BU Department of Computer Science

Using News to Predict Stock Movements

Overview

Goal: predict, with an assigned confidence value [-1,1], the 10-day returns of stocks given news and sentiment features as input (35 features).

Evaluation Metric: argmax(Sharpe ratio)

- Sharpe = Asset Return/Asset volatility.
- **Baseline**: 0.60 (current Kaggle average).

Dataset: two sets of data provided per day

- one contains daily market return data from 2009 to 2017
- second dataset contains information about news articles published about assets such as sentiment, word count etc.

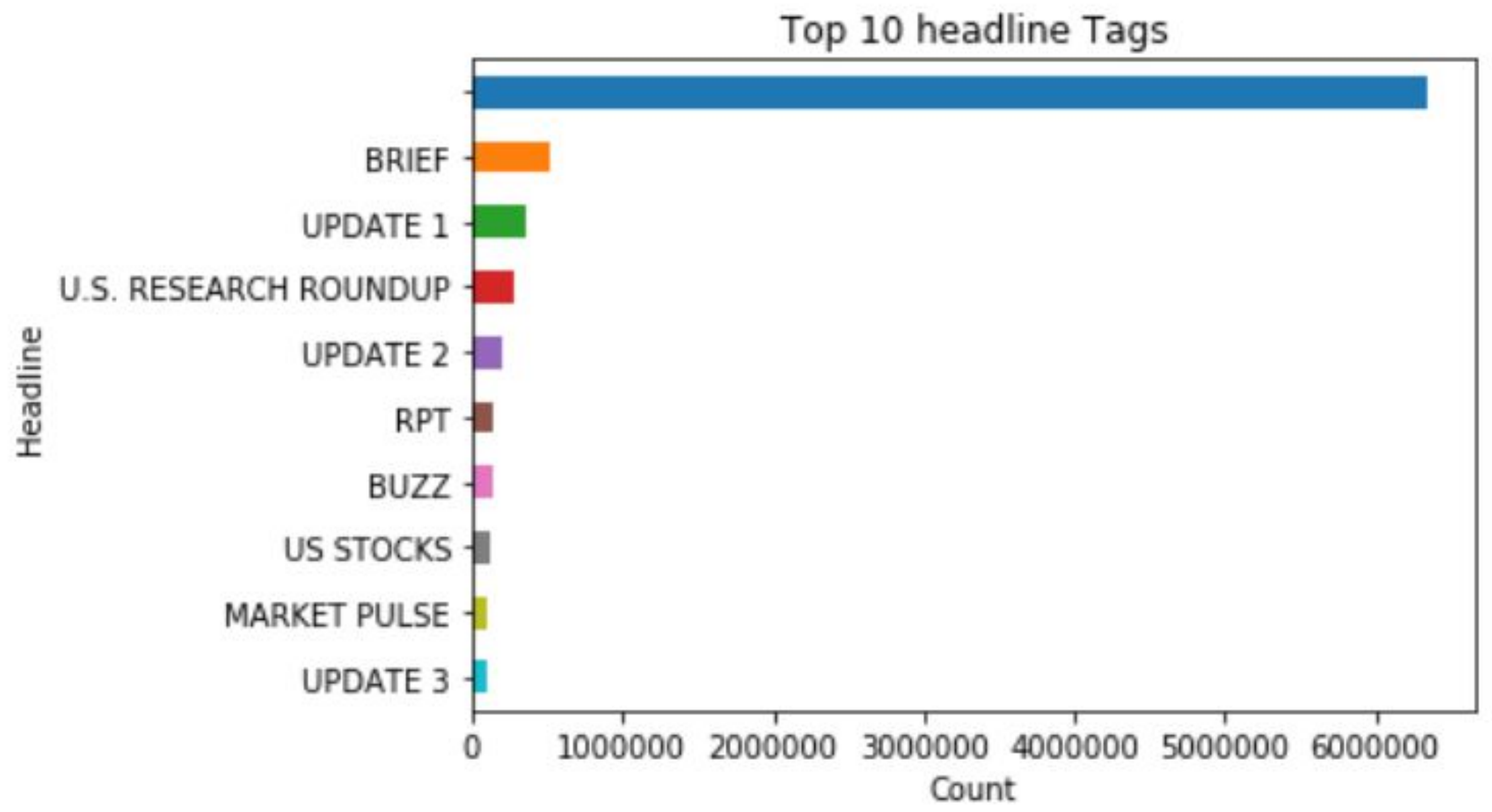
Hardware Constraints: 6 hour run time, 17gb of RAM, no GPU

