Exoplanet Hunting

Kepler Labelled Time Series Data Analysis

Reported by: Zixuan Huang, Zhilin Wang & Renee Adonteng

Outline

- **Introduction**
  - Problem statement
  - Data source
  - Code Review
- **Data preprocessing**
  - Scaling; normalization
  - Outliers removing; data balancing……
- **Models and Evaluation**
  - MLP
  - CNN
  - RNN
- **Summary**
  - Comparison; Contribution; Limitation

Introduction -- Problem statement

Background: Search for the New earths

Goal: Develop neural networks to successfully model data for the detection of exoplanets (binary target; time series features)

Framework: Keras

Introduction -- Data source

Data source: Kaggle; NASA Kepler Space Telescope

Brief description of dataset:

	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9	...	FLUX.3188	F
0	2	93.85	83.81	20.10	-26.98	-39.56	-124.71	-135.18	-96.27	-79.89	...	-78.07	
1	2	-38.88	-33.83	-58.54	-40.09	-79.31	-72.81	-86.55	-85.33	-83.97	...	-3.28	
2	2	532.64	535.92	513.73	496.92	456.45	466.00	464.50	486.39	436.56	...	-71.69	
3	2	326.52	347.39	302.35	298.13	317.74	312.70	322.33	311.31	312.42	...	5.71	
4	2	-1107.21	-1112.59	-1118.95	-1095.10	-1057.55	-1034.48	-998.34	-1022.71	-989.57	...	-594.37	
5	2	211.10	163.57	179.16	187.82	188.46	168.13	203.46	178.65	166.49	...	-98.45	

- Trainset:
  - 5087 observations
  - 3198 features
    - Column 1 - Label 0: exoplanet star
    - Label 1: non-exoplanet star
    - Column 2 - Flux values over time
- Testset:
  - 570 observations
  - 3198 features (Same as Trainset)
  - 5 confirmed exoplanet-stars
  - 565 non-exoplanets-stars

	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9	...	FLUX.3188
0	2	119.88	100.21	86.46	48.68	46.12	39.39	18.57	6.98	6.63	...	14.52
1	2	5736.59	5699.98	5717.16	5692.73	5663.83	5631.16	5626.39	5569.47	5550.44	...	-581.91
2	2	844.48	817.49	770.07	675.01	605.52	499.45	440.77	362.95	207.27	...	17.82
3	2	-826.00	-827.31	-846.12	-836.03	-745.50	-784.69	-791.22	-746.50	-709.53	...	122.34
4	2	-39.57	-15.88	-9.16	-6.37	-16.13	-24.05	-0.90	-45.20	-5.04	...	-37.87
5	1	14.28	10.63	14.56	12.42	12.07	12.92	12.27	3.19	8.47	...	3.86

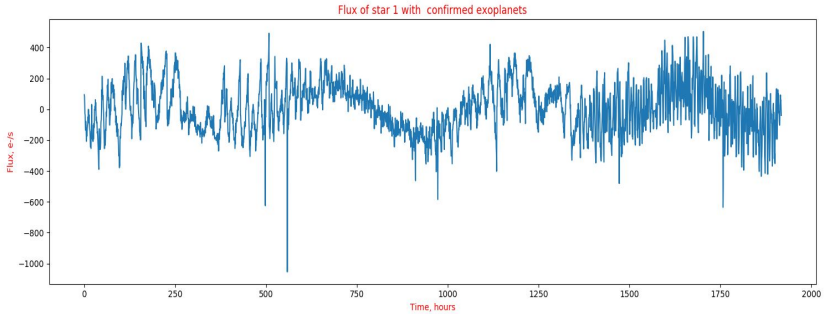
Introduction -- Code review

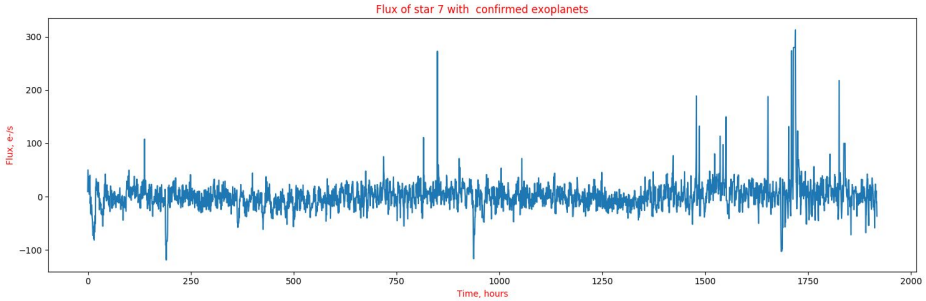
What can we do?

