Blueprint & Strategy mapping for Crowd funding Analysis

Prepared by, Nikhilesh Pattanayak

Date: 02/09/2021

Company: Technocolab

Date	Version	Written By	Reviewed By
02/09/2021	1.0	Nikhilesh Pattanayak	

Contents

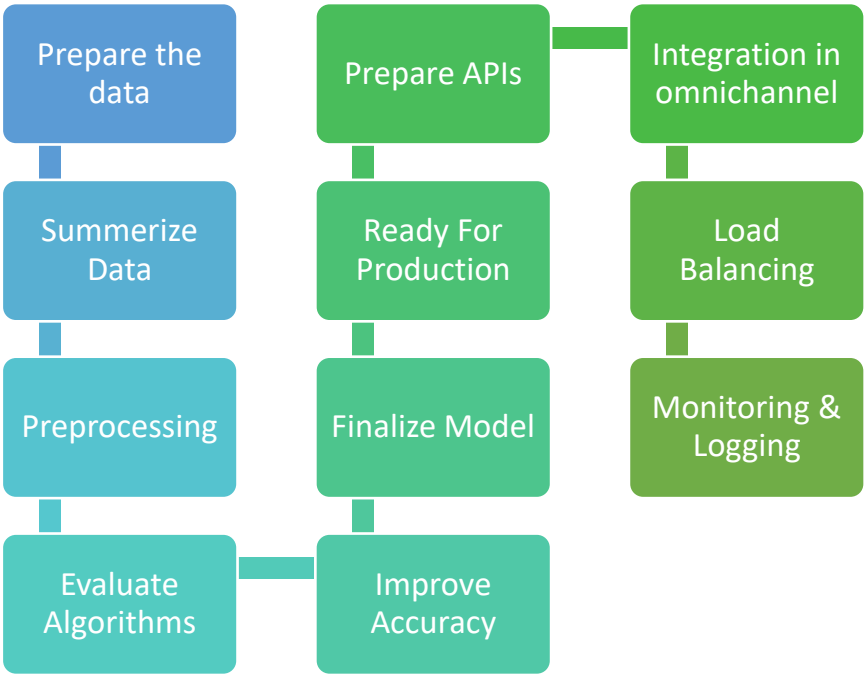
Introduction

The data set consists of last 3 years data for crowd funding events where we have some valid data to process in our codebase.

We will go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling steps that can help to improve the model performance.

Some common Python Libraries that would be need to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib



1. Prepare the data

1.1 Datasets

We have downloaded the dataset from https://webrobots.io/kickstarter-datasets/ and have downloaded sets of .csv files for our coding exercises.

1.2 Prepare the Datasets

After receiving the datasets, we will now make it ready for our operation where we can read those data sets in one instance. There are multiple ways to achieve it,

a. Convert all excel sheets to one excel sheet (recommended)
b. Read all excel sheet in a loop and convert to one Data Frame in Pandas

2. Summarize the Data

After analysing the dataset, we have found that all the labels have been given and data is understandable. Hence, we come to a conclusion that we need to **use Supervised Machine Learning Models**.

2.1 Datatype selections

Find out which columns are integer/float/objects to understand more about data

2.2 Describe the data

Find out standard deviations, mean and median and many more of the data which we need to analyse.

3. Pre-processing

data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

3.1 Errors

There is major 3 types (taking these 3 for now) of errors we find in data,

Missing Data - Missing data can be found when it is not continuously created or due to technical issues in the application

Noisy Data - This type of data is also called outliners; this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.

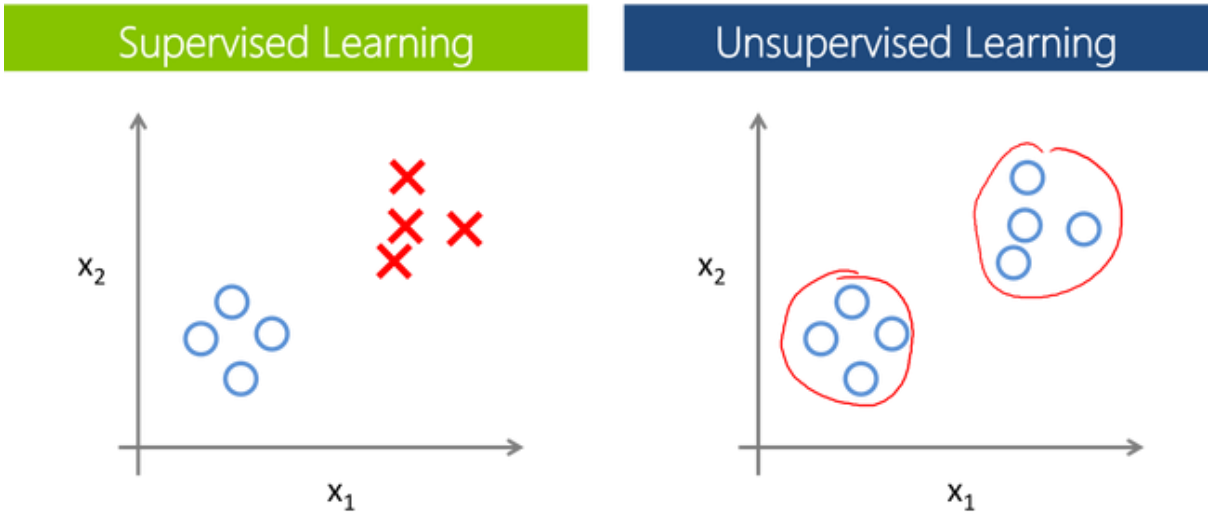
Inconsistent Data - This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

