Exoplanet Analysis

Julia Haynes

December 14, 2023

Introduction

In the rapidly evolving field of astrophysics, the study of exoplanets holds a particularly intriguing position. These celestial bodies, orbiting stars beyond our solar system, offer a unique window into the vast and varied nature of the universe. With the increasing availability of extensive datasets, such as those from the NASA Exoplanet Archive, coupled with rapidly advancing technology, we are now able to explore these distant worlds more comprehensively than ever before. This project aims to use the power of computational analysis to delve deeper into understanding exoplanet characteristics, including mass, semi-major axis, eccentricity, temperature, as well as stellar metallicity.

The motivation for this project stems from a growing interest in understanding the mysteries of exoplanetary systems and their formation, structure, and potential habitability. By analyzing relationships between various exoplanet features, this project seeks to identify patterns and anomalies in exoplanet behavior.

A key aspect of this project is the application of advanced computational techniques, a cornerstone of modern astrophysical research and a significant element of the course curriculum. Starting with linear regression models and progressing to polynomial regression, the project initially employs traditional statistical methods to identify potential relationships between variables. This foundational approach allows for an initial exploration of the dataset, setting the stage for more sophisticated analyses in the future.

As the project progresses, it adopts machine learning algorithms, including neural networks, to predict certain exoplanet characteristics based on identified correlations. This not only aligns

with the course objectives of applying computational methods to solve complex problems but also demonstrates the practical utility of machine learning in scientific inquiry. The use of neural networks, in particular, exemplifies cutting-edge computational techniques, capable of handling the vast and complex datasets typical in astrophysics.

Moreover, the project's methodology includes an examination of error margins and a discussion of the results in the context of existing astrophysical theories. This critical evaluation is crucial in scientific research, ensuring that findings are reliable.

In conclusion, this project is not only interesting due to its potential to uncover new insights into exoplanetary science but also serves as an application of computational techniques in scientific research. It aligns seamlessly with the course objectives, demonstrating the practical application of statistical and machine learning techniques in tackling complex, real-world problems in astrophysics.

LITERATURE REVIEW

From the paper titled 'Mass-Radius Relationship For Solid Exoplanets ' it is evident that highly detailed interior planet models are not always necessary to determine the bulk composition of solid exoplanets based on their mass and radius. This is because factors such as temperature structure and phase changes have a limited impact on the overall planet mass and radius. Instead, a simplified but effective approach is observed, where solid planets tend to adhere to a scaled mass-radius relationship described by $\log_{10}R_s = k1 + (1/3) * \log_{10}(M_s) - k2 * M_s^{k3}$ for a range of planetary masses up to approximately 4 times the mass of Earth ($M_p \approx 4M_\oplus$).

Furthermore, in the paper titled "Revisited Mass-Radius Relations for Exoplanets Below 120M," the utilization of a composition line representing pure water aids in distinguishing between two distinct populations of exoplanets. This innovative approach leads to the development of two new empirical Mass-Radius (M-R) relationships derived from the data, thereby enhancing our understanding of the correlation between mass and radius for exoplanets.