Data Pre-Processing

Feature Overview: we looked at different features to see if there was missing data (see graph below)

Implemented: Anomaly clipping, label encoding, feature aggregation and description, NLP{Tf-idf}, feature sampling.

Attempted: hashingVectorizer, Bag_of_words

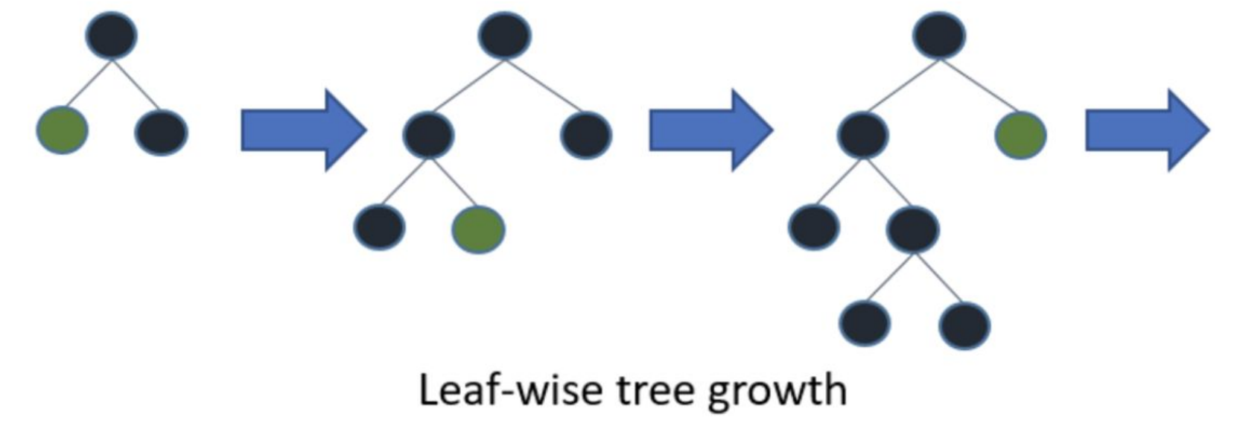


Model

Baseline: 0.60 (Kaggle average)

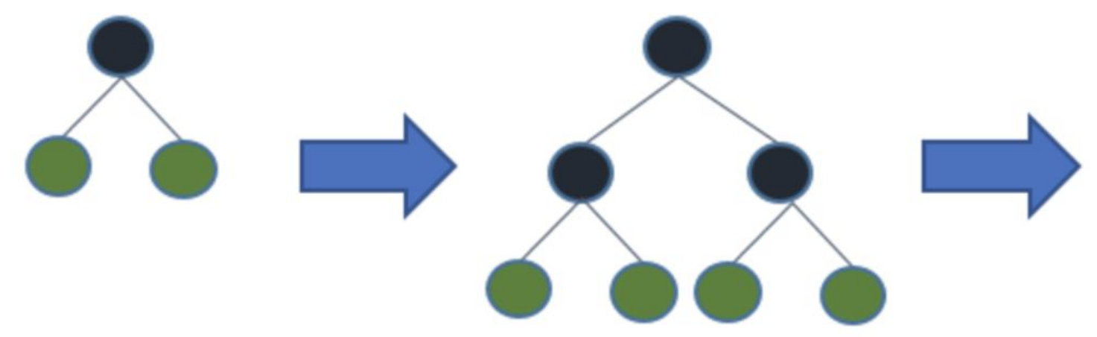
Types of Models:

- LSTM - can learn the order dependence between items in a sequence
  -slower than other models
  -infeasible due to hardware limits
- LBGM - builds model using an ensemble of weak learners
  - fast, light weight
  - uses leaf-wise growth
  - can tune parameters such as depth and number of leafs



Leaf-wise tree growth

Explains how LightGBM works



Level-wise tree growth

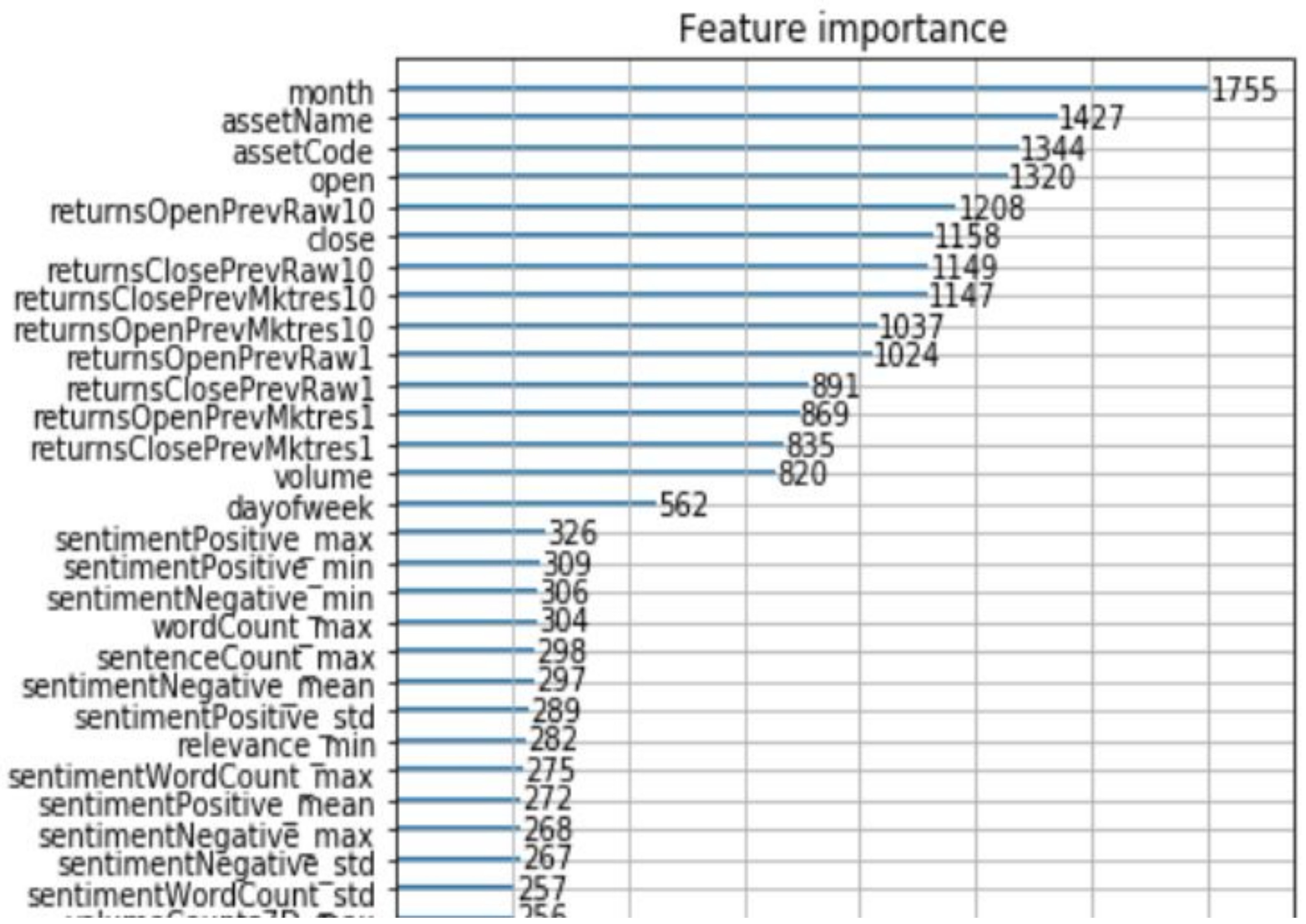
Custom scoring metric:

