

Identify Fraud from the Enron Scandal with Machine Learning

Appendix: Showing Results applying PCA with and without Feature Scaling

Simulation Run using PCA with Feature Scaling

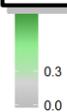
Simulation Run using PCA with Feature Scaling

All cross-examined configurations ordered by performance

All cross-examined configurations ordered by performance

Configuration	Feature Scaling	Feature Selection Setting/Metric	F1 Score	Precision	Recall	Accuracy
KNeighbors	off	FALSE	0.464	0.449	0.481	0.852
KNeighbors	on	FALSE	0.435	0.399	0.479	0.834
KNeighbors	pca	FALSE	0.37	0.351	0.392	0.823
KNeighbors	off	TRUE	0.354	0.593	0.252	0.877
KNeighbors	on	TRUE	0.288	0.318	0.262	0.827
KNeighbors	pca	TRUE	0.242	0.342	0.188	0.844
AdaBoost'ed Decision Tree	off	FALSE	0.459	0.562	0.388	0.878
AdaBoost'ed Decision Tree	on	FALSE	0.459	0.562	0.388	0.878
AdaBoost'ed Decision Tree	on	TRUE	0.41	0.494	0.35	0.866
AdaBoost'ed Decision Tree	off	TRUE	0.41	0.494	0.35	0.866
AdaBoost'ed Decision Tree	pca	FALSE	0.292	0.366	0.244	0.843
AdaBoost'ed Decision Tree	pca	TRUE	0.244	0.315	0.199	0.835
Gaussian Naive Bayes	pca	TRUE	0.384	0.411	0.362	0.846
Gaussian Naive Bayes	pca	FALSE	0.362	0.413	0.322	0.849
Gaussian Naive Bayes	off	TRUE	0.341	0.359	0.324	0.833
Gaussian Naive Bayes	on	TRUE	0.341	0.359	0.324	0.833
Gaussian Naive Bayes	off	FALSE	0.254	0.295	0.224	0.825
Gaussian Naive Bayes	on	FALSE	0.254	0.295	0.224	0.825
Support Vector Classifier	on	TRUE	0.379	0.323	0.46	0.799
Support Vector Classifier	pca	TRUE	0.369	0.266	0.603	0.726
Support Vector Classifier	on	FALSE	0.365	0.293	0.483	0.777
Support Vector Classifier	pca	FALSE	0.279	0.478	0.198	0.864
Support Vector Classifier	off	FALSE	0.279	0.478	0.198	0.864
Support Vector Classifier	off	TRUE	0.279	0.478	0.198	0.864
Adaboost'ed RandomForest	off	FALSE	0.324	0.462	0.25	0.861
Adaboost'ed RandomForest	on	FALSE	0.324	0.462	0.25	0.861
Adaboost'ed RandomForest	pca	FALSE	0.301	0.406	0.239	0.852
Adaboost'ed RandomForest	off	TRUE	0.294	0.408	0.23	0.853
Adaboost'ed RandomForest	on	TRUE	0.294	0.408	0.23	0.853
Adaboost'ed RandomForest	pca	TRUE	0.282	0.368	0.228	0.845
LDA	off	FALSE	0.324	0.437	0.258	0.857
LDA	on	FALSE	0.324	0.437	0.258	0.857
LDA	pca	TRUE	0.292	0.495	0.207	0.866
LDA	pca	FALSE	0.266	0.519	0.179	0.868
LDA	off	TRUE	0.26	0.349	0.208	0.843
LDA	on	TRUE	0.26	0.349	0.208	0.843
Logistic Regression	on	FALSE	0.312	0.414	0.251	0.853
Logistic Regression	pca	TRUE	0.265	0.545	0.175	0.871
Logistic Regression	on	TRUE	0.236	0.365	0.174	0.85
Logistic Regression	pca	FALSE	0.207	0.455	0.134	0.863
Logistic Regression	off	FALSE	0.112	0.088	0.154	0.674
Logistic Regression	off	TRUE	0.07	0.053	0.106	0.628

Configuration
 □ no feature scaling
 ○ manual feature selection
 △ MinMax feature scaling
 □ manual feature selection
 ○ PCA + MinMax feature scaling
 △ manual feature selection
 □ no feature scaling



Feature Scaling Feature Selection Setting/Metric F1 Score Precision Recall Accuracy