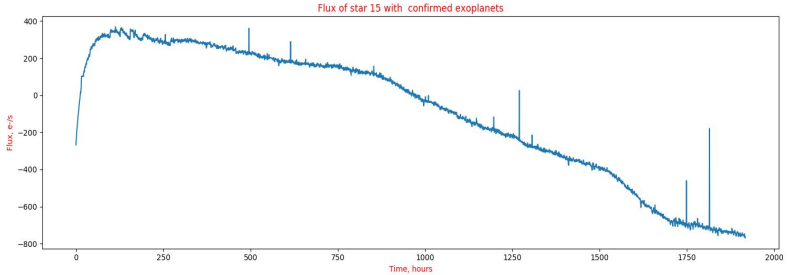
- Try different preprocessing methods
- Address overfitting problem
- Improve model performance
- Evaluate model from several perspectives

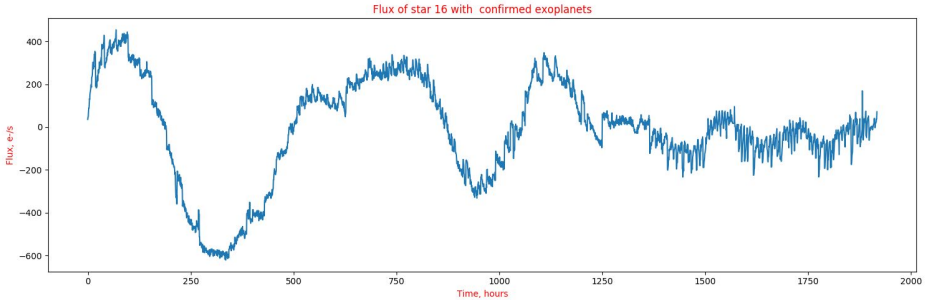
Data preprocessing

Plot features before preprocessing:





