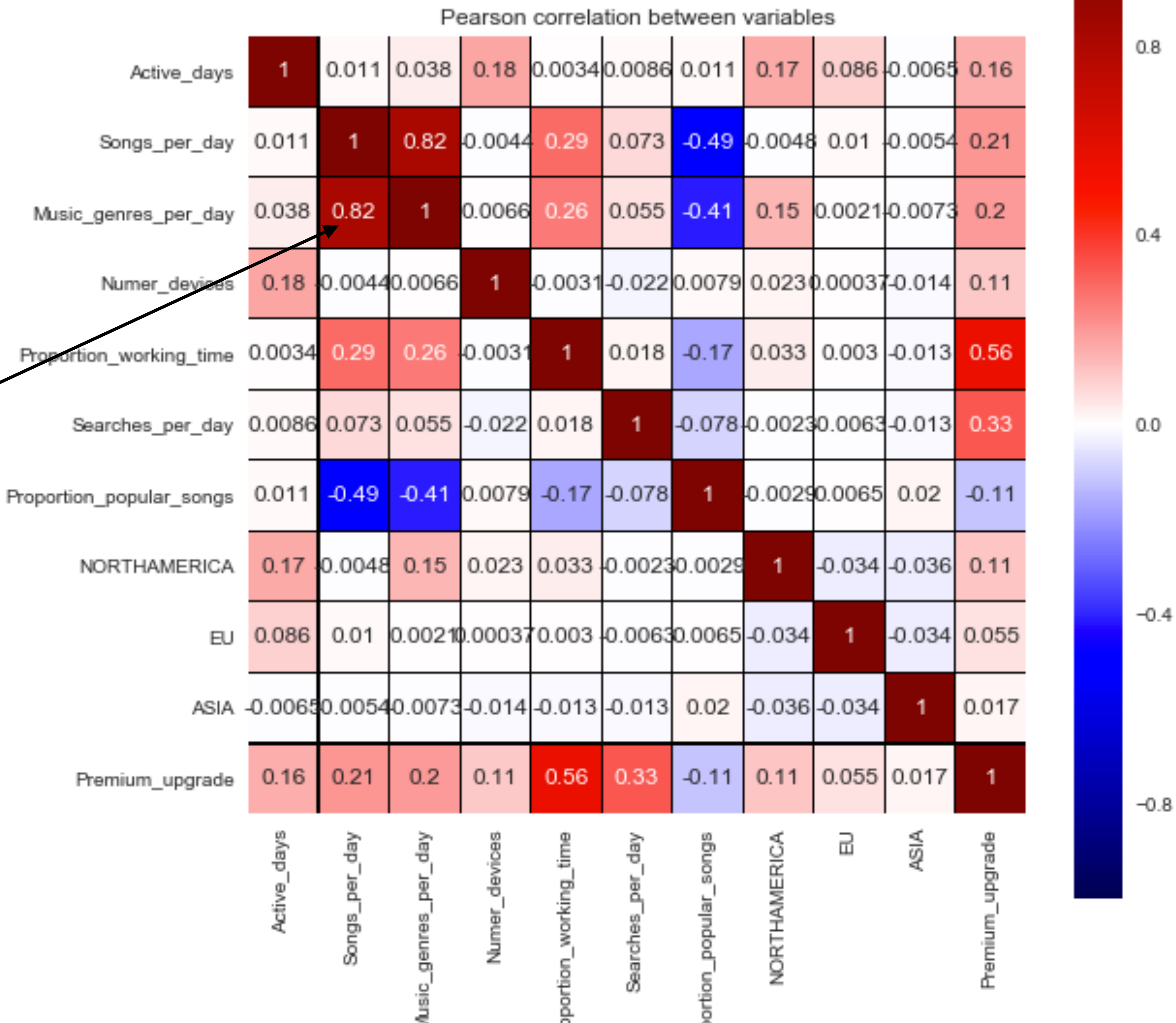
The most important variables for prediction of ”Premium upgrade” are ”Proportion working time” and ”Searches per day”



