Interstellar Explorer

Rudri Jani   Dhaval Deshkar   Riya Shah   Supan Shah

Abstract— **The task is to classify the possible exoplanets from the database retrived from the NASA's Kepler Mission. Various Machine Learning approaches are used for classification of the given data. In initial phase, having been worked on pre-processing the data source and use the feature extraction algorithms such as Logistic Regression(LR) and Support Vector Machine(SVM).**

Keywords– Classification of exoplanets, Kepler Mission, Logistic Regression, Support Vector Machine, Machine Learning, Confusion Matrix,XG Boost.

I. INTRODUCTION

An exoplanet is any planet far away from our solar system.The first exoplanet was discovered in the early 1990s. The first exoplanet was burst in the world stage was named 51 Pegasi b, Which was like "hot Jupiter" orbiting a Sun-like star far from 50 light-years. 1995 year was watershed year. Since then we have discovered thousands and more. The discovery of many of the exoplanets are so far off in a relatively small region of our galaxy, the Milky Way. We recognize from NASA's Kepler Space Telescope that there are planets more than stars in the galaxy. The size and mass plays a crucial role in determination of planets type. There are different varieties within the size/mass classifications. It is possible that critical size in planet formation: Planets that reach the size quickly attract thick atmospheres of hydrogen and helium gas, and balloon up into gaseous planets, while planets are smaller than this limit are not large enough to hold such an atmosphere and remain primarily rocky and terrestrial bodies.The planets which are in smaller size have that orbits close to their stars could be the cores of Neptune-like worlds that had their atmospheres stripped away.

In Machine learning, classification describes the predictive modeling where a class label is predicted for a given input data. Based on this idea, there are two major techniques that is used for this task: Linear Regression(LR) and Support Vector machine(SVM). Logistic Regression algorithm is used for solving the classification problems with using the statistical approach. Support Vector Machine(SVM) is used for both classification and regression problems which is based on geometrical properties of the data. This powerful techniques are efficient to determine weather it is the exoplanet or not. Fig.1 shows the random picture of exoplanets.

Fig. 1.   Exoplanets

II. LITERATURE SURVEY

A new era has unfolded in the chase for exoplanets in any case. The Modern generation of the new satellites, Such as a Kepler, have been launched in recent times to partially automate logical perceptions and data generation with an exoplanet identification proof.[1] Engineering of these satellites are not for only takes the picture but also do the process of those images using higher astronomical techniques to create such a huge collection of data with proper variety of features for identification of exoplanets. The simulated deployment of the model is apex of data pipeline, which prepares, train and test for machine learning model. The final result of machine learning modeling is introduced and summarised to show the classification of the exoplanets dimensionally over the different models constructed in the study.

The survey shows, this study attempted to create a set of candidate features after accounting for missing values, Statistical inconsistencies, correlations and bias. Here, the study shows, that while we have spotlight on making binary predictions for optimistic model comparisons, the XGBoost classifier can gives output a predicted confidence for the any given system between 1 and 0.[2]

III. IMPLEMENTATION

The project will use classical machine learning methods for the classification of planets as exoplanets or not using features extracted by the Kepler space telescope. Importing the necessary libraries for the Machine learning algorithm and working on Pre-proecessing the data.

A. Pre-processing

In the phase of data pre-processing, The data is retrived from the NASA's Exoplanet Database.Data having 8.481% of the data was null. In the given dataset, An exoplanet candidate is a probable planet found by a telescope however has not yet been demonstrated to really exist or not. It is possible for some candidates to turn out to be ”false positives.” A planet is only considered ”confirmed” once it is verified through additional observation using two other telescopes. There are so many planets candidates awaiting confirmation. But time on telescopes is considered a precious resource and it takes a lot of computing time to search the determination of the targets. This one area is where amateur scientists can work with NASA data to help filter the targets and even discover exoplanets. Where computers might miss a single transit, humans can detect little brightness dips in data that might tell us there is a planet to be found. After finding irrelevant columns from dataset it is removed. The missing values of the all the rows are removed. Next, We apply dimensionality reduction techniques to reduce the dimensions of our data. After that data is normalized. For the dataset, hot encoding is necessary for columns having categorical values. Hence, One hot encoding of categorical data is done. After that it is creating a new column on the basis of candidate and confirmed planets which is the target values. The data is splitted into train and test data in 80-20% and applied on the model. The following chart shows the flow of the task,

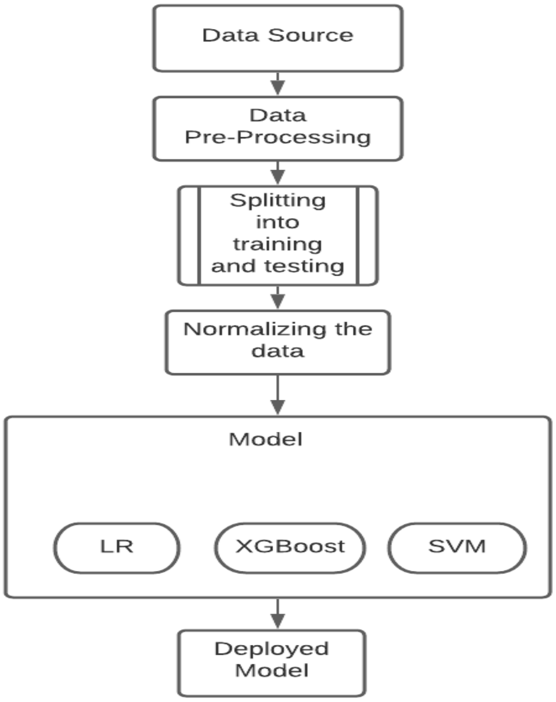

Fig. 2. Flow chart

B. Logistic Regression

Logistic Regression is significant approach of regression analysis to conduct when the dependent variables is binary. It is Predictive analysis. It is used to describe the data and to explain the relationship between one dependent binary variable and independent variables. The equation of Logistic regression :

$$Y = \exp(b0 + b1 * X)/(1 + \exp(b0 + b1 * X))$$

In this equation Y is expected output, b0 is bias and b1 is it's multiplication factor for the input X.

