Predicting valuation of ‘unicorn’ companies

Ifath Chowdhury, Daniel Fernandes, MD Moijul Islam

Abstract

By using artificial intelligence techniques, this project planned to analyse the data of over 1000 companies to eventually predict the valuation of another company given certain variables, for example the number of investors. To solve this, we used various artificial intelligence models to work on a dataset that we discovered that matched our requirements. The results that were gathered fell in line with expectations – more investors and funding allow for a company to earn a higher valuation.

1 Introduction

A “Unicorn” company is a privately held company that has a market capitalization of over 1 billion dollars. Valuation (the company’s monetary worth) is determined by investors and venture capitalists, who judge how a company will grow over time. For this reason, a company’s valuation often takes a long time to be properly forecasted. For new start-up companies, it would be advantageous to know what it exactly it takes to get a high valuation, for example how many investors to have. For this reason, this project planned to use artificial intelligence to aid this problem.

2 Related Work

There has been some previous work on building valuation prediction models for companies using Machine Learning algorithms. One example of a machine learning model to predict valuation gathers its dataset by gaining access to anonymised key performance indicators and mapping that data to funding stages using various APIs (flo.tausend, 2020). This is a key difference between our project and this example one because we have decided to borrow a publicly available dataset from Kaggle. This is because the main focus of this project is to use and learn more about different machine learning paradigms, as opposed to gathering data. For this reason, we have chosen to borrow a publicly available dataset online to put more focus on building the machine learning models that we plan to use.

3 Data

3.1 Overview

The dataset used in this experiment comes from Kaggle. It contains information about 1053 different real companies from all over the world. The dataset documents several features such as the founding year of these companies, their financial state, etc. It is for this reason that the dataset was chosen. As previously mentioned, a company’s valuation is predicted over time by observing these factors listed in the dataset – industry, financial state, etc. Therefore, we found it appropriate to run this dataset through an artificial intelligence model, that way we could gain results from real data without having to spend an excessive amount of time researching companies and assembling the dataset ourselves.

3.2 Pre-processing

In the dataset, there were many values that were null and even more values that were in the format of categorical data. The algorithms that were planned to be used on the data to achieve our results were unable to process categorical data, so it was important when we imported the dataset that any column containing categorical data was replaced with multiple Boolean columns to represent the various categories. These columns had to contain numbers so that they could be processed in the models. Null values will skew the accuracy of any artificial intelligence model, so it was imperative that the null values were dealt with. To treat the issue of null values hidden within the dataset, we removed any rows containing null values or special characters that couldn’t be processed by the algorithm. We also modified the Data value to Age so that the model could use this data efficiently.

4 Methods

For this project, we decided to use 3 different machine learning algorithms. Support Vector Machines (SVMs), Linear Regression, and Artificial Neural Networks (ANN).

4.1 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis. Generally, they are used as a classifying method. They can either be a binary classifier, or by using one of two major strategies, “one against all” and “one against one”, they can be used as a multi-class classification model that splits the problem into several binary class sub-problems. The problem we have decided to tackle, however, is not a classification problem. Our task is to predict the valuation of a company, which is a variable number that can change over time. For this reason, it cannot be classified into different classes. Therefore, we are unable to use any classification algorithms as they will not be able to process the data, which would be continuous. This makes our problem a regression problem, so it is therefore imperative that a Regression algorithm was used.

Our problem was a regression problem, so we used an SVR (Support Vector Regression) algorithm. As stated in an article on Towards Data Science, “Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as, the SVMs. Support. The basic idea behind SVR is to find the best fit line.” (Ashwin Raj,2020). SVR measures how well a model follows the trend of data by showing a regression score. The maximum value for a regression score is 1 and the minimum value is -1. A positive value means that the model follows the trend of data strongly meanwhile a negative value means that the model doesn’t follow the trend of data.

4.1 Linear Regression

Another model used to achieve results was a Linear Regression model. This model is an excellent choice for this dataset because the data values in the valuation columns are all continuous data that is not categorized into two or three distinct categories. Because of this fact, any classification algorithm would not be able to function properly as those types of algorithms are designed to handle only data which can be assorted into various categories. Regression algorithms can take on continuous data and process them to form an output; in the case of Linear Regression, the model receives the data input and analyses it to form a linear relationship between the data values. We are then able to use this linear relationship to create a graph and plot a linear line indicating the relationship between the data values on this graph. This way, the output of the model is visualized, and it is easier to observe the relationship between two different pieces of data.

The reason that a linear regression model was used instead of something else like a gradient descent model, is because it was decided that a linear regression model would be more appropriate for the situation. As a group we decided that a method of visualising the effects of various factors, such as the number of investors and the number of deal terms, would be more beneficial and allow us to see with more ease how a company can strive to achieve a higher valuation.

4.1 Artificial Neural Network (ANN)

An artificial neural network (ANN) is a computer model made up of many processing components that accept inputs and outputs based on their activation functions. that are designed to recognise patterns and are loosely modelled like the human brain. They use machine perception to understand sensory inputs, categorising or clustering raw data. All real-world data, whether images, sound, text, or time series, must be translated into the patterns they recognise, which are numerical and encoded in vectors. For this Machine Learning project, we are going to use Keras Sequential Model. Keras is a Python-based deep learning API that runs on top of the TensorFlow machine learning platform.

5 Experiments

**SVM**

For the SVR algorithm we used 3 kernels to see which gave the best score. Out of the three the one with the best score was the RBF kernel while the poly kernel gave the worst score.

Text, letter

Description automatically generated

Linear kernel - for the linear kernel we used c = 10 and gamma = 1000 for the hyperparameters. The poly model gave a regression score of -0.0038. The regression score is negative which means that this model doesn’t follow the trend of data.

Text, letter

Description automatically generated

Poly kernel - for the poly kernel we used c = 10 and gamma = 1000 for the hyperparameters. The poly model gave a regression score of -0.022. The regression score is negative which means that this model also doesn’t follow the trend of data.

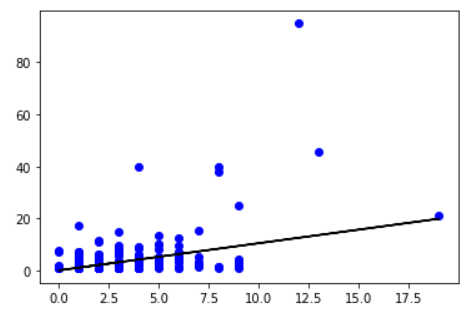
Text, letter

Description automatically generated

RBF kernel - for the RBF kernel we used c = 10 and gamma = 1000 for the hyperparameters. The poly model gave a regression score of 0.0005. The is the only model giving a positive regression score, however the score is so low that it is hard to say if the model follows the trend of the dataset.

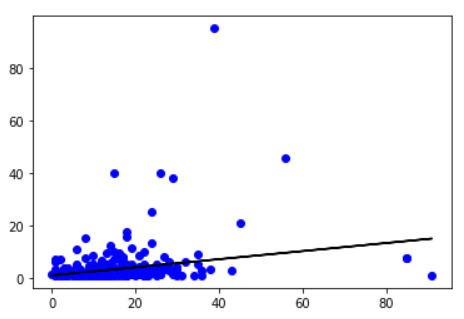
**Linear regression**

For the Linear Regression model, we sought to create a graph to visualize the results of the model in graphical form. To extensively analyse the effects of various factors on a company’s valuation, we intended to plot these different factors as input variables against the output valuation variable. The first value we plotted against the Valuation was the Deal Terms. Deal terms are a summary of the terms of whatever deals a company has made with any other company. In the context of this dataset, it is assumed that the Deal Terms column indicates how many term sheets the company has acquired. In other terms, Deal Terms represents how many deals the company has made. Before experimentation, it was predicted that the more deal terms a company has made, the higher their valuation is estimated to become. The reason for this prediction is that with more deal terms, a company can amass more connections to other companies, allowing them to create more deals that could potentially be more beneficial to the company. This hypothesis would prove to be correct when the graph of the linear regression algorithm was generated.



The graph above plots the results of the linear regression. Along the X-axis is the number of deal terms, while along the Y-axis is the valuation (measured in billions of dollars). As the graph indicates, the higher the number of deal terms, the higher the valuation will be. This falls perfectly in line with the hypothesis that we created – that more deal terms would increase the valuation.

