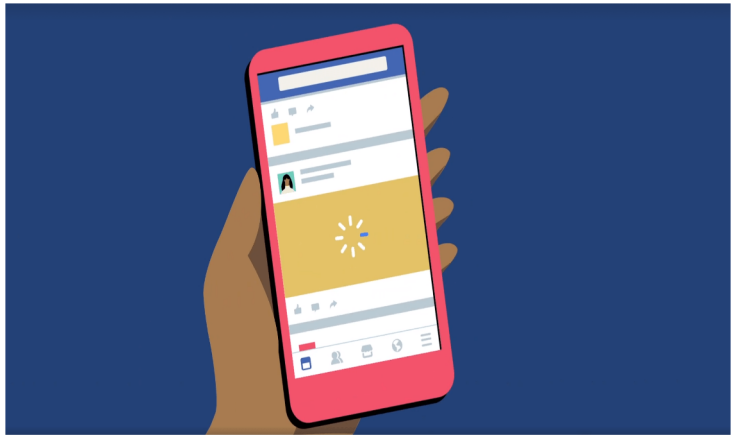
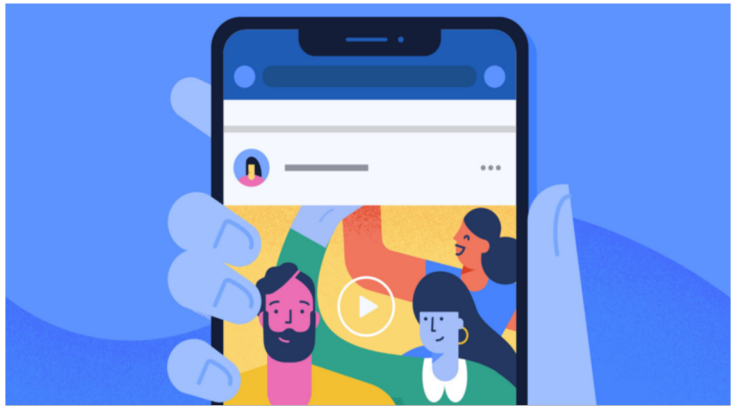
# Facebook Comment Volume Prediction

