News Popularity Production

Vinay Singh, 201402035
Ayush Joshi, 201402017
Ashutosh Ranjan, 201425031

Overview

Our project aims to develop an effective learning algorithm to predict how popular an online article (news or story) would be before its publication by analyzing several statistic characteristics extracted from it.

Project objective:
Classify articles in different classes based on how many shares (how popular) they can get.

Procedure (1/2)

1

39,000 articles taken from UCI Machine Learning Repository. Contained 58 features at first and the final no of shares associated with each of the articles. The key features can be grouped into words, links, digital media, time, keywords etc.

2

We trimmed the no. of features from 58 to 28 manually based on what we felt was important enough.

3

We extract 8 features from those 28 features using Feature selection algorithm BIC.

Best Subset Selection

Step 1

In a set of features F = {1, 2, 3, 4, ..., p} make all possible subsets of i features. Generate RSS for all of those subsets with i features and pick the one with **minimum RSS.** This gives us $F^{(i)} = \{p_1^{(i)}, p_2^{(i)}, p_3^{(i)}, p_4^{(i)}, ..., p_i^{(i)}\}$ for i = 1 to n..

Residual Sum of Squares (RSS) refers to the following formula:

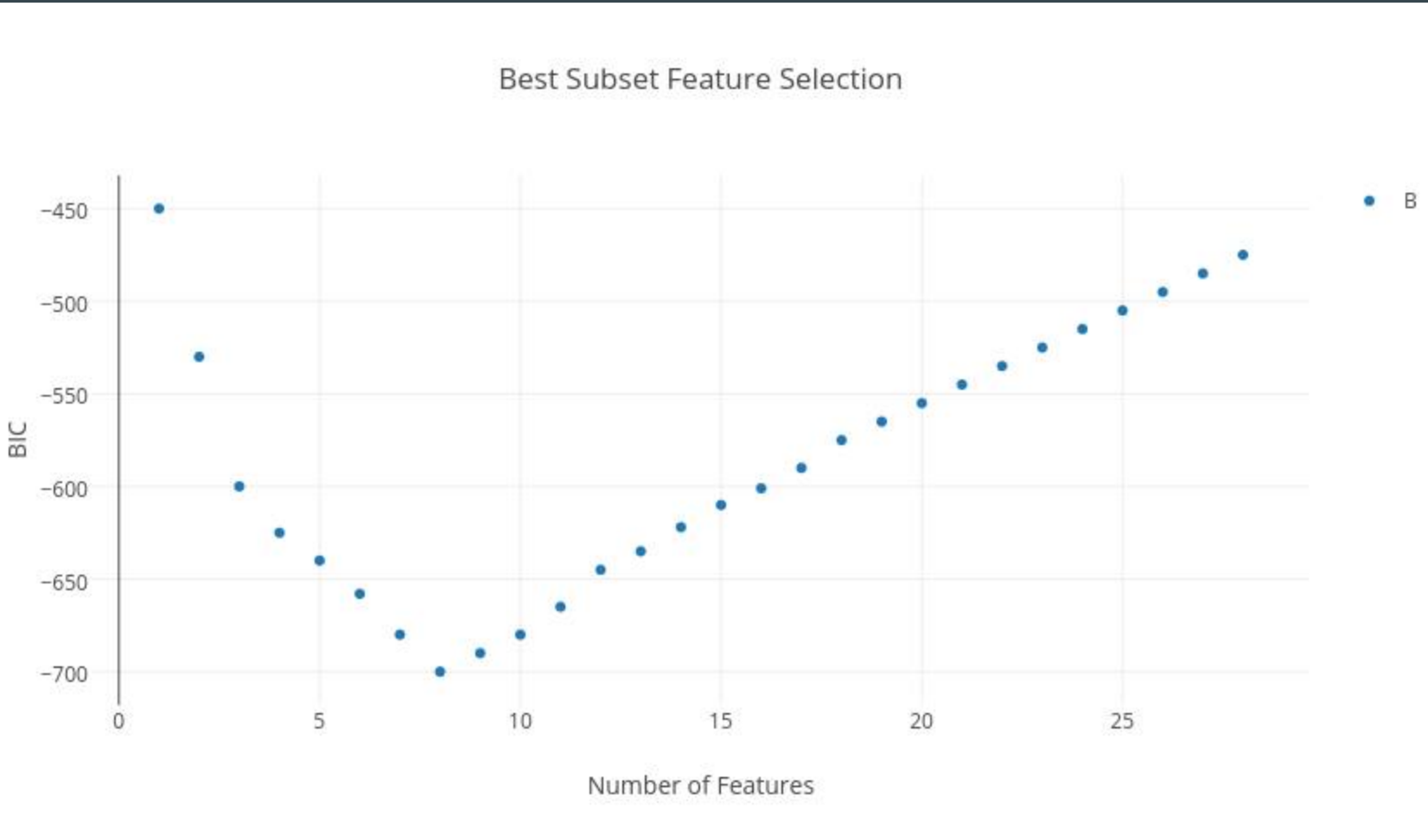
$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Best Subset Selection

