A Python module using Jupyter Notebooks to take an existing dataset available at Kaggle and undertake some data cleansing, data hard coding and data science management so it can be more useful for Machine Learning models.

Source of original dataset:

<https://www.kaggle.com/datasets/ifuurh/nasdaq100-fundamental-data>

## Introduction

The problem we are trying to solve is that there are very limited datasets on Kaggle if you wish to apply ML models to the problem of individual stock Share Price prediction using financial statement ratios as your input data.

This is a problem that needs addressing as there is a multi-billion global fundamental financial ratio investment analysis industry that is ripe for performance enhancement by Machine Learning.

We believe that the best dataset for such a purpose on Kaggle was the above dataset that we found above.

The problem with this dataset for ML model use was as follows:

* There was a number of data attributes that were not shown across each annual period. We removed data attributes that were not populated across all the annual periods.
* We filled in data that was missing and we replaced NANs and INFs with logical and reasonable fill values.
* We attached label data being 12 month ahead Share Price returns for each stock and each annual period providing this data both as discrete percentage returns and binary outperform or underperform the Nasdaq 100 index labels.

## Resulting Datasets

The resulting datasets cover 102 stocks using 39 financial ratios across both 4 and 5 year periods using two different types of labels.

In summary, this repository provides a Jupyter Notebook that shows the steps undertaken to generate:

Two datasets for 2017 to 2021 with the Y labels attached at the end column.

* labels 1 or 0: for binary outperformance against index.
* perfs labels: for actual performance for the stock for that calendar year.

And Two mote datasets for 2017 to 2020 with the same Y label data as above:

* labels 1 or 0: for binary outperformance against index.
* perfs labels: for actual performance for the stock for that calendar year.

## Usage & Contributing

At the moment the project is in development.

You can use the repository and play with the Jupyter Notebook to generate your own datasets with differing assumptions to ours.

We will then load up some ML models that we think can be the most effective at predicting 12 month forward Share Price outcomes based on the 39 financial ratios provided.

We would welcome your thoughts on our models. Even better we would welcome YOUR ideas on the best models to use to solve such a prediction problem using these datasets?

You can always help to get this problem solved. It's an open-source project after all!

## Resources

* Kaggle: <https://www.kaggle.com/datasets/ifuurh/nasdaq100-fundamental-data>
* Jupyter Notebooks: <https://jupyter.org/>
* Yfinance: <https://pypi.org/project/yfinance/>