DISCOVERGING EXOPLANETS FROM NOISY LIGHT CURVESUSING FOURIER SERIES

38428-02 Machine Learning
Prof. HyungJune Lee

Seoyeon Kim (2071008), Ganghee Lee (2076285)

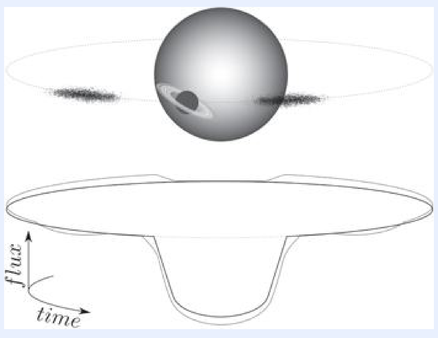
Problem Statement

Situation

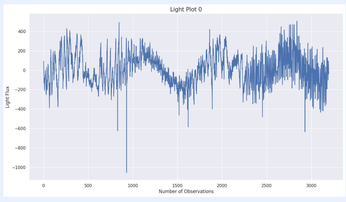
- When an exoplanet passes in front of its host star, it blocks a detectable portion of the star's light.
- Exoplanets show periodic brightness variation patterns as they orbit their host stars.
- However, when measuring the brightness of these exoplanets, significant noise is included due to interference from nearby stars, telescope sensor noise, and space environment factors.

Goal

- We aim to predict the presence of exoplanets by analyzing the brightness variations of their host stars.



Dataset



	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9	...
0	2	93.85	83.81	20.10	-26.98	-39.56	-124.71	-135.18	-96.27	-79.89	...
1	2	-38.88	-33.83	-58.54	-40.09	-79.31	-72.81	-86.55	-85.33	-83.97	...
2	2	532.64	535.92	513.73	496.92	456.45	466.00	464.50	486.39	436.56	...
3	2	326.52	347.39	302.35	298.13	317.74	312.70	322.33	311.31	312.42	...
4	2	-1107.21	-1112.59	-1118.95	-1095.10	-1057.55	-1034.48	-998.34	-1022.71	-989.57	...

- A dataset measuring brightness every 30 minutes using NASA's Kepler space telescope
- Collected brightness variations over time for approximately 5,087 stars
- The first 37 stars do not have exoplanets, while the remaining 5,050 stars do, resulting in an imbalanced dataset

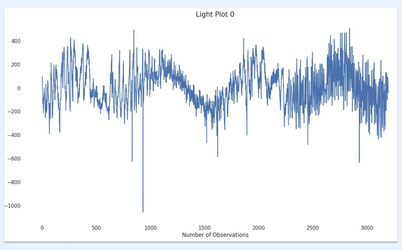
Dataset Modeling

$$F_{\text{obs}} = \bar{F}_* + \Delta F + \bar{F}_* I^{\mathcal{S}},$$

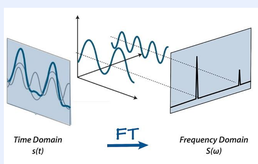
- Assuming the distribution of periodic noise

EXTRACTING PERIODIC TRANSIT SIGNALS FROM NOISY LIGHT CURVES USING FOURIER SERIES

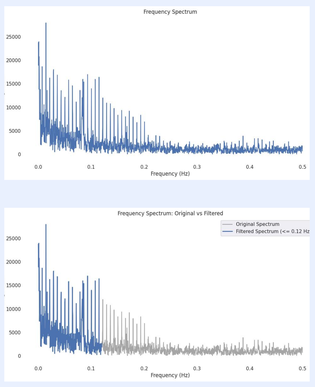
In Noisy light curve, Focusing on the low-frequency band (<1/9 Hz) highly probable with planetary transit signals



Light Flux in Time series

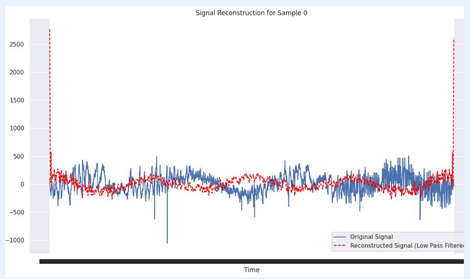


Fourier transform, FT

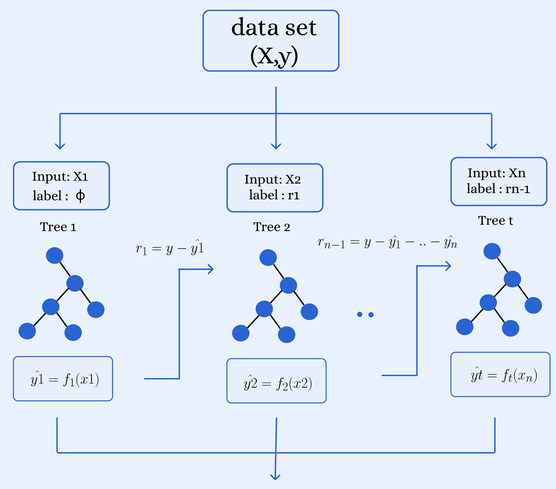


Low pass filter (<1/9 Hz)

