SENTIMENT ANALYSIS
-
AMAZON APPLIANCE REVIEWS

MILESTONE #4

TEAM: **GOURI KULKARNI JUSTINO CALVILLO**

PROBLEM STATEMENT

- To analyze customer feedback using ML. Customer buys an appliance and writes a review.
- Amazon Appliance Reviews dataset will be used to train and test ML methods
- Our goal is getting the best ML method to predict if an incoming review is either positive or negative.

AGENDA

- First Iteration: Evaluate different Models and select one.
- Second Iteration: Freeze the Model and evaluate different numbers of most important features.
- Third Iteration: Tuning the parameters
- Fourth Iteration: Freeze the selected number of features and use them in different Models

FIRST ITERATION - SUMMARY

- From 4 models to choose the best model
- Dataset has the full set of features
- Benchmark accuracy & learning curves
- Words unigram (1,1)

Review Train 800
Review Test 200

Overall Train 800
Overall Test 200

Shape of train vectors (800, 2843)
Shape of test vectors (200, 2843)

Accuracy for each model with full set of features

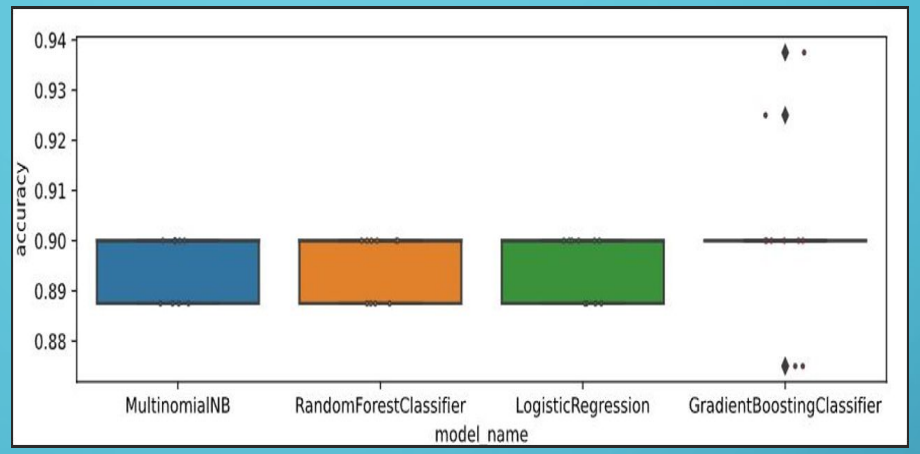
Shape of train vectors (800, 2843)
Shape of test vectors (200, 2843)

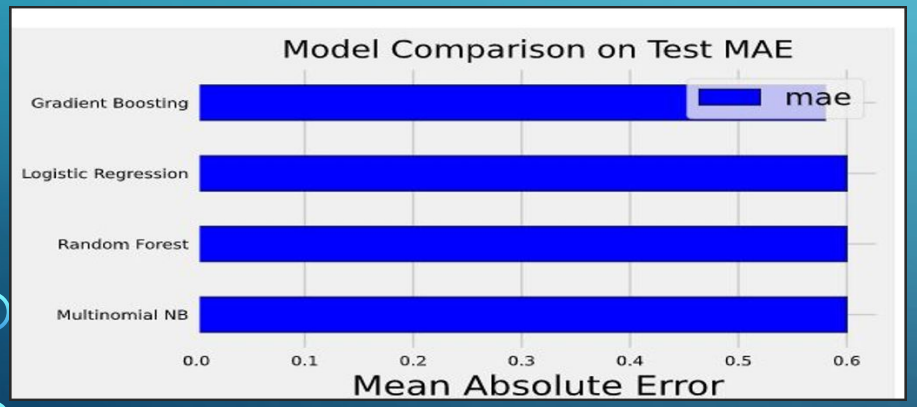
BASE ESTIMATOR RANDOM FOREST ACCURACY 86%

FIRST ITERATION

Accuracy for each model with full set of features
Shape of train vectors (800, 2843)
Shape of test vectors (200, 2843)

Accuracy score for MultinomialNB 0.85
Accuracy score for RandomForestClassifier 0.85
Accuracy score for LogisticRegression 0.85
Accuracy score for GradientBoostingClassifier 0.855

