How Facebook Applies Probability and Statistics to Empower its Features

01006719 Probability and Statistics I

Software Engineering Program,

Department of Computer Engineering,

School of Engineering, KMITL

By

65011367 Music Auyeung

Table of Contents

1. Introduction

Can you recall the latest product you bought? If so, how did you happen to learn about it? Chances are, it appeared on your social media feed, something which happens more often than one might think. Especially in present times where almost everyone and everything can be electronically connected, social media presence and publicity has seeped into the essence of relationships, marketing, business. Companies, influencers, and ordinary people are all driven to develop the abilities to incorporate their visions and creations into the lives of others, and what better way to do this than on social media. With approximately 2.989 billion users, Facebook is an example of one such platform to be taken advantage of. However, in the vast sea of online presences, how does Facebook ensure that it is able to effectively cater proper offers and content to everyone. The answer is through its carefully crafted algorithm. With probability and statistics Facebook designs and implements formulas, logistic regression models, and deep learning to manage promotional techniques and maximize reach for well-received content creators by focusing on conversion rates.

2. The Role of Ad Spend and Conversion Rates

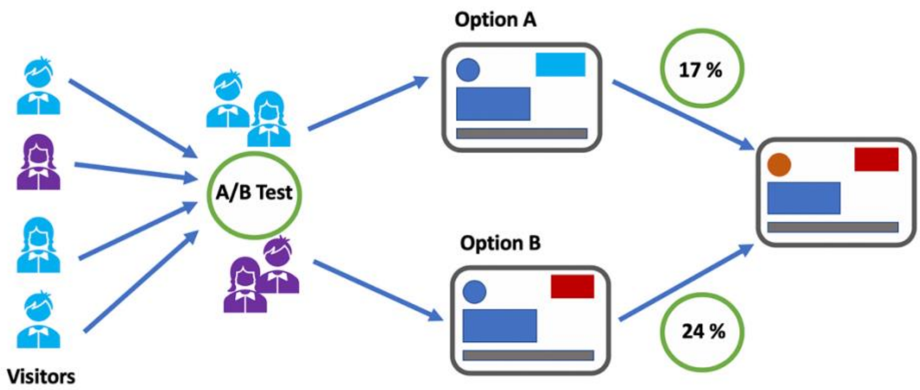
To explore how Facebook applies these systems on the platform, the concepts need to first be understood. On Facebook, advertisers or influencers have the option to pay Facebook to push their content into user's feeds and this payment falls into the category of ad spend. Ad-spend is the amount of money or resources spent on an advertising campaign which, in the modern world and on social media, may take the form of commercials, posts, or endorsements. The first step in reeling in customers is getting their attention which is why publicity is an important investment for companies. But to generate profit, social media creators cannot disregard the notion of cutting costs, specifically unnecessary costs, and ad spend is not any different. Which is why a social media page, or a business must incorporate statistics into allocating its resources in a reasonable manner. These same statistics come into play with Facebook's access to data stores and probability models. Other than ad spend, content creators also must focus on a target audience. And when it comes to Facebook, self-promotion not only performs based on how often a post is shown, but also who it is shown to. And much like ad spend, the decisions regarding these factors are influenced and calculated by statistics. Say, if an advertisement was broadcasted to every single user on Facebook, how much traction would it cause? In other words, how many people would not only view it, but interact with it? How many would feel inclined to click, or like, or comment, or share, or even make a purchase? This is measured as the conversion rate of a content. If the said ad fails to spark any attention or conversation whatsoever, could paying for the ad be considered worth it? Likewise, if an influencer's post didn't do well in terms of reception, the creator would most likely have to re-evaluate. Because when the conversion rate of a post is extremely low, creators would be paying for it with valuable resources, like the time and effort that could've been devoted to another task.

With the established significance of these thoughts, how exactly can the conversion rate or ad spend be predicted simply from statistics?