Inverse Fourier transform, IFT

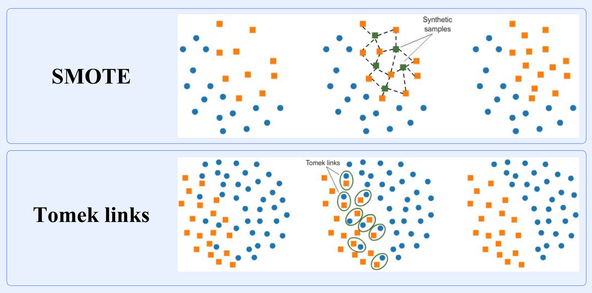


Model Architecture : XGBoost

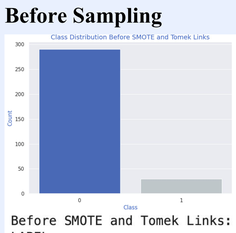


$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i), \quad f_t \in F$$

Data Sampling using SMOTE and Tomek links

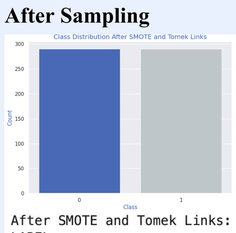


Before Sampling




```
Before SMOTE and Tomek Links:
LABEL
0    290
1     30
```

After Sampling

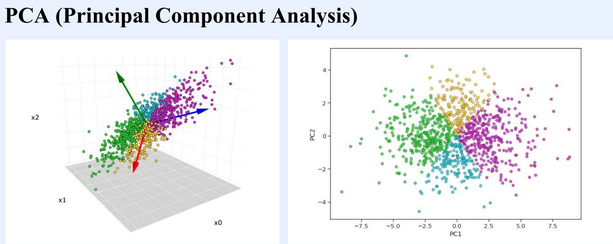



```
After SMOTE and Tomek Links:
LABEL
0    290
1    290
```

Before SMOTE and Tomek Links, the dataset was highly imbalanced, with the majority class dominating. After sampling, the dataset becomes balanced, ensuring equal representation of both classes.

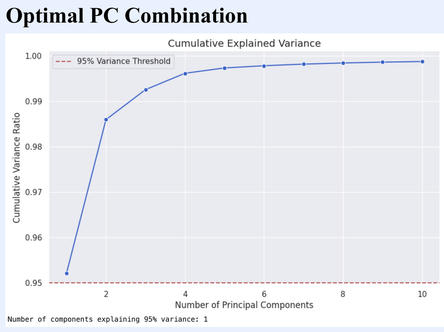
Feature Dimension Reduction using PCA

PCA (Principal Component Analysis)



By analyzing the cumulative explained variance ratio, it is confirmed that 5 PCs provide a sufficiently explanatory representation of the data. The high-dimensional FLUX data is reduced using PCA by selecting 5 principal components that retained key information while filtering out noise and irrelevant features.

Optimal PC Combination




```
Best Combination of Principal Components: ('PC2', 'PC5', 'PC7', 'PC9', 'PC10')
Best Cross-Validation Accuracy Score: 0.9362068965517242
```

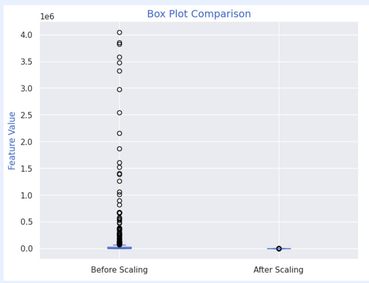
Feature Scaling using Robust Scaling

Robust Scaling

$$X_{\text{scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)}$$

Component	Definition
X_{scaled}	The scaled data value.
X	The original data value.
$\text{median}(X)$	The median (central value) of the data.
$\text{IQR}(X)$	The interquartile range: $\text{IQR} = Q3 - Q1$.
$Q3$	The third quartile (75th percentile).
$Q1$	The first quartile (25th percentile).

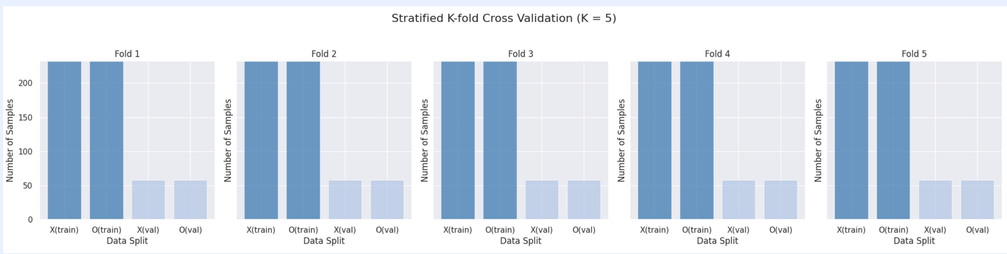
Results of Feature Scaling



Before Robust Scaling, outliers distorted the data distribution, with most values densely clustered in a narrow range. After scaling, the influence of outliers is removed, and the data is redistributed more stably based on the median and interquartile range.

