Tour Planning Advisor For Tourist Using Statistics and Machine learning

Aakash Yadav, Ms. Sheetal Rajapurkar

School of Computer Science,

MIT World Peace University,

Pune, India – 411038

Abstract – *In the 21st century whole world is exploring its tourism potential and making it as one of the major source of economic growth and for best tourism experience the planning of the tourist must also be a prime one which provides the best experience and should also be pocket friendly .This paper proposes an efficient way to save the tourist from all the mind labour of a tour planning. By taking in the input of just the source, destination and budget(or any additional information specified by the user) this model will advice the best mode to travel which including being budget friendly and time-saver will also provide the best experience of travelling using decision tree and to make it more efficient even the smallest expenditure will be taken into account and will also help with the best budget-friendly and family-friendly stays(as per the requirement) closest to the famous tourist spots using the best-fit line from linear regression model.*

Keywords: Machine-Learning; best-fit line; Linear regression; Logistic regression.

I. INTRODUCTION

For anyone belonging to a middleclass family planning a vacation is itself an experience and not a very good one there are compromises with the experience at every stage of planning , so a smart data driven planning can solve this problem as well as can boom the tourism industry and there is no better way of doing it than using supervised machine learning algorithms hence we propose a complete data driven model which will plan an end to end trip taking every need of the user into account(as specified by the user). In today's world, machine learning is a big part of our lives. We cannot even think of a world without it at the moment. It already started to take a huge portion of our day-to-day activities. Tourism in India has a strong relevance to economic development, cultural growth and national integration. India is a vast country of great beauty and diversity and her tourist potential is equally vast. With her rich cultural heritage as superbly manifest in many of the architectural wonders (palaces, temples, mosques, forts, etc), caves and prehistoric wall paintings, her widely varied topography ranging from the monotonous plains to the loftiest mountains of the world, her large climatic variations ranging from some of the wettest and the driest as well as from the hottest and the coldest parts of the world, beautiful long beaches on the sea coast, vast stretches of sands, gregarious tropical forests and above all, the great variety of the life-style, India offers an unending choice for the tourist.