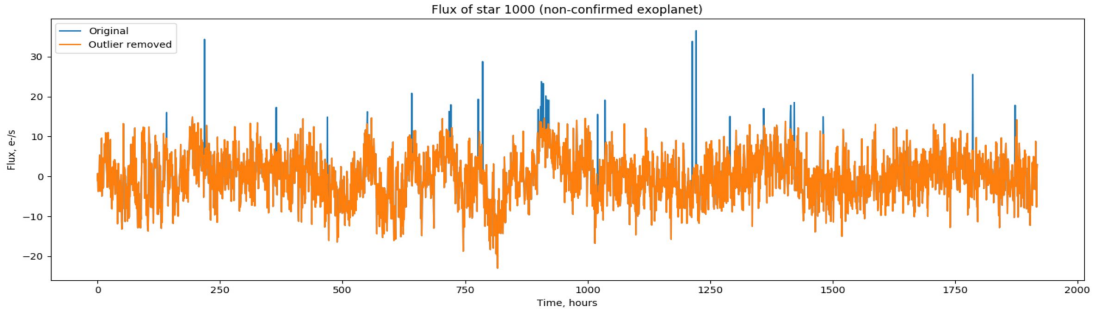


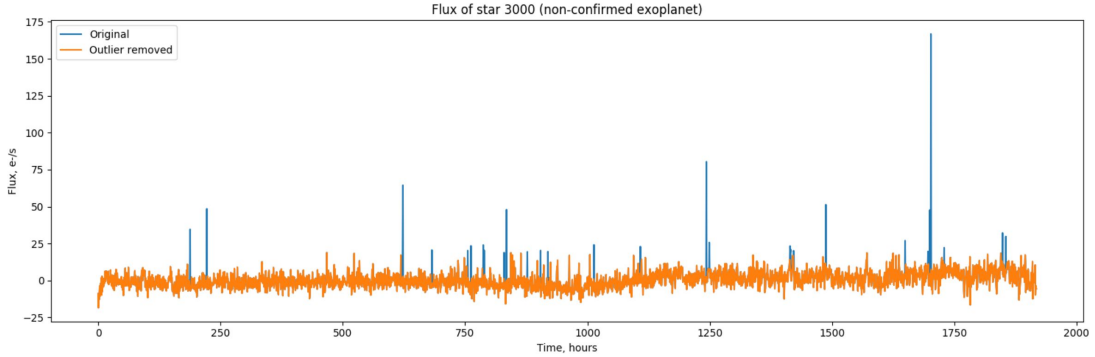


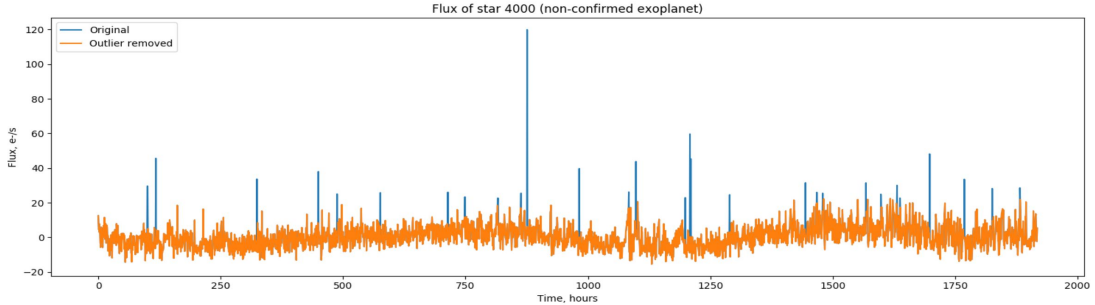
Data preprocessing

- **Outliers Removing**

  Since we were looking for dips in flux when exoplanets pass between the telescope and star, we removed any upper outliers







Data preprocessing

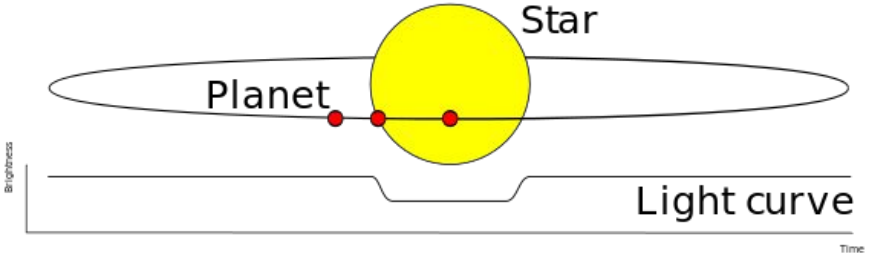
- **Data balancing**

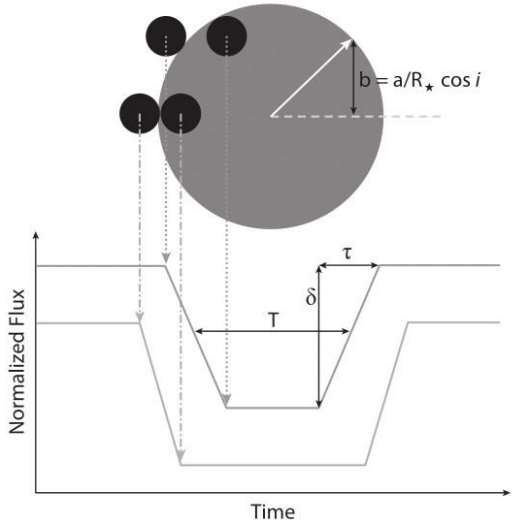
Oversampling, Undersampling

The original data is extremely imbalanced. It has thousands of negative samples but only 29 positive for training.

Fortunately, the light levels received over time by earth follow to a periodic pattern due to the orbiting feature of the planet.

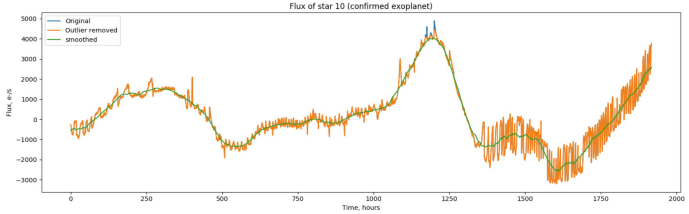
By rotating the flux of an existed exoplanet, we get a dummy exoplanet.