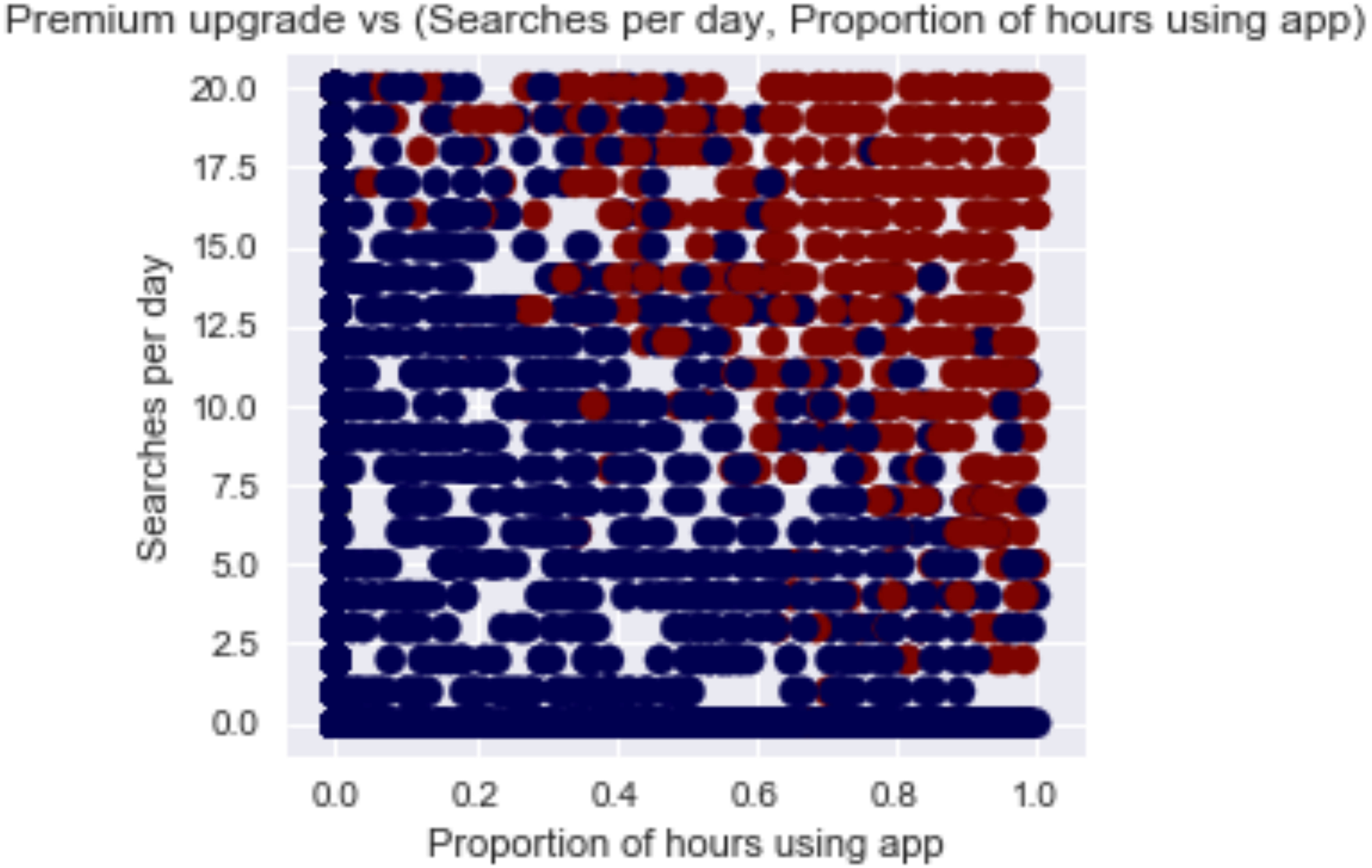
Pearson correlation between variables

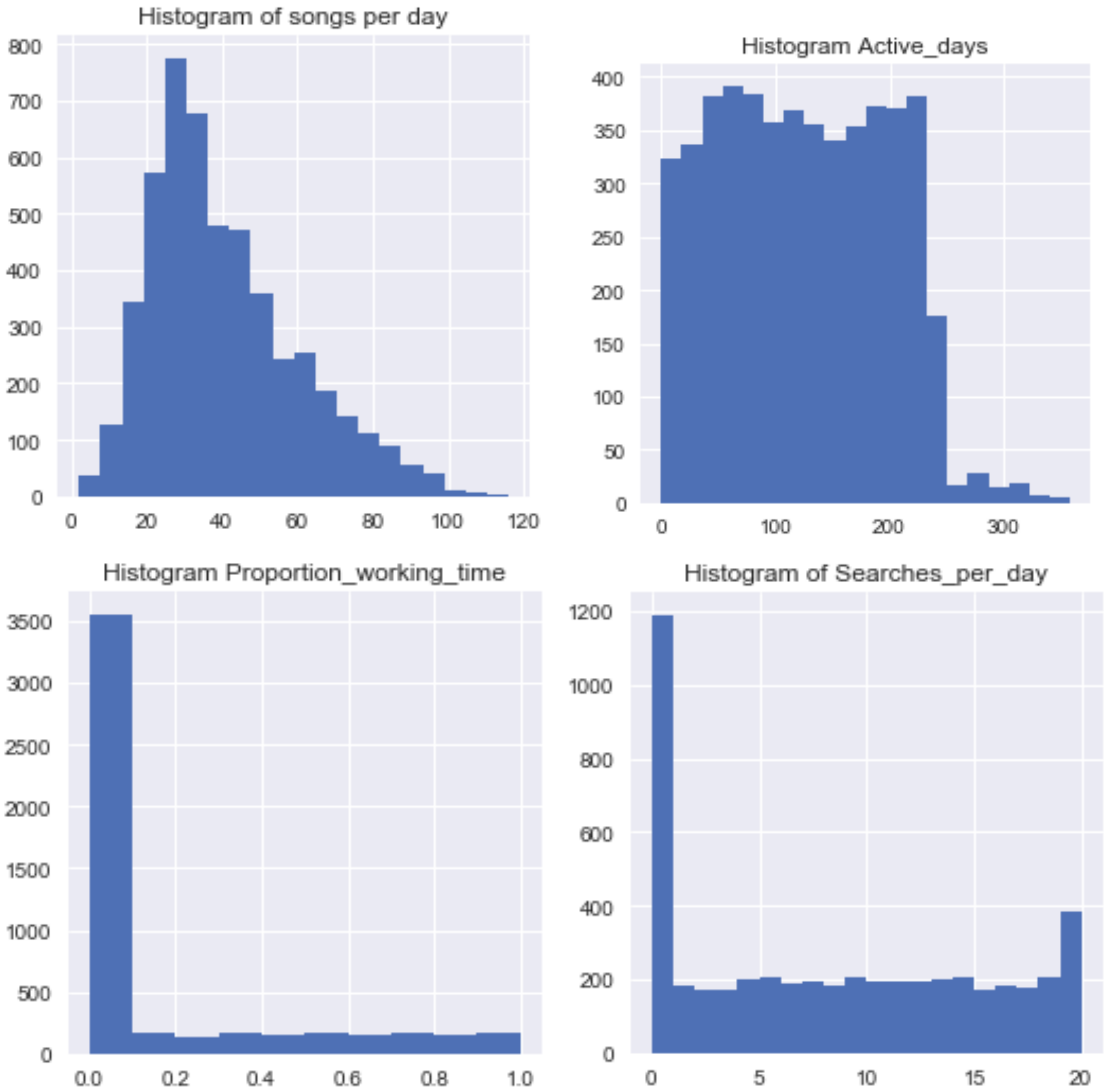
	Active_days	Songs_per_day	Music_genres_per_day	Numer_devices	Proportion_working_time	Searches_per_day	Proportion_popular_songs	NORTHAMERICA	EU	ASIA	Premium_upgrade
Active_days	1	0.011	0.038	0.18	0.0034	0.0086	0.011	0.17	0.086	-0.0065	0.16
Songs_per_day	0.011	1	0.82	-0.0044	0.29	0.073	-0.49	-0.0048	0.01	-0.0054	0.21
Music_genres_per_day	0.038	0.82	1	0.0066	0.26	0.055	-0.41	0.15	0.0021	-0.0073	0.2
Numer_devices	0.18	-0.0044	0.0066	1	-0.0031	-0.022	0.0079	0.023	0.00037	-0.014	0.11
Proportion_working_time	0.0034	0.29	0.26	-0.0031	1	0.018	-0.17	0.033	0.003	-0.013	0.56
Searches_per_day	0.0086	0.073	0.055	-0.022	0.018	1	-0.078	-0.0023	-0.0063	-0.013	0.33
Proportion_popular_songs	0.011	-0.49	-0.41	0.0079	-0.17	-0.078	1	-0.0029	0.0065	0.02	-0.11
NORTHAMERICA	0.17	-0.0048	0.15	0.023	0.033	-0.0023	-0.0029	1	-0.034	-0.036	0.11
EU	0.086	0.01	0.0021	0.00037	0.003	-0.0063	0.0065	-0.034	1	-0.034	0.055
ASIA	-0.0065	-0.0054	-0.0073	-0.014	-0.013	-0.013	0.02	-0.036	-0.034	1	0.017
Premium_upgrade	0.16	0.21	0.2	0.11	0.56	0.33	-0.11	0.11	0.055	0.017	1

