**Analyzing Company Financials to Classify Executive Titles**

Kevin W.S. Baum and Stephen F. Reagin

Shiley-Marcos School of Engineering, University of San Diego

**Section 1:**

Top executives at publicly-traded companies are among the most highly compensated workers in American society, with total compensation packages often reaching 7 figures or more on an annual basis. The subject of executive pay is also contentious in popular media, where it is often addressed in the context of pay or wealth inequality.

Calls for greater transparency of executive pay have led to disclosure requirements, including both the amounts and performance conditions tied to the pay packages for top executive officers. Publicly-traded companies must disclose, on an annual basis, the amounts paid to their top executives and the justifications for doing so. Indeed, as recently as August 2022 the SEC has adopted new rules requiring disclosure of the connection between company performance and executive pay (United States SEC, 2022).

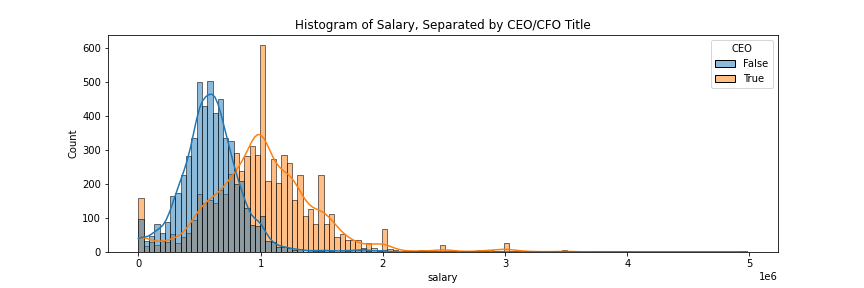
Looking through multiple such *proxy statements*, the documents in which this legally-required information is disclosed, one of the most common justifications for large executive pay packages is that higher pay should be tied to higher performance. However, performance can be defined differently for semiconductor stalwarts like Intel and AMD versus beverage companies like Keurig Dr Pepper and Coca Cola. But generally speaking, the most commonly agreed-upon metrics for overall company performance tend to focus on revenue and income growth.

A review of the academic and business literature in this space reveals common research themes: (A) The growth of executive pay over time, (Bebchuk & Grinstein, 2005), (B) The effect of increased transparency (Bruce & Skovoroda, 2015), and (C) Consultancy services advising companies on how to tie executive pay to company performance (Abbott et al., 2019).

We would like to investigate the relationship, if any, between executive pay and company performance using a data-driven approach. However, it’s not clear at the outset that the techniques of machine learning and statistical mining are the appropriate tools for such an investigation. Therefore, the subject of our report is to investigate the feasibility of using mathematical tools to answer such questions. In particular, we ask a simple question: “can machines tell the difference between CEOs and CFOs on the basis of their pay packages?”

**Figure 1**

*Histogram of Salary, Separated by CEO/CFO Title*



This phenomenon of CEOs earning more than CFOs tends to be true at most public

companies with rare exceptions, as evidenced by the histogram in Figure 1. If we have a list of pay packages given to CEOs and CFOs only for the S&P 500, the classification algorithm is relatively simple for a human observer:

1. Look at the two people listed for every company
2. See which person is paid more
3. The person with higher pay is almost always the CEO

**Section 2:**

With the course’s focus on classification and the encouragement in the instructions to do a project involving classification, we decided to create a classification model that would predict a CEO (Chief Executive Officer) versus a CFO (Chief Financial Officer). This is not a dataset that we found on Kaggle or a site that specifically curated it for the purposes of using and discovering classification techniques. Rather, it is a completely unedited data scrub using SEC data. The process to discover our question was iterative, and we relied on exploratory data analysis and combining our intuitions in order to carve our path. Our overall goal was simple: demonstrate knowledge gained in ADS502 by using key learnings from the course to apply to “wild data.”

