EAS 504 FINAL - ASSIGNMENT

By: - Katyayni Shankar Kaushik (50289158)

A.) <u>Summary of experiences over all the Lectures</u>: -

- It has been an amazing experience to know about different domains wherein data science techniques are applied. I was able to connect my use case (one, on which I'm working on as a Data Science Intern in ExxonMobil) to few other use cases that various speakers talked about.

- These sessions also gave us a glimpse of various data science problem statements, especially the most famous ones. I understand that most common type of problems are Search, Recommendations, Unsupervised and Supervised machine learning problems, Interpretation of spatial and temporal data using deep learning techniques etc.

- These problems can further be sub-categorized based on the business domain from which these problems originate.

- These sessions also explained us that with the exponential growth of data and the increased availability of computation power, implementation of data science techniques has become rather common and suppose to increase going forward.

- Availability of big-data technologies has also made processing of huge chunks of data a lot easier and has improved the turn around time to get results from them.

- Some of the sectors that are applying Data Science techniques to the best use are: - Technology, Ecommerce, Automobile, Financial Sectors (banks, stock markets etc), healthcare, Energy etc.

- We also learned about the career track and educational background of the people who conducted these sessions taking out their precious time.

- They told us about the importance of selecting the problem statement and viability of the same before working on a dataset and the very common challenges that we would face while on working on it.

- Many of them have been associated with company since the start and therefore know the in and out of how company started the data science venture and how it is benefitting them, generating more profits and being more accurate with decision making.

- We were also introduced to tools and libraries that companies usually prefer i.e.

  Languages: - Python, R programming, Mysql etc
  Libraries: - Pandas, Numpy, Scikit-Learn, Scipy, Matplotlib, Seaborn, Tensorflow, Keras, plotly NLTK etc.
  Tools: - Weka, AutoML, Tableau, Qlikview etc

B.) <u>Commonalities & Differences in lectures</u>: -

Commonalities in Lecture

- The main commonalities that I have observed among all lectures is that each and every sector is trying to implement the data science techniques to best possible use, the reason for the same is that since every company has generated so much data that it has become essential for them to apply techniques so as to understand them.

- Another similarity among all the lectures, that speakers talked about was before trying to solve any problem, a data scientist should always first question the relevance of the problem statement and whether the problem is solvable or not.

- The third similarity that I found among all the lectures was that no matter in whichever domain we are working in and whatever dataset we are working on, we always will have to manipulate the data, solve data integrity issues, outliers etc and for doing all this a general inquisitiveness in a data scientist is essential.

- We also noticed implementation same sort of techniques implemented in different domains, for e.g. in Lecture 1 we were explained how Deep learning methods such as CNN is use to identify the **License plates** and in Lecture 5 & 6 same technique is used in the **ecommerce domain** to categorize the images of different products in various categories. Deep learning techniques are also used in **manufacturing sector** (Lecture - 10) to observe the images of top surface of the metal during the amalgamation of different metals, this is done to identify any sort of impurity.

- Time Series modeling (Lecture 7 & 8) is used in various sectors such as Financial Sector, Business process services, industrial process and use same sort of models such as **ARIMA, SARIMAX** etc.

- Various other domains use Natural Language Processing to make user experience more joyful. For e.g. **HR services, Customer care services** (Lecture 5, 6 & 7) etc use NLP techniques to create ChatBots so as to make it easier for user to navigate while resolving an issue or making a query.

- I also noticed that though different domains apply different sort of techniques, most common languages and libraries used by various companies are Python, R programming, Mysql, Tableau (visualization), Pandas, Numpy, Scikit-learn, matplotlib, Scipy, Tensorflow, Keras etc.

- Above all, one rule that applies to every domain is that each and every company needs to understand the legality of the data (Lecture - 4) they are using and if the data belongs to an individual, company need to take a prior consent of those individual before using it. Company also needs to share the details of how that data is being used by them and an individual should always have the control over the data. Companies also needs to make sure that any algorithm

that they have created, should not be biased towards any section of the society and should not discriminate based on region, color, sex, age etc.

