Topics and Tools on Social Media Data Mining

CS529

Assignment 2

Report on

Understanding various link prediction methods (also known as network completion methods) and its application on various real-world problem

Bhupender (194101014)

Vandana Mishra (194101055)

Mayank Verma (194101031)

IIT GUWAHATI

Abstract:

Aim: To understand various link prediction methods (also known as network completion methods) and its application on various real-world problem.

Datasets to consider:

Foursquare Restaurant Review Dataset:

Number of Nodes: 2060

Number of existing edges: 60870

Number of non-existing edges: 60870 (randomly chosen)

Blog Catalog data:

Number of Nodes: 10312

Number of existing edges: 333983

Number of non-existing edges: 333983 (randomly chosen)

Theme A: Topological Methods (Unsupervised Approaches)

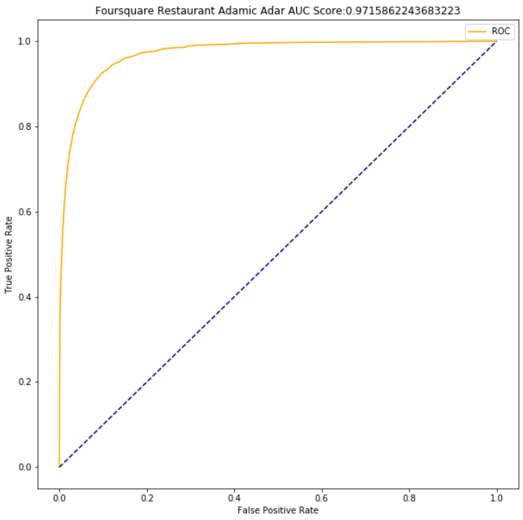
Dataset: Foursquare Restaurant Review Dataset:

