

Abstract

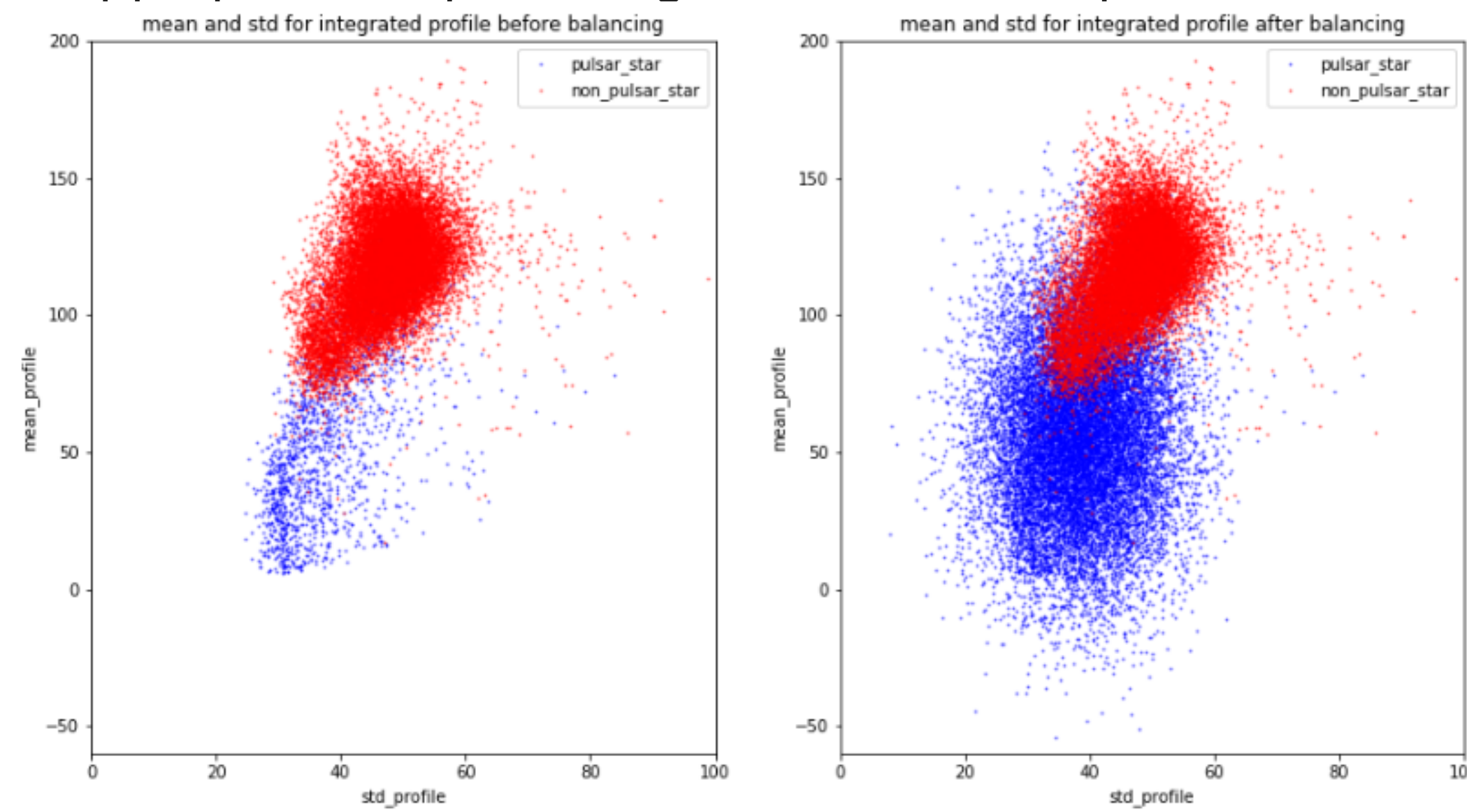
- A pulsar is a highly magnetized rotating neutron star that emits a beam of electromagnetic radiation detectable on Earth
- Astronomers and scientists use pulsars throughout our galaxy as a giant scientific instrument to directly detect gravitational waves and a very precise timepiece to keep time accurately
- Therefore, being able to identify the existence of pulsar stars with high accuracy is intriguing
- We pre-processed the HTRU2 dataset by balancing the data using statistical fake data generation
- We tried various models, including logistic regressor, random forest classifier, and convolutional neural network to predict pulsar stars, and we got high precision and recall from all 3 models

Data

- HTRU2 is a dataset which describes pulsar candidates collected during the High Time Resolution Universe Survey
- Each candidate in the dataset contains 8 attributes: Mean, Standard deviation, Excess kurtosis, and Skewness of the integrated profile, and Mean, Standard deviation, Excess kurtosis, and Skewness of the DM-SNR curve
- Candidates are labeled with the ground truth of whether a pulsar star exists or not

Features

- We have 8 features and they are all directly derived from integrated profile and DM-SNR (signal-to-noise ratio) curve
- They are radio emission-related physics property and thus are appropriate for predicting the existence of pulsar stars

