Project Title:

Predicting Construction Project Costs and Profitability

Objective:

To build a predictive model that estimates **actual cost**, **profit**, or **duration** of construction projects based on initial project details and visualize key business insights.

Dataset:

Use or generate a dataset with the following columns:

Column Name	Description
Project_ID	Unique identifier for each project
Project_Type	Type of construction (Residential, etc.)
Start_Date	Project start date
Duration	Estimated project duration (in days)
Estimated_Cost	Estimated budget for the project
Actual_Cost	Final cost after project completion
Labor_Hours	Total hours of labor involved
Materials_Cost	Cost of materials used
Profit	Estimated_Cost - Actual_Cost
Status	Project status (Planned, Ongoing, Done)

Steps in the Project:

1. Data Collection

- Use the generated or custom Excel/CSV dataset.

```
python
Copy code
import pandas as pd
df = pd.read_csv('construction_data.csv')  # or .xlsx
```

---

2. Data Preprocessing

- Handle missing values
- Convert dates and encode categorical features
- Create new features if needed (e.g., `End_Date`, `Cost_Overrun`)

```
python
Copy code
df['Start_Date'] = pd.to_datetime(df['Start_Date'])
df['Project_Type'] = df['Project_Type'].astype('category').cat.codes
df['Status'] = df['Status'].astype('category').cat.codes
```

---

3. Exploratory Data Analysis (EDA)

Visualizations:

```
python
Copy code
import matplotlib.pyplot as plt
import seaborn as sns

# Cost distribution
sns.histplot(df['Estimated_Cost'])
plt.title("Estimated Cost Distribution")

# Profit by project type
sns.boxplot(x='Project_Type', y='Profit', data=df)
```

Correlation Heatmap:

```
python
Copy code
corr = df.corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
```

---

4. Feature Engineering

- Create derived features:

```
python
Copy code
df['Cost_Overrun'] = df['Actual_Cost'] - df['Estimated_Cost']
df['Profit_Margin'] = df['Profit'] / df['Estimated_Cost']
```

---

5. Model Building

Predicting `Actual_Cost` or `Profit`

```
python
Copy code
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor

X = df[['Project_Type', 'Duration', 'Estimated_Cost', 'Labor_Hours', 'Materials_Cost']]
y = df['Actual_Cost']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor()
model.fit(X_train, y_train)
```

---

6. Model Evaluation

```
python
Copy code
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

y_pred = model.predict(X_test)
print("MAE:", mean_absolute_error(y_test, y_pred))
print("RMSE:", mean_squared_error(y_test, y_pred, squared=False))
print("R2 Score:", r2_score(y_test, y_pred))
```

---

7. Prediction on New Data

```
python
Copy code
new_data = [[1, 200, 1200000, 10000, 600000]]  # example values
predicted_cost = model.predict(new_data)
print("Predicted Actual Cost:", predicted_cost)
```

---

8. Visualizing Predictions vs Actual

```
python
Copy code
plt.figure(figsize=(10,6))
plt.plot(y_test.values, label='Actual')
plt.plot(y_pred, label='Predicted')
plt.legend()
plt.title("Actual vs Predicted Project Cost")
plt.show()
```

✓ Conclusions from the Analysis

1. Estimated Cost Distribution

- The cost histogram suggests how project budgets are distributed — possibly identifying common budget ranges or outliers.

2. Profit Trends by Project Type

- The boxplot likely reveals which project types tend to yield higher or more consistent profits.

3. Correlations

- A heatmap gives insights into relationships:
  - **Strong positive correlation** (e.g., between `Estimated Cost` and `Actual Cost`) suggests that as expected costs rise, actual costs do too.
  - **Profit** and **Profit Margin** may show positive correlation with good cost control and efficient project types.

4. Cost Overrun

- The `Cost_Overrun` feature allows analysis of how often and by how much projects exceed their estimated budgets.

5. Profit Margin Analysis

- By normalizing profit with estimated cost, the `Profit_Margin` column highlights efficiency — how much profit is earned per dollar spent.

🔮 Predictions Made

1. Actual Cost Prediction

- A **Random Forest Regressor** was trained to predict `Actual Cost` based on:
  - Project Type
  - Duration (days)
  - Estimated Cost
  - Labour Hours
  - Material Cost

This model can help project managers **estimate final costs** given initial planning data — which is essential for budgeting and controlling overruns.