KeplerInsights

Amitha, Herbert, Lucas, Michael

Background & Objective

- The scientific objective of the Kepler Mission is to explore the structure and diversity of planetary systems, using a special-purpose spacecraft to measure light variations from thousands of distant stars, looking for planetary transits.

- Kepler Objects of Interest (KOIs) are well vetted, periodic, transit-like events in the Kepler data. The Kepler Project identifies these objects from the Threshold-Crossing Events (TCE) list for further vetting. Some objects will be flagged as false positives.

Objective:

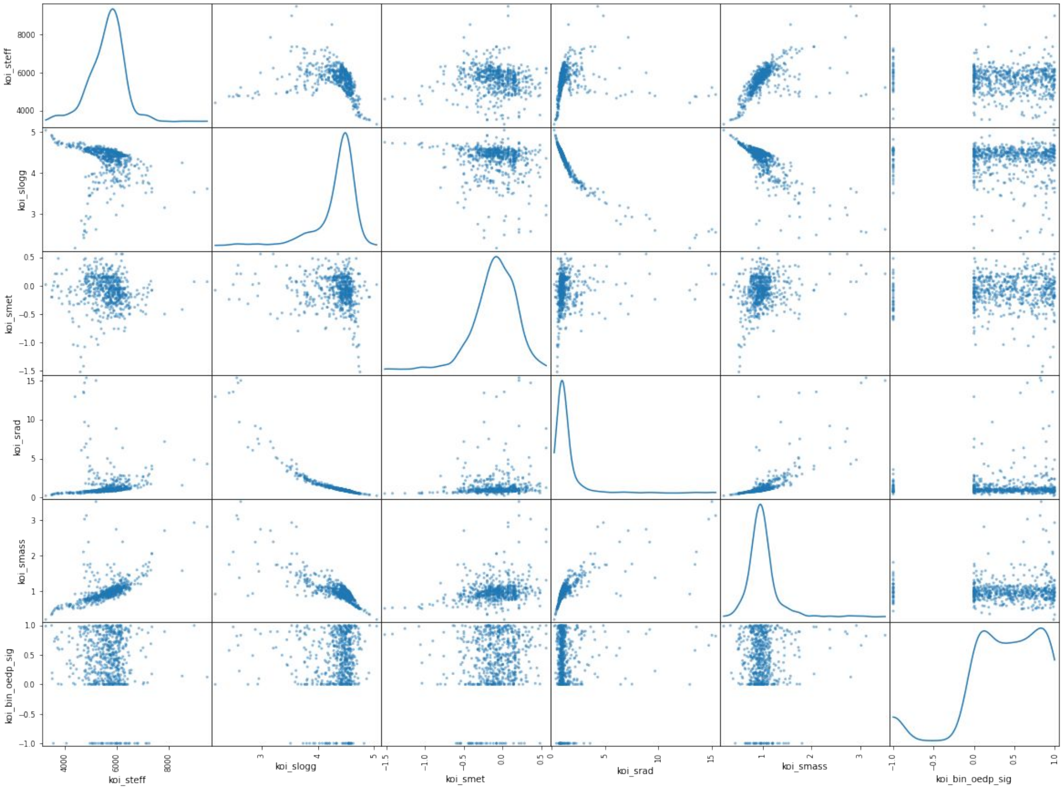
- Predict classification of KOI (exoplanet candidate or false positive)

