**The Problem**

Investing in startup companies can be risky due to ever changing market trends, management competency and technological advancement. Predicting which startup companies have potential to make profitable returns can help investors make informed decisions on their investments.

**The Goals**

Identifying IPOs that offer promising investment opportunities for investors seeking profitable returns, especially after being publicly listed on the market and has at least a 250-million-dollar valuation before selling after 262 days.

**The Criteria for success**

[Price Gain Ratio > 1]

If Price Gain Ratio > 1, it means the stock is profitable after 261 days.

If Price Gain Ratio < 1, it means the stock is unprofitable after 261 days.

**Overview of approach**

1. Defining Objectives: Goals of the analysis outlined

2. Collecting Data: Gathering relevant data which have integrity and quality

3. Cleaning Data: Addressing missing values, outliers, and inconsistencies

4. Analyzing Data: Applying appropriate analytical techniques to uncover patterns and insights.

5. Interpreting Results: Relating the findings back to the original objectives and draw meaningful conclusions.

6. Communicating Findings: Presenting the results through visualizations, reports, or presentations to stakeholders.

**Data Sources and Definitions in a Data Dictionary**

Kaggle: <https://www.kaggle.com/datasets/proselotis/financial-ipo-data/data>

Last Updated 5 Years ago

|  |  |
| --- | --- |
| **Column Headers** | **Definition** |
| Symbol | Stock Symbol |
| DaysBetterThanSP | Days the stock had a higher percentage change than the S&P500 |
| daysProfit | Days the stock was positive |
| daysProfitGrouped | A grouped form of positive days |
| Year | Year the stock went public |
| Month | Month the stock went public |
| Day | Day the stock went public |
| dayOfWeek | Day of the week the stock went public |
| closeDay0 | The value of the stock on close of first day |
| volumeDay0 | The total amount of trades on the first day |
| closeDay6 | The value of the stock on close of the day |
| volumeDay6 | The total amount of trades on the day |
| closeDay29 | The value of the stock on close of the day |
| volumeDay29 | The total amount of trades on the day |
| closeDay59 | The value of the stock on close of the day |
| volumeDay59 | The total amount of trades on the day |
| closeDay89 | The value of the stock on close of the day |
| volumeDay89 | The total amount of trades on the day |
| closeDay179 | The value of the stock on close of the day |
| volumeDay179 | The total amount of trades on the day |
| closeDay261 | The value of the stock on close of the day |
| volumeDay261 | The total amount of trades on the day |
| Name | Name of the company |
| MarketCap | Total value of a publicly traded company |
| Sector | Groups of companies that are involved in similar lines of business or economic activities |
| ipoDate | Date the stock went public |
| CEOInChargeDuringIPO | The CEO was in charged or not |
| presidentInChargeDuringIPO | The President was in charged or not |
| stateCountry | Country names |
| Revenue | Company’s total income |
| netIncome | Company’s total profit |
| employees\_x | Total number of people employed at the company |
| MarketMonthTrend | Asset’s price direction over a month |
| Market3MonthTrend | Asset’s price direction over 3 months |
| Market6MonthTrend | Asset’s price direction over 6 months |
| Profit\_Ratio | Price at day 261 divided by price at day 0 |
| Profitable | If Profit\_Ratio > 1, True  Else < 1, False |

**Context**

Initial Public Offering (IPO): When a privately held company offers shares to the public for the first time, allowing it to raise capital and become publicly traded.

