**1. Your task/problem/question, background, and motivation for this data mining project. Clearly state the problem that you have selected to investigate in this data mining project.**

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 <https://www.sec.gov/news/press-release/2022-149>

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, (<https://academic.oup.com/oxrep/article-abstract/21/2/283/450219> ), (B) The effect of increased transparency (<http://irep.ntu.ac.uk/id/eprint/35402/1/12805_Bruce.pdf> ), and (C) Consultancy services advising companies on how to tie executive pay to company performance (<https://hbr.org/2021/01/compensation-packages-that-actually-drive-performance> )

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?**

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

**3. A description of the source and the raw dataset that you perform your data mining skills.**

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 .htm files on the SEC because of the manual labor involved. We instead pulled data 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 featured 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.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

REFERENCES

[freemium paywall]

<https://hbr.org/2021/01/compensation-packages-that-actually-drive-performance>

Compensation Packages That Actually Drive Performance. Principles for designing executive pay

by Boris Groysberg, Sarah Abbott, Michael R. Marino, and Metin Aksoy

From the January–February 2021 issue of Harvard Business Review