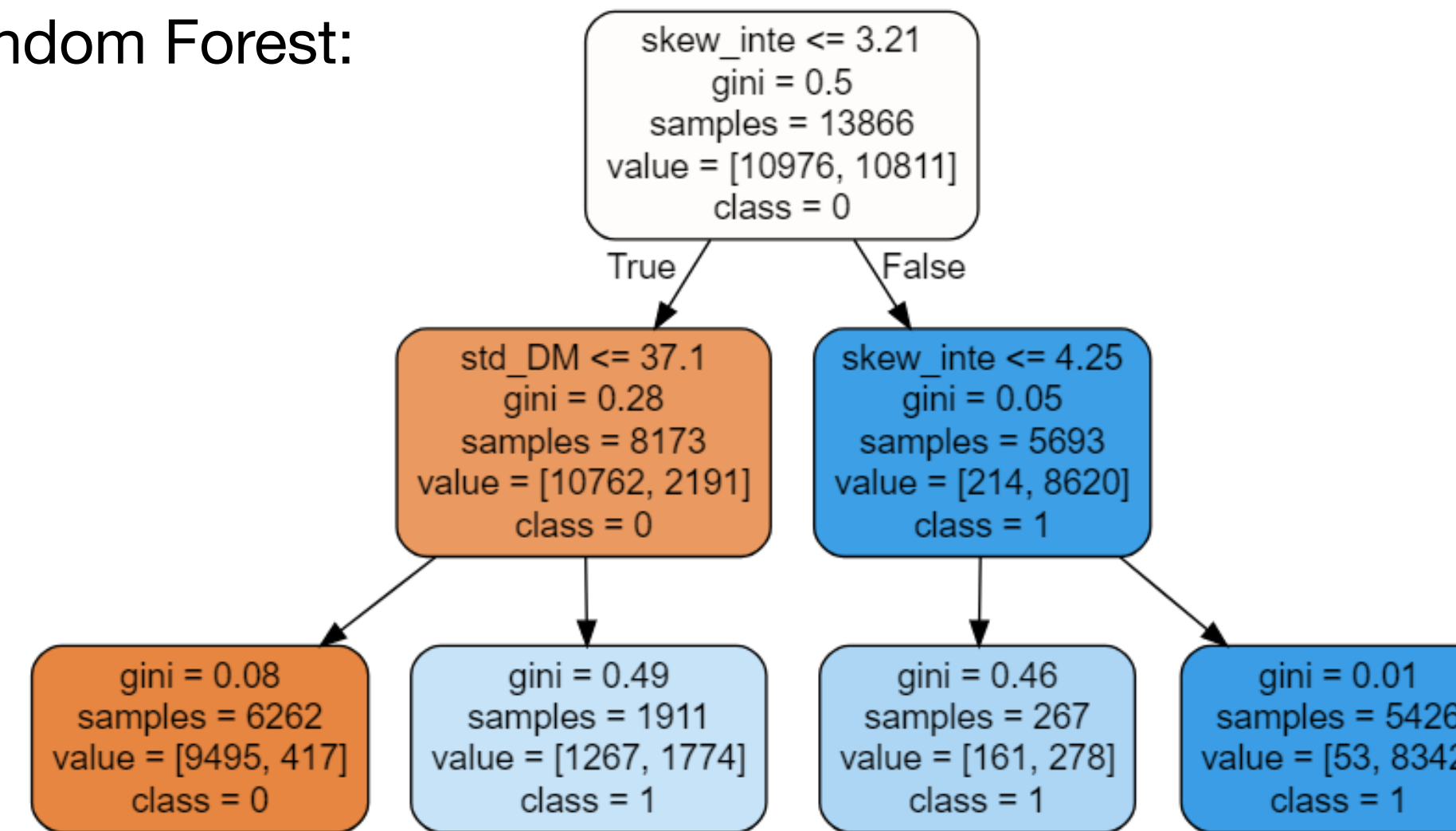


Models

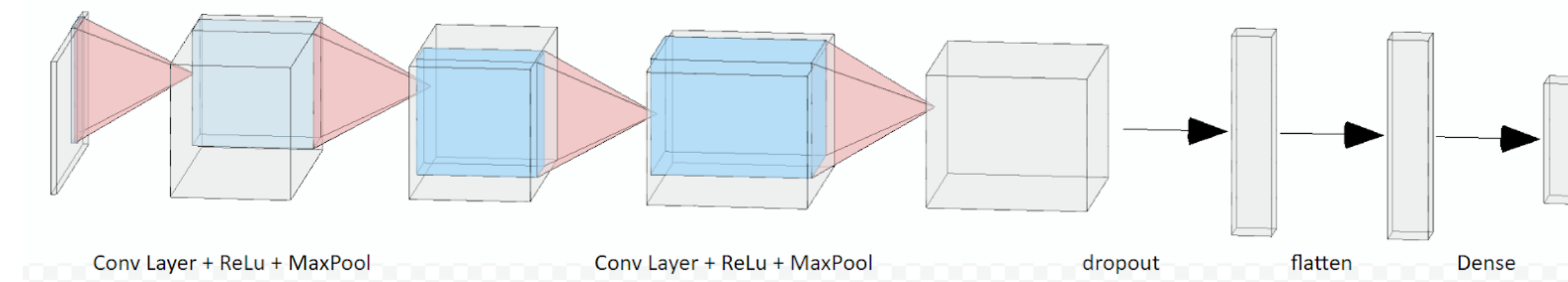
- Logistic Regression: we fit multinomial logistic regression with binary class l2 penalized logistic regression minimizes the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log(\exp(-y_i (X_i^T w + c)) + 1)$$