Step 2

Then we compare all $F^{(1)}$, $F^{(2)}$, $F^{(3)}$, ..., $F^{(p)}$ and find the subset with the minimum BIC (Bayesian Information Criterion) which is

$$BIC = \frac{1}{n}(RSS + \log(n)d\sigma^2)$$


Best Subset Feature Selection
BIC
−450
−500
−550
−600
−650
−700
0
5
10
15
20
25
Number of Features
B

Best 8 features selected

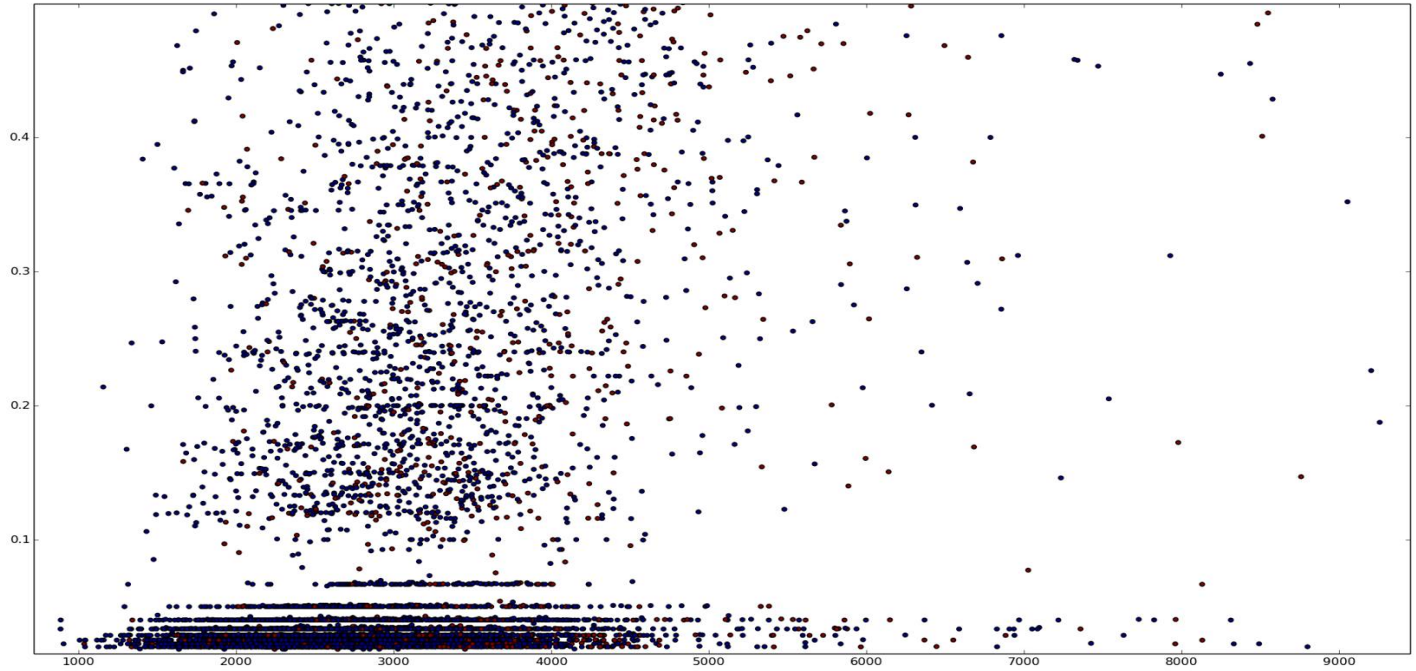
They are:

- No of links to other articles
- Article channel is entertainment
- Min shares of worst keyword
- Min shares of average keyword
- Max shares of average keyword
- Average shares of average keyword
- Min shares of referenced articles
- Closeness to top 3 LDA topic

Using cross validation technique to find correlation of individual features and final labels

```
>>> corr = df.corr()['target'][df.corr()['target'] < 1].abs()
>>> corr.sort(ascending=False)
>>> corr.head()
5    0.120266
7    0.100426
4    0.061993
3    0.055722
6    0.039791
Name: target, dtype: float64
```

Feature A vs Feature B



Procedure (2/2)

4

Out of 39000 entries, we have selected 27000 as training data and the rest 12000 as test data.

