Exoplanet Detection Using Machine Learning Models

Jainam Shah
DEPAUL UNIVERSITY
JARVIS COLLEGE OF COMPUTING AND DIGITAL MEDIA

Abstract : In recent years, the search for exoplanets-planets orbiting stars beyond our solar system-has captivated both scientist and the public, promising insights into the vast universe and the potential for life in the universe. Leveraging data from NASA's Kepler Space telescope, this study aims to enhance exoplanet detection by comparing three machine learning models: Logistic Regression, Random Forest, Light Gradient Boosting Machine(LightGBM). By evaluating these models on accuracy on the statistical metrics such as accuracy, F1-score, ROC-AUC and check for the computational time, I aim to pinpoint the most effective approach for distinguishing true exoplanets from false positives. My findings reveal that both Random forest and LightGBM significantly outperform Logistic Regression, with LightGBM excelling in prediction speed, making it ideal for real-time applications. This research not only advances the methods for exoplanet detection but also contributes to our understanding of the universe, paving the way for future discoveries.

Keywords Used:

Exoplanet Detection, Machine Learning, Kepler Space Telescope, Logistic Regression, Random Forest, LightGBM.

1. INTRODUCTION

Exoplanets, or planets that orbit stars outside our solar system, have got significant interest in recent years. The Kepler Space Telescope, launched by NASA, has been instrumental in the search for these exoplanets by analyzing light curves for the potential transits across the host stars. This method often generates, numerous false positives due to other astrophysical phenomena. Accurate detection of exoplanets is crucial for understanding planetary systems and potential for life outside of Earth.

In this study, I have focused on identifying the best machine learning model for distinguishing true exoplanets from false positives. Using data from the Kepler Space Telescope, I evaluated three models: Logistic Regression, Random Forest, and Light Gradient Boosting Machine (LightGBM). The performance of these models was compared to determine the most accurate and efficient model for exoplanet detection. The objective of this research is to enhance the accuracy of exoplanet detection methods, thereby contributing to the broader field of astronomy.

1.1 Data Source

The data used in this study is sourced from the Kepler Space Telescope, which was NASA's first mission dedicated to finding Earth-sized planets orbiting other stars in the Milky Way galaxy. Launched in 2009, the Kepler mission aimed to determine the frequency of Earth-sized planets in or near the habitable zones of their host stars. Over its operational period, Kepler monitored over 150,000 stars and discovered more than 2,600 confirmed exoplanets, significantly enhancing our understanding of planetary systems (3).

The dataset includes 9,564 observations with 50 features related to the characteristics of the observed stars and their potential planets. Key features in the dataset include:

- koi_fpflag_nt: Not Transit-Like False Positive Flag
- koi_fpflag_ss: Stellar Eclipse False Positive Flag
- koi_prad: Planetary Radius
- koi_period: Orbital Period
- koi_depth: Transit Depth

These features are crucial for identifying potential exoplanets and distinguishing them from false positives. The target variable, koi_disposition, indicates whether an observation is a confirmed exoplanet, a false positive, or a candidate (1).

1.2 Data Collection and Processing

Kepler detected exoplanets using the transit method, which involves measuring the slight dimming of a star as a planet cross in front of it. This method, while effective, often requires further validation to rule out false positives caused by other astrophysical phenomena such as binary stars or star spots (8).

Once potential exoplanet candidates were identified, they underwent a thorough validation process involving additional observations and data analysis techniques. These included:

- Ground-based follow-up observations to resolve background objects that might contaminate the signal.
- Doppler spectroscopy to measure the star's radial velocity and confirm the presence of an orbiting planet (7).

1.3 Related Work

Exoplanet detection has been a rapidly advancing field, thanks to the missions like Kepler and its successor, the Transiting Exoplanet Survey Satellite (TESS). These missions have revolutionized our understanding of planetary systems by revealing that planets are more common than stars in our own galaxy (3).

Numerous studies have utilized Kepler data to develop and refine exoplanet detection methods. For instance, the use of machine learning techniques has been explored extensively, with models like Random Forests and Gradient Boosting showing significant promise in improving detection accuracy (3). Additionally, deep learning methods, particularly convolutional neural networks (CNNs) have demonstrated higher capabilities in identifying exoplanets candidates by analyzing light curves.

