**Do you agree or disagree with this statement? Why or why not?**

I disagree with the statement but as with anything else I would say there are some caveats. Having taken MSDS 400 - Math For Data Scientists, I have to say it was not until the second assignment in this course did I really understand why linear algebra is not the solution for numerical problems like we had with the housing price competition on Kaggle.

When I was taking MSDS 400, the section I liked the most was the linear algebra topics because I always found the topic to be something I can relate to since you define your cost method and design a way to predict the cost given the features used in the equation. It seemed pretty simple in the problems we did in the course but what is really important was not the problems we did but the concepts we explored in that course. In addition, the MSDS 401 – Applied Statistics with R course really helped on the second assignment, especially the abalones exploratory data analysis assignments we did to normalize the data used to make predictions.

The second assignment is really a combination of MSDS 400, MSDS 401, and the scikit-learn libraries we learned about this week. I simply cannot imagine using straight linear methods to deal with the 163 columns of potential features that I used to predict the housing prices. And this is really the main benefit the machine learning provides over traditional linear methods.

Teboul (2018) described machine learning as not just glorified statistics but something on top that is too hard to do using the traditional linear methods. In MSDS 400 we did linear methods that had one, two, three, four, and maybe five features. In MSDS 401 we took these features and we did exploratory data analysis and applied statistical methods to normalize the data and make it more distributed to better fit a linear model.

Then came the second assignment in this course where we had 163 columns and you simply can’t easily just rely on exploratory data analysis and be given clear cut features to base your cost function on. And that is the point of machine learning! The methods we learned this week with scikit-learn provide learning methods that identify the tricky correlations in the data sets that exploratory analysis couldn’t identify, and it figures out the best correlations out of 163 possible features. We are not given or can’t easily define the cost function, rather the machine learning techniques identified the pattern on its own from the data.

However, it’s important to keep in mind that you don’t need machine learning for everything. The problems we did in MSDS 400 and 401 are easily solved with using traditional linear models and we don’t have to overcomplicate it. That is what I meant by my opening statement “with some caveats”. Machine learning is great for the very complicated models like the Kaggle competition this week and simply would be too hard to do without machine learning, however sometimes traditional linear models are just fine depending on the complexity of the problem your trying to answer.

References:

Teboul, W. (2018, July 22). *Why use machine learning instead of traditional statistics?* Medium. https://towardsdatascience.com/why-use-machine-learning-instead-of-traditional-statistics-334c2213700a

**Post a segment of your code from Assignment 1 that was either problematic for you or that you believe you handled very well. Be sure to explain the code!**

Whenever I do exploratory data analysis, I always find that I want to slice and dice the data in different ways using different views on the data. However, this can lead to what I like to call spaghetti code if you don’t use re-usable code using functions. What I think I did very well on in the first assignment is using functions throughout to look at the various ways I wanted to slice and dice the data without creating spaghetti code.

If I had spaghetti code it would be straight forward to create the initial graphs but the moment, I want to change it I would have to go to each graph and modify the code and that is very error prone and tedious. Instead I created functions throughout the assignment that I used and had a single place to modify when I wanted to make simple changes.

For example, here is a code segment of a function I used to plot a stacked bar chart of the new case trend in each continent with the stacked bars consisting of countries. I then reused this function for all the other continents instead of having separate blocks of code for each continent’s graph. To save space I only will provide a screenshot of one graph but my code for the assignment can be found on GitHub: <https://github.com/chrisfesta/NWU_MSDS422/tree/master/Assignments/Assignment%201>

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated