, time series, and deploying models to production.

Q: Tell me about a time when a model failed. What did you do?

A: During a churn project, the model failed due to campaign-driven behavior change. I added recent data, retrained weekly, and monitored prediction confidence.

Q: How do you handle pressure or urgent timelines?

A: I break tasks into blocks, prioritize impact, and communicate early with stakeholders. In the spine project, I delivered core segmentation first under time constraints.

Q: Why should we hire you for this role?

A: I bring deep ML experience, strong ownership of full-lifecycle projects, and proven results in time-sensitive environments. I'm quick to learn and adapt.

Q: Where do you see yourself in 2-3 years?

A: Leading AI projects, mentoring teams, and deploying models that solve critical industrial problems.