Machine Learning

Weekly Project Report

Quantcats

Prachee Javiya AU1841032
Kaushal Patil AU1841040
Arpitsinh Vaghela AU1841034
Vrunda Gadesha AU2049007

Tasks Performed: Week 3

01 Data labelling

02 Performing other advanced exploratory data set analysis methods

03 Testing the created dataset on fairly simple multilabel classification

- Logistic Regression
- Forests(Decision Trees),

1. Creating the final Dataset

The two datasets were merged to create the final dataset containing fundamentals + prices and labels after undergoing necessary manipulations

In [113]:
```
fundamen_data.head()
```

Out[113]:

	Ticker	SimFinId	Currency	Fiscal Year	Fiscal Period	Report Date	Publish Date	Restated Date	Shares (Basic)	Shares (Diluted)	...	Short Term Debt	Total Current Liabilities	Long Term Debt
0	A	45846	USD	2010	Q3	2010-07-31	2010-10-06	2010-10-06	347000000.0	352000000.0	...	1.501000e+09	2.917000e+09	2.177000e+09
1	A	45846	USD	2010	Q4	2010-10-31	2010-12-20	2011-12-16	344000000.0	356000000.0	...	1.501000e+09	3.083000e+09	2.190000e+09
2	A	45846	USD	2011	Q1	2011-01-31	2011-03-09	2011-03-09	347000000.0	355000000.0	...	1.000000e+06	1.406000e+09	2.138000e+09
3	A	45846	USD	2011	Q2	2011-04-30	2011-06-07	2011-06-07	347000000.0	355000000.0	...	0.000000e+00	1.592000e+09	2.144000e+09
4	A	45846	USD	2011	Q3	2011-07-31	2011-09-07	2011-09-07	348000000.0	357000000.0	...	0.000000e+00	1.505000e+09	2.168000e+09

5 rows × 30 columns

In [114]:
```
price_df.head()
```

Out[114]:

	Unnamed: 0	Ticker	Price_Present	Price_Future	Label
0	0	A	32.67	41.48	0
1	1	A	40.63	43.95	0
2	2	A	46.05	48.71	0
3	3	A	47.7	35.69	1
4	4	A	35.69	36.66	2

Tasks Performed: Week 3

Final dataset

In [129]:
```
clean_fundamen=pd.read_csv("dataset_1.csv")
clean_fundamen.describe()
```

Out[129]:

	...	Total Noncurrent Liabilities	Total Liabilities	Share Capital & Additional Paid-In Capital	Treasury Stock	Retained Earnings	Total Equity	Total Liabilities & Equity	Price_Present	Price_Future	Label
4	...	5.882800e+04	5.882800e+04	5.882800e+04	5.882800e+04	5.882800e+04	5.882800e+04	5.882800e+04	5.882800e+04	5.882800e+04	58828.000000
9	...	4.285066e+09	6.835340e+09	2.496729e+09	-1.316608e+09	2.728238e+09	3.844743e+09	1.068008e+10	2.996048e+04	2.805772e+04	0.729636
0	...	1.505207e+10	2.354231e+10	7.630913e+09	8.343771e+09	1.538363e+10	1.338728e+10	3.472629e+10	1.684409e+06	1.628787e+06	0.760094
)9	...	-2.397250e+09	-2.141441e+09	-9.446000e+08	-2.304540e+11	-9.532600e+10	-2.564100e+10	0.000000e+00	1.000000e-02	1.000000e-02	0.000000
7	...	3.647150e+07	1.337462e+08	1.448175e+08	-1.619970e+08	-6.704085e+07	1.365072e+08	3.680082e+08	1.451000e+01	1.450000e+01	0.000000
8	...	5.119960e+08	9.829910e+08	5.026950e+08	0.000000e+00	1.317635e+08	6.720960e+08	1.825716e+09	3.158000e+01	3.204000e+01	1.000000
9	...	2.693624e+09	4.292500e+09	1.768310e+09	0.000000e+00	1.355143e+09	2.517094e+09	7.101819e+09	5.924000e+01	6.063000e+01	1.000000
1	...	3.696910e+11	6.620370e+11	1.574890e+11	3.838900e+07	4.216530e+11	4.015580e+11	7.884820e+11	1.000000e+08	1.000000e+08	2.000000