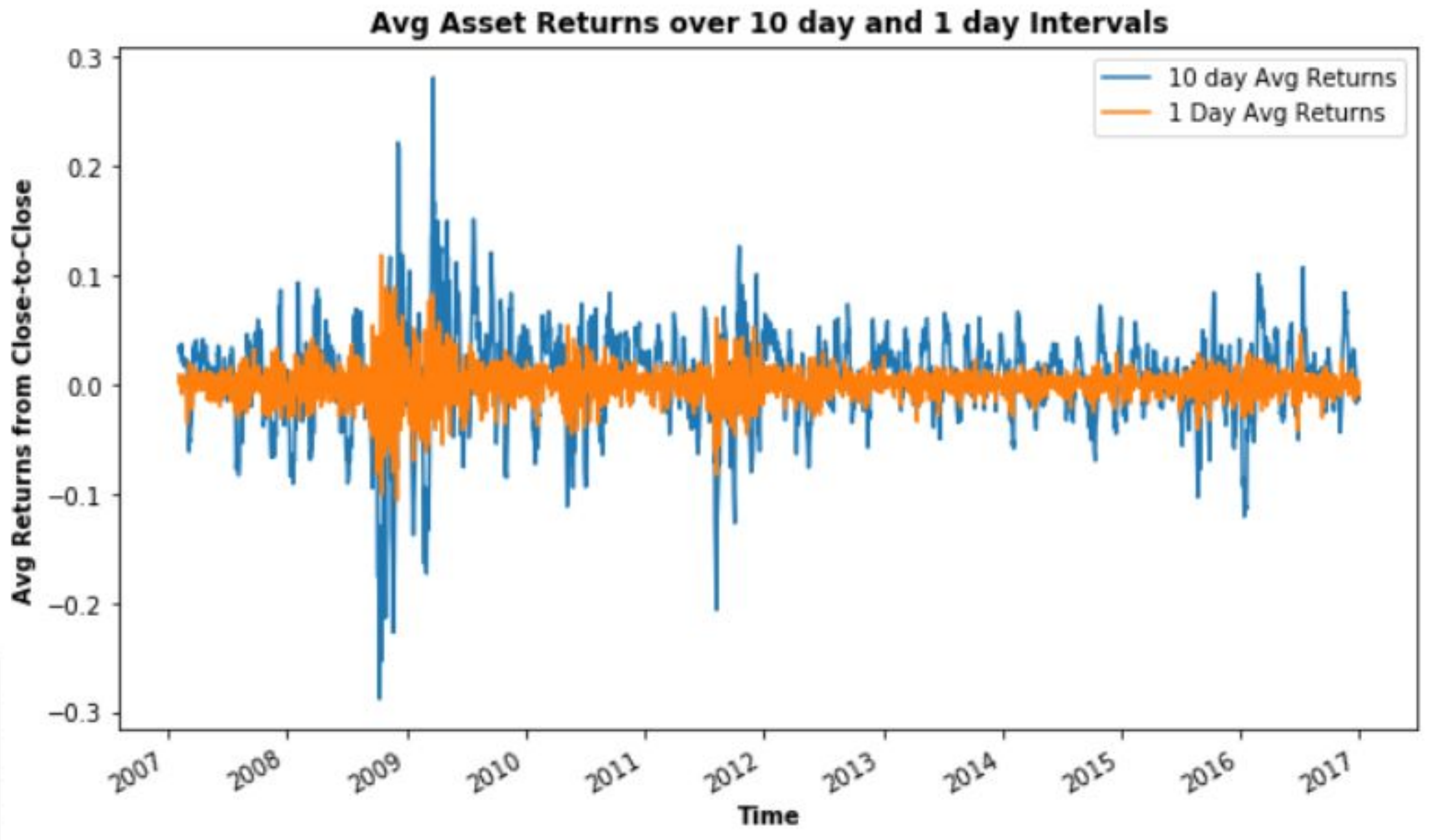
- evaluated model using a metric that divides the mean of our daily predictions by its standard deviation