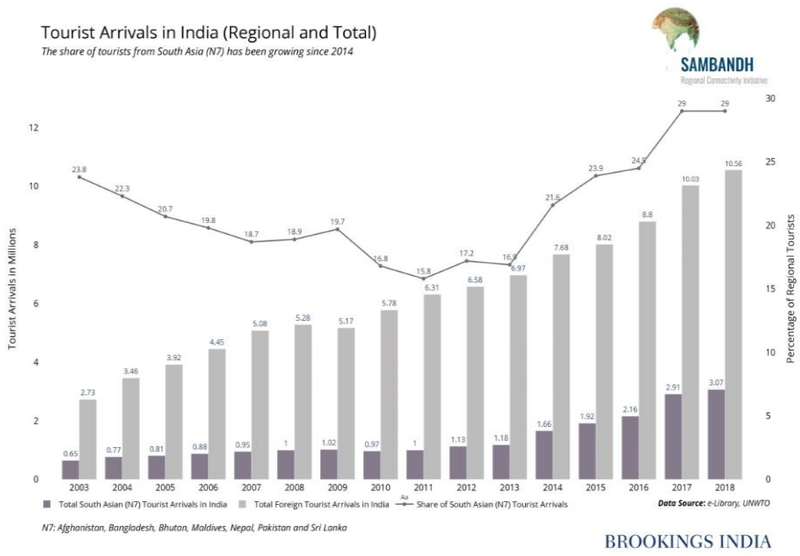
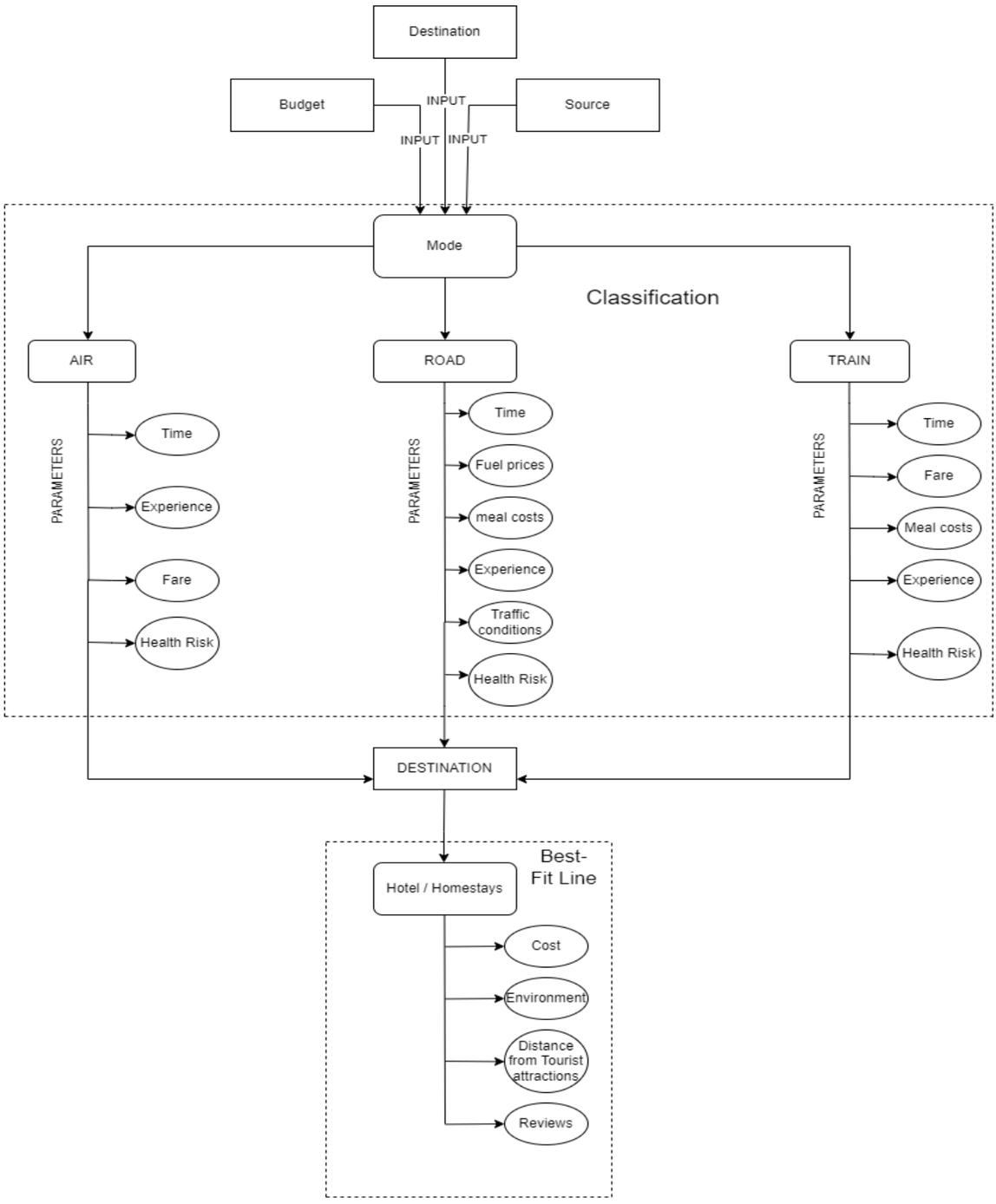


fig. 1

But after Covid-19 tourism industry took a significant hit and this model could contribute in bringing it back on track.

II. FLOW CHART



SECTION I: Mode selector

Logistic regression (3)

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common Logistic Regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

Decision Boundary

While training a classifier on a dataset, using a specific classification algorithm, it is required to define a set of hyper-planes, called Decision Boundary, that separates the data points into specific classes, where the algorithm switches from one class to another. On one side a decision boundary, a datapoints is more likely to be called as class A — on the other side of the boundary, it's more likely to be called as class B.

Let's take an example of a Logistic Regression.

The goal of logistic regression, is to figure out some way to split the datapoints to have an accurate prediction of a given observation's class using the information present in the features.

Let's suppose we define a line that describes the decision boundary. So, all of the points on one side of the boundary shall have all the datapoints belong to class A and all of the points on one side of the boundary shall have all the datapoints belong to class B.