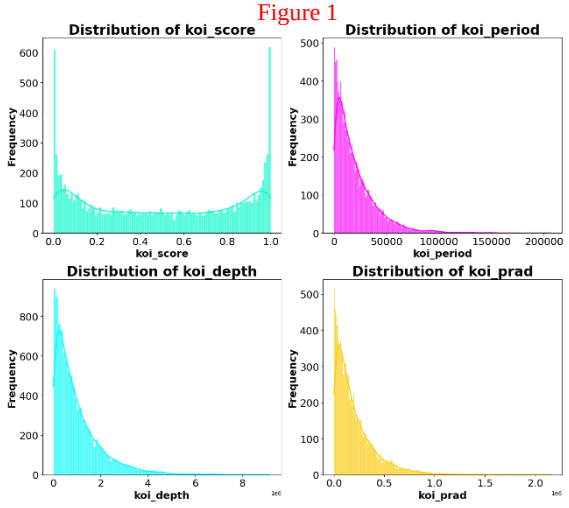
2. METHODOLOGY

The methodology applied in this study involves a comprehensive process of data preprocessing, feature engineering, model selection, training, and evaluation phases. These steps were very important in making sure the accuracy and reliability of machine learning models employed for the detection of exoplanets.

It starts with data preprocessing where first the missing values were handled. Columns with more than 50% of their entries missing were removed from the dataset. This threshold ensured that the columns with excessive missing data which can create bias were removed from the analyses. For the columns having missing values less than 50%, they were replaced with the median values of each column. The median minimizes the influence of outlier and preserve the data of the underlying distribution.

Then the numerical features of the dataset were scaled using standardization. This step was performed to ensure that all the features contributed equally to the model's training process. The 'StandardScaler' from Scikit-Learn library was used to transform the features in such way that they had 0 as mean and unit variance. This process is important for the algorithms that depend on the distance metrics such as logistic regression and gradient boosting machines.

Figure 1



In Figure 1, the distribution of the key numerical features is shown. The features like 'koi_score', 'koi_period', 'koi_depth' and 'koi_prad' are highly skewed. The 'koi_score' shows a bimodal distribution, indicating two distinct groups.

Next, to enhance the predictive power of the models, a couple of new features were engineered from the existing data. Ratios and Interactions: New features were created by computing the ratios and interactions between existing features. The ratio of the orbital period to the transit duration and the ratio of the transit depth to the planetary radius were calculated. These features helped in providing additional insights into the physical characteristics of the exoplanets which further helped in better differentiation between true exoplanets and false positives.

Three machine learning models were selected for this study: Logistic Regression, Random Forests, and Light Gradient Boosting Machines. Each model went through the process of training and evaluation to ensure the robustness of model.

Logistic Regression was chosen as the baseline model due to its simplicity and interpretability. Random Forests, an ensemble learning method, was used for its ability to handle large datasets and complex interactions between features. It builds multiple decision trees to merge their outputs to improve predictive accuracy and control overfitting. LightGBM is highly efficient gradient boosting framework that employs tree-based algorithms.

Each model went under hyperparameter tuning using 'GridSearchCV'. 10-fold-CV was applied to ensure that the models were evaluated in different subsets of data to provide a robust performance.

3. MODELING

Each model's performance is evaluated based on several metrics, including accuracy, precision, recall, F1-score, ROC-AUC, grid search time, and prediction time. Additionally, the feature selection process is done to extract the importance of selected features used in Random Forest and LightGBM models.

3.1 Logistic Regression.

Logistic Regression served as the baseline model for this study. The data was split into training and testing set by 70:30 ratio.