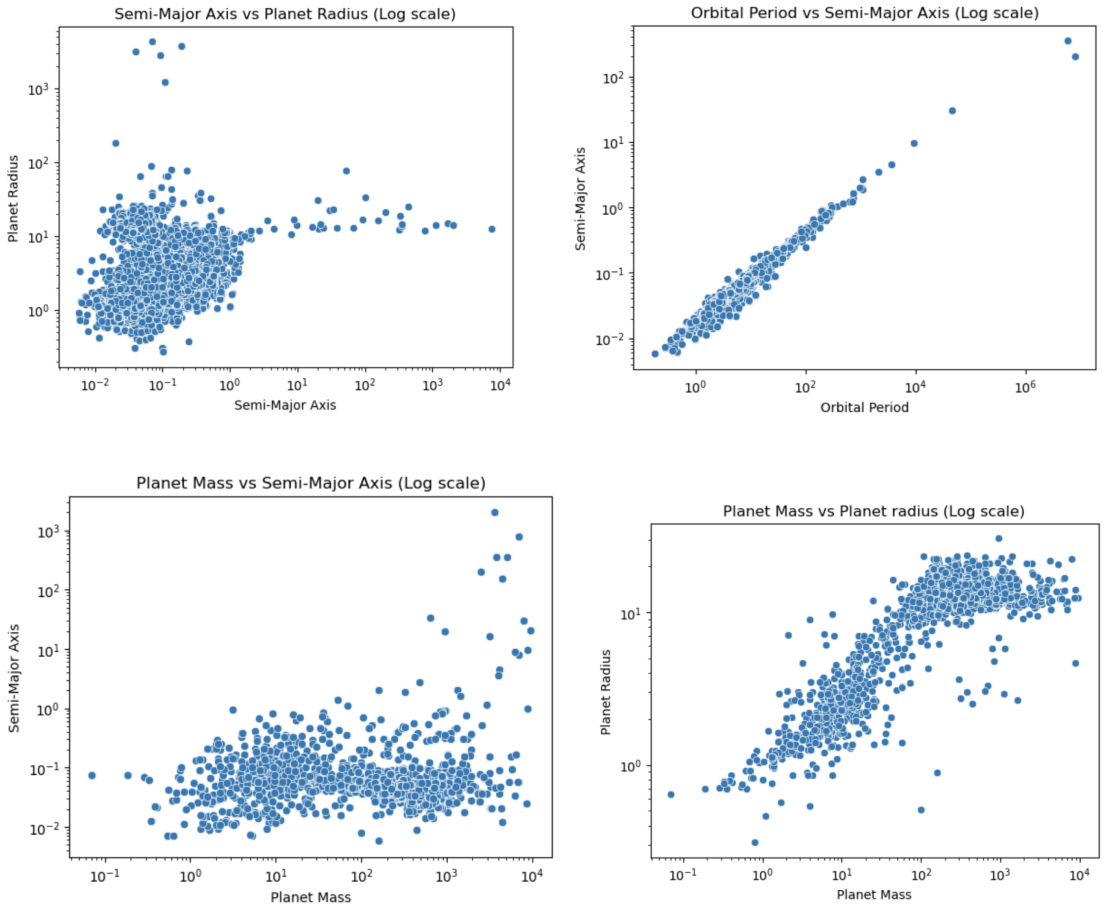
METHODS

In this project, we embarked on an exploratory analysis of exoplanet characteristics using a dataset from NASA's Exoplanet Archive. The primary focus was on establishing relationships between various exoplanetary attributes such as mass, semi-major axis, radius, eccentricity, stellar mass, stellar metallicity, and more. The data was initially cleaned by dropping missing values. I started by exploring and visualizing the data. I created histograms of variables, and analyzed which variables I thought may have a correlation. Then I made scatter plots for those variables. I analyzed the scatterplots, and determined which variables I should analyze further. For further analysis, I started by using linear and polynomial regression models. Given the non-linear nature of some relationships, machine learning techniques, including random forest and neural network regressions, were integrated to capture more patterns. We moved onto analyzing relationships between 3 or more variables, using additional methods such as the gradient boosting regression model. Moreover, we addressed the error inherent in astronomical data by implementing weights, where weights were derived from the provided error margins of the variables. This approach allowed us to account for varying levels of uncertainty across different observations. The implementation of these methodologies was conducted through

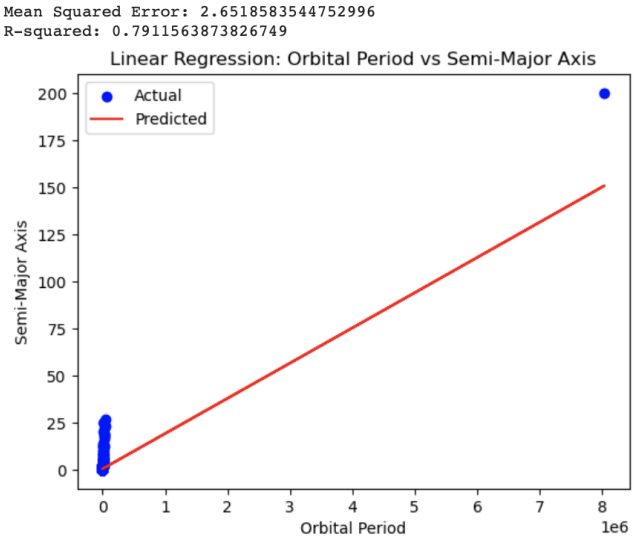
Python, utilizing libraries such as Pandas for data manipulation, Scikit-learn for machine learning models, and Matplotlib and Seaborn for data visualization.

RESULTS AND ANALYSIS

From the initial scatter plot analysis, we determined the following variables should be further analyzed: Orbital period vs Semi-Major Axis, Planet Mass vs Planet Radius, Semi-Major Axis vs Planet Radius, and Semi-Major Axis vs Planet Mass.

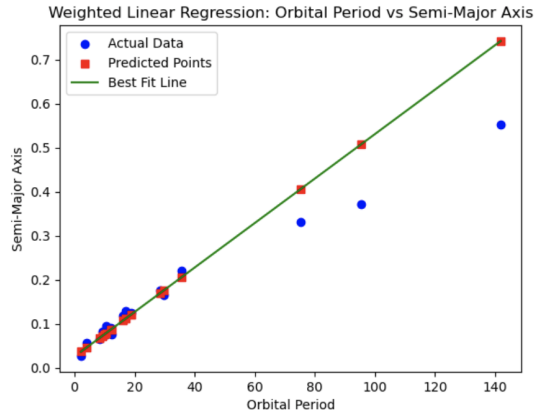


For orbital period vs semi-major axis, we found that a linear regression model worked best, giving a r-squared value of 0.79.