The median is often preferred over the mean when a dataset contains outliers or is skewed. The mean, being sensitive to extreme values, can be pulled away from the true center of the data.

|  |  |
| --- | --- |
| **Categories** | **Market Capitalization** |
| Mega-cap | $200 billion and greater |
| Big-cap | $10 billion - $200 billion |
| Mid-cap | $2 billion - $10 billion |
| Small-cap | $250 million - $2 billion |
| Micro-cap | $50 million - $250 million |
| Nano-cap | Under $50 million |

Assumptions

Include only companies with at least 250-Million-dollar valuation

Based on US market only

All sectors included

Limitations

Missed out potential high growth nano and micro-cap IPO companies

Missed out potential high growth IPO companies from some other markets

Might not be great as technology sector grows/fails faster than other sectors

**Data Processing Steps**

1. Libraries imported
2. IPO CSV imported
3. Removed duplicates using the ‘symbol’ column
4. Dropped unnecessary columns
5. Dropped all rows with nulls
6. Created 2 metric columns

-Profit ratio: Created by dividing closeDay261 by closeDay0

-Profitable: Created by checking if Profit Ratio > 1

1. Adjusted data types

-Changed all ‘-‘ with 0 under the ‘employee’ column

-Changed ‘employee’ column type to integer

-Created ‘mean\_employee\_sector’ table: Grouped by sector of median employees

-Merged ‘ipo\_data’ and ‘mean\_employee\_sector’ by sector

-Replaced all 0 in ‘employees\_X’ with values from ‘employees\_y’

-Dropped column ‘employees\_y’

-Changed ‘employees\_X’ data type to integer

-Changed ‘ipodate’ to datetime format

-Used a FOR loop statement on columns, 'CEOInChargeDuringIPO' and 'presidentInChargeDuringIPO', changed rows with ‘Yes’, ‘No’, ‘sameYear’ to 1, 0 ,0 respectively

-Replaced all strings with length of 2 in 'stateCountry' column to 'USA'

1. Filtered for companies with at least a 250-million-dollar valuation
2. Made ‘revenue’ column values into per millions
3. Removed all outliers using a FOR loop
4. Plotted histogram
5. Saved progress to CSV as ipo\_data\_cleaned.csv
6. Checked value counts for categorical columns
7. Created 1 hot encoding for 'Sector' & 'Month'
8. Calculated correlation in respect to ‘Profitable’ column

**Modelling**

1) KNN Classification

-Assigned features to ‘feature\_cols’

-Loaded data into a test split and scaler transformation with a random state of 123

