Finding Patterns in Top Financial Analysts' Predictions: the Study of a Large Collection of Research Reports

Alessandro Alviani
School of Mathematics, Computer Science and Engineering
City, University of London

Abstract — The presented work analyzes a large collection of financial research reports with the aim to identify top performing financial analysts and find patterns that described their predictions. A dataset containing 277,886 research reports collected over the last ten years was first augmented with companies' close values and ESG scores and then mined following the main principles of Data Science.
It was possible to identify fifteen financial analysts, active in the last three years, with a prediction success rate equal to or greater than 75% (vs. all-analysts mean: 57%). These top analysts mostly recommended buying or holding large market capitalization companies listed in the US markets and which primarily operated in the Health Technology (33%) and Technology Services (22.3%) sectors. There was no statistically significant difference in ESG risk scores of recommended companies versus the rest. Although the top analysts' accuracy was consistently higher than their peers, a pattern of worse performance during market sell-offs was still observed.

I. INTRODUCTION

The first archaic forms of stock markets go as far back as the early 13th century. These emerged in France ("bourse") and Italy ("borsa") and very quickly reached the neighboring countries and spread across Europe [1]. Stock markets are a place for companies to raise money, mainly to grow their business or repay debt, and for private investors to invest their money in the hope of a good return. Due to its highly lucrative nature, ever since its origin, there has been a strong interest in finding ways to predict the stock market and identify those companies which might provide the greatest returns. For example, data from the "2016 Survey of Consumer Finances (SCF)" [2] shows the broad appeal of stock market investment, with 52% of all families in the United States investing some proportion of their income.
Banks are deeply involved in the markets and, through their financial analysts, publish thousands of reports every month. These reports take the form of 12-month performance analysis on specific companies and provide expert advice aimed to guide institutions and private investors in making investment decisions that should generate alpha. However, despite all the best efforts and advancements in technologies, accurately predicting the markets remains an unsolved challenge and over time, inaccurate analysis has resulted in considerable adverse financial outcomes.
As the interest in the stock market has increased over the years, so has the role of Data Science in this domain. Following this trend, the presented work uses some of the main principles of this intradisciplinary field to analyze 277,886 financial reports collected over the last ten years with the aim to find patterns in top financial analysts' predictions.

II. ANALYTICAL QUESTIONS AND DATA

A. Data Received

The initial data was provided by *"Upside Technologies"*, a fintech start-up company based in London.
The data comprised two datasets:

1. *"research_report_data_all.csv"* : 277,886 rows and 9 columns. Each row was a financial research report and the columns its summarizing features (i.e., bank, author, company, rating, target price). This dataset contained the research reports of 1,961 financial analysts, working for 7 different banks and writing on 12,903 companies.
The data frame's features are presented below (Figure 1):

Figure 1 – Research Reports Dataset

research_report_id	research_provider_id	lead_author_id	symbol	research_report_title
39e96cfb-812b-414...	47898792-7a28-4914-b...	73e6854c-ba82-...	AAPL US	Apple Inc. "UBS Evide...

research_rating	research_publish_date	research_target_date	research_target_price
HOLD	2020-12-14	2021-12-14	115.0

2. *"equity_reference_data_all.csv"* : 64,342 rows and 9 columns. This was a table containing the references for 64,342 companies (i.e., name, country, sector, market capitalization size) . All the features are presented below (Figure 2) :

Figure 2 – Equity References Dataset

identifier	symbol	name	country	currency
MH33D6-R	AAPL US	Apple Inc.	US	USD

sector	industry	market_cap_bn	market_cap_classification
Electronic Technology	Telecommunications Equipment	2.456383e+09	Mega Cap

B. Data Preparation

The received datasets contained a wealth of information but, to answer the questions proposed in this study, the data had to undergo an extensive preparation. Firstly, the companies' symbols were mapped to match reports and references with Yahoo Finance's tickers. The two datasets were subsequently merged on the key feature *"symbol"* and augmented with companies' environmental, social and governance (ESG) risk scores and market close prices scraped from Yahoo Finance (YF). Further, the original analysts' research target prices were adjusted by stock splits to make them compatible with the current market's close values.
Finally, analysts' and banks' names (originally masked) were mapped to more readable dummy variables.
The resulting dataset can be outlined as follows:

frame = research reports + companies' references + companies' close values + ESG risk scores

C. Research questions

Upside Technologies offers a mobile platform that provides smart insights into the stock market.

The original datasets were offered as a starting point to a work aimed to identify patterns in the most accurate predictions that, when recognized in future research reports, could be turned into a sound trading advice to their customers.

The proposed research questions were formulated as below:

1. Can analysts that perform significantly better than others be identified?
2. Do features such as sector, country or company size influence the ratings received?
3. Are forecasts influenced by specific external events?
4. Is there any correlation between ESG values and analysts' ratings?

