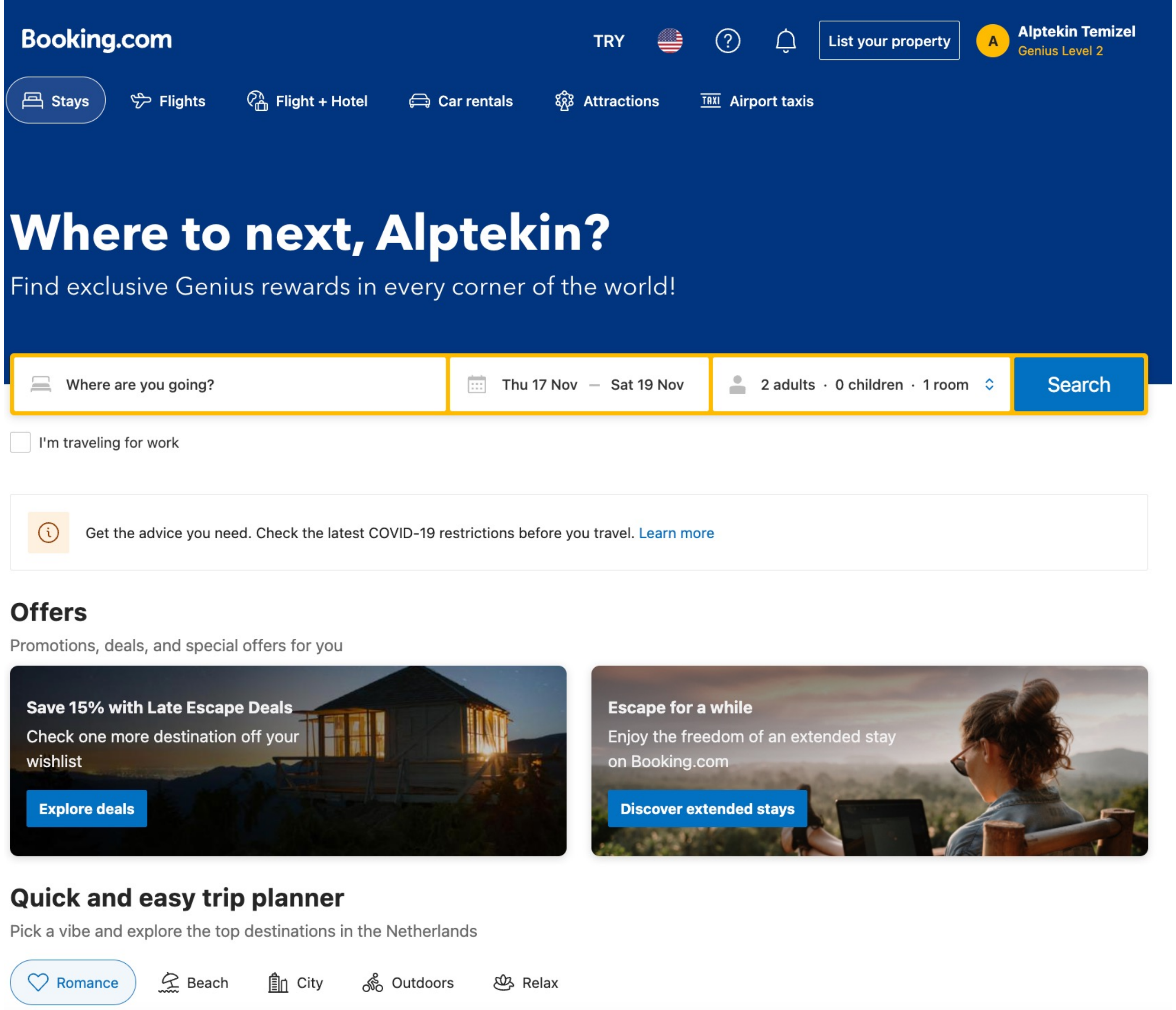
150 successful machine learning models: 6 lessons learned at Booking.com

Bernardi, L., Mavridis, T. and Estevez, P., 2019, July. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1743-1751).



Booking.com

TRY

List your property

A Alptekin Temizel
Genius Level 2

Stays · Flights · Flight + Hotel · Car rentals · Attractions · Airport taxis

Where to next, Alptekin?

Find exclusive Genius rewards in every corner of the world!

Where are you going? | Thu 17 Nov — Sat 19 Nov | 2 adults · 0 children · 1 room | Search

□ I'm traveling for work

Get the advice you need. Check the latest COVID-19 restrictions before you travel. Learn more

Offers

Promotions, deals, and special offers for you

Save 15% with Late Escape Deals
Check one more destination off your wishlist
Explore deals

Escape for a while
Enjoy the freedom of an extended stay on Booking.com
Discover extended stays

Quick and easy trip planner

Pick a vibe and explore the top destinations in the Netherlands

Romance · Beach · City · Outdoors · Relax

Take Home Lessons

1. Projects introducing machine learned models deliver strong business value
2. Model performance is not the same as business performance
3. Be clear about the problem you are trying to solve
4. Prediction serving latency matters
5. Get early feedback on model quality
6. Test the business impact of your models using randomised controlled trials (follows from #2)

“We found that driving true business impact is amazingly hard,

plus it is difficult to isolate and understand the connection between efforts on modeling and the observed impact…

Our main conclusion is that an iterative, hypothesis driven process, integrated with other disciplines was fundamental to build 150 successful products enabled by machine learning.”

Context

Delivering a great experience to their users is made challenging by a number of factors:

- High Stakes: Booking a stay at the wrong place is much worse than streaming a movie you don’t like!
- Infinitesimal Queries: User’s provide little information about what they are really looking for when booking a trip
- Complex Items: There is a lot of rich information available regarding accommodations, which can be overwhelming for users
- Constrained Supply: The supply of accommodation is limited, and changing prices impact guest preferences
- Continuous cold start: Guest preferences may change each time they use the platform (if e.g. booking only once or twice per year)

Different Models

- There are around 150 models in production.
- Some models are very *specific*, focusing on a particular use case in a particular context (e.g. recommendations tailored for one point in the funnel)
- Other models act as a *semantic layer*, modelling concepts that can be generally useful in many contexts.
  - For example, a model indicating how flexible a user is with respect to the destination of their trip.