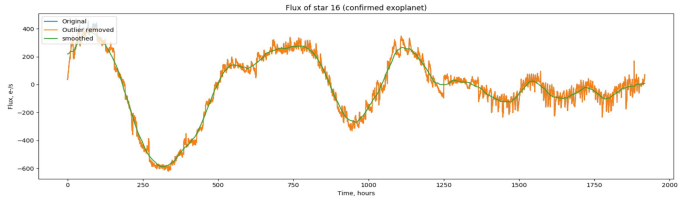


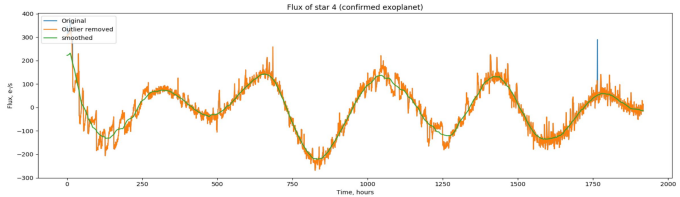


Data preprocessing

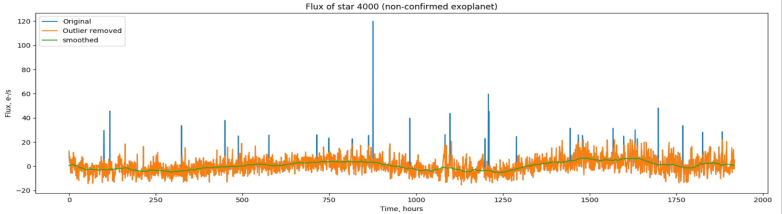
Exoplanet

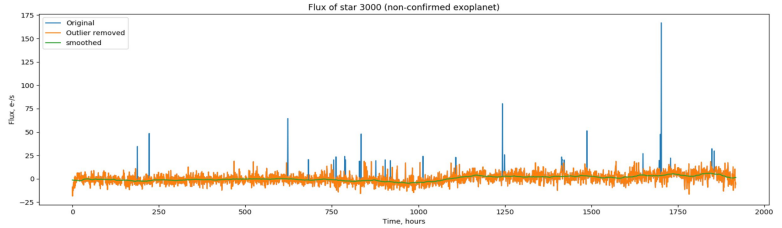


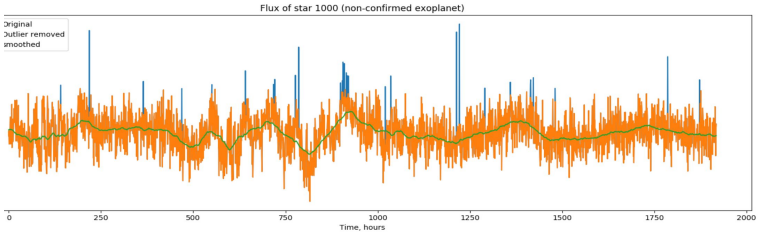




Non-Exoplanet





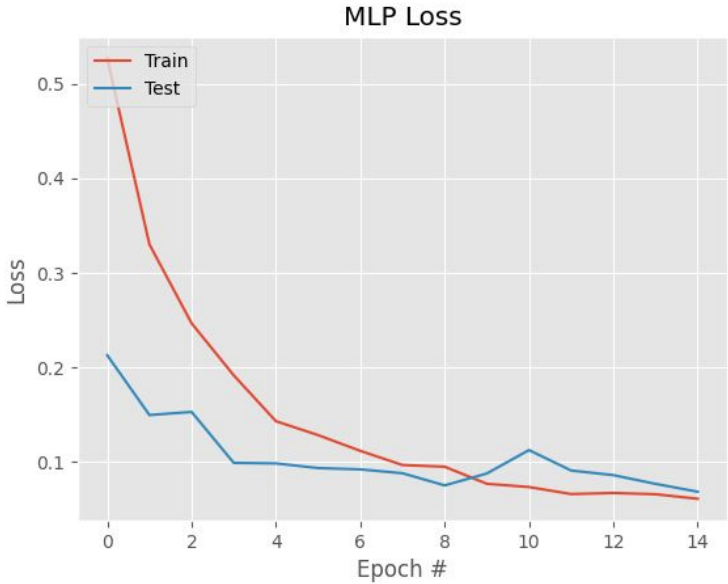


Models and Evaluation - MLP

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 64)	204672
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2080
dropout_2 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 8)	264
dropout_3 (Dropout)	(None, 8)	0
dense_4 (Dense)	(None, 1)	9

