How Booking.com leverages its online data to get grip on the customer service workload forecast

Improving operations by driver-based behavioral forecasting

Daan de Bruin
Senior Analyst
MIcompany

Roland Tabor
Partner
MIcompany

Wynfrtih Meijwes
Program Manager
MIcompany

Erik Benson
Manager Customer Service
Workforce Management
Booking.com

Booking.com

MI COMPANY

How Booking.com leverages its online data to get grip on the customer service workload forecast

Improving operations by driver-based behavioral forecasting

Handling a million bookings per day in more than 40 languages. Booking.com is the global market leader in the travel industry, handling over a million bookings per day. To optimize the way it handles the millions of customer interactions per day, Booking.com is well known to be one of the innovators in using Big Data Analytics. The recommendations you get as a customer, the web-site design and functionality, has all been optimized using the data available. A critical customer interaction platform is the Customer Service centre, since many customers or prospects still have a need to interact over the phone or mail. Weekly, Booking.com has to handle more than a million questions, requests and complaints. To do so, the Booking.com Customer Service (CS) centre consists of thousands of FTE of CS agents speaking 43 different languages. All of these agents need to be planned and scheduled properly. The question is how.

This challenge is not trivial, since Customer Service traffic follows a different pattern than web site visits, and there is very significant variance in Customer Service workload per language per day and across weeks. Even regional events in a country can have significant impact on the daily workload. Therefore, next Tuesday's Customer Service workload can easily be more than double than that of this Tuesday's workload. The workload forecast is of utter importance, since shortage of agent capacity leads to long waiting times, low net promotor scores, missed revenues and stressed agents. Whereas agent overcapacity leads to unnecessary high labour costs and bored agents. With thousands of FTEs a 10% over forecast will very substantially impact Booking.com's operating expenses.

Booking.com currently hires and schedules its agents 13 weeks ahead, using a daily workload forecast per language. This forecast was often off by more than 25% for difficult to forecast languages. Especially the smaller, fast growing languages turn out to be especially difficult to forecast properly. As a result, the benefits of improving forecast accuracy are high: reducing the average forecast error to 5 percent results in more satisfied customers and a potential of saving many millions of yearly costs savings. However, the challenge is significant. Over the past years, several attempts to implement a more accurate forecast failed in the complexity.

> Erik Benson, head of customer services workflow management: "We were frustrated with the stability and accuracy of our CS forecast. Some of the inputs into our forecasting approach were still largely informed by a family of spreadsheets, manually maintained. We were aware of the potential to improve these forecasts, however two major "black-box model" attempts failed in the complexity involved. We chose to work with MIcompany because of their relevant experience, focus on getting impact, and on its commitment to build up our capabilities."

The challenge: drive more accuracy, while building up a system that people understand and adopt

So why did Booking.com struggle with getting grip on the Customer Service workload forecast? This all seems surprising since the rise of readily available Big Data software opens up the opportunity for do-it-yourself forecasting. The amount of customer and open source data available is rapidly increasing, while open source tools like R and Python provide all relevant statistical and machine learning packages with forecasting applications within easy reach for all analysts. Furthermore, Booking.com is known to attract the most talented computer scientists and statisticians. So, with all the data and technology that is available, why did Booking's previous forecasting projects result in a highly complex model, that lacked accuracy, and was delivered way past required deadlines?

First, in our experience many companies suffer from relatively poor performing forecasts, often in a much more simple and stable business environment compared to Booking.com. Efforts to improve these forecasts are often unsuccessful since they lack an integral model but consists of many different spreadsheets that serve as inputs to the forecast engine.

Secondly, data scientist that are in love with advanced Big Data Analytics methodologies, tend to make forecasts overly complex, while failing to really understand the drivers of consumer behavior. The key question for Booking.com is which customers call at what moment and why? Advanced analytics allows analysts to ignore these drivers, and focus more on the technicalities of a machine learning approach. However, even with advanced analytical methodologies, still the understanding of customer behavior is key for accurate forecasts, and explaining changes in trends.

Third, forecasting models need to get used and adopted by management. Analyst sometimes fail to appreciate this point, believing everything is about the strength of the algorithm. However full management acceptance of a model and trust in the output is necessary for a model to be adopted by an organisation.

