IST 707 Applied Machine Learning
By Prof. Kelvin King

Assignment 1

Task 1: What, why, and where?
Task 2: Practice Your Critical Thinking and Writing
Task 3: Machine Learning and Limitations

Submitted by:
Samarth Sandesh Mengji
SUID: 718473878
NetID: smengji@syr.edu

Date of Submission: 9/10/2023
Syracuse University – School of Information Studies

Introduction of the Assignment: In this assignment, we go into the world of Machine learning-its benefits over traditional methods, applications in business and its limitations. We also try to practice our critical thinking and writing by researching two news articles on Google flu trend.

Task 1: What, why and where?

What is Machine Learning?
Machine learning is a part of Artificial Intelligence where we train and develop models and algorithms to predict the upcoming thing and make decisions based on those predictions by providing data to that model. The word 'machine learning' itself defines that the machine learns from the data and depicts human-like behavior and intelligence. In simple words, the machine tries to mimic a human by considering the data that is being fed to it.

Where can you use it?
Machine learning can be used in different industries and domains where there is a need to make decisions based on huge data or a need for process improvements or developing new products. There are many more applications where machine learning can be applied in different sectors as well. Some infamous examples of machine learning would be forecasting market conditions and supply chain challenges, hiring, and recruiting, monitoring physical assets, making long-term strategic decisions. Even at a university, machine learning can be used to predict a student's success in coursework by tracking and monitoring his/her progress. This can help the faculty make right informed decisions to support students.

Why should we use it?
Machine learning provides a wide range of advantages from automating decision-making process to providing real time insights and predictions. It can provide more and more accurate results each time as it learns from previous data which will eventually lead to better business decisions and more ROI. Using machine learning, we can optimize all processes involved in a project to make use of least resources and provide the maximum output. It has a lot of advantages over traditional methods of analysis and predictions as it can handle complex and big datasets. **Some notable numbers from the article, "Notes from the AI frontier", Machine Learning techniques have the potential to create between $3.5 trillion and $5.8 trillion in value annually across 9 business functions in 19 industries.** Therefore, we should make use of ML.

What benefits does it have over other traditional analysis, prediction, and decision-making methods
Machine learning can scale up quickly, adapt to changing constraints, provide high accuracy, and efficiency. ML can handle large and complex datasets (labeled and unlabeled data) to extract valuable business insights which would be highly impossible to do with traditional analysis methods. As highlighted in the McKinsey's article of deep learning, the ability to adapt to changing trends and to continuously learn on that are one of the benefits of ML over traditional methods, "**Deep learning's capacity to analyze very large amounts of high dimensional data can take existing preventive maintenance systems to a new level**". Machine learning can also help in automating tedious manual and repetitive tasks which reduce the use of resource utilization and proper allocation could be achieved. Right decisions could be made by creating

simulations before implementing into business resulting in more accurate decision making and predictions. These are some of the key benefits of machine learning over traditional approach.

What business problems are machine learning and artificial intelligence most suited to?
As stated in the McKinsey's article, "Two-thirds of the opportunities to use AI are in improving the performance of existing analytics use cases". Business problems like process improvement, demand forecasting, forecasting market conditions, optimizing supply chains, detection of fraud are few use cases where machine learning can be implemented. Machine learning are most suited where there is a need for finding trends and complex patterns in huge complex datasets and where there is a need to automate processes for efficient data-based decision making.

Task 2: Practice Your Critical Thinking and Writing

Critique: "Google Flu Trends: The Limits of Big Data" (The New York Times):
The New York times article from March 2014 criticizes the Google Flu Trends project which was initiated to detect flu outbreaks by using data generated by people's search on google. The article talks about the inaccuracy of GFT and various factors influencing the model. The problem with GFT was that it was solely based on the keywords which people searched which could have been easily manipulated by different things like false and over coverage by media channels (One of the things mentioned in the article). Media coverage increases the keywords model picks up for prediction resulting in providing inaccurate results. Not only GFT was inaccurate in its predictions but also it was not transparent for outsiders to calculate the accuracy of the model. "Using GFT as a standalone flu detector would be questionable", said Google in the article. They talk about different challenges that they face while using big data for flu prediction when the source of that data is questionable.

Defense: "In Defense of Google Flu Trends" (The Atlantic):
The Atlantic article talks in defense of GFT. Even though GFT was inaccurate in its predictions in its early stages, it can be seen as a room for improvement and advancement in using such tools and real time big data for predicting such things. The article talks about the limited resources and surveillance we have in traditional methods. There is a lot of delay in reporting outbreaks in traditional methods as compared to GFT. GFT opens a new door in predicting flu outbreak and how it can be improved over time. GFT can be seen as big step in using google search data (Real time big data) in public health domain. In summary, the article gives us a new perspective of using real time big data and the importance of continuous improvement.

Thoughts on Debate:
Even though GFT failed at predicting in the start, it can be seen as new breakthrough of using real time search query big data in the public health domain. I think the defense article makes sense that continuous improvement will eventually lead to better predictions which would have many benefits over the traditional surveillance methods. There has to multiple layers of verification along with GFT to properly predict a flu outbreak. By using GFT in conjunction with other resources, we accurately predict next outbreak. In summary, properly addressing the problems and opportunities in GFT and traditional methods, we can work on improving GFT and

predict flu outbreaks accurately.

Task 3: Machine Learning and Limitations

Three Machine learning limitations are as follows:

1. **Data Quality and Quantity**
Limitation: As studied in our previous task, biased data will provide with inaccurate results and so will insufficient data.
Overcome: Data collection is one the first and most important things in developing ML models. Using different data preprocessing techniques to remove the outliers, biases, null values will increase the quality of the data.

2. **Lack of Common sense in Machines**
Limitation: Machine lack common sense and sense of interpretability. Making sense contextually can be challenging for machines since they look for patterns in data.
Overcome: Overcoming this limitation is one of the challenges in ML. We can simplify the model by using decision tree or linear regression, but they have their own limitations. On researching more on this point, I found LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) models which provide explanation about the decision and prediction that model makes but I cannot comment on their accuracy.

3. **Overfitting**
Limitation: Models will work well on the training data by overfitting the training data but might not perform similar on the testing or unseen dataset.
Overcome: There are various techniques in ML which can prevent the overfitting of data like L1, L2 regularization, bagging and boosting or hyperparameter tuning.

Some more limitations include lack of sense in judging Ethics and morality as Professor King explained in class about discrimination of race, color, or religion by machine learning models. Those are one of the examples of limitations of ML.

Website Citations:

https://www.sciencedirect.com/science/article/abs/pii/S0278612522001066
https://archive.nytimes.com/bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/
https://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/