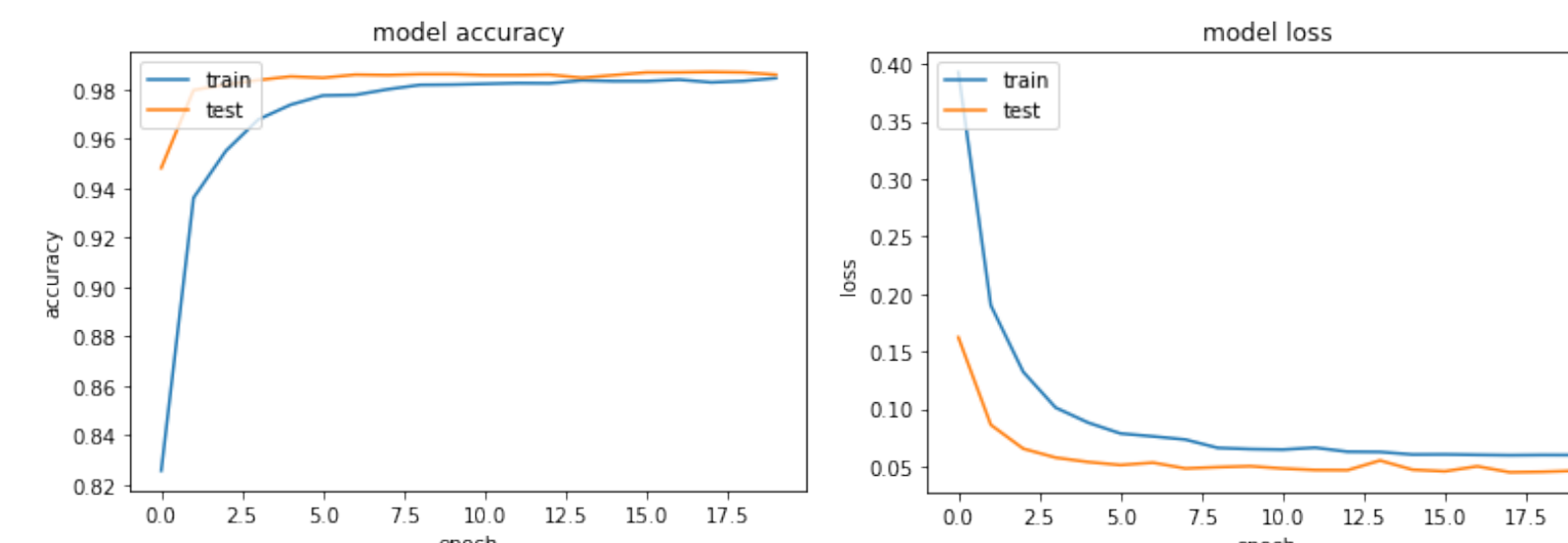
- Random Forest:



- Convolutional Neural Network: Our CNN model uses Stochastic gradient descent (SGD) algorithm to minimize the training loss. In each iteration, the weight matrix is updated according to the following formula: $\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)})$, where the η is the step size and J is the loss function. For J, the model uses binary cross entropy to measure loss over iterations: $-(y \log(p) + (1 - y) \log(1 - p))$, where p is the prediction and y is either 1 or 0.



- The model accuracy and loss for our CNN model is as below:



Results

	Accuracy	Negative Precision	Positive Precision	Negative Recall	Positive Recall
Logistic Regression	0.975585	0.9707667	0.9805367	0.9808657	0.9702693
Random Forest	0.980896	0.9806907	0.9811038	0.9812372	0.9805535
CNN	0.982108	0.9725105	0.9921726	0.9923834	0.9717652

Discussion

- Initially, our models' test accuracies were very high because the data was unbalanced (90% with -1 and 10% with +1 labels). As a result, we have relatively low precision and recall for Pulsars (+1 label)
- After sampling more data with +1 data, our models' performance dropped as expected. After fine tuning the models, we were able to achieve similar accuracy as what we had for the unbalanced data and higher precision and recall
- We found that excess kurtosis is the feature that contributes the most in identifying a Pulsar, but using just kurtosis for integrated profile and DM-SNR curve does not improve the performance
- We expected Logistic regression and random forest to be capable of identifying a Pulsar because these two models have gained popularity for binary classification tasks. However, we did not expect CNN, which is well-known for image classification, to have an impressive performance.

Future Works

- Build a better dataset, include more features that are key representations of Pulsars and is larger and balanced
- Use better balancing methods like mc-stan to generate better distributed fake data

References

- R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656