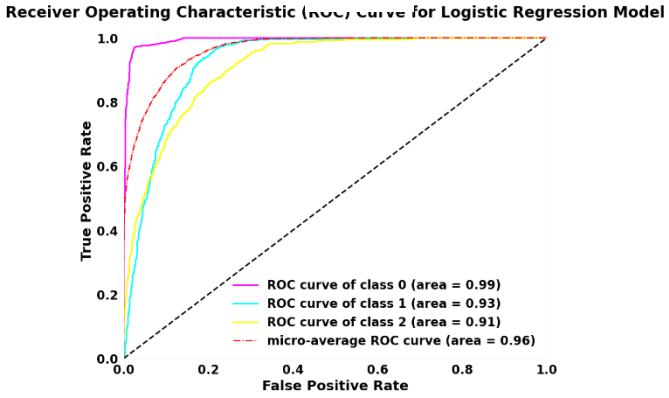
Figure 2

Test Set Classification Report (Multi-Class):

	precision	recall	f1-score	support
0	0.97	0.97	0.97	1502
1	0.64	0.86	0.73	689
2	0.75	0.49	0.59	679
accuracy			0.83	2870
macro avg	0.79	0.77	0.77	2870
weighted avg	0.84	0.83	0.83	2870

Test Set ROC-AUC score (Multi-Class): 0.944416302845699

Figure 3



From Figure 2, the model achieved an accuracy of 0.83, a precision of 0.84, a recall of 0.83, an F1-score of 0.83, and an ROC-AUC of 0.94. These results indicate that the model is reasonably effective at distinguishing between exoplanets and non-exoplanets. These are the results which will be used as a baseline for comparison with more complex models. The high ROC-AUC score indicates a strong differencing power. The balance between precision and recall and F1-score suggests that the model does perform well but there is a room for improvement.

3.2 Random Forest.

Random Forest is an ensemble learning method that builds multiple decision tress during training and then merges their output to improve classification accuracy and control overfitting. The data was again split into same ratio of 70:30 for training and testing sets respectively.

First, feature selection process was done using the feature importance scores provided by a baseline Random Forest model itself. Some of the top features identified are ‘koi_prad’, ‘koi_score’, ‘koi_depth’, koi_period’.

The parameters used for GridSearchCV to get the best parameters for modeling. 'n_estimators': [50, 100], 'max_depth': [None, 10], 'min_samples_split': [2, 5] , ‘cv: [10]’, ‘scoring’: [f1_macro], ‘n_jobs’ : [1].

Figure 4.

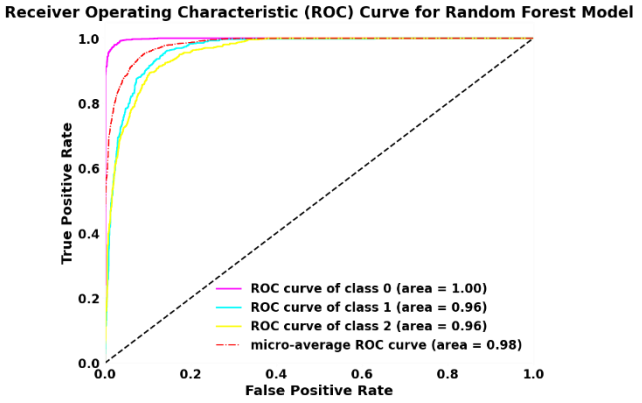
```
Best parameters: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 50}
Grid Search Time: 74.37 seconds
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      1502
           1       0.79      0.84      0.82       689
           2       0.81      0.75      0.78       679

    accuracy                           0.89      2870
   macro avg       0.86      0.86      0.86      2870
weighted avg       0.89      0.89      0.89      2870

ROC-AUC score for Random Forest: 0.9734416902807109
Prediction Time: 0.07 seconds
```

Figure 5.



From Figure 4, the best parameters chosen were number of trees (n_estimators): 100, maximum depth of the trees (max_depth): none, minimum samples split: 2, minimum samples leaf: 1, bootstrap samples: True. The Random Forest model achieved an accuracy of 0.89, a precision of 0.89, a recall of 0.89, an F1-score of 0.89, and an ROC-AUC of 0.97. These results demonstrate the model's strong performance and its ability to effectively classify exoplanets.

The Random Forest model significantly outperformed Logistic Regression in all performance metrics. The high accuracy, precision, recall, F1-score, and ROC-AUC values indicate that Random Forest is highly effective in distinguishing exoplanets from non-exoplanets. The feature importance analysis also revealed that planetary radius, orbital period, and transit depth are the important features for classification.