-Found the K value for KNN classification model

-Plotted a graph showing testing and training error

-Calculated Feature Importance

-Checked the results from the model to the null accuracy

-Created a confusion matrix

-Modification of KNN Model adjusting features

-Created pipeline

-Grid searched over number of features (k) and KNN params

-Found which features gives the best accuracy

-Replaced features with new columns selected

-Checked the results from the model to the null accuracy

2) Decision Tree Classification

-Decision tree classification model loaded with original features selected

-Checked the results from the model to the null accuracy

-Calculated Feature Importance

-Created a confusion matrix

-Used GridSearch and SelectKBest to find the most optimal input features

-Got SelectKBest

-Replaced features with new columns selected

-Checked the results from the model to the null accuracy

-Created a confusion matrix

3) Logistic Regression Classification

-Logistic Regression classification model loaded with original features selected

-Calculated Feature Importance

-Checked the results from the model to the null accuracy

-Created a confusion matrix

-Used GridSearch to find the most optimal input features

-Checked the results from the model to the null accuracy

**Patterns, Trends and Insights**

Introduction

Total number of companies: 484 companies

Total number of countries: 15

Total number of sectors: 12

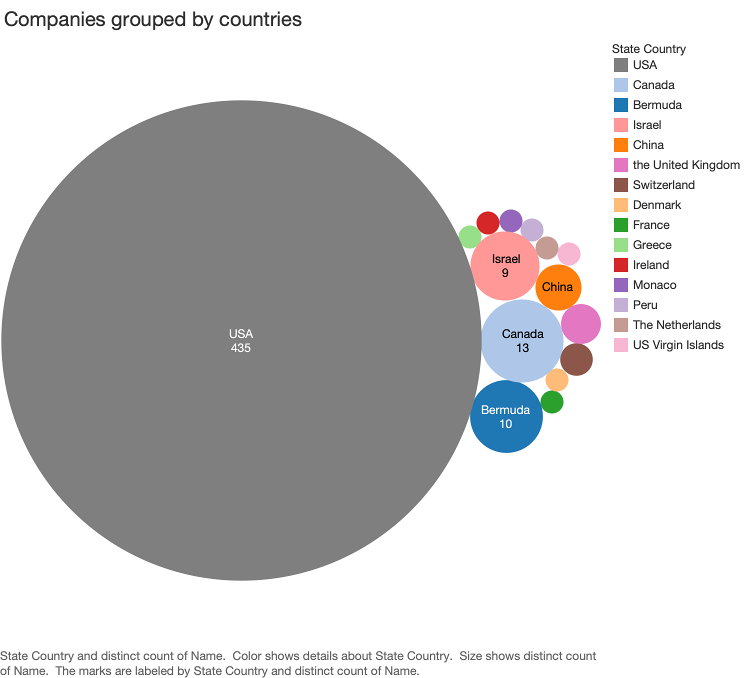
Years Included: 1996 - 2017

Average closing price for stocks on day 1: $15.13

Average days stocks do better than the S&P 500: 121 days

Before Dataset cleaning: 3762 rows and 1664 columns

After Dataset cleaning: 485 rows and 37 columns



USA has the most companies in the US market of 435, with the next biggest being Canada of 13 companies.

A screenshot of a chart

AI-generated content may be incorrect.

Health Care, Consumer Services and Finance Sector has the most companies with Transportation being the lowest.

A graph with a line

AI-generated content may be incorrect.

In 2014, the number of companies that IPO, peaked at 51 but dropped sharply after that.

A graph with numbers and bars

AI-generated content may be incorrect.

A graph of a price

AI-generated content may be incorrect.

These are the top 10 companies that have the highest and lowest opening prices. From the Top 10 companies with the highest opening price chart, 4 companies come from the healthcare sector.

The company with the highest opening price is Aduro Biotech of $42 and the company with the lowest opening price is Cronos Group of $0.17, with the average opening price being $15.13

A graph with numbers and a line

AI-generated content may be incorrect.

The most popular day of the week to IPO is on Thursday, with 0 companies that IPO on a Friday.

A screenshot of a graph

AI-generated content may be incorrect.

 59.8% of companies are profitable after 262 days on being on the stock market.

A diagram of company's company's company's company's company's company's company's company's company's company's company's company's

AI-generated content may be incorrect.

These are the top 10 companies that beat the S&P 500. 3 of the 10 companies comes from the Energy Sector

A diagram of company's company's company's company's company's company's company's company's company's company's company's company's

AI-generated content may be incorrect.

These are the top 10 companies that least beat the S&P 500. 5 of the 10 companies comes from the Consumer Services Sector.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

These 2 charts show the companies in different sectors by their Market Capital, Employee count, Revenue and Income.

Public Utilities has the highest Market Capital with Transportation having the lowest Market Capital, with it have one of the most employee centric sectors. Even though Transportation has the highest revenue, they have one of the lowest Profits, with Health Sector having negative profits. This might be due to healthcare not earning enough revenue.

A graph of different colored bars

AI-generated content may be incorrect.

Companies in the transportation sector experience negative median market month trends over the 6 months since they IPO.

A graph with blue and brown bars

AI-generated content may be incorrect.

These are the 4 sectors that have a higher chance of getting an unprofitable return after selling them after 262 days.

A graph of a bar chart

AI-generated content may be incorrect.

These are the 8 sectors that have a higher chance of getting a profitable return after selling them after 262 days.

**Predictive Modelling**

Correlation Table

|  |  |
| --- | --- |
| Column Name | Correlation |
| daysProfit | 0.704692 |
| Profit\_Ratio | 0.546951 |
| closeDay261 | 0.481725 |
| DaysBetterThanSP | 0.349080 |
| closeDay179 | 0.299088 |
| netIncome | 0.139873 |
| Year | 0.103277 |
| MarketCap | 0.102705 |
| employees\_x | 0.100750 |
| closeDay6 | -0.103770 |
| Market3MonthTrend | -0.106430 |
| volumeDay0 | -0.113906 |
| Market6MonthTrend | -0.114429 |
| MarketMonthTrend | -0.120170 |
| closeDay0 | -0.129045 |

Correlation helps determine if changes in one variable are associated with changes in another, which ranges from -1 to +1, indicates the strength and direction of the relationship.

