Machine Learning

Weekly Project Report

Quantcats

Prachee Javiya AU1841032

Kaushal Patil AU1841040

Arpitsinh Vaghela AU1841034

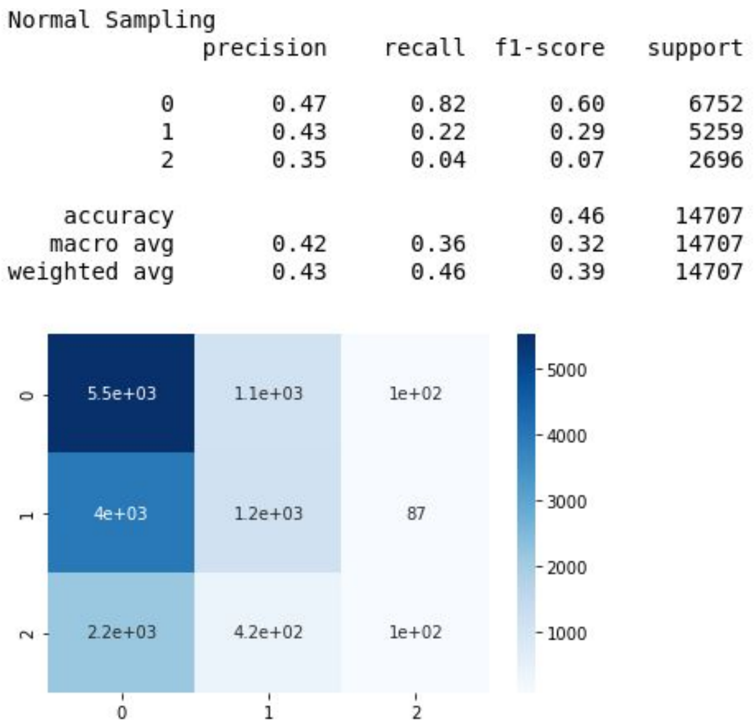
Vrunda Gadesha AU2049007

Tasks Performed: Week 7

- Attempt to improve performance using resampling i.e, oversampling and undersampling of the data.
- On training the Random Forest Classifier without resampling, the f1-score showed that the model was biased and performed badly on class_1 and class_2. Furthermore it is known that the training sample of these two classes are lesser than that of class_0.
- Two attempts were made, one using **oversampling class_1 and class_2** and one using **undersampling on class_0** and **oversampling on class_2**.
- We take a decision tree with three classes buy sell and hold and try to kill nodes that might be overfitting/underfitting by assigning a 4th label called IDK to the classification task. This way we can make the tree more dynamic.

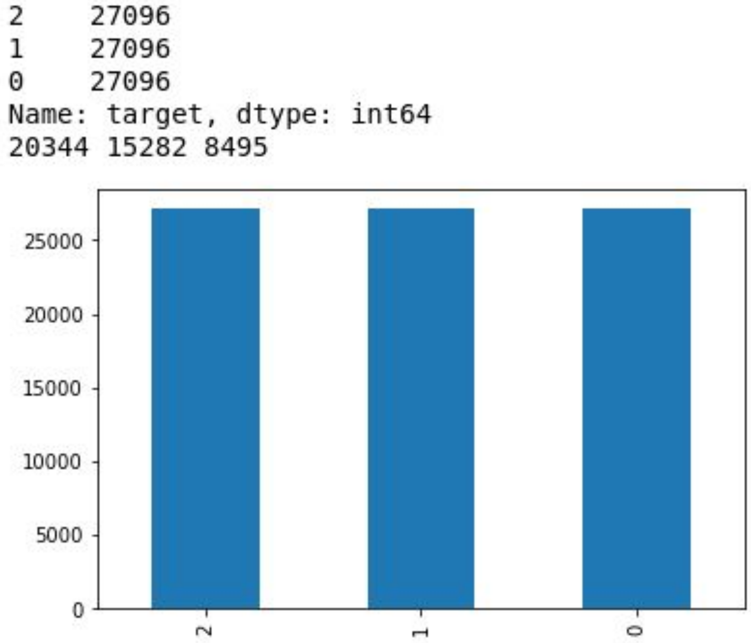
- We take such a tree with simple classes and convert some classes to classify as "IDK" ("I don't know")

Outcomes: Week 7



```
Normal Sampling
              precision    recall  f1-score   support

           0       0.47      0.82      0.60      6752
           1       0.43      0.22      0.07      5259
           2       0.35      0.04      0.07      2696

    accuracy                           0.46     14707
   macro avg       0.42      0.36      0.32     14707
weighted avg       0.43      0.46      0.39     14707
```