extremely correlated

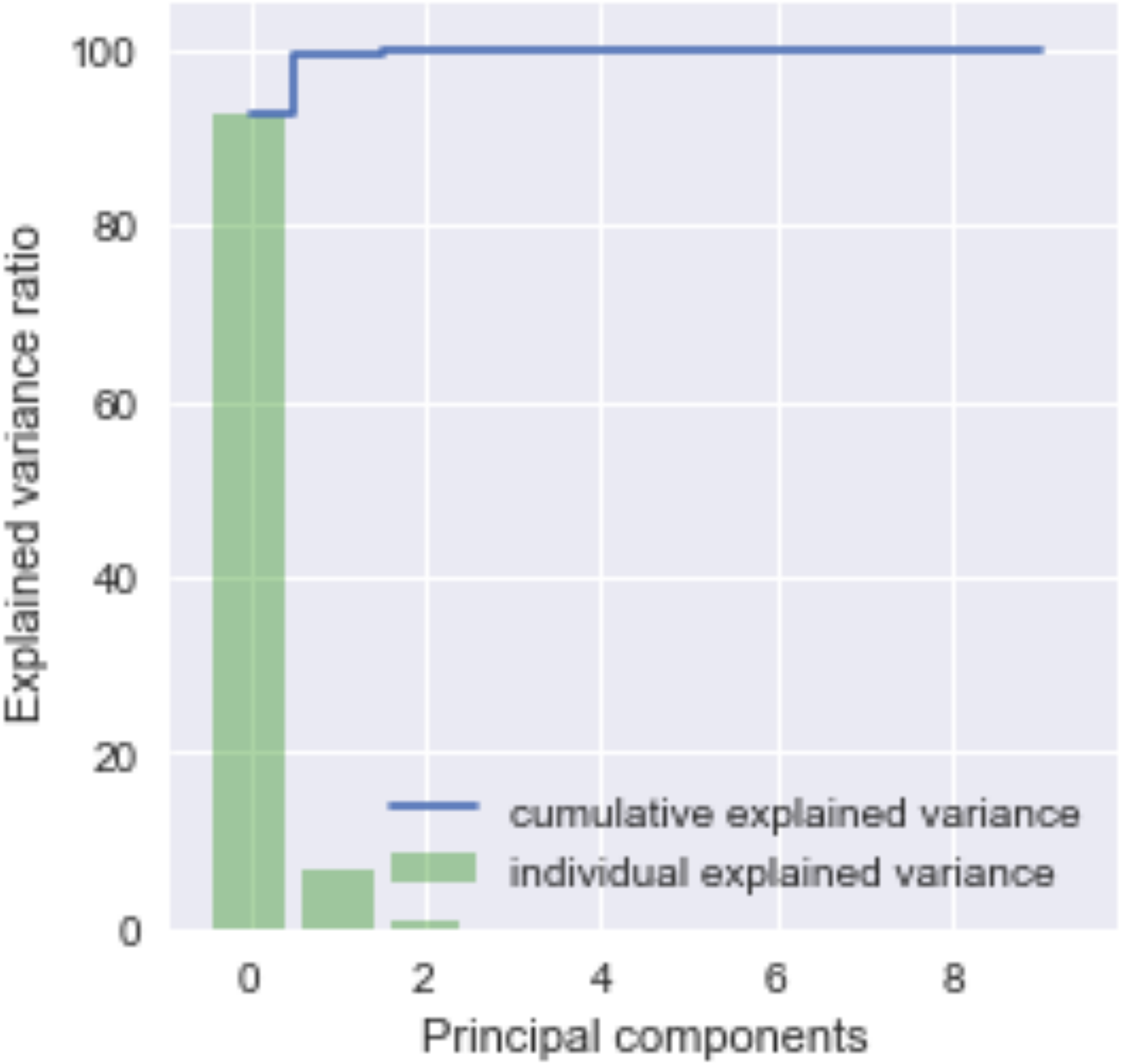
Blues → non-premium users

Reds → premium users

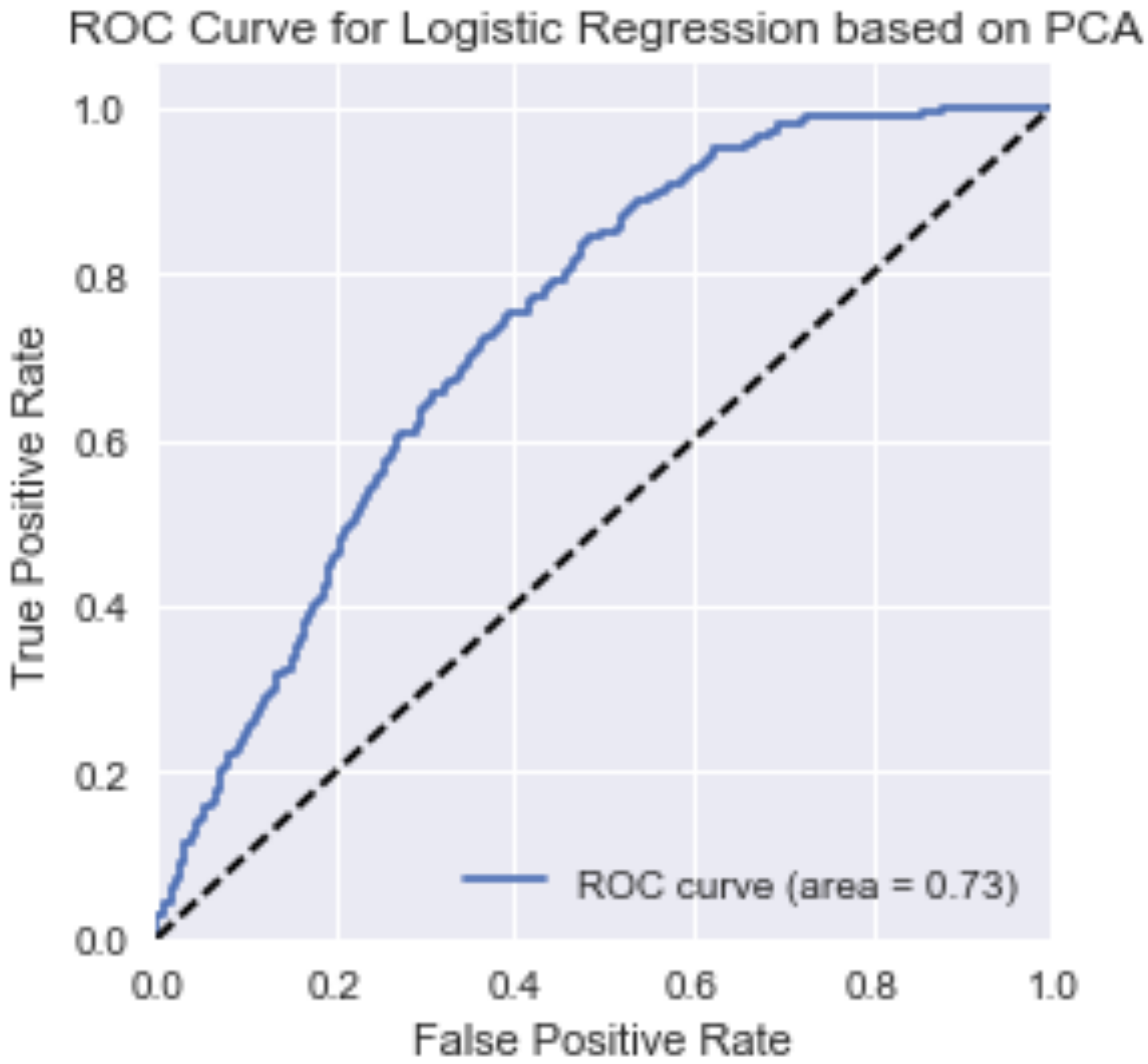


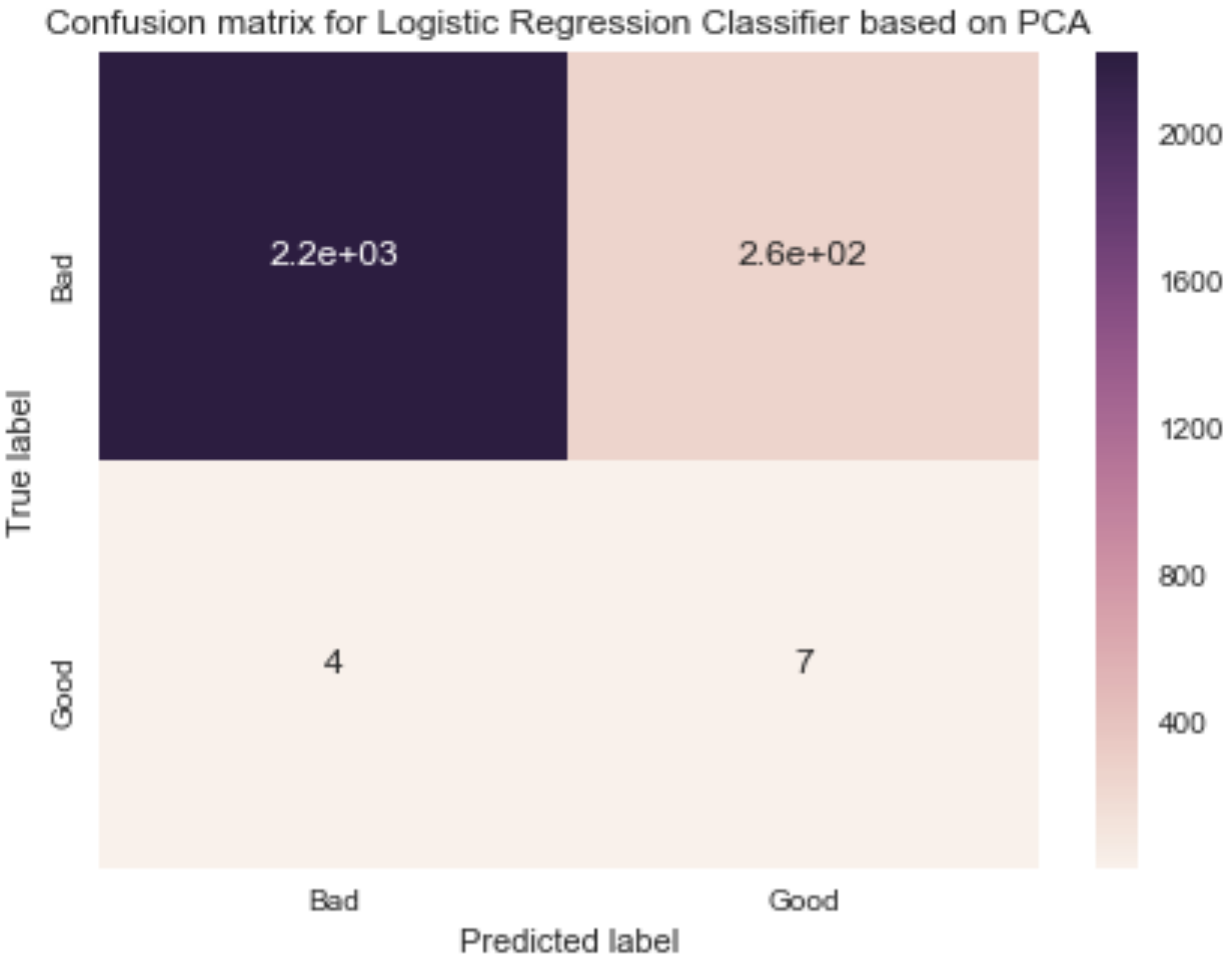

Histogram of songs per day
800
700
600
500
400
300
200
100
0
0
20
40
60
80
100
120
Histogram Active_days
400
350
300
250
200
150
100
50
0
0
100
200
300