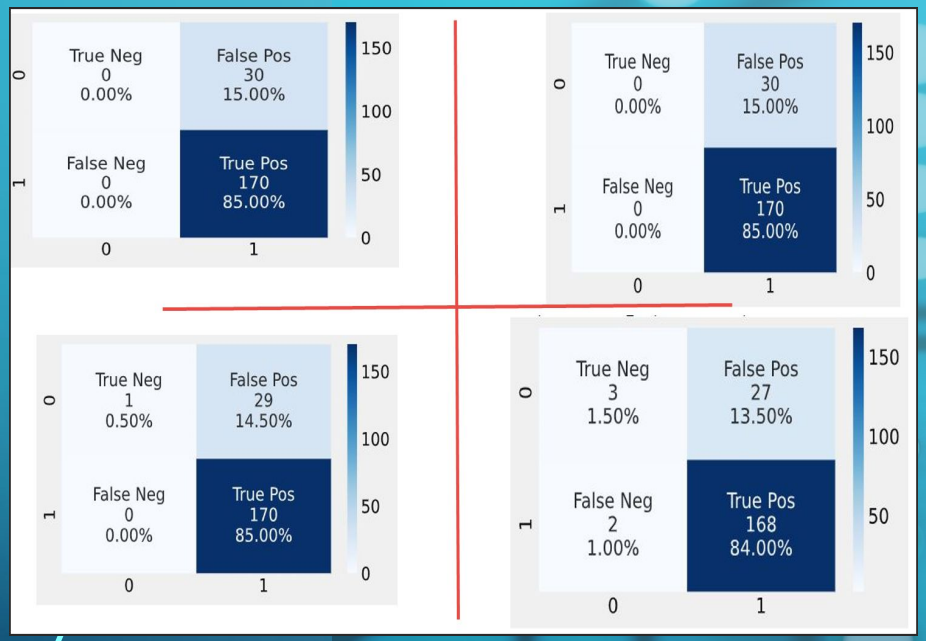




MAE for each model with full set of features
Shape of train vectors (800, 2843)
Shape of test vectors (200, 2843)

MAE for MultinomialNB 0.6
MAE for RandomForestClassifier 0.6
MAE for LogisticRegression 0.6
MAE for GradientBoostingClassifier 0.58

FIRST ITERATION




```
CR for each model with full set of features
Shape of train vectors  (800, 2843)

Shape of test vectors  (200, 2843)

Classification Report for  MultinomialNB
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00        30
         5.0       0.85      1.00      0.92       170

    accuracy                           0.85       200
   macro avg       0.42      0.50      0.46       200
weighted avg       0.72      0.85      0.78       200

Classification Report for  RandomForestClassifier
              precision    recall  f1-score   support

         1.0       1.00      0.03      0.06        30
         5.0       0.85      1.00      0.92       170

    accuracy                           0.85       200
   macro avg       0.93      0.52      0.49       200
weighted avg       0.88      0.85      0.79       200

Classification Report for  LogisticRegression
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00        30
         5.0       0.85      1.00      0.92       170

    accuracy                           0.85       200
   macro avg       0.42      0.50      0.46       200
weighted avg       0.72      0.85      0.78       200

Classification Report for  GradientBoostingClassifier
precision    recall  f1-score   support

         1.0       0.60      0.10      0.17        30
         5.0       0.86      0.99      0.92       170

    accuracy                           0.85       200
   macro avg       0.73      0.54      0.55       200
weighted avg       0.82      0.85      0.81       200
```

