Spinoff Project Summary and Takeaways

Use Case/Primary Question:

Longtime professional investor Joel Greenblatt’s book *You Can Be a Stock Market Genius* endorsed investing in spinoff stocks. These are stocks that previously were just one piece of a larger company but are “spun off” to stand on their own in the financial marketplace. Greenblatt specifically encourages researching these opportunities and selectively investing in the best ones. Out of personal curiosity I wanted to break down the data on these stocks to see if Mr. Greenblatt was right about this specific type of stock being capable of consistently outstanding returns. I have invested in one myself, and to date it has earned 85%. But we need a much bigger sample size to make any reliable conclusions. This project’s results represent a hopeful answer.

Secondary Question:

As I progressed in this project I decided a nice additional wrinkle would be determining if these investments would be affordable for the average person. Stocks can get very expensive, but many of us don’t have lots of money to spend on investing in stocks outside of our other means of saving and investing (e.g. 401ks and other similar retirement vehicles)

Assumptions and Set Up:

* Good investment means a return of at least 10% over a 1-year period. This was initially 15%, but that created imbalanced outcomes in the dataset that were difficult to compensate for. 10% created a fairly even balance of outcomes in the data. 10% is also generally assumed to be the long-term average rate of return for the stock market as a whole. So if your holdings can beat the average, I’d say that’s a good investment. 15% was an arbitrary number I chose to represent a good investment using 10% as an anchor.
* Prices used to calculate the return are generally the opening price of the first day closest to the one-year mark from the day that the stock started trading on its own and the opening price of the closest trading day to the same date the next year. Financial convention often assumes prices used in return calculations are the closing price, but I used opening prices since those are inherently easier to trade on than closing prices. Since the return is being calculated for only one time period, the type of return calculation (e.g. arithmetic, geometric, CAGR, etc) doesn’t matter.
* The cutoff date for spinoffs is 7/31/2019 so we can get a full year of pricing data. Joel Greenblatt says to buy about a year after the spinoff to give the selling pressure time to abate
* If a spinoff was later acquired or merged, then it was generally dropped from the dataset due to difficulty finding pricing data. The idea for this project was to determine what drives returns for companies that run on their own for the long-term, not companies that at some point become acquisitions or involved in mergers
* Type of company was determined by sector. This was the most effective way to standardize firms by their line of business while keeping the dataset relatively compact
* The proxy for “affordable” was decided as an initial price of $100 per share

Data Collection

* Gathered pricing data from wherever I could. Mostly used Yahoo Finance, but also used Stooq, Google Finance, and Barchart. I generally got the sector information from MarketWatch, with an assist from the official NAICS site with translating some of the sector names from MarketWatch to the official ones. Pricing data is generally easy to obtain because the dates that you want are uniform for all stocks, so the only requirement is to get the tickers and write a quick script to call one of the financial information vendors that has an API with Python. That wasn’t the case here. The 1-year return period was different for every single security because their spinoff dates, which were the anchors for the 1-year return start and end dates, were all different.
* Got the spinoff raw data from stockspinoffs.com
* Got the insider buying vs selling data from openinsiders.com, which I used a web-scraper for

Data Cleaning

* Changed spinoff company tickers and/or names for 10-15% of the data set. The information was outdated. Companies changed their names.
* Generally I used the trading start date as the anchor to decide the first date included in the one-year return calculation. But occasionally I used the given spinoff date as the anchor date because the spinoff was really the original business for the firm that was now being spun off to make more room for a newer, more promising business.
* Some of the stocks were foreign and denominated in Danish krone, Euros, Canadian dollars, or Australian dollars. I converted those currencies on those dates into USD because this project was intended to determine whether these stocks would be affordable and good investments for the average US investor. These FX conversions were also important for the good investments/bad investment classification. The returns on these foreign stocks often changed significantly going from the functional currency to the USD.
* Removed an extreme P0 outlier that was messing up all the distribution visuals. The P0 ranged from .04 to 33,000. I removed the 33,000 record because .04 was much closer to the other prices in the data set than 33,000.

Feature Engineering

* I realized in the middle of the project that due to the unique and sometimes extremely complex aspects of each spinoff it would be very hard to generate lots of meaningful features. My attempts to find something meaningful and doable resulted in the following engineered features:
  + P0 bucket
  + Insider trading margin/ratio
* Engineered a variable to quantify whether insiders bought or sold more stock by value during the return calculation period. Sometimes executives’ stock buying patterns in their employers can portend good and bad things. This is because (1) they are likely to know something about the company that the public doesn’t – good or bad – and (2) buying or selling by itself in large enough volume can move the stock price. The hope was that a net insider trading margin by value might correlate with the return, given that the insider trading activity and price movement occurred during the same windows. Unfortunately the correlation was nonexistent.
* Engineered another variable to group the starting prices into buckets to see if this classification had any relationship to the good investment/bad investment decision. The buckets were based on IQR. Bucket 1 was less than 25th percentile. Bucket 2 was between 25th and 75th inclusive. Bucket 3 was greater than 75th percentile. Unfortunately this feature also had virtually no correlation. You can see representative scatterplots below

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Machine Learning Application

* The goal was to determine if spinoffs are on, average, a good investment. “Good investment” is a binary outcome. Either the investment is good, or it isn’t. Therefore I used logistic regression for the ML model.
* The dataframe that I apply the model to had the following driver variables: P0, 1-Year Return, and dummy variable for sector. The response variable was Good Investment/Decision

Results:

* As you can see from the regression summary results below, only the insider trading margin was statistically significant, but its coefficient was so small as to be worthless. The only other variable that is even close to statistical significance is the Financials sector dummy variable. That outcome was expected after I investigated correlation among the variables extensively and found nothing meaningful, but I still wanted to run the model to show I can do that, and to show I can interpret statistical output.