And lastly, we believe that the value of the model does not lie in its initial performance. The value lies in the integration of the model in a system that allows for continuous accuracy monitoring and improvement. When the system is right and self-correcting, the forecast's accuracy and trust grows over time. When it is a one-shot model, you can only wait until troubles arise. At most companies, however, this fundamental principle of prediction modelling is poorly understood.

Our joint Booking.com – MIcompany approach

So what would be a good way forward for Booking.com? At the end of 2015 MIcompany and Booking.com set up a joint project team to answer this question. The ambition of the project was to develop an approach to decrease the average "13-weeks ahead daily forecast" error by 50%. The conclusion from the joint team was that we needed a project approach that would avoid the pitfalls as addressed above. Hence, the scope would be limited to two languages initially, we would go for a best effort in a first wave of modelling with a fixed timing, and support Booking.com with building a system that would allow for further forecast roll-outs and forecast improvements.

The agreed approach was inspired by our principles for effective forecasting that we distilled from our own experiences, and the global forecasting literature. (see frame to the left).

These principles are:

1. Understand the dynamics: Invest in truly understanding the business dynamics that explain the patterns, in a hypothesis driven approach.

2. Minimize the number of variables: Complexity often leads to worse forecasts; smartly combine data and select only the most important drivers.

3. Get your statistics right: Thoroughly test predictive properties of different forecasting techniques and manage outliers by identifying and understanding them.

4. Build a system, not a model: Start with a small scope and work quickly towards a minimum viable product, creating focus and momentum. Then build a system that not only runs the forecasts, but also integrates a learning loop to improve from the forecasts deviations.

We will illustrate these principles, with the Booking.com project approach.

Principle 1. Understand the dynamics

Let's say in 2015 we have 100K incoming calls on a certain day and in 2016 300K for that same day. What should be the expected number of calls for 2017? You could extrapolate the trend, but a more clever solution would start with understanding why we had 100K in 2015 and 300K in 2016. In other words, find the drivers of the dynamics. Getting insights into these drivers and thus customer dynamics is the most important and crucial part in designing an appropriate model.

To find out what was driving the workload, we started with decomposing it in phone contacts and emails. The phone workload we modelled as the number of incoming phone calls times the average handling time (AHT) of the agent. As an example, in Figure 1 we see that the Contact Volume (calls + emails) had both a strong trend and a seasonal pattern. In contrast, the AHT has less variation and no clear yearly or weekly seasonality or trend.

We identified the key drivers in a structured hypothesis driven approach, starting with a long list of more than fifty drivers, using input from field experts like business owners, call centre agents and schedulers. This list was later condensed by conducting relatively basic

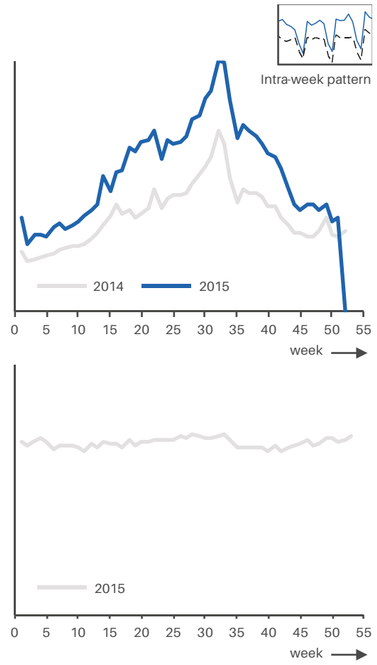

Figure 1. Seasonal pattern in Contact Volume and Average Handling Time

correlation analyses. Through getting a basic picture of correlation between drivers of Contact Volume (i) and AHT (ii), we immediately could weed out some variables that have no predictive power and therefore focus on the more important ones:

i. Drivers of Contact Volume

The Contact Volume is driven by, firstly, the amount of known active bookings today and its "booking phases". Typically, around the reservation and around the moment of check-in and check-out of their stay, bookers are likely to contact the Customer Service centre. For instance, around the moment of booking, a young family wonders whether a cot is available in the hotel room and after the check-out a dissatisfied business man may share his complaints. Moreover, the bookers in between making the reservation and checkin in also tend to contact. For example, the summer peak of contacts is not only caused by many check-ins or check-outs, but mostly by bookers who are in the weeks before their check-in. The amount of 'pending check-ins' builds up in the spring towards the summer.