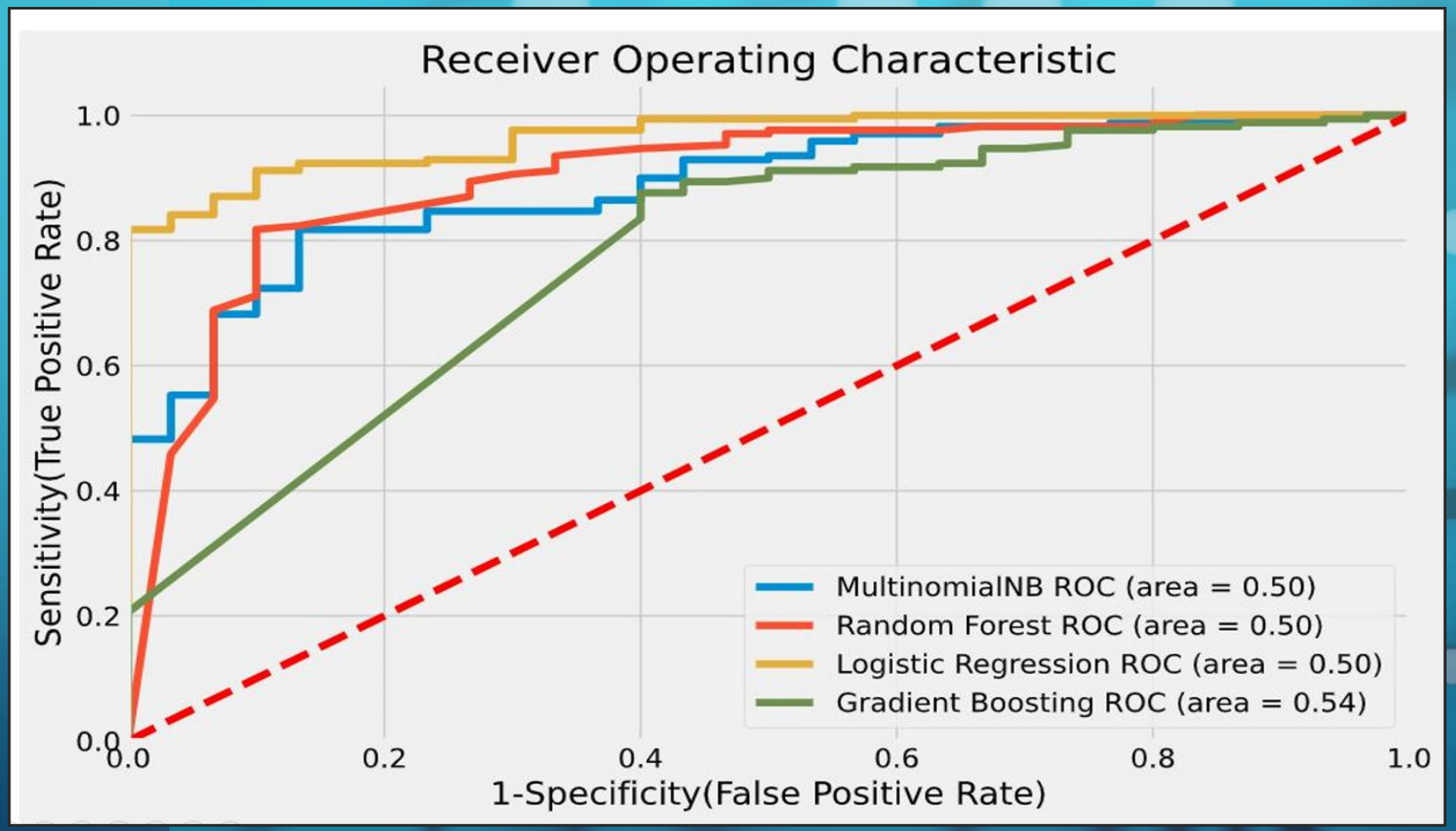
- Accuracy tp + tn/ all
- Precision Proportion of predictions actually correct
- TP/ All positives

- RF chosen because it gave a good precision score on both good and bad reviews
- When it predicts, it is correct most of the time

- As our classes are balanced via random oversampling , and a true positive or true negative review prediction could impact the business , we use accuracy

- Had we used the unbalanced set, we could have used f1 score but it would have resulted in biased predictions

