Università
Ca'Foscari
Venezia

Artificial Intelligence and Data Engineering

Statistical Inference and Learning

Project: **'Kepler Objects of Interest: Candidate Prediction'**

Professor
Cristiano Varin

Author
Davide Tonetto
Student ID 884585

Academic year
2024/2025

Contents

1 Dataset overview **1**
- 1.1 Introduction . . . 1
- 1.2 An overview of the dataset . . . 2
  - 1.2.1 Kepler Object of Interest (KOI) . . . 2
  - 1.2.2 Feature Description . . . 3
  - 1.2.3 Target Variable Description . . . 4

2 Analysis **5**
- 2.1 Data preprocessing . . . 5
  - 2.1.1 Correlation matrix . . . 5
- 2.2 Visualizations . . . 5
- 2.3 Principal Component Analysis . . . 7
- 2.4 Models . . . 8
  - 2.4.1 Data preprocessing . . . 8
  - 2.4.2 Outliers and Correlated Features . . . 8

3 Conclusions **10**
- 3.1 Original Features vs. PCA Features . . . 10
- 3.2 Performance Among Original Feature Models . . . 10
- 3.3 Overall Conclusions . . . 11

Chapter 1

Dataset overview

1.1 Introduction

The dataset used for this project is the **Kepler Object of Interest (KOI)** dataset, which is a collection of observations of exoplanets provided by NASA's Exoplanet Archive. In particular, the dataset selected for this statistical analysis aimed at predicting exoplanet status is the Q1-Q17 Data Release 25 (DR 25) Kepler Objects of Interest (KOI) table. This specific table was chosen due to several key characteristics that make it particularly well-suited for developing robust statistical models:

1. **Uniform Data Processing**: The table is derived from Data Release 25, which represents the final, uniform processing of the entire primary Kepler mission dataset (Quarters 1 through 17). This ensures consistency in the underlying light curve data used for analysis.
2. **Automated and Uniform Vetting**: A critical feature of the DR 25 KOI table is its use of a fully automated dispositioning process, known as the Kepler Robovetter. This algorithm applies a consistent set of rules and metrics to uniformly classify each Threshold Crossing Event (TCE) as either a Planetary Candidate (PC) or a False Positive (FP).
3. **Designed for Statistical Analysis**: As explicitly stated in its documentation, this catalog was generated with the primary goal of enabling statistical analyses, such as the calculation of exoplanetary occurrence rates. The emphasis was placed on uniformity and automated, repeatable classification rather than maximizing the accuracy for every single individual object (which might involve manual intervention or external data, as seen in the DR25 Supplemental table).
4. **Homogeneity**: The automated and uniform vetting process yields a homogenous catalog suitable for robust statistical analyses. This internal consistency is crucial for training reliable statistical or machine learning models, as the classification criteria are applied consistently across all entries designated as KOIs within this table.

Therefore, the Q1-Q17 DR 25 KOI table was selected because its **rigorous, automated, and uniform classification methodology provides the most suitable foundation for building a statistical model** intended to predict whether an object is likely an exoplanet based on the parameters derived consistently from the Kepler pipeline and vetting process. This minimizes biases that could arise from aggregating data processed or vetted using different methods over time, as found in the Cumulative table, or from incorporating non-uniform manual assessments, as in the Supplemental table.

1.2 An overview of the dataset

Let's start by understanding what the dataset contains, in particular, what are KOI and the details of each column used in the analysis.

1.2.1 Kepler Object of Interest (KOI)

A **Kepler Object of Interest (KOI)** is a target star observed by NASA's Kepler space telescope that exhibits transit-like signals in its photometric light curve data. These signals meet specific criteria suggesting they might be caused by an object passing in front of (transiting) the star, with characteristics initially consistent with those of an exoplanet.

Specifically, a target star receives a KOI designation when:

- Its Kepler light curve shows at least one sequence of periodic, transit-like dips in brightness.
- The detected signal appears to be of astrophysical origin, rather than being caused by instrumental effects or stellar variability that only mimics a transit.
- The signal's properties (like shape, duration, and depth) are initially consistent with the hypothesis of a planet passing in front of its host star.

