BU CS506 Spring 2018 Kaggle Competition

Bote Robert Xu brx@bu.edu
Github, Kaggle username: bidoai

For this homework I use a simple linear regression model that takes as input HelpfulnessDenominator, HelpfulnessNumerator, Summary, Text, HelpfulnessNumerator / HelpfulnessDenominator, sentiment polarity, and sentiment subjectivity, and makes predictions about the score. I then take the predictions made and round them up to the nearest integer. To analyze the sentiment of the text data, I use TextBlob, a Python library for processing textual data built on NLTK and pattern. I combine the summary and text of every review, and analyze it with TextBlob. TextBlob returns a sentiment polarity score, which is a number between -1 and 1, and a subjectivity score between 0 and 1. I also messed around with using Watson's Natural Language Understanding service, because it can also return the mood (anger, disgust, fear, joy, and sadness) of text data. However, I am limited to 30, 000 requests per month so I was not able to analyze all 500, 000 reviews. I also tried incorporating the product type into my model, but there around 7000 different products and creating 7000 indicator variables would make the model too complex. The summary of the linear model is as follows:

```
1 results = model.fit()
2 results.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.305
Model:	OLS	**Adj. R-squared:**	0.305
Method:	Least Squares	**F-statistic:**	4.039e+04
Date:	Sat, 05 May 2018	**Prob (F-statistic):**	0.00
Time:	14:53:54	**Log-Likelihood:**	-6.9484e+05
No. Observations:	460804	**AIC:**	1.390e+06
Df Residuals:	460798	**BIC:**	1.390e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.6244	0.007	516.504	0.000	3.611	3.638
x1	-0.1524	0.001	-175.213	0.000	-0.154	-0.151
x2	0.1584	0.001	167.601	0.000	0.157	0.160
x3	0.0849	0.004	23.550	0.000	0.078	0.092
x4	2.7351	0.007	367.932	0.000	2.721	2.750
x5	-0.2548	0.012	-20.631	0.000	-0.279	-0.231

Omnibus:	58908.722	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	111019.023
Skew:	-0.829	**Prob(JB):**	0.00
Kurtosis:	4.742	**Cond. No.**	102.

As we can see, the x variables in this model are all pretty significant and have extremely low p-values, therefore, I continued to use all these variables when making predictions in the final model. Using this model, I feed in my test data and make predictions on the scores. Since this is a linear regression model, the score predictions are floats, therefore I round every prediction to its nearest integer. Using 5-fold cross validation, this model gives an average MSE value of 1.276. When submitted to Kaggle, this method returns a RMSE value of 1.1276. In addition to a linear model, I also tried a regularized linear regression model with various ridge parameters. However, these model were not as good and gave a RMSE of over 3.