FIRST ITERATION

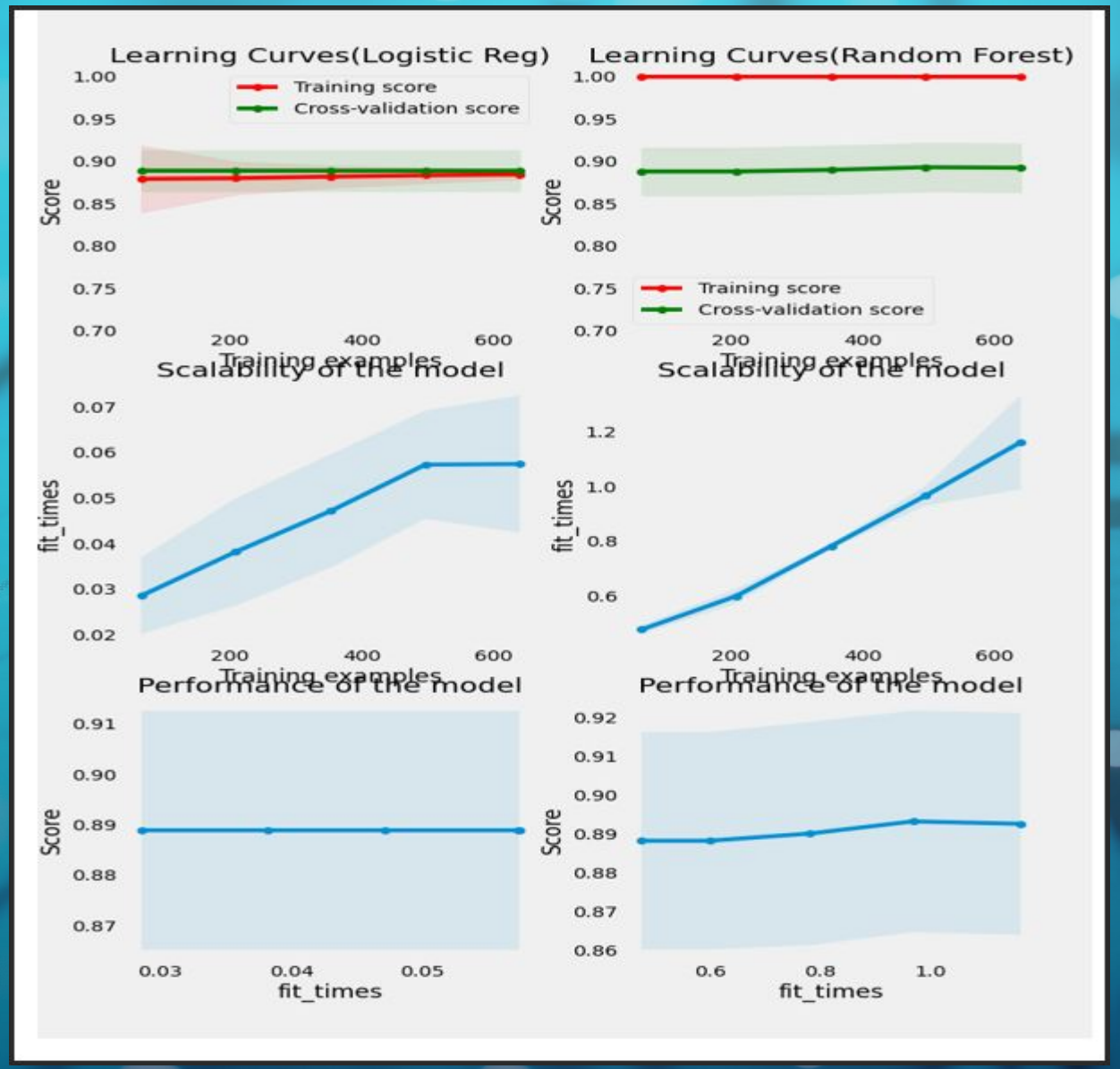


Tradeoff between tpr and (1- fpr) Or sensitivity and specificity

ROC does not depend on class distribution unlike f1 score

Though Random forest scored lower than LR , LR may perform worse

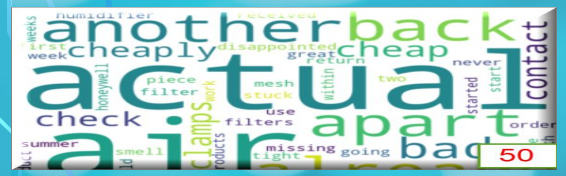
FIRST ITERATION



As lr and rf results were close for AUC and F1 we went on to plot the learning curves
The results were similar

