Difference between CatBoost and XGBoost.

Feature	CatBoost	XGBoost
Algorithm	Uses ordered boosting to reduce overfitting	Uses gradient boosting with decision trees
Handling Categorical Features	Handles categorical variables natively without encoding	Requires one-hot or label encoding for categorical features
Training Speed	Faster when handling categorical data	Slower with categorical features due to encoding
Overfitting Prevention	Uses ordered boosting to prevent target leakage	Uses regularization (L1, L2) to reduce overfitting
Performance on Small Datasets	Performs well due to efficient feature handling	Requires careful hyperparameter tuning for small datasets

What is the reason for the model to predict the particular output(Yes or No).

A model predicts "Yes" or "No" based on the learned patterns in the dataset. It considers:

- **Feature Importance**: Some features have more influence on predictions.
- **Decision Boundaries**: It calculates probabilities using learned weights.
- **Thresholding**: If the probability is above a threshold (e.g., 0.5), it predicts "Yes"; otherwise, "No".
- **Feature Contributions**: Using SHAP or LIME, we can analyze how much each feature contributes to the final decision.

Analayse the column(Monthly Rate, Monthly Income, Hourly Rate, Daily Income), to understand the column.

The issue occurs because:

- **Irregular Work Hours**: Employees may work variable hours, so hourly and daily earnings don’t always sum up logically.
- **Missing Data or Errors**: Possible data entry mistakes or missing records.

How to Overcome This Problem?

- **Check Data Consistency**: Ensure correct data entry and calculations.
- **Feature Engineering**: Create derived features like "Effective Hourly Rate" (Monthly Income / Hours Worked).
- **Normalization**: Scale the data to reduce inconsistencies.

What is correlation? What did you find in correlation mapping?

Correlation measures the relationship between two numerical variables.

Values range from **-1 (strong negative correlation)** to **1 (strong positive correlation)**.

In correlation mapping:

- A strong correlation suggests dependent relationships (e.g., high "Monthly Income" correlates with "Monthly Rate").
- No correlation suggests independent variables.

What does value count means and how does it work.

value_counts() is a Pandas method used to count occurrences of unique values in a column.

What method did you use to remove the columns from the dataset.

value_counts() can help identify columns with a single dominant value, which may not be useful for modeling.

What are outliers. How do you handle outliers.

Outliers are extreme values that deviate significantly from most data points.

Handling Outliers:

- **Z-score method**: Remove values beyond 3 standard deviations.
- **IQR (Interquartile Range) method**: Remove values outside 1.5 * IQR.
- **Transformation**: Use log or Box-Cox transformations to reduce skewness.
- **Capping**: Replace extreme values with the upper/lower percentile.

What is standardisation? Why we need it.

Standardization scales data to have a mean of 0 and a standard deviation of 1.

Why Needed?

- Ensures all features contribute equally to model performance.
- Improves convergence speed in gradient-based optimization (e.g., logistic regression, neural networks).

What is one hot encoding. What is label encoding. When we need to use each one.

One-Hot Encoding (OHE): Converts categorical variables into multiple binary columns.

Label Encoding: Assigns unique numerical values to each category.

What is oversampling.

Oversampling is increasing the number of samples in the minority class to balance the dataset.

Example: If Class "No" has 500 samples and "Yes" has 100, oversampling increases "Yes" to 500.

Methods: Random Oversampling, SMOTE.

What is undersampling.

Undersampling reduces the number of majority class samples to balance the dataset.

Example: If "No" has 500 samples and "Yes" has 100, undersampling reduces "No" to 100.

Risk: Loss of valuable information.

Difference between oversamplig and undersampling.

Aspect	Oversampling	Undersampling
Method	Adds synthetic/minority samples	Removes majority samples
Goal	Balance the dataset by increasing minority samples	Balance by reducing majority samples
Risk	Overfitting	Loss of information

What is SMOTE?

Synthetic Minority Over-sampling Technique (SMOTE) is an advanced oversampling method.

Instead of duplicating samples, it creates **synthetic samples** by interpolating between existing minority class samples.

What is Precision, Recall. Difference between them.

Metric	Formula	Meaning	When to Focus on it?
Precision	TP / (TP + FP)	Measures how many positive predictions were correct	When false positives are costly (e.g., fraud detection)
Recall	TP / (TP + FN)	Measures how many actual positives were correctly identified	When false negatives are costly (e.g., medical diagnosis)