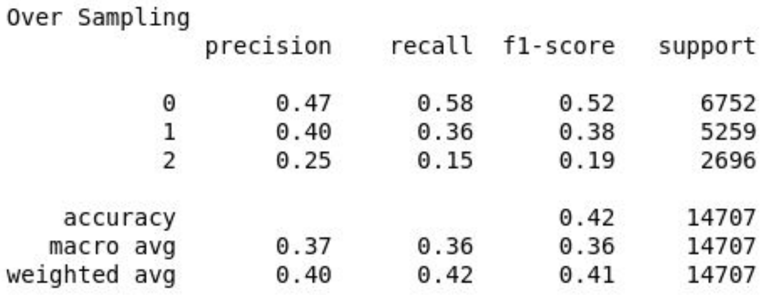
Original score of Random Forest Classifier



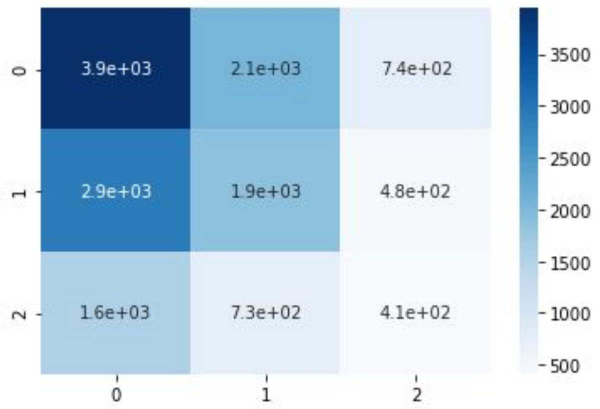
```
2    27096
1    27096
0    27096
Name: target, dtype: int64
20344 15282 8495
```

Over Sampling class_1 and class_2

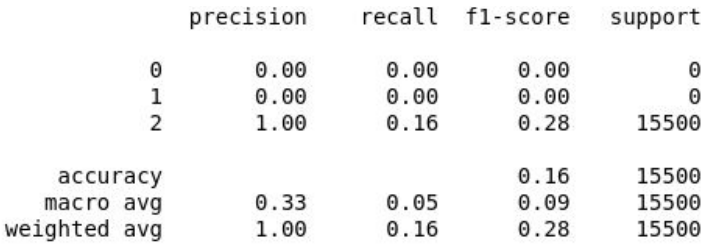
Outcomes: Week 7



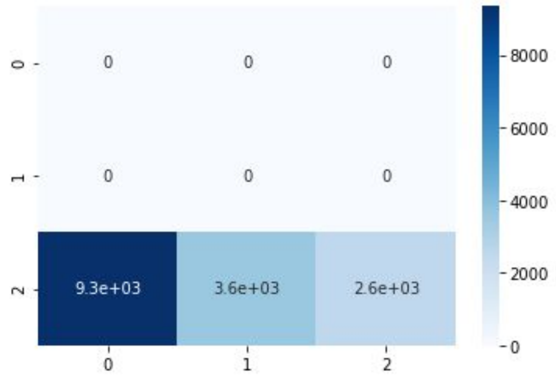
```
Over Sampling
              precision    recall  f1-score   support

           0       0.47      0.58      0.52      6752
           1       0.40      0.36      0.38      5259
           2       0.25      0.15      0.19      2696

    accuracy                           0.42     14707
   macro avg       0.37      0.36      0.36     14707
weighted avg       0.40      0.42      0.41     14707
```