In Figure 2a we see that property type drives the number of contacts per booking too. Apartment bookers are three times more likely to contact than hotel bookers. The increasing share of apartment bookings partly drives the upward trend in number of contacts. Other booking characteristics like the trip length and group size also influence the chance of contacting.

Third, the type of (holi)day matters. Weekends and certain holiday periods like Christmas and Easter result in less contacts than we would expect based on the number of bookings. Last, changes in opening hours result in trend breaks. For example, the opening hours of a specific Asian Customer Service phone line varied dramatically, between 8 hours a day and 24 hours a day during a year.

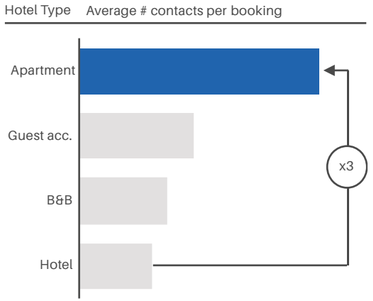

Figure 2a. Insight Contact Volume
Average # contacts per booking per hotel type

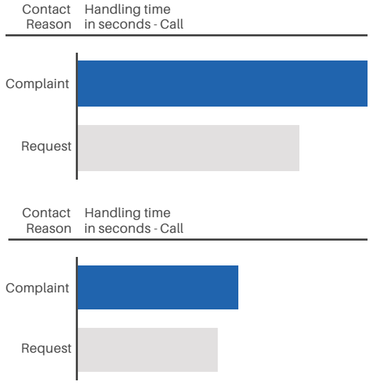

Figure 2b. Insight AHT
Average handling time per contact reason, call and mail

ii. Drivers of AHT

The average handling time (AHT) is driven by factors other than volume. Figure 2b shows, in general, that calls have a higher AHT than e-mails. Diving deeper into the data we get some more interesting insights: First, a complaint takes much longer than a request, most probably this is caused by the more actions the agent has to do to solve the issue. Secondly, the language match between the customer and the agent drives AHT; when an agent speaks in English with a non-English customer, the calls take much less time. Last, the AHT is influenced by the type of agent. An inexperienced agent for example takes much more time to handle a request or complaint than an experienced agents.

Principle 2. Minimize the number of variables

The result of the analysis phase was a deep understanding of the causal drivers of contact volume and handling time and thus the workload. However, when a predictive model has too many drivers and parameters, there

is a substantial risk that the model does not fit the 'true' relationships, but fits the random variations in the data. The in-sample fit will then be very good, but out of sample the model will forecast badly. This is called overfit and is one of the main reasons why predictions of models fail.

Therefore, we asked ourselves the question; how could we condense the number of variables, and as such simplify the model and reduce the risk of over-fit? What we found out was that we could translate the analysis results in a minimum number of appropriate drivers of workload that can be forecasted 13 weeks ahead. Especially for the contact volume model this was an important part of the project. We constructed one driver that combines all the relevant information on the number and distribution of active bookings; we named this driver as the "normalized bookings driver" as seen in Figure 3.

The normalized booking driver is a weighted average of all (predicted) active bookings per booking phase, where the weights are determined by the historical probability of contacting in a specific phase for a specific segment. This normalized booking driver turns out to be highly predictive for the contact volume. We also constructed one driver for the (special) days, such as bank holidays or major events. For each day in the future we determined the most appropriate historical reference day. This reference day is usually the same day of the week at the same week last year. However, when for example today is Christmas Day, we selected Christmas Day last year as the reference day. The resulting 'contacts last year at reference day' was one of the main drivers of the model.

Principle 3. Get your statistics right

So now we have a strong grip on the drivers, but we lack a perspective on the right statistical model that translates these drivers into the overall workload forecasts. We took this challenge in two steps: manage outliers and choose the right modelling technique. Outliers. Historical shocks (outliers) can have a huge impact on the forecast and its accuracy. For example, in October 2015, the number of Asian calls about website problems was extremely high. After a detailed assessment this turned out to be an incident that we should exclude from the data. For the handling time, the outlier detection was even more important. Certain contacts (by chance) took several hours to handle, whereas data issues caused other contacts to take 0 seconds. Excluding this 'noise' from the data dramatically improved the forecast.