S(z)=1/(1+e^-z)

- S(z) = Output between 0 and 1 (probability estimate)
- z = Input to the function (z = mx + b)
- e = Base of natural log

Our current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class (A/B), we select a threshold value or tipping point above which we will classify values into class A and below which we classify values into class B.

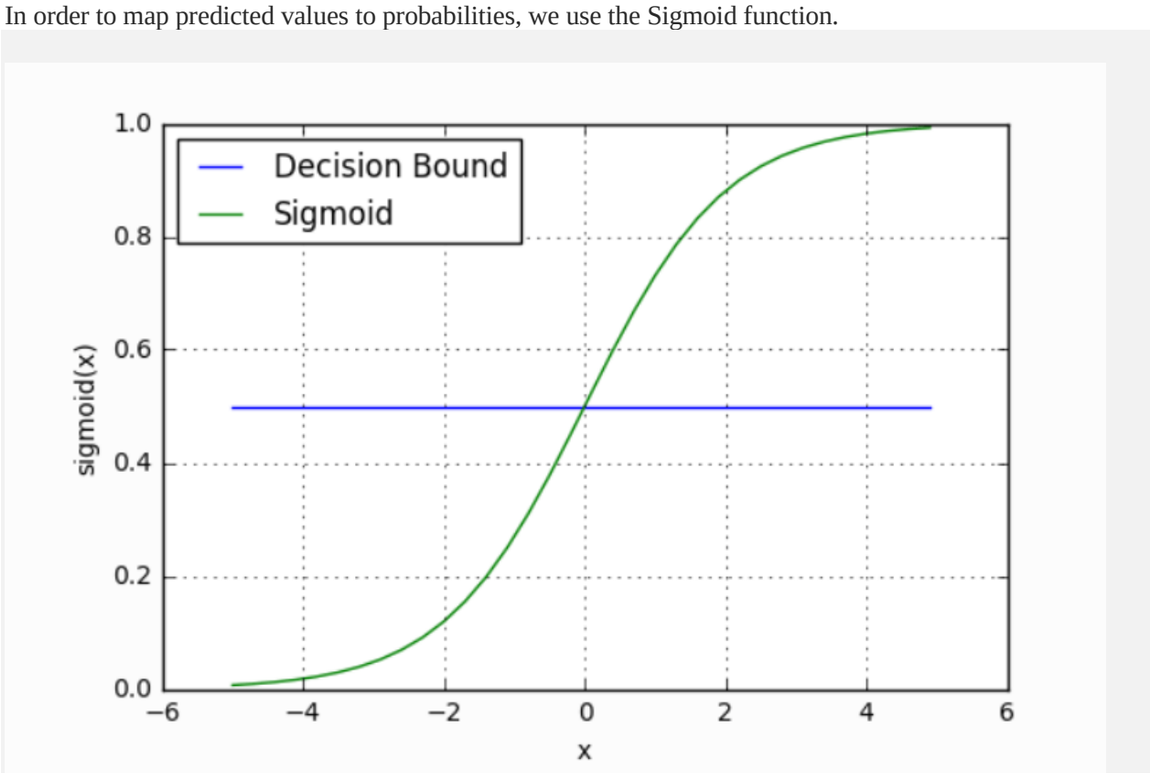
p>=0.5, class=A

p<=0.5, class=B

If our threshold was .5 and our prediction function returned .7, we would classify this observation belongs to class A. If our prediction was .2 we would classify the observation belongs to class B.

So, line with 0.5 is called the decision boundary.

In order to map predicted values to probabilities, we use the Sigmoid function.



Application on the model

As there are three modes of travel available we will use the concept of logistic regression of multiclass Classification based on the training set available the model will take parameter as independent variables and the final mode as dependent variable.

After training by examining the parameter value it will classify the dependent variable(mode).

