Assignment 3: Feature Engineering on Financial Statements and Financial Ratios

Overview

Weight: 5%

Learning objective: Develop an understanding of financial statements and financial ratios

The goals of this assignment are for the students to

- Learn about corporate income statements, cash flow statements, and balance sheets.
- Understand important financial ratios based on key variables.
- Perform feature engineering using financial ratios for corporate finance ML tasks.
- Please note that the most important part of the assignment is not the programming or data analysis. The primary learning should be to identify patterns in your analysis and draw meaningful conclusions from those patterns.

Required Files for Submission

- One jupyter notebook that includes all the data wrangling and computations
- A report with all the plots and analyses
- Note: Please do **not** submit data

Data

In order to reduce the time demands for the assignment, we have uploaded the required dataset to OneDrive: Link

- You need to use data from 1996 to 2022
- “gvkey” is unique identifier for a company in COMPUSTAT data
- “fyear” is financial year of the company
- Description of all the variable is provided in *COMPUSTAT_Variable_Definition.pdf*

Assignment

- See a basic corporate finance textbook (for example, Berk and Demarzo), if you would like some background on corporate income statements, cash flow statements, and balance sheets. These would be useful for this and future assignments, including those on trading strategies.
- Optional Step: You don’t need to read it, but in case you want more background, check Frank and Goyal (2003) paper in Assignments (only Tables 1,2, 8, 9, and 10 of the paper). Note that the variable names in the paper are in OLD format. The variable names in the EXCEL spreadsheet and dataset are current. Translation table posted has the variable translations.
- **Qualitative Feature Selection** From all the variables in the *Assignment_FinRatios_Tables.xlsx* (Table 1, 2, 8, 9, and 10) file
    - Make a list of 20-25 variables that you believe would be most valuable for a corporate finance prediction task (e.g., bankruptcy prediction).
        - from each group that is included in the assignment pick at least a few variables. For example
            - income-sheet
            - balance-sheet
            - cash-flow statement
            - financial ratios: liquidity, leverage, etc.,
    - for each selected variable, write a 1-2 line explanation on why you think it will be useful
- **Compute** all the ratios that you have selected
- **Quantitative Feature Selection**
    - Calculate the correlation matrix for the variables you selected.
        - would you group them differently?
        - Based on the analyses what variables you would think of dropping? Provide reasoning

  - Get NBER recession data. Compute the correlation matrix for NBER recession = 1 and NBER recession = 0. What do you notice?

- **Faceted Histograms**: Create separate histograms for each year and place them side by side using a facet grid.
  - What do you surmise from the distribution indicates about the evolution of public firm's financials over time (across histograms) and in the cross-section (within histogram)?
  - Do this analysis for all selected features

- **Feature Scaling**: Based on your observation about the distribution of each feature, think about which features need feature scaling.
  - standard feature scaling methods are:
    - Normalization: Scale features to a range (e.g., 0 to 1)
    - Standardization: Transform features to have zero mean and unit variance
    - Log Transformations: Apply log transformations to skewed features to reduce the impact of extreme values

- **Principal Component Analysis (PCA)**
  - Perform PCA on the selected variables for the last two years of data
  - How many principal components are enough to cover 90% of the variance in the data

Python Functions and Libraries that might be useful for the Assignment

This section provides an overview of important Python functions and libraries that will assist you in completing various tasks of the assignment.

1. Feature Scaling

- **sklearn.preprocessing**: For feature scaling (normalization, standardization, etc.).
  - `StandardScaler()` – Standardize features (zero mean, unit variance).
  - `MinMaxScaler()` – Normalize features to a range (0, 1).

```python
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df[['feature1', 'feature2']])
```

2. Correlation Matrix

- **pandas.corr()**: Use this to calculate the correlation matrix between selected financial ratios.
- **matplotlib** or **seaborn**: Visualize the correlation matrix.
  - `heatmap()` (from seaborn) – Create heatmaps for visualizing the correlation matrix.

```
import seaborn as sns
import matplotlib.pyplot as plt
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True)
```

3. Histograms and Faceted Grid Plots

- **seaborn.FacetGrid**: Plot faceted histograms for financial ratios over different years to explore trends.
- **matplotlib.pyplot.hist()**: Use this function to create individual histograms for each feature.

```
import seaborn as sns
facet = sns.FacetGrid(df, col="fyear")
facet.map(plt.hist, "financial_ratio", bins=20)
```

4. Principal Component Analysis (PCA)

- **sklearn.decomposition.PCA**: Perform PCA to reduce dimensionality and identify principal components.
  - `PCA().fit()` – Fit the PCA model on selected features.
  - `explained_variance_ratio_` – Determine how many components are needed to cover 90% of the variance.

```
from sklearn.decomposition import PCA
pca = PCA(n_components=0.90)  # Keep 90% variance
pca.fit(df_scaled)
explained_variance = pca.explained_variance_ratio_
```