Identifying Exomoons : Identifications of Exoplanet Orbitals through estimating the Orbital Semi Major Axis using Machine Learning

Name:	**Himadri Sonowal**
Registration No./Roll No.:	20128
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	17 August 2023
Date of Submission:	20 November 2023

1 Introduction

The project aims to develop a machine learning-based solution to classify exo-moons based on the orbital semi-major axis. It combines aspects of data preprocessing, feature engineering, machine learning modeling, and performance evaluation to achieve the goal. The main objective is to use machine learning to build a regression model to predict the semi-major axis of exoplanet orbit collecting data from NASA archive.

- exoplanet trn data.csv: This file contains training feature vectors. It has 17969 rows (data points) and 288 columns (features).
- exoplanet trn data targets.csv: This file contains training target values (semi major axis) . It has 17969 rows (data points) and 2 columns (index and semi major axis).
- exoplanet tst data.csv: This file contains test feature vectors. It has 1997 rows (data points) and 288 columns (features).
- exoplanet data desc.txt: This is a text file that contains the description of all 288 features available in our dataset.

2 Plans

The github repository can be found here. In the context of data processing for a machine learning task, it appears that there were four data files involved:

1. Training Data: This dataset contains 17,969 instances, and each instance has 288 features.
2. Train Labels: This dataset contains the corresponding labels for each of the instances in the training dataset. It is used to supervise the training of machine learning models, allowing them to learn patterns and make predictions.
3. Test Data: This dataset contains 1,997 instances, and like the training data, each instance has 288 features. However, unlike the training dataset, the test dataset does not include labels. The goal is to use trained machine learning models to predict labels for these instances.

3 Missing Data Feature Removal

1. Determining the overall count of null values for each attribute in the dataset. 2.Organize the features in a descending order based on the quantity of null values they possess. 3.Choose the top features that surpass a specified threshold of null values. 4. Eliminate these features from the dataset and generate a new dataset named tr05.

In a few 'numerical' features, the presence of non-numeric values among some instances led to misclassifying them as Categorical Columns. This issue has been addressed in the subsequent section.

4 Converting Categorical Data to Numerical for Model Run

Due to the presence of categorical features with class names represented as strings, which cannot be directly utilized in a regression model, we transformed these distinct 'string' values into numerical classifiers.

1. Formulate a catalog of categorical features. 2. Attribute a numerical value to each unique class within the categorical feature.

5 Outlier Removal

Numerous columns in the dataset exhibit significant disparities in data values, spanning from the micro to mega range. This deviation far exceeds the acceptable range of standard variance. Consequently, we eliminated outliers with the aim of enhancing correlation.

Instances beyond the specified threshold were removed, where the outlier value is greater than or equal to $\mu \pm (10 \times)$, with μ representing the mean of the data and denoting the variance of the dataset.

Following this operation, certain feature vectors became null, prompting their removal as well

6 Correlation Calculation

Correlation computations were performed for each feature in the training set, leading to the generation of a correlation map (C-map) that visually represents features exhibiting the highest correlation with the target variable.

The Orbital Period and its associated errors were identified as having the most substantial correlation with the target variable.

7 Feature Engineering

Utilizing Kepler's third law, we implemented specific mathematical transformations on the orbital period, resulting in the creation of two new feature vectors. These newly generated features were subsequently incorporated into the training dataset to enhance the overall training quality. The correlation analysis revealed that these additional features exhibited a significantly high correlation with the existing variables.

8 Modelling

The models run on the dataset were : 1. Linear Regression 2. Decision Tree 3. Support Vector Regression 4. Random Forest Regression

The that models were performed on the training set and we calculated the RMSE values for each of the training models.

The best-performing model is identified to be Random Forest Regression as it provided the minimum error. It was then applied to the test data to predict the final prediction that is the orbital semi-major axis for each instance.

The final values of the target variable as predicted by the Random Forest were saved into a text file named, predicted semi major axis.txt.

Once the best-performing model is identified, it will be applied to the test data to predict the orbital semi-major axis for each instance.

9 Results and Discussion

We used "Mean Squared Error" to calculate the average of the squared differences between each predicted values and actual values in the dataset. The goal was to minimize the error to improve the accuracy of predictions.

Hyperparameter tuning for different models are :

Linear Regression: 0.0617095 Decision Tree Regression: 0.0131003 Support Vector Regression: 0.0105230 Random Forest Regression : 0.009371719665190195

The R2 (coefficient of determination) scores for various regression models are as follows:

Linear Regression: 0.9096 Decision Tree Regression: 0.9808 Support Vector Regression: 0.9846 Random Forest Regression: 0.9894

The R2 score is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model. A higher R2 score indicates a better fit of the model to the data. In this context, the Random Forest Regression model has the highest R2 score, suggesting that it performs exceptionally well in explaining the variability in the target variable.

The method demonstrates it's abilities in handling missing data, outliers, and inaccurately classified categorical data, making it versatile for cleaning diverse research datasets. It autonomously cleans new data and provides predictions for the semi-major axis of exoplanets. Importantly, the method requires no prior domain knowledge, autonomously learning relevant features for regression problems without bias toward any specific feature.

However, the main limitation lies in the high time complexity, demanding substantial computational power for parameter tuning, particularly on large exoplanet datasets. The current hyperparameter tuning involves hand-picked parameters, which could benefit from automation. Future improvements could explore alternative feature selection metrics such as p-values and incorporate ensemble techniques, utilizing multiple regression models for enhanced predictive performance. Additionally, the method's potential expands to exoplanet habitability analysis by considering eccentricity and habitable zone information, presenting an exciting avenue for further exploration and application.

References

Exoplanet Archive NASA

Ethem Alpaydin-Introduction to Machine Learning-The MIT Press (2014).pdf