Creating Decision boundary in Multiclass-Classification:

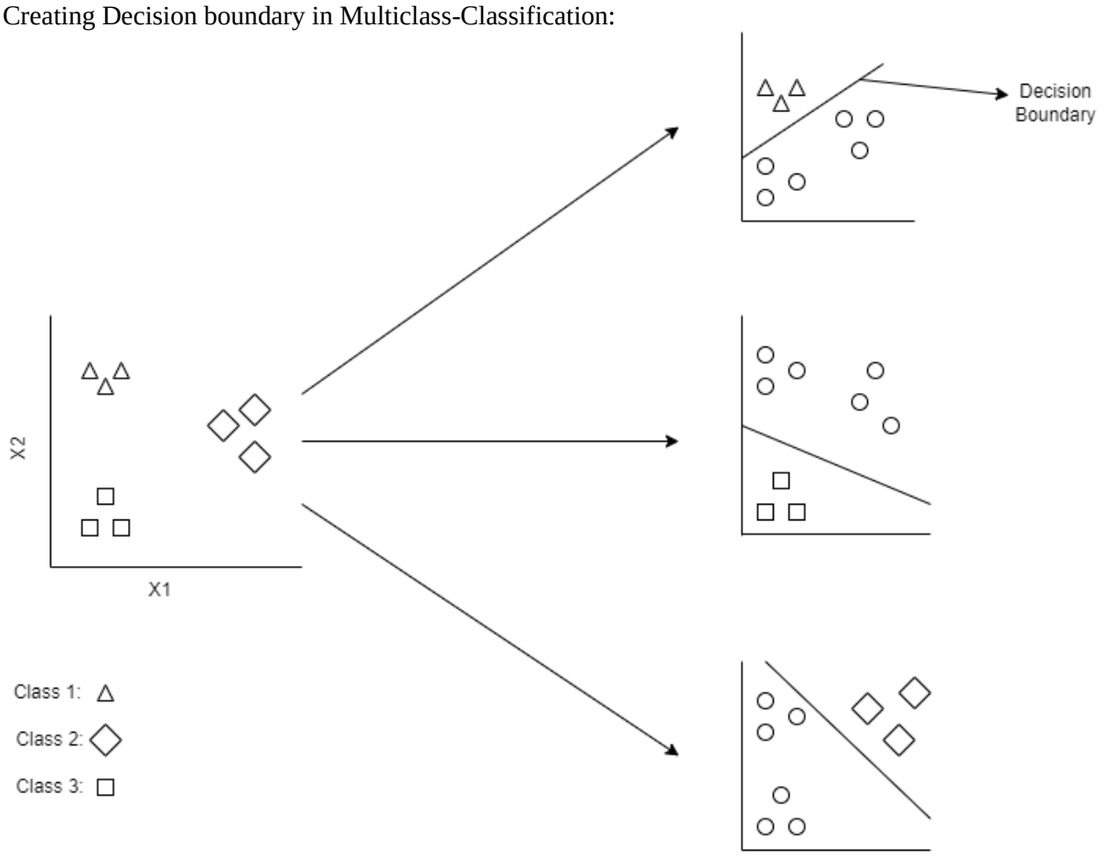


Fig. 2

Now the question arises how to get the dataset for training?

One way is to take a direct survey from the public that is a great way but very time consuming

So the other way is creating a function and putting the value of parameters according to different modes in that function.

Example:

To Reach a certain destination: for a Family of 4 members and budget for travelling is 40,000(including return journey).

	Time	Cost	Experience	Health Risk
Flight	2 hrs	30,000	5	9
Road	25 hrs	32,000	8	4
Train	24 hrs	22,000	4	9

Let Priority: Cost(50%) > health risk(30%) > experience(15%) > time(5%)

Function = weight(time(in days)) + weight(cost(in lacs)) + (1- (weight(experience/10))) + weight(health risk/10)

Weights to these parameters will be assigned according to the priority list

And the mode which will give minimum value for this function will be the desired class

For flight:

2 hrs = 0.08 days

30000 = 0.30 lacs

Experience = 5/10 = 0.5

Health risk = 9/10 = 0.9

F(x) = 0.05*0.08 + 0.5*0.30 + 0.15*(1-0.5) + 0.3*0.9 = 0.004 + 0.15 + 0.925 + 0.27 = 1.349

For road:

25 hrs = 1.04 days

32000 = 0.32 lacs

Experience = 8/10 = 0.8

Health risk = 4/10 = 0.4

Similarly for road f(x) = 1.212

For train:

24 hrs = 1 day

22000 = 0.22 lacs

Experience = 4/10 = 0.4

Health risk = 9/10 = 0.9

Similarly for train f(x) = 1.790

As we can see there is not much difference but in this case since road travel gives the minimum value for the function this example will belong to the class of road travel, similarly we can create training dataset for the classification model with different values of the parameter and train the data accordingly.