Next step :
Iteration 2 -feature selection and model interpretation

SECOND ITERATION – FEATURE SELECTION



50

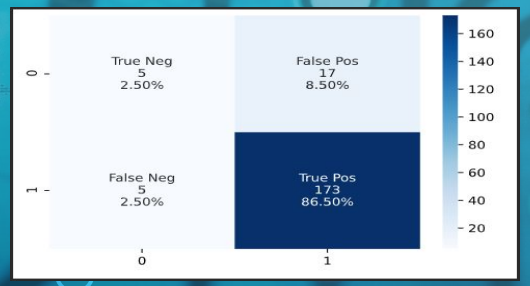
	precision	recall	f1-score	support
1	0.50	0.18	0.27	22
5	0.91	0.98	0.94	178
accuracy			0.89	200
macro avg	0.70	0.58	0.60	200
weighted avg	0.86	0.89	0.87	200

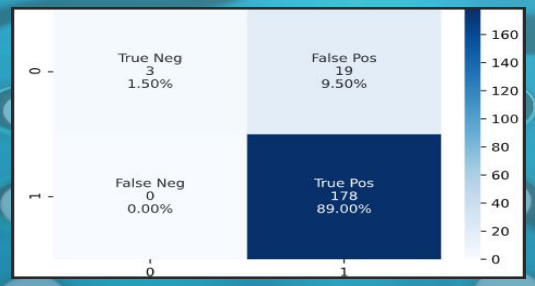
250

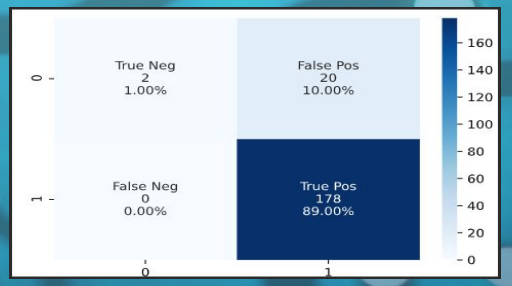
	precision	recall	f1-score	support
1	1.00	0.14	0.24	22
5	0.90	1.00	0.95	178
accuracy			0.91	200
macro avg	0.95	0.57	0.59	200
weighted avg	0.91	0.91	0.87	200

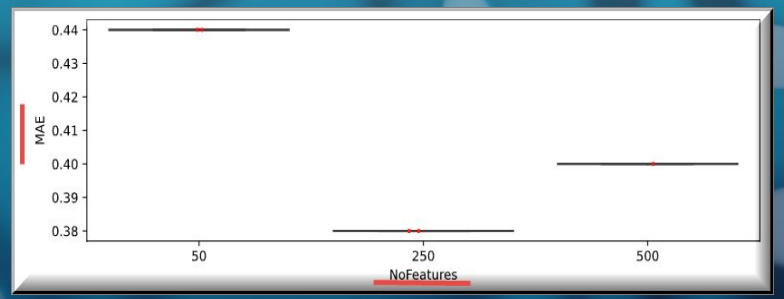
500

	precision	recall	f1-score	support
1	1.00	0.09	0.17	22
5	0.90	1.00	0.95	178
accuracy			0.90	200
macro avg	0.95	0.55	0.56	200
weighted avg	0.91	0.90	0.86	200









- SELECTED MODEL: RANDOM FOREST

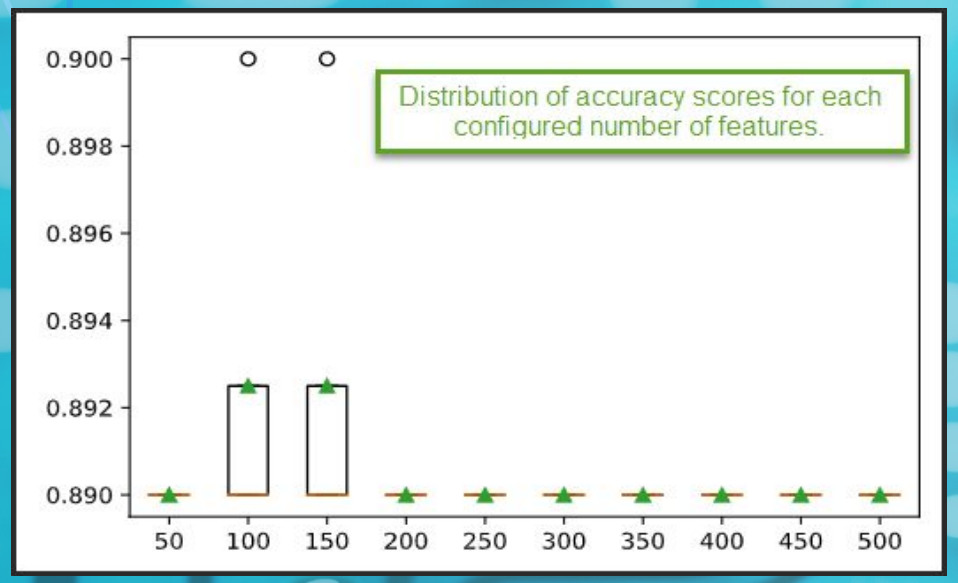
  SELECTED FEATURES: 50, 250, 500 FOR RFE