We are using Sigmoid Function, Cost Function and Gradient Descent for the Logistic regression

Sigmoid Function:

$$\sigma(x) = 1/(1 + \exp(-x)$$

Cost Function:

$$J = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log(a^{(i)}) + (1 - y^{(i)}) \log(1 - a^{(i)})$$

Gradient Descent:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} f(\theta_0, \theta_1) \text{ for j=0,1}$$

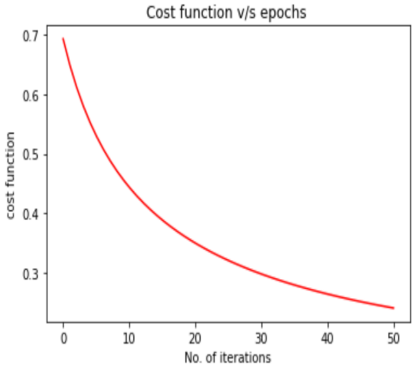

Fig. 3. Output: Cost Function

Output: It will gives the smooth curve.

C. Support Vector Machine

Support vector machine (SVM) is powerful and flexible class of supervised algorithms for both classification and regression.

D. XG Boost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. The engineering behind this algorithm is for efficiency of compute time and memory resources. A goal of this algorithm is implementation of gradient boosting decision tree algorithm.

Here, we are taking few features for understanding the features and we applied it with XG Boost algorithm. Features used like: disposition score,planetary radius,TEC planet number,Stellar Temperature,Transit Signal-to-noise ratio.

following figure shows XGboost Decision tree,

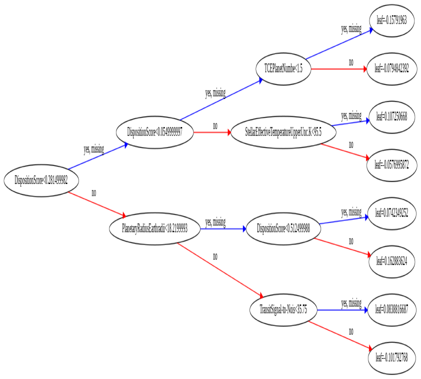

Fig. 4. XG Boost Decision tree

IV. RESULTS

After taking the features, We applied the XG Boost algorithm on it. After that we subset our datato these few features and comparing the models. Following Figure gives the brief idea about the XG Boost training and testing model.

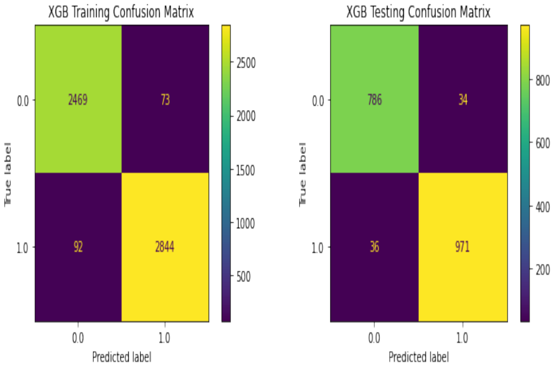

Fig. 5. XG Boost Training and Testing model

Here, the figure shows the comparision between the withouth selecting the features and selection of the features applied through the XGBoost algorithm.

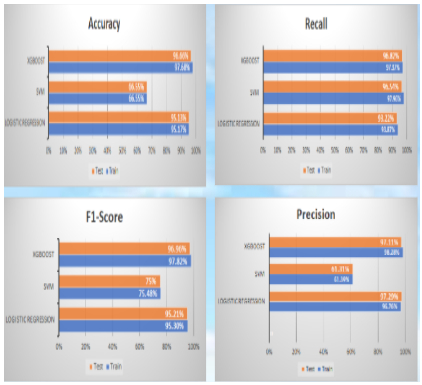

Fig. 6. Comparision between the models

V. Conclusion

Hence, the result shows in this project we used different models to get better accuracy and f1-score so we can achieve better results for classifying exoplanets. To conclude, XGBoost model performs way better on unseen data as it gives 96.66% test accuracy and 96.96% test f1-score. As final part, we deployed the model on IBM cloud. We generates deployed model through API inference and we construct the final web based service.

The final code is available at `https://colab.research.google.com/drive/182VPMbN5D1EqTFl1pVQuHR91KJ6MsP3t?usp=sharing`.
The final web service is available at `https://interstellar-explorers-app.herokuapp.com/`.
The Github repository is available at `https://github.com/Supan14/CSE523-Machine-Learning-Interstellar-Explorers`

References

[1] G. C. Sturrock, B. Manry, and S. Rafiqi, "Machine learning pipeline for exoplanet classification," *SMU Data Science Review*, vol. 2, no. 1, p. 9, 2019.

[2] D. Tamayo, A. Silburt, D. Valencia, K. Menou, M. Ali-Dib, C. Petrovich, C. X. Huang, H. Rein, C. Van Laerhoven, A. Paradise *et al.*, "A machine learns to predict the stability of tightly packed planetary systems," *The Astrophysical Journal Letters*, vol. 832, no. 2, p. L22, 2016.