Model Families

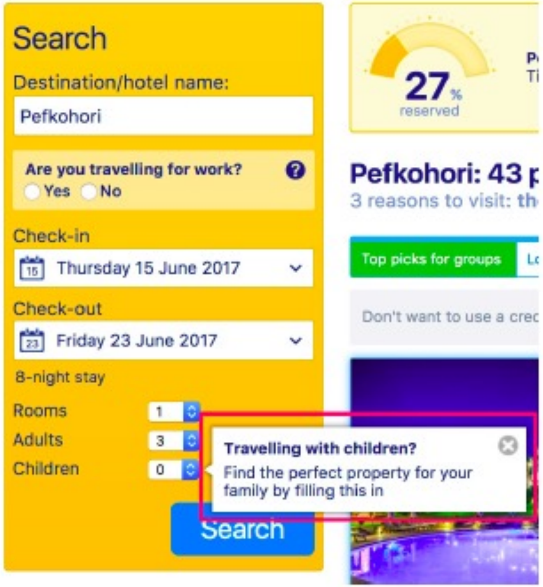
The models deployed at Booking.com can be grouped into six broad categories:

- **Traveller preferences models** operate in the semantic layer*, and make broad predictions about user preferences (e.g., degree of flexibility)
  - If a user is flexible, dates recommendations might be relevant in some situations,
  - If the user is not flexible, date recommendations might turn out distracting and confusing, and are therefore not displayed.

*A semantic layer is **a business representation of corporate data that helps end users access data autonomously using common business terms**. By using common business terms, rather than data language, to access, manipulate, and organize information, a semantic layer simplifies the complexity of business data. By using common business terms, rather than data language, to access, manipulate, and organize information, a semantic layer simplifies the complexity of business data.

Model Families

- **Traveller context models**, also semantic, which predictions about the context in which a trip is taking place (e.g. with family, with friends, for business, …).
    - *Usually, Family Travellers forget to fill in the number of children they travel with, going through a big part of the shopping process only to find out that the chosen property is out of availability for their children. The Family Traveller Model is used to remind the user to fill in the children information as early in the experience as possible, hopefully, removing frustration.*



Search

Destination/hotel name: Pefkohori

Are you travelling for work? ○ Yes ○ No

Check-in: Thursday 15 June 2017

Check-out: Friday 23 June 2017

8-night stay

Rooms: 1

Adults: 3

Children: 0

Travelling with children?
Find the perfect property for your family by filling this in

27% reserved

Pefkohori: 43 p

3 reasons to visit: th

Don't want to use a crec

Model Families

- **Item space navigation models** which track what a user browses to inform recommendations both the the user’s history and the catalog as a whole.
  - Most users who browse the inventory navigate through several supplementary and complementary options and items, such as dates, properties, policies, locations, etc.
  - In order to make a choice, they need to keep track of the options they have seen, while exploring neighbouring ones and trying to make a purchase decision.
  - Item space navigation models both feed from this process and try to guide it. They treat different actions, like scrolling, clicking, sorting, filtering etc., as implicit feedback about the user preferences.
  - These signals can then be used to facilitate access to the most relevant items in the user history, as well as to surface other relevant items in our inventory.

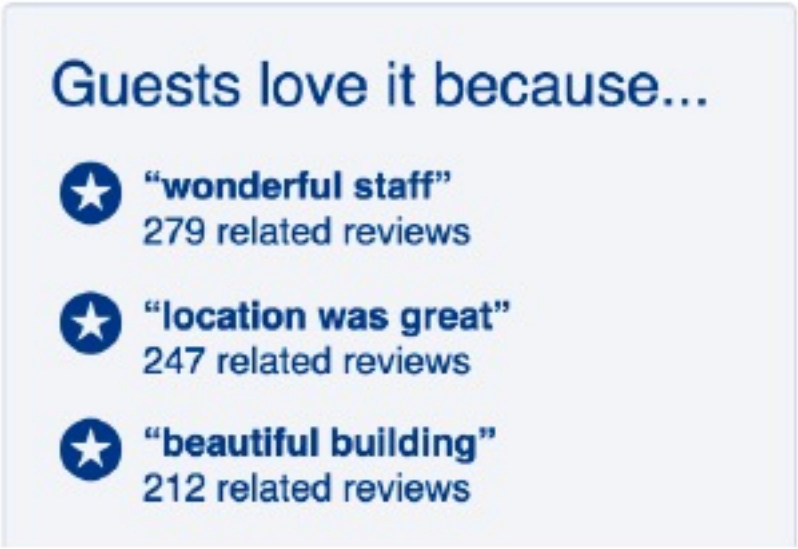
Different Models

- **User interface optimisation models** optimise elements of the UI such as background images, font sizes, buttons etc.
    - “we found that it is hardly the case that one specific value is optimal across the board, so our models consider context and user information to decide the best user interface.”

Different Models

- **Content curation models** curate human-generated content such as reviews to decide which ones to show

- A Machine Learning model "curates" reviews, constructing brief and representative summaries of the outstanding aspects of an accommodation.



Guests love it because...

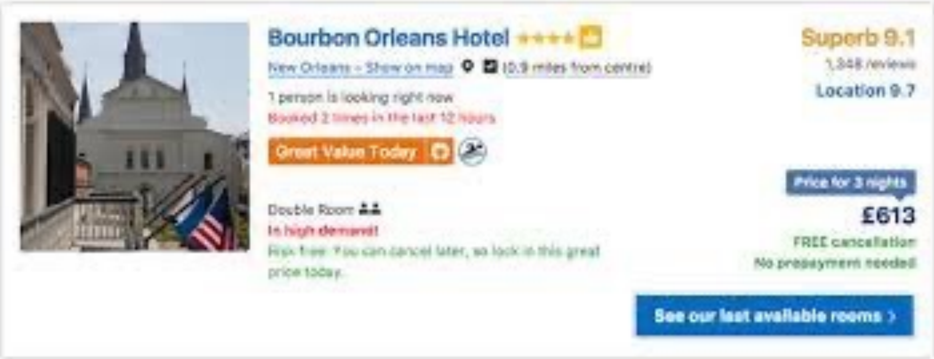
- **“wonderful staff”**
  279 related reviews
- **“location was great”**
  247 related reviews
- **“beautiful building”**
  212 related reviews

Different Models

- Content describing destinations, landmarks, accommodations, etc., comes in different formats like free text, structured surveys and photos; and from different sources like accommodation managers, guests, and public databases.
- It has huge potential since it can be used to attract and advertise guests to specific cities, dates or even properties, but it is also very complex noisy and vast, making it hard to be consumed by users.
- Content Curation is the process of making content accessible to humans. For example, they have collected over 171M reviews in more than 1.5M properties, which contain highly valuable information about the service a particular accommodation provides and a very rich source of selling points.