Each potential transiting object identified is assigned a unique KOI name, typically in the format KIC_Number.XX or KOI-Number.XX (e.g., K00752.01 or KOI-752.01). The integer part refers to the target star (often linked to its Kepler Input Catalog or KIC number), and the two-digit decimal part identifies a specific transit signature associated with that star (allowing for multi-planet systems).

It is crucial to understand that a KOI designation represents a *candidate* detection. It signifies that an interesting signal worthy of further investigation has been found, but it does not guarantee the presence of a planet. Every KOI must undergo a rigorous vetting process, utilizing detailed analysis of the Kepler pixel data, light curves, and often supplementary follow-up observations from other telescopes. This vetting aims to classify each KOI into more definitive categories:

- **Planetary Candidate (PC):** A KOI that successfully passes all vetting tests designed to rule out common false positive scenarios. It remains consistent with being a genuine planet, pending further confirmation.
- **False Positive (FP):** A KOI whose signal is ultimately determined to be caused by something other than a transiting planet. Common sources of false positives include eclipsing binary star systems (within the target aperture or in the background), stellar variability (like starspots), or instrumental artifacts.
- **Confirmed Planet:** A Planetary Candidate that has been statistically validated or independently confirmed as a planet through methods such as measuring its mass (via radial velocity or transit timing variations) or achieving extremely high confidence through statistical analysis. Confirmed planets typically receive official names (e.g., Kepler-186f).

The various KOI catalogs released by the Kepler mission team are foundational

datasets for the exoplanet community. They provide lists of targets for detailed follow-up observations and serve as the basis for statistical studies aimed at understanding the prevalence and characteristics of planets beyond our solar system (planet occurrence rates).

1.2.2 Feature Description

The dataset contains 8054 rows and 142 columns, each row representing a different object and each column representing a different parameter. The following is a list of the most important columns and their description:

- `koi_period` - **Orbital Period (days):** Time interval between consecutive transits.
- `koi_duration` - **Transit Duration (hours):** Duration of the transit event from first to last contact.
- `koi_depth` - **Transit Depth (ppm):** Maximum fractional decrease in stellar flux during transit (parts per million).
- `koi_prad` - **Planetary Radius (Earth radii):** Estimated radius of the KOI ($R_\oplus$).
- `koi_teq` - **Equilibrium Temperature (Kelvin):** Estimated temperature assuming typical albedo and heat redistribution (K).
- `koi_insol` - **Insolation Flux (Earth flux):** Incident stellar flux relative to Earth.
- `koi_model_snr` - **Transit Signal-to-Noise Ratio:** SNR of the transit detection from model fit.
- `koi_steff` - **Stellar Effective Temperature (Kelvin):** Host star's photospheric temperature (K).
- `koi_slogg` - **Stellar Surface Gravity (log10(cm s^{-2})):** Log base-10 of the star's surface gravity.
- `koi_srad` - **Stellar Radius (solar radii):** Host star's estimated photospheric radius ($R_\odot$).
- `koi_smass` - **Stellar Mass (solar mass):** Host star's estimated mass ($M_\odot$).
- `koi_impact` - **Impact Parameter:** Normalized projected distance between star and KOI centers at mid-transit (Dimensionless).
- `koi_ror` - **Planet-Star Radius Ratio:** Ratio of KOI radius to stellar radius (Dimensionless).
- `koi_srho` - **Fitted Stellar Density (g cm^{-3}):** Mean stellar density inferred from transit shape (g cm^{-3}).
- `koi_sma` - **Orbit Semi-Major Axis (AU):** Half the longest diameter of the orbit (AU).

- `koi_incl` - **Inclination (degrees):** Angle between the orbital plane and the plane of the sky.
- `koi_dor` - **Planet-Star Distance over Star Radius:** Distance at mid-transit normalized by stellar radius (Dimensionless).
- `koi_ldm_coeff1` - **Limb Darkening Coefficient 1:** First coefficient for the limb darkening model used in the fit.
- `koi_ldm_coeff2` - **Limb Darkening Coefficient 2:** Second coefficient for the limb darkening model used in the fit.
- `koi_smet` - **Stellar Metallicity ([Fe/H]):** Log base-10 of the Fe/H ratio relative to solar (dex).