Over Sampling
class_1 and class_2

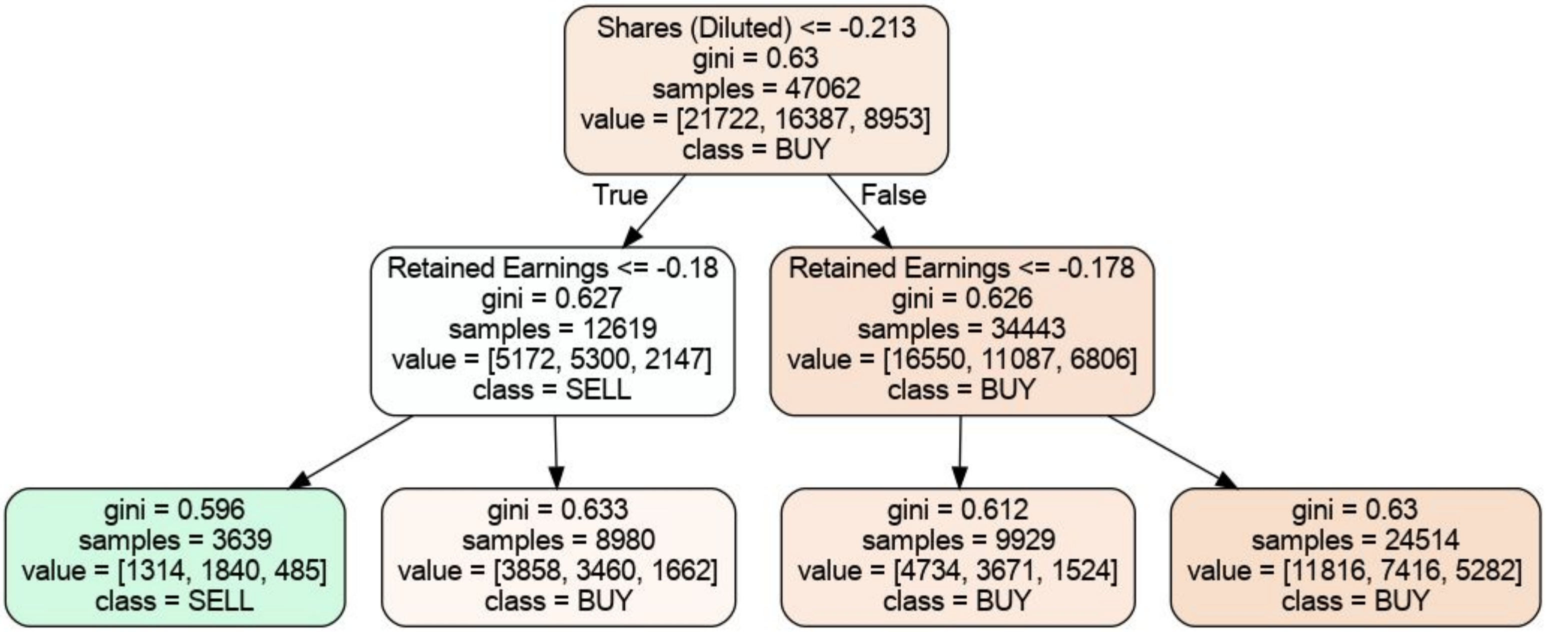


```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         0
           1       0.00      0.00      0.00         0
           2       1.00      0.16      0.28     15500

    accuracy                           0.16     15500
   macro avg       0.33      0.05      0.09     15500
weighted avg       1.00      0.16      0.28     15500
```

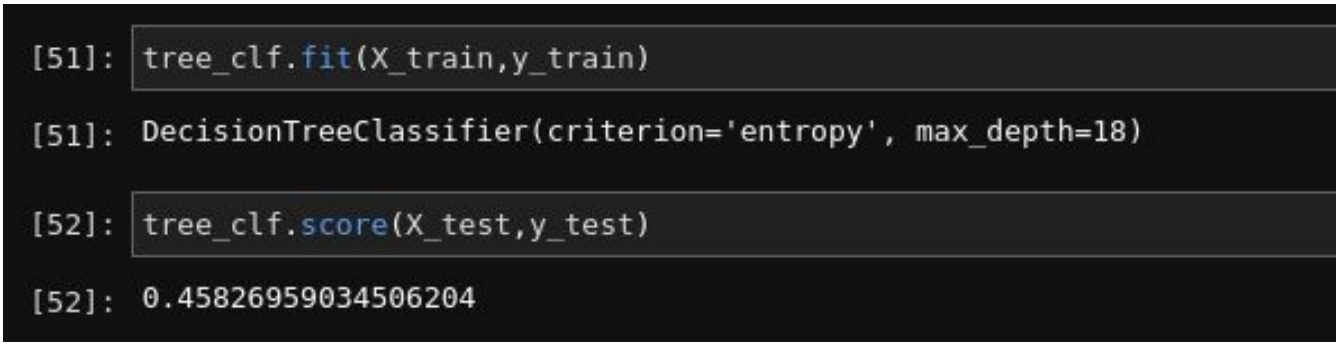


Over Sampling
class_1 and class_2

Outcomes: Week 7



Metrics before the addition of extra class:

