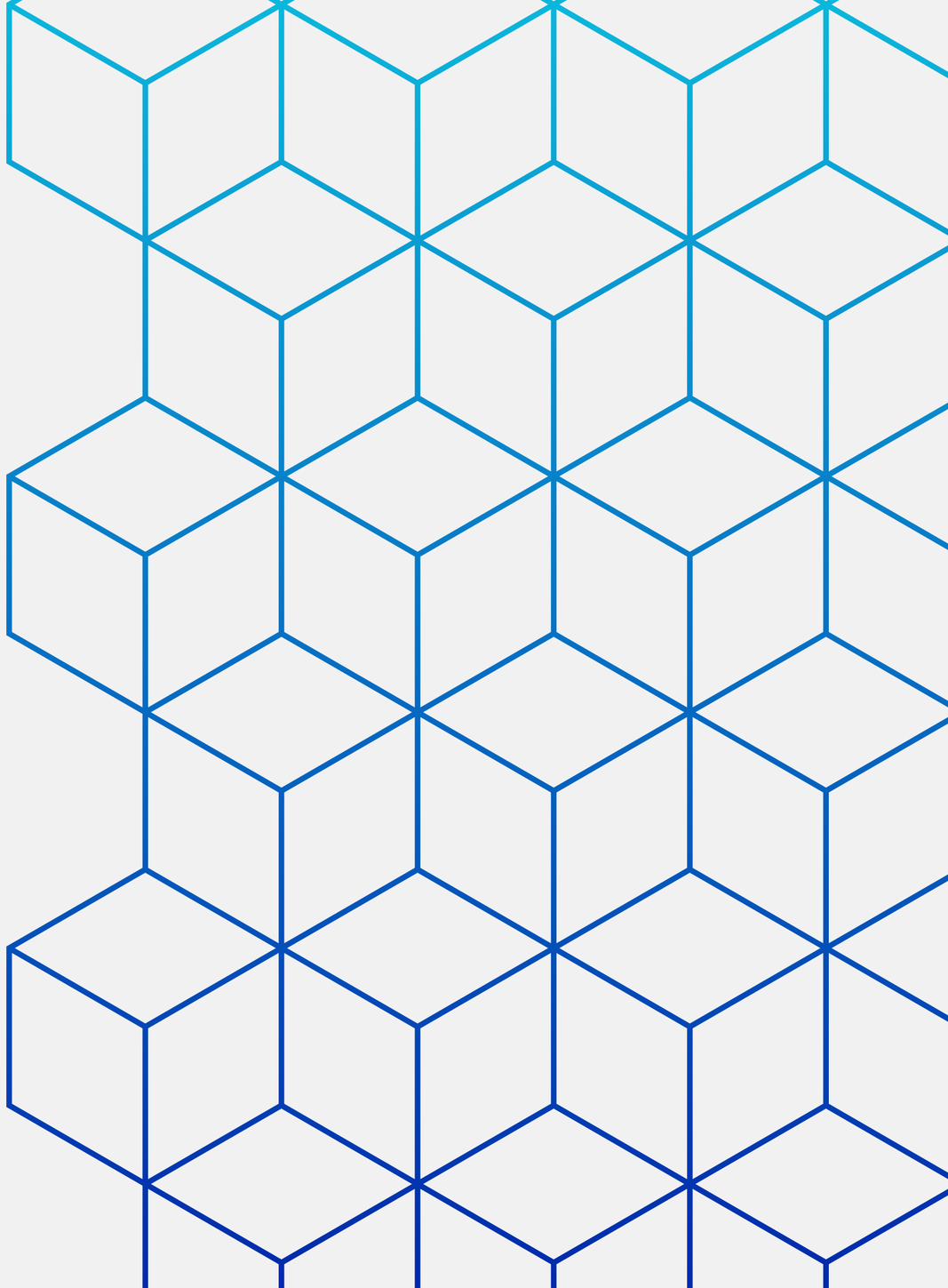
Predicting funding of a Plaksha startup based on twitter and market sentiment

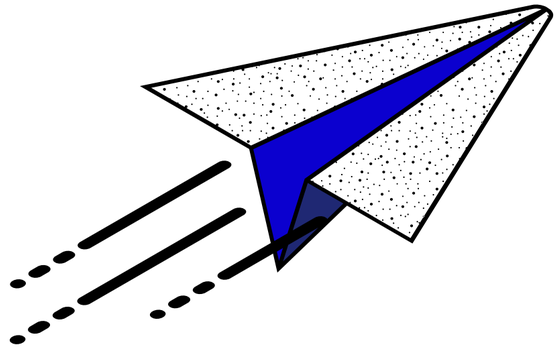
Anshul Rana, Nikita Thomas, Sauhard Sharma