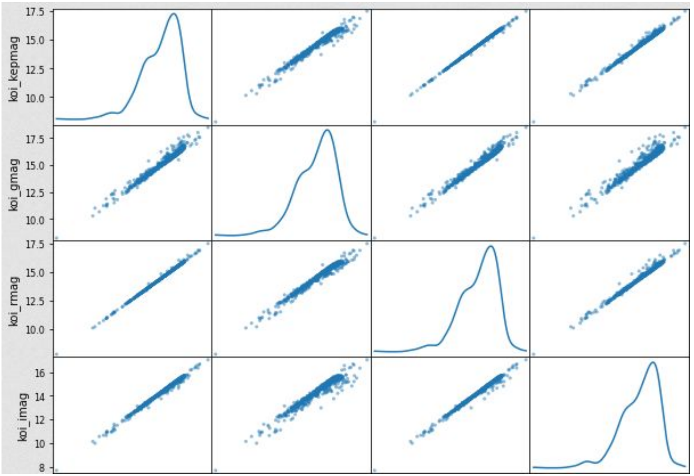
Data Exploration

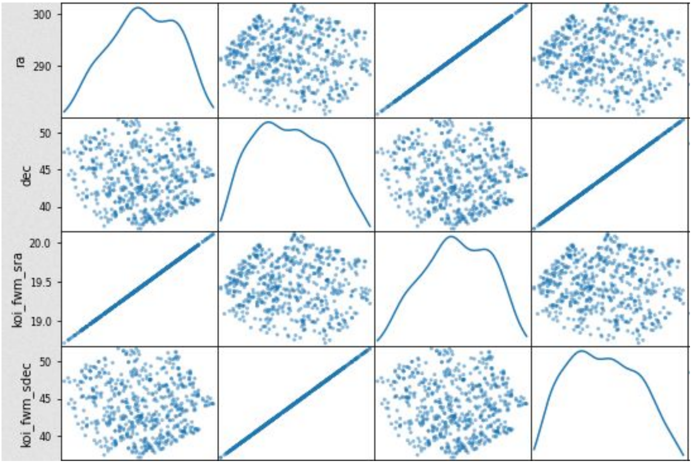
- Dataset with 82 features, 9564 instances
  - Identification Columns, Exoplanet Archive Information, Project Disposition Columns
  - Transit Properties, Threshold-Crossing Event (TCE) Information, Stellar Parameters, Kepler Input Catalog (KIC) Parameters, Pixel-Based KOI Vetting Statistics

- Plot scatter matrix of features to evaluate data distributions, trends and understand relationship with other features
  - Mix of distributions: majority left/right skewed, some normally distributed
  - Some outliers observed

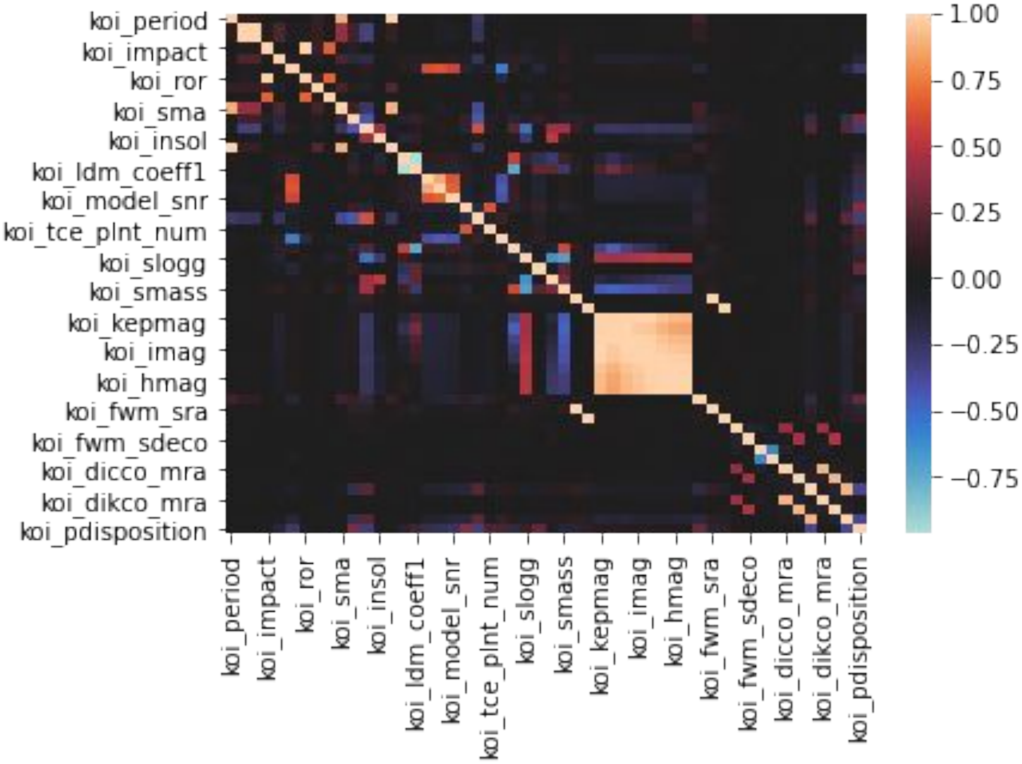
- Plot correlation to identify redundant features
  - High correlation found in subset of features: KIC parameters