1.Adamic Adar

FPR TPR ROC AUC=0.971

Precision=0.937

Recall=0.616

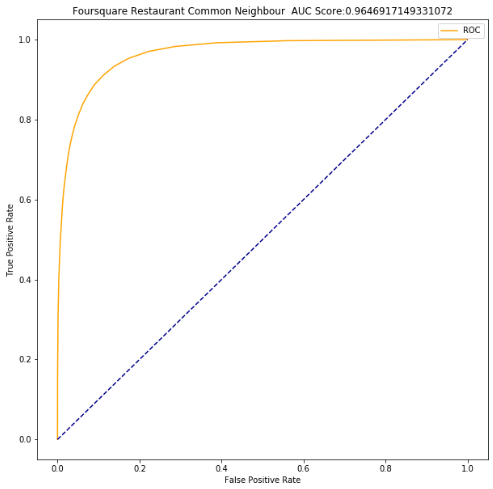


2.Common Neighbour

FPR TPR ROC AUC=0.9646

Precision=0.98122

Recall=0.18399

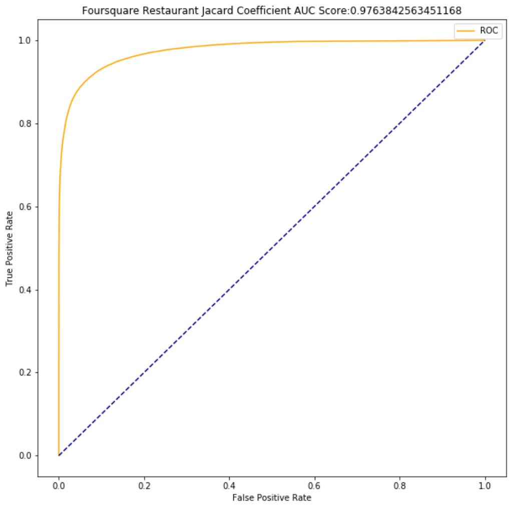


3.Jacard Coefficient

AUC=0.976

Precision=0.9668

Recall=0.4593

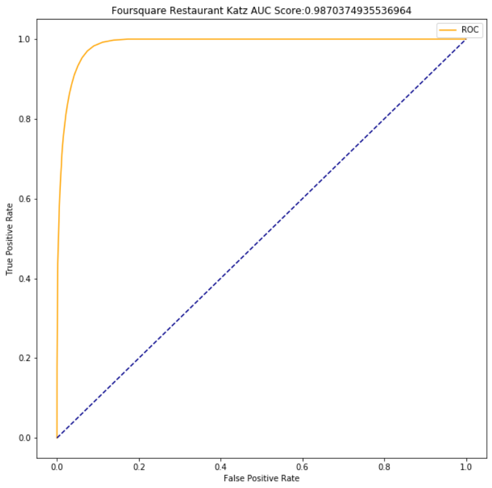


4.Katz

AUC=0.987

Precision=0.9897

Recall=0.2234

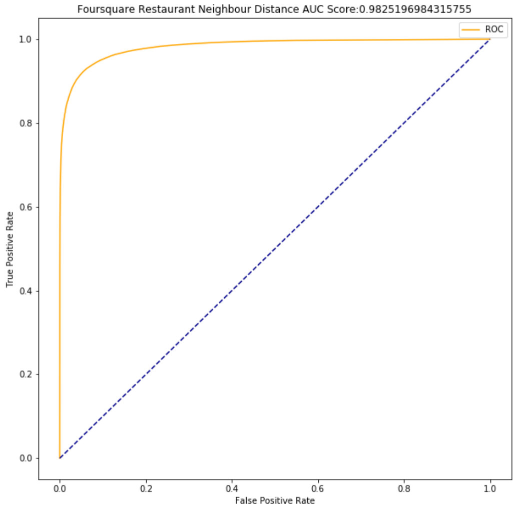


5.Neighbour Distance

AUC=0.982

Precision=0.9517

Recall=0.5840

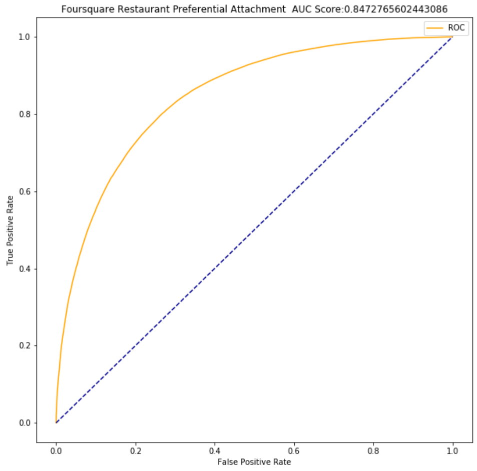


6.Preferential Attachment

AUC=0.8472

Precision=0.8649

Recall=0.3965

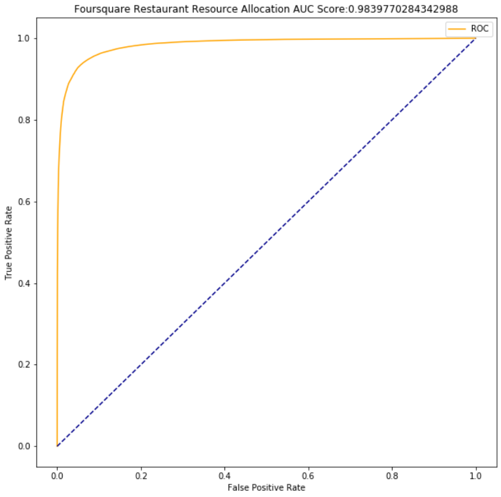


7.Resource Allocation

AUC=0.9839

Precision=0.9470

Recall=0.6260

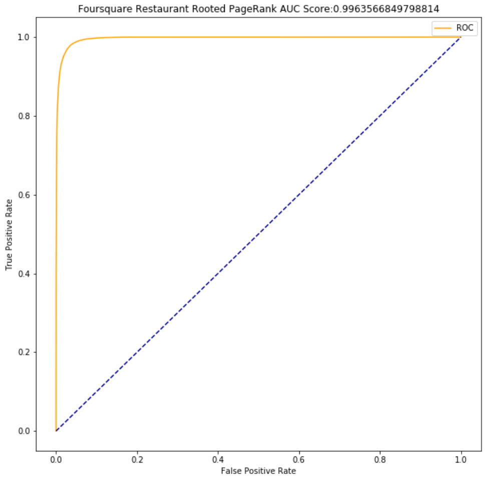


8.Rooted PageRank

AUC=0.9963

Precision=0.9803

Recall=0.5844

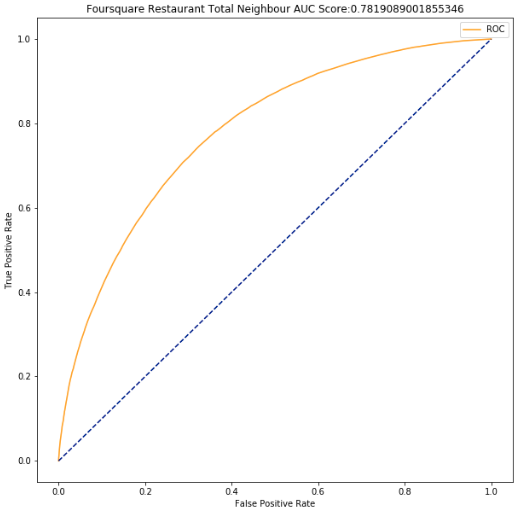


9.Total Neighbour

AUC=0.7819

Precision=0.8299

Recall=0.3003



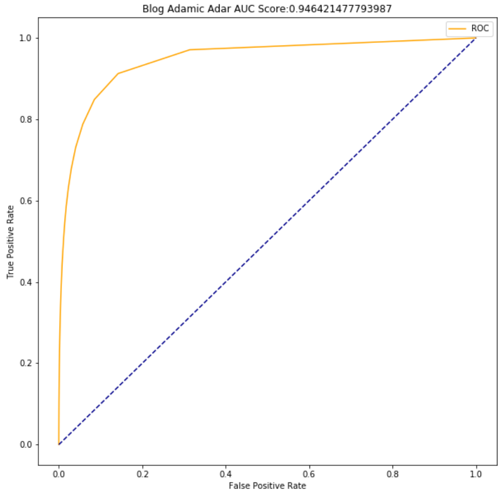
Dataset: Blog Catalog data

1.Adamic Adar

AUC=0.9464

Precision=0.9391

Recall=0.5698

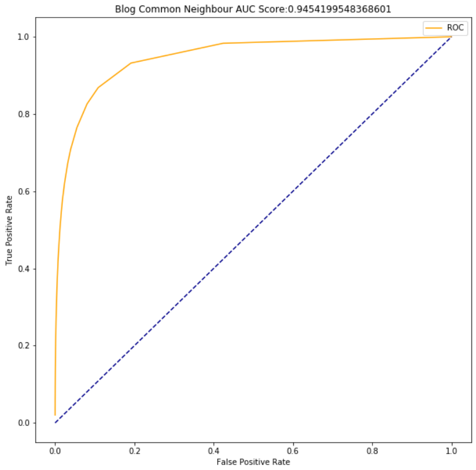


2.Common Neighbours

AUC=0.9454

Precision=0.9956

Recall=0.0473

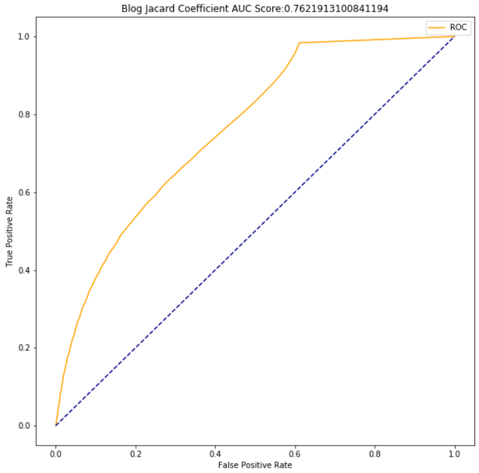


3.Jacard Coefficient

AUC=0.762

Precision=0.7442

Recall=0.4724

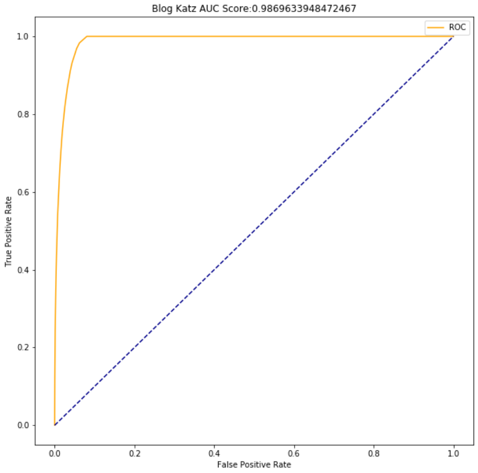


4.Katz

AUC=0.9869

Precision=0.9976

Recall=0.060

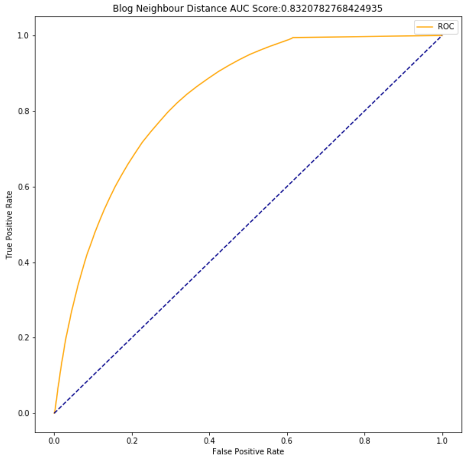


5.Neighbour Distance

AUC=0.8320

Precision=0.8060

Recall=0.4458

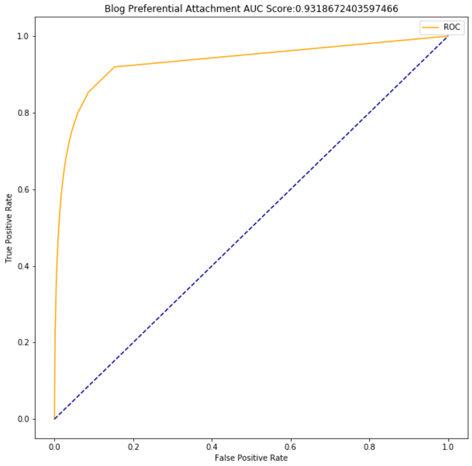


6.Preferential Attachment

AUC=0.9318

Precision=0.9737

Recall=0.3406

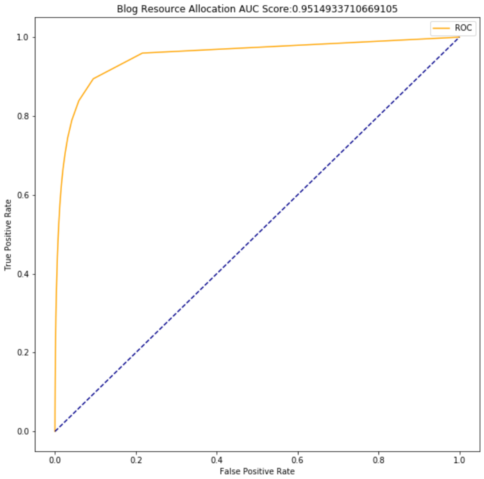


7.Resource Allocation

AUC=0.9514

Precision=0.94517

Recall=0.5759

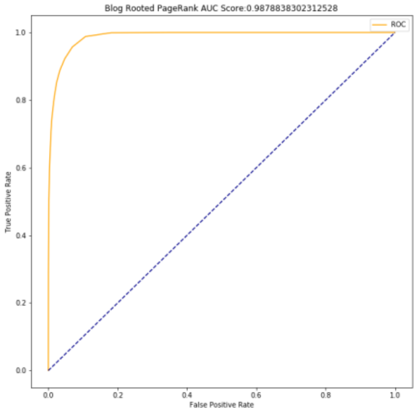


8.Rooted PageRank

AUC=0.9878

Precision=0.9718

Recall=0.5802

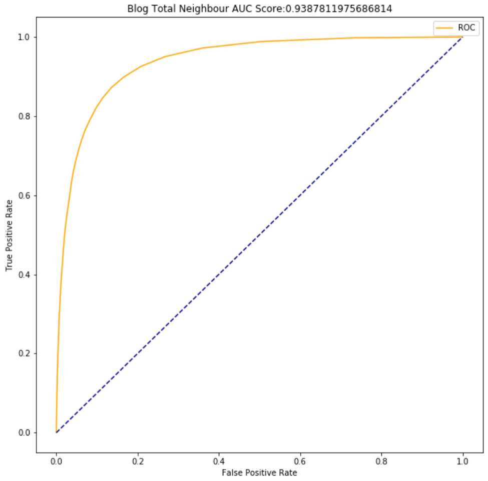


9.Total Neighbours

AUC=0.9387

Precision=0.9651

Recall=0.2569



Theme B: Classification Models (Supervised Approaches)

