Final Report

EAS 504

Anish Shah

A) Experiences:

- Before the course started, didn't have much idea of how the course will be, just knew the structure. But after completing this course last week, I can confidently say that got to a learn lot of new things which you don't get to know from the people who actually work as a data scientist.
- Learnt about the legal aspects of data science from a person who actually engages in litigation and legal advocacy on issues such as technology and privacy, which can't be learned in technical education until you step into industry or deal with legal matters.
- Various new disciplines such as Computational Advertising about which there is little knowledge available on how and what exactly goes on inside, and how complex the architecture is, was a surprise to know.
- We learnt many things such as how we should identify a business case before deciding on a solution.
- We should try different tools and understand their limitations before settling on a single tool.
- We should let data tell its own story and how machine learning methods are useful for learning structure.
- Interpreting and communicating the structure is the final step in the process.

B) Carefully, draws out commonalities across lectures and differences. Please be specific and cite the lectures when you bring out these commonalities.

- Each lecture gave a different perspective on various applications of data science. Every lecturer had a very different and diverse background. Best thing to know was how everyone was passionate about more and more data and that really was inspiring to know to what it takes to be successful in this area of work.
- Each company or rather two people from the same company(eBay) also had different types of problems, and use cases specific to different domains.
- Every company had almost same core values related to trust, competence, being together, looking for solutions and not scapegoats, being adaptable and ready to change, etc. which really is a foundation of a good company.
- Every lecturer gave us an idea that we should be passionate about what we're working on so that work is an enjoyable experience.
- Skills such as Natural Language Processing(NLP), deep learning, Text mining were the most used techniques in all the lectures.
- Programming language being used didn't matter as similar jobs were carried out in different programming languages by different companies. For example, lecturers from Lecture 1 used MATLAB for model development, R was used lecturer from Lecture 2 for building their models and many others used Python to do the same.
- Spark and MapReduce were the most common data processing tools used.

C) Comment on ideas introduced in the cross-cutting theme lectures e.g. lectures on trust, privacy, ethics and legal aspects.

- Data science and "big data" are reshaping many aspects of commerce and society. We got an overview of the major ethical and legal problems data scientist are likely to encounter.
- A data scientist working on some projects such as criminal injustice has to be very careful while developing their models and presenting the results of the same. The algorithm may exacerbate inequalities and disparities that already exist on the police force. For example, Machine Bias. Machine Bias takes various forms. One of the most prominent examples involves the use of machine learning systems to make judgements about individual people or groups of people. Many times, when used in the field of criminal justice, some machine learning models have been shown to assume higher crime rates for individuals based on superficial data such as ethnicity or location.
- Social media algorithms are designed in such way that they keep you hooked on to their website to generate more revenue. But sometimes this same algorithm could cause polarization of views for the user as the algorithm keeps optimizing and showing the user what he previously saw, eventually influencing his view about a certain topic instead of exposing him to different views on the same topic to get a broader idea. We as data scientists or developers need to be very careful about such issues as for short term it might reap benefits but in long term, it may cause public outrage. For example, Huge outrage was caused over Facebook when people thought it was influencing their views for voting a particular voter in an election.
- Data science methods involve human decisions at every stage. So sometimes those decisions can encode bias inadvertently or maybe intentionally.
- While doing a feature analysis, some features in the dataset may be 'proxies' for protected characteristics, because they correlate highly with those characteristics. ML/ statistical processes inadvertently reintroduce influence of characteristics like race, sex, orientation, gender even though they aren't in the data set.
- To overcome these challenges or to avoid them we should not intentionally treat people differently on the basis of a protected characteristic, we should make sure our analytics are closely linked to a legitimate business need, we should sometimes talk to the on-ground experts and identify the potential oversights and we should be wary of replicating and amplifying unfairness or discrimination encoded in the dataset.
- Privacy in our context means control over information about oneself. Most data science methods require lots of data about people if you're making decisions about people. The responsible people should be able to protect this data and avoid breaking any privacy laws which are taken very seriously in this age of technology.
- Some of the ideas that need to be practiced for maintaining the privacy of the users are
    - Notice: where you inform the user of what is being collected and how it will be used.