Histogram Proportion_working_time
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Histogram of Searches_per_day
1200
1000
800
600
400
200
0
0
5
10
15
20

First two principal components explain more than 99% of the variance

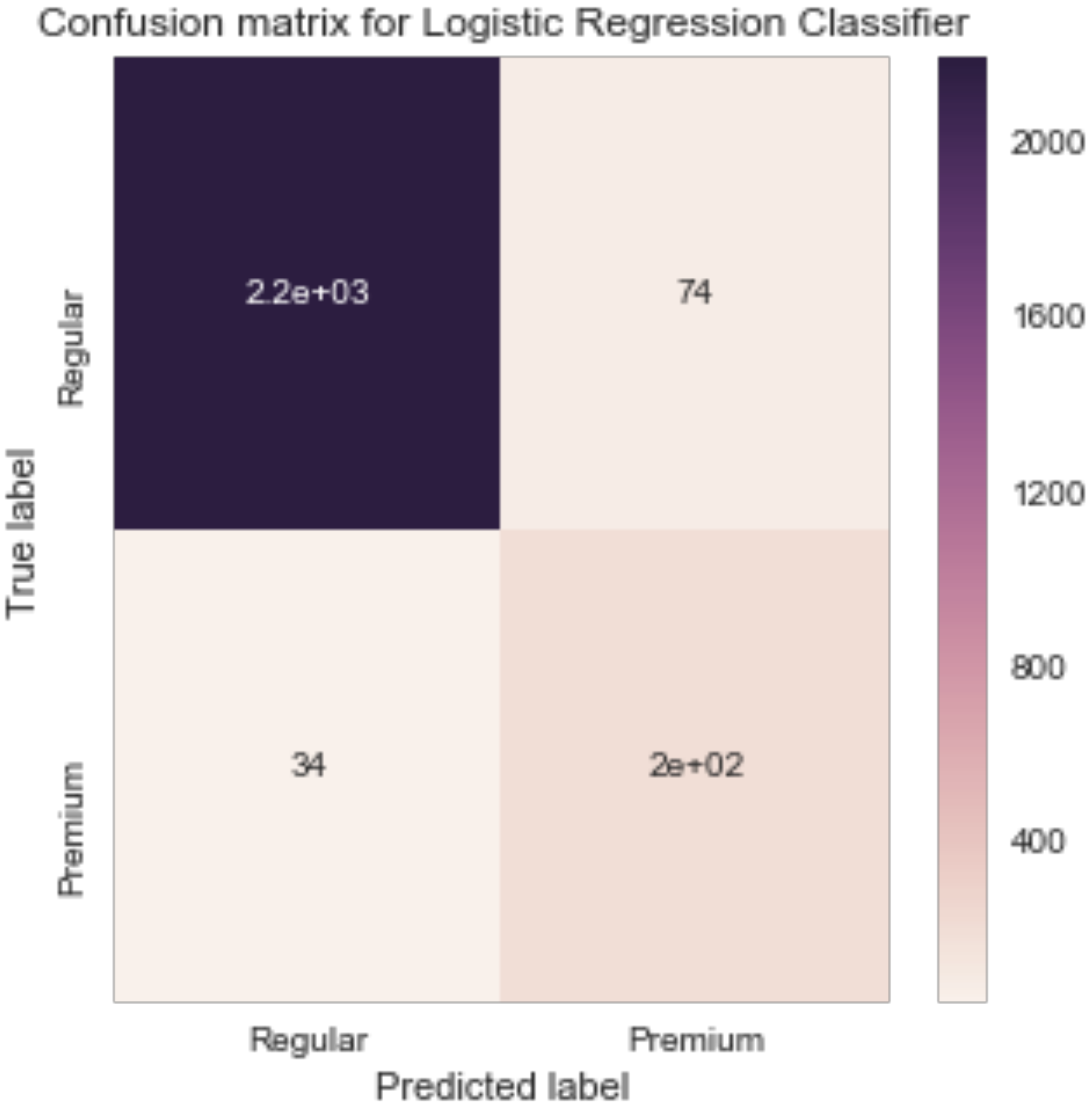


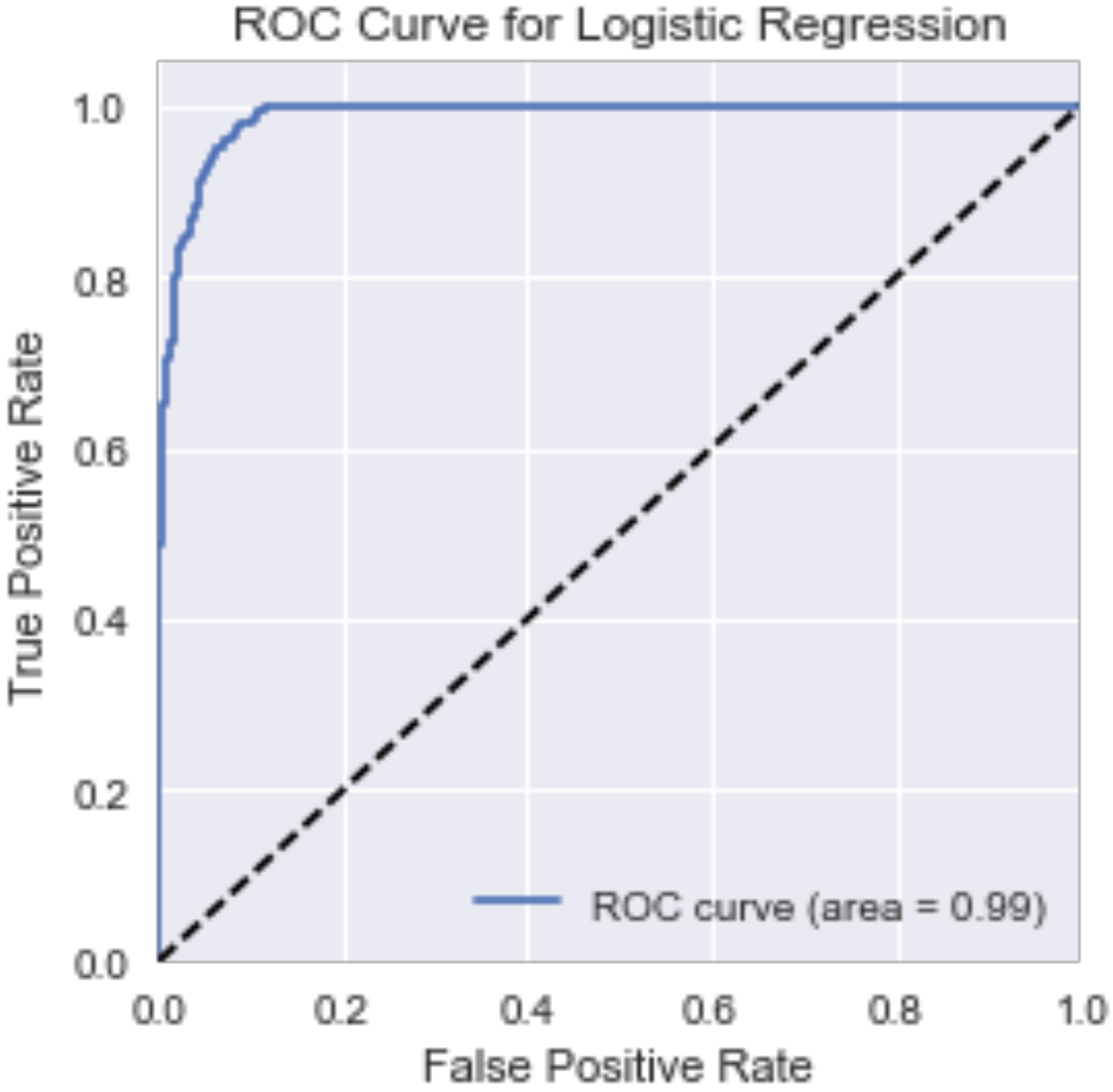
Not very good predictor for Logistic Regression