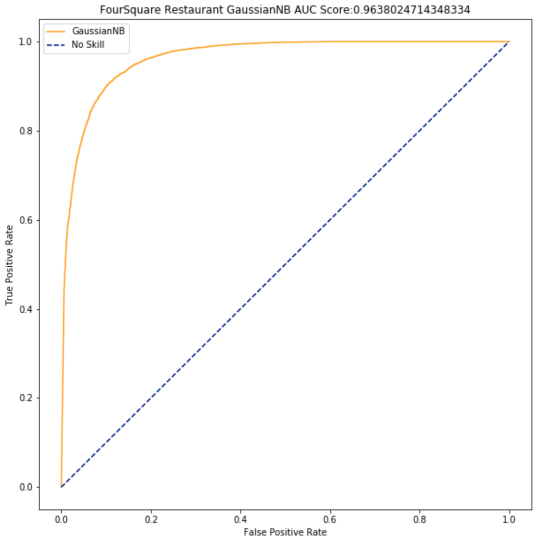
Part I: Explicit Features

Dataset: Foursquare Restaurant Review Dataset

1. **Naive Bayes:**

1.1 ROC:

AUC Score: 0.964



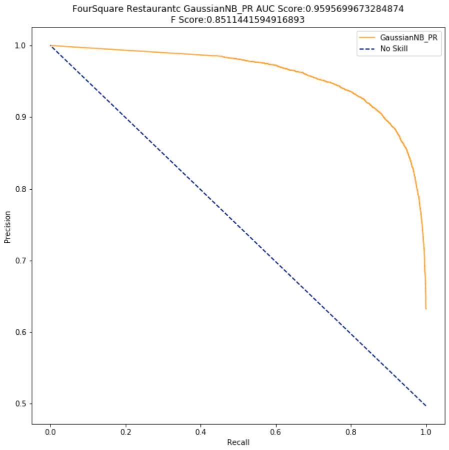
1.2 Precision- Recall Curve:

Precision = 0.842036930187154

Recall = 0.8535681195353716

No Skill: f1=0.000 AUC=0.748

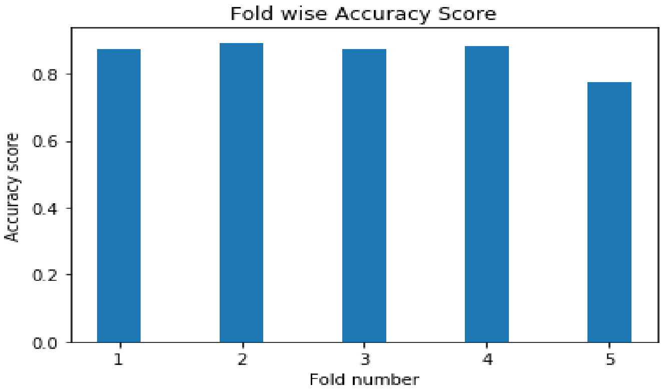
GaussianNB : f1=0.851 AUC=0.960



1.3 Five-Fold Accuracy Score:

Fold 1 = 0.87550134, Fold2 = 0.89217079, Fold3 = 0.87395555, Fold4=0.88260361 Fold5 = 0.77281083

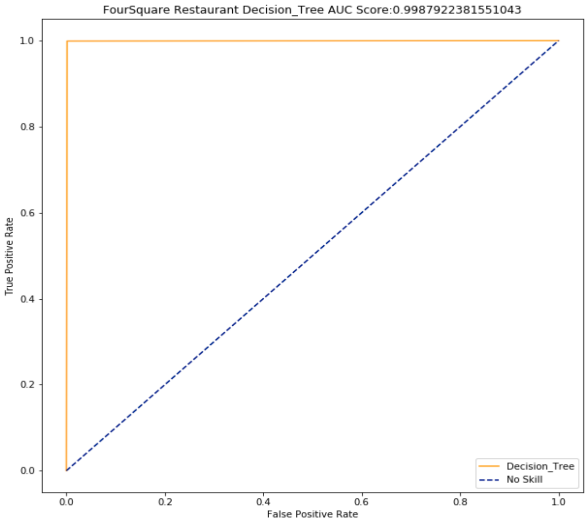
Mean Accuracy: 0.859408422459893



2. Decision Tree (ID3):

2.1 ROC:

AUC Score: 0.999



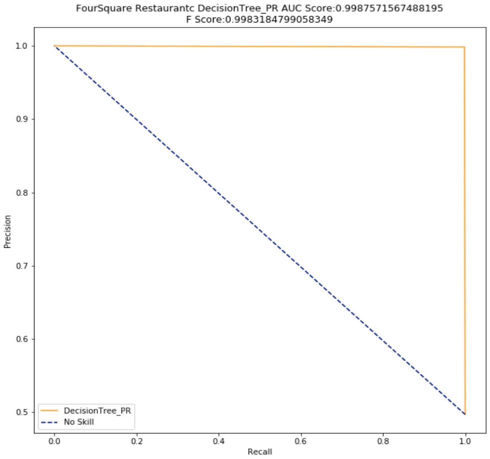
2.2 Precision- Recall Curve:

Precision = 0. 666078184110971

Recall = 0. 8317842073757643

No Skill: f1=0.000 AUC=0.748

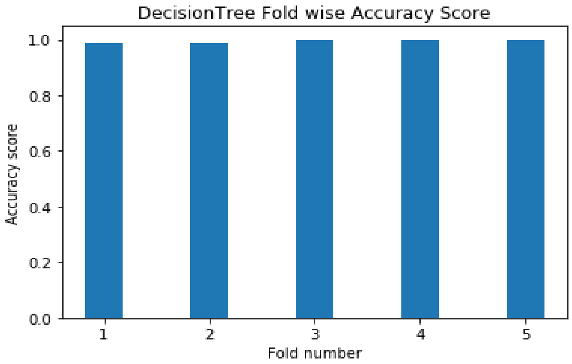
Decision Tree: f1=0. 998 AUC=0. 999



2.3 Five-Fold Accuracy Score:

Fold 1 = 0. 98943015, Fold2 = 0. 99093416, Fold3 = 0. 99853777, Fold4=0.99920622 Fold5 = 0. 99958222

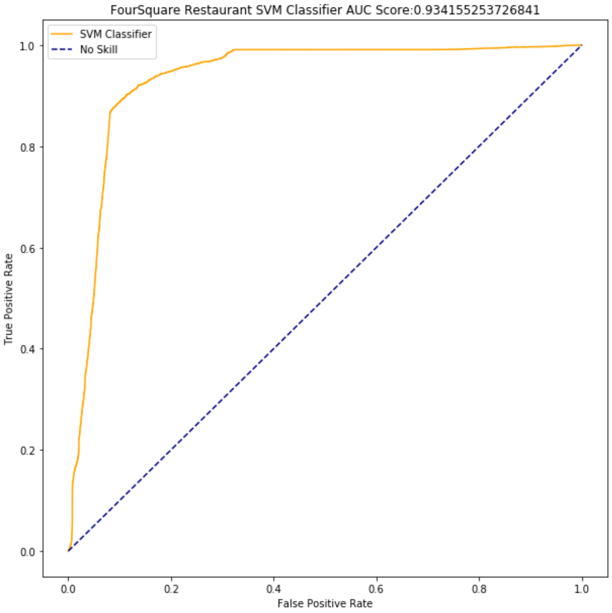
Mean Accuracy: 0. 995538101604278



3. SVM(SVC):

3.1 ROC:

AUC Score: 0.934



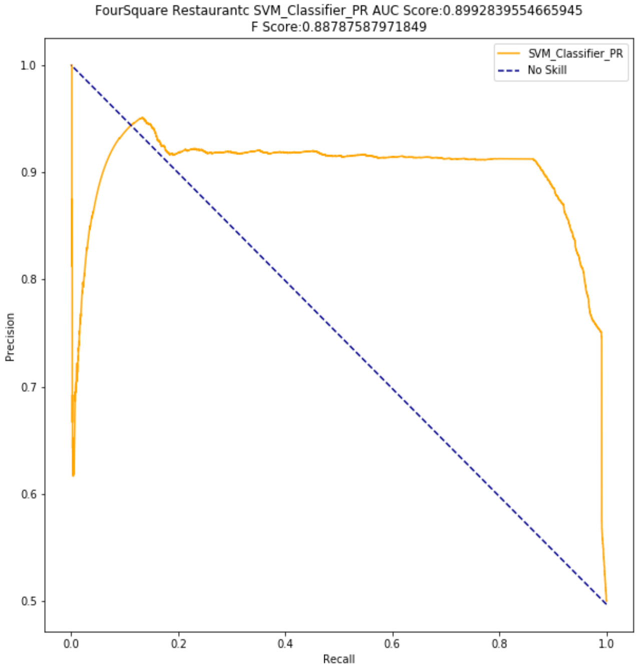
3.2 Precision- Recall Curve:

Precision = 0. 7124517310609713

Recall = 0. 7875836485242941

No Skill: f1=0.000 AUC=0.748

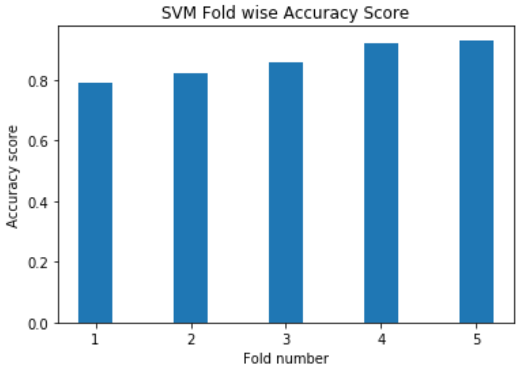
SVM Classifier: f1=0.888 auc=0.899



3.3 Five-Fold Accuracy Score:

Fold 1 = 0. 78939672, Fold2 = 0. 82398897, Fold3 = 0. 85833055, Fold4=0. 91945187 Fold5=0. 93185996

Mean Accuracy: 0. 864605614973262



4. Gradient Boosting:

Learning rate: 0.05Accuracy score (training): 0.984Accuracy score (validation): 0.982

Learning rate: 0.075Accuracy score (training): 0.986Accuracy score (validation): 0.984

Learning rate: 0.1Accuracy score (training): 0.988Accuracy score (validation): 0.987

Learning rate: 0.25Accuracy score (training): 0.997Accuracy score (validation): 0.995

Learning rate: 0.5Accuracy score (training): 0.999Accuracy score (validation): 0.998

Learning rate: 0.75Accuracy score (training): 0.998Accuracy score (validation): 0.998

Learning rate: 1Accuracy score (training): 0.998Accuracy score (validation): 0.997

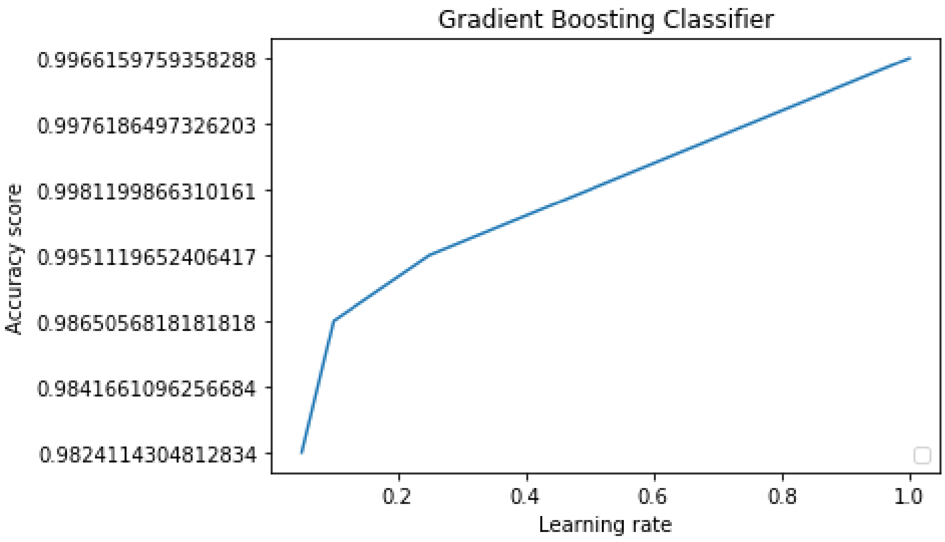
Confusion Matrix:

```
[[12      11946]

 [ 11    11760]]
```

Classification Report

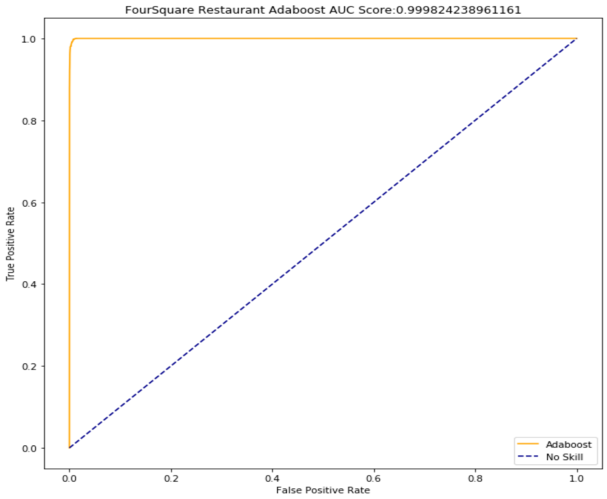
	Precision	Recall	f1-score	Support
0	1.0	1.0	1.0	12165
1	1.0	1.0	1.0	11771
accuracy			1.0	23936
Macro Avg	1.0	1.0	1.0	23936
Weighted Avg	1.0	1.0	1.0	23936



5. <u>Adaboost (n_estimators=30)</u>

5.1 ROC:

AUC Score: 1.000



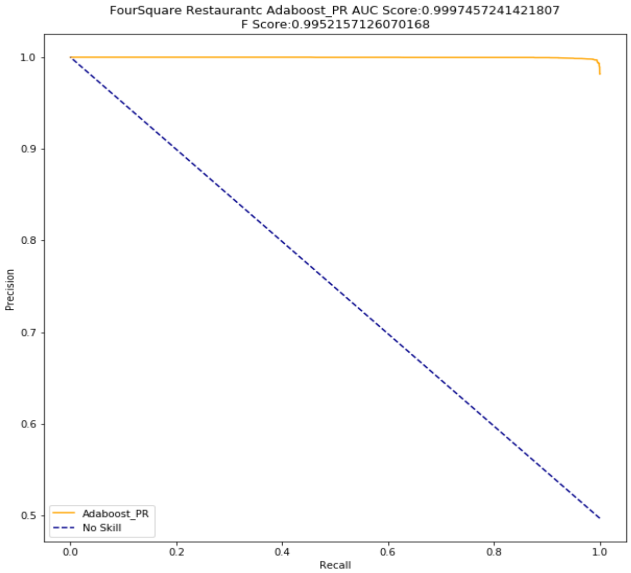
5.2 Precision-Recall

Precision = 0. 7227184949596495

Recall = 0. 9984003354570029

No Skill: f1=0.000 AUC=0.748

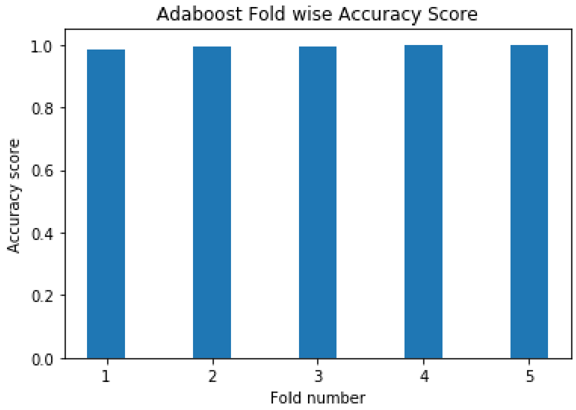
Adaboost : f1=0. 995 AUC= 1.000



5.3 **Five-Fold Accuracy Score:**

Fold 1 = 0. 98429144, Fold2 = 0. 9912266, Fold3 = 0. 99544619, Fold4=0.99837066 Fold5 = 0. 99983289