Total params: 207,025
Trainable params: 207,025
Non-trainable params: 0

Loss Plot



Dense 64, 32, 8, 1
Dropout 0.25 between each layer

Models and Evaluation - MLP

Optimizer = Adam, Batch Size = 32, Epochs = 15

Early Stopping

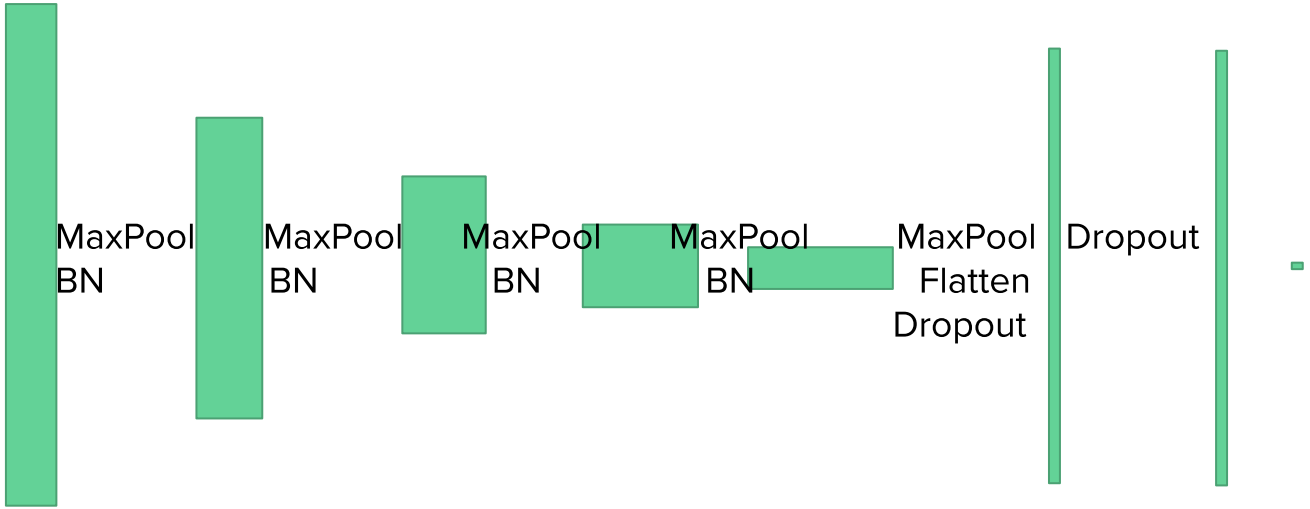
Model Checkpoint

Evaluation:

		precision	recall	f1-score	support
Exoplanet	0	0.99	0.99	0.99	565
Non-exoplanet	1	0.29	0.40	0.33	5
accuracy				0.99	570
macro avg		0.64	0.70	0.66	570
weighted avg		0.99	0.99	0.99	570

Models and Evaluation -- CNN

Model structure



3197x8 conv1d | 1583x16 conv1d | 781x32 conv1d | 380x64 conv1d | 180x128 conv1d | 64 dense | 64 dense | 1 dense