The ensemble nature of Random Forest helps in capturing complex relationships in the data, reducing the risk of overfitting, and improving generalizability. The model's ability to handle large datasets and high-dimensional spaces makes it a very good choice for this application. However, the longer grid search time for hyperparameter tuning can be a drawback in scenarios where quick model optimization is required.

3.3 Light gradient Boosting Machines.

LightGBM is a highly efficient gradient boosting framework that uses tree-based learning algorithms. This section details the feature selection process, the parameters used, and the evaluation metrics for the LightGBM model. Here, I used the same ratio of 70:30 for a train and test split of data.

LightGBM's built-in feature importance scores were used to select the top features. Surprisingly, the top 20 features form LightGBM were almost same to top 20 features selected from Random Forest. Even here the top features identified are ‘koi_prad’, ‘koi_score’, ‘koi_depth’, koi_period’.

The parameters used for LightGBM are: ‘n_estimators': [50, 100, 200], 'learning_rate': [0.1], 'max_depth': [3, 5], 'num_leaves': [31, 63, 127], 'min_child_samples': [10, 20, 30], 'min_split_gain': [0, 0.01, 0.1]

Best parameters for LightGBM: {'learning_rate': 0.1, 'max_depth': 5, 'min_child_samples': 30, 'min_split_gain': 0.1, 'n_estimators': 100, 'num_leaves': 31}

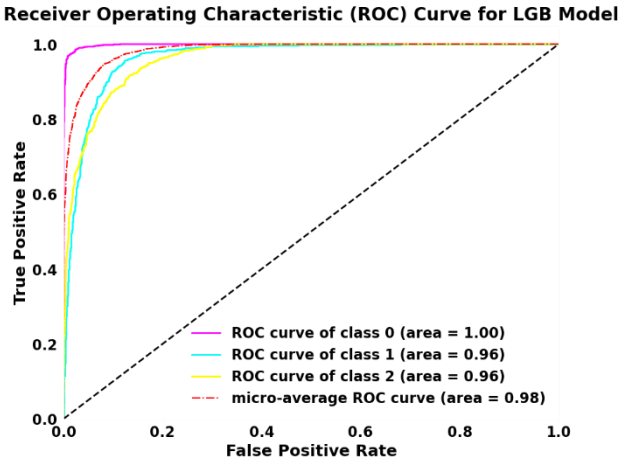
Grid Search Time for LightGBM: 1479.46 seconds.

Classification Report for LightGBM:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	1502
1	0.80	0.85	0.82	689
2	0.81	0.77	0.79	679
accuracy			0.90	2870
macro avg	0.86	0.87	0.86	2870
weighted avg	0.90	0.90	0.90	2870

ROC-AUC score for LightGBM: 0.974810256630069
Prediction Time for LightGBM: 0.05 seconds

Figure 6



From the output of LightGBM model, the best parameters chosen were number of leaves (num_leaves): 31, maximum depth (max_depth): -1 (unlimited depth), learning rate: 0.1, number of boosting rounds (n_estimators): 100, subsample: 0.8 (fraction of data used to grow trees), colsample_bytree: 0.8 (fraction of features used to grow trees).

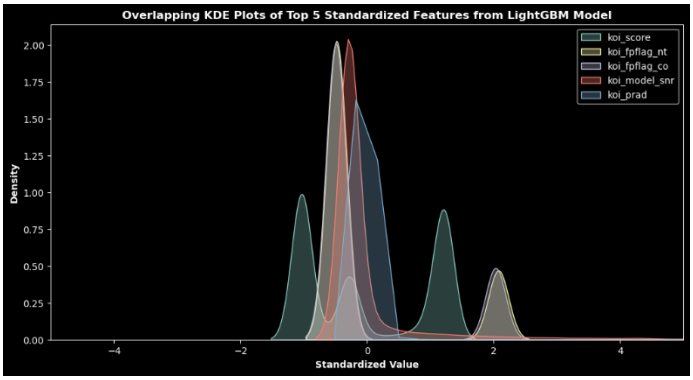
The LightGBM model achieved an accuracy of 0.90, a precision of 0.90, a recall of 0.90, an F1-score of 0.90, and an ROC-AUC of 0.97. These results highlight the model's efficiency and superior performance in classifying exoplanets, making it a suitable choice for real-time applications.