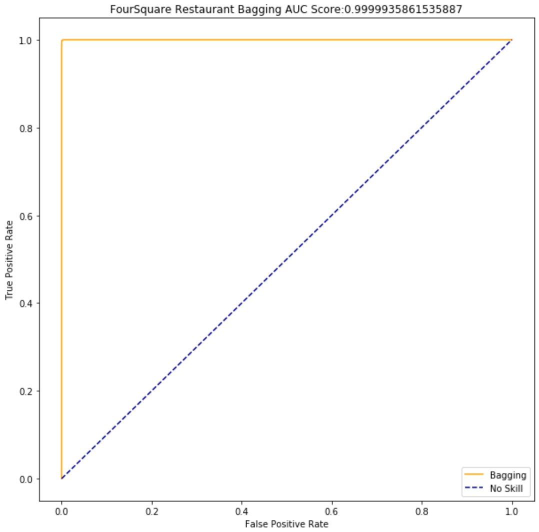
Mean Accuracy: 0. 9938335561497327



6. <u>Bagging Classifier</u>

6.1 ROC:

AUC Score: 1.000



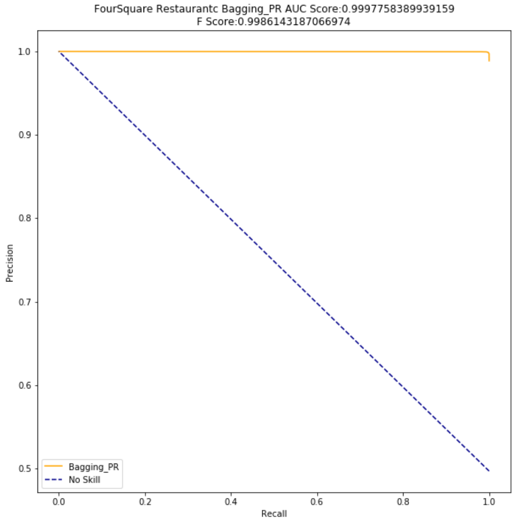
6.2 Precision-Recall

Precision = 0. 897208911307272

Recall = 0. 9969494766771161

No Skill: f1=0.000 AUC=0.748

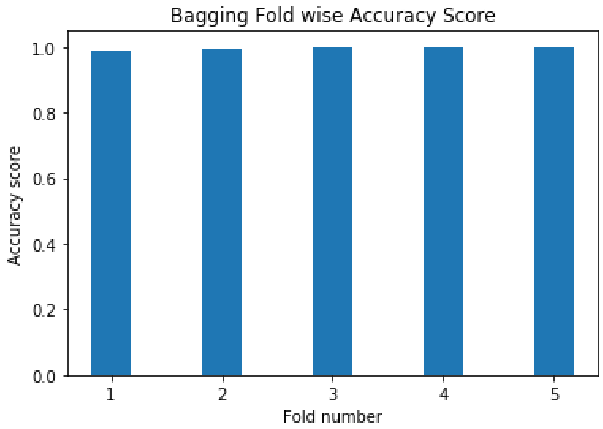
Bagging: f1=0. 999 AUC= 1.000



6.3 Five-Fold Accuracy Score:

Fold1 = 0. 98901237, Fold2 = 0. 99339906, Fold3 = 0. 99908088, Fold4=0. 99949866 Fold5=0. 99991644

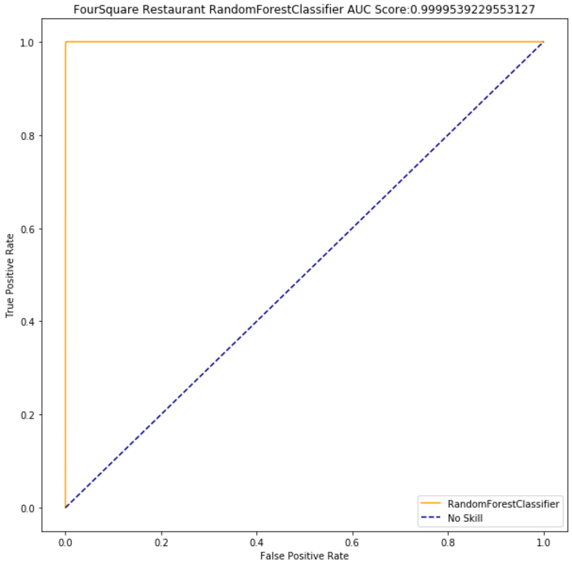
Mean Accuracy: 0. 9961814839572192



7. Random Forest Classifier

7.1 ROC:

AUC Score: 1.000



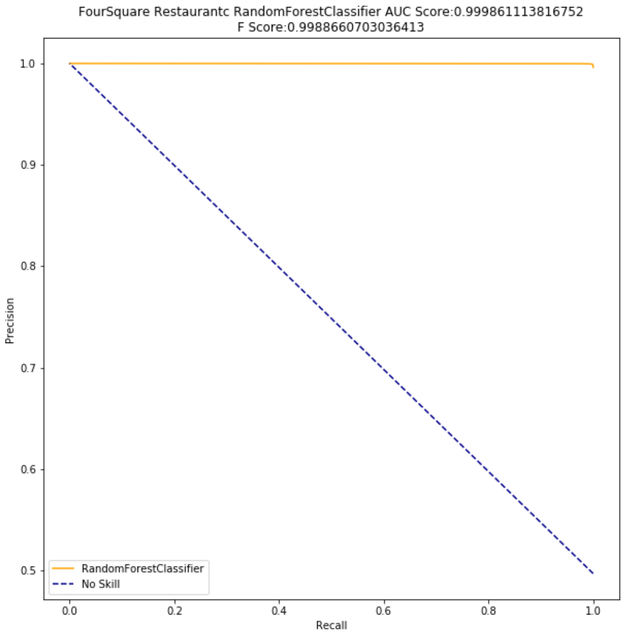
7.2 Precision-Recall

Precision = 0. 8700714585960487

Recall = 0. 9986042840841911

No Skill: f1=0.000 AUC=0.748

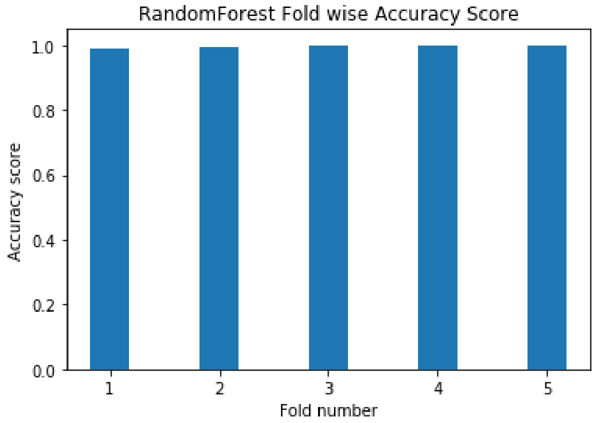
Random Forest Classifier: f1=0.999 AUC=1.000



7.3 Five-Fold Accuracy Score:

Fold1 = 0. 98963904, Fold2 = 0. 99373329, Fold3 = 0. 99857955, Fold4=0. 99916444 Fold5=0. 99983289

Mean Accuracy: 0. 9961898395721924

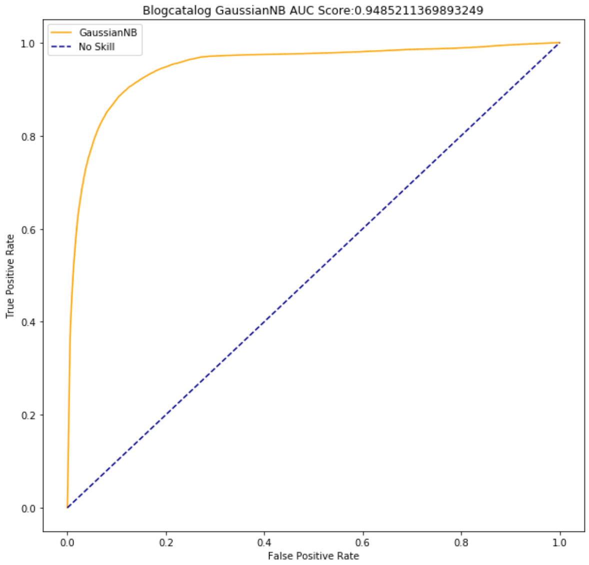


Dataset: blog catalog Review Dataset:

1. Naive Bayes:

1.1 ROC:

AUC Score: 0.949



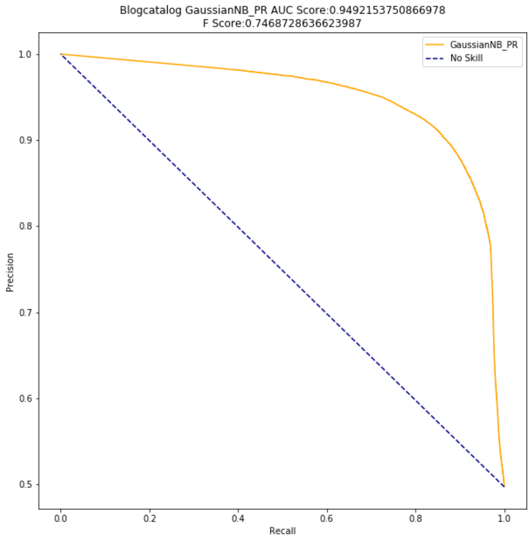
1.2 Precision-Recall

Precision = 0. 8535422766277322

Recall = 0. 7747774111083952

No Skill: f1=0.000 AUC=0. 749

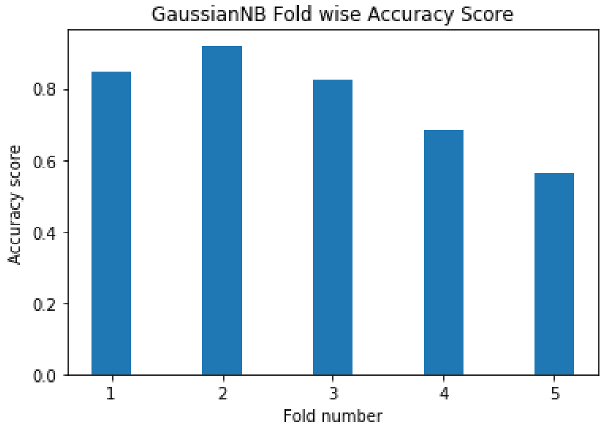
Random Forest Classifier: f1=0. 747 AUC= 949



