**✅ Part 2: Cheat Sheet to Revise Before Interview**

**Machine Learning Workflow**

* Data Acquisition → EDA → Feature Engineering → Model Selection → Training → Tuning → Validation → Deployment
* Use examples like Chest X-ray, British Gas Churn Model

**Key Algorithms**

* Regression: Linear, Lasso, Ridge
* Classification: SVM, Random Forest, XGBoost
* Clustering: K-Means, DBSCAN
* Deep Learning: CNN, LSTM
* Explain LSTM and Time Series concepts clearly

**Time Series**

* Techniques: Rolling mean, differencing, seasonal decomposition
* Forecasting: ARIMA, Prophet, LSTM
* Challenge: Handling missing data, multivariate data

**Projects to Pitch**

* Chest X-ray (CV + deep learning)
* British Gas Churn Model (EDA → NLP → Clustering → Forecasting)
* Tender Summarisation (LLM + RAG)
* Spine Segmentation (CV + Integration into workflow)

**Visualization**

* Show knowledge of Power BI, Tableau, and use Matplotlib/Seaborn daily

**Cloud & DevOps**

* Say: "I have experience deploying models on Linux servers. While I haven’t worked extensively with cloud, I am confident in adapting to AWS or Azure and have cleared Azure AI-900 certification."

**Soft Skills**

* Stakeholder communication
* Leading cross-functional teams
* Training juniors (show leadership)

**✅ Part 3: Mock Interview Questions (30-min, DS Role)**

**🔹 Technical Questions**

1. Walk me through a complete ML project you've led.
2. How do you handle missing data in time series?
3. When would you use SVM vs Decision Tree?
4. What is LSTM? How is it better than traditional RNN?
5. Tell me about your experience with NLP. How did you apply it?
6. What are SHAP and LIME? When would you use each?
7. How do you validate a model for deployment in production?
8. Explain a time when you had to choose between multiple algorithms.
9. What steps do you take to make a model interpretable for business?
10. Describe your approach to model monitoring post-deployment.

**🔹 Behavioral / Scenario-Based**

1. Tell me about a time you disagreed with a stakeholder. How did you handle it?
2. How do you prioritize tasks across multiple projects?
3. Have you trained or mentored team members?
4. What would you do if your model performed poorly in real-time after deployment?

**🔹 Role-Specific**

1. How comfortable are you with industrial IoT data or sensor logs?
2. Have you deployed any models on cloud or edge devices?
3. If given large time series from a manufacturing line, how would you start?

**✅ Part 4: Sample Answer Templates (Pick-N-Use)**

**Q: Walk me through a complete ML project you've led.**

*"Sure. One of the major projects I led was chest X-ray pathology detection. We started with dataset curation from multiple hospitals. I led EDA, image preprocessing, and applied CNN-based models. After tuning and validating across 5 pathologies, we hit over 93% accuracy. I deployed the model on a Linux server and integrated it into the clinical workflow, reducing diagnosis time by 80%."*

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

*"First, I identify the nature of missingness: random or systematic. Then depending on use case, I use forward-fill, interpolation, or predictive imputation. If gaps are too wide, I exclude or model separately. I ensure the method preserves the trend and seasonality."*

**Q: What is LSTM and when do you use it?**

*"LSTM is a type of RNN designed to capture long-term dependencies in sequence data. It uses gates to control memory, avoiding vanishing gradients. I used LSTM while working on service demand forecasting and explored it in my medical AI work where sequential signals matter."*

**Q: Have you worked with cloud ML tools like AWS/Azure?**

*"While I haven’t deployed on cloud yet, I’ve deployed models on Linux servers and handled REST APIs. I’m certified in Azure AI-900 and familiar with concepts like blob storage, model endpoints, and pipelines. I'm confident in adapting quickly if given a cloud project."*

Would you like me to now generate a **one-pager cheat sheet PDF** or **roleplay as interviewer and start the mock interview**? Let me know your preferred next step.

**📄 Data Scientist Interview Cheat Sheet - Vivek Choudhari**

**1. Machine Learning Workflow**

* Data Collection → EDA → Feature Engineering → Model Selection → Training → Tuning → Validation → Deployment
* Projects: Chest X-ray, British Gas Churn Prediction, Spine Segmentation, Tender Summarisation (LLM + RAG)

**2. Time Series & ML Algorithms**

* Time Series: Rolling Mean, Differencing, Seasonal Decomposition, LSTM, ARIMA
* ML: SVM, Decision Trees, Random Forest, XGBoost, KMeans, DBSCAN
* Deep Learning: CNNs (Image), LSTM (Sequential), YOLO (Object Detection)

**3. Model Interpretability**

* SHAP & LIME for feature importance
* Choose interpretable models when business stakeholders need clarity

**4. Visualization Tools**

* Python: Matplotlib, Seaborn
* BI Tools: Power BI, Tableau
* Use-cases: Clustering dashboards, Churn prediction reports

**5. Deployment & Infrastructure**

* Linux server deployment (YOLO, CNN)
* Azure AI-900 certified – cloud-ready mindset
* Monitor models post-deployment and retrain on drift

**6. GenAI & NLP**

* LLM + RAG for Tender Summarisation
* Medical report chatbot with OLLAMA Gemma
* Tools: Embeddings, Vector DBs, Streamlit UI

**7. Behavioral Strengths**

* Leadership: Managed full AI model lifecycle, mentored juniors
* Adaptability: Transitioned from energy domain to healthcare AI
* Ownership: Delivered full-stack ML solutions end-to-end

## PART 1: ✅ Must-Have Technical Questions & Answers

### ****1. Walk me through a machine learning project you led.****

**Answer:**

"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%. This was a full-lifecycle project — from idea to deployment."