Logistic Regression works much better using all original variables



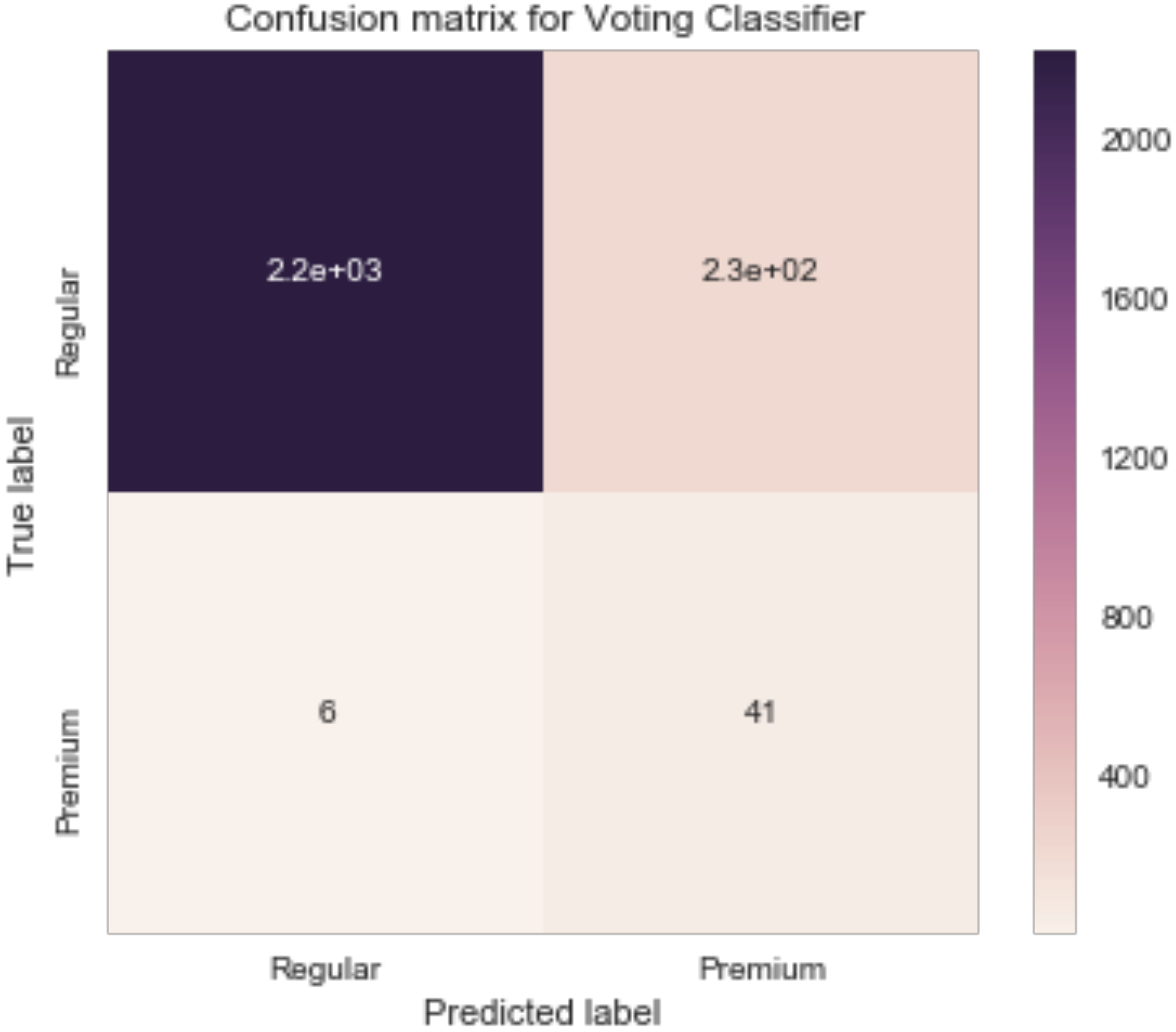


accuracy score = 0.957079823506

precision score = 0.728624535316

recall score = 0.852173913043

kNN+ SVM + Decision Tree



