🏦 Machine Learning Bankruptcy Predictions

Author: Riddhi Maiti
Objective: Predict whether a company is likely to go bankrupt using financial indicators, aiding risk assessment and early warning systems.

---

1. 📂 Data Overview

- **Source**: `data.csv`
- **Target Variable**: `Bankrupt?` (Binary: 1 for bankruptcy, 0 otherwise)
- **No Missing Values**: Verified using `df.info()`
- **Features**: Various financial ratios and metrics (e.g., ROA, net income to total assets, debt ratio)

---

2. 📊 Exploratory Data Analysis (EDA)

Class Distribution:

- Imbalanced dataset, with fewer instances of bankrupt companies
- Visualized using `sns.countplot`

Feature Distributions:

- Used `df.hist()` for full feature visualization
- Scatter plots analyzed bankruptcy relationship with:
  - **ROA (A)**: Before interest and after tax
  - **ROA (B)**: Before interest and depreciation after tax
  - **Net Income to Total Assets**
  - **Debt Ratio (%)**

Correlation Matrix:

- Large heatmap visualizes correlation between all features
- Used to assess which variables relate most strongly with bankruptcy

---

3. 🧪 Dimensionality Reduction (Correlation-based)

- Selected features with absolute correlation > 0.2 with the target
- Created a filtered dataset `df_filt` with these high-impact features

- Visualized correlation matrix among selected variables

Top Correlated Features:

- `Net Income to Total Assets`
- `ROA(A)`
- `ROA(B)`
- `Debt Ratio %`
- Others based on correlation threshold

---

4. 🧼 Data Preprocessing

- **Train-Test Split**: 70/30 using `train_test_split`
- **Feature Scaling**: Applied `StandardScaler` to normalize the feature space

---

5. 🤖 Machine Learning Models

Models Used:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Support Vector Classifier (SVC)
- Decision Tree Classifier
- Random Forest Classifier

Evaluation Metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **Confusion Matrix**

These metrics help assess model performance, especially with class imbalance.

---

6. 📌 Key Observations

- The dataset is **highly imbalanced**, which may affect model performance
- A few financial indicators (like ROA and debt ratios) have **strong predictive power**
- Dimensionality reduction based on correlation helped simplify the problem
- Standardization was critical due to varying scales of financial ratios
- Performance evaluation likely needed **class balancing** or **threshold tuning**

✅ Summary

This notebook outlines a well-structured pipeline to predict company bankruptcy using machine learning, focusing on:

- Careful EDA
- Correlation-driven feature selection
- Preprocessing and scaling
- Benchmarking multiple ML models

It provides a strong foundation for implementing early warning systems in financial risk management.

📊 Credit Risk Assessment

Author: Riddhi Maiti
Objective: Build a machine learning model to predict the likelihood of loan default, enabling financial institutions to make informed lending decisions.

---

1. 🗂️ Data Overview

Dataset:

- File: `credit_risk_dataset.csv`
- Goal: Classify loans as default (`loan_status = 1`) or non-default (`loan_status = 0`)

Key Features:

- `person_age`: Age of the individual
- `person_emp_length`: Employment length
- `person_income`: Annual income
- `loan_int_rate`: Loan interest rate
- `loan_status`: Target variable (default/no default)
- `loan_grade`, `loan_intent`, `cb_person_cred_hist_length`: Other contextual features

---

2. 🔍 Exploratory Data Analysis (EDA)

Distribution & Relationships:

- **Pairplot** (`sns.pairplot`): Visualized feature interactions and separation by `loan_status`
- **Correlation Heatmap**: Found strongest relationships (e.g., `person_emp_length` vs `person_age`)
- **Loan Intent Analysis**: Visualized default rates across different loan intents (e.g., EDUCATION, PERSONAL)
- **Credit History Distribution**: Most users had credit histories shorter than a few months
- **Class Imbalance**: Around **25%** of loans were defaults, indicating moderate imbalance

---

3. 🧼 Data Preprocessing

Missing Value Handling:

- **`person_emp_length`**:
  - Imputed using a **linear regression approach** based on correlation with `person_age`
  - Formula used: `person_emp_length ≈ 0.163 * person_age`
- **`loan_int_rate`**:
  - Missing values filled with **mean interest rate per loan grade**
  - Grade-based averages used:
    - A: 7.33, B: 11.00, C: 13.46, D: 15.36, E: 17.01, F: 18.61, G: 20.25

Post-Imputation:

- Confirmed no missing values remained

---

4. 🏗️ Feature Engineering & Modeling

Anticipated Steps:

- Encoding categorical features
- Feature scaling
- Train-test split
- Classification model training (e.g., Logistic Regression, Random Forest)
- Evaluation (accuracy, confusion matrix, ROC AUC, etc.)
- Feature importance analysis

---

5. 📌 Key Observations

- `loan_status` distribution is imbalanced → may need techniques like SMOTE or class weights
- Employment length is moderately correlated with age → useful for imputation
- Interest rates vary significantly across loan grades → meaningful feature
- Visualizations are well-structured to understand feature relationships

---

✅ Summary

This notebook demonstrates a robust approach to preparing data for credit risk modeling, including effective visualization, missing value imputation using domain-aware heuristics, and careful EDA. Once modeling and evaluation are added, this pipeline will offer valuable insights into risk prediction.