III. ANALYSIS

A. Data cleaning

1. Missing values

The original datasets had a negligible amount of missing data (< 0.1%). After the merge and scraping the ESG and close values from YF the resulting dataset presented 2.7% of missing values in the companies' references. These were research reports that could not be linked to any references hence they were removed from the dataset. With ESG scores being a relatively new indicator, only about half of the companies could be matched with them. Due to their specificity, these ratings could not be imputed, therefore they were left in the dataset as NaN (47%).

2. Duplicates

Duplicates were only found in the original "research_reports" data. They were either different versions of the same report ("updates", "reviews", "drafts") published within a few days of each other, or identical reports published under a different ID.

To make sure these duplicates (or "semi-duplicates") would not bias the analysis towards repeated research reports, they were dropped according to the following logic:

if author has published the same *target price* on the same *symbol* within the same *month*: keep only the first report

This operation removed 26,537 duplicates.

3. Outliers

Several outliers were identified through the exploratory visual analysis, particularly in the *target_price* column. When the close values were plotted against the target prices the outliers showed mostly along the x-axis. From a closer analysis they appeared to be corrupted numbers, not possible to impute without making strong assumptions, therefore they were dropped using a robust logic [3]:

if *target price* is three or more standard deviations away from the *actual close* value's median: drop *target price*.

This operation removed 13,767 outliers.

Figure 3 and 4 show plots of *actual_close* against *target_price* on different zoom levels. The fist two subplots, at low alphas, presented horizontal lines that highlighted some of these outliers. These became more visible in subplots 3 and 4, where they appeared along the x-axis.

Figure 3 – Outliers on different zoom levels (before)

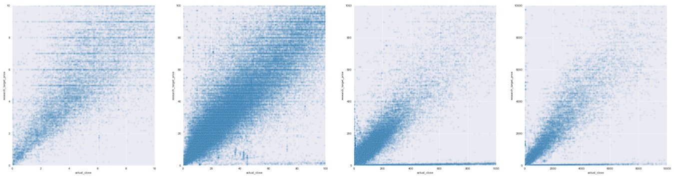
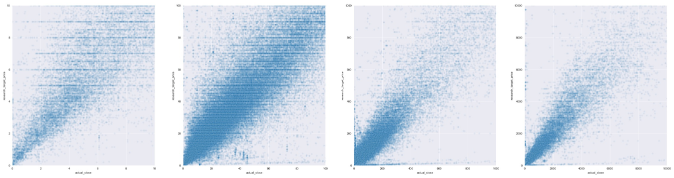


Figure 4 – Outliers on different zoom levels (after)



The adopted cleaning strategy presented noticeable improvements, especially in subplots 3 and 4, whilst subplots 1 and 2 still showed those horizonal lines. Given the use of median and standard deviation to impute the outliers, those horizontal lines could be explained by fixed predictions on stocks that had a high variance.

B. New feature derivation

Data derivation was one of the end goals of all the data preparation of this work. *"prediction_result",* critically, expressed whether the predictions made by the analysts were correct or not. It took a binary form, *"wrong prediction", "correct prediction",* and was derived by comparing the companies' close values with their target price and *rating*. The derivation logic is outlined below:

if *"close_target_date"* < *"research_target_price*" and *"research_rating"* == *"BUY": "prediction_result"* = 0 (*wrong*)

if *"close_target_date"* >*"research_target_ price"* and *"research_rating"* == *"BUY": "prediction result"* = 1 (*right*)

The data offered eight possible combinations given by the four ratings: "Strong Buy", "Buy", "Hold", "Sell" and the two outcomes: "Right", "Wrong".

Each report was successfully labeled and the feature *"prediction_result"* was used in most of the following analysis to rank the analysts and their predictions.

C. Construction of models

WordCloud was a valuable tool to compare the most used words in the titles of the *"right"* and *"wrong"* research reports. With the use of custom "stop-words", non-distinguishing features and features common to both classes were progressively removed. This process was carried on until interesting words appeared and revealed insightful differences between the titles of the two classes of research reports.

The "wrong" research reports' WordCloud is presented in Figure 5.

Figure 5 – WordCloud of "wrong" research reports' titles

The "wrong" research reports showed, as the main word, "Sell", which did not appear in the titles of the correct research reports. This insight was used to discover and validate one of the later findings: top analysts, in their correct predictions, most often recommend buying or holding (aggregated: 90%) and very rarely selling (10%), whilst in their wrong predictions they recommended selling 24% of the times.

Clustering, associated with multiple correspondence analysis (MCA)[4][5], was used to try to identify similarities between categorical companies' features, namely: industry, country and size. It was first applied to all the companies during the initial exploratory data analysis (figure 6) and later to the subset of companies researched by the top performing analysts (figure 7). The latter clusters were also compared to the analysts' prediction outcomes. Despite some weak patterns being observed, these were not particularly insightful to the analysis.