Differences in Lecture: -

- One important difference that I noticed is that same data science technique can be applied in multiple domains differently and therefore it is very important to know the nuances of various businesses so as to apply the data science methodology in that domain, in the best possible way. **Domain (Business) Knowledge is very important**.

- I understand that Ecommerce (Lecture 1,5,7,), in comparison to other domains, is the one in which data science techniques are most widely applied and there are various set of data science applications and not just one for e.g. **Search, Recommendations for query understanding and autocompletion and Deep Learning models for categorization of products based on image data, NLP techniques for building chatbots for customer care** etc.

- Many domains such as Ecommerce and Financial (Lecture 5, 7 & 8) sector generate huge amounts of data in comparison to other domains and thus these domains also implement technologies like Big Data, AWS to make processing faster, more optimized and make best use available technologies.

- In lecture-9 we got to know about **Electronic Design Automation**, where we got to know that machine learning based approaches are not always used to predict the performance of the manufactured electronic product, sometimes it better used to test the design level framework of a chipset by analyzing the timing performance prediction of a particular design. This technique is used by a company named Synopsis.

- Some domains such as metallurgy, industrial sector etc , we read in Lecture (3, 9 & 10), may requires knowledge that is specific to that domain apart from the knowledge of statistical modeling, probability etc that is required for applying data science techniques. For e.g. metallurgy requires knowledge of chemistry and industrial sector requires some knowledge of physics. It is essential to have some knowledge of these domains too because without the basic understanding of the subjects, it would be very difficult to work on such datasets from these domains.

- One important difference that I noticed while going through all the lectures is that data science laws (Lecture - 4) may be applied in a different way in different companies since they are working on different problem statements. For e.g. ecommerce, financial, healthcare data has the potential to use personal data and therefore an individual's consent is necessary whereas domain like energy, defense, government has classified data and should be worked upon under legal framework.

C.) Ideas Introduced in Cross-Cutting Theme Lecture

- Apart from what we have learned about different domains, where data science has its applications, we should always ensure few things such as is the data being used in the right way, if it is a private data does a user know about it and do we have his/ her permission to use that data etc.

- This is one very important topic that was discussed in Lecture 4 by Professor Jonathan Maines, who is an Assistant Clinical Professor in UB School of Law and teaches about free speech, privacy and anti-discrimination.

- Professor Jonathan Maines talks about trust, privacy, ethics and legal aspects from the perspective of data usage. He explains us the importance of fairness & transparency in data science applications being built and also about legality and ensuring personal privacy while working on a data.

- He gave us an example of a Financial company i.e. CompuCredit that use to cut credit score of an individual based on personal data like marriage counseling, therapy, frequency of visits to clubs etc, CompuCredit did this without the knowledge of a customer as a result of which a lawsuit was filed against it.

- He also talked about the **bias introduced in the model** when wrong data is used. To give an example, LAPD used a data science model that predicted whether an individual is accused in criminal charges or not. Since the model was built using features like Age, color etc therefore is was flawed and many a times framed innocent people as accused.

- These examples explained us that we should take consent before working and on a private data and any data science application that is being used should not be biased towards a particular section of the society and should be cause any sort of discrimination.

- Professor also talked about different type of data related laws to ensure privacy, anti-discrimination, transparency etc for e.g. laws relevant to algorithm fairness are: -

    - Federal Trade Commission (FTC)
    - Fair Credit Reporting Act (FCRA)
    - EU General Data Privacy regulation (GDPR)

- It is also important to share details with an individual and letting them know what data is being used, while building an application. People should also have control over their privacy, thus allowing what data they want to share.

D.) Case Studies: -

1.) Ecommerce (eBay)

Mr. Manoj Kumar Rangasamy works as a Tech Lead in eBay and he conducted the 5th lecture wherein he talked about the applications of data science in eBay. He gave us a brief insight of how data science techniques are used in ecommerce domain to improvise both Buyers and Sellers experience. Following are things he talked about: -

Seller experience

Search: -

is one of the most important application of data science in ecommerce domain. It eases of the accessibility of the website for buyers and sellers. Buyers and Seller experience can be defined as follows: -

Buyers

Buyer experience is be aimed at making transactions more comfortable for individual who visit website to make purchases. Thus, buyers are provided with features such as intent classification, query categorization, query auto-completion, recommendations based on relative/ comparative/ structured searches, spell correction recommendation for buyers to search for the correct product or the one they are looking for.