With this in mind, we started with two datasets, one for executive pay, and the other for company financial data, extracting 10 years of quarterly trailing 12 month data from current S&P companies into a Pandas data frame. We then grouped the data into average values for every year based on its ticker. For example, Apple will have average values specific to its financials for each year over the last 10 years. Next, we moved to the executive pay dataset, and extracted the executive information for those individuals within the S and P 500 ticker list over the last 10 years. After this, we combined the two separate data frames into one by joining them via their tickers. Now that we had our main data frame ready, we performed cleaning, pre-processing, and exploratory data analysis (EDA) before sub-setting the data frames by CEO and CFO. Once this was done, we concatenated them together, essentially pasting one on top of the other. Once this data frame was ready, we partitioned the data.Next, we next scaled the data because of the large numbers involved when summing over ten years. We proceeded to run CART, C5.0, random forest, logistic regression, naïve bayes, K Neighbors, and support vector machine models and scored them with evaluation metrics.

**Section 3:**

The ultimate source of all executive compensation information is the United States Securities and Exchange Commission (SEC). Every publicly-traded company must file an annual proxy statement with the SEC for shareholders to review, among other things, the total compensation packages paid to the CEO, CFO, and next three most highly-paid officers, a group is often referred to as the “Top 5 executives.” Of particular importance is the “Summary Compensation Table” which is also required by the SEC, and for all Top 5 officers the company must disclose three years of:

• Salary

• Performance bonus (“Non-Equity Incentive Compensation”)

• Stock and Option awards

• Off-cycle or one-time bonuses

• Other services like personal security, air travel, etc.

However, this data is inconvenient to source directly from .html files on the SEC because of the manual labor involved. We instead pulled data from sec-api.io, a third-party source, using API calls under a paid academic license. This raw dataset includes more than 170,000 observations of Summary Compensation Table data rows for companies in the Russell 3000, dating from 2006 to 2022. We also feature engineered several boolean title columns (CEO, CFO, Interim) to flag specific roles. We further narrowed our universe to include only the 485 members of the S&P 500 for which we have approximately 28,000 reliable data observations.

We sourced company financial information from stockrow.com using their free Excel-based worksheets of Income Statement information, specifically targeting Revenue and Net Income.

**Table 1**

*Descriptive Statistics of Raw Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Observations | Median | Mean |
| salary | 28440 | $618K | 708 |
| bonus | 28440 | $0 | 247 |
| stockAwards | 28440 | $1,425K | 2788 |
| optionAwards | 28440 | $93K | 5133 |
| nonEquityIncentiveCompensation | 28440 | $612K | 1054 |
| otherCompensation | 28440 | $61K | 370 |
| total | 28440 | $4083K | 6460 |
| Revenue | 28391 | $8.79B | 22.0B |
| Net Income Common | 28391 | $787M | 1.99B |

**Section 4:**

The first step was to get the data we needed from the SEC API for executive pay, and from Stockrow.com for the company financials. After gathering the data from both sources, we decided we wanted to use only S&P 500 data and so we also loaded a data source containing the 503 tickers. We matched the data from the company financials to only return data for each ticker in the list of 503. For each ticker in the list, we created CSV files and then combined all the CSV files into a Pandas data frame to perform manipulations on. Now that our data frame was ready, we narrowed the fields into just ones we wanted: Ticker, Date, Revenue, Gross Profit, Operating Income, Income after Tax, and Net Income Common. Additionally, we created a datetime object from the “date” feature and extracted the year out of it. This let us ultimately break down the information by yearly basis so that we could establish yearly trends.

Next, we grouped the data by ticker (representative of company) and year and got the mean values for each ticker per year. To make sure this was done correctly, we manually checked “AAPL,” Apple, and found that the numbers matched when we did an independent search online. From here, we were ready to join in the executive pay data set and did so based on the ticker; only executive pay information for the S&P 500 companies were included. Note that not all companies were in the executive pay dataset, and after merging them together only 485 companies remained with 18 not present in the executive pay from the SEC API.

For data cleaning, we noted some duplicate records and performed a standard drop\_duplicates() method to eliminate these duplicates. The company Agilent had duplicate records that were duplicates in reality, but had only one slight difference that caused them to get past the drop\_duplicates method. We eliminated these duplicates by deleting the column that was supposed to contain full names but instead contained the words “human resources” and “group.” We noted about 200 outliers in the executive pay data and dropped these as well. One such outlier was a salary of 18 billion US dollars in one year. Additionally, there were also some missing values in the company financials. For instance, 73 of the 485 companies did not have revenue data in 2012 and 14 did not have revenue data in 2021. We decided not to impute values or drop these companies because our classification model was classifying executives as either CEOs or CFOs. The more important metrics, in our eyes, were individual pay statistics.