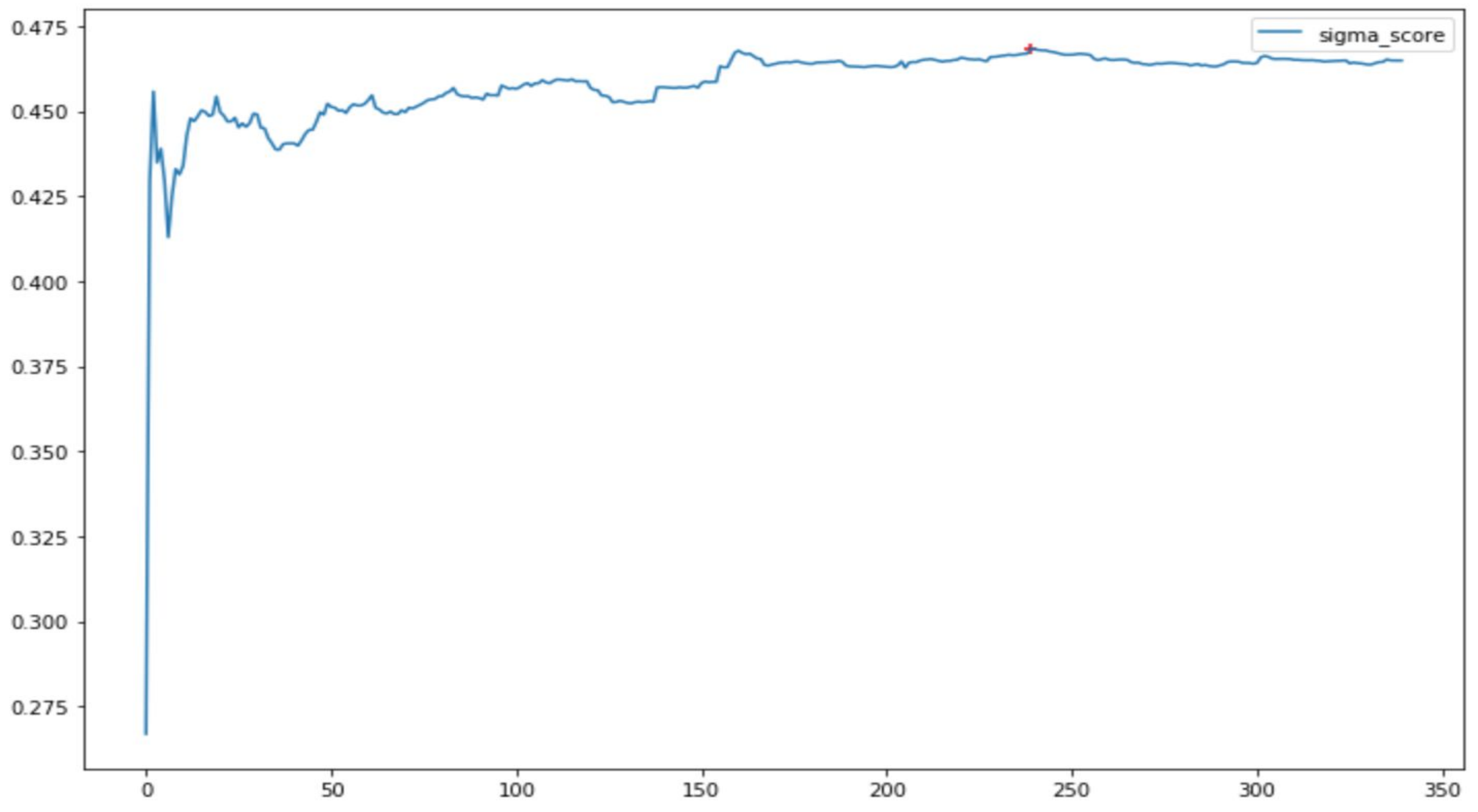
Feature Analysis & Preliminary Results

Feature Selection: plotted feature importance to help us fine tune our features

- found outliers in asset returns during the financial crisis and excluded them from our training data