After finding the correlation of the features to what stocks are profitable, these are the 13 features that are chosen.

**1. K-Nearest Neighbors (KNN) Classification**

Below shows the training and testing error, which finds the best K with the least error to be 0.

A graph with a line graph and a line graph

AI-generated content may be incorrect.

**1st model: 8 Neighbors | 0.844 Accuracy**

Null accuracy: 0.557 Accuracy

The First model gave an accuracy score of 0.844 which is better than the null accuracy of 0.557.

A diagram of a confused matrix

AI-generated content may be incorrect.

Shown above is the confusion matrix. This shows that 43 companies were predicted correctly to be profitable while 60 companies were also predicted correctly to be unprofitable. However, 11 companies were profitable but predicted unprofitable and 8 companies were predicted as profitable but was unprofitable.

Next, a second model was made using Gridsearch to find the best features and parameters to put into the model.

**2nd model: 6 Neighbors | 0.787 Accuracy**

Selected features: ['daysProfit', 'DaysBetterThanSP', 'closeDay179', 'netIncome', 'Market3MonthTrend', 'volumeDay0', 'Market6MonthTrend', 'MarketMonthTrend', 'closeDay0']

Best KNN parameters:

- 'algorithm': 'auto'

- 'leaf\_size': 5,

- 'n\_neighbors': 8,

- 'p': 1, 'weights': 'uniform'

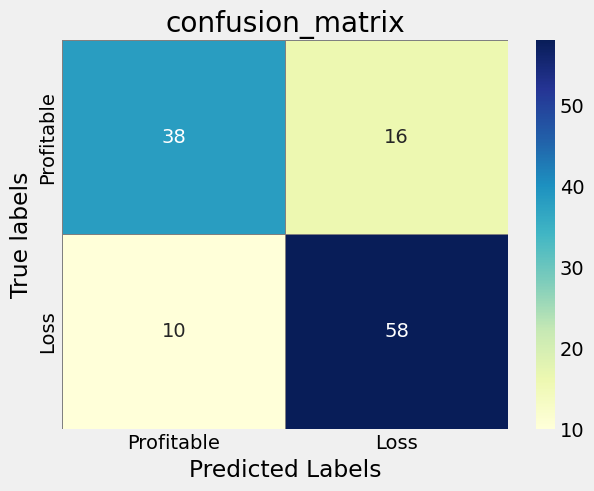
In conclusion, the 1st KNN model was the better choice.

**2. Decision Tree Classification**

**1st model: 0.852 Accuracy**

Null accuracy: 0.557 Accuracy

The 1st Model gave an accuracy score of 0.852 which is better than the null accuracy of 0.557.



Shown above is the confusion matrix. This shows that 38 companies were predicted correctly to be profitable while 58 companies were also predicted correctly to be unprofitable. However, 16 companies were profitable but predicted unprofitable and 10 companies were predicted as profitable but was unprofitable.

Next, a second model was made using Gridsearch to find the best features and parameters to put into the model.

**2nd model: 0.877**

Selected features: ['daysProfit', 'DaysBetterThanSP']

Best Decision Tree Parameters:

* 'dt\_\_ccp\_alpha': 0.0
* 'dt\_\_max\_depth': 5
* 'dt\_\_max\_features': 'sqrt'
* 'dt\_\_min\_samples\_leaf': 10
* 'dt\_\_min\_samples\_split': 2
* 'select\_\_k': 2

A chart with different colored squares

AI-generated content may be incorrect.

The Confusion Matrix is the same as the 1st model. Therefore the 2nd model is preferred over the 1st model as it has a higher accuracy.