Problem statement & Motivation

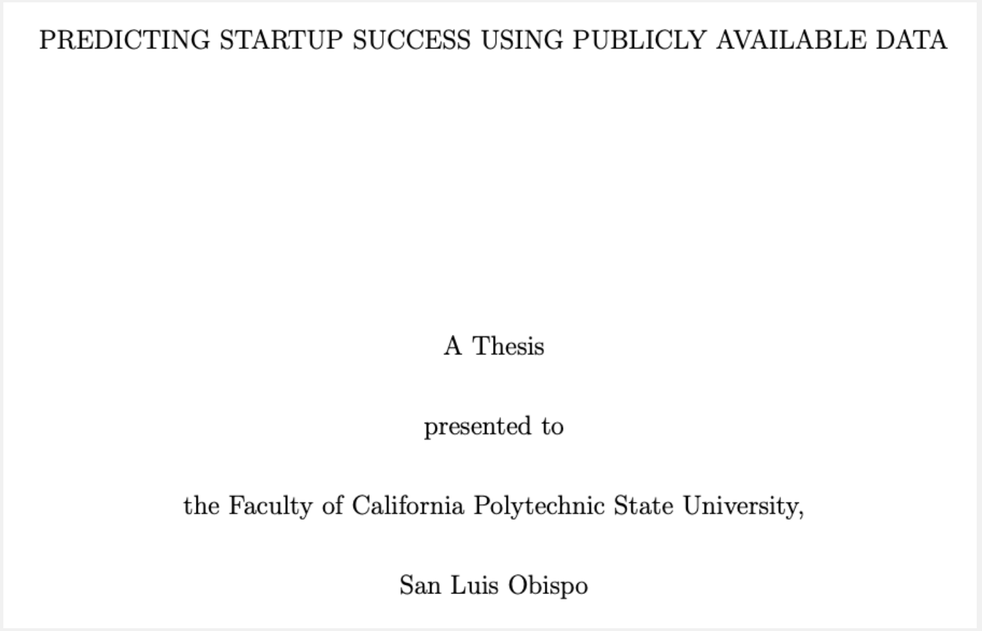
Only 1 in 12 startups obtain funding[1], this competitive nature of financing creates a need for startups to know the likelihood of receiving funding. With social media sentiment influencing investor decisions by 13.5%[2], twitter could be a viable option in gauging the markets views.

Studies have observed the effect of social media presence on the success of a startup, yet they fail to account for changes in market. A 1% increase in S&P 500 is associated with a 0.5% increase in VC funding[3]. We fill this gap by predicting which industries are facing a boom and adjusting likelihood accordingly.



1. Startup Genome
2. Journal of Business Venturing Insights
3. University of California, Berkeley

Literature Review



PREDICTING STARTUP SUCCESS USING PUBLICLY AVAILABLE DATA

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

Investigates whether the “wisdom of crowds” can be captured through online, publicly available data and combined with hard, factual data to improve model performance.

Analysed volumetric data and sentiment data from twitter & other sources.
The feature set on startup companies includes general company data, previous funding rounds, published news articles, Google Search results, and Twitter data.

Failed to account for the Market trends

Literature Review

Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy

Torben Antretter [a], Ivo Blohm [b], Dietmar Grichnik [a], Joakim Wincent [a c d]

Show more

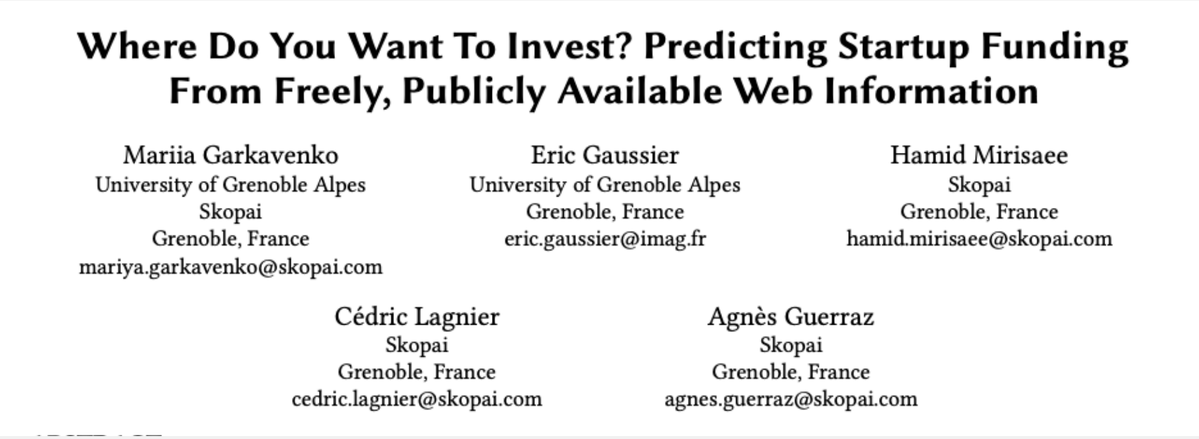
+ Add to Mendeley Share Cite

Relied solely on twitter data to predict if a startup would survive after 5 years. Made use of Random Forrest & XGBoost(Gradient Boosting) for the model. They achieved a recall of 86%, showing the importance of twitter in success prediction.

