Results

80/20 split

Note that tfidf is incorrect here need to correct it

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | Data Points | Featurization | MSE Value | Processing Time  (sec) |
|  |  |  |  |  |
| Regression Tree |  |  |  |  |
| Regression Tree | 26,548 | Bernoulli | 2.4538 | 919 |
| Regression Tree | 26,548 | tf-idf | 2.4057 | 1064 |
| Regression Tree | 26,548 | Multinomial | 2.4057 | 1147 |
|  |  |  |  |  |
| Linear CCA |  |  |  |  |
| Linear CCA | 26,548 | Bernoulli | 2.1571 | 24 |
| Linear CCA | 26,548 | tf-idf | 2.2181 | 24 |
| Linear CCA | 26,548 | Multinomial | 2.2181 | 25 |
|  |  |  |  |  |
| Adjusted CCA\* |  |  |  |  |
| Adjusted CCA | 26,548 | Bernoulli | 1.68578 | 26 |
| Adjusted CCA | 26,548 | tf-idf | 1.691519 | 29 |
| Adjusted CCA | 26,548 | Multinomial | 1.691519 | 26 |
|  |  |  |  |  |
| Naïve Bayes (double check) |  |  |  |  |
| Naïve Bayes | 26,548 | Multinomial | 2.2629001 | 9 |
| Naïve Bayes | 26,548 | Bernoulli | 2.12768361 | 9 |
|  |  |  |  |  |
| Linear Regression (Baseline) |  |  |  |  |
| Linear Regression | 26,548 | Bernoulli | 5.5778 | 78 |
| Linear Regression | 26,548 | tf-idf | 14.589 | 79 |
| Linear Regression | 26,548 | Multinomial | 14.589 | 79 |
|  |  |  |  |  |
| SVM (using LibSVM) |  |  |  |  |
| SVM | 26,548 | Bernoulli | 2.0354048 | 2589 |
| SVM | 26,548 | tf-idf | 2.16007532 | 4526 |
| SVM | 26,548 | Multinomial | 2.13540489 | 3265 Accuracy = 78.9077% |
|  |  |  |  |  |
| SVM (using LibSVM) |  |  |  |  |
| SVM | 800 | Bernoulli | 2.75625 | 1.5299 Accuracy = 68.125% |
| SVM | 800 | tf-idf | 2.537500 | 1.5907 |
| SVM | 800 | Multinomial | 2.7500000 | 1.543 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Note that CCA is significantly faster while providing lower error

\* In the adjusted CCA we map all the values above 5 to 5 and all the values below 1 to 1