1.3 Five-Fold Accuracy Score:

Fold1 = 0. 84809947, Fold2 = 0. 91963711, Fold3 = 0. 82487237, Fold4=0. 68138062 Fold5=0. 56191988

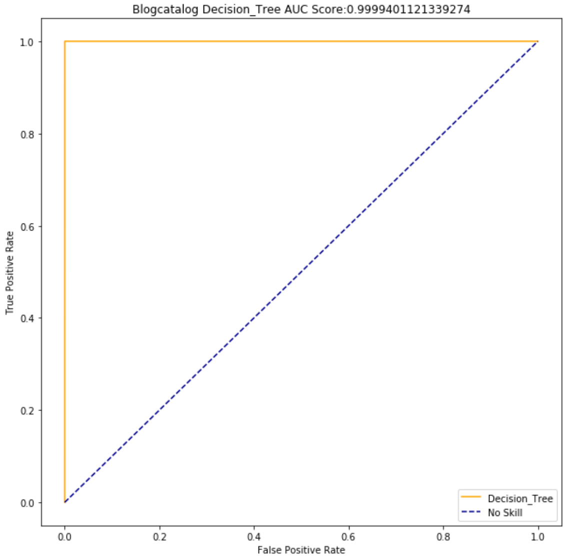
Mean Accuracy: 0. 76718188930789



2. Decision tree

2.1 ROC:

AUC Score: 1.000



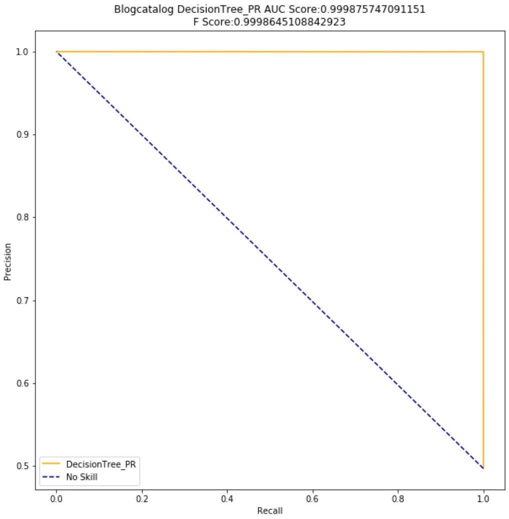
2.2 Precision-Recall

Precision = 0. 666651610960554

Recall = 0. 8323174073930408

No Skill: f1=0.000 AUC=0. 749

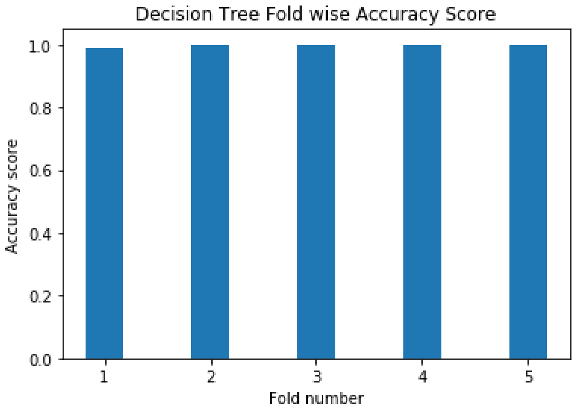
DecisionTree: f1=1.000 AUC=1.000



2.3 Five-Fold Accuracy Score:

Fold 1 = 0. 98936329, Fold2 = 0. 99986526, Fold3 = 0. 99742503, Fold4=0. 99992515 Fold5 = 1.0

Mean Accuracy: 0. 997315747489094



3. Gradient Boosting:

Learning rate: 0.05Accuracy score (training): 0.987Accuracy score (validation): 0.987

Learning rate: 0.075Accuracy score (training): 0.991Accuracy score (validation): 0.991

Learning rate: 0.1Accuracy score (training): 0.991Accuracy score (validation): 0.991

Learning rate: 0.25Accuracy score (training): 1.000Accuracy score (validation): 0.999

Learning rate: 0.5Accuracy score (training): 1.000Accuracy score (validation): 1.000

Learning rate: 0.75Accuracy score (training): 1.000Accuracy score (validation): 1.000

Learning rate: 1Accuracy score (training): 1.000Accuracy score (validation): 1.000

Confusion Matrix:

```
[[66717        18]

[   3        66856]]
```

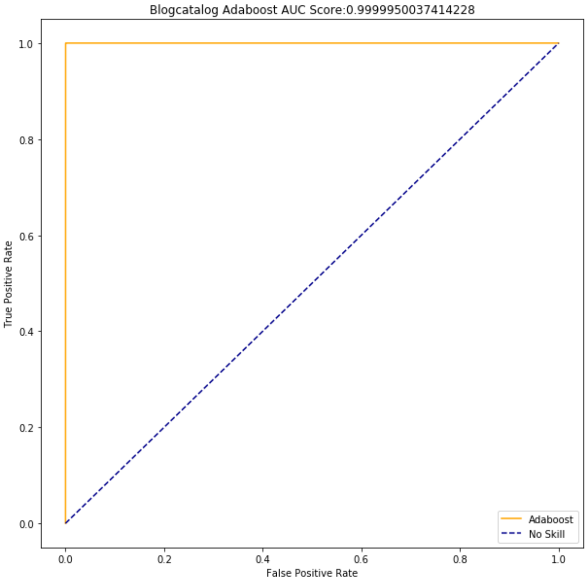
Classification Report

	Precision	Recall	f1-score	Support
0	1.0	1.0	1.0	66735
1	1.0	1.0	1.0	66859
accuracy			1.0	133594
Macro Avg	1.0	1.0	1.0	133594
Weighted Avg	1.0	1.0	1.0	133594