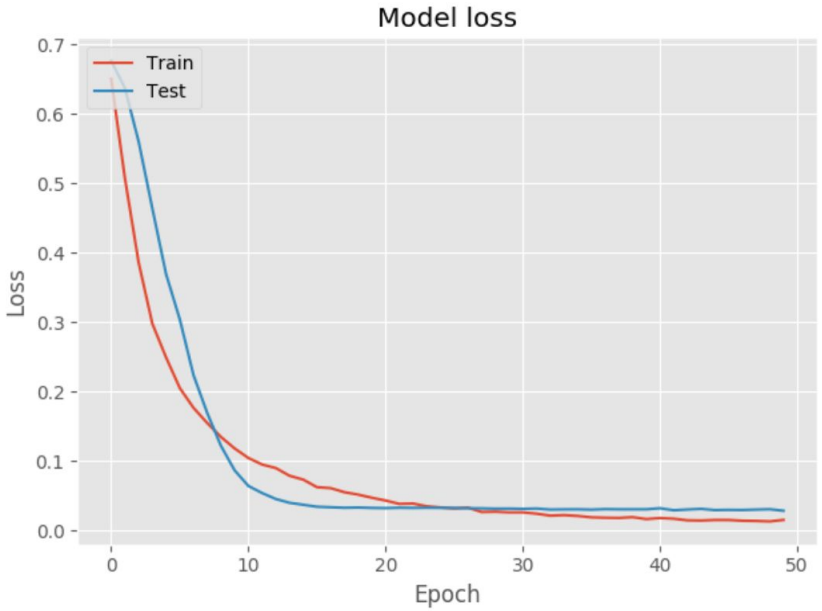
Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 3187, 8)	96
max_pooling1d_1 (MaxPooling1	(None, 1593, 8)	0
batch_normalization_1 (Batch	(None, 1593, 8)	32
conv1d_2 (Conv1D)	(None, 1583, 16)	1424
max_pooling1d_2 (MaxPooling1	(None, 791, 16)	0
batch_normalization_2 (Batch	(None, 791, 16)	64
conv1d_3 (Conv1D)	(None, 781, 32)	5664
max_pooling1d_3 (MaxPooling1	(None, 390, 32)	0
batch_normalization_3 (Batch	(None, 390, 32)	128
conv1d_4 (Conv1D)	(None, 380, 64)	22592
max_pooling1d_4 (MaxPooling1	(None, 190, 64)	0
batch_normalization_4 (Batch	(None, 190, 64)	256
conv1d_5 (Conv1D)	(None, 180, 128)	90240
max_pooling1d_5 (MaxPooling1	(None, 90, 128)	0
flatten_1 (Flatten)	(None, 11520)	0
dropout_1 (Dropout)	(None, 11520)	0
dense_1 (Dense)	(None, 64)	737344
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 64)	4160
dense_3 (Dense)	(None, 1)	65

Total params: 862,065
Trainable params: 861,825
Non-trainable params: 240

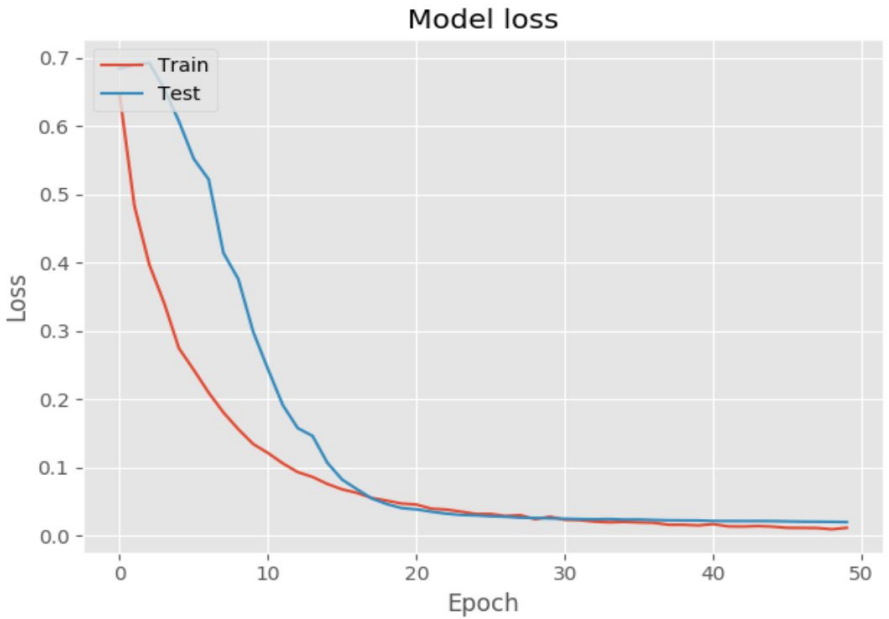
Models and Evaluation -- CNN

Loss Plot

Training all ‘0’ samples



About 70% ‘0’ samples



Models and Evaluation -- CNN

Optimizer: SGD Batch Size: 512 Epochs: 50
Early stopping;
Model Checkpoint

Evaluation:

		precision	recall	f1-score	support
Exoplanet	0	1.00	1.00	1.00	565
Non-exoplanet	1	0.83	1.00	0.91	5
accuracy				1.00	570
macro avg		0.92	1.00	0.95	570
weighted avg		1.00	1.00	1.00	570

Models and Evaluation -- RNN

Model Structure (pseudocode):

INPUT

X = GRU (4 layers)

Y = CNN (4 Conv1d + MaxPool + BN) + Dropout + Relu

ADD = concatenate ([x , y])