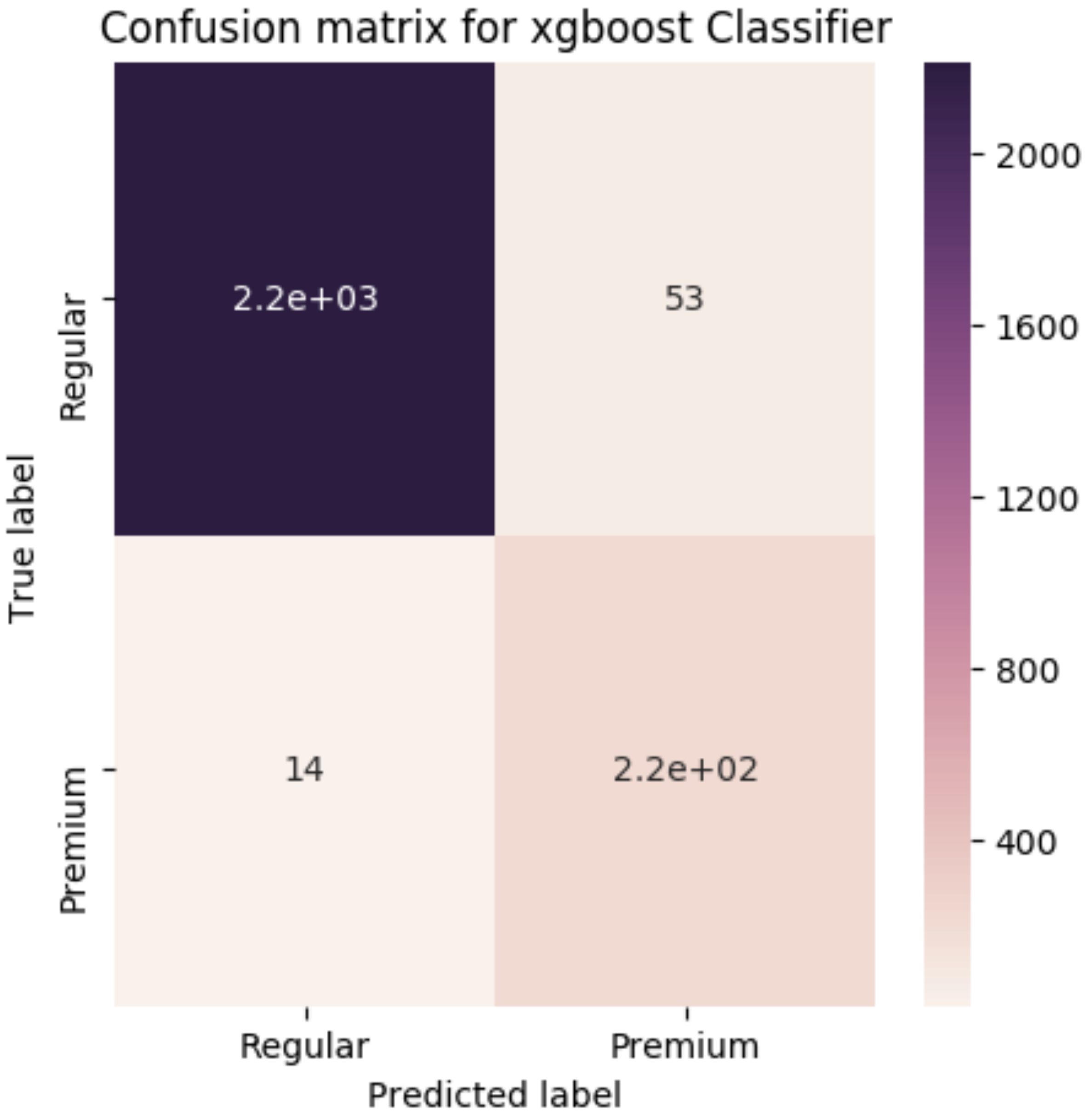
accuracy score = 0.906137184116

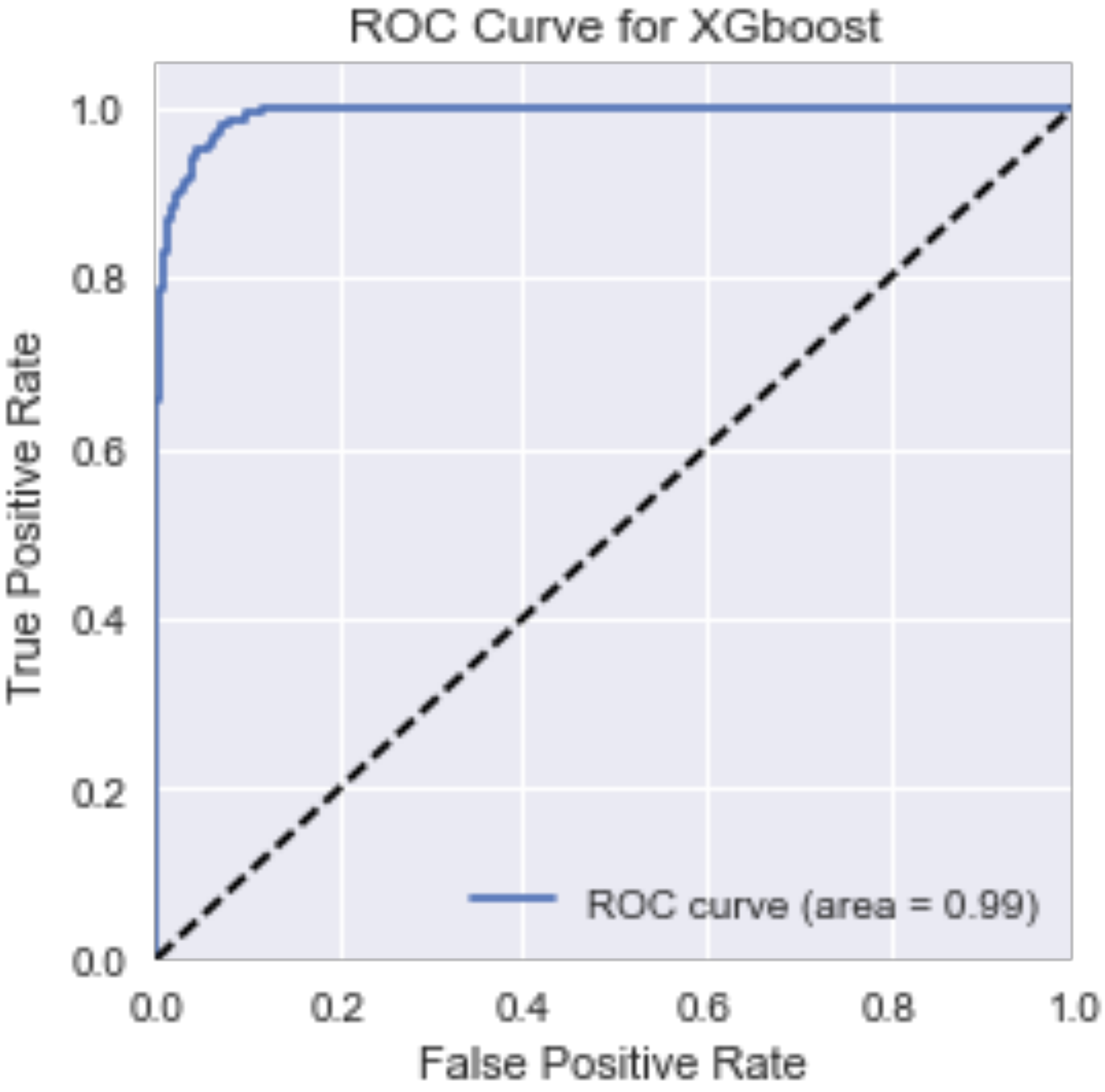
precision score = 0.152416356877

recall score = 0.872340425532

cross_val_score = 0.905173909564

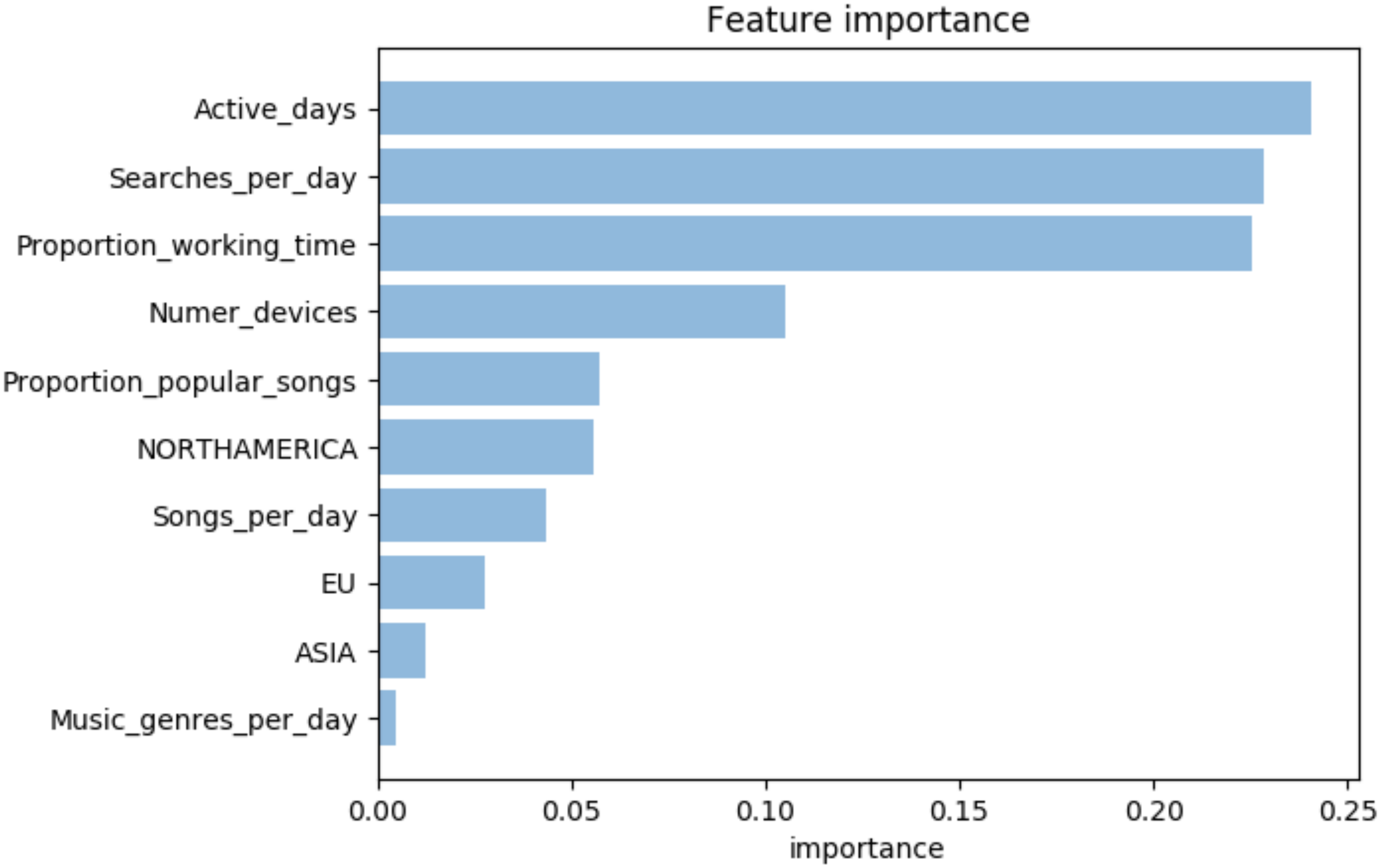
