Startup Investments

Predicting the success of the startup companies using Machine Learning

GROUP 6

INTRODUCTION

In the last few decades, startups have turned age-old industries on their heads, solved big problems with the click of a button, and have managed to cash in big on their products and services – if they're successful. A startup is traditionally defined as a newly established private company (< 5 - 10 years old), that is designed to scale very quickly. Most startups kick off as very small operations while they develop their initial idea, and then seek additional funding from venture capitalists and angel investors as they build out their businesses. Companies can receive many different types of funding – Angel, Pre-Seed, Seed, Venture, Round A, Round B, Round C.

We explored the world's largest structured database for start-ups – provided by the website CrunchBase.com, with the objective of building a predictive model, through supervised learning, to accurately classify which start-ups are successful and which aren't.

The success of a start-up is commonly defined as a two-way strategy as a company can either have an IPO (Public Initial Offering) by going to a public stock market, allowing its shareholders to sell shares to the public, or be acquired or merged (M&A) with another company where those who have previously invested receive immediate cash in return for their shares. This process is often designated as an exit strategy.

Previous studies tend to focus primarily on managerial features and often overview the impact of financial features related with funding (specially from Venture Capital funds). So, we predict the success of a startup company based on market, region and financial variables. This prediction will help investors to get an idea whether investing in a startup will be successful or not?

DATA

The dataset was downloaded from the following kaggle link:
https://www.kaggle.com/arindam235/startup-investments-crunchbase

The dataset related to startup investments in 116 countries has been collected from the CrunchBase database. We considered only the startup companies based in USA.

The factors we considered - the type of market, the state, the region the startup belongs to, total funding (in USD), different types of funding – seed, angel, crowdfunding etc., founded date, last funding date, operating status.

Both an IPO (Initial Public Offering) and a process of M&A (Mergers & Acquisitions) are the critical events that classify a start-up as successful.

We created the “post_success” column which indicates if the startup was successful or not successful based on these rules – if the status of the company is “acquired” or if the post IPO equity is greater than 0 or if the status of the company is “acquired” and the post IPO equity is greater than 0.

KEY INFORMATION IN THE DATASET

- **name:** Name of the startup company
- **category_list:** Category the startup belongs to
- **market:** Type of market the startup belongs to
- **funding_total_usd:** Total funding the startup received
- **status:** Status of the startup – acquired, operational, closed
- **state_code:** State in which the startup was founded in
- **region:** Region in which the startup was founded in
- **city:** City in which the startup was founded in
- **funding_rounds:** Number of funding rounds the startup went through
- **founded_at:** Day on which the startup was founded
- **founded_month:** Month in which the startup was founded
- **founded_quarter:** Quarter in which the startup was founded – Q1, Q2, Q3, Q4
- **founded_year:** Year in which the startup was founded
- **first_funding_day:** Day on which the startup started receiving funding
- **first_funding_month:** Month in which the startup started receiving funding
- **first_funding_year:** Year in which the startup started receiving funding
- **last_funding_day:** Day on which the startup stopped receiving funding
- **last_funding_month:** Month in which the startup stopped receiving funding
- **last_funding_year:** Year in which the startup stopped receiving funding
- **seed:** Seed rounds are among the first rounds of funding a company will receive, generally while the company is young and working to gain traction. A seed round typically comes after an angel round (if applicable) and before a company’s Series A round.
- **venture:** Venture funding refers to an investment that comes from a venture capital firm and describes Series A, Series B, and later rounds. This funding type is used for any funding round that is clearly a venture round but where the series has not been specified.
- **equity_crowdfunding:** Equity crowdfunding platforms allow individual users to invest in companies in exchange for equity. Typically, on these platforms the investors invest small amounts of money, though syndicates are formed to allow an

individual to take a lead on evaluating an investment and pooling funding from a group of individual investors.

