Northeastern University
DS4400 Machine Learning
Spring 2023

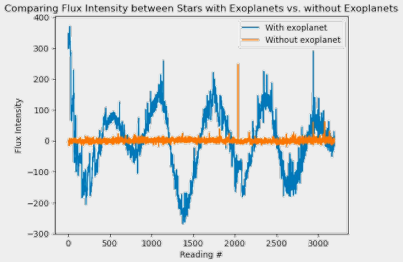
Kepler Exoplanet Exploration

Ayan Chowdhury, Anika Das, Gary Shetye, Grace Yang

Abstract: The Kepler space telescope, launched by NASA in 2009, had a mission to discover planets outside of our solar system, known as exoplanets. In this project, we used k-Nearest Neighbors (KNN), logistic regression, and recurrent neural networks (RNN) to predict if a given star has an exoplanet by using the light intensity (flux) data gathered by the Kepler space telescope. After preprocessing and hyperparameter tuning, we achieved accuracies of 98.3%, 57.3% and 36.1% for KNN, logistic regression, and RNN respectively, indicating that the KNN model is optimal for this project's exoplanet identification.

Background:

- Kepler Mission data measures light intensity (flux) of distant stars
- When exoplanets orbit around a star, slight dimming occurs → transit period
- Identify trends in flux over time to determine whether transit periods are occurring intermittently



Dataset:

- Flux measurements of stars over a period of time
- Source: Kaggle (originally NASA)
- Features: Flux light intensity emitted from star at a given time of measurement
- Label: binary label of "1" - star does not have exoplanet, or "2" - star with exoplanet
- Data is imbalanced and does not follow a normal distribution → use SMOTE for generating minority class samples

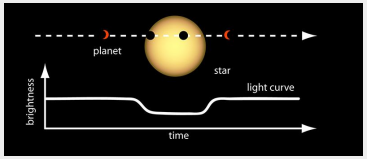


Fig. 1 Light flux decreases when an exoplanet orbits around it

Models & Rationale	Hyperparameters	Accuracy on Testing Data Using Best Hyperparameters
KNN: Non-parametric: makes no assumptions about data	**Distance Function:** [euclidean, manhattan] **# Neighbors:** [3, 5, 7]	**Distance**: Manhattan **# of Neighbors**: 5 **Accuracy:** 98.3%
Logistic Regression: Ability to model binary outcomes with no linear relationship	**Penalty:** [L1, L2] **Solver:** [liblinear, saga]	**Penalty**: L1 **Solver**: liblinear **Accuracy**: 57.3%
Recurrent Neural Networks Handles time-series data well with memory cells + hidden layers	**Training Epochs:** [3, 5, 7] **Learning Rate:** [0.01, 0.001] **LSTM Cells:** [3, 7]	**Training Epochs:** 5 **Learning Rate:** 0.01 **LSTM Cells:** 7 **Accuracy:** 36.1%

Conclusion:

- Conducted supervised learning using KNN, logistic regression, and RNN, which had k-fold average accuracies of 98.3%, 57.3%, and 36.1%, respectively.

Future Directions:

- Use Kepler light curve data in tandem to examine transitory periods.
- Focus on hyperparameter tuning get the highest sensitivity without using the SMOTE function

Acknowledgments: We would like to thank Professor Rachlin for his guidance during this project.