Twitter has been found to be a news source for 69% of its users

(S. Atske. News on twitter: Consumed by most users and trusted by many, Apr 2022.)

Literature Review



Where Do You Want To Invest? Predicting Startup Funding From Freely, Publicly Available Web Information

Mariia Garkavenko
University of Grenoble Alpes
Skopai
Grenoble, France
mariya.garkavenko@skopai.com

Eric Gaussier
University of Grenoble Alpes
Grenoble, France
eric.gaussier@imag.fr

Hamid Mirisaee
Skopai
Grenoble, France
hamid.mirisaee@skopai.com

Cédric Lagnier
Skopai
Grenoble, France
cedric.lagnier@skopai.com

Agnès Guerraz
Skopai
Grenoble, France
agnes.guerraz@skopai.com

Stated that features extracted from Twitter have been shown to be crucial for predictive models, especially for predicting the ability to raise funds.

While studies have evaluated the success of a startup based on the location, companies features like no. of employees etc. one of the reasons it fails includes poor product-market fit

Extracted general, financial, social network presence features (like if they have accounts with youtube, blogs etc)

Aimed to use a simple model (CatBoost) by using publicly available web data

Literature Review

Conference Paper

Predicting Startup Crowdfunding Success through Longitudinal Social Engagement Analysis

November 2017

DOI:10.1145/3132847.3132908

Conference: the 2017 ACM

Authors:

Qizhen Zhang
University of Pennsylvania

Tengyuan Ye

Meryem Essaidi

Shivani Agarwal

shows that the social engagement of a startup, like the number of tweets and the number of followers, has a **significant correlation with its success in receiving crowdfunding.** suggests that the activity of a startup on social media such as Facebook and Twitter heavily boosts the odds of receiving crowdfunds.

AnglelList (a crowdfunding investment platform), Facebook and Twitter were chosen as data sources, where AnglelList has been used to identify the startups which received crowdfunds. It could reach an accuracy of 84% in predicting crowdfunding events

Work Pre-Midsem

Funding

Scraped funding data for 3,500+ indian startups from trak.in

Preprocessed currency, time period of receiving funding, investment type

Twitter

Market

Work Pre-Midsem

Funding

Twitter

Scraped a companys tweets,
# of likes, retweets,
comments from 2015

pre-processed tweets
(stopwords, emojis, links,
usernames, non-english
languages)

performed sentiment
analysis using NLTK library

Market

Work Pre-Midsem

Funding

Twitter

Market

Scraped the Open Price, High, Low, Closing Price and Adjusted Close for companies in each sector relevant for each major

Calculating aggregate closing price to simulate index fund

Sentiment Analysis Model

Model description

- Sequential model in Keras, a linear stack of layers.
- Input layer: Dense (50 units, ReLU activation, L2 regularization).
- First Dropout layer (0.4 dropout rate).
- Second Dense layer (50 units, ReLU activation, L2 regularization).
- Second Dropout layer (0.3 dropout rate).
- Third Dense layer (50 units, ReLU activation).
- Output layer: Dense (1 unit, sigmoid activation).

Training

- 50 epochs, batch size of 32
- Incorporates L2 regularization, dropout, and early stopping for overfitting prevention.
- L2 regularization penalizes large weights.
- Dropout introduces randomness to prevent overfitting.
- Early stopping finds an optimal number of epochs.

```
loss: 0.2589 - accuracy: 0.9542 - val_loss: 0.5287 - val_accuracy: 0.8630
```

Sentiment Scores

For each tweet, the sentiment score was obtained by taking mean of all tokens in tweet.
The sentiment score for each token was obtained through the model developed previously.

```
input_indices = text_to_index("This movie is fantastic", index, max_len=10000)
print("The induvidual scores per word is : ", model.predict(vectorize(input_indices)))
print("The mean sentiment score is : ",np.mean(model.predict(vectorize(input_indices))))
```

✓ 0.0s

```
1/1 [==============================] - 0s 9ms/step
The induvidual scores per word is :  [[0.47816375]
 [0.53930837]
 [0.53032327]
 [0.4912861 ]]
1/1 [==============================] - 0s 9ms/step
The mean sentiment score is :  0.5097704
```