Model Evaluation using Stratified K-Fold Cross Validation

Stratified K-Fold Cross Validation



Stratified K-Fold is used to split the data into 5 folds, ensuring that the class proportions are maintained in each fold. Each fold is used once as validation data, and training-validation is repeated to calculate the average performance, thereby evaluating the model's generalization capability.

Result

1) Pure Data

모델	데이터 전처리 기법	데이터 불균형 기법	정확도 (Accuracy)	정밀도 (정상 샘플)	재현율(정상 샘플)	F1 점수(정상 샘플)	검증 기법
XGBoost	X	전체 데이터셋 n = 5050 + 37					StratifiedKFold k_fold = 2
		언더샘플링 (n=800)	0.95	0.50	0.10	0.17	train_test_split, random_state= 2
		언더샘플링 (n=200)	0.84	1.00	0.38	0.56	train_test_split t_random_stat e=2
		언더샘플링 (n=74)	0.63	0.75	0.55	0.63	train_test_split, random_state= 2
		오버샘플링 (n=400) 양성 샘플 10배 증강 0 : 275 1 : 250	0.90	0.67	0.33	0.44	train_test_split, random_state= 2
		scale_pos_weight(400/37) n=400 weight = 10	0.90	0.67	0.33	0.44	train_test_split, random_state= 2

2) Fourier Transform

데이터 전처리 기법	데이터 불균형 기법	정확도 (Accuracy)	정밀도 (Precision)(양성 샘플)	재현율(양성 샘플)	F1 점수	검증 기법
푸리에 변환 + 저주파(< 1/9 Hz) 필터링	전체 데이터셋 n = 5050 + 37	0.99	1.00	0.14	0.39	Stratifi k_fold
	언더샘플링 (n=400)	0.92	0.58	0.30	0.39	Stratifi k_fold
	언더샘플링 (n=200)	0.82	0.52	0.43	0.47	Stratifi k_fold
	언더샘플링 (n=800)	0.96	**1.00**	**0.67**	0.80	train_t random 2
	언더샘플링 (n=200)	**0.99**	**1.00**	**0.92**	**0.96**	**train_t t_rando e=2**
	언더샘플링 (n=74)	0.80	0.36	0.83	0.5	train_t random 2
	오버샘플링 (n=400) 양성 샘플 10배 증강 0 : 275 1 : 250	0.93	0.86	0.50	0.63	train_t random 2
	scale_pos_weight(400/37) n=400 weight = 10	0.94	1.00	0.44	0.62	train_t random 2
	scale_pos_weight weight = 100	0.94	0.88	0.58	0.70	train_t random 2

3) Fourier Transform + α

데이터 전처리 기법	데이터 불균형 기법	모델	재현율(정상샘플)	재현율(양성샘플)	검증 기법	하이퍼이 퍼능
X	SMOTE, Tomek Links, RobustScaling, PCA	XGBoost	0.88	0.29	Stratified K-Fold Cross Validation	
X	X	XGBoost + RandomizedSearchCV	0.99	0.14	Stratified K-Fold Cross Validation	
푸리에, RobustScaling	SMOTE, Tomek Links, PCA	XGBoost	0.89	0.71	Stratified K-Fold Cross Validation	
푸리에 + 저주파만 RobustScaling,	SMOTE, Tomek Links, PCA	XGBoost	0.84	0.43	Stratified K-Fold Cross Validation	
푸리에 + RobustScaling+ PCA	SMOTE, Tomek Links	XGBoost	0.86	0.86	Stratified K-Fold Cross Validation	Ramndomized Search 튜닝 결과 (subsample: 0.8, reg_lambda: 0.5, reg_alpha: 1, n_estimators: 200, min_child_weight: 1, max_depth: 5, learning_rate: 0.2, colsample_bytree: 0.8)
푸리에 + RobustScaling	SMOTE, Tomek Links	LSTM	0.92	0.14	단순 분할	
푸리에 + RobustScaling + PCA	SMOTE, Tomek Links	LSTM	0.88	0.43	Stratified K-Fold Cross Validation	
푸리에 + RobustScaling + PCA,	SMOTE, Tomek Links	GRU	0.90	0.29	단순 분할	
푸리에 + RobustScaling + PCA,	SMOTE, Tomek Links,	GRU	0.97	0.43	Stratified K-Fold Cross Validation	

Conclusion

- With pure data, focusing on precision typically achieved 50-70%, while focusing on recall achieved 60-100%
- Using Fourier transform, model has improved for the most part
- The model focusing on low-frequency signals through Fourier Transform demonstrated better performance
- There is a some correlation between the low-frequency signal band and the presence of exoplanets
- Understanding the limitations of imbalanced data, the model's performance is up to maximum of 70% recall
- With only 37 positive exoplanets, It is insufficient to build a robust machine learning model for discovering other planets

References

EXTRACTING PERIODIC TRANSIT SIGNALS FROM NOISY LIGHT CURVES USING FOURIER SERIES
https://www.kaggle.com/code/marcinrutecki/smote-and-tomek-links-for-imbalanced-data
https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d
https://www.geeksforgeeks.org/feature-scaling-part-3/

Ops

https://github.com/ML-Ewha-24-2/Discovering_Exoplanets