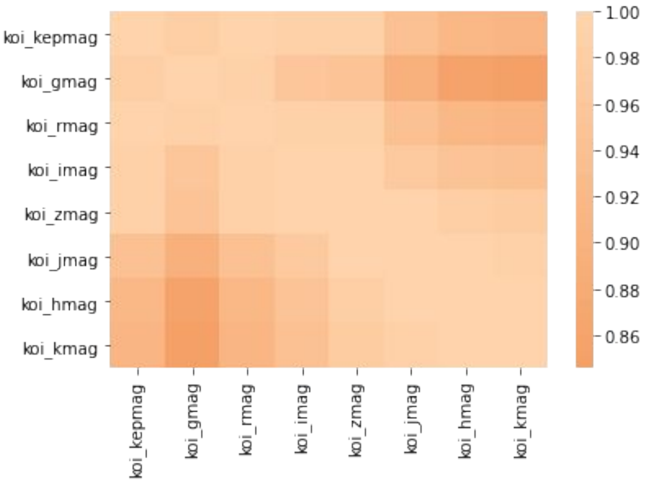


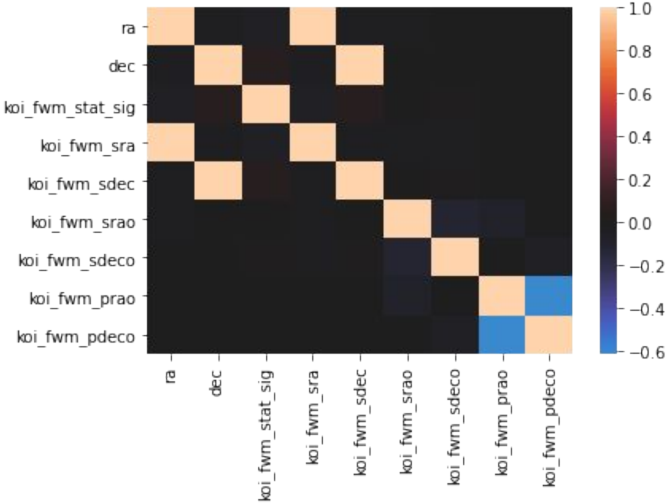




Feature distributions







Correlation analysis

Data Preparation

- Removed identification and documentation information from dataset
  - Identification Columns, Exoplanet Archive Information, Project Disposition Columns