4. Adaboost (n_estimators=500)

4.1 ROC:

AUC Score: 0.999



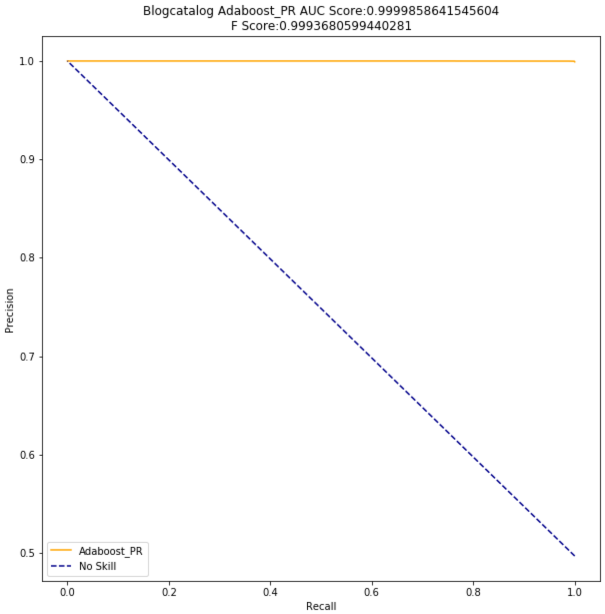
4.2 Precision-Recall

Precision = 0. 7579183830896619

Recall = 0. 9999549883198281

No Skill: f1=0.000 AUC=0. 749

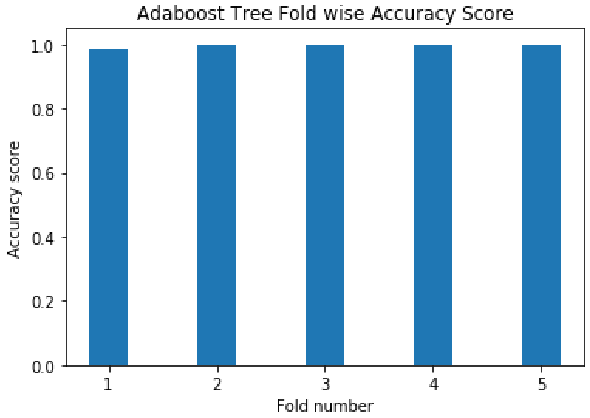
Adaboost: f1=1.000 AUC=1.000



4.3 Five-Fold Accuracy Score:

Fold 1 = 0. 98547839, Fold2 = 0. 99982784, Fold3 = 1.0, Fold4=1.0 Fold5 = 1.0

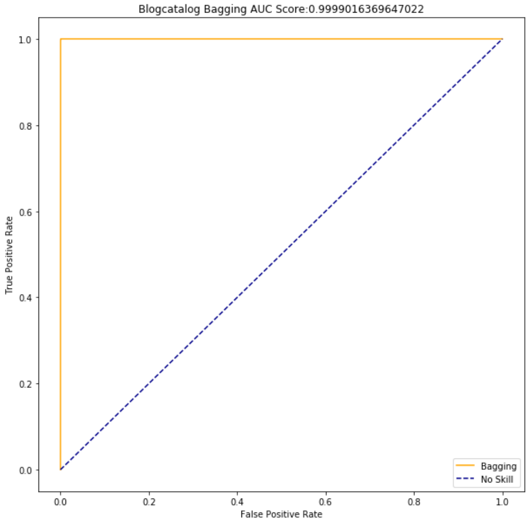
Mean Accuracy: 0. 99706124526550 59



5. **Bagging:**

5.1 ROC:

AUC Score: 0.999



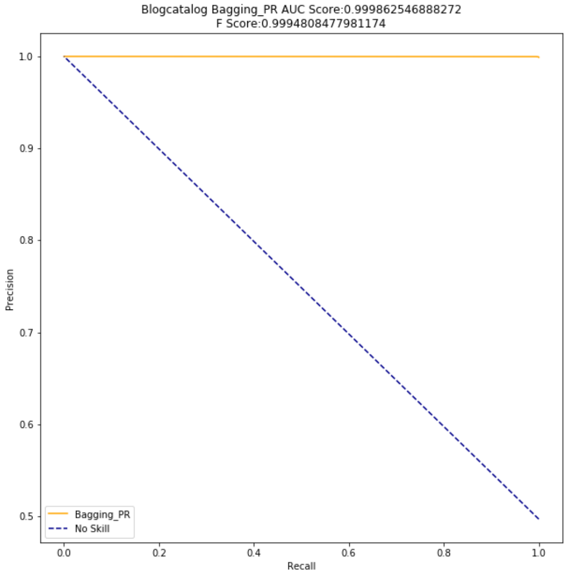
5.2 Precision-Recall

Precision = 0. 7989792231255646

Recall = 0. 9995212591817081

No Skill: f1=0.000 AUC=0. 749

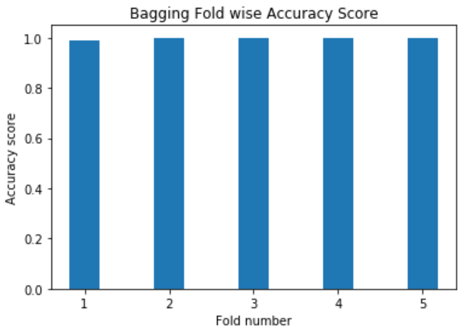
Bagging: f1=0.999 AUC=1.000



5.3 Five-Fold Accuracy Score:

Fold 1 = 0. 98842014, Fold2 = 0. 99981287, Fold3 = 0.99997754, Fold4=0.9999476, Fold5=1.0

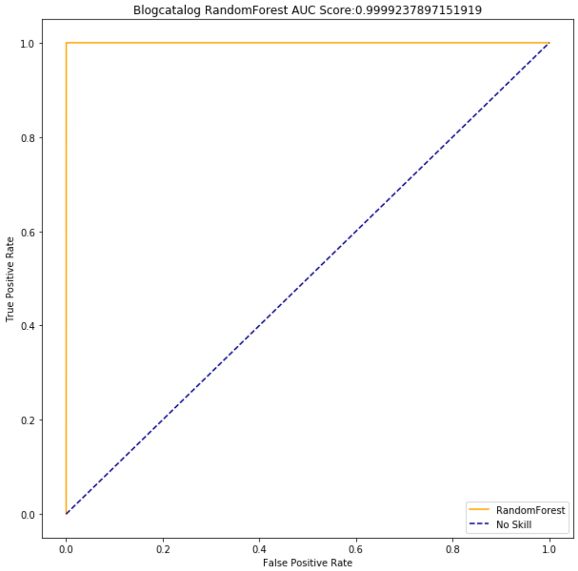
Mean Accuracy: 0. 9976316300061431



6. Random Forest Classifier

6.1 ROC:

AUC Score: 0.999



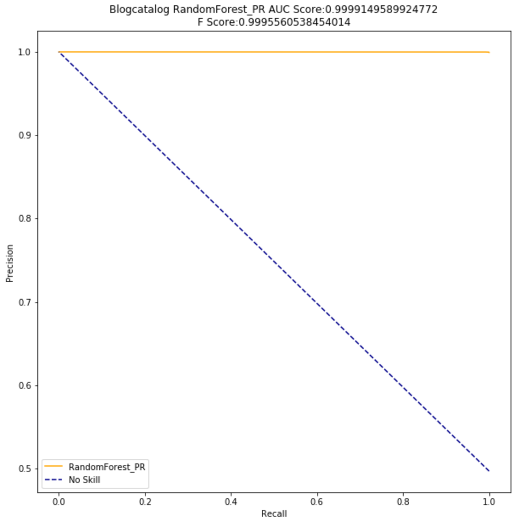
6.2 Precision-Recall

Precision = 0. 79826859379970491

Recall = 0. 9996626155342883

No Skill: f1=0.000 AUC=0. 749

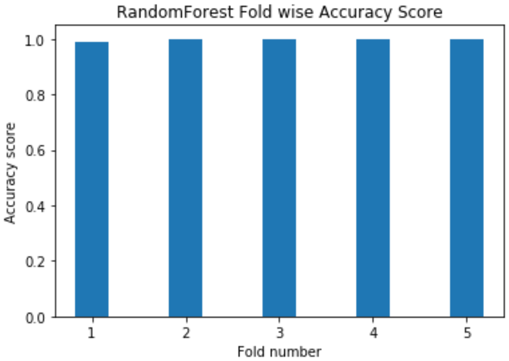
Random Forest: f1=1.000 AUC=1.000



6.3 Five-Fold Accuracy Score:

Fold 1 = 0. 98842014, Fold2 = 0. 99987275, Fold3 = 0. 99990269, Fold4=0.99999251, Fold5=1.0

Mean Accuracy: 0. 99775439014481O5

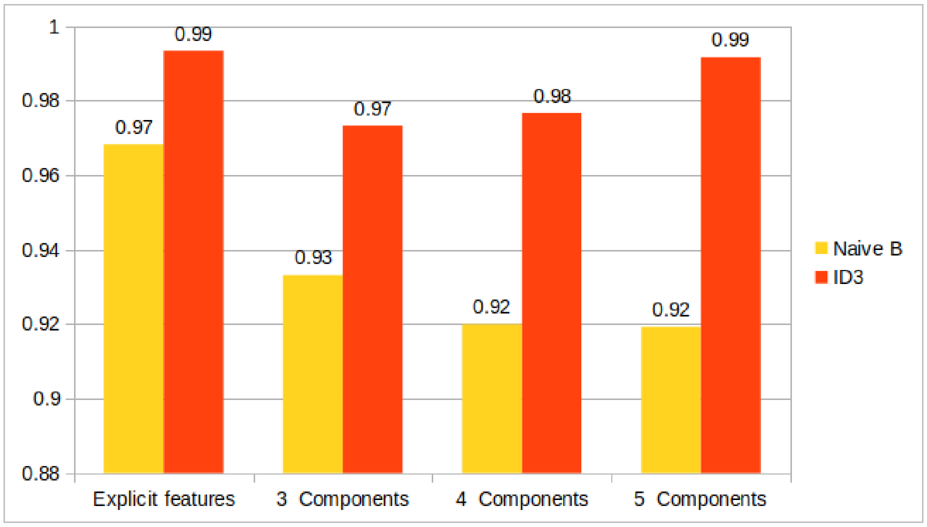


Part II: Embedding Features

Feature embedding is one of solutions for overfitting problem of a model. Overfitting is modelling error which occurs when a function is too closely fit to a limited set of data points. Also embedding is done to reduce the features space. By doing so, we can execute the same model with approx. same performance having less features.