Seller

Seller experience is aimed at making it easier for an individual to sell their products online. Thus, they require different set of features when compared to buyer. For e.g. comparative auto- pricing feature helps the seller to make an informed decision while setting the price of a product. Top products recommendations help sellers, trying to sell quality products, make their listings more visible on website. Spell correction feature can also be used to improve seller experience. Sometimes seller list their products with incorrect spelling, this feature ensure that products are listed correctly.

There are many types of searches that ecommerce website implements to make the buyer/ seller experience much more joyful, following are they: -

- **Text Search**
- **Faceted Search**
- **Image Search**
- **Voice Search**
- **Conversational Search**
- **Recommendations**

These searches have many applications for Query Auto-Completion, Query Understanding, Query Categorization: -

Query Autocompletion: -

Whenever a customer arrives on the website and tries to search for a product, it is very important to ensure he/ she should reach their product quickly. To ensure this, one way is to suggest the customer what he/ she is looking for. Whenever customer types few characters of the product he/ she is searching for, recommendations should be made of the likely product he is looking for, this reduces the time of the search.

Query Categorization: -

Whenever customer searches for anything thing on website, it is very important to understand under what category does the search product comes under. Reason to do this is to make the search result much faster and more accurate. Imagine if a website won't classify categories for the searches and anytime a new user would come and search for a new product, website would have to scroll through millions and billions of products listed on the website. This would make the search so much delayed that website would hang and experience delivered to the customer would not be good enough.

To improve this Deep Semantic Similarity Models/ Convolutional Latent Semantic Models are used to improve the ranking for a given query in the document.

Query Understanding**: -**

also known as intent classification is to understand the query/ search made by the user/ customer. Many a times customers also feed wrong name of the product they are looking for. They should be suggested the correct name so as to make their search much easier.

This is done by using recommenders' systems, this methodology calculates the probabilistic similarity of products using Pearson's coefficient (there are other ways too, to calculate the similarity). Using clustering techniques, categories are defined for the products and using recommenders' systems, similar products are displayed to customers upon searching for a particular product.

Impact on the Organization

This improves the turnaround time of search result and also make customers experience much more joyful by letting him search for his/her desirable product in a much faster and easier way. The whole idea is to improvise the customer experience so that he/ she can reach the product, they are looking for, quickly and make a purchase from the website. Easier it is for a customer to search more likely it is for the company to make a sale.

2.) Transport (State Transport Authorities)

This use case has multiple applications and was discussed by Mr. Sriganesh Madhavanath in Lecture 1. He discussed implementation of Data Science in Transport sector and gave various examples, following are they: -

Characterizing connectivity in Public Transportation: -

- Transit Network Connectivity deals with monitoring the movement of transport (e.g. buses) in an area and looks at the problem from 2 perspective i.e. Supply and Demand. Supply is related to provider and considers spatial/ temporal measures, Demand is related to Consumer. End goal is to improve the functioning and optimize the cost so as enhance the provider and consumer experience.
- Many significant features are taken into consideration such as physical network, service schedule, reliability etc. Relevant information like walking time, waiting time, travel time, buffer time, in-vehicle time, stop-level connectivity, stops, routes, service schedule etc is considered to optimize the movement of vehicles.
- The idea is to optimize the network in such a way that people don't have to wait more than the designated time of arrival of buses. To achieve this deep learning techniques are used for e.g. Convolutional Neural Networks for Spatial Data and Recurrent Neural Networks for Temporal Data. Using such techniques movement of the transport is observed and gaps in the network are identified.
- This is a research-based use case and upon applying data science techniques and after doing some visualization, it is observed that if all the services ran on schedule there will be an overall improvement of 62% in connectivity.

