Will a start-up succeed or fail?

Our word of the day is “Startup”.

The term “startup” has been bandied around with increasing frequency over the past few years to describe scrappy young ventures. Start-up means “the act or an instance of setting in operation or motion” or “a fledgling business enterprise”.

Startups are usually small and initially financed and operated by a handful of founders or one individual. These companies offer a product or service that is not currently being offered elsewhere in the market, or that the founders believe is being offered in an inferior manner.

Startups have a high failure rate, would-be investors should consider not just the idea, but the management team's experience. Potential investors should also not invest money that they cannot afford to lose in startups. Finally, investors should develop an exit strategy, because until they sell, any profits exist only on paper.

Logistic Regression on startup dataset

1. Problem Statement

The goal is to **predict** whether a start-up will succeed or fail.

2. Data Loading and Description

The dataset consists of the information about the startup company’s locations, founders. Investment and many more details which will be used for making a perfect startup model.

The dataset comprises of **234** observations of **51** columns. Below is a table showing the names of all the columns and their description.

Variables	Description
Dependent	0- Company Failed, 1- Company is successful
Company_Location	Location of Headquarter of company
Company_raising_fund	If company has been raising funds recently

Company_Industry_count	Number of Industry company is catering to
Company_mobile_app	If company has mobile application
Company_investor_count_seed	Number of investor in seed funding
Company_investor_count_Angel_VC	Number of investor in Angel or VC funding
Company_cofounders_count	Number of cofounders
Company_advisors_count	Number of company advisors
Company_senior_team_count	Number of top management employees
Company_top_Angel_VC_funding	If company has been funded by top Angel or VC funds
Company_repeat_investors_count	Number of investors who invested 2nd or more times
Founders_top_company_experience	If founder or co-founder has worked top tech companies(Google, Microsoft, Amazon, FB, Twitter, IBM, yahoo and Oracle)
Founders_previous_company_employee_count	Average employee size of previous company of founders
Founders_startup_experience	If founders or co-founders have previously worked in a startup
Founders_big_5_experience	If founders or co-founders have previously worked in a any of big 5 consulting firm
Company_business_model	If company business model is B2B, B2C or both
Founders_experience	Average experience of founders
Founders_global_exposure	If founders have global exposure (worked outside the country of their education)
Founders_Industry_exposure	Industry exposure of founders
Founder_education	Founder highest education category
Founder_university_quality	Categorization of Ranking of Founder educational institute
Founders_Popularity	If founder is renown in professional circle
Founders_fortune1000_company_score	Score of experience of working in fortune 1000 company
Founders_profile_similarity	Profile similarity score of founders
Founders_publications	Number of publications by founders