**3. Logistic Regression Classification**

**1st model: 0.893 Accuracy**

Null accuracy: 0.557 Accuracy

The 1st Model gave an accuracy score of 0.893 which is better than the null accuracy of 0.557.

A chart with different colored squares

AI-generated content may be incorrect.

Shown above is the confusion matrix. This shows that 38 companies were predicted correctly to be profitable while 58 companies were also predicted correctly to be unprofitable. However, 16 companies were profitable but predicted unprofitable and 10 companies were predicted as profitable but was unprofitable.

Next, a second model was made using Gridsearch to find the best features and parameters to put into the model.

**2nd model: 0.885 Accuracy**

Best Logistic Regression parameters:

'C': 10

'solver': 'liblinear'

The 1st model was chosen as it has a higher accuracy than the 2nd model.

**Comparison between the Models**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **K-Nearest Neighbors** | **Decision Tree** | Logistic Regression |
| **Accuracy** | **0.844** | **0.877** | **0.893** |
| **True Positives** | **43** | **38** | **38** |
| **True Negatives** | **60** | **58** | **58** |
| **False Positives** | **8** | **16** | **16** |
| **False Negatives** | **11** | **10** | **10** |

Logistic Regression (highest accuracy: 89.3%)

K-Nearest Neighbors (lowest FP = 8) (highest TP = 43)

Logistic Regression is chosen for balanced and reliable predictions of profitable stocks. It gives a strong overall performance and avoids the costliest errors without sacrificing too many opportunities.

**Feature of Importance**

Below are the features of importance on the 3 models

|  |  |  |
| --- | --- | --- |
| **Rank** | **KNN Features** | **Feature Importance** |
| 1 | daysProfit | 0.144904 |
| 2 | DaysBetterThanSP | 0.025069 |
| 3 | closeDay0 | 0.011019 |
| 4 | closeDay6 | 0.010468 |
| 5 | MarketCap | 0.007438 |
| 6 | closeDay179 | 0.003581 |
| 7 | Year | 0.002204 |
| 8 | Market6MonthTrend | -0.000275 |
| 9 | Market3MonthTrend | -0.002204 |
| 10 | volumeDay0 | -0.003306 |
| 11 | employees\_x | -0.005785 |
| 12 | MarketMonthTrend | -0.006887 |
| 13 | netIncome | -0.007989 |

|  |  |  |
| --- | --- | --- |
| **Rank** | **DT Features** | **Feature Importance** |
| 1 | daysProfit | 0.79686 |
| 2 | DaysBetterThanSP | 0.20314 |

|  |  |  |
| --- | --- | --- |
| **Rank** | **LR Features** | **Feature Importance** |
| 1 | daysProfit | 1.470206 |
| 2 | closeDay6 | 1.295600 |
| 3 | DaysBetterThanSP | 0.658048 |
| 4 | closeDay179 | 0.557961 |
| 5 | closeDay0 | 0.410486 |
| 6 | netIncome | 0.314902 |
| 7 | MarketCap | 0.191450 |
| 8 | Year | 0.175671 |
| 9 | Market6MonthTrend | 0.147758 |
| 10 | MarketMonthTrend | 0.139165 |
| 11 | Market3MonthTrend | 0.094513 |
| 12 | volumeDay0 | 0.073432 |
| 13 | employees\_x | 0.051647 |

**Recommendations**

Avoid investing in companies in sectors like Capital Goods, Consumer Durables, Finance and Transportation.

Investing in companies in sectors like Basic Industries, Consumer Non-Durables, Consumer Services, Energy, Health Care, Miscellaneous, Public Utilities and Technology.

Use Logistic Regression to predict if a stock is profitable or not using features like:

Days Profit, Profit Ratio, Close Day 261, Days Better Than SP, Close Day 179, Net Income, Year, Market Cap, Employees, Close Day 6, Market 3Month Trend, volume Day 0, Market 6Month Trend, Market Month Trend and close Day 0.

Further analysis on features could be done with more advanced models to classify stocks.