1.2.3 Target Variable Description

`koi_pdisposition` (**Disposition Using Kepler Data (Pipeline Disposition)**) is the target variable used for classification in this analysis. It represents the classification assigned to the Kepler Object of Interest (KOI) by the Kepler data processing pipeline and associated automated vetting procedures (like the Robovetter for DR25). It indicates the pipeline's assessment of the most probable physical nature of the transit-like signal based solely on the analysis of Kepler photometric data. The typical values are:

- **CANDIDATE:** The signal passed the automated vetting tests designed to identify common false positives. It remains consistent with the hypothesis of a transiting planet based on the pipeline's analysis.
- **FALSE POSITIVE:** The signal failed at least one vetting test, suggesting it is likely caused by phenomena other than a transiting planet (e.g., an eclipsing binary star system, instrumental artifact, background contamination).

This disposition may differ from the final classification in the Exoplanet Archive (`koi_disposition`), which incorporates additional information, including human vetting and follow-up observations. For modeling based purely on pipeline-derived features, `koi_pdisposition` is often used as the ground truth label.

Chapter 2

Analysis

2.1 Data preprocessing

In the `data_preparation.rmd` file, the dataset is cleaned from rows with a high rate of missing values and factor variables are prepared for further analysis. Then in the `data_visualization.rmd` file, the dataset is visualized to understand the distribution of the variables and the relationship between them.

2.1.1 Correlation matrix

The following is the correlation matrix of the used columns for the analysis. The correlation matrix shows us that there are some strong relationships between some variables and will help us to understand the dataset better.

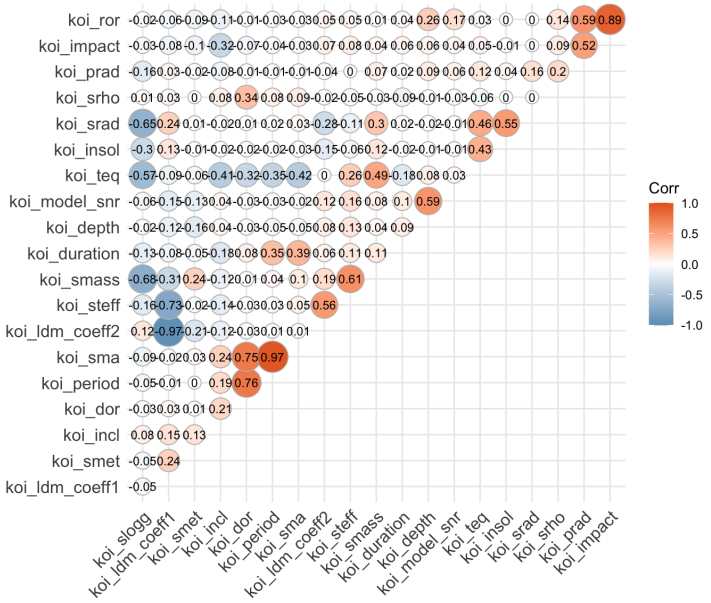


Figure 2.1: Correlation Matrix

2.2 Visualizations

The following plots illustrate the distribution of the variables and their relationships based on the used columns for analysis.

(a) Orbital Period vs. Planetary Radius

(b) Transit Depth vs. Model Signal-to-Noise Ratio

(c) Insolation Flux vs. Planetary Radius

(d) Stellar Metallicity vs. Planetary Radius

(e) Distribution of Orbital Periods

(f) Distribution of Pipeline Dispositions

(g) Distribution of Planetary Radius

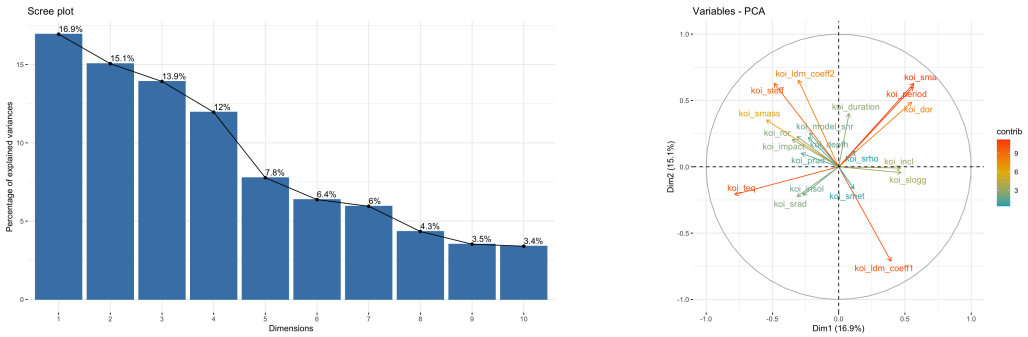
Figure 2.2: Distribution of variables and relationships between them.