- **undisclosed:** Undisclosed amount on the last funding date.
- **convertible_note:** A convertible note is an ‘in-between’ round funding to help companies hold over until they want to raise their next round of funding. When they raise the next round, this note ‘converts’ with a discount at the price of the new round. You will typically see convertible notes after a company raises, for example, a Series A round but does not yet want to raise a Series B round.
- **debt_financing:** In a debt round, an investor lends money to a company, and the company promises to repay the debt with added interest.
- **angel:** An angel round is typically a small round designed to get a new company off the ground. Investors in an angel round include individual angel investors, angel investor groups, friends, and family.
- **grant:** A grant is when a company, investor, or government agency provides capital to a company without taking an equity stake in the company.
- **private_equity:** A private equity round is led by a private equity firm or a hedge fund and is a late stage round. It is a less risky investment because the company is more firmly established, and the rounds are typically upwards of $50M.
- **post_ipo_equity:** A post-IPO equity round takes place when firms invest in a company after the company has already gone public.
- **post_ipo_debt:** A post-IPO debt round takes place when firms loan a company money after the company has already gone public. Similar to debt financing, a company will promise to repay the principal as well as added interest on the debt.
- **secondary_market:** A secondary market transaction is a fundraising event in which one investor purchases shares of stock in a company from other, existing shareholders rather than from the company directly. These transactions often occur when a private company becomes highly valuable and early stage investors or employees want to earn a profit on their investment, and these transactions are rarely announced or publicized.
- **product_crowdfunding:** In a product crowdfunding round, a company will provide its product, which is often still in development, in exchange for capital. This kind of round is also typically completed on a funding platform.
- **round_A:** Round A is a funding round for earlier stage companies and range on average between $1M–$30M.
- **round_B:** Round B is a funding round for earlier stage companies and range on average between $1M–$30M.
- **round_C:** Round C is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.
- **round_D:** Round D is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.
- **round_E:** Round E is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.
- **round_F:** Round F is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.
- **round_G:** Round G is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.
- **round_H:** Round H is a funding round for later stage and more established companies. These rounds are usually $10M+ and are often much larger.

- **post_success:** Indicates if the startup was successful or not successful based on the rules mentioned above.

DATA PREPROCESSING

- On initial observation, the dataset downloaded from Kaggle had a lot of missing data. So, we removed rows which had missing data in one or more columns.

- The values in the column **funding_total_usd** had commas. We removed commas for the values.

- The **market** column had left and right trailing spaces for each keyword. So, we trimmed it using trimws() function.

- We didn't require **homepage URL** of the startup company, **X** which is the ID, **permalink** - permanent link in the crunchbase, **country code**. We are considering only startups in United States of America (USA).

- We split columns which had dates in the format (YY/MM/DD) like **first_funding_day**, **first_funding_month**, **first_funding_year**, **last_funding_day**, **last_funding_month**, **last_funding_year**.

- There are more than 32 markets, regions and states. We binary encoded the values in **market**, **region**, **state_code**. Then, digits from that binary string are split into separate columns. This is useful for machine learning. Better encoding of categorical data can mean better model performance.

- We selected data from year 1990 until 2014. We converted all the categorical variables to factors. We also converted the data types from "num" to "logical" (Boolean) for the binary encoded columns.

- In the end, we got a cleansed dataset which was stored as an RDS file.

CORRELATION ANALYSIS

From the last matrix, we can see that

- funding_rounds
- founded_at
- founded_year
- seed
- venture
- post_ipo_equity
- round_A
- round_B

- round_C
- first_funding_day
- first_funding_year
- last_funding_month
- last_funding_year

are all significantly associated with outcome variable **post_success** (p-values < 0.01).

PREDICTION MODELS

- We predicted if the startup was successful using different prediction models and chose the best model by observing the performance metrics.

- By using this model, we can explain or predict if the startup will be successful or not from the other factors it is dependent on.

- We took the performance metrics (Accuracy, Kappa, Precision, Recall, Specificity, F1 score, AUC, Balanced Accuracy) to check the performance of the model.

- For prediction and performance measurement, we used classification models instead of regression models because our variable **post_success** to be predicted is a categorical response variable.

- We used 3 algorithms, namely the **KNN**, **logistic regression** and the **gradient boosting model** (**GBM**) to train 3 different machine learning models. The features obtained during the data preprocessing phase were used to fit the 3 different models.

Models	**AUC**
Logistic Model	0.995
Gradient Boosting Machine	1.000
KNN	0.533

- The **AUC score** for each of the machine learning models were compared and the model with the maximum AUC score is considered for the given dataset.

- Based on AUC score, we can observe that our best model is **Gradient Boosting Machine (GBM). Logistic Regression (GLM)** is the second-best model.

RESULTS AND CONCLUSIONS

- California (CA) state has the highest number of startups and the New York (NY) state is in the second place.
- Most of the startups are in the field of software, Biotechnology, Health Care, E-Commerce.
- Regions like San Francisco Bay Area, New York City, Boston, Los Angeles and Washington DC have higher number of startups compared to other regions.
- Startups in the field of software, advertising and biotechnology have more success rate when compared to others.
- Successful companies are backed by Venture capital. Even top companies like Facebook, YouTube, Google are backed by venture capital in the initial stages.
- Every year few successful companies are acquired, whereas as few are closed and the remaining stay operated.
- More number of funding rounds, seed amount, venture capital, post IPO equity contributes to the success of a startup.
- More funding in the initial rounds A, B,C can contribute to the success of a startup.
- We used 3 algorithms, namely the **KNN**, **logistic regression** and the **gradient boosting model** (**GBM**) to train 3 different machine learning models. Based, on AUC score we concluded that GBM is the best model.