Different Models

- **Content augmentation models** compute additional information about elements of a trip, such as which options are currently great value, or how prices in an area are trending.
    - ○ "Great Value Today" icons highlight properties offering an outstanding value for the price they are asking, as compared to other available options.
    - ○ A machine learning model analyses the value proposition of millions of properties and the transactions and ratings of millions of users and selects the subset of properties with a "Great Value" offer.



Bourbon Orleans Hotel ★★★★

New Orleans – Show on map (0.9 miles from centre)

1 person is looking right now

Booked 2 times in the last 12 hours

Great Value Today

Double Room

In high demand!

Risk free: You can cancel later, so lock in this great price today.

Superb 9.1

1,348 reviews

Location 9.7

Price for 3 nights

£613

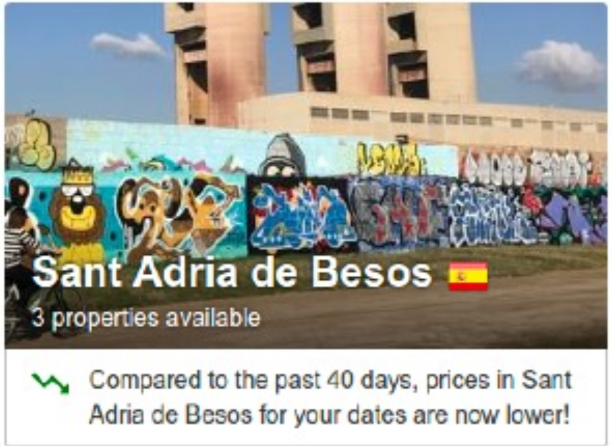
FREE cancellation

No prepayment needed

See our last available rooms ›

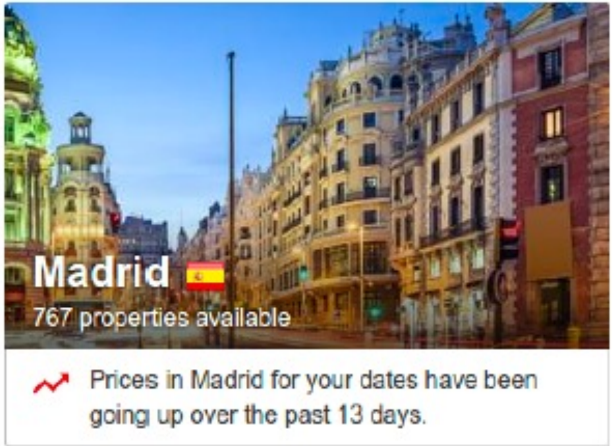
Different Models

- **Content augmentation models**
    - Price Trends: Depending on the anticipation of the reservation, the specific travelling dates and the destination, among other aspects, prices display different dynamics.
    - Since they have access to thousands of reservations in each city every day, they can build an accurate model of the price trend of a city for a given time and travelling dates.
    - When the model finds a specific trend, users is informed to help them make a better decision, either by encouraging them to choose a destination and dates that look like an opportunity, or discouraging particular options in favor of others.
    - Note that in this case, the augmented item is not an accommodation but a destination


Sant Adria de Besos

3 properties available

Compared to the past 40 days, prices in Sant Adria de Besos for your dates are now lower!


Madrid

767 properties available

Prices in Madrid for your dates have been going up over the past 13 days.

Different Models

- Models in this family derive attributes of a property, destination or even specific dates, augmenting the explicit service offer.
- Content Augmentation differs from Content Curation in that curation is about making already existing content easily accessible by users whereas augmentation is about enriching an existing entity using data from many others.

Lesson 1: projects introducing machine learned models deliver strong business value

- All of these families of models have provided business value at Booking.com.
- Moreover, compared to other *successful* projects that have been deployed but did not use machine learning, the machine learning based projects tend to deliver *higher returns*.