PCA

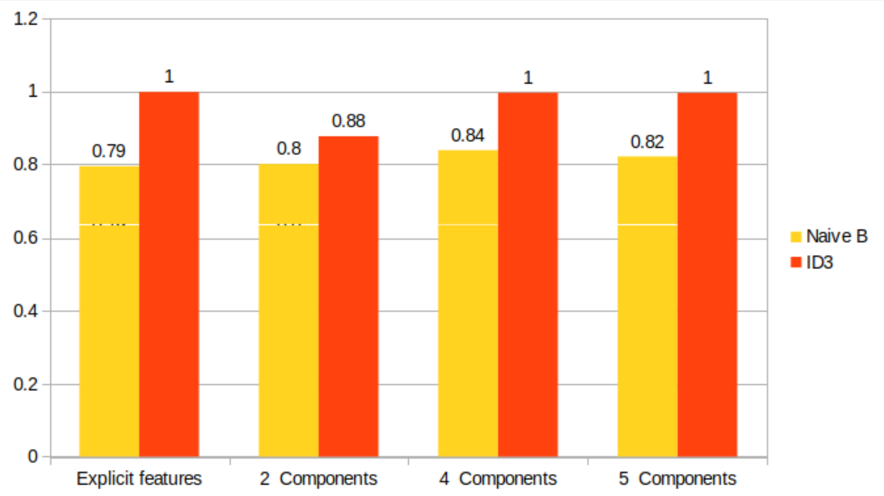
Dataset: Foursquare Restaurant Review Dataset.



Accuracy Vs. different no of reduced features of PCA (FOURSQUARE)

As seen above, Accuracy on Y-Axis and features on X-Axis. It is clearly seen that embedding or feature reduction can be applied on original feature space. Although PCA use all features to reduce all into less dimensional feature space. In Naive Bayes with minute change in accuracy its model completed task in less time comparatively. So as in ID3/decision tree. The given dataset was of Foursquare restaurant review.

Dataset: Blog Catalog data

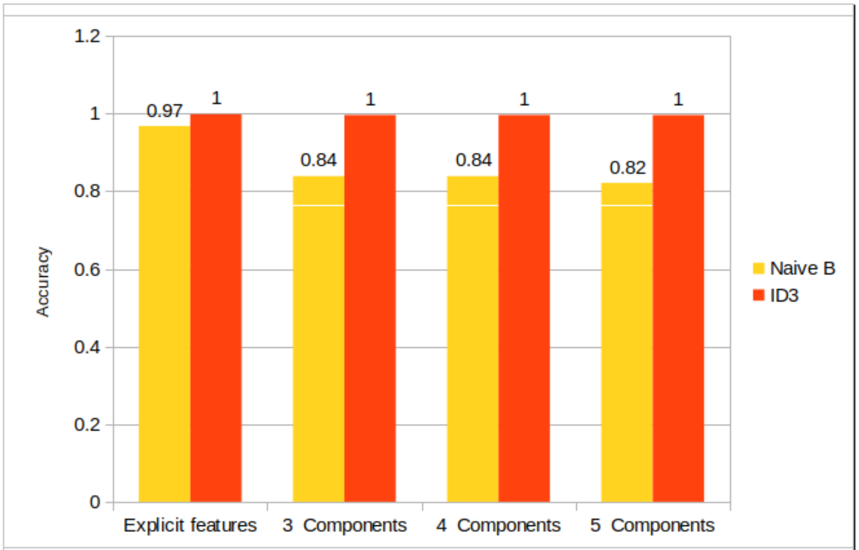


Accuracy Vs. different no of reduced features of PCA (BLOG)

As seen above, Accuracy on Y-Axis and features on X-Axis. It is clearly seen that embedding or feature reduction can be applied on original feature space. Although PCA use all features to reduce all into less dimensional feature space. In Naive Bayes with minute change in accuracy its model completed task in less time comparatively. So as in ID3/decision tree. The given dataset was of Blog catalog dataset.

SVD

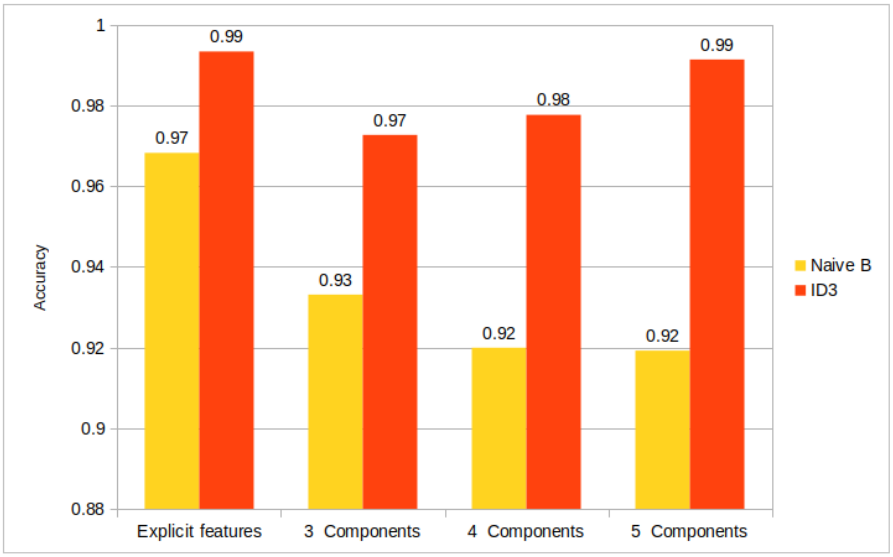
Dataset: Blog Catalog data



Accuracy Vs. different no of reduced features of SVD (BLOG)

As seen above, Accuracy on Y-Axis and features on X-Axis. It is clearly seen that embedding or feature reduction can be applied on original feature space. Although SVD use all features to reduce all into less dimensional feature space. In Naive Bayes with minute change in accuracy its model completed task in less time comparatively. So as in ID3/decision tree. The given dataset was of Blog catalog dataset.

Dataset: Foursquare Restaurant Review Dataset.



Accuracy Vs. different no of reduced features of SVD (FOURSQUARE)

As seen above, Accuracy on Y-Axis and features on X-Axis. It is clearly seen that embedding or feature reduction can be applied on original feature space. Although SVD use all features to reduce all into less dimensional feature space. In Naive Bayes with minute change in accuracy its model completed task in less time comparatively. So as in ID3/decision tree. The given dataset was of Foursquare restaurant review.

Deepwalk

DECISION TREE

Results, using embeddings of dimensionality 64

Train percent: 0.2

micro: 0.11637480724142461, macro: 0.08283903269480884

NAIVE BAYES

Results, using embeddings of dimensionality 64

Train percent: 0.2

micro: 0.33620449970549005, macro: 0.22687296932567674

Theme C: Network Destruction

In this part of assignment, we have tasked give an algorithm to destruct the given networks using various topological link prediction methods and compare which is more efficient for the purpose.

1. Basic principle for purpose of network destruction:
2. Find most important nodes/link within a network.
3. Delete that node/link and delete all other node adjacent to it.
4. Repeat above steps until size of giant cluster decreases.
5. Print the number of edges deleted.
6. Repeat all steps until network vanished.
7. Select the criteria for which dismantle is fastest with the least number of edges deleted.
8. PR gives the best result as can be seen from the below observation.

We implied this procedure with different topological methods and the results are given below:

Network: Foursquare Restaurant Review Dataset.

	CN	***AA***	***JC***	***Katz***	***ND***	***PA***	***RA***	***TN***	***PR***
1st itr	3790	4080	16850	3790	10100	33420	6730	720	20
2nd itr	3920	6840	25710	3920	13890	43630	8630	1260	160
3rd itr	3950	13470	27540	3950	15470	45160	9380	3390	230
4th itr	6740	14000	29180	6740	15720	46520	10320	3980	280
5th itr	13610	14530	30300	13610	16010	46540	17480	5530	300

Network: Blog Catalog Dataset.

	CN	***AA***	***JC***	***Katz***	***ND***	***PA***	***RA***	***TN***	***PR***
1st itr	256300	260100	218600	256300	182300	262200	161300	4200	200
2nd itr	257800	263200	220900	257800	217300	262500	163700	4300	900
3rd itr	258300	263400	224000	258300	230600	262800	193900	4400	3200
4th itr	258500	263600	224300	258500	233200	265800	194300	7800	3800
5th itr	260500	264500	225000	260500	236200	266900	228400	8000	5200

The above results are in tabular form which shows number of iterations in which size of giant cluster decreased comparatively. As shown above we can clearly see that Rooted PageRank gives us the most efficient program for network destruction among all other features.

Rooted PageRank gives best results for our algorithm.

Abbreviations Used:

CN: Common Neighbour

AA: Adamic Adar

JC: Jaccard Coefficient

ND: Neighbour Distance

PA : Preferential Attachment

RA: Resource Allocation

TN: Total Neighbour

PR: Page Rank

PCA: Principle Component Analysis

SVM: Support Vector Machine

SVD: Single Vector Decomposition

ROC: Receiver Operating characteristic Curve

AUC: Area Under Curve

P-R: Precision Recall

FPR: False Positive Rate

TPR: True Positive Rate

CV: Cross Validation