Dimitrije Adzic
Behzad Pouyanfar



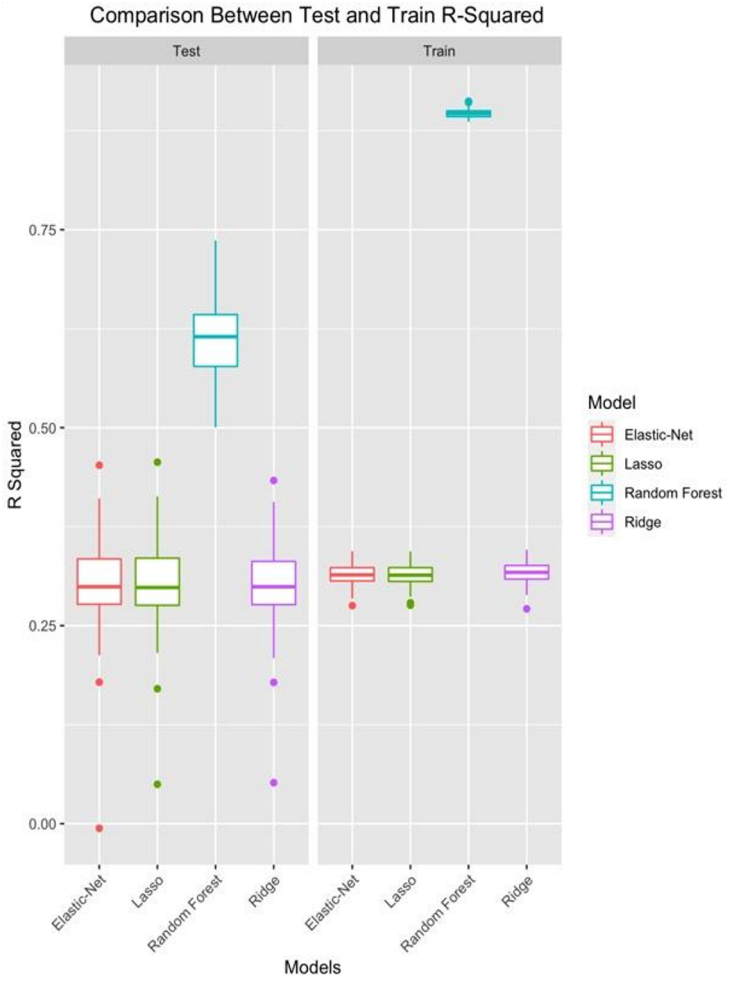


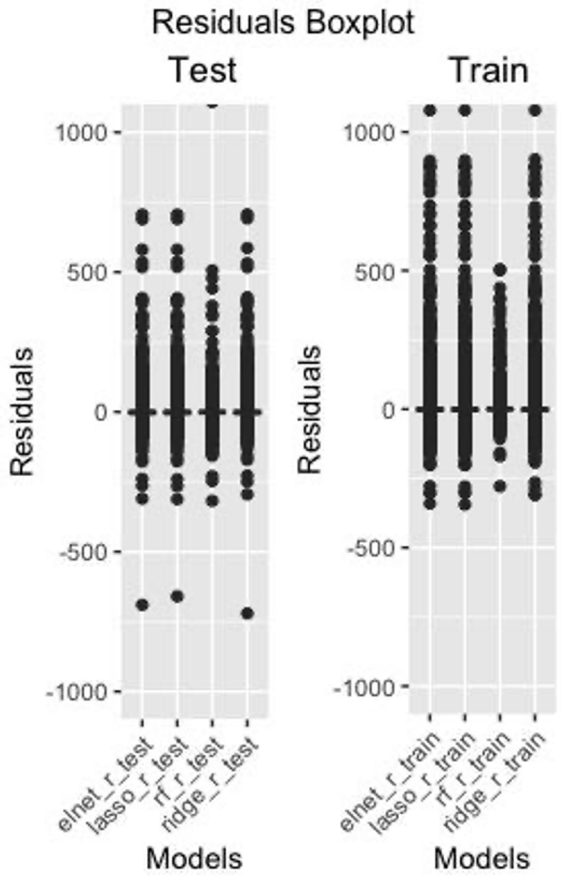
- The increasing use of social networking services had drawn the public attention explosively from last 15 years
- The merging up of physical things with the social networking services had enabled the conversion of routine objects into information appliances
- These services are acting like a multi-tool with daily applications like: advertisement, news, communication, banking, commenting, marketing
- These all services have one thing in common that is daily huge content generation, that is more likely to be stored on hadoop clusters. As in Facebook, 500+ terabytes of new data ingested into the databases every day, 100+ petabytes of disk space in one of FB's largest Hadoop (HDFS) clusters and there is 2.5 billion content items shared per day
- Methods that we will be using are: Ridge, Lasso, Elastic Net, Random Forest (500-trees)
- There are 40949 observations with 54 features. Our response variable is number of comments in a facebook post

**Data Source:** https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset
**Paper:** https://ijssst.info/Vol-16/No-5/paper16.pdf

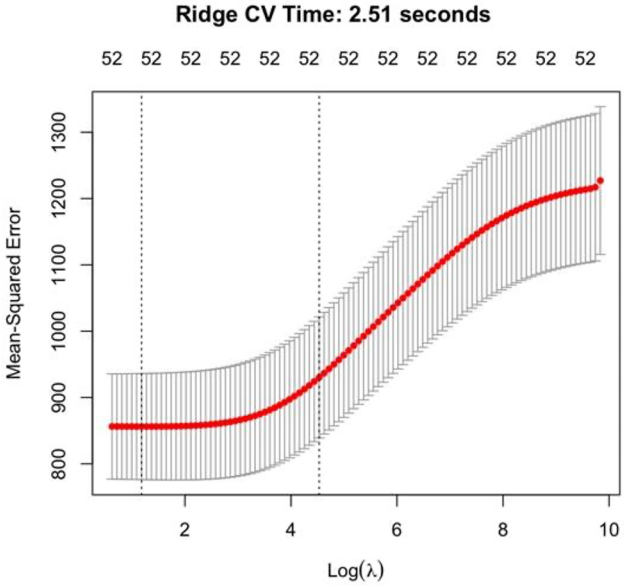
# Boxplots of 100 Simulations for Test and Train

