Moneyball: Time Machine Data Set using GPT-3

Grace Sodunke[1] and Yigit Ihlamur[2]

[1]University of Oxford, UK
[2]Vela Partners, US

March 2023

Abstract

The aim of this project was to evaluate the potential of GPT-3, a large language model released by OpenAI, to augment data set features used to make investment decisions. Features relating to the background of founders and conditions of the startups when founded were engineered through an automated pipeline using GPT-3. Subsequently, performance of the logistic regression model, known as Moneyball, was measured using receiver operating characteristic (ROC) curves and confusion matrices. We find that language model generated features provide a promising technique for improved prediction models, resulting in True Positive rates above 75%.

1 Motivation & Overview

Vela Partners, also referred to as Ventech, focuses on developing algorithms that utilize machine learning and advanced data analysis to forecast the success of startups. Ascertaining feasible business ideas and forecasting the potential success of startups during their initial stages is a challenging task because of the limited availability of data and the absence of conventional metrics for measuring company success such as sales revenue, profit margins, and year-on-year sales growth.

Before the emergence of large language models, it was challenging to acquire a comprehensive understanding of the founders' background and the degree of competition during startup inception. Past studies have demonstrated that this information can serve as a valuable predictor of future startup success. Hence, we developed a novel dataset by utilizing GPT-3's assessment of founders' LinkedIn profiles and historical company data. This facilitated a more refined feature engineering process, resulting in a better comprehension of long-term startup performance.

In Section 2, we present the architecture of the recently created dataset. Subsequently, in Section 3, we provide a detailed discussion of each feature, including its generation using GPT-3 and the outcomes obtained. Section 4 showcases the predictive modeling results derived through logistic regression. This section also delves deeper into model performance evaluation by analyzing ROC curves that determine the trade-offs between true positives and false negatives at varying decision thresholds. Finally, in Section 5, we highlight the potential directions for this project, focusing on prompt engineering.

2 Structure of the data set

The data set consists of three newly generated features, each created by prompting GPT-3 to make conclusions from existing data. The columns in the data set are as follows:

- **org_uuid** The strings that represent unique identifiers for the companies included in the initial spreadsheet of both successful and unsuccessful companies.
- **org_name** The company names that correspond to each uuid.
- **how_many_founders_studied_at_top_tier_institutions** This column stores the number of founders that studied at top institutions for each startup, according to analysis by GPT-3.
- **how_many_founders_worked_at_top_tier_companies** This column counts the number of founders that worked at a top company during the years that the company was relevant, based on evaluation by GPT-3.
- **level_of_competition_in_founding_year** This column describes whether the startup had high (1) or low (0) competition when the it was founded, based on market analysis by GPT-3.

3 Data exploration & feature engineering

3.1 How many founders studied at top institutions

Our initial exploratory analysis focuses on the reputation of institutions that the founders of successful and unsuccessful companies studied at. For each founder in the data set, GPT-3 was queried with an identical prompt to give a 'True' or 'False' value. Context was provided to improve accuracy: *"You are a Gartner market research analyst and you should share your opinion as 'True' or 'False'"*. Where education data isn't available, the value 'None' is stored. The total for each startup is then stored in the new data set.

3.2 How many founders studied at top companies

A similar approach to the first feature was taken to total the number of founders that worked at leading companies. It is also important to consider the years each founder worked there, for example working at Yahoo in 2004 is a great signal but not now. Identical context to the previous feature was provided when prompting GPT-3, returning either 'True' or 'False' or 'Not sure' as a placeholder value. The total for each startup is stored in the final data set.

3.3 Level of competition when startup was founded

For this final feature, we use GPT-3 to evaluate whether each startup had many competitors when it was founded. This required sufficient context about the market the startup is in, as well as conditions in the founding year. In each prompt, a company description is provided to the model, as well as the founding date, and either 'Low competition' or 'High competition' is returned.

4 Predictive modelling

In order to measure prediction accuracy, we use a classification model to predict the success of companies using the following features, modelled as a pandas data frame:

- Number of founders that studied at top tier institutions
- Number of founders that worked at top tier companies
- Level of market competition when the startup was founded.

4.1 Logistic regression

To classify the data with logistic regression, we represent the data set as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P \in \mathbb{R}^3$ where P represents the size of the data set. The objective of the model is then to map these vectors of inputs to their correct ground truth labels $y_1, \ldots, y_P \in \{0, 1\}$.

Metric	Model Result
True Positive Rate	.778
False Negative Rate	.222
True Negative Rate	.622
False Positive Rate	.378
Precision	.333

Table 1: Table of logistic regression performance metrics

Our logistic regression model achieves this goal by mapping inputs through a weight vector $\boldsymbol{w} \in \mathbb{R}^3$. The prediction is then generated as $\hat{y} = \sigma\left(\mathbf{w}^T\mathbf{x}\right)$ where $\sigma()$ represents the sigmoid activation function whose range is $(0, 1)$. To find the best classification performance, we then minimised the binary cross-entropy loss between predictions and labels as shown in Equation 1. We used 80% of data for model training and we used 20% of the data for model testing to avoid over-fitting on the data set.

$$L_{BCE} = \frac{1}{P}\sum_{i=1}^{P} -\left(y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right) \tag{1}$$

We utilized the L-BFGS algorithm to minimize the objective function stated in our problem specification. With this approach, we achieved an accuracy rate of about 65%, significantly surpassing a random baseline model. Key performance metrics are presented in Table 1. The precision achieved was 33%, which was calculated using Equation 2.