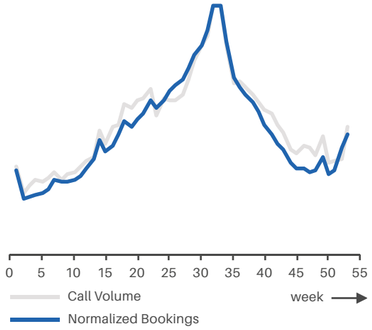

Figure 3. Weekly Call Volume and Normalized Bookings

Modelling technique. Oncewe had a cleaned set of data, and a very solid perspective on the key drivers, we were keen to use the technique that would give us the most accurate and stable forecasts. But we also aspired to have a technique that provides results we can interpret and understand (no black box) and could be easily implemented and further developed. In order to find out which model would fit our purpose best, we estimated (trained) our model using all the data up until October 2015, and tested the predictive properties of our model on a sample from the end of 2015 and beginning of 2016. We thoroughly tested the predictive power of various models, both in the short term (1 to 2 weeks ahead) and in the long term (13 weeks ahead).

Daan de Bruin, senior analyst at MIcompany: "Based on our joint team's experience, we selected four different modelling techniques. These where techniques that are academically sound, but also applicable in business. We considered standard linear regressions, two versions of exponential smoothing (the Holt-Winters model and TBATS model) and ARIMA(X) models. The ARIMAX model turned out to be the best choice; it was the most accurate, especially in the short term, while we can still understand and explain what the model is doing"

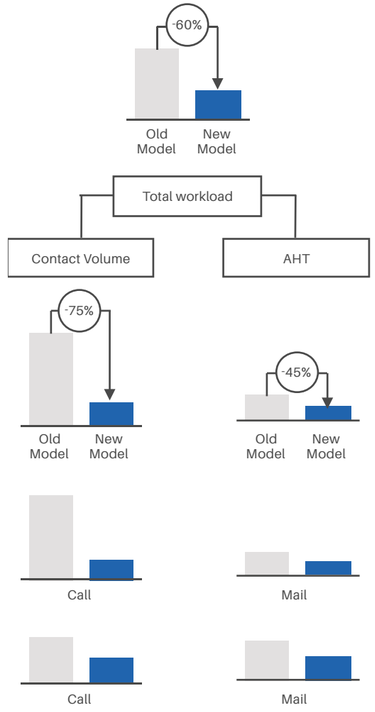

Figure 4. Mape Tree example language

Principle 4. Build a system, not a model

So we understand the dynamics and have a model that gives accurate forecasts. Are we done now? The last principle is about the importance of integrating the model within a system, instead of only building a model. We need a system that not only runs the forecasts, but also enables Booking.com's data scientists to constantly check upon the model's performance, identify and understand forecast errors and update and improve the model. Without such a system, the model will be inaccurate within a year. At most companies, creating this system is an undervalued but critical step.

Before building the system, we deliberately implemented a test and learn approach. We chose two pilot languages to start with; one European and one Asian language. These were important (growing) languages with a high forecasting inaccuracy, with dynamics that were clearly different. We chose only two in order to be able to thoroughly understand the dynamics of these two languages, believing these findings would become the first basis for an iterative roll-out approach. From the very beginning of the project, we aimed for finishing a minimum viable version 1.0. When we showed that this first version was already a major improvement over the existing forecast model, we gained momentum and enthusiasm within the organisation. Besides that, this approach helped the project team to stay focused and work via a structured approach, constantly having the end goal in mind.

After finishing the 1.0 version, one part of the project team worked on the technical implementation and automation of the forecast, i.e. building the system. The system was built such that every forecast is automatically written to a database, together with all the drivers that were used in the forecast. On top of that, a layer was built that automatically calculates and visualizes the forecasts' (in)accuracies and explains these through the model's driver realizations. The generic set up of the system was such that it could be easily extended to other languages and forecasting problems. The rest of the team iteratively improved upon the model, by learning from the deviations of the forecast from the actual workload. For example, both the normalized booking driver and the outlier removal went through various iterations, improving model performance.

"We strive to learn and to solve problems for our partners and guests at Booking.com. The iterative way of working with MIcompany and the project structure helped us to stay focused, deliver a good result in time and learn the essence (or AB) of successful project management. And the uniqueness of how we, at Booking.com, approach challenges and the way our project partner adapted to us, showed not only that we choose the right company to work with but that our work culture is something others can learn from too."