Figure 6 – Clusters of companies by industry, country and size

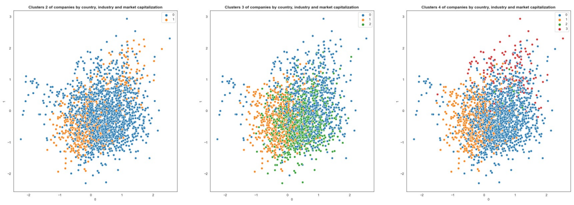
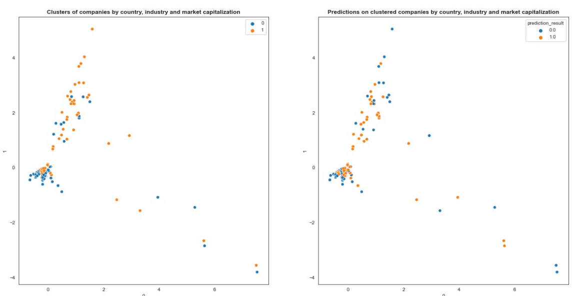


Figure 7 – Clusters of companies by industry, country and size



D. Validation of results

The various findings in this study were validated with the use of comparisons (often to external data), visualizations and statistical measures of significance. Visualizing and comparing were used in the analysis of the outliers, where the analysts' predictions were plotted against the companies' close values (scraped from YF).

Similarly, the scraped S&P 500 index was plotted against the analysts' predictions' accuracy (calculated with a derived feature) to compare and validate its trend.

WordCloud, histograms and computational methods showed the differences in "buy" and "sell" ratings between different types of research reports.

When plots were not robust enough (e.g., in the comparison of the ESG risk scores' means) suitable statistical significance tests, such as Cohen's d [6], were used.

IV. FINDINGS

The work was carried out as an iterative process where the findings in one section informed others and often led to further re-exploration. This paper presents such work in an ordered fashion to facilitate the reading.

1. *Can analysts that performed significantly better than others be identified?*

Analysts that performed significantly better than others were identified. These top-performing analysts were defined as those who had made at least 50 correct predictions and had a success rate equal to or greater than 75% (against the all-analysts average of 56.8%).

Out of 1,386 analysts, 41 all-time best performers emerged, of which 15 were active in the last three years.

2. Did features such as sector, country or company size influence the ratings received?

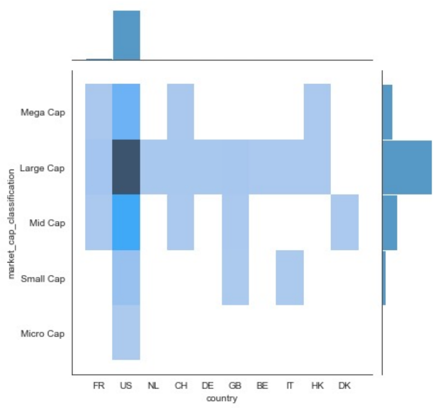
Sector - The top analysts focused their predictions on those sectors that kept performing well despite market crises and even the pandemic. These sectors were Health Technology (33%), Technology Services (20.2%), Electronic Technology (10.5%) and Health Services (10.5%), which may be further grouped in Health and Technology (aggregated: 74.2%).

Industries – More granular than sectors, industries showed a prevalence of Medical Specialties (26%), followed by Packaged Software (8%) and Semiconductors (7%).

Country and company size - Large market capitalization companies were the most prevalent in all the data (44%) and even more in the top-analysts' research reports (63%). Similarly, the US was the most represented country in the data (43.5%) but became dominant (90%) in the top analysts' research reports.

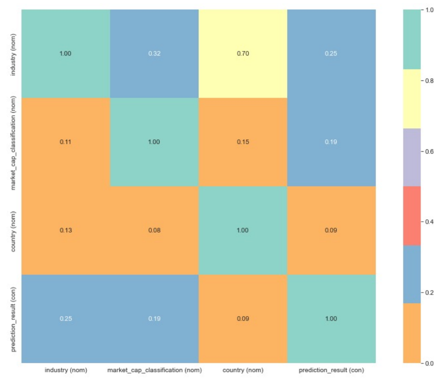
Figure 8 highlights the size of these features in the top analysts' research reports:

Figure 8 – Joint histogram plot of countries and market cap sizes



The correlation between those features and the predictions' outcomes (*"right"*, *"wrong")* were moderate (industry: 0.25, size: 0.19, country: 0.09) (figure 9).