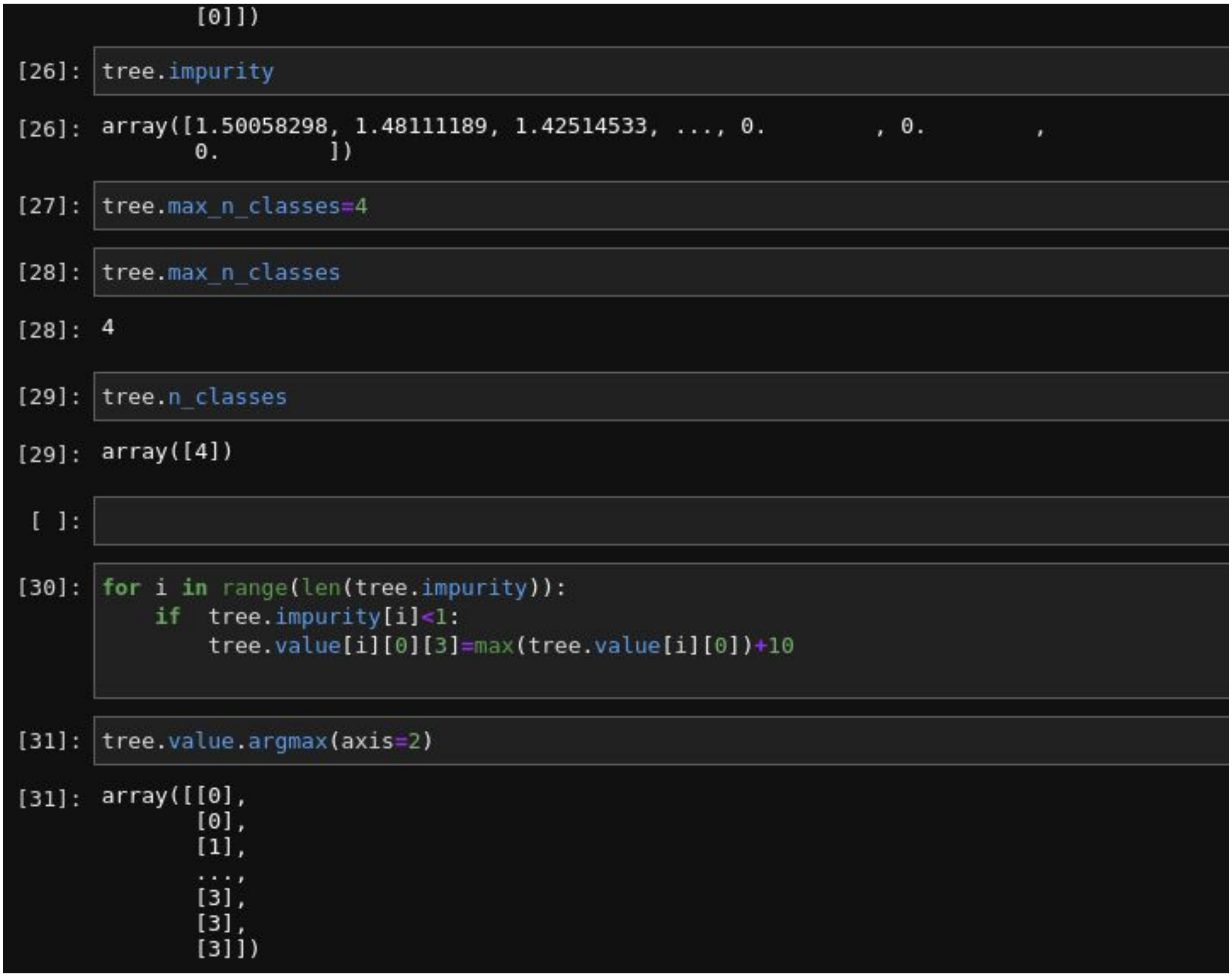


```
[51]: tree_clf.fit(X_train,y_train)

[51]: DecisionTreeClassifier(criterion='entropy', max_depth=18)

[52]: tree_clf.score(X_test,y_test)

[52]: 0.45826959034506204
```

Outcomes: Week 7

Addition of extra class:

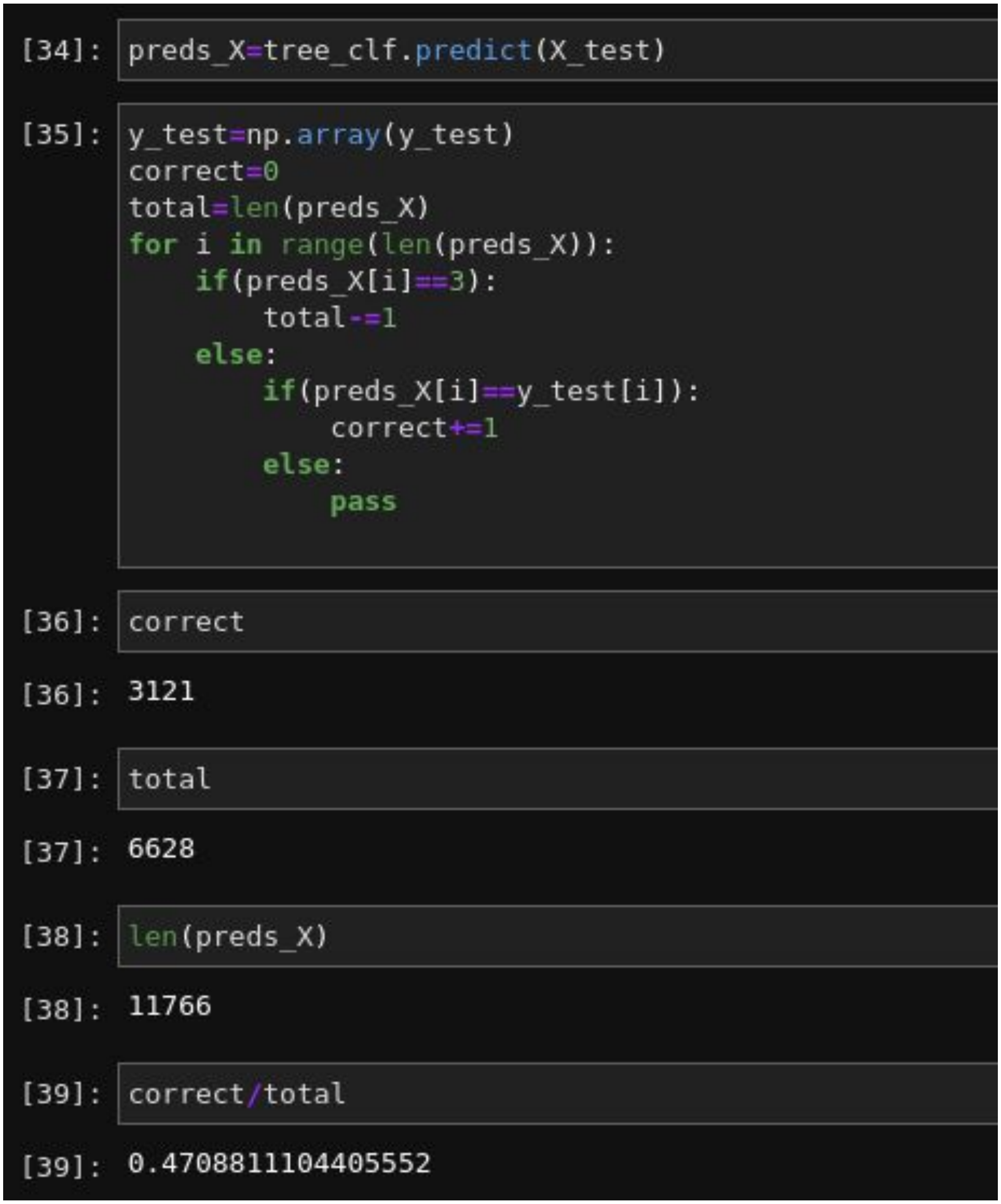
1. Find impurity values of all nodes
2. Kill certain nodes that you feel are overfitting/underfitting
3. Assign label 3 to these nodes



```
             [0]])

[26]: tree.impurity

[26]: array([1.50058298, 1.48111189, 1.42514533, ..., 0.        , 0.        ,
             0.        ])

[27]: tree.max_n_classes=4

[28]: tree.max_n_classes

[28]: 4

[29]: tree.n_classes

[29]: array([4])

[ ]:

[30]: for i in range(len(tree.impurity)):
          if  tree.impurity[i]<1:
              tree.value[i][0][3]=max(tree.value[i][0])+10

[31]: tree.value.argmax(axis=2)

[31]: array([[0],
             [0],
             [1],
             ...,
             [3],
             [3],
             [3]])
```

Outcomes: Week 7

Metrics after killing some nodes of the tree:



```
[34]: preds_X=tree_clf.predict(X_test)

[35]: y_test=np.array(y_test)
      correct=0
      total=len(preds_X)
      for i in range(len(preds_X)):
          if(preds_X[i]==3):
              total-=1
          else:
              if(preds_X[i]==y_test[i]):
                  correct+=1
              else:
                  pass

[36]: correct

[36]: 3121

[37]: total

[37]: 6628

[38]: len(preds_X)

[38]: 11766

[39]: correct/total

[39]: 0.4708811104405552
```

Upcoming Week

01 Sampling dataset with 6% tolerance

02 Compiling results