Use faker library and generate datasets

1) Dataset of apartment rentals with features: square_feet, bedrooms, age_years, and target variable rent_price.

1. Split the data into 80% training and 20% test sets.
2. Train a Linear Regression model
3. Make predictions on the test set
4. Calculate and print R^2, MAE, MSE, and RMSE

P.S. Make sure that you understand the concepts both theoretically and conceptually.

2) Dataset predicting house prices with these features:

square_feet (range: 500 - 5000). lot_size_acres (range: 0.1 - 2.5)

num_bedrooms (range: 1 - 6). distance_to_downtown_km (range: 0.5 - 45)

1. Train a Linear Regression model WITHOUT scaling

2. Print the coefficients for each feature

3. Calculate test R^2

Now with scaling

1.Apply StandardScaler to the features

2. Train a Linear Regression model on scaled data

3. Print the coefficients for each feature

4. Calculate test R^2

3) Real estate company that wants to predict house prices. You have a dataset with features including square footage, number of bedrooms, house age, distance to city center, and neighborhood type (categorical).

Train three different models:

Model A: Simple linear regression using only square footage

Model B: Multiple linear regression using all features

Model C: Polynomial regression (degree 3) on square footage plus other features

For each model, calculate training R^2, test R^2, and RMSE.

3.2 Perform 5-fold cross-validation on three models and get these results:

Model B: fold scores [0.78, 0.80, 0.79, 0.81, 0.77], mean = 0.79, std = 0.015

Model C: fold scores [0.85, 0.83, 0.91, 0.68, 0.88], mean = 0.83, std = 0.085

Lasso: fold scores [0.81, 0.80, 0.82, 0.79, 0.81], mean = 0.81, std = 0.011

Write your answers:

1. Why is cross-validation more reliable than a single train-test split?
2. Model C has the highest mean score but much higher standard deviation. What does this tell you about the model's stability? Which model would you choose for production and why?
3. After selecting your final model using cross-validation, you train it on the full training set and evaluate on a held-out test set. You get test $R^2 = 0.77$. Your cross-validation mean was 0.81. Should you be concerned? Why might these numbers differ?
4. Write the code to perform 5-fold cross-validation and print the scores for each fold.

3.3 Final model has $R^2 = 0.80$ and RMSE = \$45,000. The average house price in your dataset is \$350,000.

1. Explain what $R^2 = 0.80$ means in practical terms that a non-technical real estate manager would understand. Is this model accurate enough for making business decisions? What additional information would you need to determine this?
2. The RMSE is \$45,000. Explain what this metric means and whether you think this level of error is acceptable for predicting house prices
3. After deploying your model for 6 months, you notice the RMSE has increased from \$45,000 to \$65,000. What are three possible reasons why a model's performance degrades over time.

Quiz is mainly focused on concepts and theory, but you'll also need some estimation and calculations. Reviewing each slide in details on ML, Regression, and Classification would be the ideal way to study.