**Part 1: What is overfitting? Why is it a special problem when working with machine learning (data-adaptive) models? How can we protect against overfitting?**

Our textbook does a great job describing overfitting. Another term for overfitting is overgeneralizing and it is when our model performs well on training data, but it does not perform well when put in practice. This is a special problem in machine learning because some models can detect subtle patterns in the data, and if too noisy or small, then the model will detect patterns in the noise (Géron, 2017).

Some possible solutions to prevent overfitting are:

* Simplifying the model by selecting fewer parameters which reduce the number of attributes in the training data
* Getting a larger training data set that better represents the population the model is applied to
* Reduce the noise in the training set. A good way to do this is to fix data errors and remove outliers

(Géron, 2017)

At my company, there was a group of individuals who suffered recently with overfitting when they attempted to do a data analysis on a survey they sent. The survey had only three questions

1. Have you personally encountered internal pages in our knowledgebase with incorrect or stale information?
2. Second question assumed I selected yes and proceeded to ask a question about stale data and there was no path if I thought the knowledge base was great
3. Third question suffered the same as the second

This survey was also sent to a special focus group of technologists who are known to be “complainers” and the authors of the survey were a member of… as opposed to a sample that better represents the entire company’s technology employees.

This was overfitted because the training data was too small, and the survey suffered from interviewer bias. Therefore, applying any analysis or models on the results would overgeneralize and perform bad. This is similar to the example in our textbook where it mentioned: if you visit a country and the taxi driver rips you off, you might be tempted to say all the taxi drivers in that country are thieves when that is not the case (Géron, 2017).

Nevertheless, I gave those employees an earful ☺

Lastly, in our first assignment doing a data analysis on COVID-19, I also was puzzled on some of the visualizations I created because it had noise and data errors. For example, there was a country that reported negative numbers, and this required me to do a data cleanup.

All of these are examples of overfitting.

**Part 2: Every week, I want you to share Python coding as a Team. To start, go to Kaggle.com and sign up for the Titanic: Machine Learning Through Disaster competition.** [**https://www.kaggle.com/c/titanic**](https://www.kaggle.com/c/titanic)

This is a very interesting data set. Who hasn’t seen the Titanic movie!

Code on GitHub: <https://github.com/chrisfesta/NWU_MSDS422/blob/master/Module%201%20-%20Disussion/Titanic%20Challenge.ipynb>

To start I made some observations on the data:

A screenshot of a cell phone

Description automatically generated

* There are a total of 891 passengers in our training set
* The Age feature is missing approximately 19.8% of its values. Age is usually an important feature to survival, so I imputed the data with the median
* The Cabin feature is missing approximately 77.1% of its values. This may be an important feature so for now I imputed the data as a category of Unknown and will decide to drop this feature or use it later
* The Embarked feature is missing 0.22% of its values. I did not do anything with this for now and will look at it later if it becomes an important feature

***Exploration 1: What percent of males and females survived?***

A screenshot of a cell phone

Description automatically generated

***Conclusion:***

As predicted, females had a higher survival rate than males

**Exploration 2: Did class have a major impact on survival?**

A screenshot of a cell phone

Description automatically generated

***Conclusion:***

As predicted, first class passengers had a higher survival rate. Morally I do not agree with this but this is indicative of the decade titanic occurred in

**Exploration 3: What is the distribution of age of all the passengers?**

A screenshot of a social media post

Description automatically generated

***Conclusion:***

The age of each passenger is within the middle-age group of 30-40. Keep in mind that I imputed this data using the median on 19.8% of the passengers where it was missing.

References:

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (First edition). O’Reilly Media.