The other value I ran through the linear regression model was the number of investors. Investors are people who fund a company so that they can eventually return a profit. These investors, while gaining their own profit in the long term, give companies a higher fund to spend on new things. This new funding from the investors could earn companies a higher valuation, so it was our hypothesis that the more investors are involved in a company, the higher the valuation would become. The below graph plots the line of linear regression calculated by the model. The X-axis is the input number of investors, while the Y-axis is the output valuation number measured in billions of dollars. The graph clearly proves our hypothesis – In the graph it is shown that as the number of investors along the bottom increases, the valuation would also increase along with it, thus proving our hypothesis.



**ANN**

To get The best possible out come from our Artificial Neural Network we used Keras Sequential Model. This Model works as finding an optimal set of weights (and biases) where each single weight and bias is used for computation between layers. The parameter set of a network is composed of all weights and biases in a model. Here “optimal” means that they have the minimum loss defined by a model and the algorithm tries to update the weights repeatedly.

The entire training procedure can be summarized to repeat following loop for the fixed number of iterations, defined by epochs and sample size.

* Draw a fixed size of mini-batch training sample (batch size) and this mini-batch sample enters the first layer of the network.
* Along the entire network, transform input data to high-level features by multiplying weights and going through activation functions.
* Calculate the mean loss using true and estimated values of output on the batch.
* Compute a gradient of the loss to each weight, on the way back from output layer to input layer, and update the weights in the direction of decreasing the loss.
* Draw a next mini-batch sample and iterate.

We used 21 input Layers and 15 hidden layers and 1 output layer to maximise the accuracy. We used loss='mean\_squared\_error', because It avoids and rectifies vanishing gradient problems. And with all hyperparameter tuning our ANN Model produces results with 30% accuracy.

6 Conclusion

From the results we can conclude that the SVR models do not produce strong results that follow a trend in the dataset even though we thought they would. This could have been caused by various factors, one of them being the dataset. Most of the information in the dataset was written in categorical format. As mentioned previously, categorical data cannot be analysed by a machine learning algorithm because they are only able to work with numbers. Due to this, we were forced to change most of the dataset and edit in many columns containing values of 1 or 0 to represent the many categories. In the end, there were an extensive number of columns in the dataset. This may have confused the model, resulting in the failure of the SVR model. Another factor in the failure of SVR is our lack of knowledge on the subject. We did not have much experience in creating a regression model using SVMs, so we had to learn how to create an SVR model as we took on this project. As a result, our inexperience with SVRs made it difficult to create the model and produce good results. We also didn’t have experience creating an ANN model which took lots of time to learn.

However, the linear regression model was still able to gather results. The graphs produced by the model helped us understand more about companies and how they would go about increasing their valuation. It was discovered that by increasing the number of investors and the number of deal terms, a company would be able to garner a higher valuation. This understanding is very beneficial for future start-up companies because the linear regression model and the graphs it created provides these new start-up companies with a goal to strive for if they want to be successful.

Artificial Neural Network produces the best among all three models but still not usable in real world problem. with 30 percent accuracy we can say the model follows the Trend of data, but the rate is very low. We tuned various hyperparameter but we came to a conclusion that for this particular dataset this is not the model to use. One great option will be “Gradient Boosting Regressor” as at the end of ANN Model we tried a “GradientBoostingRegressor” model which produced 74% accuracy which is can use used in real world to predict Company Valuation

To improve on this project in the future, we would choose a different dataset, or perhaps even create our own dataset by gathering data using various technologies. As previously mentioned, the dataset was a large hinderance in this project because we did not notice that there were many null values and many other values that had to be changed from a categorical format to a numerical format. This made the data harder to use in the model and it also made it harder to plot the data on a graph. Another factor we hope to improve upon in future is our lack of knowledge and experience in certain machine learning paradigms, in particular the SVR algorithm. Our lack of knowledge and experience made it difficult to fully utilise the capabilities of Support Vector Machines. As a result, the regression scores that we achieved were quite poor, resulting in us not being able to extract much knowledge about unicorn companies and valuations from the Support Vector Regressor algorithm.

References

Ashwin, R. (2020) Unlocking the True Power of Support Vector Regression, *Towards Data Science* [online], 1, pp.1. Available from:<https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0>[Accessed 3 May 2022]

Flo.tausend (2020) Applied Machine Learning Models for Improved Startup Valuation, *Towards Data Science* [online], 1, pp.1-3. Available from: <https://towardsdatascience.com/applied-machine-learning-models-for-improved-startup-valuation-a7b180fee127> [Accessed 4 May 2022]