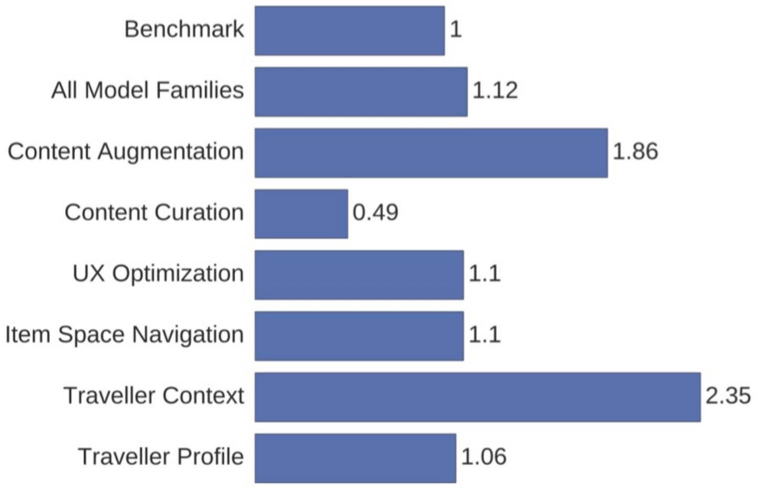


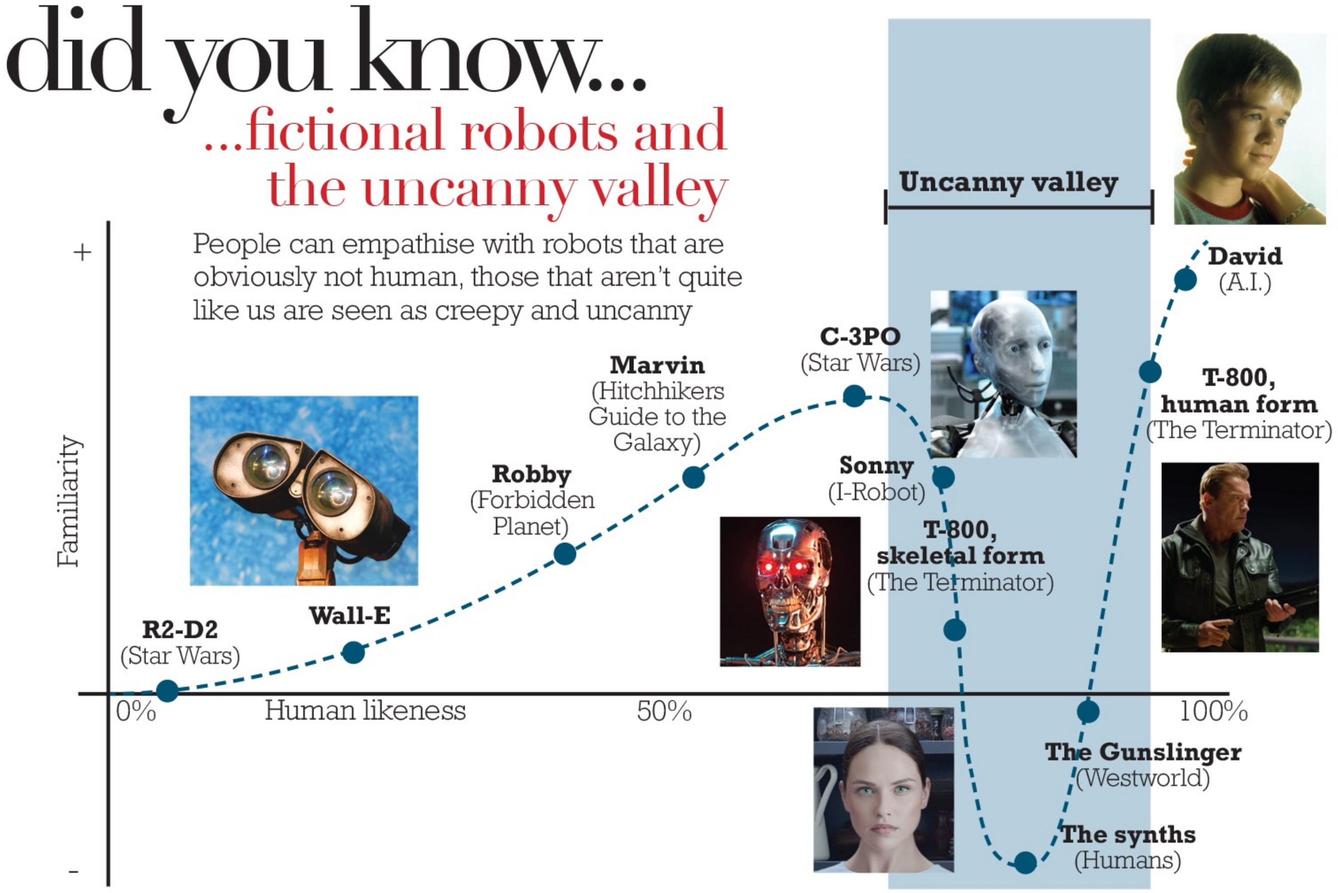
Figure 2: Model Families Business Impact relative to median impact.

Lesson 2: Model performance is not the same as business performance

- An interesting finding is that increasing the performance of a model does not necessarily translate into a gain in [business] value
- This could be for a number of reasons including
  - saturation of business value – you cannot drive value from model performance gains indefinitely, gains in performance produce little or no value gain

  - segment saturation due to smaller populations being exposed to a treatment – when testing a new model against a baseline they apply ”triggered analysis” to make sure only the users exposed to change are considered (users for which the models disagree) (as the old and new models are largely in agreement);

  - over-optimisation on a proxy metric (e.g. clicks) that fails to convert into the desired business metric (e.g. conversion);

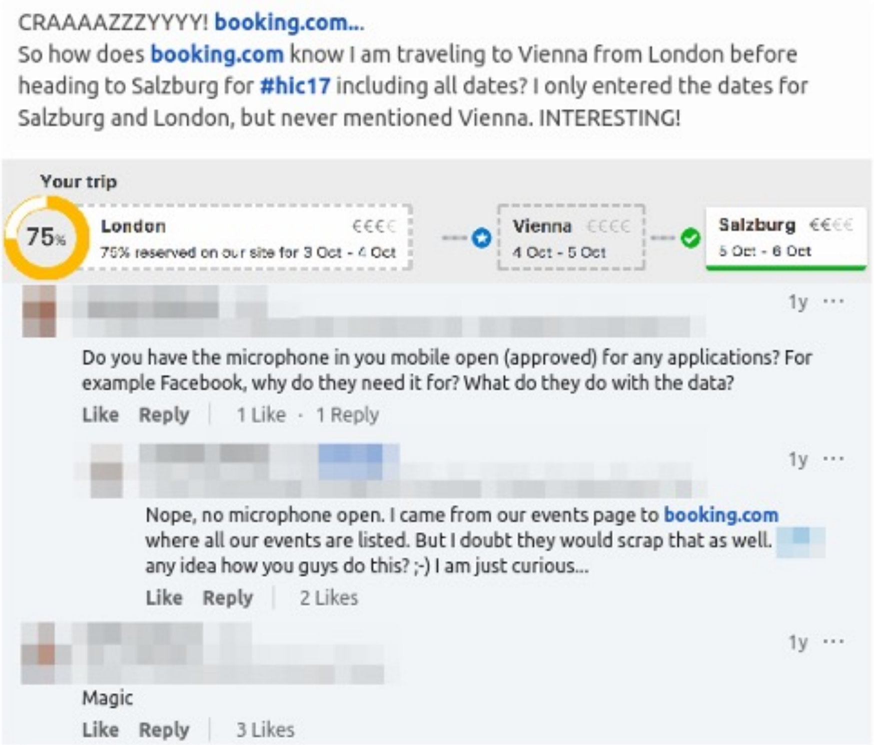


did you know...