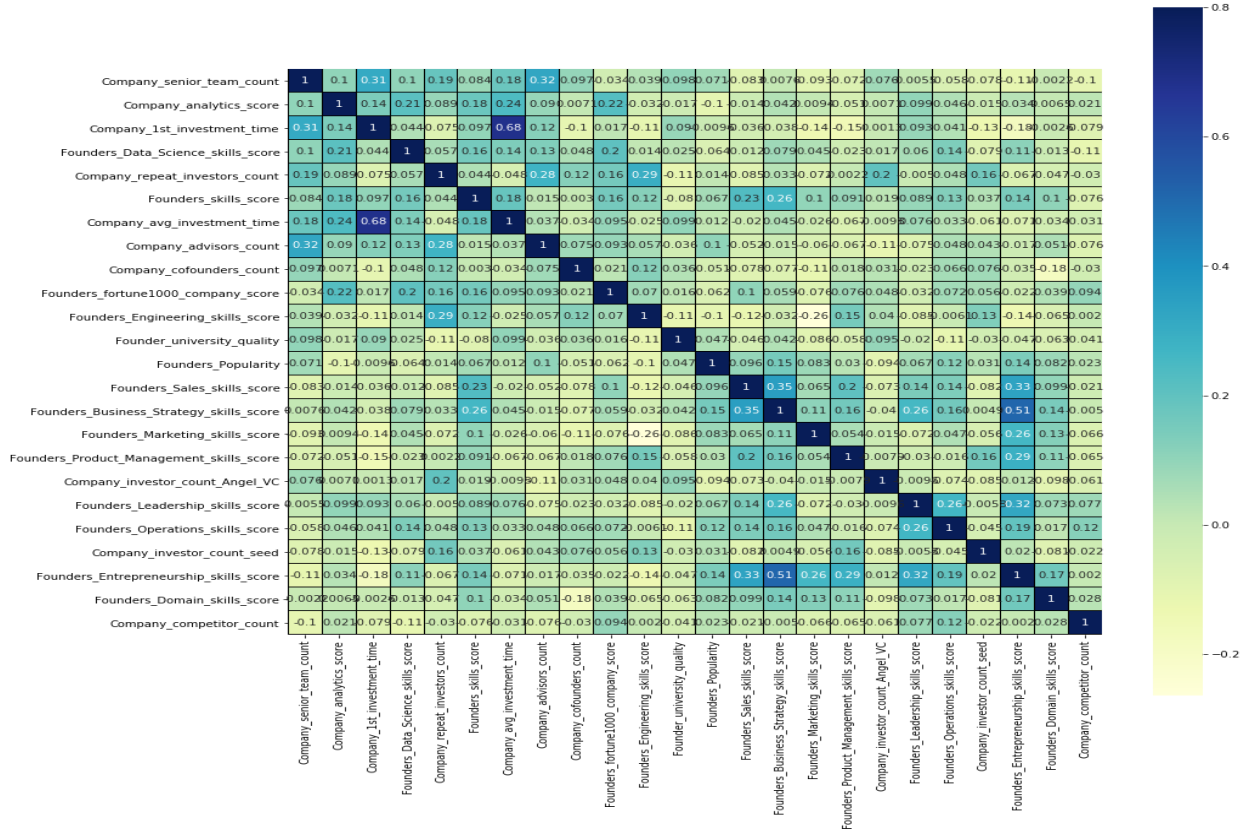
Founders_skills_score	Overall skill score of founders
Founders_Entrepreneurship_skills_score	Entrepreneurship skill score of founders
Founders_Operations_skills_score	Operational skill score of founders
Founders_Engineering_skills_score	Engineering skill score of founders
Founders_Marketing_skills_score	Marketing skill score of founders
Founders_Leadership_skills_score	Leadership skill score of founders
Founders_Data_Science_skills_score	Data science skill score of founders
Founders_Business_Strategy_skills_score	Business strategy skill score of founders
Founders_Product_Management_skills_score	Product management skill score of founders
Founders_Sales_skills_score	Sales skill score of founders
Founders_Domain_skills_score	Domain knowledge score of founders
Company_incubation_investor	If company has been funded by top incubators
Company_competitor_count	Number of direct competitors of company
Company_1st_investment_time	Time in months to get 1st investment
Company_avg_investment_time	Average time in months between multiple rounds of investment
Company_crowdsourcing	If company is crowd sourcing related
Company_crowdfunding	If company is crowd funding related
Company_big_data	If company is big data related
Company_analytics_score	Analytics score of company (level of analytics they are doing)
Company_Product_or_service	If company is product or service based
Company_subscription_offering	If company is offering subscription
Founder_highest_degree_type	Founder highest education type
Company_difficulty_obtaining_workforce	level of difficulty in obtaining workforce
Company_Founder_Patent	If company or founders have patent

3. Preprocessing the data

- Check the columns present in the Dictionary data
- Check the columns present in the train data
- Check the shape of Dataset
- Check the shape of test Data
- Combining the train and test dataset.