For exploratory data analysis (EDA), we aimed to find patterns in the pay of CEOs over the last 10 years.

**Figure 2**

*CEO Salary Percentiles 2012-2021*

Chart, line chart

Description automatically generated

Figure 2 measured the percentiles of salaries for CEOs on the x-axis with the salary number on the y-axis. This means if a salary is the highest in the dataset, it will be in the 99th percentile, and if a salary is the lowest in the dataset, it will be in the 1st percentile. From this visualization, we can see that there is a consistent slope until roughly the 80th percentile mark is reached. This means that the increase in income from say the 40th percentile to the 60th percentile is reliable and predictable. This is not the case after about the 80th percentile however, where we see a steepening of the slope of the line for all years. This indicates that those CEOs above the 80th percentile earn much than those below them. A comparable increase in earner percentile, say 20% increase, from the 40th to 60th is much less in real dollar value than a 20% increase from the 70th percentile earner to the 90th percentile earner. Additionally, we highlight the lines for 2020 and 2021 to show the difference between these two years and note that 2021 is much higher for salary than 2020 for any given percentile chosen.

We have also created boxplots, violin plots, kernel density estimate (KDE) plots, a color mapped kernel density estimate (KDE) plot, and bar plots to investigate a multitude of factors. Our focus was on the changing CEO pay landscape, and we used visualizations to investigate changes in salary, stock awards, non-equity incentive compensation, and total compensation. We found a trend of CEO salaries rising in the last 16 years (since 2006), and the most visually telling representation of this trend was the KDE plot.

**Figure 3**

*KDE Plot for CEO Salary Under $3M*

*A picture containing background pattern

Description automatically generated*

Figure 3 displays the density of incomes from 2006-2022 with blue marking the earlier years and red marking the more recent years, with a clear trend of rising salaries during this time.

When analyzing stock awards, we see a surprising widening of box spread in 2021, with very few outliers relative to prior years.

**Figure 4**

*Stock Awards Boxplot Distribution*

A picture containing table

Description automatically generated

Figure 4 displays the overall trend of widening interquartile ranges between 2012 and 2021, which may be due to stock awards being a more common form of compensation now compared with years past. Such a short box with extreme positive skewness in 2012 indicates that a high proportion of CEOs had low stock options as part of their compensation package. In fact, for all years the data is very positively skewed, which indicates that the mean is larger than the median and is heavily impacted by outlying very large values. This is a logical conclusion based on the observation that the top 20% of CEO earners are making much more than the lower 80%. The large stock awards going to these top earners drags up the mean and results in a high positive skewness.

Finally, it is worth noting that companies vary wildly in how they pay their top earners. For instance, Microsoft has a very exaggerated change the year that they elected to have Satya Nadella become the new CEO in 2014. In years prior, the CEO position had collecting hardly any compensation from the company. This is the way that Bill Gates and Steve Ballmer, the prior CEOs, had decided to run the business as original founders. But when it was time to make a change, the chief executive position started to make thousands of times the amount as before. Microsoft is only one example, and it is conceivable that other CEOs that founded their companies likewise have decided to have very low compensation as well.

**Section 5:**

Our basic goal for model selection was to follow the textbook’s advice to select the “to select the model that shows lowest generalization error rate” (Tan et al., 2019). We considered other possibilities for optimal selection, such as minimizing the amounts of false positives or false negatives. Ultimately, we felt that a generalized accuracy measure would work well for this type of data, as there is not a need to be extra conservative like there would be for, say, diagnosing whether a patient’s health indicators classify them as having a disease or not having it. In such a case, it would be better to have false positives (patient does not really have the disease but tests say they do) rather than false negatives (patient really has the disease but tests say they do not).