The training data has been selected randomly so as to contain as many different types of entries as possible.

5

Types of classification done:

- 2-class classification (High and Low)
- 3-class classification (High, Moderate and Low)

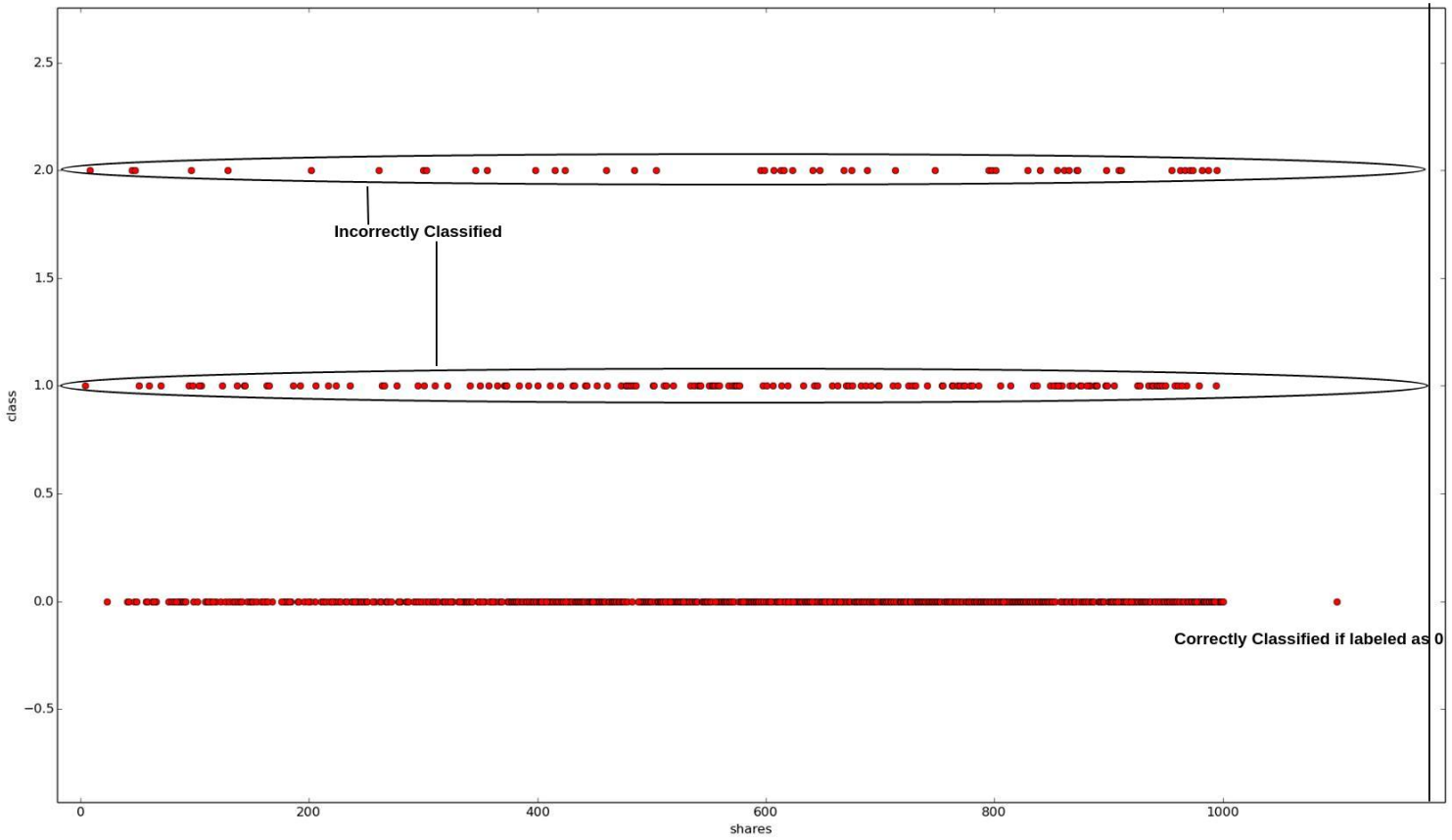
6

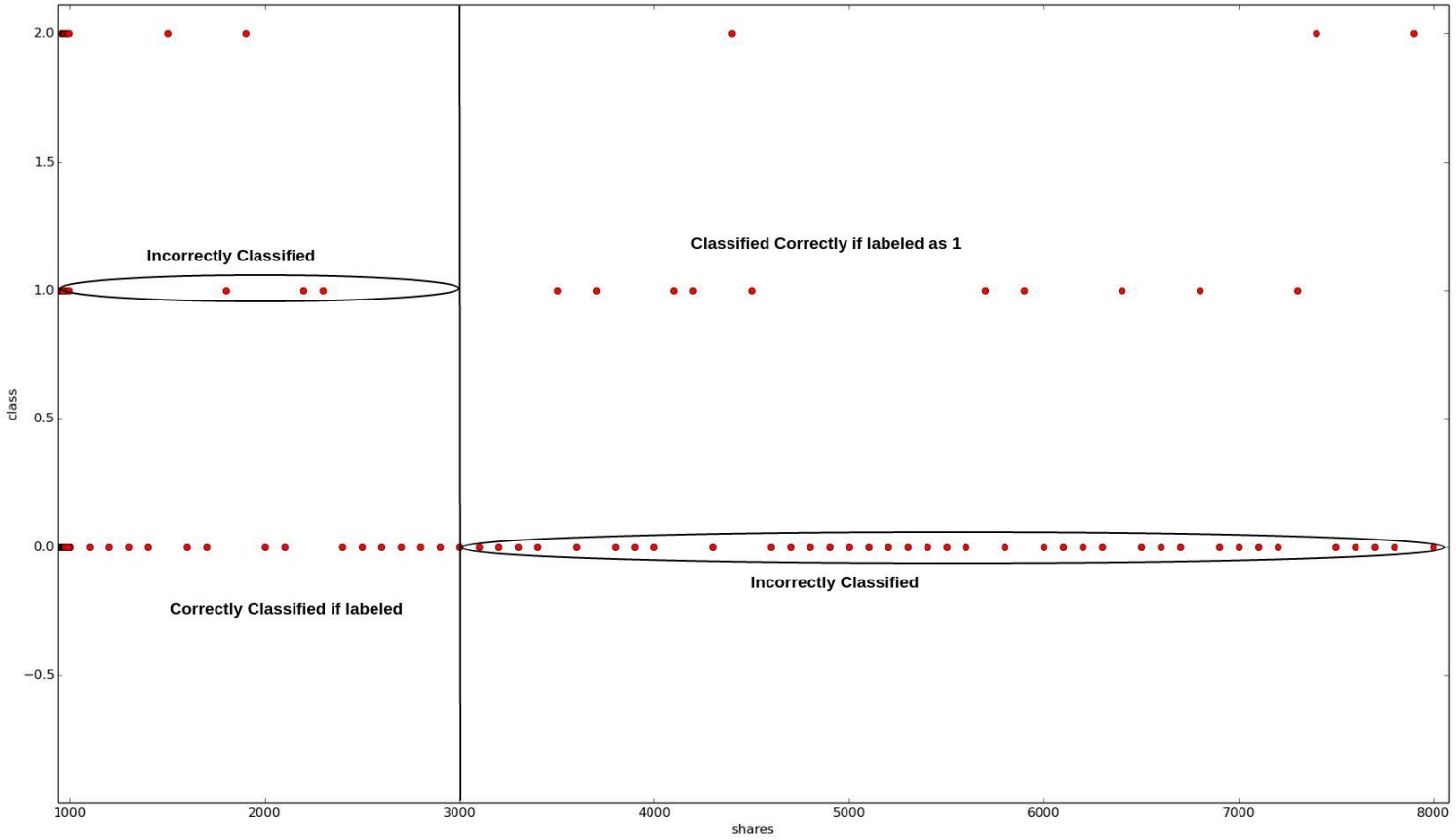
Algorithms used:

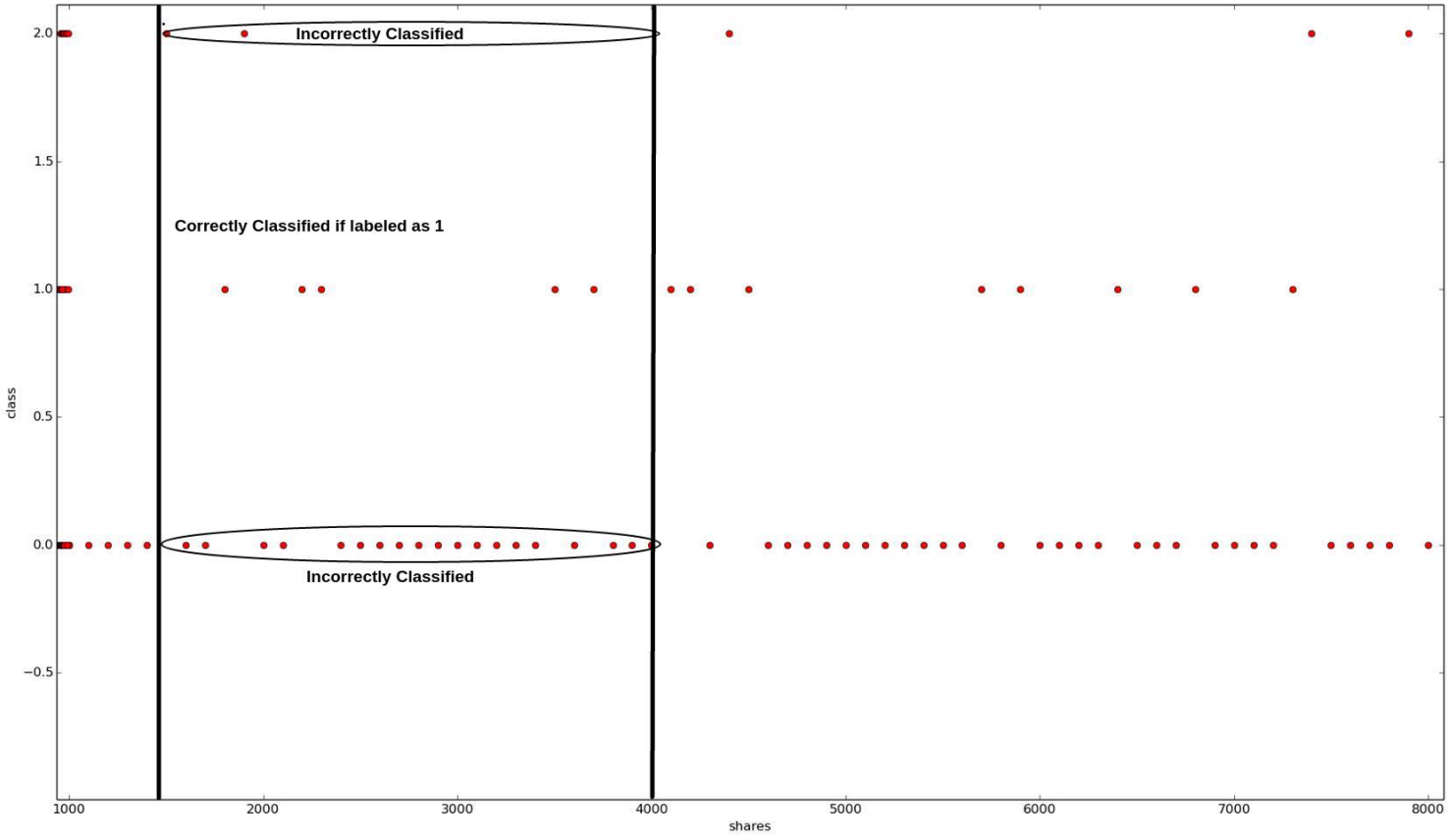
- Kernel SVM (RBF)
- Random Forest

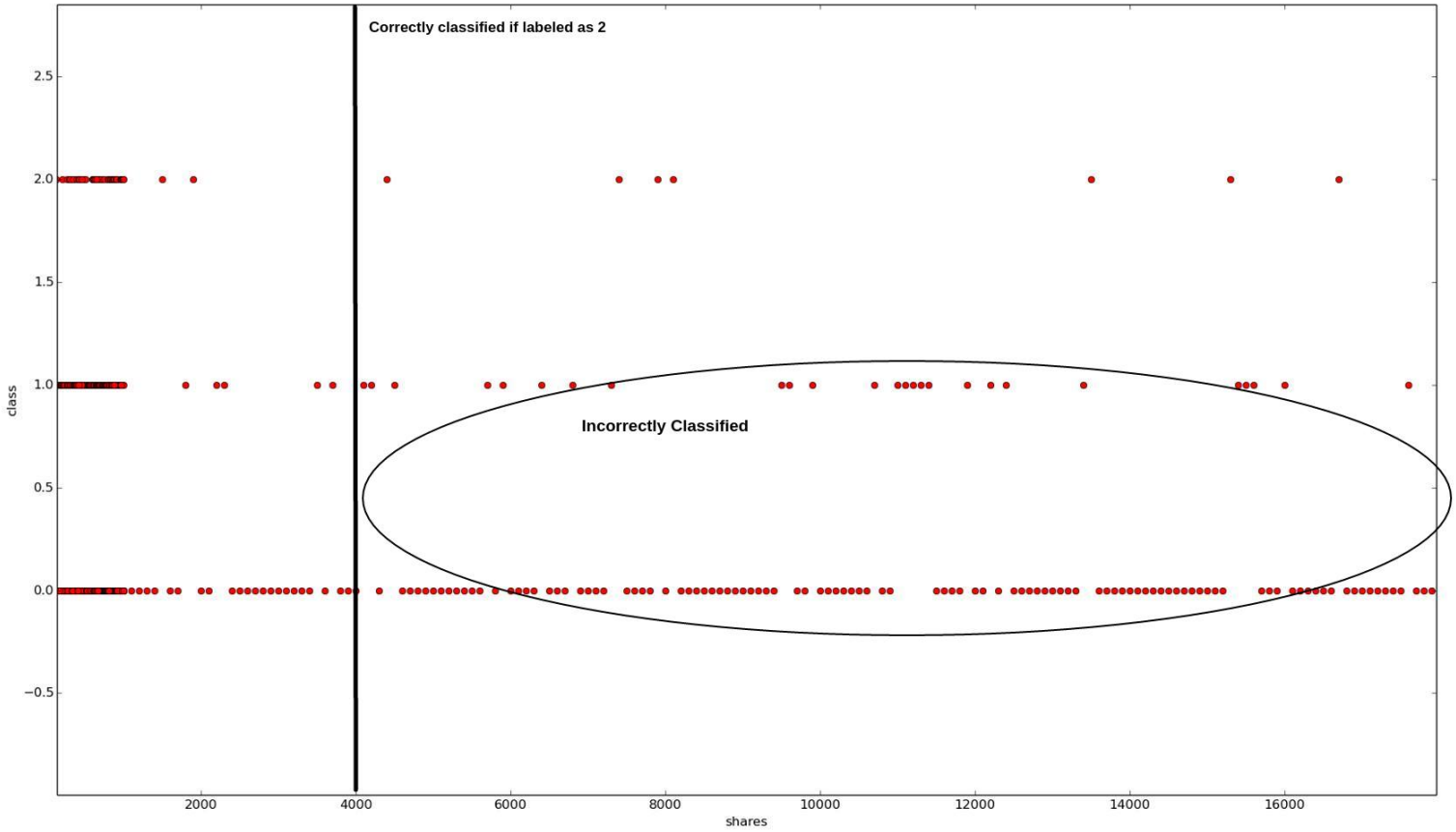
Results

2-class classification	• Kernel SVM • Accuracy: 76-78%
2-class classification	• Random Forest • Accuracy: 78-80%
3-class classification	• Kernel SVM • Accuracy: 53-55%
3-class classification	• Random Forest • Accuracy: 54-56%


Incorrectly Classified
Correctly Classified if labeled as 0
class
shares
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000


Incorrectly Classified
Classified Correctly if labeled as 1
Incorrectly Classified
Correctly Classified if labeled
class
shares
2.0
1.5
1.0
0.5
0.0
−0.5
1000
2000
3000
4000
5000
6000
7000
8000


Incorrectly Classified
Correctly Classified if labeled as 1
Incorrectly Classified
class
shares
2.0
1.5
1.0
0.5
0.0
−0.5
1000
2000
3000
4000
5000
6000
7000
8000


Correctly classified if labeled as 2
Incorrectly Classified
class
shares
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
2000
4000
6000
8000
10000
12000
14000
16000

Thank you