4. Establishing a figure between all the features using a heatmap.

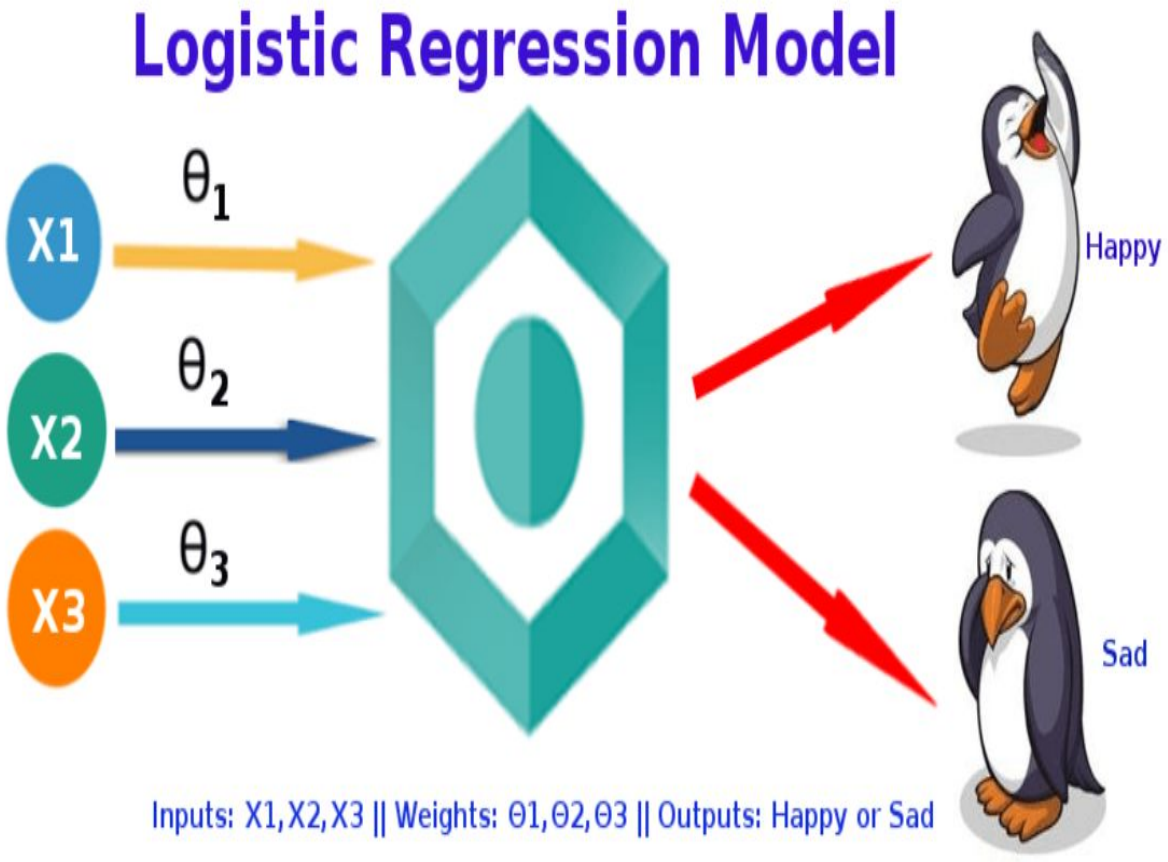
- Preparing X and y using pandas
- Splitting X and y into training and test datasets.
- Check the shape of the X and y of train dataset.
- Check the shape of X and y of test dataset.



5. Logistic Regression

5.1 Introduction to Logistic Regression:

Logistic regression is a technique used for solving the **classification problem**.
And Classification is nothing but a problem of **identifying** to which of a set of **categories** a new observation belongs, on the basis of *training dataset* containing observations (or instances) whose categorical membership is known.
For example to predict:
Whether an email is a spam (1) or not (0) or,
Whether the tumor is malignant (1) or not (0)
Below is the pictorial representation of a basic logistic regression model to classify a set of images into *happy or sad.*



5.2 Logistic regression in scikit-learn.

To apply any machine learning algorithm on your dataset, basically there are 4 steps:

1. Load the algorithm
2. Instantiate and Fit the model to the training dataset
3. Prediction on the test set
4. Calculating the accuracy of the model

The code block given below shows how these steps are carried out:

```
from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression()

logreg.fit(X_train, y_train)

accuracy_score(y_test,y_pred_test))
```

6. Model evaluation

Error is the *deviation* of the values *predicted* by the model with the *true* values.
We will use the **accuracy score __ and __confusion matrix** for evaluation.

- Model Evaluation using accuracy classification score
- Model Evaluation using the confusion matrix

7. Results

After building the initial models, I picked the most promising ones and tuned their various hyperparameters to find the one with the most predictive power. Considering the use case, interpretability was also a priority for me, so I was more focused on logistic regression and tree-based models, which are highly interpretable.

The chosen model is logistic regression which calculates a company's chances of success as a function of the features described earlier.