Hackerrank Predict Email Opens

Train data:

1. Correlation between mail category and “opened”
2. Correlation between recent activities and “opened”
3. Correlation between last login and “opened”
4. Same users appearing => group by users?

Regardless of columns I chose, most of them show 50~60% correlation between each column and “opened” column. It really depends on users

Data Processing Steps:

1. Convert “TimeZone” by dividing each value by 3600 (hour level)
2. Create “Forum” column by adding “forum_comments_count”, “forum_count”, “forum_expert_count” and “forum_questions_count”
3. I tried other columns for “submission”, “ipn”, and “weekday” (when email was sent), but they did not have strong enough correlation
4. Ignore all rows with invalid “TimeZone” and “Forum”
5. Y-value is “opened” column

Model:

1. “RandomForestClassifier” was used over “logisticRegression” based on the nature of data set (average of sub-samples were important factor for this problem)