**Executive Summary:**

The objective of this report is to present our research project titled "An Empirical Examination of Board Types in Companies Based in the United States." The primary aim of this project is to investigate the structural implications of different board categories within companies and develop predictive models to classify them into specific categories, namely "ARISTOCRATIC," "DEMOCRATIC," "MONARCHY," "OLIGARCHY," "SOCIALIST," "STANDARD," and "TOTALITARIAN." This comprehensive report will encompass diverse facets of our study, including literature reviews, research methodologies, data processing, descriptive analysis, modeling techniques, and more.

To accomplish our project goal, our research will encompass four key phases:

Phase 1: We will establish a set of assumptions to overcome any limitations present in the original data sets.

Phase 2: Data Processing and Feature Engineering Given that the data provided by our sponsor originates from various sources and is structured at different levels (company and director levels), thorough data processing is essential prior to analysis.

Phase 3: Descriptive Analysis Once Phase 2 is complete, we will conduct statistical analyses and employ visualizations to compare and showcase the distributions of various board categories, including "ARISTOCRATIC," "DEMOCRATIC," "MONARCHY," "OLIGARCHY," "SOCIALIST," "STANDARD," and "TOTALITARIAN."

Phase 4: Modeling Analysis In this final phase, we will develop predictive models to address the primary objective outlined by our sponsor.

During the early stage of analysis, a decision was made to home in on all companies, since historical data regarding directors' influence is exclusively available for this group. Consequently, the EDA segment portrays an analysis strictly related to companies to define board categories. To facilitate this, we procured a spreadsheet encompassing the catalog of all companies listed by Free Float Media, which was subsequently utilized to categorize companies within the primary document, retaining only the relevant ones. This categorization was accomplished by merging files using the Inner Join feature, centered on the companies' ticker symbols. The revelations from this research will offer meaningful perspectives to organizations, shareholders, and other stakeholders, elucidating the effect of directors' identities on the efficiency of the board, strategic decision making, risk mitigation, and the comprehensive performance of the corporation.

Subsequently, we employed visualization techniques to better comprehend the critical features in the primary document related to the board of directors, thereby gaining deeper insights. Figure 1 showcases a dashboard report that delivers a comprehensive view of the companies and directors. The number of companies that remain post-data transformation primarily belong to the financial, industrial, and IT sectors. The prevalent form of ownership is principal shareholder, trailed by widely held, and then controlled ownership.

While the formats of the provided data sets are largely analogous—each director or company possesses a single record in the original data, which includes statistics on the board of directors, performance details, and so forth—the data concerning influence and interconnections amongst directors follows distinct structures. Initially, we restructure the data pertaining to the highest influence percentage by transposing the values from separate rows for individual directors into distinct columns. The feature engineering phase of our study is threefold: 1) crafting the target variable (categorizing the Board of Directors based on directors' statistics); 2) pinpointing directors' concurrent directorships spanning multiple companies; 3) recognizing the historical and current roles (CEO, lead director, and chairman) held by individual directors across various companies.

In response to our sponsor's request for the most accurate model to predict potential departures due to fraudulent activities, we are tasked with constructing multiple models. Rather than manually assembling each model, we opted to employ the Auto ML technique, which automates the process by swiftly constructing various ML models. In this scenario, we utilized numpy and scikit libraries for the classification problem, facilitating the creation of multiple models and proceeding with the most efficient one, showcasing the optimal parameters applicable to the data. This not only accelerates the process but also enhances performance accuracy by mitigating potential human errors or biases, thereby concentrating on the actual problem at hand. Moreover, it incorporates a setup function that permits the setting and modification of a wide range of parameters as needed, such as eradicating multicollinearity, normalizing, imputing, encoding, rectifying class imbalance, and even stipulating the method.

In conclusion, among all evaluated models, the Random Forest exhibits superior performance. We propose the utilization of Random Forest, in conjunction with Gradient Boosting, for this data set. The latter method demonstrates an overall superior accuracy rate. Particularly in logistic regression, the proportional representation of insider influence bears a high coefficient. This secondary model imparts a unique perspective on the data. Ultimately, despite the numerous parallels drawn with the initial approach, it introduces an alternative thought process and data analysis strategy that corroborates our model construction and investigative efforts.

To further improve the predictive model, the following aspects should be considered:

**Addressing Imbalanced Distributions:** Overcoming the imbalance in independent variables' distributions.

**Variable Control:** Controlling variables related to sectors, capital levels, and other relevant factors. In our data manipulation process, we've confined our application to directors associated with companies incorporated in the US to prevent data attrition and potential discrepancies in board management strategies across diverse nations.

**Combination of Processing Strategies:** Exploring the combination of two different strategies for processing IPS (Independent Variable Set) in various structures.

**Refining Assumptions and Target-Identifying Algorithm:** Modifying assumptions and the algorithm used to identify the target, incorporating additional and more detailed data sources.