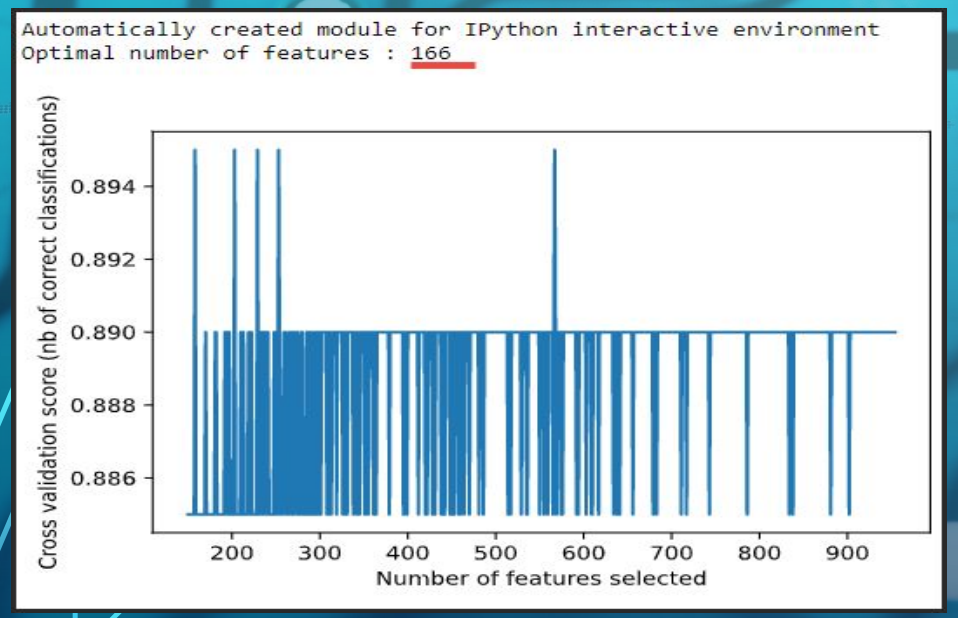
  DATASET: 1000

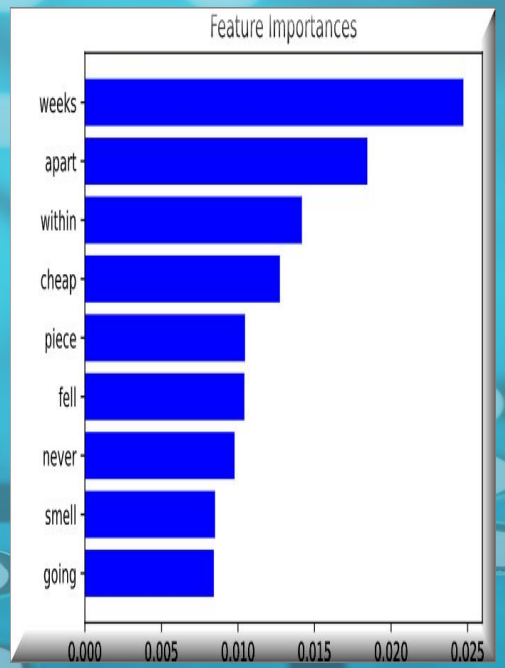
- Conclusions:

- For the Random Forest 250 is the best number of features taken from the RFE
- Accuracy is the highest, MAE the lowest.