Automated License Plate Reidentification: -

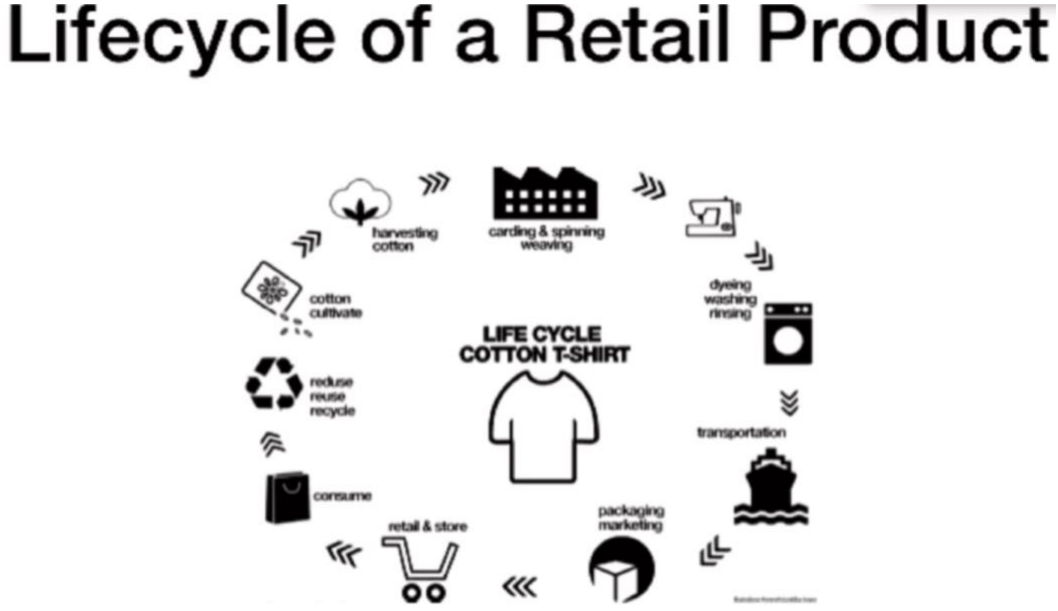
- Another use of Data Science Techniques applied in Transport domain is to read the license plate of vehicles, running on highways and charge them toll accordingly. This removes the need of physical toll booth structures that slows the movement of traffic.
- This technique is already being applied in some of the New York State Highways wherein an individual doesn't have to pass through a physical Toll Booth. There are structures on the highways that have camera installed on it and the camera has the capacity to read the license plates of cars passing through the highways.
- Various techniques such **SURF, SIFT features, CNN, Metric Learning etc** is used to achieve this automation.
- **Deep Learning based models** are used to read the license plate of car, whose images are captured through camera. Model has been tested on huge number of cars passing with the speed of 170 mile/hr and showed an accuracy of more than 95%.

Impact on the State Transport Authorities: -

Upon applying deep learning techniques i.e. CNN & RNN, New York State transport Authority was able to achieve automation and improvement in services by optimizing the functioning. It also improved the transport movement in network thereby generating more revenue.

3.) Life-Cycle of Retail Product

This case study was discussed by Mr. Anurag Bharadwaj in Lecture 7. He currently works as professor in Northeastern University Silicon Valley and has worked as a Director in eBay. He has done is PhD in Computer Sciences from University at Buffalo.



Data Science plays a huge role in the life-cycle of a retail product being manufactured. Starting from the procurement of raw material to manufacturing/ operations and then logistics. Effective management during the lifecycle, helps planner to do an effective pricing of a product and also helps them decide, whether/ where to stock the product.

Speaker gives an example of application of data science in life cycle of apple products. Company implement **Visualization techniques** to understand the stock level of raw material/ finished products and implements **Machine Learning techniques** to predict how the quantity of goods required. This helps company make important decision such as how much raw material is required, from where should company source raw materials, how much space is available to stock the goods based on consumption and shell life of a product. Another example that speaker gave is of lifecycle of cotton clothing, based on which company decides cost of raw material, logistics, packaging, marketing and storage requirements.

<u>Impact on Retail Companies</u>: -

Implementing data science techniques in Life Cycle of manufacturing retail product helps optimize the supply chain network thereby reducing the cost in every domain of the life cycle. More the network is optimized, less expense it is and more profit generated in the end.