Introduction to Data Science and Analytics

Group Number 1 / Delivery 6

19 Mayl 2019

1 Feature Selection

Table 1: First 15 Word For Mutual Information Gain

Word	***Mutual Information Gain Value***
good	0.050022
great	0.047929
movie	0.039749
film	0.036147
phone	0.035585
one	0.031655
like	0.027275
food	0.026628
time	0.025886
place	0.025647
service	0.024317
really	0.023111
bad	0.021199
well	0.019981
even	0.017811

This 15 words are the most common words in our dataset.

Table 2: Last 15 Word For Mutual Information Gain

Word	***Mutual Information Gain Value***
11	0.000231
1199	0.000231
15lb	0.000231
18	0.000231
18th	0.000231
1928	0.000231
1947	0.000231
1948	0.000231
1949	0.000231
1971	0.000231
1973	0.000231
1979	0.000231
1980	0.000231
1986	0.000231
1995	0.000231

This 15 words are the least common words in our dataset.

Table 3: First 15 chi-square:

Word	***Chi-square value***
great	0.047929
bad	0.021199
excellent	0.011919
good	0.050022
works	0.012124
love	0.014241
waste	0.007679
nice	0.012371
worst	0.009088
terrible	0.007444
poor	0.006045
delicious	0.005344
awesome	0.004643
money	0.007198
horrible	0.004410

This 15 words are the most common words in our dataset.

Table 4: Last 15 chi-square:

Word	***Chi-square value***
minute	0.000924
given	0.001848
verizon	0.002079
evening	0.000462
morgan	0.000462
production	0.001155
wine	0.000924
eggs	0.000462
internet	0.001617
replace	0.000924
heard	0.000924
comments	0.000462
ends	0.000462
stage	0.000462
central	0.000462

This 15 words are the least common words in our dataset.

Two selection method have different formulas. But they both find common words. They're not the same words, but they're all are the most common words in our dataset.

2 Algorithms

We did 10 tests for each algorithm. Because we wanted our results to be more reliable. We split our dataset into 10 part and we did training and testing on each of them

Table 5: Algorithm name: Decision Tree:

K-fold	***Accuracy***	***AUC***
0	0.727	0.732
1	0.740	0.738
2	0.777	0.786
3	0.770	0.773
4	0.783	0.784
5	0.757	0.757
6	0.743	0.743
7	0.770	0.773
8	0.750	0.747
9	0.720	0.743

Table 6: Algorithm name: Naive Bayes

K-fold	***Accuracy***	***AUC***
0	0.803	0.804
1	0.843	0.843
2	0.793	0.802
3	0.837	0.835
4	0.853	0.851
5	0.767	0.767
6	0.810	0.811
7	0.820	0.820
8	0.813	0.812
9	0.733	0.769

Table 7: Algorithm name: k-NN neighbor number is 1

K-fold	***Accuracy***	***AUC***
0	0.633	0.576
1	0.620	0.570
2	0.510	0.574
3	0.567	0.603
4	0.673	0.688
5	0.623	0.619
6	0.653	0.643
7	0.737	0.733
8	0.630	0.666
9	0.667	0.569

Table 8: Algorithm name: k-NN neighbor number is 2

K-fold	***Accuracy***	***AUC***
0	0.617	0.550
1	0.587	0.530
2	0.443	0.521
3	0.530	0.578
4	0.593	0.616
5	0.583	0.578
6	0.610	0.597
7	0.693	0.711
8	0.550	0.601
9	0.687	0.571

Table 9: Decision Tree with remove words occurring less than 5 times

K-fold	***Accuracy***	***AUC***
0	0.680	0.679
1	0.740	0.739
2	0.727	0.737
3	0.763	0.768
4	0.743	0.744
5	0.727	0.726
6	0.753	0.752
7	0.740	0.744
8	0.773	0.772
9	0.710	0.725

Table 10: Naive Bayes with remove words occurring less than 5 times

K-fold	***Accuracy***	***AUC***
0	0.737	0.741
1	0.777	0.781
2	0.723	0.731
3	0.797	0.795
4	0.803	0.801
5	0.733	0.734
6	0.770	0.771
7	0.793	0.793
8	0.807	0.809
9	0.733	0.763

Table 11: k-NN neighbor number is 1 with remove words occurring less than 5 times

K-fold	***Accuracy***	***AUC***
0	0.620	0.641
1	0.600	0.618
2	0.700	0.682
3	0.680	0.668
4	0.673	0.666
5	0.597	0.598
6	0.637	0.640
7	0.680	0.657
8	0.740	0.724
9	0.617	0.654

Table 12: k-NN neighbor number is 2 with remove words occurring less than 5 times

K-fold	***Accuracy***	***AUC***
0	0.650	0.617
1	0.567	0.573
2	0.687	0.687
3	0.637	0.638
4	0.663	0.661
5	0.577	0.576
6	0.600	0.601
7	0.673	0.694
8	0.703	0.727
9	0.730	0.680

3 Mean Accuracy

Table 13: Mean Acuracy

Algorithm	***Accuracy***	***AUC***
Decision Tree	0.754	0.757
Naive Bayes	0.807	0.811
k-NN neighbor number is 1	0.631	0.624
k-NN neighbor number is 2	0.589	0.585
Decision Tree with remove words occurring less than 5 times	0.736	0.739
Naive Bayes with remove words occurring less than 5 times	0.767	0.772
k-NN neighbor number is 1 with remove words occurring less than 5 times	0.654	0.655
k-NN neighbor number is 2 with remove words occurring less than 5 times	0.649	0.645

We can see that k-NN algorithm with neighbour number is 2 has the lowest accuracy and AUC value. Naive Bayes algorithm has the highest accuracy and AUC value.
When k-NN algorithm neighbour number increases, accuracy and AUC value is decreases.
k-nn Algorithm gives higher accuracy and AUC values when we remove the words occurring less than 5 times. But the Decision Tree and Naive Bayes gives lower accuracy and AUC values when we remove the words occurring less than 5 times.