Tasks Performed: Week 3

2. Data Cleaning and Exploration

```
In [122]: clean_fundamen.count()

Out[122]: Ticker                                              63000
          SimFinId                                            63000
          Currency                                            63000
          Fiscal Year                                         63000
          Fiscal Period                                       63000
          Report Date                                         63000
          Publish Date                                        63000
          Restated Date                                       63000
          Shares (Basic)                                      62367
          Shares (Diluted)                                    62367
          Cash, Cash Equivalents & Short Term Investments     62794
          Accounts & Notes Receivable                         57619
          Inventories                                         43426
          Total Current Assets                                62980
          Property, Plant & Equipment, Net                    61782
          Long Term Investments & Receivables                 17907
          Other Long Term Assets                              62238
          Total Noncurrent Assets                             62728
          Total Assets                                        63000
          Payables & Accruals                                 62714
          Short Term Debt                                     41580
          Total Current Liabilities                           62996
          Long Term Debt                                      49689
          Total Noncurrent Liabilities                        62257
          Total Liabilities                                   63000
          Share Capital & Additional Paid-In Capital          61967
          Treasury Stock                                      29730
          Retained Earnings                                   60111
          Total Equity                                        62999
          Total Liabilities & Equity                          63000
          Price_Present                                       58835
          Price_Future                                        58829
          Label                                               63000
          dtype: int64
```

Tasks Performed: Week 3

```
In [127]: clean_fundamen.count()
```

```
Out[127]: Ticker                                                58828
          SimFinId                                              58828
          Currency                                              58828
          Fiscal Year                                           58828
          Fiscal Period                                         58828
          Report Date                                           58828
          Publish Date                                          58828
          Restated Date                                         58828
          Shares (Basic)                                        58828
          Shares (Diluted)                                      58828
          Cash, Cash Equivalents & Short Term Investments       58828
          Accounts & Notes Receivable                           58828
          Inventories                                           58828
          Total Current Assets                                  58828
          Property, Plant & Equipment, Net                      58828
          Long Term Investments & Receivables                   58828
          Other Long Term Assets                                58828
          Total Noncurrent Assets                               58828
          Total Assets                                          58828
          Payables & Accruals                                   58828
          Short Term Debt                                       58828
          Total Current Liabilities                             58828
          Long Term Debt                                        58828
          Total Noncurrent Liabilities                          58828
          Total Liabilities                                     58828
          Share Capital & Additional Paid-In Capital            58828
          Treasury Stock                                        58828
          Retained Earnings                                     58828
          Total Equity                                          58828
          Total Liabilities & Equity                            58828
          Price_Present                                         58828
          Price_Future                                          58828
          Label                                                 58828
          dtype: int64
```

Tasks Performed: Week 3

Normalized dataset

In [137]:
```
clean_fundamen.head()
```

Out[137]:

ublish ate	Restated Date	Shares (Basic)	Shares (Diluted)	...	Total Current Liabilities	Long Term Debt	Total Noncurrent Liabilities	Total Liabilities	Share Capital & Additional Paid-In Capital	Treasury Stock	Retained Earnings	Total Equity	Total Liabilities & Equity	Price_Present
)10-10-06	2010-10-06	0.013189	0.013262	...	0.009165	0.006773	0.015508	0.012695	0.055579	0.965186	0.190487	0.066599	0.011541	3.266000e-07
)10-12-20	2011-12-16	0.013075	0.013413	...	0.009686	0.006809	0.015518	0.012950	0.055888	0.964960	0.191052	0.067596	0.012297	4.062000e-07
)11-03-09	2011-03-09	0.013189	0.013376	...	0.004417	0.006664	0.015309	0.010308	0.056873	0.963789	0.191426	0.067837	0.010202	4.604000e-07
)11-06-07	2011-06-07	0.013189	0.013376	...	0.005002	0.006681	0.014763	0.010283	0.057378	0.963789	0.191812	0.069293	0.010969	4.769000e-07
)11-09-07	2011-09-07	0.013227	0.013451	...	0.004728	0.006748	0.014634	0.010079	0.058060	0.962956	0.192451	0.069853	0.011101	3.568000e-07