Although this is just an alternative the best dataset for training can only be received from real world.

Section II: Best Stays

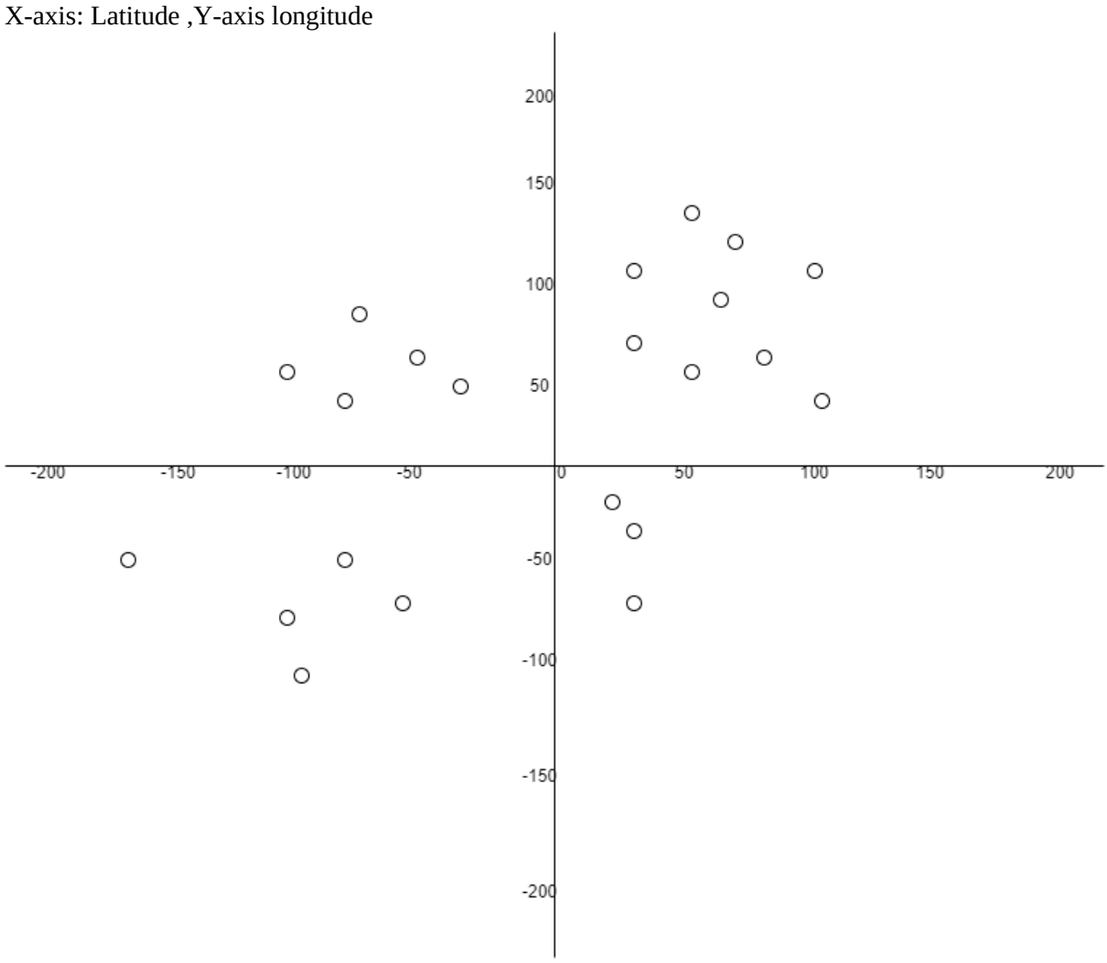
To create this feature we will use the concept of best-fit line in linear regression

Line of best fit refers to a line through a scatter plot of data points that best expresses the relationship between those points. Statisticians typically use the least squares method to arrive at the geometric equation for the line, either though manual calculations or regression analysis software.

