START UP ANALYSIS

Personal Details

Name: Mandla Dharani

University/College: Malla Reddy Engineering College for women

Email: mandladharanireddy@gmail.com(connect with me on linkedin)

Telephone:(+91) 8688437406

Country of Residence: India

Timezone: IST (GMT + 05:30)

Primary Language: English

I am a second year undergraduate student pursuing B.Tech. in Computer science and engineering at Malla Reddy Engineering college for women, Hyderabad.
I am keenly interested to develop real-time projects using different technologies like AI and ML. I had undergone multiple courses and worked on different projects to polish my skills. I am Skilled in Python, C, Data Structures, Algorithms, Flask, HTML, and Problem Solving. I have participated in 8 hackathons, solving different problems, and exploring new fields.
If I am selected, I shall be able to work around 40 hrs a week on the project, though am open to putting in more effort if the work requires.

Personal Projects:

1.Skin Disease prediction using CNN

2. Heart Disease prediction using ML

3.Spam Classifier

4. Aadhar Card Management using C

Project:

Abstract:

Determining the valuation of an early-stage Startup is in most cases very challenging due limited historical data, little to no existing revenues, market uncertainty and many more. Traditional valuation techniques, such as Discounted Cash Flow (DCF) or Multiples (CCA), therefore often lead to inappropriate results. On the other hand, alternative valuation methods remain subject to an individual's subjective assessment and a black box for others.Therefore, the underlying study leverages machine learning algorithms to predict fair, data-driven and comprehensible startup valuations. Three different data sources are merged and applied to three regression models. The regression models' performance is compared on defined measurements and the final outcome is a continuous numeric value — the predicted startup valuation.

Technical Details:

In order to build a machine learning model and get significant results, it is essential to have a solid database. In this case the final analytical data set is a combination of public and private, anonymised data sources. Some of the steps attain the model are:

1.Data Preprocessing:

Most of the effort had to be spent merging the different data sources together and cleaning them.In the very beginning an overview of the different tables and information was generated with a simple database client DBEAVER. After that relevant information was extracted, transformed and loaded mainly with the help of Python programming.

2.Normalisation & Scaling:

An important step to prepare the dataset for our machine learning algorithms was normalisation. The idea is to bring the numeric values in a common scale. This is necessary as we have headcount and revenue as two features. While revenue may take a numeric value with 6 digits, it is very unlikely that the headcount has values above 100. The features have a very different range and revenue would influence our model more only due to its higher values.The MinMaxScaler is very sensitive to outliers. If your data has some severe outliers, please take care of that beforehand. The range of the normalized data is usually between zero and one as seen in the code above.

3.Applying Regression Models:

Within the project there were three regression models applied to the dataset. They were chosen according to their characteristics and usual purpose of application. Each regression model was evaluated with the help of three indicators: Mean Absolute Error, Root Mean Squared Error and R-Squared.

4.Linear Regression:

A linear regression describes the process of finding a straight line that is as close as possible to the given data points. The model then tries to minimize the squared errors. As it is assumed that the company valuation increases with better performance gradually, the linear regression is applied to the given dataset. Furthermore, as there are multiple numerous features given in the dataset for this project, it is not a simple but a multiple linear regression. As this project is set up with Python, the module statsmodels is used to perform regression and evaluate the predictive results. While there are several classes available, the ordinary least squares method was chosen as it is mostly used for linear relationships in the training data.

5.Poly Regression:

The polynomial regression is performed by two steps. First of all, the data has to be transformed into a polynomial. A degree of one would create a simple linear regression. To create a polynomial regression the degree must be higher than one and chosen carefully, because if the value is chosen too high, the model may be overfitted.

6.Neural Network Regression:

Neural networks are well known for outperforming traditional learning algorithms, especially when it comes to image recognition or classification problems. Applying neural networks to regression problems is not found that often in existing literature. This probably results from the fact that a neural network for regression may be considered as over engineered for a regression, which is in many cases simpler and does not require that much computational power. But on the other hand, it is also argued that deep neural networks mimic the human brain and therefore perform better at complex tasks. In conclusion to the fact that the prevailing dataset is very complex and has many features, the idea is to apply a neural network for performance improvements competing with the previously applied traditional models.

7.Evaluating Model Performance:

R2, MAE and RMSE are used to determine the overall performance of each model and compare them. , interpreting the coefficients helps to find the drivers of a model. Before interpreting the results, it can be seen that the polynomial regression was outperforming the Neural Network Regression and the Linear Regression.

Conclusion:

The underlying applied machine learning model takes only structured performance data into consideration. Important factors that usually drive a startup valuation enormously, such as management team experience, technological advantage, operating market growth and many more, are not yet considered. This also explains the moderate r-squared value. It needs to be stated that there is no standard guideline defined for the level of acceptance when it comes to r-squared - Especially when the model mimics human behaviour.