**Data Description and Pre-Processing**

**1. Overview of the Dataset**

The data sampled for this investigation include financial and macroeconomic information obtained from sources such as the company’s annual reports, other information from the stock market, and other macroeconomic factors. The emphasis is placed on a luxury segment at the company, specifically Hermès, and the industry level, that of the fashion business for 2019-2023. Included are details of the company and the firm’s account, changes in shares' prices, and details of other factors significant to the luxury fashion industry (Altman, 2018). These are the fundamentals of establishing an ML model for predicting financial risks in the current sector.

This data is used to measure the financial susceptibility of luxurious fashion companies and verify the relevance of certain financial ratios. As metrics data, one has revenue, net income, total assets, debt-to-equity ratio, non-metric variables, inflation rate, and GDP (Damodaran, 2020). It is convenient to use various financial indicators to receive as much information as possible about the financial position of these enterprises and increase the model limit.

**2. Financial Indicators and Their Importance**

Some commonly utilized model variables for predicting a firm’s financial distress consist of several financial ratios and KPIs included in the dataset. Some of them are profitability ratios like return on assets and return on equity or net worth, liquidities ratios like current and quick ratios, leverage ratios like debtors to equity, interest coverage, and market performance ratios like price earning ratio and stock price volatility. These have been used in other prior studies to assess the degree of the firm’s financial risk and to identify its ability to forecast bankruptcy.

As highlighted by Altman (2018) and Zopounidis and Doumpos (2021), there is a shared focus on financial ratios in the appraisal of the financial stability of the business corporations. Working capital, retained earnings, and EBIT are the three primary parameters that still make up the framework for the Z-score model that predicts financial distress. Thus, based on incorporating these traditional financial ratios with advanced machine learning methods, this research shall seek to improve the overall risk forecasting in the luxury fashion industry.

**3. Data Collection and Sources**

For this reason, data was collected from different sources to enhance its reliability and accuracy. By analyzing the annual financial statements of Hermès and Prada, company-specific financial information was gathered, and stock market information was retrieved from Yahoo and Bloomberg. Macroeconomic data was obtained from the world bank and international monetary fund sources that offered a clear insight into the external environment for analysis of the financial performance (Chen, Liu & Wang, 2021).

In my case, data collection was also a challenge, mainly due to various sources' disagreement in financial reporting format. Even if using IFRS and GAAP as a basis for accounting makes the process regulated and comparable, the values may differ from company to company or country regarding the metrics presented. To cater to this problem, various data normalization methods were adopted to ensure that the monetary values adopted a normalized form (James et al., 2021). Further, data gaps were also imputed statistically by employing mean imputation and K-Nearest Neighbors (KNN).

**4. Pre-Processing Techniques**

Several cleaning activities are sometimes done to the data. They include: Essentially, raw financial data from various sources had noise, missing observations, and did not follow consistent structures, rendering it necessary to clean the data. The first one addressed the issue of missing values by performing multiple imputations. Regarding numerical data, the missing values were imputed using mean or median depending on the distribution. In contrast, the missing values were imputed for categorical data using mode substitution, as Friedman (2001) stated.

Normalization and feature scaling were done on the training data to standardize financial ratios and stock prices. Mean and standard deviation normalizations were avoided, so the Min-Max normalization method was applied to normalize the data because financial ratios based on the financials and stock price fluctuate on different scales (revenue is in million USD, whereas stock price is in double value). This step is crucial in the case of machine learning models, especially distance-based algorithms like k-means clustering and support vector machines (Bishop, 2006). As a result of normalizing the data, the models achieved better efficiency and interpretability of results.

**5. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was done to understand the financial dataset better and check for aberrations, correlations, or trends. Fundamental analysis of mean values showed that these two business entities are not analogously positioned: Hermès demonstrated particular profitability and reasonable debt indicators, while Prada displayed poor financial performance figures. Standard tools such as historical plots, box and whisker plots, and correlation heat maps were used to establish relationships between financial factors (Lundberg & Lee, 2017).

As previous research shows, correlation analysis confirmed a high relationship between leverage ratios and financial distress (Altman, 2018). For example, a high debt-to-equity ratio was constantly correlated with financial fragility, indicating the importance of the feature when using machine learning algorithms. Thus, choosing only a subset of features was accomplished by using feature selection methods like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), which increases the efficiency of the models (Breiman, 2001).

**6. Data Challenges and Solutions**

As for data preprocessing issues, class imbalance was the main problem in constructing the financial distress classification model. In the following sections, some forms of distress were uncommon in luxury fashion companies because those companies tend to be financially sound. To address this problem, measures have been taken using Artificial data generation practices, for example, SMOTE, which comes in handy to help overcome issues to do with class imbalance (Chen, Liu & Wang, 2021).

One of the difficulties was addressing cases where some financial indicators were outliers, and their presence distorted performance estimation. Elimination was also conducted with the help of the IQR and z-score techniques used in detecting outliers. Furthermore, the skewness transformation was done on certain variables by choosing the appropriate logarithmic transformation to rectify skewed distributions and enhance the data and model efficiency. (Friedman, 2001).

**7. Feature Engineering and Selection**

Feature selection was one of the strategies that boosted the model's performance to greater efficacy. Some new composite indices developed through the failure-prediction formula include, for example, the Altman Z-score, which is made of ratios able to measure bankruptcy probability. In addition, new features defined after some period were constructed using moving average and volatility indexes (Zopounidis & Doumpos, 2021).

The authors carried out the selection feature based on statistical measures and other methods to apply machine learning algorithms. About these features, chi-square tests were used to identify the chi-square value for independence of categorical data with the class. In contrast, for numerical datasets, mutual information was used to determine the mutual information between the features and the class. RFE and GBDT were also applied in this research, as noted by James et al. (2021).

**8. Data Partitioning and Model Preparation**

Data stratification was performed to eliminate any outliers in the dataset, where 8% were for training, 1% were for validation, and the remaining 1% were for the test data set. The sampling technique used during the workshop’s procurement was stratified sampling to get an appropriate respondent in the case of financial distress. Another method used in the study includes k-fold cross-validation, which helps make the model more general. (Breiman, 2001).

Grid search and random search methods were used to determine the optimal values of the learning rate, depth of the trees, and regulation coefficients to select the best parameters for the ML models. This last dataset was then translated into a model-friendly format to allow easy direct implementation of the models (Bishop, 2006).

**9. Final Data Readiness and Validation**

Sampling results were cleaned and processed, and validity was established by using statistical checks and graphical exploratory data analysis. It was also confirmed that the skewness of quantitative variables was insignificant, and equal variances could be assumed based on Levene’s test. The final dataset collected was then formatted into structured forms in preparation for the model and the predictive analysis (See Friedman, 2001).

This is proposed to be achieved through complex preprocessing approaches with an appropriate selection of methodologies for assessing the financial risks for luxury fashion firms. Consequently, the structured data preparation can be considered as providing additional safety to the models and helps understand the results and other interpretations of the models. (Lundberg & Lee, 2017).