EAS 504 - Applications of Data Science : Industry Overview
Assignment 2

Name : Aboli Rawanhate
UB Number : 50374341

Q1. Describe the market sector or sub-space covered in this lecture.

Computational advertising was the market segment covered in Sharat Chikkerur's presentation. Computational advertising is a new field that combines computer science, economics, and machine learning. Advertisements account for 90% of the revenue of firms like Google and Facebook. As a result, questions such as which advertisements to display the user, who should see the ads, and when should the ads be shown arise. Informational retrieval would determine when to show which ad, and a recommendation system would select the target demographic for commercials, among other things. This industry is well-known for its focused and individualized advertising, which was previously unheard of. Search ads and display ads are two types of advertisements. Search Engines has dominance on Search ads whereas 50% ads come from display ads which are mixed in content. Ads on the internet and on mobile devices are on the rise. Advertisers who want to display their advertising and publishers who want to display them on their platform have developed a network. Dynamic pricing has been implemented based on parameters such as when advertising is displayed and the duration of the content, among others. So, it is the major reason responsible for making machine learning mainstream.

Q2. What data science related skills and technologies are commonly used in this sector?

This industry also employs a wide range of data science skills and technology. In this field, anything from a simple classification method to a complicated algorithm can be employed to address problems. The abilities of Data Science will be determined by the sector in which it will be used.
Below problems will be solved by using different Machine Learning techniques.-

1. Target audience
2. Ranking Users
3. Bidding
4. Optimisation
5. Budget and Pacing
6. Dynamic Pricing
7. Personalize content
8. Latency

- Generalized linear models such as linear, logistic, SVM, and quantile regressions are used for value estimate and audience targeting. To evaluate bid, for throughput and latency it is used.
- Recommender systems are used to present adverts in a tailored manner.

- Distributed computing is employed when dealing with a large audience on the internet. Big-Data technologies such as Hadoop and Spark are being implemented in real time.
- Information retrieval - When should which ad be presented, i.e. when there will be a smaller target audience and when there will be a larger audience?
- For tailored ad display, recommender systems are utilized.
- To group similar algorithms, clustering methods such as latent Dirichlet allocation or k-means are used.
- Text mining and search advertising use tools like TF-IDF, wor2vec, and Bag of Words.
- LDA variational approach is used when we don't know the topic distribution for a document then we find the best solution for the solvable problem which is close to the really complicated problem.
- In LDA, the value of alpha is adjusted according to the document specificity. Example, a document with a large number of topics will have high value of alpha else alpha value will be less.

Q3. How are data and computing related methods used in typical workflows in this sector? Illustrate with an example.

Data and Computing related methods :

- For throughput, latency and bid evaluation Linear/Log linear or Poisson Regression is used.
- For data with outliers, quantile regression is used instead of linear regression.
- Feature hashing used to reduce the dimensions of sparse features. Hash is one-to-one so it reduces sparsity.
- Ranking of the ads calculated using expectation formula using price per click and probable cost per click.
- Vowpal Wabbit(VW) is adaptive and for every dimension it will compute square gradient so we don't have to normalize.
- It scales based on previous gradient value.
- VW is useful for all dynamic ranges. It has namespaces and connections which makes indicating cross product simple.
- VW doesn't load large amounts of data in memory like scikit learn so efficiently.

Some entities involved in the workflow :

- Advertiser - Agencies or businesses who want to buy adverts to show to users.
- Publisher - A company who has access to a large number of people through platforms such as search engines or any website with multiple users.
- Demand Side Platform(DSP) - Serves as an agent for advertisers, ensuring that their ad spend is optimized. Similar to Robinhood's automated trading (e-trade). They will speak with many exchangers in order to obtain the stock at the specified price.They are likely to precisely value every possibility to make money by showing ads using machine learning.
- Supply side platform - Technology businesses that understand the marketplace and work with publishers to maximize income are known as supply side platforms.

- Hedge funds/brokers at wall street responsible for picking stock, entering or existing stock to maximize return
- Ad networks - A place where publishers subscribe to ad networks, which are open to all advertisers and allow them to bid on any ad.
- Ad Exchanges - A platform where buyer meets seller.
- Data Aggregator - Collects data from all the sources and analyzes the opportunity to get more from advertisement.

Q4. What are the Data Science related challenges one might encounter in this domain

