Summary:
"150 Successful Machine Learning Models: 6 Lessons Learned at Booking.com"

Alicja Wilk
2/12/2019

6 main points we learned from Booking.com about Applied Machine Learning & e-commerce.

There have been plenty of breakthroughs in Machine Learning in the last few years. Today we will discuss how Machine Learning can impact the e-commerce business. This summary touches on what is not talked enough in AI and ML today – how can the business value be driven in the real world?

Booking.com is one of the largest travel agent websites. I highly encourage to go and play with the website a little before proceeding further. Now, let's dive into the paper!

In the introduction the authors talk about Booking.com's challenges on their website:

For example, if your stay is unsatisfactory at the destination, there is not an easy way to undo the reservation once you arrive (**high stakes).** While searching, users barely specify exactly what they are looking for (**Infinitesimal (math.so small, there is no way to measure) queries).** The multi-dimensional space of bookable items is not trivial to navigate (**complex items).**

Dynamic pricing model and limitations in accommodations cannot be neglected while designing machine learning models (**constrained supply).** Also, people often only travel once or twice a year which means that it is hard to provide that personalized UX **(continuous cold start challenge).**

There is a huge amount of content from hotels and guest reviews, the ability to sort through all the data and successfully exploit the content in an accessible way with information that is relevant to the end user is crucial **(content overload).**

The authors note multiple times that driving actual business impact in ecommerce through ML is tough, and many before them have noted that there are not enough papers where real results were achieved.

The first part of the paper talks about the inception phase in Machine Learning.

Inception phase involves building the business cases, hypothesis, product ideas and how ML fits as part of the possible solution.

- There are models that fit a specific spectrum (optimize an element of user interface) and due to this specificity, we can tune the models and have big business impact that is limited to a few use cases
- On the other side there are models that act as semantic layer and give involved stakeholders of the organization platform to add new features and generate personalization

On average, sematic models generate twice as many use cases as the specialized models. Machine Learning Models Families to address the challenges:

- ***Traveler Preference Models*** *– measures flexibility of a user and returns page results accordingly (semantic layer models)*
- ***Traveler Context Models*** *– discovering the context/purpose of a trip as early in the process as possible (semantic layer)*
- ***Item Space Navigation Models*** *– navigation on the website (scrolling, clicking, viewing etc.) is treated as guidance to help the user in facilitating access to relevant items*
- ***User Interface Optimization Models*** *– deciding on best user interface for a particular user (font size, background colors, images, presence vs. absence of a visual element)*
- ***Content Curation*** *– content comes from many sources like texts, reviews, photos etc. and is posted by many users, giving us vast and complex space. This model makes the content accessible to users. The model 'curates' the content and translates into a summary.*
- ***Content Augmentation*** *– using existing content and augmenting the service offer based on given attributes. Examples: Great Value: highlights properties offering outstanding value compared to usual and Price Trends: informing the user based on trends if today is best time to book*

All families of the models provide a positive contribution and value to the business collectively and separately. Often models become a foundation for another product. All models showed positive impact relative to the median impact.

Part 2 discusses quantifying the model performance.

The value that the models bring is estimated using Randomized Controlled Trails (RCTs) and business metrics like cancellations, customer service tickets and conversion.

Finding: Increasing performance does not necessarily translates into increase in value. There is no correlation between offline performance gain and business value gain. This interesting discovery can be explained by:

- **Value Performance Saturation** – at some point the business value vs. model performance curve saturates
- **Segment Saturation** – the population of the treatment group of users decreases as we test more new users that are exposed to the change.
- **Uncanny Valley Effect** – the more the model knows the 'creepier' it becomes to some users
- **Proxy Over – optimization** – increasing click through rate but not conversion, as user clicks more through different options there are too many options and user is left confused.

Part 3 of 6 talks about designing a problem before solving it.

The problem construction process takes business case and outputs a modeling problem. Usually we need to construct target variable and observation phase. The authors mention that *Learning Difficulty* and *Selection Bias* are often encountered in this stage and must be addressed. At this stage models can be iterated over into more complex or improved upon. Very often the problem is uncovered AFTER changing the set-up of the model, and the true value is unraveled.

Part 4 talks about how we must minimize latency and the time is money principle during deployment.

High latency is not accepted well by users and waiting longer than 3 sec for website to load can create frustration and disengagement. Some techniques used to minimize latency:

- ◊ Model Redundancy
- ◊ In- house developed Linear engine: highly tuned to min. prediction time: Naïve Bayes. GLMs, k-NNs, Matrix Factorization.
- ◊ Sparse Models – the less parameters the less time to compute

Part 5 talks about the monitoring stage and red flags

Challenge of output monitoring is having ***incomplete feedback:*** even if a prediction was made for an individual (whether or not they will add a special feature at the end of the booking). Unless they actually book, we will not know the true label of the prediction or in other cases, we will not know for days which creates ***delayed feedback.***

Figure 7: Examples of Response Distribution Charts

The authors note that metrics like *confusion matrix* are not appropriate. To address quality of the model Booking.com uses Response Distribution Chart (RDC).

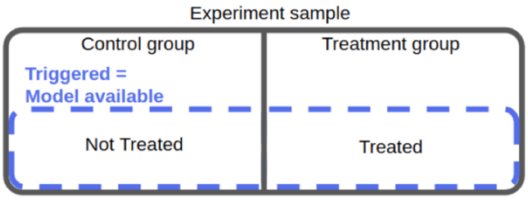

Figure 8: Experiment design for selective triggering.

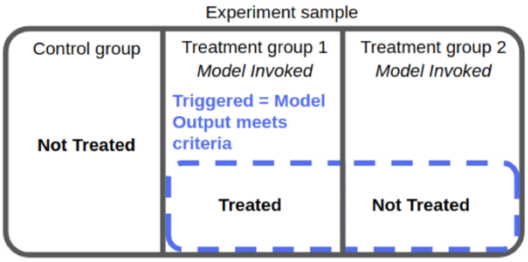

Figure 9: Experiment design for model-output dependent triggering and control for performance impact.

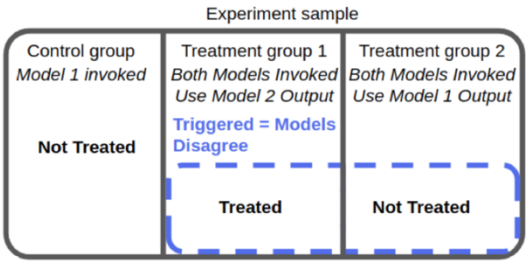

Figure 10: Experiment design for comparing models.

The takeaway: if "a model cannot assign different scores to different classes then it is most likely failing at discriminating one from another, small changes in the score should not change the predicted class".

Part 6 talks about the evaluation phase and best practices.

Booking.com has built their own experimentation platform of Randomized Controlled Trails where everyone can test hypothesis. In the evaluation phase, Booking.com uses Selective Triggering, Model- Output dependent triggering and designed an experiment that compares models to isolate the real business value and impact. (Figure 10)

This paper gave us an inside look into Booking.com and their data science environment. We learned about challenges, deployments, and evaluation of ML models that create real business impact.

If you liked this summary please let me know, also list any other research papers you would like me to go over!

- Alicja