Let’s take an example:

Plot the coordinates of famous tourist places in any area

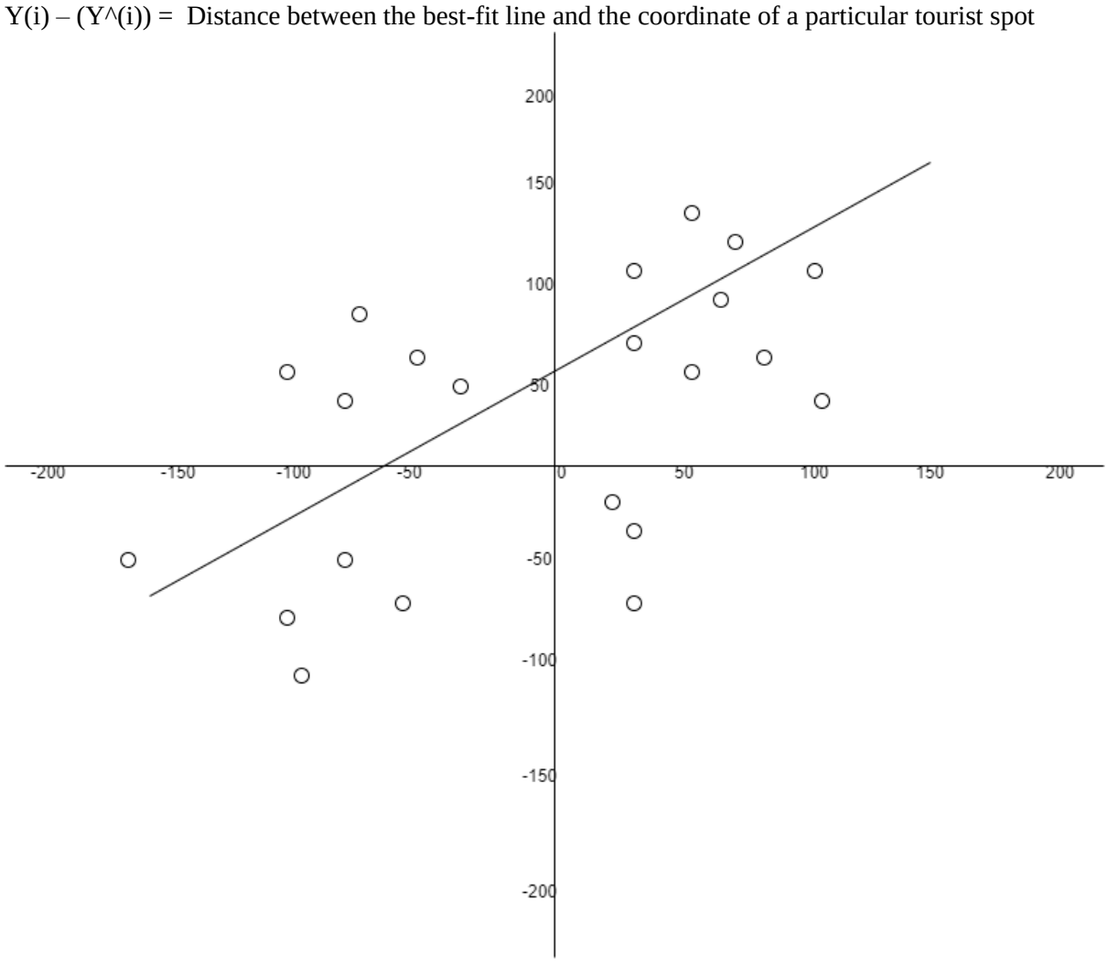
X-axis: Latitude ,Y-axis longitude



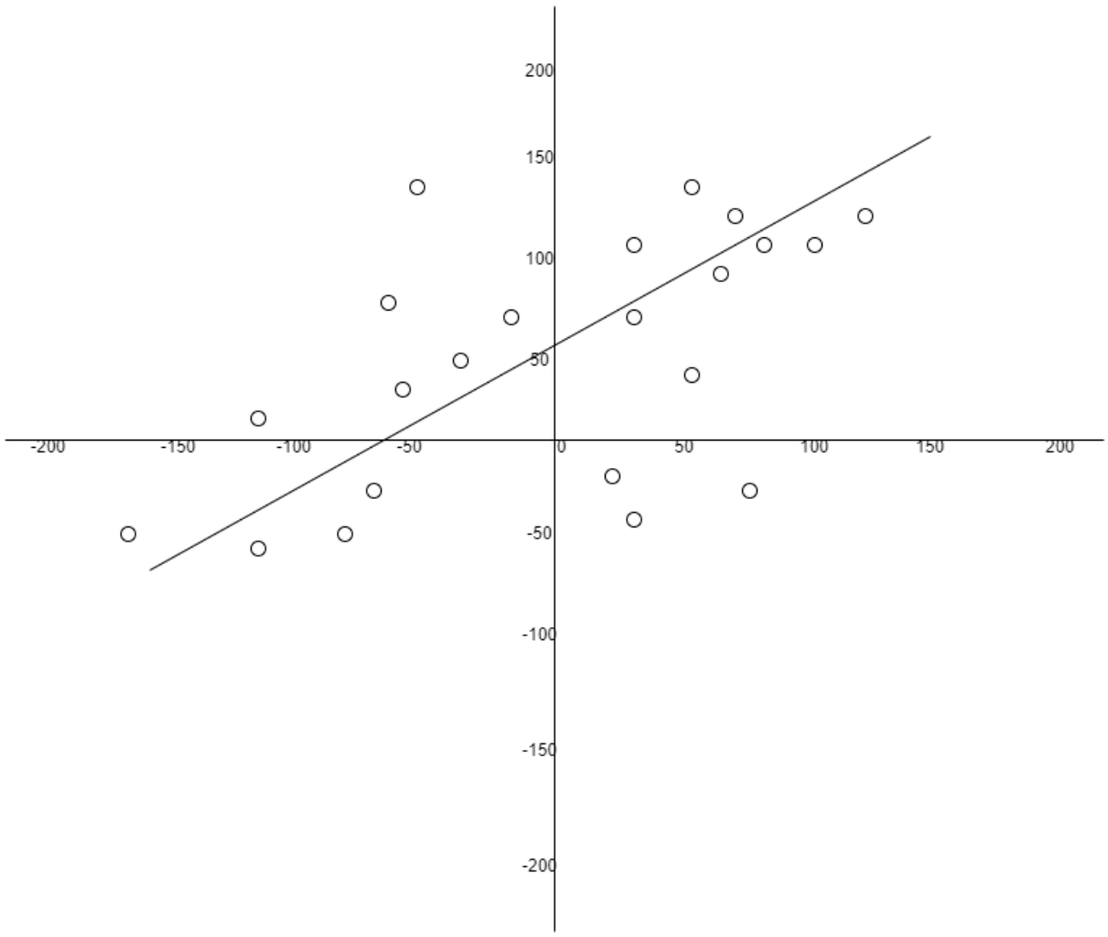
Draw a line that best fits the plot(minimizes the mean square error)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Y(i) – (Y^(i)) = Distance between the best-fit line and the coordinate of a particular tourist spot



Now keeping the best fit line the same instead of tourist spots let’s Plot the stays



By calculating the distance between the best fit line and the stay and arranging them in ascending order and pick out the top 10 stays and as in the priority list cost matters the most

Again on the basis of cost arrange those top 10 stays in ascending order and pick out top 5

And at the final stage based on the reviews this time arrange those 5 stays in descending order pick out the top 2 stays and let the user choose between those two.

CONCLUSION AND FUTURE WORK

Our paper proposed a Machine Learning Model which can efficiently provide the best help in planning a trip for any family. Using the classification algorithm and the concept of best fit-line in linear regression algorithm the model is trained without any complexity in the training dataset.

In future work now current model is only useful till reaching the final destination only helping with best-mode and best stays but using more data related to traffic conditions ad crowd intensity(using satellites and smartphones) a model can be created which will be able to provide advices to the user regarding the best mode and time to explore the destination on any given date in the future and this will also be helpful in providing the best date to road travel.

References

Fig. 1 - (https://www.brookings.edu/research/travel-south-asia-indias-tourism-connectivity-with-the-region/)

Fig. 2 – (https://www.coursera.org/learn/machine-learning/) (By Andrew Ng)

(3) - https://medium.com/analytics-vidhya/decision-boundary-for-classifiers-an-introduction-cc67c6d3da0e