- **Orbital Period vs. Planetary Radius:** The plot shows planet radius against orbital period, highlighting known exoplanet populations and false positives.
- **Transit Depth vs. Model Signal-to-Noise Ratio:** Transit depth and SNR are positively correlated, with deeper transits being easier to detect. The plot visualizes false positives and candidates spanning a wide range of depths and SNRs.
- **Insolation Flux vs. Planetary Radius:** The plot examines the relationship between planetary energy receipt, size, and insolation, potentially identifying false positives based on these parameters.
- **Stellar Metallicity vs. Planetary Radius:** The plot investigates the relationship between planet size and star metallicity, examining if CANDIDATEs and FALSE POSITIVEs exhibit distinct trends in this parameter space.
- **Distribution of Orbital Periods:** Most detected KOIs have short orbital periods due to detection bias.
- **Distribution of Pipeline Dispositions:** The plot shows the balance between candidate and false positive KOIs in the Kepler dataset, which is important for model building and evaluation.

- **Distribution of Planetary Radius:** The histogram shows planet candidate sizes, with peaks for common types and a potential dip around 1.5-2 Earth radii. Detection biases influence the distribution.

2.3 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform a large set of variables into a smaller set of uncorrelated variables called principal components. These principal components are linear combinations of the original variables and capture the maximum amount of variance in the data. PCA is particularly useful when dealing with high-dimensional data, where many variables are correlated. The analysis of the PCA results is performed in the `data_visualization.rmd` file and then used to build the models in the `model_pca.rmd` file.



(a) PCA eigenvalues

(b) PCA Loadings

Figure 2.3: PCA Results

Figure 2.3-(a) shows that the first two principal components explain only 32% of the variance, suggesting a complex data structure requiring 11 PCA to capture over 90% of the variance.

Figure 2.3-(b) shows the loadings of the first two principal components allowing us to understand the relationship between the original variables and the principal components. In general, the loadings analyses report the following:

- **PC1 (17% variance):** Captures the relationship between orbital characteristics and temperature, showing high positive loadings for orbital parameters (period, semi-major axis, planet-star distance ratio) and negative loadings for equilibrium temperature. Moderate contributions from stellar properties.
- **PC2 (15% variance):** Primarily represents orbital parameters and stellar temperature relationships, with positive loadings for the orbital period, semi-major axis, and stellar effective temperature, while showing contrasting patterns in limb darkening coefficients.
- **PC3 (14% variance):** Dominated by stellar characteristics, particularly contrasting stellar radius, and insolation (positive loadings) with surface gravity (negative loading). Orbital parameters show moderate influence.
- **PC4 (12% variance):** Reflects planetary size and transit geometry, with

strong negative loadings for planetary radius, radius ratio, and impact parameter.

- **PC5 (8% variance):** Represents transit signal characteristics, showing strong negative loadings for transit depth and model signal-to-noise ratio.
- **Later PCs:** Capture more subtle relationships:
  - PC6: Transit duration and stellar density
  - PC7: Insolation and metallicity relationships
  - PC19/20: Specific period-axis relationships and limb darkening effects

These components suggest that the main sources of variation in the dataset are related to transit signal strength, stellar properties, transit geometry, and orbital characteristics.

2.4 Models

Many different models were tested. The following table shows the results of the models tested by comparing their accuracy, AUC (Area Under the ROC Curve), sensitivity, and specificity (see files `model.rmd` and `model_pca.rmd`).