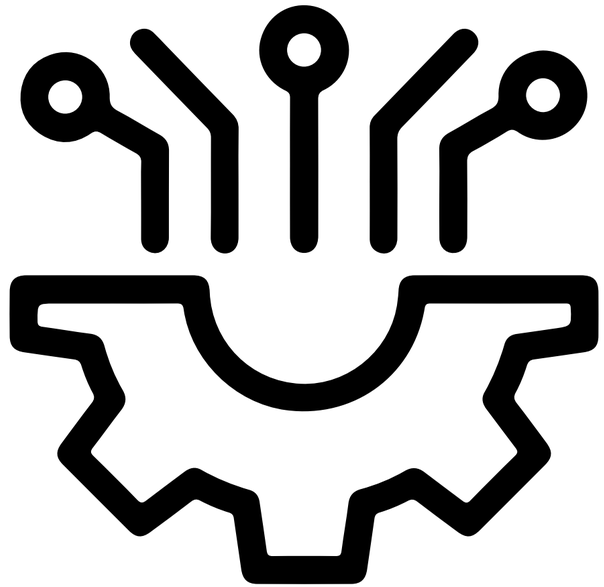
Final Model

Feature Engineering:

- Polynomial features: Polynomial features of degree 2 were created from the original features. This allows the model to capture non-linear relationships between features.

Feature Selection:

- Recursive Feature Elimination with Cross-Validation (RFECV): RFECV was used to select the most relevant features for the model. It recursively removes features and uses cross-validation to evaluate the impact on model performance.



Model Selection:

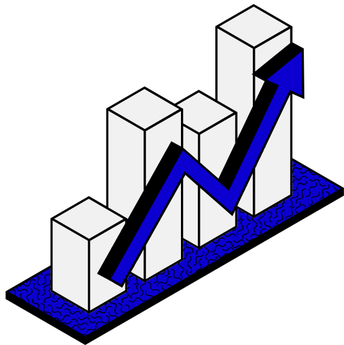
- Random Forest Classifier: A Random Forest classifier was chosen as the model. Random Forest is an ensemble learning method known for its ability to handle complex relationships and provide feature importances.

Hyperparameter Tuning:

- GridSearchCV. Grid search helps find the best combination of hyperparameters for the Random Forest model.

Model Training and Evaluation:

- Model Fitting: The Random Forest model was trained on the training data.
- Cross-Validation: Cross-validation was performed using a 5-fold cross-validation strategy to assess model generalization.
- Prediction: The model was used to make predictions on the test data.



Market Data’s role

we took a slight deviation from our initial plan

We considered GDP and FDI as indicators of market sentiment

Performance Metrics

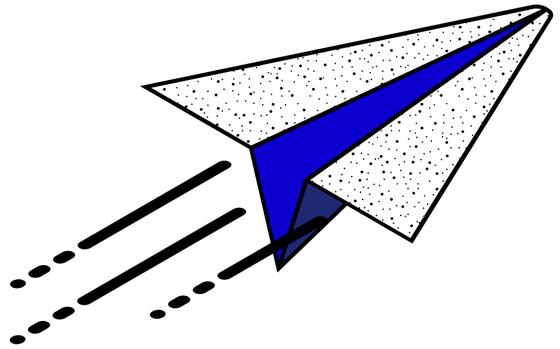
In order to test the accuracy of the model, we used f1 as a performance metric. We preferred a higher precision metric due to the riskiness of investing in a false positive company.

```
Cross-validated scores: [0.41666667 0.5        0.08333333 0.72727273 0.72727273]
Mean CV score: 0.4909090909090909
Accuracy of the Random Forest model: 0.8076923076923077
Classification Report:
              precision    recall  f1-score   support

          No       0.83      0.56      0.67         9
         Yes       0.80      0.94      0.86        17

    accuracy                           0.81        26
   macro avg       0.82      0.75      0.77        26
weighted avg       0.81      0.81      0.80        26
```

Deployability of Model

Since Plaksha University promotes a dynamic entrepreneurial culture, encouraging students to launch their own ventures, and with funding being considered as one of the key metrics for assessing the success of a startup. Our solution can help students gauge their **likelihood of receiving funding** and the **importance** that a companies social/**twitter activity** has on it allowing them to grow accordingly.

One of the challenges would be the **limited twitter data** (since the tweets would be over a small period of time) and the **changing markets.**



Limitations

- Limitation in twitter data due to paywalls faced while scraping. Although we were able to obtain funding data for over 3,500+ we were only able to find and scrape twitter handles for approximately 80 companies.

- Limitation in complexity of model due to hardware constraint on laptops. The BERT model we had additionally trained for the sentiment analysis took too long to run in order to have a high accuracy

Thank you

Funding Data EDA

Investment Trends

From yearly and quarterly trends, we can infer the growth and fluctuations in the startup investment market over the years.