3. A/B Testing and Bayesian Bandits Paradigm

3.1 A/B Testing to Compare Ad Performance

If the question is of which advertisement has the potential to perform better, the simple solution seems to be to directly compare the performance of two alternatives to determine the more well-received one. This is also known as "A/B Testing". Show two ads and compare how much interaction each one receives. Using the probability of the ads doing well or otherwise seems reasonable enough for the process of selecting a preferable option. However, recall that the results are still based on a probability, which at the same time, implies its uncertainties. For example, according to A/B testing results, nine times out of ten, ad A will perform better than ad B. But the probability of ad B performing better still exists. Trends and preferences are constantly changing over time, now more than ever, with the rise of consumerism. Completely disregarding even, the "losing" probability can still mean lost opportunity which is a risk that doesn't have to be taken. Because instead of picking one ad to show over the other or showing both equally, alongside the straightforward statistics, Facebook also takes into consideration crucial prior information as it adopts the "Bayesian Bandits Paradigm" to forecast the relevant effectiveness of the released content.



(https://towardsdatascience.com/how-to-conduct-a-b-testing-3076074a8458)

3.2 Introduction to the Bayesian Bandits Approach

The Bayesian Bandits approach offers an upgraded alternative solution to the problem. Rather than picking one ad and discarding the other, instead, consider their success rates in comparison to one another and show them accordingly to the compared rates. In simpler terms, if an ad does better than its competitor, instead of showing only this ad, simply show it more than the option deemed inferior in performance. So, if the ad does substantially better than its competitor, show it substantially more. Likewise, if it does only slightly better, then show it only slightly more. To measure exactly how much "better" ads perform, Facebook needs to collect the key performance indicator of each ad. In other words, how well it is doing in the selected field of priority. For instance, if a business was looking to make purchases, the priority may be the number of purchases per thousand impressions, also known as the PPM. Given the two ads have the required threshold number of past purchases such that the

PPM could be accurately computed. Then, the ads could be shown more often or less often based on the resulting comparative PPM ratios.

4. The Relevancy Score Algorithm

Now that may sound easier said than done, how can a social media platform or the software behind it completely comprehend the human concept of what it means to prefer one thing over the other? Moreso, how can it be done time and time again for every single user of the platform with its broad and expansive population worldwide, with different personalities, cultures, backgrounds, and all the abstract ingredients that make up a person and their passions. After all, it's not like Facebook is concerned with discovering a user’s latest movie obsession, or favorite cartoon character, or group of best friends, is it? That’s the kind of information you gain from casual conversations at the dinner table or from seeing an acrylic keychain bought at a merch store hanging from a backpack. But that’s also the kind of information that guides a birthday gift purchase for a friend or the latest release to indulge in, in other words, it’s the information that can determine exactly what people are most likely to buy into, literally and figuratively. Which is why, whether it is done unknowingly or fully aware, that’s exactly the kind of information that users constantly hand over to Facebook themselves. Since it’s not like the employees of Facebook are interviewing users one by one. Even if that were the case, it would be much more time consuming. One indisputable fact is that Facebook holds a vast collection of statistics and information on its users. With the modern advances in technology and data analysis, Facebook can handle these very details and data to predict which content can be benefited from according to the latest trends or the long-time ideologies that continue to seep into seemingly superficial social media interactions.

4.1 The Relevancy Score Formula

In such a manner, Facebook engineers were able to design an even more specific "Relevancy Score" algorithm represented by this formula:

$$\mathbf{R(u, c) = \alpha * I(u, c) + \beta * P(u , t) + \gamma * E(c)}$$

In this formula, R refers to the content's "Relevancy Score", u refers to a specific user, and c refers to the content of interest. Symbols α, β, and γ, represent the weights which determine the relative importance of each signal, and the weights are developed over time using information obtained from both the user's own habits and past interactions on the platform as well as the general trends throughout the entire platform's community. α serves as a factor to I(u, c) which describes the user's interaction with the content's source or creator, which may be the user's friends, pages that the user follows or frequently views, or groups that the user has joined. The importance of this is that it gives a general idea of the user's relationship with the content's origin which may be an indicator of its credibility or importance to the user. P(u, t), where t stands for the type of content in this case, is multiplied by β and serves as representation of the user's preference to the type of content. Again, this is based on user, u's, past interactions with contents of the type, t. Finally, with γ as a coefficient, E(c) which denotes the engagement score of the content is determined by the different other users' interaction with the content. Liking, commenting, or sharing the post indicate the popularity level and quality of the post from the general communities' perspective. Moreover, it is also considered whether the user has a certain relationship with interactors. With these three ranking components, the algorithm decides which content is worth showing than others.