Model	Acc	Sens	Spec	AUC
GLM	.819	.764	.873	.888
GAM	.827	.789	.865	.893
GLM Int.	.750	.899	.603	.751
Lasso	.809	.720	.896	.882
Ridge	.811	.730	.890	.883
RF PCA	.627	.914	.330	.697
GLM PCA	.725	.464	.979	.869
GAM PCA	.621	.238	.993	.733

Figure 2.4: Models Performance Comparison

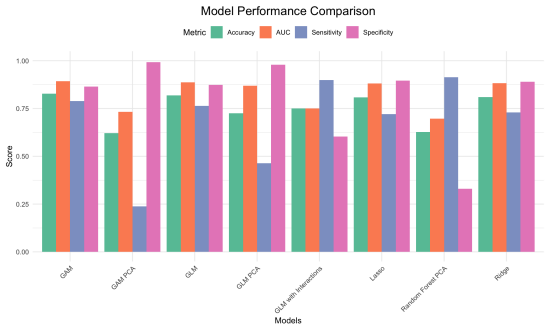


Figure 2.5: Visual Comparison of Models Performance

2.4.1 Data preprocessing

The **null values** in the dataset were **imputed** after splitting the dataset into train and test sets by applying simple median/mode imputation. This simple but effective approach was chosen because it was the most efficient and because the number of missing values was small for the interested columns.

Also, the numerical columns were **scaled** before training the models. Also in this case the scaler is trained on the train set and then applied to the test set to avoid data leakage.

2.4.2 Outliers and Correlated Features

For some models, the **outliers** were **removed** from the dataset by computing the Cook's Distance for each observation and removing those with a Cook's Distance

greater than 0.5. This approach was chosen because it was the most efficient and because the number of outliers was small for the interested columns.

The **correlated features** were **removed** from some models by computing the VIF (Variance Inflation Factor) for each feature and removing those with a VIF greater than 10.

Chapter 3

Conclusions

3.1 Original Features vs. PCA Features

Models trained directly on the original (scaled) features generally demonstrated superior performance compared to those trained on PCA features in this analysis. The GLM, GAM, Lasso, and Ridge models using original features all achieved AUCs above 0.88, whereas the models using PCA features had lower AUCs (0.70-0.87) and often exhibited extreme trade-offs between Sensitivity and Specificity, potentially influenced by the definition of the positive class during evaluation. This suggests that PCA, as implemented here, might have resulted in some information loss detrimental to predictive performance for this specific task compared to using the original, interpretable features.

3.2 Performance Among Original Feature Models

- **GAM**: The Generalized Additive Model (GAM) emerged as the top performer, achieving the highest AUC (0.8930) and Accuracy (0.8274). It also showed a good balance between Sensitivity (0.7894) and Specificity (0.8648). This suggests that capturing non-linear relationships using smooth functions, as identified in the residual analysis, provided a tangible benefit over the standard linear GLM.
- **Baseline GLM**: the standard GLM performed strongly, with an AUC (0.8875) and Accuracy (0.8193) only slightly below the GAM. It represents a solid baseline.
- **Regularized Models (Lasso & Ridge)**: Lasso (AUC 0.8816) and Ridge (AUC 0.8828) performed very similarly to each other and slightly below the baseline GLM and GAM in terms of AUC and Accuracy. However, they achieved the highest Specificity values (0.8958 and 0.8896, respectively). This indicates that regularization helped in correctly identifying the negative class (likely candidate if false positive was positive), potentially by handling collinearity or simplifying the model, but at the cost of slightly lower Sensitivity compared to GLM/GAM.
- **GLM with Interactions**: The specific interaction terms added in this test significantly degraded performance (AUC 0.7510, Accuracy 0.7498), particularly hurting Specificity. This suggests the chosen interactions were not beneficial and may have led to overfitting or were not representative of the true underlying relationships.

3.3 Overall Conclusions

Based on these results, the GAM using the original scaled features stands out as the most promising model, offering the best balance of overall predictive power (highest AUC and Accuracy) while effectively modeling the non-linearities present in the data.

The standard GLM, Ridge, and Lasso models also provide competitive performance and could be considered strong alternatives, especially if model simplicity (Lasso potentially performs feature selection) or high Specificity (Ridge/Lasso) is prioritized. Further tuning of the prediction probability threshold for any of these top models (GAM, GLM, Ridge, Lasso) could optimize the balance between Sensitivity and Specificity based on specific application requirements (e.g., minimizing false negatives vs. minimizing false positives). The PCA-based approaches and the tested interaction GLM appear less effective based on this comparison.