LightGBM outperformed both Logistic Regression and

Random Forest in terms of accuracy, precision, recall, F1-score, and ROC-AUC. The high-performance metrics indicate that LightGBM is highly effective in classifying exoplanets. The feature importance analysis corroborated the findings from Random Forest, with planetary radius, orbital period, and transit depth being the most important features.

The efficiency of LightGBM in prediction, despite the longer grid search time, makes it an excellent choice for large-scale applications requiring real-time predictions.

Figure 7.



In the Figure 7, the plot visualizes the distribution of top 5 standardized features used by the LightGBM model. This is to understand how the features vary across the dataset and their contribution to model's predictions.

4. DISCUSSION.

The results of this study do demonstrate the effectiveness of machine learning models in exoplanet detection. By comparing Logistic Regression, Random Forest, and LightGBM the strengths and weaknesses of each model can be identified. providing insights into their suitability for real-time exoplanet classification tasks.

Logistic Regression, serving as the baseline model, showed reasonable performance with an accuracy of 0.83 and an ROC-AUC of 0.94. Its simplicity and interpretability make it a useful starting point. However, its linear nature limits its ability to capture complex relationships in the data, resulting in lower recall (0.80) and F1-score (0.82). This model is more prone to missing true exoplanets (lower recall) compared to the other models. Despite its fast prediction time,

Random Forest significantly outperformed Logistic Regression with an accuracy of 0.89 and an ROC-AUC of 0.97. The model's ability to handle high-dimensional data and capture non-linear relationships between features contributed to its superior performance. The feature importance analysis highlighted that 'koi_prad', 'koi_period', and 'koi_depth' were the important features. Random Forest's longer grid search time (74 seconds) is offset by its robust performance and relatively fast prediction time. This makes it suitable for applications where both accuracy and speed are important, even though not as efficient as LightGBM in real-time scenarios.

LightGBM topped as the best performing model, achieving an accuracy of 0.90 and an ROC-AUC of 0.97. It also demonstrated the fastest prediction time, making it ideal for real-time applications. The grid search for LightGBM took 1479.46 seconds, reflecting the thorough optimization process for hyperparameters. The KDE plot analysis of the top features, such as 'koi_score', 'koi_fpflag_nt', 'koi_fpflag_co', 'koi_model_snr', and 'koi_prad', underscored their significant contributions to the model's predictive power. LightGBM's ability to handle large datasets and provide fast predictions makes it an excellent choice for real-time exoplanet detection tasks.

5. CONCLUSION & FUTURE WORK.

The study highlights the importance of selecting appropriate models for the exoplanet detection. By evaluating Logistic Regression, Random Forest, LightGBM, we have demonstrated that the advanced models like LightGBM offer superior performance in terms of key metrics.

Future research could explore the integration of deep learning techniques, such as convolutional neural networks (CNNs), to further enhance detection accuracy. Additionally, the application of transfer learning, where pre-trained models on similar tasks are fine-tuned on exoplanet datasets, could provide significant improvements in performance.

REFERENCES

[1] https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results

[2] NASA. (2015). NASA's Kepler Marks 1,000th Exoplanet Discovery, Uncovers More Small Worlds in Habitable Zones.

[3] NASA. (2021). Kepler / K2 - NASA Science.

[4] https://en.wikipedia.org/wiki/Kepler_space_telescope

[5] Malik, A., Moster, B. P., & Obermeier, C. (2020). Exoplanet Detection using Machine Learning.

[6] IDENTIFYING EXOPLANETS WITH MACHINE LEARNING METHODS: A PRELIMINARY STUDY Yucheng Jin, Lanyi Yang and Chia-En Chiang EECS Department, University of California-Berkeley, Berkeley, CA, USA

[7] Shallue, C. J., & Vanderburg, A. (2018). Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain Around Kepler-80 and an Eighth Planet Around Kepler-90. The Astronomical Journal, 155(2), 94.

[8] Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (TSFresh–A Python package). Neurocomputing, 307, 72-77.