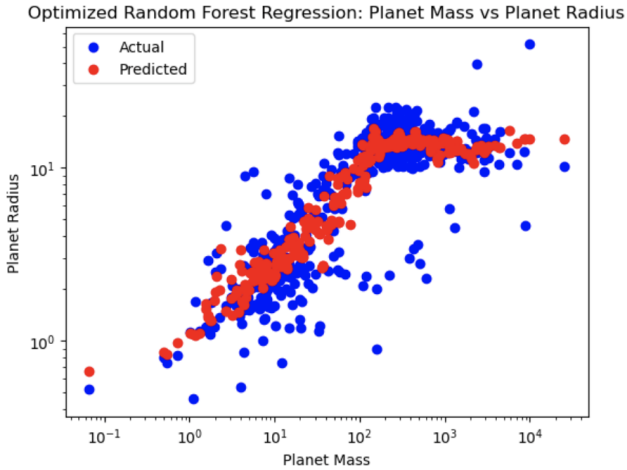
This makes sense, as the orbital period and semi-major axis are related by Kepler's third law. However, this changed when we accounted for the errors using weights. Values with greater error are given a lower weight. When accounting for the error, the r-squared value for the linear regression model improved to 0.82. The polynomial regression model improved from -0.42 to 0.88, making the second degree polynomial regression model more accurate. While this may seem surprising, the polynomial regression model shows an almost linear fit, which reflects Kepler’s third law. Using weights to account for the error resulted in models that better fit the data and what we expect.

Weighted Mean Squared Error: 0.000361516172976902
Weighted R-squared: 0.8248211578162223



For planet mass vs planet radius, we found that the optimized random forest regression technique worked best, giving a r-squared value of 0.67. This follows what previous papers have concluded, confirming the correlation between mass and radius. There were no significant correlations for semi-major axis vs radius, and semi-major axis vs mass.

Mean Squared Error: 13.492908684681879
R-squared: 0.6732797296291344



Next we analyzed correlations between 3 or more variables. For the first analysis, the predictor variables were radius and semi-major axis, and the response variable was planet mass. There was very little correlation, with the gradient boosting regression having the greatest r-squared value

of 0.09. The second analysis had the predictor variables radius, mass, and temperature, and eccentricity as the response variable. There was little correlation, with the random forest regression model giving an r-squared value of 0.15.

The third analysis was with the predictor variables: planet radius, semi-major axis, planet mass, stellar metallicity, with the response variable being planet temperature. The gradient boosting regression analysis gave a r-squared value of 0.935, which is an extremely high correlation. This shows a complex interplay between the predictor variables and the planet temperature. This shows that there are multiple factors that impact an exoplanet's temperature, and highlights the importance of observing the effects of multiple variables.

For the next analysis, the predictor variables were radius, semi-major axis, planet temperature, and eccentricity. The response variable was mass. This showed to have some correlation, with the random forest regression analysis giving a r-squared value of 0.37.

CONCLUSION

In conclusion, this research project has yielded valuable insights into the relationships between various key parameters of exoplanets and their physical characteristics. What we have learned is that while some correlations between specific variables and exoplanet properties are evident, the complexity of exoplanetary systems often leads to nuanced and multifaceted relationships. Notably, we observed strong correlations in certain cases, such as the remarkable R-squared value of 0.935 for the gradient boosting regression model, indicating an intricate relationship between planet radius, semi-major axis, mass, stellar metallicity, and planet temperature.

While we found many variables to not be correlated, the project can be considered a success in achieving its primary goal of exploring correlations and relationships among the selected variables. It provided valuable insights into the interdependencies of these parameters, enhancing understanding of exoplanetary science.

While the project succeeded in its primary objectives, it also revealed opportunities for further exploration and refinement. Future applications and extensions of this research could involve further analyzing the correlation between mass and radius. The linear regression analysis returned a r-squared value of 0.67. While this is significant, it could be improved. Previous literature suggested a different relationship between mass and radius for different sizes of exoplanets. In the future, we can separate the data by mass, and fit different relations based on the mass. Perhaps it could also give us more accurate results for other relationships.

While using weights to account for errors was very useful for linear and polynomial regression models, in the future it would be helpful to find a way to account for errors in the more complex machine learning techniques I implemented.

In the future, we want to incorporate other datasets that include other interesting information, such as atmospheric composition, chemical composition, and additional host star characteristics. It would also be interesting to create a classification system based on observed correlations, aiding in the characterization of exoplanets. For a complete and thorough analysis, it would be useful to further analyze relations that didn’t seem to have any correlation when reflected on a scatter plot. It would also be useful to explore more relations between multiple predictor variables.

Bibliography:

1. Otegi, J. F., et al. “Revisited Mass-Radius Relations for Exoplanets below 120 $M_{\oplus}$.” *Astronomy & Astrophysics*, EDP Sciences, 4 Feb. 2020, www.aanda.org/articles/aa/full_html/2020/02/aa36482-19/aa36482-19.html .
2. Seager, S, et al. “Mass-Radius Relationships for Solid Exoplanets - Iopscience.” *The Astrophysical Journal*, iopscience.iop.org/article/10.1086/521346/pdf. Accessed 15 Dec. 2023.