Vladimir Sterngold, Senior Business Analyst at Booking.com:

"Find the right variables"

Although everyone has the same data, this doesn't mean everyone creates the same forecasting drivers. Often smartly selecting, transforming and combining data is the key for making predictive drivers.
Signal and the Noise, Nate Silver

"Avoid overfitting"

One of the most important pitfalls when forecasting is considering random variations over time (noise) as trends or seasonality's that will return next period. The future is like the past but never the same. Adding irrelevant predictors can make forecasts worse because the coefficients fitted to them add random variation tot the subsequent predictions.
The Problem of Overfitting, Douglas M. Hawkins

"Detect outliers"

The best model can return very bad forecasts when outliers in the data are not properly handled. This is especially important in forecasting, since often quite a small number of data points is used compared to modeling for example on customer data.
Superforecasting: The Art and Science of Prediction, Philip Tetlock and Dan Garder

"Be conservative"

It is in the nature of man to extrapolate (exponential) trends based on very small number of observations, while often time series return to normal averages.
Golden rule of forecasting: Be conservative, JS Armstrong, KC Green and A Graefe

The Results

In Figure 4 we see that the total workload forecast error decreased by 60 percent for the pilot European language. For the Asian pilot language, it decreased by 50 percent. The large workload improvement is mainly driven by the Contact Volume Forecast (a 75 percent reduction). The handling time forecast also improved substantially (60 percent).

When we zoom in on the model's predictions (Figure 5), we see that the new model is spot on for most of the days, and especially outperforms the old model during the special days, e.g. Christmas, New Year's Day, and Epiphany.

Furthermore, the method developed turned out to be generally applicable to all other Customer Service languages. Out of the 43 languages, only 2 didn't improve. These 2 remaining languages are the first ones to be improved upon in the test & learn approach. The outstanding results for all the other languages firmly validates the chosen approach, drivers and model techniques.

To conclude

To conclude, forecasting is not magic. We have all the data and tooling right under our nose, but success lies first in properly understanding customer behavior that explains the data. Subsequently translating this understanding into a minimum number of predictive variables, applying the appropriate statistical models, and building a system to continuously test and improve your model is the final critical step

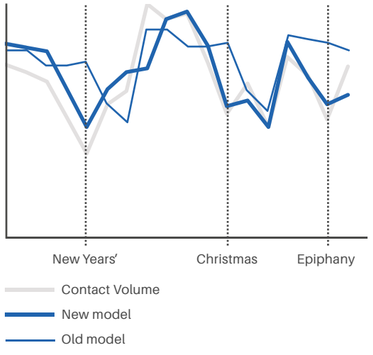

Figure 5. Old model, New model and Actual contact volume

"The results of the forecast project exceed our expectations. The model is much more accurate and the project gave us fundamental insight in our customer's behavior. When we scaled up the model to all other 41 languages, all but 2 language forecasts improved. It is a major step-up in the way we manage our Customer Service centre, and has also brought us a new set of capabilities that make this impact lasting. What was the result or our joint approach to the forecasting challenge? We measured the results of our model by the mean of the absolute % error (MAPE) of the forecast."

Erik Benson, Head of Customer Services at Booking.com

1. An intuitive – but rather lazy – answer would be either 500K (200K extra each year) or 900K (times three each year).
2. A booking is active between the day the reservation is made up until several months after the check-out.
3. We distinguish the phases pre-check-in, check-in day, during trip, check-out day and post-check out.
4. This driver can be best explained in an example: imagine today it is the 1st of January 2016 and we want to make a forecast for the 1st of March 2016. At this moment, 100 customers already made a booking with a check-in date for the 1st of March. Based on last year we expect these 100 check-ins to quadruple, since a lot of customers book less in advance. So, we expect to have 400 check-ins by the 1st of March. We also expect: 500 check-outs, 200 bookers that are one day before check-in (checking in on March 2nd), 100 bookers that are one day after check-out, etc. Generally, we can predict the number of active bookings for all booking phases. We can also predict how these bookings will be distributed over customer segments and booking types like hotels and apartments.
5. The correlation between the driver and the contact volume is above 0.90.
6. The correlation between this driver and the contact volume was 0.96.
7. This means if the forecast is 950 and the realisation is 1000, the error is 50 and the % error is 5%. The MAPE is the average of the % error over the whole forecasted period.