Kneighbors:
w/o scaling better, Precision++, Recall-

AdaBoost'ed DT:
w/o scaling better, Precision++, Recall+

Gaussian NB:
w/o scaling better, Precision+, Recall+

Support Vector Cl.:
with scaling better, does not converge w/o scaling

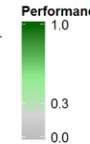
LDA:
w/o scaling better, Precision++, Recall+

Logistic Regression:
w/o scaling better, Precision-, Recall++

AdaBoost'ed RF:
w/o scaling Precision+, Recall+

Configuration	Feature Scaling	Feature Selection Setting/Metric	F1 Score	Precision	Recall	Accuracy
KNeighbors	pca	FALSE	0.487	0.666	0.384	0.892
KNeighbors	off	FALSE	0.464	0.449	0.481	0.852
KNeighbors	on	FALSE	0.435	0.399	0.479	0.834
KNeighbors	off	TRUE	0.354	0.593	0.252	0.877
KNeighbors	pca	TRUE	0.349	0.575	0.25	0.875
KNeighbors	on	TRUE	0.288	0.318	0.262	0.827
AdaBoost'ed Decision Tree	on	FALSE	0.459	0.562	0.388	0.878
AdaBoost'ed Decision Tree	off	FALSE	0.459	0.562	0.388	0.878
AdaBoost'ed Decision Tree	off	TRUE	0.41	0.494	0.35	0.866
AdaBoost'ed Decision Tree	on	TRUE	0.41	0.494	0.35	0.866
AdaBoost'ed Decision Tree	pca	FALSE	0.332	0.423	0.273	0.853
AdaBoost'ed Decision Tree	pca	TRUE	0.33	0.405	0.278	0.849
Gaussian Naive Bayes	pca	FALSE	0.416	0.451	0.387	0.855
Gaussian Naive Bayes	pca	TRUE	0.358	0.45	0.297	0.858
Gaussian Naive Bayes	on	TRUE	0.341	0.359	0.324	0.833
Gaussian Naive Bayes	off	TRUE	0.341	0.359	0.324	0.833
Gaussian Naive Bayes	off	FALSE	0.254	0.295	0.224	0.825
Gaussian Naive Bayes	on	FALSE	0.254	0.295	0.224	0.825
Support Vector Classifier	on	TRUE	0.379	0.323	0.46	0.799
Support Vector Classifier	on	FALSE	0.365	0.293	0.483	0.775
Support Vector Classifier	off	FALSE	0.365	0.293	0.483	0.775
Support Vector Classifier	pca	FALSE	0.365	0.293	0.483	0.775
Support Vector Classifier	pca	TRUE	0.365	0.293	0.483	0.775
Support Vector Classifier	off	TRUE	0.365	0.293	0.483	0.775
LDA	pca	FALSE	0.365	0.561	0.27	0.875
LDA	off	FALSE	0.324	0.437	0.258	0.857
LDA	on	FALSE	0.324	0.437	0.258	0.857
LDA	pca	TRUE	0.315	0.507	0.228	0.868
LDA	off	TRUE	0.26	0.349	0.208	0.843
LDA	on	TRUE	0.26	0.349	0.208	0.843
Logistic Regression	pca	FALSE	0.354	0.266	0.53	0.742
Logistic Regression	on	FALSE	0.312	0.414	0.251	0.853
Logistic Regression	pca	TRUE	0.239	0.199	0.3	0.746
Logistic Regression	on	TRUE	0.236	0.365	0.174	0.85
Logistic Regression	off	FALSE	0.112	0.088	0.154	0.674
Logistic Regression	off	TRUE	0.07	0.053	0.106	0.628
AdaBoost'ed RandomForest	pca	FALSE	0.325	0.427	0.262	0.855
AdaBoost'ed RandomForest	off	FALSE	0.324	0.462	0.25	0.861
AdaBoost'ed RandomForest	on	FALSE	0.324	0.462	0.25	0.861
AdaBoost'ed RandomForest	off	TRUE	0.294	0.408	0.23	0.853
AdaBoost'ed RandomForest	on	TRUE	0.294	0.408	0.23	0.853
AdaBoost'ed RandomForest	pca	TRUE	0.284	0.387	0.224	0.849

Configuration
 □ no feature scaling
 ○ manual feature selection
 △ MinMax feature scaling
 □ manual feature selection
 ○ PCA + MinMax feature scaling
 △ manual feature selection
 □ no feature scaling



Feature Scaling Feature Selection Setting/Metric F1 Score Precision Recall Accuracy

Finding: Using Feature Scaling in conjunction with PCA is only useful for Support Vector Classifiers. And also there only in a limited way, since PCA still does not improve performance significantly.