3.2 Rectification

1. Conversion of data: As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.
2. Ignoring the missing values: Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset.
3. Filling the missing values: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.
4. Machine learning: If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.
5. Outlier's detection: There are some error data that might be present in our data set that deviates drastically from other observations in a data set.

4. Evaluate Algorithm

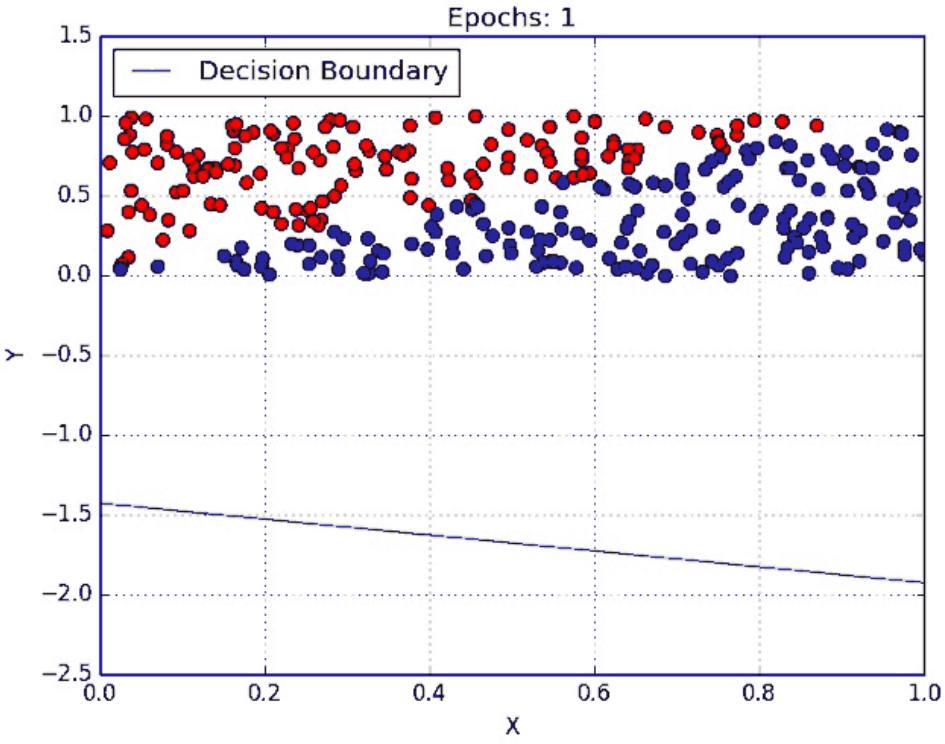
As in point 2, we have found that the dataset we have has labelled data. So, we need to use supervised models for better results. Below is an example of supervised learning



Now we need to find out which type of supervised learning we need to use. Please find the types of supervised learning below,