Best algorithm is xgboost


Confusion matrix for xgboost Classifier

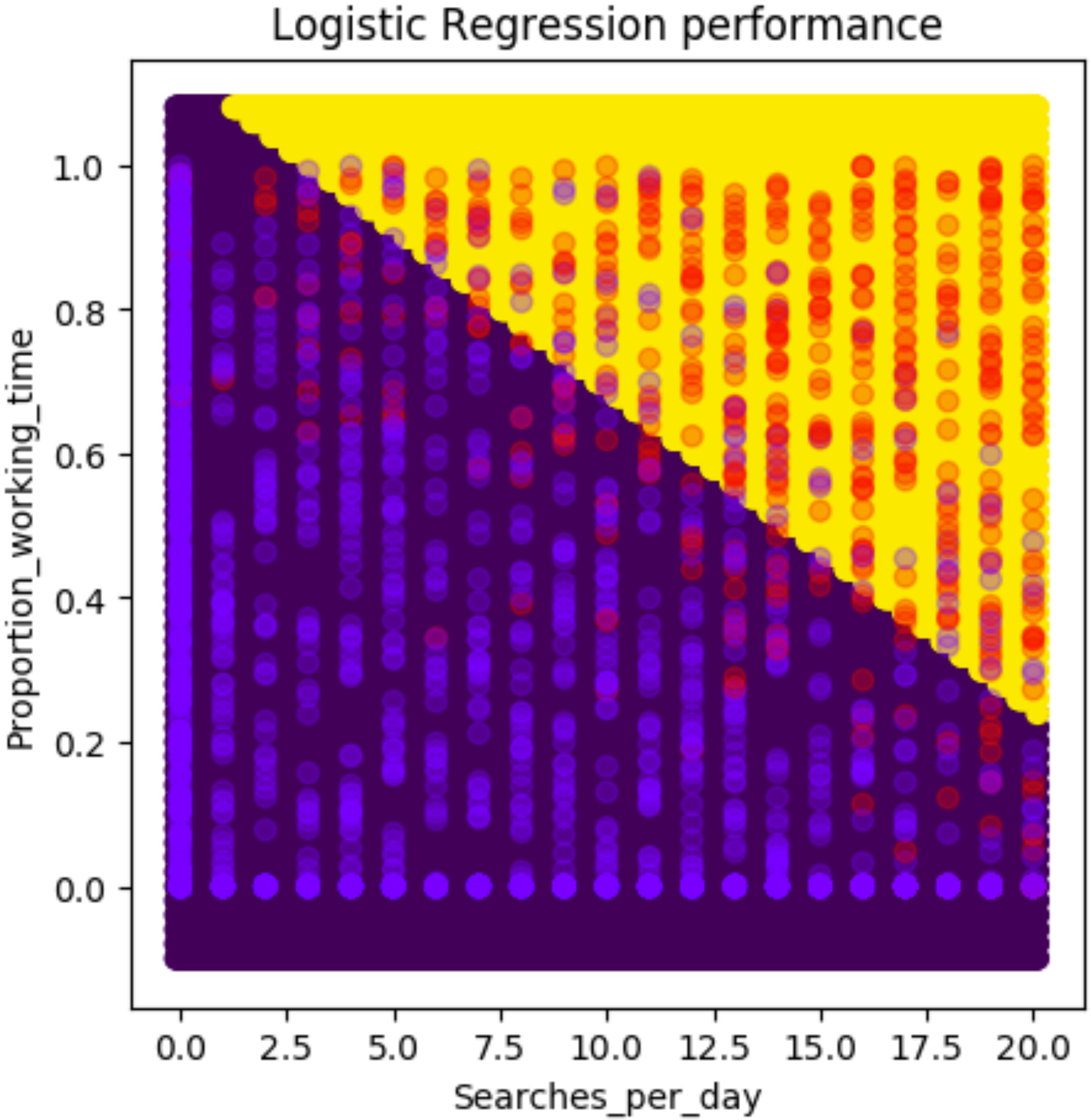

ROC Curve for XGboost

accuracy score = 0.973926995588
precision score = 0.810408921933
recall score = 0.939655172414
cross_val_score = 0.97321392557