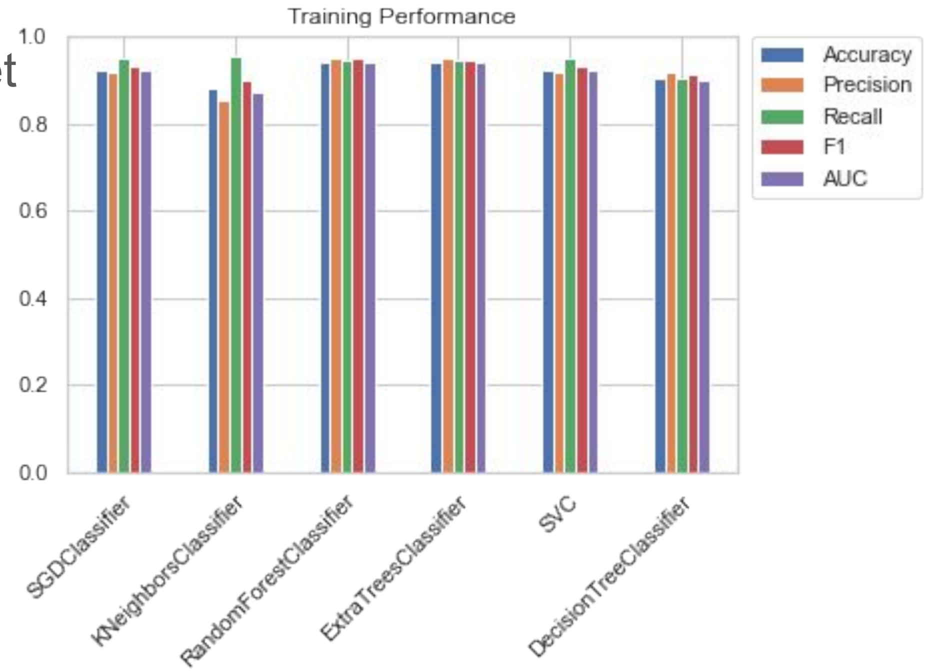
- Examine rows with missing data, evaluate options for dataset
  - Remove all rows with any missing values
  - Remove rows with a threshold of missing values, impute remainder
  - Impute all missing values

Modelling

Train on default models with clean dataset

Performance ranking based on accuracy:

- Random Forest
- Extra Trees/Linear
- SVC
- Decision Tree
- KNN



In [24]:
```
# 51 feature dataset - no missing values
default_metrics = pd.DataFrame(fit_metrics, columns=metric_names, index=model_names)
default_metrics
```

Out[24]:

	Accuracy	Precision	Recall	F1	AUC
SGDClassifier	0.933433	0.942590	0.938830	0.940706	0.932663
KNeighborsClassifier	0.875841	0.843091	0.957447	0.896638	0.864193
RandomForestClassifier	0.941660	0.948138	0.948138	0.948138	0.940736
ExtraTreesClassifier	0.938669	0.950269	0.940160	0.945187	0.938456
SVC	0.918474	0.912709	0.945479	0.928805	0.914620
DecisionTreeClassifier	0.902767	0.920270	0.905585	0.912869	0.902365

Tuning

- Hyperparameter tuning, apply grid search with cross-validation
  - KNN, SVC, Random Forest, Extra Trees

- Train model with different datasets
  - Imputed data, replace missing with 0s
  - Drop redundant features (based on correlation analysis)

- Apply dimensionality reduction
  - PCA on all features
  - PCA on subset of highly linear features (based on pairwise distributions)

Models	Grid Search options	Best parameters	Best score (accuracy)
KNN	'weights': ['uniform', 'distance'] 'neighbors': [5,10,20,30] leaf_size = [5,10,30,50] p=[1,2]	'n_neighbors': 5 'weights': 'uniform' ‘leaf_size’ = ‘5’ p=’1’	0.88312
SVC	'kernel' : ['linear', 'rbf', 'poly'] 'degree' : [0, 1, 2, 3, 4, 5, 6] 'C' : [1,5,10,1000]	'C': 5 'degree': 0 'kernel': 'rbf'	0.91898
Random Forest	'n_estimators' : [500, 1000, 1500] 'max_leaf_nodes' : [15, 20, 25] 'max_depth' : range(8, 11)	'max_depth': 9 'max_leaf_nodes': 25 'n_estimators': 500	0.91711
Extra Trees	'n_estimators': [75, 100, 125, 150] 'max_depth': [30, 35, 40, 45] 'min_samples_split': [10, 15, 25, 45]	'max_depth': 35 'min_samples_split': 20 'n_estimators': 100	0.92385

Hyperparameter tuning results

Conclusion

- Model tuning did not lead to significant improvements in performance
    - Best model is Random Forest (based on accuracy score)
- Similar for models trained with imputed data and redundant features removed
    - Imputed data resulted in much lower performance compared to baseline
    - Redundant features removed resulted in better performance, but still lower than baseline
- Performance with dimensionality reduction
    - PCA on all features resulted in lower performance compared to baseline
    - PCA on subset resulted in better performance, similar to redundant features removed

Data acknowledgement

This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program.