Machine Learning

Weekly Project Report

Quantcats

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

Tasks Performed: Week 4

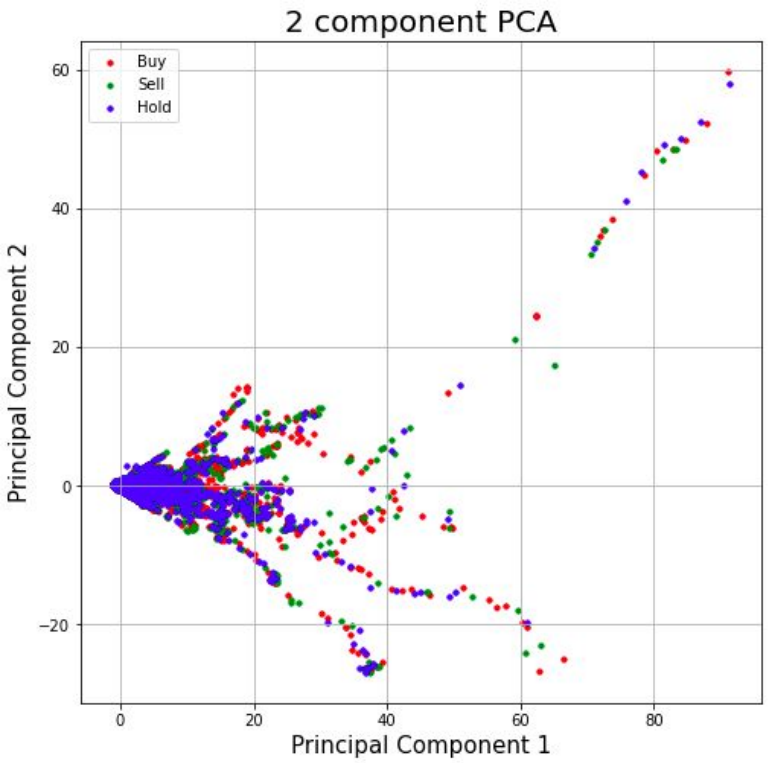
01 PCA analysis
02 K means clustering
03 Binary classification

1. Principal component analysis

Data was trained on the numeric columns of labelled dataset

	principal component 1	principal component 2	target_numeric	target
0	0.210449	0.043768	0	Buy
1	0.267291	0.050526	0	Buy
2	0.071494	-0.233925	0	Buy
3	0.126114	-0.227153	1	Sell
4	0.146986	-0.243517	2	Hold
...	...	...	...	...
58823	-1.039740	0.078976	1	Sell
58824	-1.039956	0.079430	0	Buy
58825	-1.037820	0.079236	0	Buy
58826	-1.034459	0.079764	0	Buy
58827	-1.028980	0.081075	0	Buy