SECOND ITERATION – FEATURE SELECTION







- SELECTED MODEL: RANDOM FOREST

  SELECTED FEATURES: 50, 250, 500 FOR RFE

  DATASET: 1000

- Conclusions:

- RFE was run for a range (50,500) features
- Best accuracy happened for 150 features but this result was not the same as the 91 getting with 250

THIRD ITERATION

Base model RF on the full feature set

```
The accuracy of the classifier on the train set was: 100.0
The accuracy of the classifier on the test set was: 86.0
```

Run 1 192 candidates	Run 2 192 candidates
N_estimators 10	N_estimators 10
Max_features sqrt	Max_features sqrt
Cv 2	CV 10
The accuracy on the train set was: 91.25 The accuracy on the test set was: 86.0	The accuracy on the train set was: 89.5 The accuracy on the test set was: 85.0

Increasing folds did no help with the accuracy

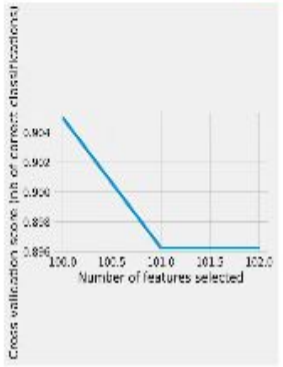
- We tuned the model, cross validated the results, evaluated the performance on the train and test set
- Grid search CV to validation to estimate model accuracy on unseen data
- Results with 10 estimators and two different folds

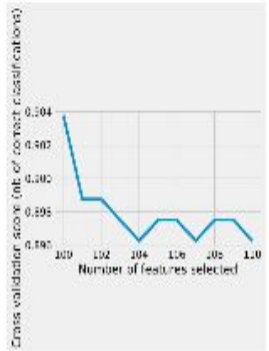
THIRD ITERATION

Base model RF - running RFECV

N_estimators 100	N_estimators 100
RFECV step size .5	RFECV step size .1
CV method StratifiedKFold (5)	CV method StratifiedKFold (5)
Minimum features to select 100	Minimum features to select 100

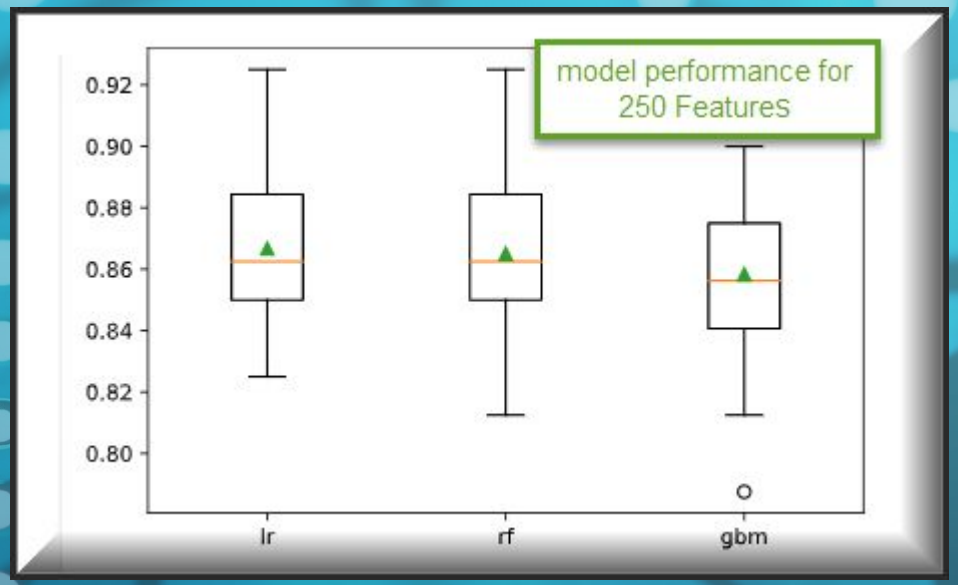
Increased number of estimators from 10 to 100





- Feature selection to drop unwanted features

FORTH ITERATION



- RFE with 250 features running with wrapped algorithms
- And three ML methods

THANK YOU