a. Classification
b. Regression

4.1 Classification



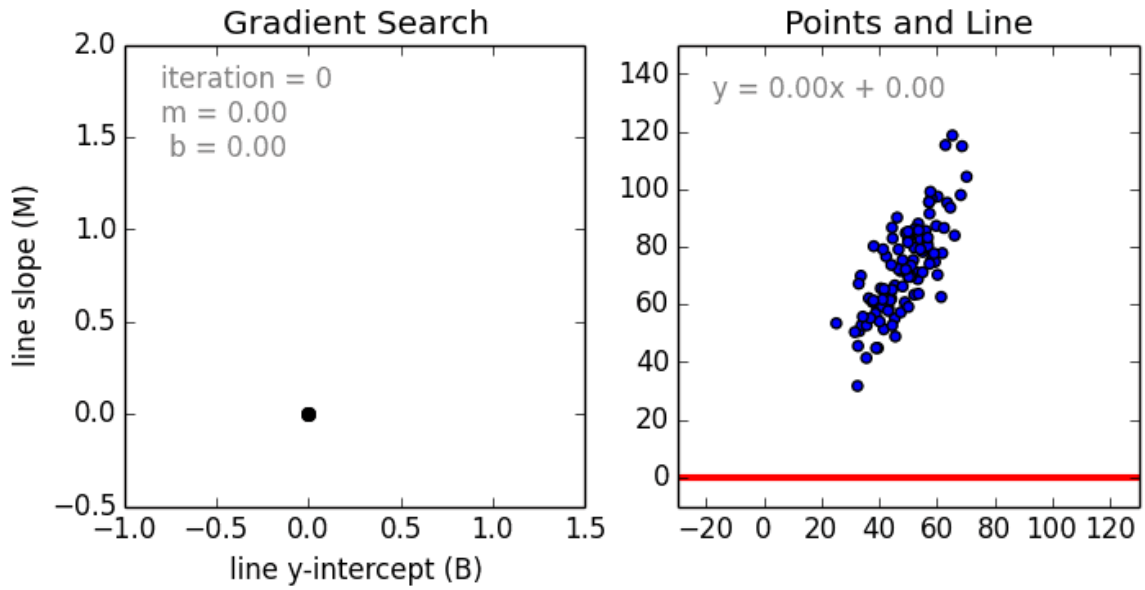
As shown in the above representation, we have 2 classes which are plotted on the graph i.e. red and blue which can be represented as 'setosa flower' and 'versicolor flower', we can image the X-axis as ther 'Sepal Width' and the Y-axis as the 'Sepal Length', so we try to create the best fit line that separates both classes of flowers.

These some most used classification algorithms.

- **K-Nearest Neighbor**
- **Naive Bayes**
- **Decision Trees/Random Forest**
- **Support Vector Machine**
- **Logistic Regression**

So, we need to test out with all the above algorithms to check which model fits the data.

4.2 Regression



As shown in the above representation, we can imagine that the graph's X-axis is the 'Test scores' and the Y-axis represents 'IQ'. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn't present in the given data.

These some most used regression algorithms.

- Linear Regression
- Support Vector Regression
- Decision Tress/Random Forest
- Gaussian Progresses Regression
- Ensemble Methods

So, we need to test out with all the above algorithms to check which model fits the data.

5. Improve accuracy and choose best fitted model

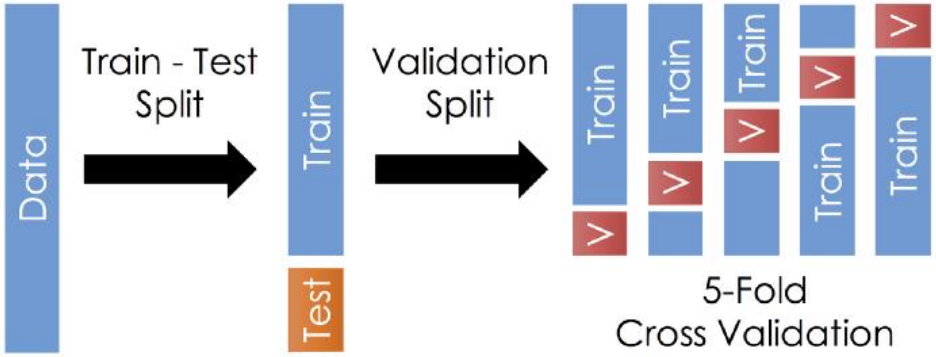
Once we have the F1 score along with p0, p1, r0, r1 scores, we can then finalize the better fitted model for our dataset. If there are possibilities of improving the accuracy by doing augmentation, we will have to implement the augmentation strategy to improve the accuracy for a given algorithm.

Please find the below calculation for accuracy score,

$$Accuracy = (True\ Positives + True\ Negatives) / (Total\ number\ of\ classes)$$

While doing the accuracy test, we need to run some certain events to make sure the model does not over fit or under fit.

NOTE: Please find the training set below:



6. Ready for production

Once the model selection is there, we need to create training data set for production use. There are multiple ways to create training data models using pickles or h5 data files.

Once the models are ready, we need to create modular codebase for easy integration. So, we need to put in an OOPs concept by creating classes and functions for the same.

7. Creating APIs

Once the code is ready for production use, we need to create apis for multiplatform use. There are several options in market to populate our codebase into an API such as,

1. Flask
2. Django
3. Fastapi

After doing some research, its best to use fast api for api development as it provides a faster development area with api documentation handy.

8. Maintenance and Monitoring

TBD

Conclusion

We have covered all the process involved in the selection of best suited model out of all supervised learning models. The tasks need to be assigned such a way that we can cover all the model algorithms which has been mentioned in this document to achieve a better accuracy.

Name: Nikhilesh Pattanayak

Date: 02/09/2021

Designation: Intern