4.2 The Feed and Reels Formula

Another example of the same practice can be seen in the Facebook Feed and Reels Algorithm. Although, it shares a similar goal with the previous formula. This mathematical algorithm serves the purpose of determining the order in which contents are presented on each user's feed. The general formula goes as follows:

$$\mathbf{R\_feed(u, c) = \alpha * I(u, c) + \beta * P(u, t) + \gamma * E(c) + \delta * I\_feed + \varepsilon * P\_feed}$$

While the formula follows in the footsteps of its predecessor with the familiar ranking signals, there are several additions to the algorithm. The additions to this formula are the symbols δ and ε, corresponding coefficients of the variables, I_feed and P_feed. The former, I_feed refers to the inventory components, while the latter represents the predictions components which are based on logistic regressions models as well as neural networks which are both part of machine and deep learning which will be explored later.

Generalizing the formulas can provide an overview of the algorithm, however, what does it all mean in practice? A Relevancy Score is presented to an advertiser as a rating from one to ten, which can serve as an estimated measure of performance according to the responses of targeted audiences. And as mentioned previously, the metrics are only accurate once a certain number of audiences have been reached. In this case, that number is 500 as the Relevancy Score details only become available to a Facebook advertiser once the ad receives over 500 impressions. Since the Relevancy Score is a number scale from one to ten, if the score is above the average score of five, the advertiser can benefit from a lower cost per click cost to be paid to Facebook or a higher click through rate which means the ad would be shown more often.

5. The Quality Ranking Scheme

And as the world develops in technological and analytical skills, Facebook has endless room for growth. In 2021, it was established that the Relevancy Score system was to be left behind, with Facebook instead opting for the Quality Ranking scheme. Like its ancestor, the Quality Ranking seeks to determine the quality of a content based on how well-received it is amongst audiences, this time with the metrics: Quality Ranking, Engagement Rate Ranking, and Conversion Ranking.

5.1 Quality Ranking, Engagement Rate Ranking, Conversion Ranking

The Quality Ranking metric is a representation of the quality of an ad compared with the quality of related ads which are aimed towards a particular demographic. Likewise, the Engagement Rate Ranking compares the content's engagement with the engagements of similar contents with like audiences. Lastly, the Conversion Ranking of an ad, like its name suggests, is measured by the number of conversions that a content can generate as opposed to the number of conversions that other similar contents can generate. And this time, the criteria are divided into Above Average, Average, and Below Average. If the content's performance falls into the bottom ten to thirty-five percent of all ads of the same category and target audience, the content is considered Below Average and will be treated poorly compared to its fellow ads.

6. User Data and Personalization

In previously mentioned algorithms, target audiences are an important key to consider when evaluating the quality of a content. Facebook determines this information about a user to personalize the program, but it also uses this information for targeting contents that are likely to do well at users. This information includes but is not limited to posts which the user has liked, comments the user has made, the user's interaction with content creators, friends, followers, pages, and groups. Facebook also gathers personal information such as age, gender, and location.

6.1 Use of Demographic Data, Preferences, and Past Interactions

All this information is used to train a logistic regression model, which predicts how the user will respond to content based on their demographics. From the given example formula:

$$\mathbf{p(y=|1|x) = |1| / (|1| + exp(-(b0 + b1*x1 + b2*x2 + ...\ bn*xn)))}$$

where p(y = |1|x) denotes whether or not a user is likely to support a political candidate, which can also apply to representing whether or not a user is likely to respond well to a certain post. x1, x2, and x3 stand for the user's demographic information, age, gender, and location, respectively. And b0, b1, b2, b3 are the values that have been calculated by the logistic regression model which has to be trained beforehand by the information store provided to Facebook by users. Like the coefficients in the Relevancy Score formula, these indicate the importance of each factor accordingly. Again, this formula resembles another formula for feed ranking because it takes knowledge about the user into consideration in the equation and employs the use of neural networks. The formula combines multiple predictions to come up with a single resulting score, V_{ijt}:

$$\mathbf{V_{ijt} = w_{ijt1}Y_{ijt1} + w_{ijt2}Y_{ijt2} + \ldots + w_{ijtk}Y_{ijtk}}$$