58828 rows × 4 columns



K means clustering

Final dataset

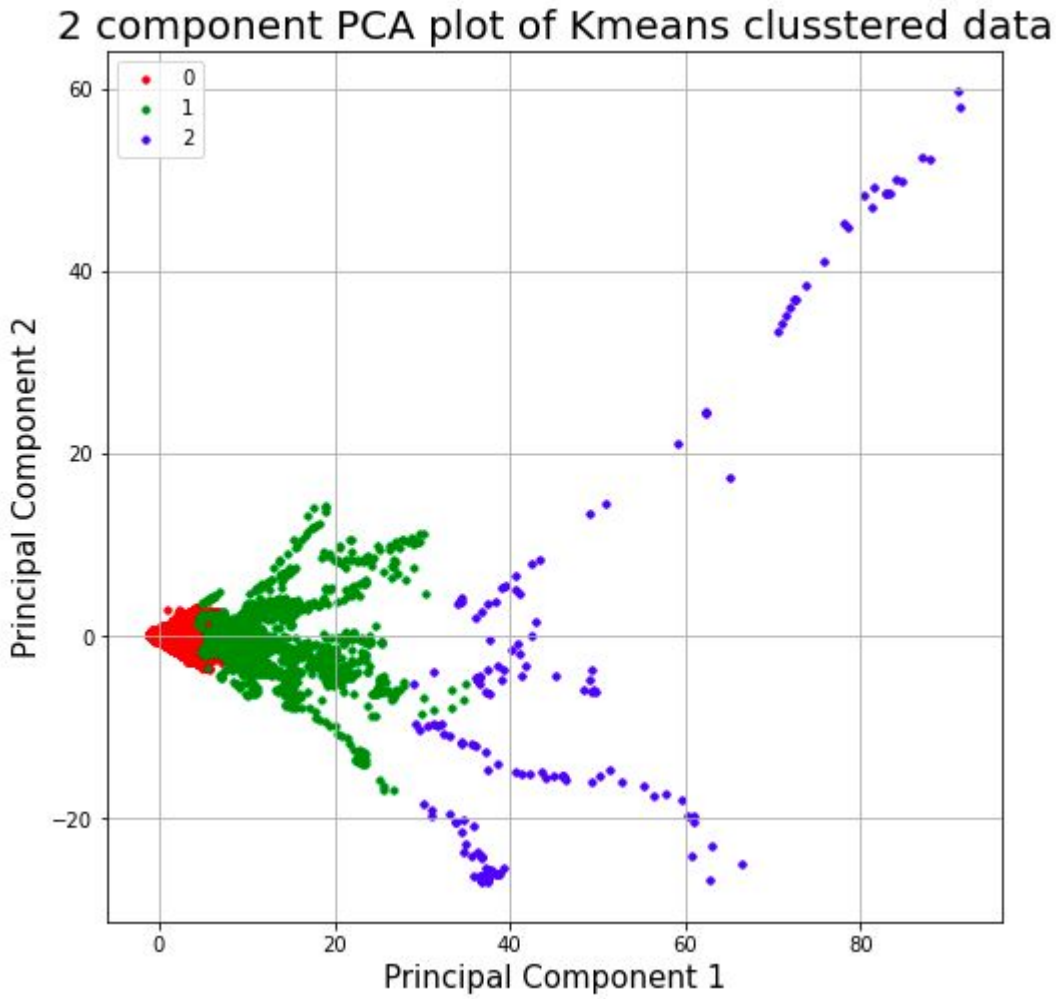
```
[72]: kmeans = KMeans(n_clusters=3, random_state=0).fit(X)

[73]: pcadf["kmean_target_numeric"]=kmeans.labels_

[74]: pcadf
```

[74]:

	principal component 1	principal component 2	target_numeric	target	kmean_target_numeric
0	0.210449	0.043768	0	Buy	0
1	0.267291	0.050526	0	Buy	0
2	0.071494	-0.233925	0	Buy	0
3	0.126114	-0.227153	1	Sell	0
4	0.146986	-0.243517	2	Hold	0
...	...	...	...	...	...
58823	-1.039740	0.078976	1	Sell	0
58824	-1.039956	0.079430	0	Buy	0
58825	-1.037820	0.079236	0	Buy	0
58826	-1.034459	0.079764	0	Buy	0
58827	-1.028980	0.081075	0	Buy	0

58828 rows × 5 columns

Output - K means clustering




```
[77]: correlation = pcadf["target_numeric"]. corr(pcadf["kmean_target_numeric"])

[78]: correlation

[78]: 0.016179456580554825
```

Binary Classification

Classified labelled dataset into three classes.

[6]:

	Label	Price_Future	class_0	class_1	class_2
0	0	41.48	1	0	0
1	0	43.95	1	0	0
2	0	48.71	1	0	0
3	1	35.69	0	1	0
4	2	36.66	0	0	1
...	...	...	...	...	...
58823	1	21.06	0	1	0
58824	0	25.38	1	0	0
58825	0	31.68	1	0	0
58826	0	38.29	1	0	0
58827	0	41.49	1	0	0

58828 rows × 5 columns

[71]:
```
model = MultiBinaryModel()
model.fit(X_train,y_train)
model.score(X_train,y_train)
```

[71]:
```
[0.5395899911606717, 0.6508295369551914, 0.8098014550894133]
```

Outcomes

Model Training

- PCA and Kmeans clustering

  Clustering does not give us an optimal output. Hence this model is dropped

- Binary classification on individual labels work better. Scores for each label - 0.53,0.65,0.80

Upcoming Week

01 K nearest neighbours
02 Softmax regression
03 Stacking multi binary classifier
04 Trying Multiple Model Stacking (Random Forest + Multinomial Softmax Regression)
05 Trying Gradient Boosting Methods, XGBoost Classification.
06 Trying out probabilistic models.