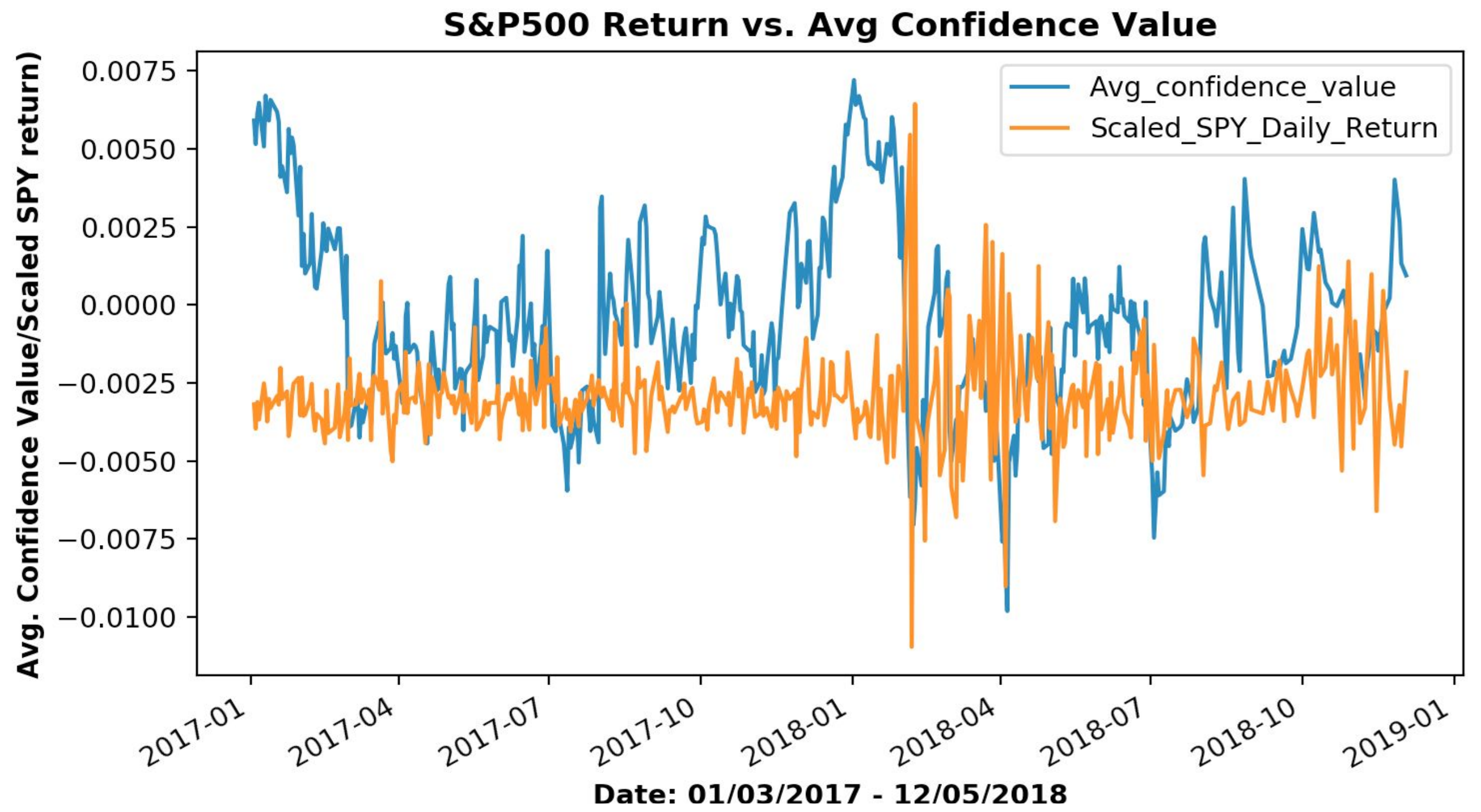


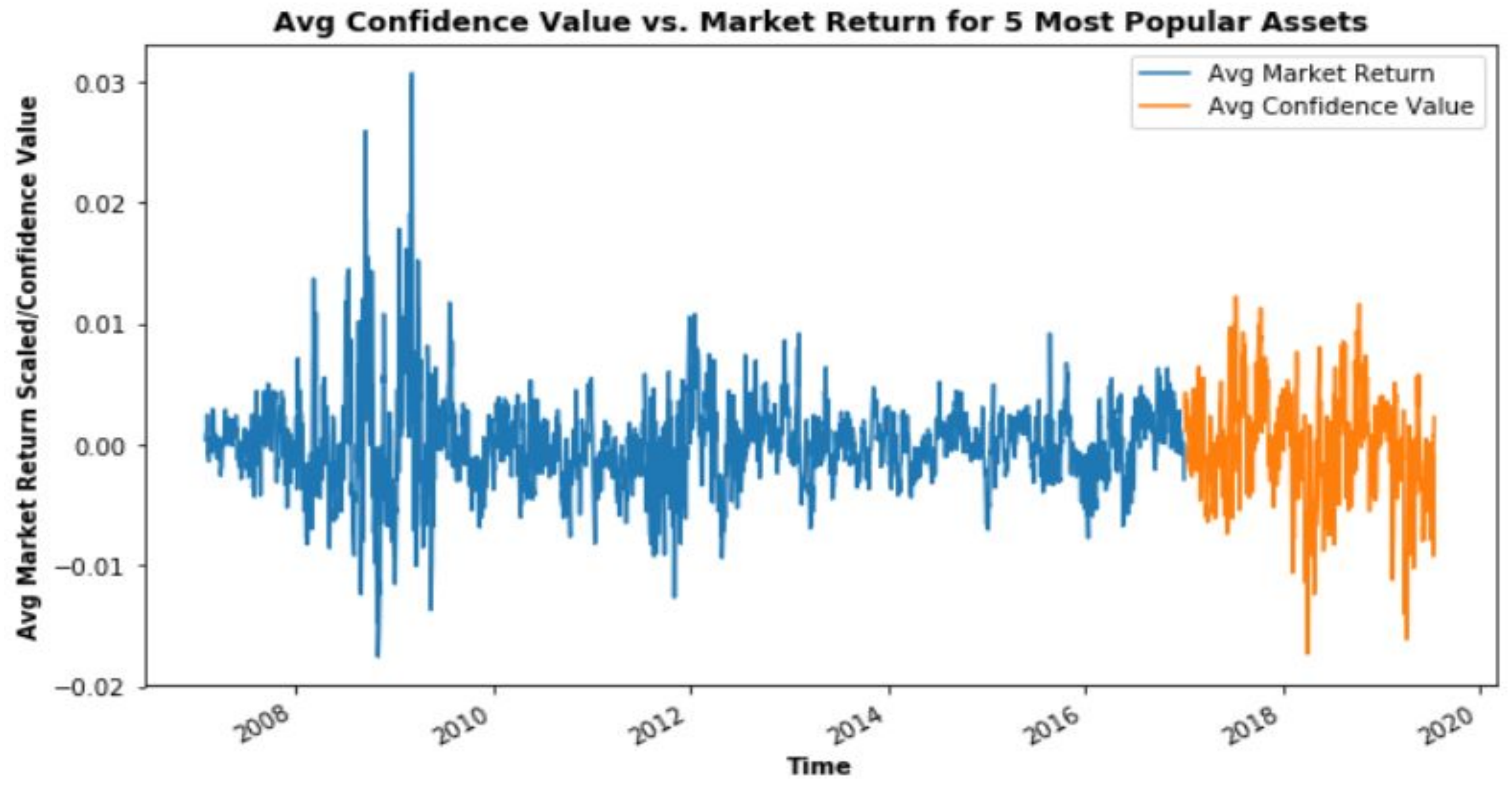
Results & Analysis

Final performance metric: 0.619

Kaggle leaderboard (12/8/18): 1,196/2,165

Graphical Analysis: the graphs below compare historical market returns to our predicted confidence values





Group 11: Nicole Mis, Jon Hale, Jorge Nario, Duruvan Saravanan

bu.edu/cs @BUCompSci