...fictional robots and the uncanny valley

People can empathise with robots that are obviously not human, those that aren't quite like us are seen as creepy and uncanny

https://engtechmag.wordpress.com/2015/06/16/why-real-robots-will-never-look-like-the-synths-in-humans/



CRAAAAZZZYYYY! booking.com...
So how does booking.com know I am traveling to Vienna from London before heading to Salzburg for #hic17 including all dates? I only entered the dates for Salzburg and London, but never mentioned Vienna. INTERESTING!

Your trip

75% London €€€€ 75% reserved on our site for 3 Oct - 4 Oct

Vienna €€€€ 4 Oct - 5 Oct

Salzburg €€€€ 5 Oct - 6 Oct

1y ···

Do you have the microphone in you mobile open (approved) for any applications? For example Facebook, why do they need it for? What do they do with the data?

Like Reply | 1 Like · 1 Reply

1y ···

Nope, no microphone open. I came from our events page to booking.com where all our events are listed. But I doubt they would scrap that as well. any idea how you guys do this? ;-) I am just curious...

Like Reply | 2 Likes

1y ···

Magic

Like Reply | 3 Likes

Figure 5: Uncanny valley: People not always react positively to accurate predictions (destination recommender using Markov chains).

As models become better and better, they know more about the users and can predict very well what the user is about to do. This might be unsettling for some of the customers and may translate to a negative effect on value.

Lesson 3: Clear Problem Definition

- Before you start building models, it is worth spending time carefully constructing a definition of the problem you are trying to solve.
- The Problem Construction Process takes as input a business case or concept and outputs a well-defined modeling problem (usually a supervised machine learning problem), such that a good solution effectively models the given business case or concept.
- The point(s) at which the prediction needs to be made are often given, which fixes the feature space universe. Yet, the target variable and the observation space are not always given and need to be carefully constructed.

Lesson 3: Clear Problem Definition

- For example, “dates flexibility” model is where we want to know the flexibility of the users every time a search request is submitted.

- Though, what is flexibility?
  - User is considering more alternative dates than a typical user?
  - The dates the user will end up booking are different to the ones he/she is looking at right now?
  - User is willing to change dates but only for a much better deal?

- We could
  - Learn to predict how many dates the user will consider by applying regression to a specific dataset composed by users as observations
  - Estimate the probability of changing dates by solving a classification problem, where the observations are searches.

Lesson 3: Clear Problem Definition

- Some of the most powerful improvements come not from improving a model in the context of a given setup, but changing the setup itself.
- For example, changing a user preference model based on click data to a natural language processing problem based on guest review data.
- In general we found that often the best problem is not the one that comes to mind immediately and that changing the set up is a very effective way to unlock value.

Lesson 3: Clear Problem Definition

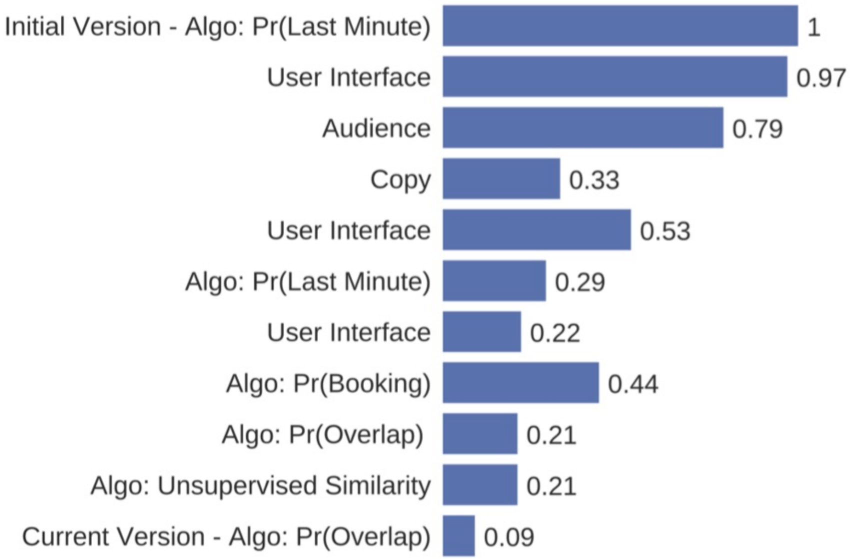


Figure 3: A sequence of experiments on a Recommendations Product. Each experiment tests a new version focusing on the indicated discipline or ML Problem Setup. The length of the bar is the observed impact relative to the first version (all statistically significant)

- There are 6 successful algorithm iterations and 4 different setups.
- Pr(Last Minute) classifies users into Last Minute or not
- Pr(Booking) is a conversion model
- Pr(Overlap) models the probability of a user making 2 reservations with overlapping stay dates
- Unsupervised Similarity models the similarity of destinations

Lesson 4: prediction serving latency matters

- In the context of Information Retrieval and Recommender Systems, it is well known that high latency has a negative impact on user behavior.
- In an experiment introducing synthetic latency, Booking.com found that an increase of about 30% in latency cost about 0.5% in conversion rates
- Hypothesis: “Decreasing latency can produce a gain in conversion”

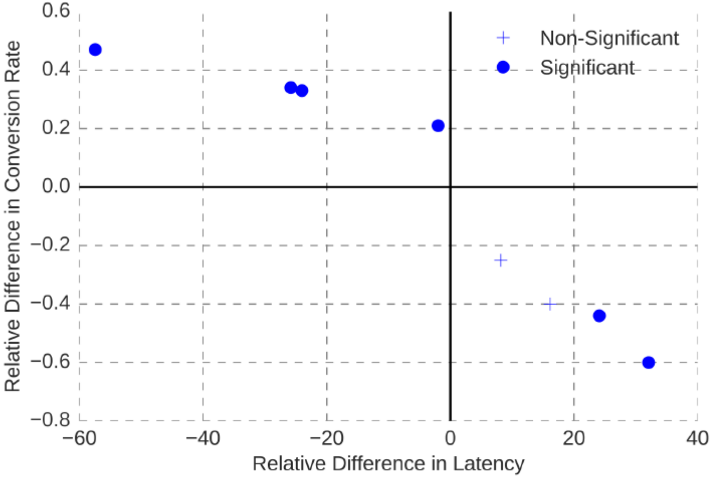


Figure 6: Impact of latency on conversion rate

Lesson 4: prediction serving latency matters

- To minimise the latency:
  - Model Redundancy: Model copies are distributed across a cluster to respond as many predictions as requested – scale horizontally to deal with large traffic
  - In-house developed linear prediction engine: Highly tuned to minimize prediction time – serves all models reducible to inner products (such as, Naïve Bayes, generalized linear models, k-NN with cosine or Euclidean distance, matrix factorization)
  - Sparse models: Model with less parameters require less computation
  - Precomputation and caching: When the feature space is small, all predictions can be stored in a distributed key-value store. When it is too big, cache frequent requests in memory.
  - Bulking: When many request per prediction is required, bulk them together in a single request.
  - Minimum Feature Transformations: ?