Figure 9 – Correlation Matrix



3. Is forecasts' accuracy influenced by specific external events?

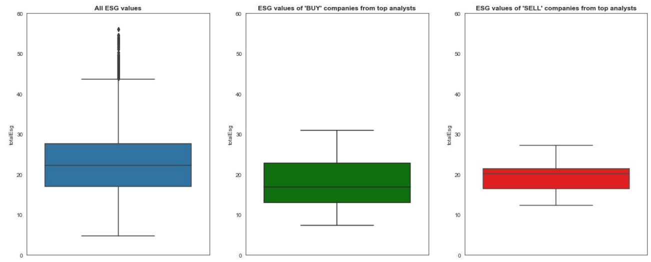
The analysts' predictions were compared to the monthly S&P 500 performance, a commonly referenced market index that responds to external events [7].
External macro-events seemed to have a substantial impact on forecasts' accuracy, with big swings in the market resulting in an average deterioration of analysts' performance of around 20%. All the analysts "suffered" the stock market crises (2016, 2019, 2020), with their prediction accuracy dropping each time the market plunged.
Although the top analysts' accuracy was consistently higher than their peers, this pattern of worse performance during market sell-offs was still prevalent.
All the analysts achieved their best prediction rates during bull markets that demonstrated consistent upward price moves.
Figure 10 shows the above-mentioned aspects of analysts' performance when benchmarked against the SP500.

Figure 10 – "Top analysts' predictions' success rate vs. SP500"

4. Do ESG values influence the analysts' ratings ?

Although top-performing analysts recommended buying or holding companies with mean ESG risk scores lower than the mean on all companies (18.03 vs. 22.77) and companies they rated "Sell" (18.03 vs. 18.91) (figure 11) these differences in means were not statistically significant (Cohen's d: 0.02, and 0.002) given the variance of the data.

Figure 11 – "ESG risk scores' means"



V. CONCLUSIONS

It was possible to identify fifteen analysts who, in the last three years, made between 50 and 427 predictions with a success rate equal to or greater than 75%.
66.3% of these top analysts' research reports were published by Bank 2 and 33.7% by Bank 1.
50% of their research reports recommended buying, 40% holding and only 10% selling.
Their recommendations focused on large market-capitalization companies (63%), predominantly based in the US (90%) and operating in the Health and Technological sectors (aggregated: 74%).
The top-performing analysts' accuracy fluctuated between 60% and 90%, with a mean of 78%, with peaks during consistent upward market trends and lows during market crashes.
There was no statistically significant difference between the ESG risk scores of recommended "buy" companies and the average company in the data.
It was possible to identify the top-performing analysts' names and the 136 companies they recommend to buy.

VI. FUTURE WORK

Further improvements in the tickers' mapping and manual labeling of the remaining ones would have avoided significant data loss and would be advisable for more in-depth analysis.
Likewise, more in-depth work on the outliers could likely recover a significant amount of data.
The dataset lends itself to be augmented with additional features on companies' references and particularly external events. World events, important announcements, inflation and interest rates would broaden the range of the analysis and likely lead to discover meaningful patterns overlooked in this study. These features would also work for the possible build of a machine learning forecasting model.
Scraping data from Yahoo Finance was an effective way to augment the dataset but it turned out to be very slow due to the Yahoo Query Language (YQL) public API blocking large volumes of synchronized requests. YQL private API (with OAuth authentication) could resolve this issue and reduce scraping times drastically.
Finally, focusing the study on specific subsets (i.e., times or industries) would allow more granular findings on topics of interest.

REFERENCES

[1 https://en.wikipedia.org/wiki/Stock_market#History
[2] https://www.federalreserve.gov/econres/scfindex.htm
[3] Osborne, Jason W., and Amy Overbay. "The power of outliers (and why researchers should always check for them)." Practical assessment, research & evaluation 9.6 (2004): 1-12.
[4] Broeksema, Bertjan & Telea, Alexandru & Baudel, Thomas. (2013). Visual Analysis of Multi-Dimensional Categorical Data Sets. Computer Graphics Forum. 32. 10.1111/cgf.12194.
[5] Patrício Soares Costa, Nadine Correia Santos, Pedro Cunha, Jorge Cotter, Nuno Sousa, "The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing", Journal of Aging Research, vol. 2013, Article ID 302163, 12 pages, 2013. https://doi.org/10.1155/2013/302163.
[6] Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4:863. doi: 10.3389/fpsyg.2013.00863
[7] Liu, Chongda & Wang, Jihua & Xiao, Di & Liang, Qi. (2016). Forecasting S&P 500 Stock Index Using Statistical Learning Models. Open Journal of Statistics. 06. 1067-1075. 10.4236/ojs.2016.66086.

Word Count		
Abstract	156	+ 6
Introduction	278	- 22
Analytical questions and data	375	- 225
Analysis	887	- 113
Findings, conclusions and future work	839	+ 239
Total	**2535**	**- 115**