### ****2. How do you handle missing values in time series data?****

**Answer:**

"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."

### ****3. What is LSTM and how is it used in time series prediction?****

**Answer:**

"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.  
I’ve used it in sequential data settings such as service demand forecasting."

### ****4. When would you use SVM vs Decision Trees?****

**Answer:**

"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.  
I often start with both and use cross-validation to choose based on performance and explainability needs."

### ****5. Explain SHAP and LIME — when would you use each?****

**Answer:**

"Both are model explainability tools.  
SHAP uses game theory to explain global and local predictions, providing consistent feature importance.  
LIME explains individual predictions by fitting local interpretable models.  
I use SHAP when the stakeholder needs feature contribution across the dataset. LIME when I need to explain why a single prediction was made."

### ****6. How do you deploy a model into production?****

**Answer:**

"I package the model using a framework like Flask or FastAPI.  
Then deploy on a Linux server or Docker container, exposing an API endpoint.  
For model monitoring, I track input drift, prediction confidence, and performance metrics.  
In my chest X-ray project, we deployed directly on an Ubuntu machine with real-time inference from DICOM viewers."

### ****7. How do you ensure a model is production-ready?****

**Answer:**

* It must generalize well on unseen data.
* Metrics like F1-score, precision, and AUC must meet business benchmarks.
* I test it for edge cases, latency, and memory usage.
* I also ensure version control, logging, and exception handling are in place.

## PART 2: 📊 Time Series Focused Questions

### ****8. What challenges arise in industrial time series modeling?****

**Answer:**

* Missing or irregular data points
* Sensor drift or noise
* High correlation among variables
* Multivariate dependencies
* Cold-start issues for new equipment  
  I use LSTM, Prophet, and feature lagging with rolling stats to address these.

### ****9. How do you preprocess sensor data for ML modeling?****

**Answer:**

* Convert timestamp to datetime
* Create lag features and rolling statistics (mean, std, min/max)
* Normalize the data if using neural networks
* Handle missing values using forward-fill or interpolation
* Resample if needed (hourly/daily)

### ****10. What metrics do you use to evaluate time series models?****

**Answer:**

* MAE (Mean Absolute Error): Simple and interpretable
* RMSE: Penalizes large errors
* MAPE: Expresses error in percentage (but beware of zeros)
* R²: If regression-like models are used

## PART 3: ☁️ Cloud, Tools & Infra

### ****11. Have you deployed ML models on cloud?****

**Answer:**

"I’ve deployed models on Linux servers using Flask APIs.  
While I haven’t done full production deployment on AWS or Azure yet, I’m Azure AI-900 certified, understand Blob Storage, ML endpoints, and CI/CD basics. I’m confident in adapting to cloud deployment pipelines quickly."

### ****12. What visualization tools do you use?****

**Answer:**

* Python: Matplotlib, Seaborn for EDA and result analysis
* BI Tools: Power BI and Tableau for dashboards
* Used them in churn analysis, customer segmentation, and project reporting to stakeholders

## PART 4: 🤝 Behavioral / HR Questions

### ****13. Tell me about yourself.****

**Answer:**

"I’m a Data Scientist with 4+ years of experience in machine learning and AI, with a strong focus on healthcare and utility sectors.  
I’ve worked on projects involving deep learning, time series, computer vision, and NLP — like detecting diseases from X-rays, automating tender document summarization, and forecasting churn.  
I enjoy solving real-world problems using AI and aim to build solutions that reduce manual effort and improve decision-making."

### ****14. Tell me about a time when a model failed. What did you do?****

**Answer:**

"During a churn prediction project, my model performed well in testing but failed post-deployment.  
The issue was concept drift due to a change in customer behavior during a campaign.  
I handled it by adding recent campaign data, retraining the model weekly, and deploying alerts for prediction confidence."

### ****15. How do you handle pressure or urgent timelines?****

**Answer:**

"I break the task into smaller blocks and prioritize based on business impact.  
I communicate proactively with the team or stakeholders if trade-offs are needed.  
In the spine project, under tight deadlines, I focused only on core vertebrae segmentation first, then expanded to full pipeline."

### ****16. Why should we hire you for this role?****

**Answer:**

"Because I bring a rare combination of deep ML experience and full lifecycle ownership — from EDA to deployment.  
I’ve worked in time-sensitive healthcare settings, handled sensor-like imaging data, and implemented both deep learning and classical ML.  
I am also a fast learner, certified in cloud, and eager to take on industrial ML challenges."

### ****17. Where do you see yourself in 2-3 years?****

**Answer:**

"I see myself leading ML projects end-to-end, mentoring junior data scientists, and contributing to scalable AI systems that solve real industrial problems.  
I aim to deepen my expertise in time series, optimization, and edge deployment."