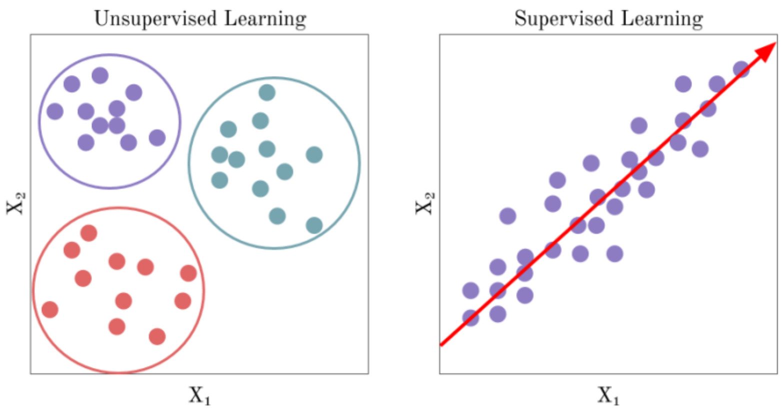
V_{ijt} is the combined prediction value and Y_{ijtk} is the prediction value of each post based on the post characteristics such as the type of post or the relationship between the post's origin and the user. w_{ijtk} serves as a weight of sorts for measuring the importance of a user's response. For example, if a user is historically more likely to like, then liking is considered to have a higher weight than say, commenting or sharing.

7. Deep Text and Language Understanding

As shown in the above examples, other than logistic regression models which can be trained and developed using data, Facebook also uses neural networks and intelligent computer systems. Another such way that Facebook uses this is through development of Deep Text. The purpose of the Deep Text is to be able to accurately understand human language with all its depth and nuance. This doesn’t refer to having the meaning of words or phrases, but also means being able to interpret the meaning of colloquial language and context clues. In social media terms, Deep Text is another way of gathering information by studying several thousands of posts and comments simultaneously and continuously.

7.1 How Deep Text Develops Understanding of Human Language

Deep Text is developed using unsupervised neural networks that digest unlabeled datasets to understand texts found in posts and comments. Since the datasets are unlabeled, the learning engine must self-learn the meaning of each word, each phrase, and each thought it comes across through experience and trial-and-error. Additionally, the neural network uses word embeddings to link related terms together, resulting in clusters of information which provide context for each other in practice. Although this process takes more time compared to supervised learning with predefined labeled datasets, it grants the learning engine freedom to explore, resulting in the capabilities of understanding the user's communication as well as the sentiments present between the lines. Deep Text can be used to collect user data as it can interpret the user's expressed thoughts and opinions which can be inferred towards preferences or inclinations to be applied in previous examples or it can be used for the automation of management such as removal of questionable content such as explicit content or hate speech which are direct violations of Facebook's policies.



(https://towardsdatascience.com/a-brief-introduction-to-unsupervised-learning-20db46445283)

8. Summary

To sum it all up, Facebook not only uses probability and statistics to learn more about users for several purposes such as targeted advertisements, feed personalization, and experience cultivation but also use logistic regression models and neural networks to develop learning engines which are able to master human language with accuracy and automatically enforce policies with these engines. This shows that although seemingly static, given the proper techniques and usage, data, probability, and statistics can become flexible tools that know no bounds and will indisputably continue to be incorporated into technology and human life.

References

- https://mobiledevmemo.com/bayesian-bandits-behind-the-scenes-of-facebooks-spend-allocation-decisioning/
- https://www.linkedin.com/pulse/mastering-facebook-algorithm-comprehensive-guide-insights-patel
- https://www.digitopia.agency/blog/understanding-facebooks-new-quality-ranking
- https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/
- https://www.networkworld.com/article/3077998/understanding-deep-text-facebooks-text-understanding-engine.html
- https://www.thegazette.news/features/how-algorithms-will-change-our-lives/
- https://towardsdatascience.com/a-brief-introduction-to-unsupervised-learning-20db46445283