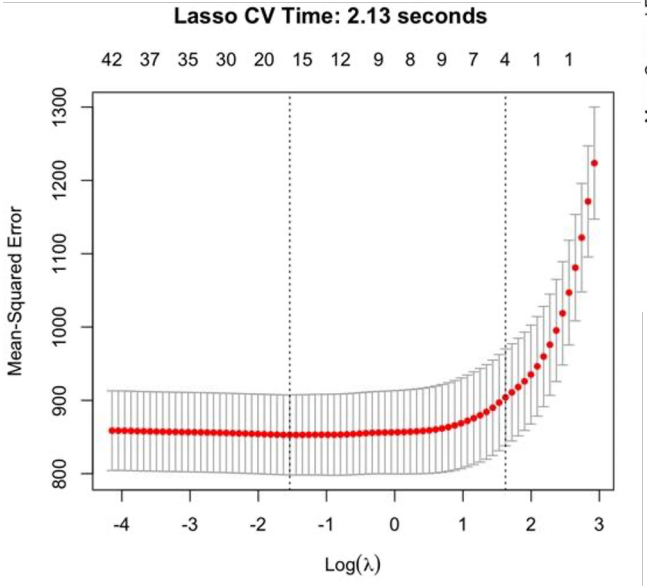
| | Train $R^2$ (Avg) | Test $R^2$ (Avg) |
|---|---|---|
| **Ridge** | 0.3166 | 0.3004 |
| **Lasso** | 0.3141 | 0.3015 |
| **Elastic Net** | 0.3144 | 0.3015 |
| **Random Forest** | 0.8967 | 0.6117 |