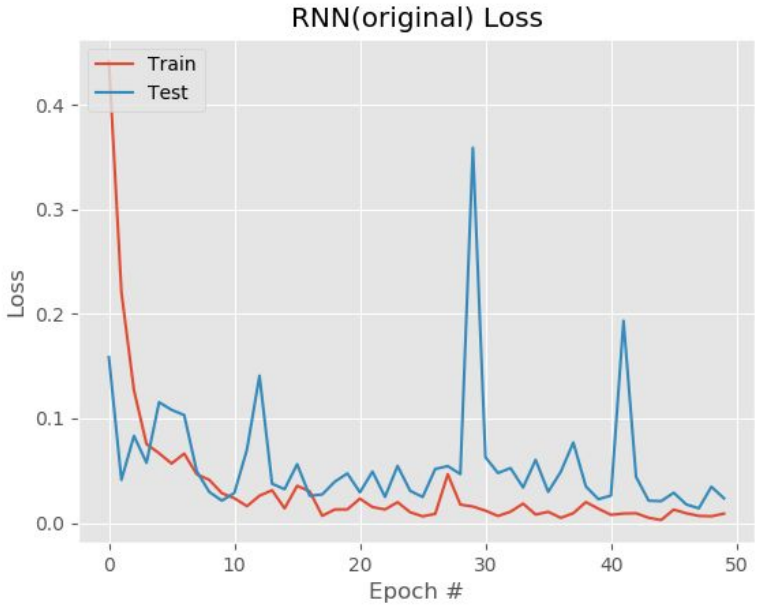
OUTPUT = ADD + Relu + Sigmoid

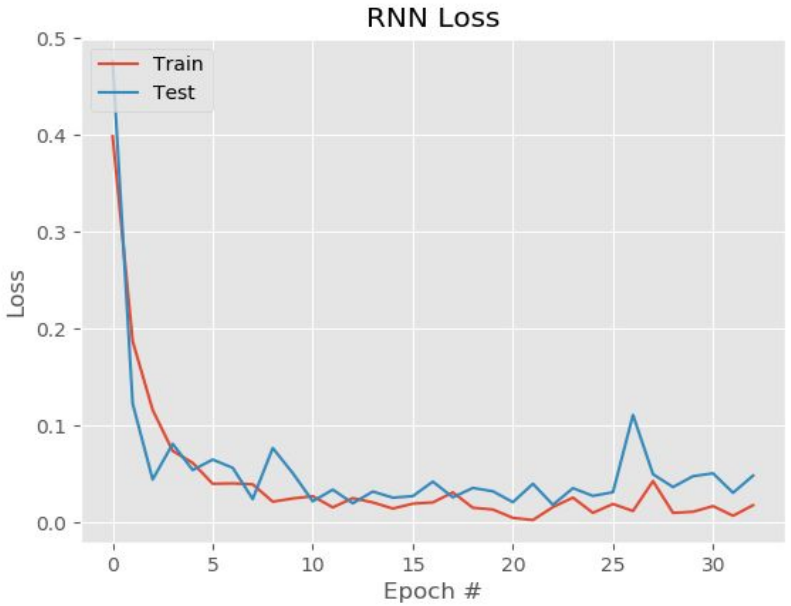
Model (INPUT, OUTPUT)

Models and Evaluation -- RNN

Original model:
High F1 score; test loss is fluctuated; might overfitting

Loss:





Models and Evaluation -- RNN

Batch generator(batch=32, equal number of class 0 and class 1)
Hyperparameter (GRU, SGD, lr= 0.01);
Early stopping;

Evaluation:

		precision	recall	f1-score	support
Exoplanet	0.0	1.00	1.00	1.00	565
Non-exoplanet	1.0	0.67	0.80	0.73	5
accuracy				0.99	570
macro avg		0.83	0.90	0.86	570
weighted avg		1.00	0.99	0.99	570

Summary

- Compare three models:
  - F1 score: CNN performs the best
- Compare with previous works:
  - Contribution:
    - Compared different ways of balancing data;
    - Found smoothing is not suitable;
    - Paid more attention to the loss value and addressed overfitting problem;
  - Limitation/ Need to improve:
    - Which indicator is the best to evaluate this dataset? (F1 ➜ test loss ➜ both?)
    - Use K-fold validation to assess performance of model
    - Classify the features, stratify train set and test set