Lesson 5: get early feedback on model quality

- When models are serving requests, it is crucial to monitor the quality of their output but this poses two main challenges:
  - Incomplete feedback: In many cases true labels cannot be observed. For example, a model that predicts whether a customer will ask for a “special request”. Predictions are used while the user shops but a true label is only assigned if user actually makes a booking, as special request is filled in at actual booking time.
  - Delayed feedback: True label is observed many days or weeks after the prediction. For example, a model that predicts whether a user will submit a review or not. This model might be used at shopping time but the true label will only be observed after the stay.
- In such cases label-dependent metrics such as precision, recall are not applicable.
  - What can we say about the quality of a model but just looking at the predictions it makes when serving?

Lesson 5: get early feedback on model quality

- For binary classifiers: analyze distribution of responses generated by the model based on Response Distribution Chart (RDC) – histogram of the output of the model.

- RDC of an ideal model should have one peak at 0 and one peak at 1 (height of which are relative to the class proportion)

- “*Smooth bimodal distributions with one clear stable point are signs of a model that successfully distinguishes two classes*.”

- Other shapes can be indicative of a model that is struggling.

Lesson 5: get early feedback on model quality

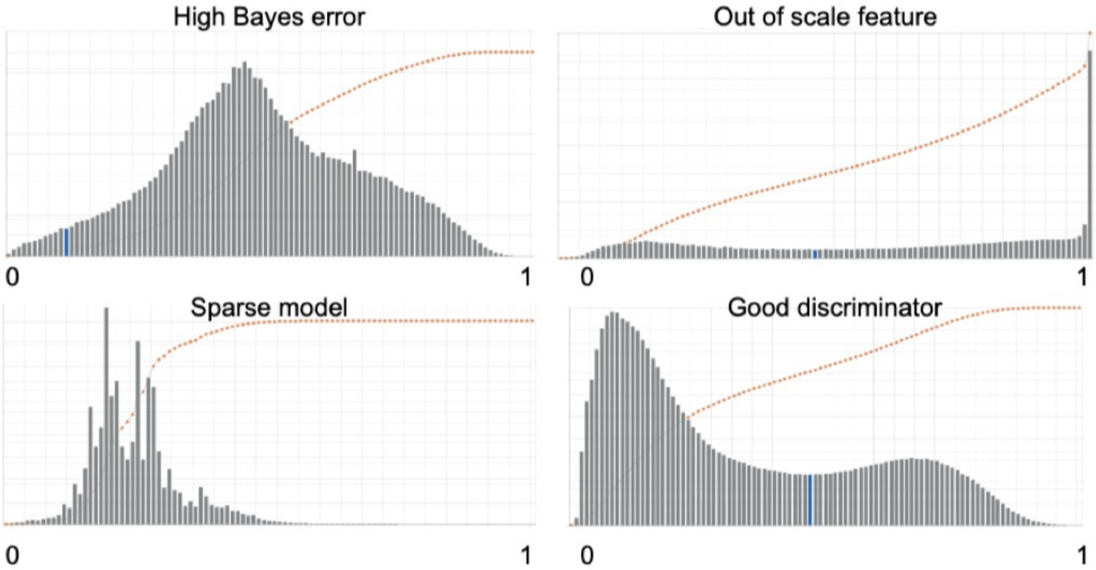


Figure 7: Examples of Response Distribution Charts

A smooth unimodal distribution with a central mode might indicate high bias in the model or high Bayes error in the data.

Lesson 5: get early feedback on model quality

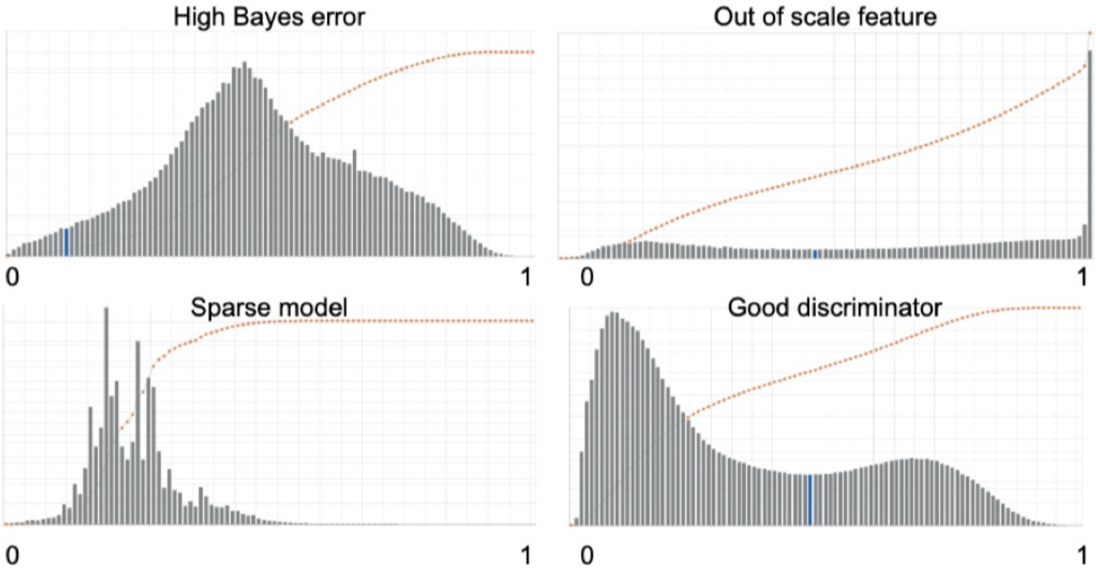


Figure 7: Examples of Response Distribution Charts

An extreme, high frequency mode might indicate defects in the feature layer like wrong scaling or false outliers in the training data.

