■ Taxi Fare Prediction – Viva Cheat Sheet

1. Project Overview
The goal of this project is to predict the taxi fare amount based on trip features like pickup/dropoff location, distance, duration, passenger count, and time-based features. I built a regression model that learns the relationship between these features and the fare, using various machine learning algorithms.

2. Steps in ML Workflow
1. Data Understanding – exploring the dataset columns and meaning.
2. Data Cleaning – handling missing values, duplicates, and invalid entries.
3. Feature Engineering – creating features like distance, duration, and time features.
4. EDA – visualizing distributions, correlations, outliers.
5. Outlier Handling – capping extreme values.
6. Skewness Handling – applying log transformations.
7. Model Building – testing Linear, Ridge, Lasso, Random Forest, Gradient Boosting.
8. Evaluation – using R^2, MAE, MSE, RMSE.
9. Model Tuning – optimizing with GridSearchCV.
10. Final Model Selection – choosing the one with best performance.

3. EDA (Exploratory Data Analysis)
I used plots and statistics to understand the data distribution. Total_amount and trip_distance showed positive skewness, meaning some extreme values exist. I also checked correlations and trends like distance vs fare, duration vs fare, and time-of-day vs fare.

4. Feature Engineering
Created new features like:
• trip_distance_km – using Haversine formula
• trip_duration_hr – difference between dropoff and pickup
• pickup_hour, pickup_day – from timestamps

5. Outlier Handling
Outliers can affect model accuracy. Identified using boxplots and IQR method, then capped extreme values instead of removing them to preserve data.

6. Skewness Handling
Applied log transformation to reduce skewness and make data more normally distributed, helping linear models perform better.

7. Model Building
Trained multiple models: Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting. Compared all using evaluation metrics.

8. Model Evaluation Metrics
• R^2: measures how much variance is explained by the model.
• MAE: Mean Absolute Error – average absolute difference.
• MSE: Mean Squared Error – penalizes larger errors.
• RMSE: Root Mean Squared Error – square root of MSE.
Higher R^2 and lower MAE/RMSE indicate better performance.

9. Best Model
Among all models, Gradient Boosting Regressor gave the best performance with high R^2 and low error values, so it was selected as the final model.

10. Key Theory Questions

Q: What is Linear Regression?
A simple algorithm that finds the best-fitting straight line to predict target values by minimizing sum of squared errors.

Q: What are Ridge and Lasso?
Ridge adds L2 regularization (shrinks coefficients). Lasso adds L1 (shrinks & removes features). Both prevent overfitting.

Q: What is Random Forest?
An ensemble model that builds multiple decision trees and averages predictions. It uses Bagging and handles non-linear data.

Q: What is Gradient Boosting?
An ensemble method that builds trees sequentially, where each new tree fixes errors of the previous one.

Q: What is R^2?
Coefficient of Determination – measures how much variance is explained by the model (0 to 1).

Q: Difference between Linear and Random Forest?
Linear: linear relationships, interpretable.
Random Forest: non-linear, more accurate, less interpretable.

11. Conclusion
Successfully built a taxi fare prediction model using Gradient Boosting Regressor, which gave the best accuracy. This project demonstrated skills in EDA, feature engineering, outlier handling, model building, and evaluation.