Detecting Expolanets using Unsupervised Deep Learning

Naveen Mathew Nathan Sathiyanathan[1]
Department of Statistics, University of Illinois at Urbana-Champaign

(Dated: 21 April 2019)

Abstract goes here

I. INTRODUCTION

Give some introduction

II. LITERATURE SURVEY

Talk about paper which uses CNN, which is supervised learning. Also talk about classical methods: BLS, TLS.

III. DATA DESCRIPTION

Describe Kepler data

- Variable: Description.
- Next variable: Next description.

IV. EXPLORATORY ANALYSIS

Speak about number of files, number of stars, number of planets, summary of number of planets around each star. Data is noisy. Talk about median subtraction. Talk about missing data. Add plot

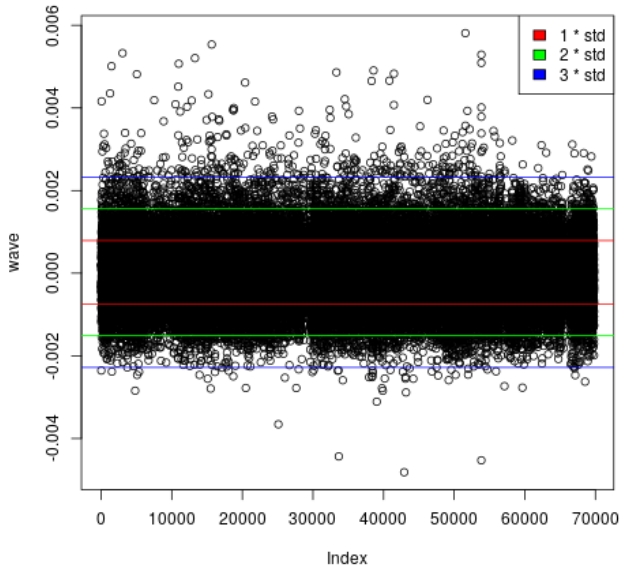

FIG. 1. Residual Flux vs Time.

Summarize few findings

V. PREPROCESSING

Talk about capping, time series interpolation for filling missing values.

A. Capping

Something

B. Imputing Missing Values

Something

VI. MODELING

Why did you choose the model?

A. Hyper-Parameters

Some hyperparameters

VII. RESULTS AND CONCLUSIONS

Show some good and bad examples. Mention how the results can be improved. Remember, this is completely unsupervised! It can be improved drastically with supervised learning.

Dataset	Model	Metric	Value
Test	Logistic Regression	AUC	0.504
Test	Logistic Regression	Accuracy	0.535
Test	Random Forest	AUC	0.797
Test	Random Forest	Accuracy	0.73
Test	Stacked Ensemble	AUC	0.881
Test	Stacked Ensemble	Accuracy	0.779

APPENDIX

- Model building

```
save_plot <- function(y_pred, y, out_file) {
  sqr_error <- (y_pred - y)^2
  avg_error <- mean(sqr_error)
  sd_error <- sd(sqr_error)
  idx <- sqr_error > avg_error + 2 * sd_error
  png(out_file, width = 1366, height = 768)
  plot(y_test, col = idx + 1)
  dev.off()
}

seq_len <- 10
```

- Preprocessing.

```
shp <- c(seq_len - 1, 1)

reduceLr <- callback_reduce_lr_on_plateau(
  monitor = "val_loss", factor = sqrt(0.1),
  patience = 2)
# Early stopping is not working for some reason
# earlyStopping <- EarlyStopping(monitor = "val_acc
    ", min_delta = -0.001,
#                                patience = 2)

files <- list.files(path = "data/", pattern = "*.
    tbl", full.names = T)
system("sh counts.sh")
text <- system("sh counts_df.sh", intern = T)
text <- t(sapply(strsplit(text, "\t"), function(
    elem) c(elem[1], elem[2])))
text <- data.frame(text)
text$X2 <- as.numeric(text$X2)
out_files <- sapply(strsplit(files, "/"), function(
    strs) strs[length(strs)])
out_files <- sapply(strsplit(out_files, "\\."),
    function(strs) strs[1])
dir.create("plots", showWarnings = F)
setwd("plots")
dir.create("learning_curve", showWarnings = F)
dir.create("test_pred_plot", showWarnings = F)
```

- The Data Source
  https://www.kaggle.com/mczielinski/bitcoin-historical-data
- References
  https://www.sciencedirect.com/science/article/pii/B9780128021170000011?via%3Dihub

  Advances in Financial Machine Learning by Marcos Lopez de Prado
- Text Sources:
  - CNBC
  - BBC
  - The Verge
  - Guardian
  - Bloomberg
  - Coindesk