S&P 500 Stock Analysis: Predicting High-Value Securities Using Machine Learning

Our study dives into analyzing the S&P 500 stock market data to predict whether stocks will trade above or below the median price point. This is particularly interesting because accurately classifying stock price levels could help investors make better-informed decisions about their portfolios.

Looking at the data we gathered from 503 different companies in the S&P 500, we noticed some pretty striking patterns. The tech giants really dominate the top spots - Apple's sitting at a massive $3.75T market cap, with NVIDIA and Microsoft right behind. What's really interesting is how different the companies are in size - some have just 28 employees while others, like the tech giants, employ hundreds of thousands.

We decided to tackle this problem using two different approaches. First, we went with logistic regression, which is one of those fundamental techniques we covered early in the course. It's like using a basic tool that still gets the job done well. Then we added bootstrap resampling to the mix - that's one of the more advanced techniques we learned about later. The cool thing about combining these methods is that we get both the straightforward interpretability of logistic regression and the robust validation that bootstrap provides.

The results were pretty interesting. Our model hit about 70.65% accuracy in predicting whether a stock would be above or below the median price. That might not sound mind-blowing, but in the stock market, being right 70% of the time is actually pretty solid. What really caught our attention was how market cap turned out to be super important ($p < 0.001$) while revenue growth, surprisingly, didn't matter much statistically.

Looking at the numbers more closely, we saw some clear patterns. The market cap and EBITDA had this strong relationship with stock prices, which makes sense - bigger, more profitable companies tend to have higher stock prices. But here's the weird thing - EBITDA showed a negative correlation, which wasn't what we expected at all.

The bootstrap analysis gave us an RMSE of 0.47 and an R-squared of 0.13. In plain English, this means our model's okay at the prediction job, but there's definitely room for improvement. It's like we've got the basic picture right, but we're missing some of the finer details.

What does all this mean for real-world applications? Well, investors could use this to get a quick read on whether a stock might be overvalued or undervalued. Portfolio managers might want to look more closely at market cap when making decisions - our analysis suggests it's more important than some might think.

We've got to be honest about the limitations though. The market's way more complicated than just these few factors we looked at. There's all sorts of stuff we couldn't include -

market sentiment, economic indicators, global events. That's probably why our R-squared isn't higher.

Looking ahead, we could make this model better by adding time series data, bringing in some macroeconomic factors, maybe even throwing in some sentiment analysis from social media. But even with its current limitations, our model gives us a decent framework for thinking about stock price levels.

This project really shows how combining old-school statistical methods with newer techniques can give us useful insights into stock market behavior. It might not tell us exactly what stocks to buy, but it definitely adds another tool to the investor's toolkit.