Table

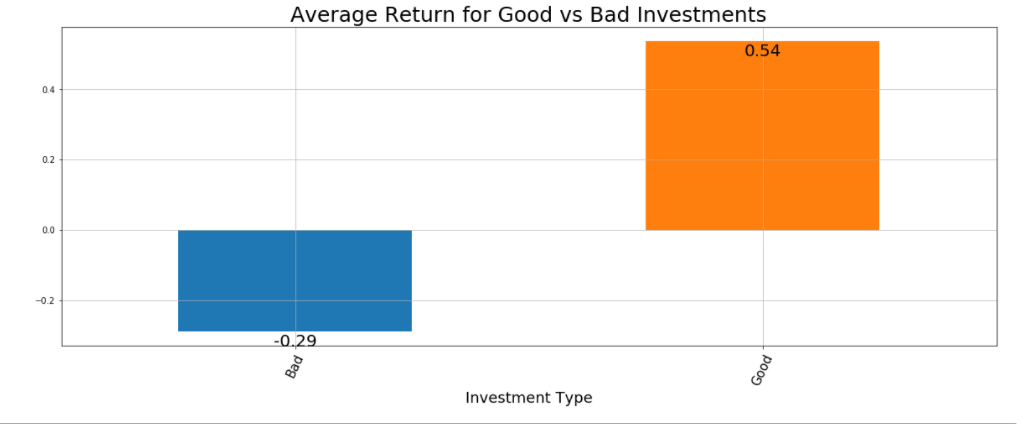
Description automatically generated

* Below is the classification report. The average F1 score is 62%, which is better than random chance. I’m surprised that the F1 score is as high as it is given that only one variable was statistically significant and even that one had such a small coefficient that it would hardly contribute to price movements anyway

Table

Description automatically generated

* The good investments tend to be really good, and the bad ones tend to be really good.



* The good investments are concentrated in more defensive sectors like Consumer Staples and Healthcare. The bad ones are concentrated in cyclical sectors like Energy and Financials. This makes sense. Firms in defensive sectors are more likely to pursue spinoffs because the whole firm is doing well and executives are confident that the spinoff will do fine on its own. Firms in cyclical sectors are more likely to pursue spinoffs because part of the firm is dragging the rest of the firm down, and the spinoff is a way to take out the trash. Note - the utilities probability of a good investment in the sector is an outlier because there was only one utilities firm in the data set.

Chart, waterfall chart

Description automatically generated

Chart, bar chart

Description automatically generated

* There were ultimately 81 stocks out of 176 in the data set that were both good investments and affordable – 46% of the data set – and those stocks had an average return of 54%! I don’t have a graph for this one because I think the numbers speak for themselves

Learning Takeaways:

* Gained an appreciation for what it takes to build a dataset almost from scratch, even though this was a fairly small one. On a related note, I learned how important but also challenging data cleaning can be.
* Importance of having domain knowledge so you can leverage it to help determine the scope of the project. I considered including data on mergers and acquisitions too, but determined it would take too long and also kind of miss the point. Investing in merging companies or acquisition targets is generally not known as a way to find value in the stock market. Spinoffs are. My goal, leveraging my domain knowledge, was to determine if spinoffs really were a good investment on average for average investors. I also removed some data points because further research into the transactions suggested to me that the fact pattern for these firms didn’t really match the typical spinoff
* How different file formats can help facilitate a project. I used downloadable CSV files from Yahoo Finance to populate the price data. I tried using Pandas with Stooq but the date was automatically populated in the dataframe as the index, and there was no good way to transfer the data from the index to a column
* Importance of clearly defining the parameters and assumptions you use, and keeping them handy for reference as you work. I collected around 20% of my pricing and dates data before realizing I was using dates that I didn’t want
* Learned how to work around limitations inherent in data sources. The date and pricing data was very much available for just viewing, but not so much for collecting in one place. First I tried adjusting the website settings to download all pricing data in a CSV file, and after download I’d remove all records except for the two dates I wanted. I’d copy the data over, then save the curated CSVs for reference. This process was very slow and tedious. Then I realized that for most entities I could adjust the download settings to show me the first trading day for a given security. Then I’d use that date to determine the first and last dates to collect prices for determining the return. It was faster and easier.
* Learned better how to preprocess data. I first tried feeding the model the floating-point returns for the target variable. Given that it was a logistic regression model, it didn’t take those values. So I quickly converted the raw returns to a good/bad investment dummy variable and regressed on that.
* Got a better feel for how to create specific features for a final dataset that you run the model against with custom functions