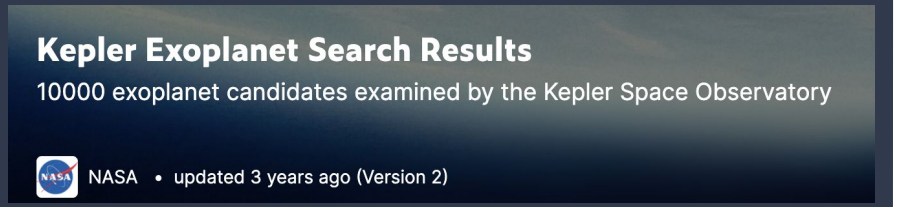
Classifying exoplanets with Kepler data

Trevor Santiago

Data set



Kepler Objects of Interest Data from Kaggle

- Target feature: koi_pdisposition
  - CONFIRMED v. FALSE POSITIVE
- Other interesting features:
  - koi_period: orbital period
  - koi_duration: transit duration
  - koi_kepmag: estimated star magnitude in Kepler band
  - koi_srad: star radius

Data source

Purpose

Space is cool!

How accurately can we predict whether one of the objects of interest will be confirmed an exoplanet or a false positive?

Which features are most important for making these predictions?

Procedure

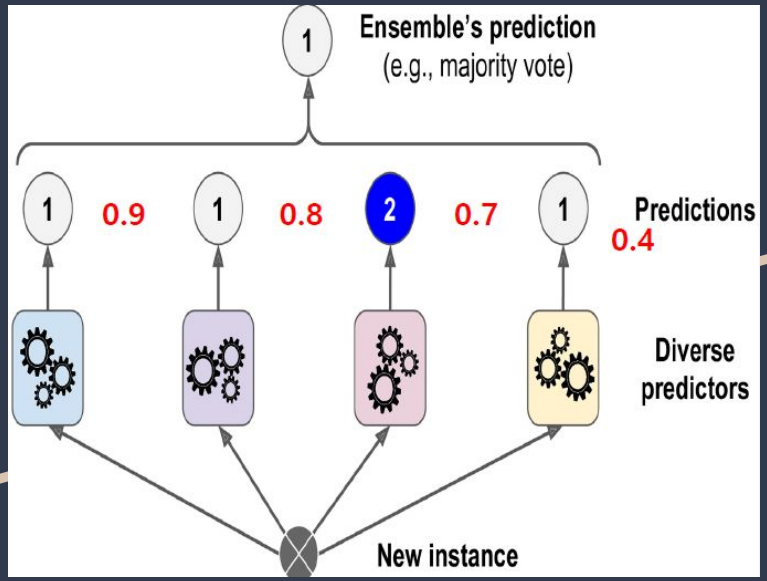
- Split data (3-way)
- Define base pipeline

```
num_prep = Pipeline([
    ('imp', SimpleImputer(strategy='median')),
    ('scale', RobustScaler())
])

# Bin koi_kepmag
binner = ColumnTransformer([('bin', KBinsDiscretizer(), [0])], remainder='passthrough')
cat_prep = Pipeline([
    ('imp', SimpleImputer(strategy='most_frequent')),
    ('bin', binner)
])

prepper = ColumnTransformer([
    ('numeric', num_prep, num_cols),
    ('catergorical', cat_prep, cat_cols)
], remainder='drop')
```

Procedure (Cont.)



- Fit Baseline model (KNN)
- Explored bagging/voting ensembling methods with additional simple models
- Best validation metrics from starter models:
  - Model: VotingClassifier with knn, decision tree, logistic regression
  - Accuracy: 80.8%
  - Log-loss: 0.43038

Image Source

Procedure (Cont.)

- Explored more robust models like RandomForest and xgboost with automated hyperparameter search
- Final model: StackingClassifier
  - Base learners: same 3 simpler algorithms from voting
  - Meta-learner: Best estimator output from Cross-validated hyperparameter search

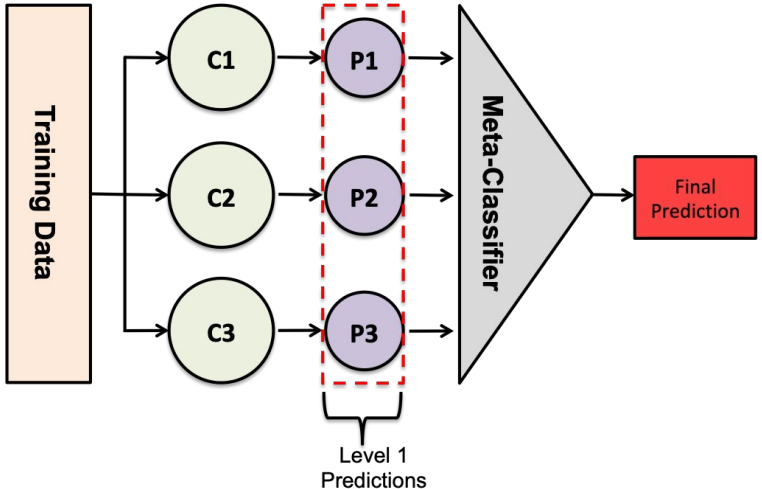


Image Source

Results

- Final model metrics on hold-out test set:
  - Accuracy: 85.9%
  - Log-loss: 0.32342

- Top 3 important features using permutation importance:
  - koi_duration
  - koi_model_snr: normalized transit depth (percent flux blocked by object)
  - koi_period