- Finding a best match between a given user and context and the pool of advertisements.
- Basic problem is determining the target audience (who to show advertising to), the target platform (where to show ads), and the amount of money to invest (how much to pay for ads). The goal is to find a good match between the user and the advertiser in a specific situation.
- Pricing (Auction) - In an ideal world, a VCG auction would be employed to make it fair for all bidders, but it is difficult to solve and costly; for 'n' bidders, the computational cost of optimization is the square of the dimensions.
- Quality - Ad blindness and ad diversity are examples of subjective quality. When commercials become repetitious, the audience may overlook or skip them, thus it's crucial to know when to show them. This is known as ad blindness. Another option is to use a utility function to show ads only when income exceeds cost. Instead of exhibiting the same repeated ad, advertisements diversity involves showing a variety of commercials to attract the audience's interest.
- Bidding and value estimation - The best strategy to bid is to bid what the publisher's value is worth in the long run. This problem can be tackled by using general setups to develop machine learning models such as linear, log-liner, logistic, and poisson regression, based on the available data and desired value estimation.
- Budgeting and Pacing — The primary purpose is to disperse spend in order to maximize income, and to do so, a bid computing feedback system such as a PID controller or a water-level based controller is utilized to change bids over time.
- Ranking and Targeting (who and what should be shown) – Targeting can be thought of as a search issue in which we strive to locate the best users for a specific ad based on its features and context. Ranking is the process of selecting the best ad for a particular user and context; the best choice is the one that earns the most money; the most common metric used is a function of revenue per click and cost per click.

Q5. What do you find interesting about the nature of data science opportunities in this domain?

The most intriguing aspect is that the use of data science in this subject is unrestricted. In this industry, different approaches to different difficulties are taken. We've already looked at the various machine learning techniques utilized in this industry, such as linear regression, Poisson regression, quantitative methods, and so on. The employment of these algorithms opens up a plethora of possibilities in this subject. I found factored value formulation to be really beneficial.
Big firms like Facebook and Google, which are leaders in this field, collect a massive quantity of data in order to offer the best ads. Working with such vast amounts of data necessitates a great deal of problem solving, which is their primary source of money.

The transmission of data science knowledge and skills leads to increased applicability as the sector grows. The computational advertising market is quite huge and developing fast, as ads account for about 90% of revenue for key platforms like Google, Facebook, and others. Various data science approaches are utilized to solve challenges such as targeted distribution, personalized content, dynamic pricing, and value estimation. Vowpal Wabbit is a well-known library among computer advertising, and it was created to handle any difficulty that a firm might have. On top of these algorithms, real-time bidding strategies are utilized to acquire advertising in order to maximize income. VW implementation is also very easy.

Q.6 The roles of Demand Side Platforms, Supply Side Platforms, Ad Networks and Ad Exchanges and how data science plays a role in online advertising.

Demand Side Platforms(DSP) - As an agent, he is in charge of maximizing a company's ad investment. They are likely to precisely value every possibility to make money by showing adverts using machine learning. They are compensated based on how well they perform, thus the better they perform, the more they get paid. They must ensure that the predicted return is greater than the cost of the investment. Ads are designed to reach the greatest number of relevant users possible. As a result, the revenue and DSPs bid for that specific user, determining the amount at which the ad should be priced. DSPs enable programmatic advertising when used in conjunction with supply-side platforms.
Supply Side Platform (SSP) - Technology businesses that understand the market and collaborate with publishers get the most money. Companies that collaborate with publishers to optimize income know the market and where we can sell advertisements. On the publisher side, a supply-side platform connects them to ad networks and exchanges, which connect them to demand-side platforms (DSP) on the advertiser side.
Ad Networks - Advertisers and publishers use ad networks as a conduit between them. They can be described as internet platforms that gather revenue from publishers and organize it into various formats. Buyers can communicate their needs to the ad network, and the ad network will match those needs to the publisher's inventory. Some ad networks have a very specialized focus, such as inventory scale or cost, or solely a specific demographic of users.
Ad Exchanges - This industry's marketplace is the ad exchange. It is an online marketplace where both supply and demand individuals may directly buy and sell inventories. On any network, there is no need. Ad Exchange auctions its inventory directly using Real Time Bidding technology.

Data Science role in online advertising:

The application of data science to online advertising is improving the power of decision-making. When we examine the computational advertising process, we can observe that each stage requires some sort of judgment.

We must make decisions at every level, from targeting a certain audience with likes and clicks to demographics. Demand Side Platforms (DSPs) are used to optimize spend and are responsible for appropriately valuing each ad in order to generate revenue.The term "supply side platform" refers to technology companies that assist publishers in maximizing income by selecting the best bidding value. Whether it's a Demand side platform or supply side platform, ad exchange, or ad network, each function needs to be able to make decisions while on the job.

One of the most essential roles of data science is the use of machine learning for designing real-time bidding algorithms, monitoring client behavior, and generating money in online advertising. Recommendation system does the user advertisement mapping.

Q.7 Comment on the role of stochastic gradient methods in ML applications

Gradient descent selects a location at random and optimizes an algorithm by minimizing loss and updating weights after each iteration. Because this can be exceedingly computationally expensive when dealing with large amounts of data, stochastic gradient descent is used. SGD updates parameters using approximate gradient estimates derived from sections of the training data. Because subsets of training data are used, the computational complexity is reduced, and the issue solution is optimized. Stochastic Gradient is commonly utilized in SVMs, regression problems, and other applications. Random probability is stochastic. As a result, in Stochastic Gradient Descent, one data sample is randomly chosen for each iteration, and weights are calculated for that single sample. It will minimize the computation. Instead of validating all training data to get the direction of slope, we can analyze and reach a global minima.

Q8. Multiple choice questions -

Q1. E
Q2. D
Q3. D
Q4. A
Q5. E