Lesson 5: get early feedback on model quality

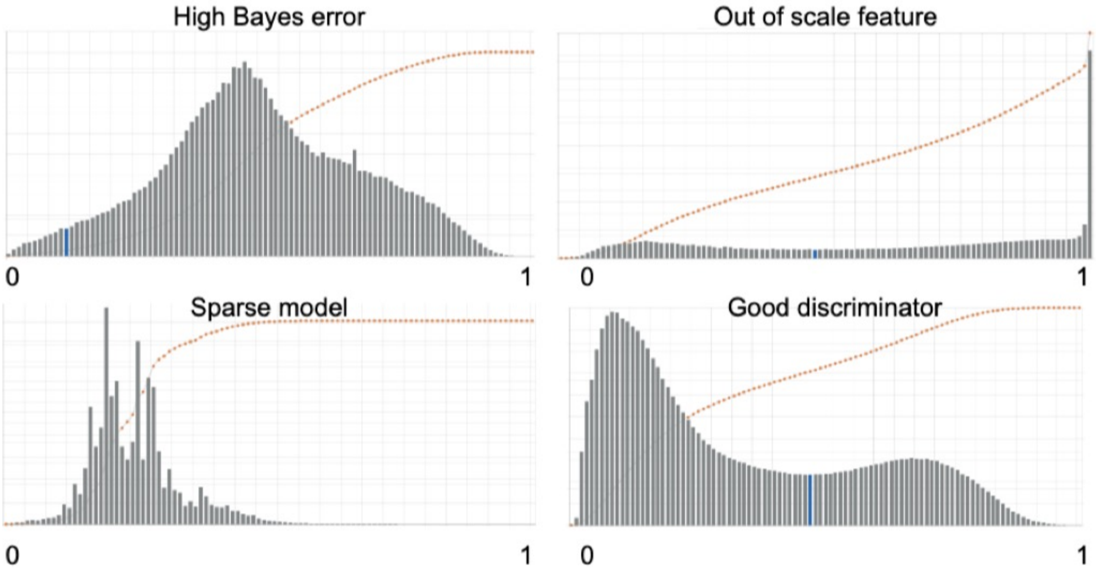


Figure 7: Examples of Response Distribution Charts

Non-smooth, noisy distributions point to too excessively sparse models.

Lesson 5: get early feedback on model quality

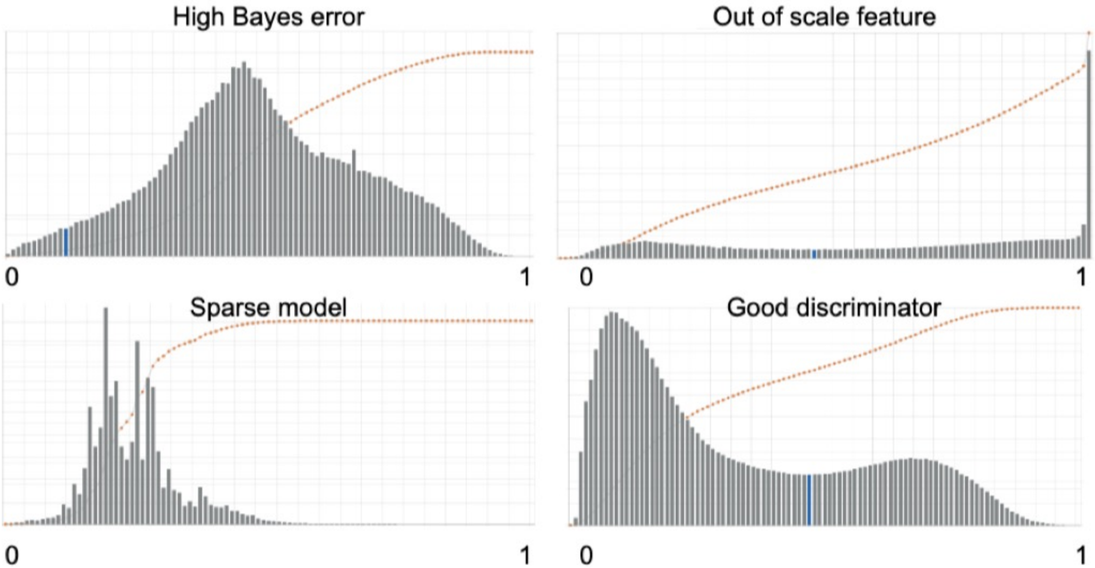


Figure 7: Examples of Response Distribution Charts

Smooth bimodal distributions with one clear stable point are signs that a model successfully distinguishes two classes.

Lesson 6: test the business impact of your models

- Experimentations are done through: “Randomized Controlled Trials (RCT)”
- In-house experimentation platform enables everybody to run experiments to test hypotheses and assess the impact of their ideas.
- “The large majority of the successful use cases of machine learning studied in this work have been enabled by sophisticated experiment designs, either to guide the development process or in order to detect their impact.”

Lesson 6: test the business impact of your models

- Selective Triggering: In a standard RCT, the population is divided into control and treatment groups.
  - Control groups are exposed to no change
  - Treatment groups are exposed to change
- However, in many cases, not all subjects are eligible to be treated and the eligibility criteria are unknown at assignment time.
- This is often the case for ML models, since they may require specific features to be available. The subjects assigned to a group but not treated add noise to the sample, diluting the observed effect, reducing the statistical power and inflating the False Discovery Rate.

Lesson 6: test the business impact of your models

To deal with this, “Triggered Analysis” is used where the only treatable (triggered) subjects in both groups are analysed.