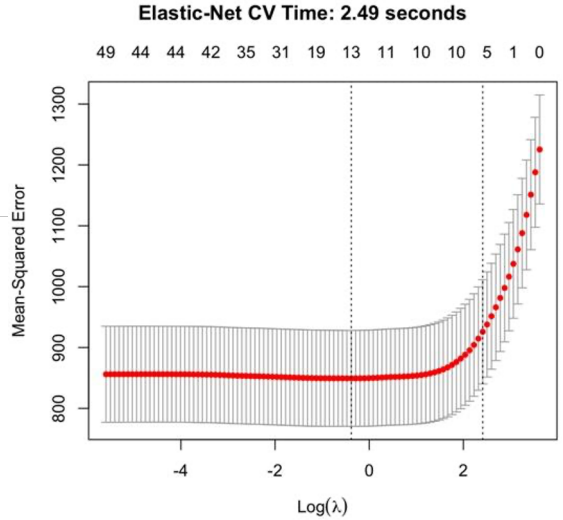


# Cross-Validation Curves



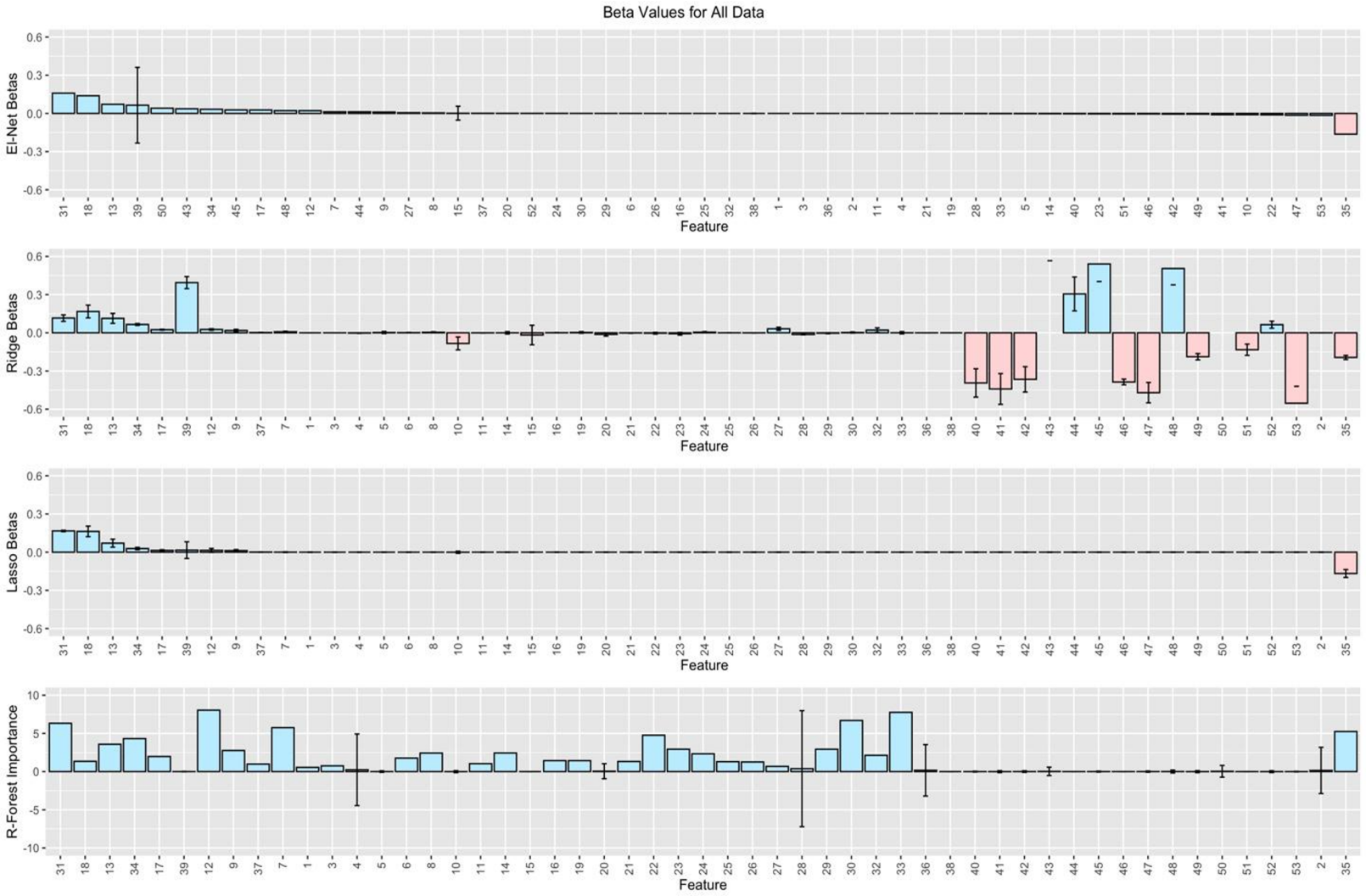



# Test R-Squared Confidence Interval and Time

| | Test R-Squared Lower Bound<br>For 90% confidence interval | Test R-Squared Upper Bound<br>For 90% confidence interval | CV Time<br>(Average) | Single Fitting Time<br>(Average) |
|---|---|---|---|---|
| **Ridge** | 0.2982 | 0.3026 | 1.76 seconds | 0.1421 seconds |
| **Lasso** | 0.2993 | 0.3038 | 1.66 seconds | 0.1394 seconds |
| **Elastic-Net** | 0.2994 | 0.3037 | 1.69 seconds | 0.1243 seconds |
| **Random Forest** | 0.6108 | 0.6126 | NA | 9.0020 minutes |

# Coefficients

Elastic-Net and Lasso agree more or less on coefficients while Ridge includes more variables




Random Forest with mtry =7 ntree =500

# To Conclude:

- Our results for this specific dataset shows that Random Forest does a significantly better job than linear models. However, it takes much longer.
- It is possible that for different data points the gap between R-Squared in linear models versus Random forest would be smaller. In this scenario it justifies using a linear model which takes significantly less time.