To evaluate our models, we partitioned the data to “assess the performance of a learned model on a labeled test set has not been used at any stage of model selection” (Tan et al., 2019). We felt that the data set was sufficiently small to follow the books advice for smaller or less complex datasets, where “one should retain sufficient records for accurate assessment, so that the training sets would contain only 50–67% of the original data” (Tan et al., 2019). Our main concern was balancing the proportion so that the training model could effectively learn the behavior but not be so large that the testing model’s error metric was unreliable because it was not able to compute over enough instances. Ultimately, we decided on a 70% partition into the training set and 30% into the test set.

We tried different models to see which would produce the best results for reducing our error rate, producing the best accuracy for true positives. The model which we felt was best for this purpose was the logistic regression model.

**Figure 5**

Chart, scatter chart

Description automatically generated*Chart, scatter chart

Description automatically generatedComparison of Classification Prediction Versus Actual*

Many of the models tended to overfit the training data. For instance, random forest produced a higher overall accuracy, but we felt it overfit the training data with an accuracy of 95% on the training data with 84% on the testing data. The CART and C5.0 models had lower accuracy for both training and testing than the logistic regression model, while also being slightly overfit in the training data. Naïve Bayes, K Neighbors, and SVM produced good results, but also had more of a difference in accuracy between the training and testing datasets than the logistic regression model. Ultimately, logistic regression works well for the general purpose for our model because it is a “discriminative model for classification that directly computes the poster[ior] probabilities without making any assumption about the class conditional probabilities” making it “generic” and “applicable in diverse applications” (Tan et al., 2019). While all of the models in Table 2 are an improvement of the baseline 52% accuracy if all cases of CEO are set to positive, logistic regression was our choice.

**Table 2**

*Comparison of Model Accuracy Results*

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Data Accuracy** | **Test Data Accuracy** |
| CART | 87% | 83% |
| C5.0 | 87% | 80% |
| Random Forest | 95% | 84% |
| Logistic Regression | 90% | 89% |
| Naïve Bayes | 87% | 83% |
| KNeighbors | 91% | 87% |
| Support Vector Machine | 91% | 88% |

**Section 6:**

Classifying CEOs and CFOs is only scratching the surface of the type of information this data source is capable of. Our initial goal was to solve the more politically hard-hitting question of, “do executives actually get paid based on their companies do or do they reward themselves no matter what the company does?” We wanted to create a classification model that would predict if a CEOs or CFOs compensation trends would correctly classify if a company was a rapid grower, flatliner, or decliner. Even though the 2010s was a very prosperous time in the American economy, we found that about 70 of the companies in the S&P 500 had declining revenues during the period. One of the beautiful elements of the economy though is that growth is limitless, while the maximum amount of decline possible is just 100%. Thus, due to the fantastic success of companies like Tesla and Amazon, the growth of the overall economy far outweighed the companies bringing down the overall index.

Beyond classification, the data set also has potential for linear regression projects, as well as unsupervised algorithms like clustering. For example, a linear regression project could investigate which factors are most important to overall compensation for executives. Variables such as company revenue, company net income, position title, age, location, and many more could provide indicators as to which factors affect executive pay statistics the most. Clustering algorithms could be used to find trends that are worth exploring for supervised algorithms that are not known from an initial view of the data.

**References**

Abbott, S., Aksoy, M., Groysberg, B., & Marino, M.R. (2019, February). Compensation Packages That Actually Drive Performance. Principles for designing executive pay. *Harvard Business Review*. <https://hbr.org/2021/01/compensation-packages-that-actually-drive-performance>

Bebchuk, L., & Grinstein, Y. (2005). The Growth of Executive Pay. *Oxford Review of Economic Policy*, *21(2)*, 283–303. <https://doi.org/10.1093/oxrep/gri017>

Bruce, A., & Skovoroda, R. (2015). *The Empirical Literature of Executive Pay: Context, the Pay-Performance Issueand Future Directions.* Nottingham University Business School. <http://irep.ntu.ac.uk/id/eprint/35402/1/12805_Bruce.pdf>

Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

United States Securities and Exchange Commission. (2022, August 25). *SEC Adopts Pay Versus Performance Disclosure Rules.* <https://www.sec.gov/news/press-release/2022-149>