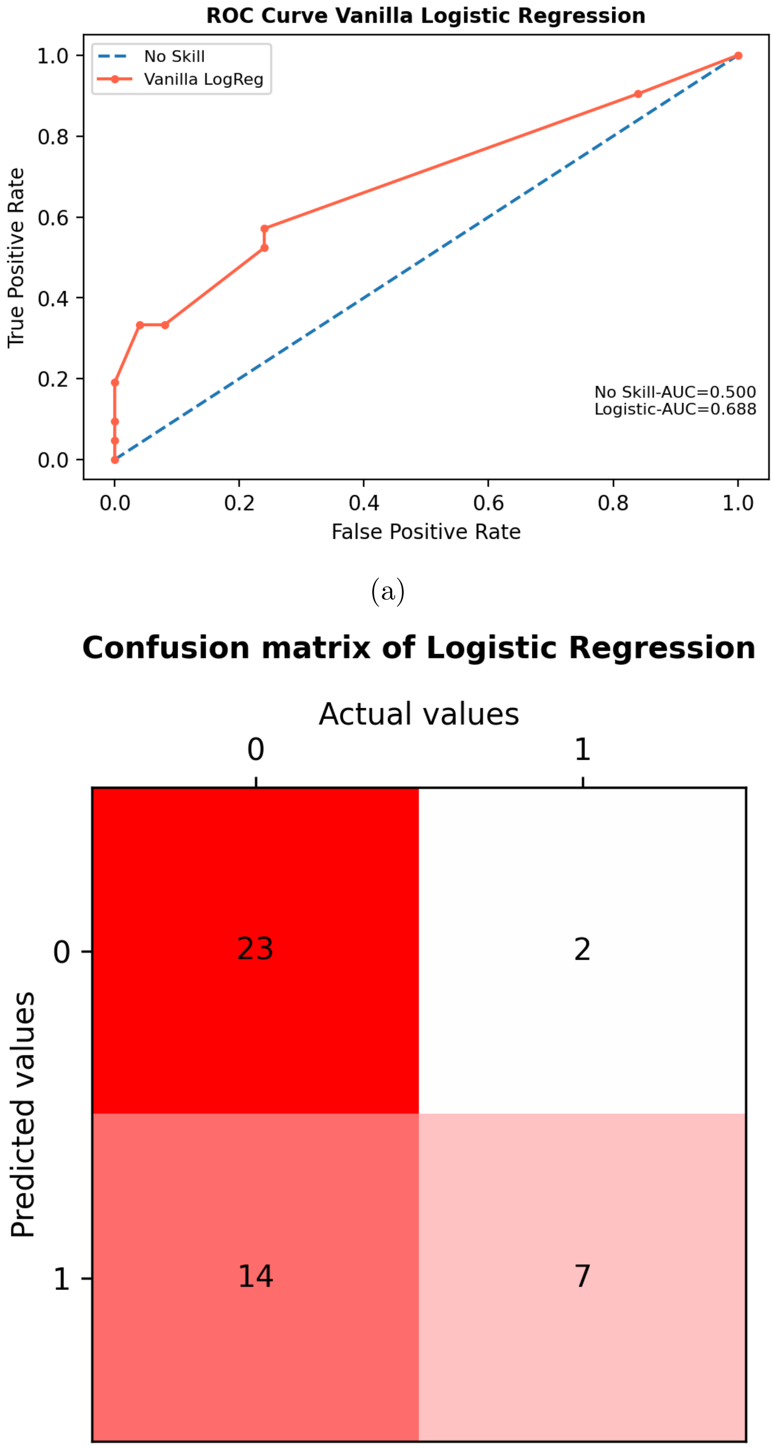
$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{ False Positives (FP)}} \tag{2}$$

For further analysis, we plot a receiver operating characteristic (ROC) curve that displays the trade-off between True and False Positives as a function of decision thresholds applied to model output probabilities. The results are shown in Figure 1a. In addition, we plot a confusion matrix that shows classification performance on the test set, in order to visually analyse error metrics (Figure 1b).

5 Future directions

5.1 Proof of concept

This analysis established a correlation between GPT-3 generated features on the background of founders and future startup success. As LLMs increase in capability, this approach will become more promising and enable a wide variety of features to be engineered. A current constraint is the rate limit of requests to the API of 25,000 tokens/min, which meant that about 200 companies could be analysed in



	0	1
0	23	2
1	14	7

(a)

(b)

Figure 1: (a) ROC curve for logistic regression. The blue dotted line represents a baseline classifier that predicts all companies as unsuccessful. (b) Confusion matrix for logistic regression on the test set

the time limit for this project. Utilizing more data points would lead to a much more complex and accurate analysis of the relative success of different startups, and potentially improve model accuracy.

5.2 Prompt engineering

A key future direction for this project is to design improved prompts for GPT-3 / GPT-4 that gives more detailed, precise and accurate results. Some considerations include:

- Determining the right amount of context that gives precise results
- Further checks to ensure the model isn't hallucinating facts
- Parsing through output to check whether out of range data is returned, and how to interpret in each case.

Furthermore, additional exploratory analysis with the data set can involve:

- Additional features to refine the background picture of each founder: how many founders are professors, how many founders have relevant work experience in their field, etc
- Classification using different neural network architectures and a comparison of True Positive Rates
- Measuring the effect of hyper-parameters on the behaviour of each model.