In [139]:
```
clean_fundamen.describe()
```

Out[139]:

	SimFinId	Fiscal Year	Shares (Basic)	Shares (Diluted)	Cash, Cash Equivalents & Short Term Investments	Accounts & Notes Receivable	Inventories	Total Current Assets	Property, Plant & Equipment, Net	Long Term Investments & Receivables	...
count	5.882800e+04	58828.000000	58828.000000	58828.000000	58828.000000	58828.000000	58828.000000	58828.000000	58828.000000	58828.000000	...
mean	4.561002e+05	2014.762477	0.009883	0.009982	0.008535	0.002388	0.007961	0.006935	0.026874	0.002203	...
std	3.094505e+05	2.876963	0.041161	0.041600	0.043004	0.017644	0.030639	0.029270	0.045300	0.023822	...
min	1.800000e+01	1999.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	1.792010e+05	2013.000000	0.001035	0.001034	0.000234	0.000043	0.000000	0.000313	0.014806	0.000000	...
50%	3.781880e+05	2015.000000	0.002430	0.002443	0.000995	0.000315	0.000564	0.001136	0.015641	0.000000	...
75%	7.055880e+05	2017.000000	0.006842	0.006869	0.003721	0.001384	0.004888	0.004118	0.019949	0.000016	...
max	1.085359e+06	2020.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...

8 rows × 25 columns

Tasks Performed: Week 3

Example: Exploring the labels

```
In [132]: LabnF.head()
```

Out[132]:

	Label	Price_Future
0	0	41.48
1	0	43.95
2	0	48.71
3	1	35.69
4	2	36.66

```
In [133]: LabnF.describe()
```

Out[133]:

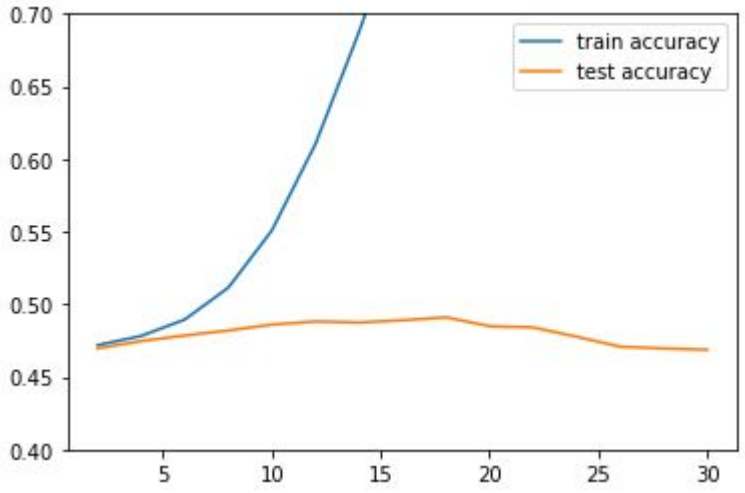
	Label	Price_Future
count	58828.000000	5.882800e+04
mean	0.729636	2.805772e+04
std	0.760094	1.628787e+06
min	0.000000	1.000000e-02
25%	0.000000	1.450000e+01
50%	1.000000	3.204000e+01
75%	1.000000	6.063000e+01
max	2.000000	1.000000e+08

Testing out Various Models

Logistic Regression

1. Normal data: score **0.454086489426**
2. After normalizing the data: score **0.461208948119**
3. After standardizing the data: score **0.4661895695927**

Random Forest [Best non-overfitting model max_depth=18]
Best score **0.490000089426 [Test]**
Best score **0.8512532620743 [Train]**



Depth vs accuracy

Outcomes

Labelling the Dataset

The final dataset is ready and normalized and standardized versions for the same were created and used for model training.

Model Training

- *Logistic regression* gives almost 45%~46% accuracy on normal, normalized and standardized datasets.

- *Random forest*, a depth of around 5 gives a good generalization , the later depths leads to overfitting the model with higher train accuracy but low test accuracy.

Upcoming Week

01 Performing PCA and plotting random points in 2 and 3 dimensions to observe how different classes cluster.

02 Tackling the problem as binary classification for each class individually and finding insights from it.

03 Finding correlation of each feature individually with the labels.

04 Trying out other models and methods of dataset analysis.