TCDS-20
FINAL PROJECT
IRIT ORGIL BRONER

“Any sufficiently advanced technology is indistinguishable from magic”

— Arthur C. Clarke

“Ginny!" said Mr. Weasley, flabbergasted. "Haven't I taught you anything? What have I always told you? Never trust anything that can think for itself if you can't see where it keeps its brain”

— J.K. Rowling, Harry Potter and the chamber of secrets

This is not a slide show, but rather a guide or a cover letter, to accompany the following notebooks:

Wine Dataset - EDA-Copy2

Wine Dataset - EDA-Copy4

Wine Dataset - EDA-Copy22

Wine Dataset - EDA-Copy6

Wine Dataset - EDA-Copy65

Wine Dataset - EDA-Copy9

Basic editing of the data included clean up, removing NAs, removing duplicates and irrelevant columns. Regularly you would want to preserve as much of the data as you can, and even try and extract more features from it. here, I experimented with the data, choosing the columns by their type – numeric, categorical or text.

In a few cases, the categorical data was treated as text. In others, it was left out altogether.

Running time was an issue, making it hard to perform grid searches, and making it an easy decision to let go of data.

I ran KNN, with non-impressive results, as expected. Following code example from class, I cleaned the data from stop words and stemmed the words, then practiced some grid search too. Notebook [2].

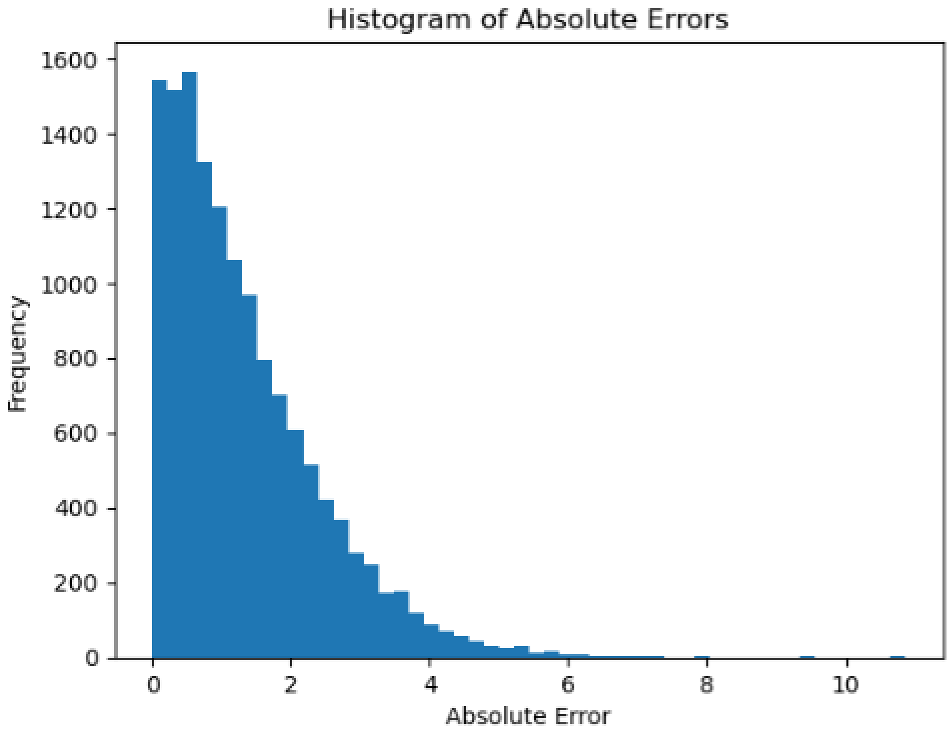
Notebook [4]: Moving on to another experiment, this time using only tee textual description. Tried a couple of variations. The only interesting thing to watch was how the model converged after 7 epochs (out of 10).

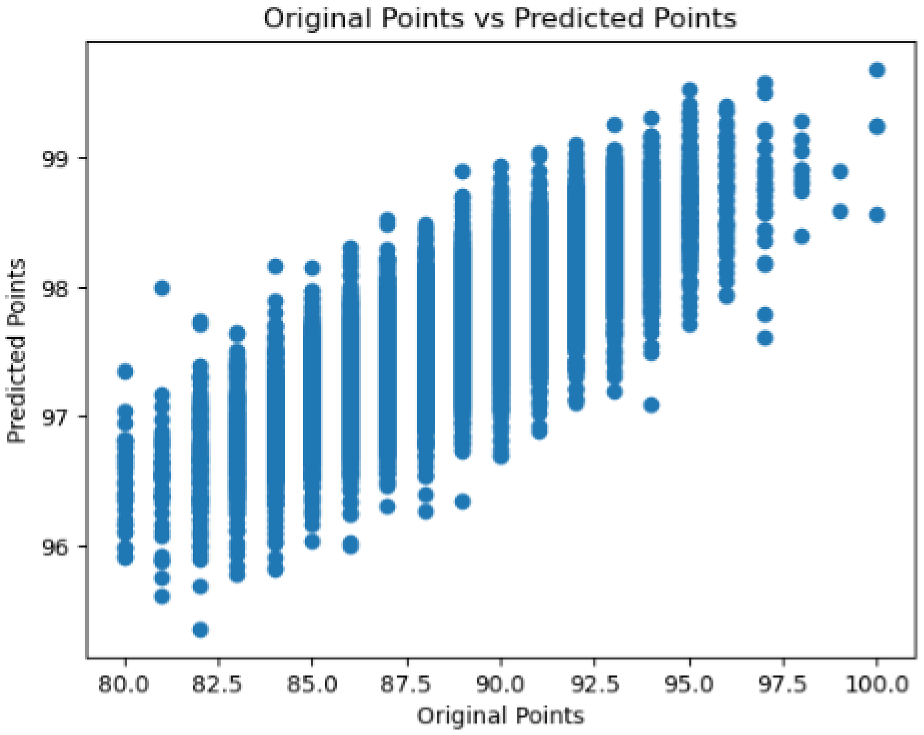
Notebook [22]: Several trials. Different algorithms, different inputs, yielding the first interesting results. I printed graphs and tables of predicted vs original points to see the results. Dealt with overfitting.

Notebook [6] -More versions and small changes in column-choice and models.

Notebook [65] -best results so far:

Mean Absolute Error: 1.32


Histogram of Absolute Errors
Frequency
1600
1400
1200
1000
800
600
400
200
0
0
2
4
6
8
10
Absolute Error


Original Points vs Predicted Points
Predicted Points
99
98
97
96
80.0
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
Original Points

There were many considerations trying to choose the participating data columns, sometimes contradicting ones: at the beginning I thought that, real life logic-wise, “winery” contains the data for region and country as well, making those columns redundant. Later, I thought it might lead me to overfitting, being so many of them, regarding data size.

It what obvious that the price was, If not correlative, at least monotonic with the score. Yet, I found it less effective.

Had there been no size limits, I would be happy to throw all the data at the model, using all the columns, extracting new features, adding new columns such as log(price).

The “whatever works” approach, the concept of black box algorithms, is fascinating, especially since it is, of course, anything but. It is only the complexity that makes it so.

I certainly want to learn more.

After trying most of what we saw in class, I had the not so good idea to try BERT.

After the necessary installations, when it did not work, I used a small sample of the data (2000 lines, and later 1000). I tried GOOGLE COLAB.

I once got it to run, with very bad results.

That took a lot of time. The only good things I can say about it is that it was interesting, and I would have done better with proper tools.

Some of the unsuccessful trials can be seen in notebook [9].

THANK YOU!