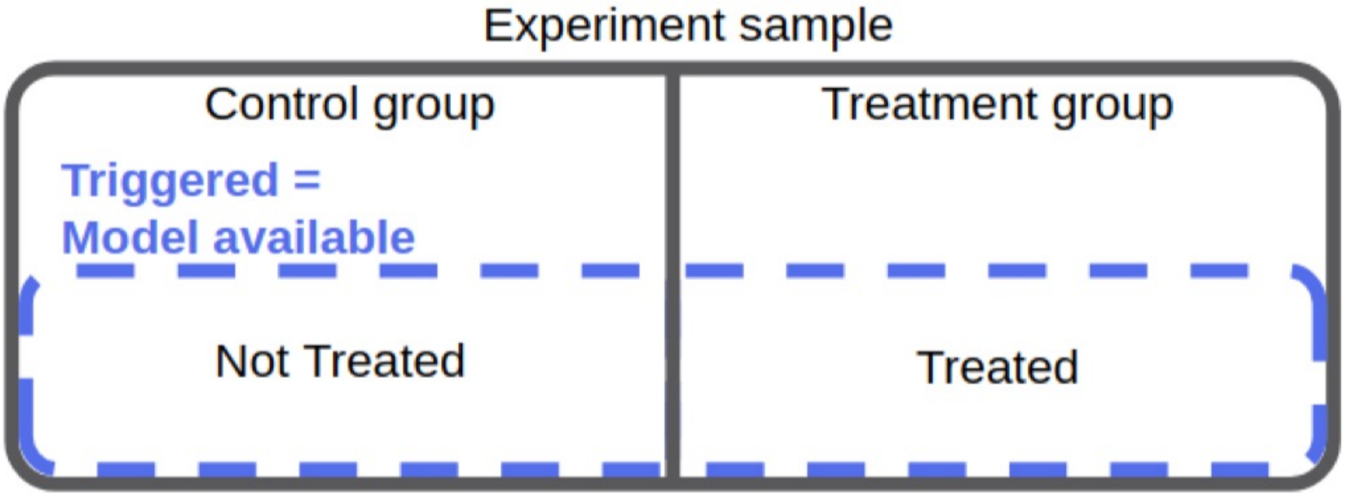


Figure 8: Experiment design for selective triggering.

Lesson 6: test the business impact of your models

- Even when all the model requirements are met, the treatment criteria might depend on the model output.
    - For instance, a block with alternative destinations are only shown to users identified as destination flexible by the model.
- In such cases, some users are not exposed to any treatment, diluting the observed effect. Previous setup cannot be used since in the control group, the output of the model is not known and therefore cannot condition the triggering.
- Modifying the control group to call the model is not advised, since this group is also used as a safety net to detect problems with the experiment setup, and in such cases all the traffic can be directed to control group while investigating the issue.

Lesson 6: test the business impact of your models

The setup for model output dependent triggering requires an experiment with 3 groups. Control group *C* is exposed to no change

Treatment groups T1 and T2 invoke the model and check the triggering criteria, but only in T1 triggered users are exposed to change. In T2 users are not exposed to any change regardless of the model output.

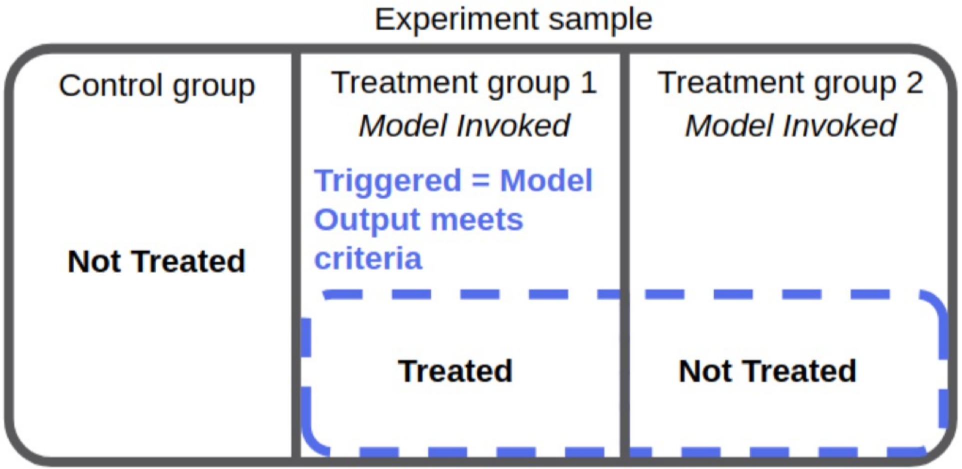


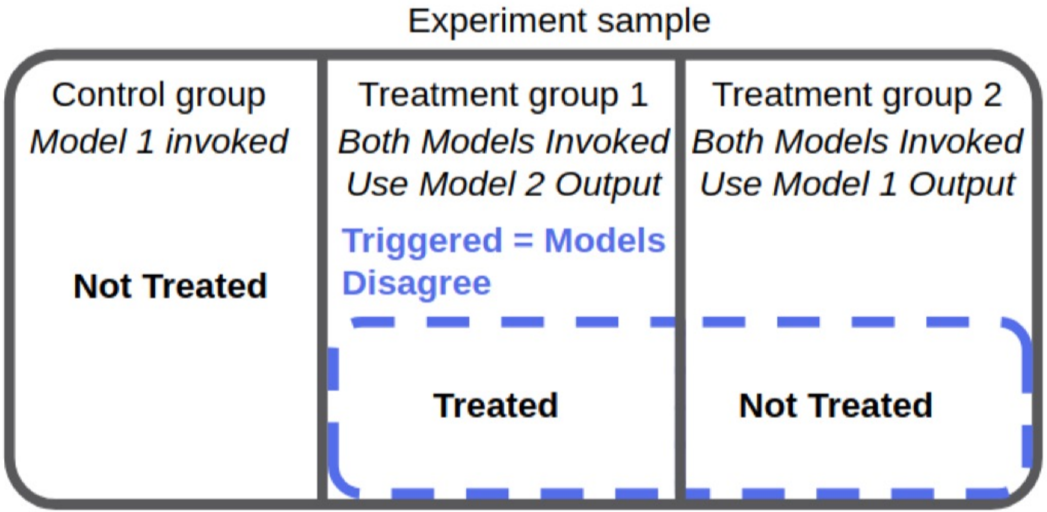
Figure 9: Experiment design for model-output dependent triggering and control for performance impact.

Lesson 6: test the business impact of your models

When comparing models the situations where the two models disagree are of interet.

Control group that invokes and uses the output from model 1. This is the current baseline and safety net

Triggering condition is *models disagree* (outputs from both models are required in T1 and T2)



Experiment sample

Control group
Model 1 invoked
Not Treated

Treatment group 1
Both Models Invoked
Use Model 2 Output
Triggered = Models Disagree
Treated

Treatment group 2
Both Models Invoked
Use Model 1 Output
Not Treated

Figure 10: Experiment design for comparing models.