    - Consent: The user must agree to the collection and use of the data being collected.
    - Access/Correction: The user should have access to his data that is being collected so that if there are any errors he can appeal.
    - Integrity/Security: This data should be kept safe and accurate, and secure against a breach.
    - Enforcement: There should be an effective mechanism for enforcing these rights

- Other important ideas to follow for data protection are
    - We should take the privacy policies seriously.
    - We should think hard every time about whether we actually need to collect that data and for what reason before starting data collection.
    - We have to make sure that users can access, delete or correct their own data.
    - Private data of users should be encrypted all the time.

D) Pick 3 case studies from different lectures and write in some detail about each bringing out the data science methodologies and impact on the organization.

1. Prognostics for Low-cost Devices:
    - The main idea behind the use case is to collect the streams of data from the machines and use it to draw some insights using Machine Learning.
    - Survival analysis which is used by Insurance companies is an application of this use case.
    - One approach used in this use case was to screen all machine faults for association with service calls. Identify relevant predictive indicators for device subsystems and service call problem descriptions.
    - Another approach used was to coerce the problem into a classification problem for predicting the likelihood of a service call within a certain time window with the use of the random forest classification algorithm.
    - Random Forest model was a good final model as it gave higher accuracy (80%).
    - Techniques such as Random Forests, Gaussian Mixture Models, Decision Trees and other regression models are also used.
    - Data Processing was carried out using Spark. R language was used for building the respective models.
    - Data and computing methods used in the organizational flow really depends on the type of data (speech, tabular, text) we have.
    - Neural networks are used on tabular data, CNN's on image processing related tasks, Long Short-Term Memory on speech and text data is used.
    - BPS typically deals with core businesses. Instead of becoming third party service provider they kind of become your business partners.

- BPS focuses on how to add significant value, or direct business impact, beyond the bottom line. It also focuses on optimizing how knowledge workers and advanced algorithms can be combined to automate key operational activities.
- As the speaker mentioned the health care industry is going to have an exponential of growth with projected market value being over 5.5 Trillion by 2030. So, there is a huge opportunity for growth here as the usage of BPS will also grow with the industry.

2. Price Optimization:
    - Dynamic pricing allows large and small companies to improve their margins quickly. The Price here is flexible based on demand, supply, competition price, subsidiary product prices. Prices may even change from customer to customer based on their purchase habits.
    - Dynamic pricing enables suppliers to be more flexible and adjusts to prices to be more personalized.
    - There are various benefits of price optimization such as more precise SKU level prices, faster response to demand fluctuations
    - Line search algorithm. Linear programming equation, constraint optimization is used to put an objective function to maximize your revenue where we find optimal price to sell a product to a customer and drive the sales.
    - This domain of e-commerce retail uses various Machine learning, AI and deep learning techniques throughout the product lifecycle, STL decomposition for statistical forecasting, A/B testing for reminder emails. Technologies include data parallel processing tools such as Spark and native Map Reduce jobs.
    - The data science prediction function is used for better inventory management. Inventory management also consists of forecasting, demand planning, demand modeling, machine learning.
    - As more and more people are taking their buying habits online, the e-commerce industry is growing at a fast pace. The more people spent their time on these websites the more data we have to work on. So, with this increasing data, we can build better models and recommender systems in our existing system. Better algorithms and AI techniques will result in better accuracy which will, in the end, give us more users getting converted to buyers resulting in growth in revenue.

3. Aftermarket Demand Forecasting:
    - The first step in developing a predictive model for aftermarket demand forecasting is establishing demand. We also got to know that the forecast is always wrong.
    - Time Series for Demand was created using the actual customer contracts data.
    - For building the predictive model Bayesian Machine Learning techniques such as Gaussian process was used. Using this algorithm, a structure was learnt from the data.
    - The posterior sample obtained using the model seems to predict identical results to historical data which tells us about consistency.

- MOOG's manufacturing and procurement workflows were reviewed and business data related to Shipment/Order information, Part Information, Manufacturing/Work Orders, Inventory, Vendor/Purchase Orders was identified and mapped to support the project objectives.
- Skills such as Exploratory Data Analysis, Compromise Programming, Bayesian Machine Learning, Gaussian Process Learning and Extrapolation, Bayesian Network, Text Mining for finding Word Co-Occurrences and Topical analysis using latent Dirichlet allocation are used.
- Technologies such as MATLAB for mathematical calculations and model development, SQL for understanding and preparing the data for analysis using SQL queries and data aggregation is used.
- Collecting more data as the current models were just trained on the 17 months data and training the models on the larger dataset so that we can see how the model accuracy changes from the previous version and therefore making changing and keep developing new models to improve accuracy and insights.