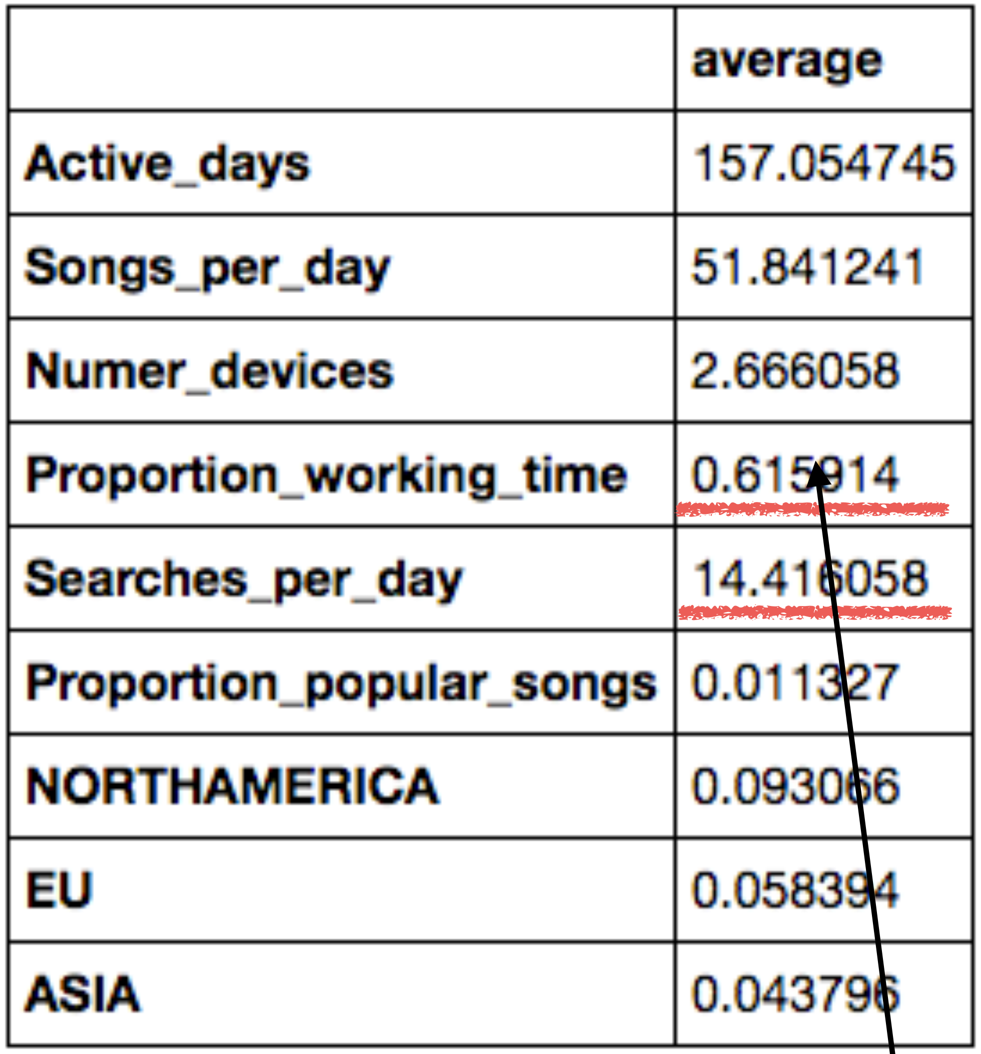
xgboost also gives us importance of each feature



Visualizing how Logit classifier works with most important two-variables



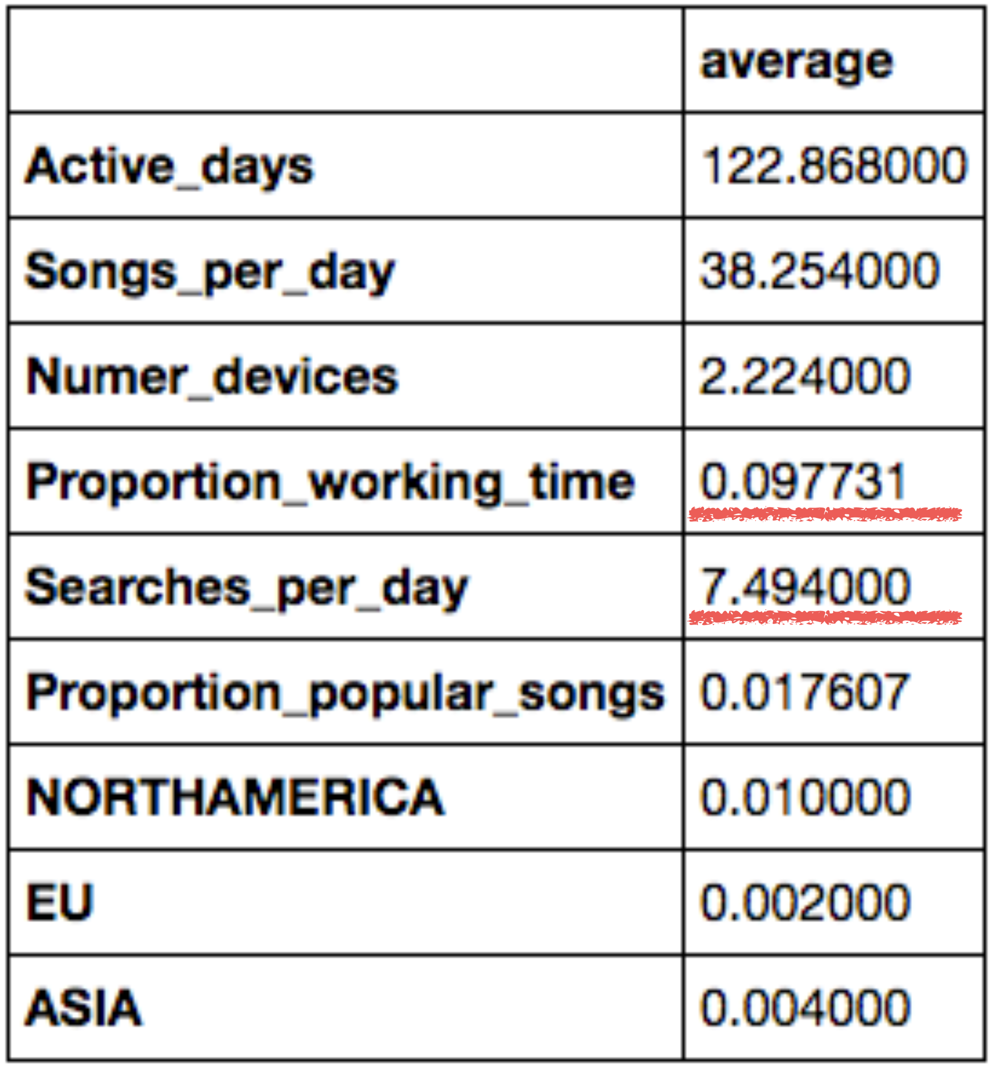
Premium users



	average
Active_days	157.054745
Songs_per_day	51.841241
Numer_devices	2.666058
Proportion_working_time	0.615914
Searches_per_day	14.416058
Proportion_popular_songs	0.011327
NORTHAMERICA	0.093066
EU	0.058394
ASIA	0.043796

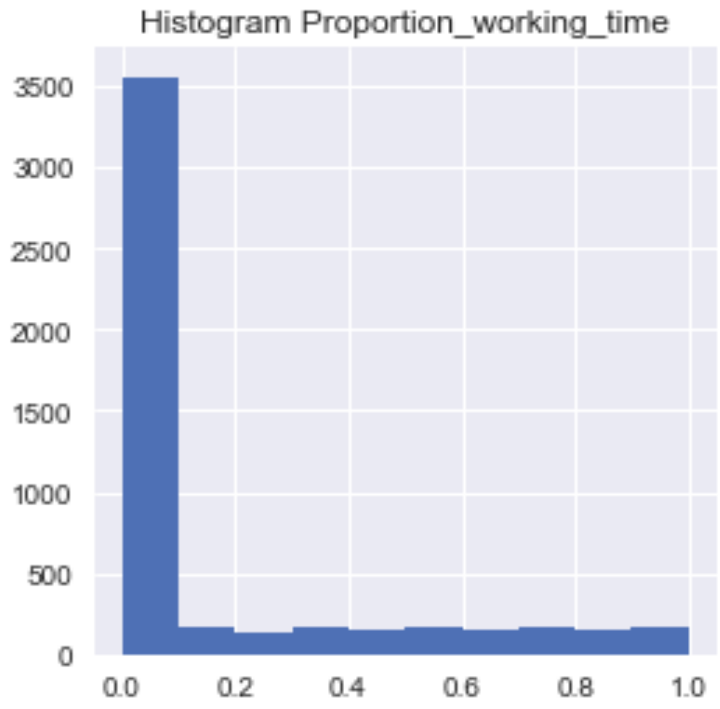
What kind of jobs do this people work at?. I would like more info on their workplace and job-status

Candidates for upgrade



	average
Active_days	122.868000
Songs_per_day	38.254000
Numer_devices	2.224000
Proportion_working_time	0.097731
Searches_per_day	7.494000
Proportion_popular_songs	0.017607
NORTHAMERICA	0.010000
EU	0.002000
ASIA	0.004000

selected from highest probabilities


Histogram Proportion_working_time