Smart Send Time: How our data science team developed our automated send time optimization feature

Author: Christina Dedrick

Claps: 187

Date: Aug 16, 2019

Klaviyo released Smart Send Time, our send time optimization feature, last week. Smart Send Time makes it easy for brands to send their emails at the best time for their recipients.

In this post, we will take you through the journey of how we worked alongside 35 of our brands as developed Smart Send Time. Weâ€™ll show how and why it works. Weâ€™ll show why it outperforms industry standard personalized send time strategies and show the data weâ€™re standing on when we say we can help a brand achieve a +10% lift in open rate. Itâ€™s a long story, so in modern fashion, hereâ€™s the TL;DR version before we dive in.

Key findings:

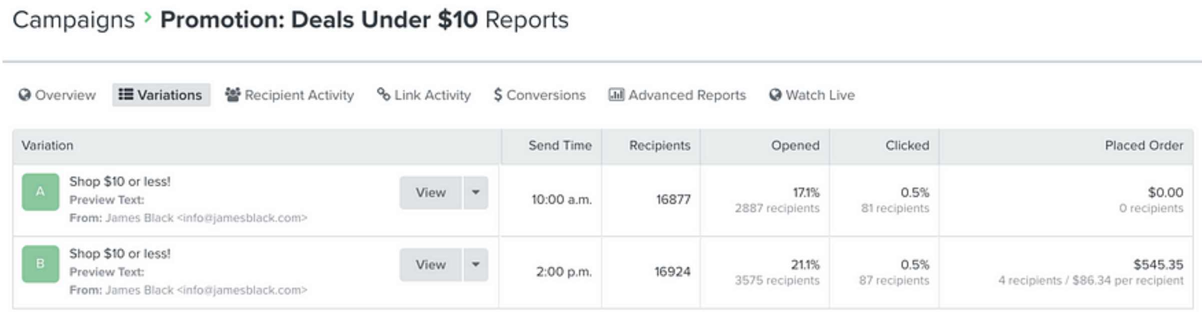
- Sending email at different times of the day leads to different open rates
- Using historical sent/opened data to make either list wide or personalized predictions does not provide substantial lift. Why?
- Missing data â€" Most companies have never tested an afternoon or evening send time, so we canâ€™t make any inferences about those times of the day
- Survivor bias â€" Most opens happen shortly after an email is sent out, so we canâ€™t transfer learnings and make inferences about how recipients interact with different send times without directly testing them
- Overfitting â€" Models tend to overfit to just a few data points that are not representative of a recipientâ€™s true activity
- Klaviyoâ€™s Smart Send Time lets us collect unbiased data about how recipients react to different send times by using automated A/B testing, the only way to do a true comparison of different send times.
- Using an exploratory send to learn about recipients, we are able to find a list wide optimal send time for a brand in just a few sends
- Smart Send Time focuses sending in on the optimal send time but continuously tracks and monitors recipients to make sure weâ€™re always sending at the best time
- Smart Send Time has been in private beta since April. 15 brands tried it out and saw an average of a +10% lift in open rate between their previous business practice and their optimal send time.

Why send time optimization?

We chose to research send time optimization because we heard about it from Klaviyo customers. Customers wondered if data science research had solved the question about how time of day affected recipient engagement. They were eager for best time of day practices in maximizing engagement. Weâ€™ve also been surveying customers at this yearâ€™s Klaviyo workshops, and itâ€™s consistently one of the most popular new feature ideas.

Can varying send time create lift?

The first step in our research journey was to see if optimizing send time was something that provided value for customers. Before investing time and research into optimizing send time, we wanted to confirm that time of day had an effect on email performance. So, we looked for campaigns where customers had A/B tested sending time.



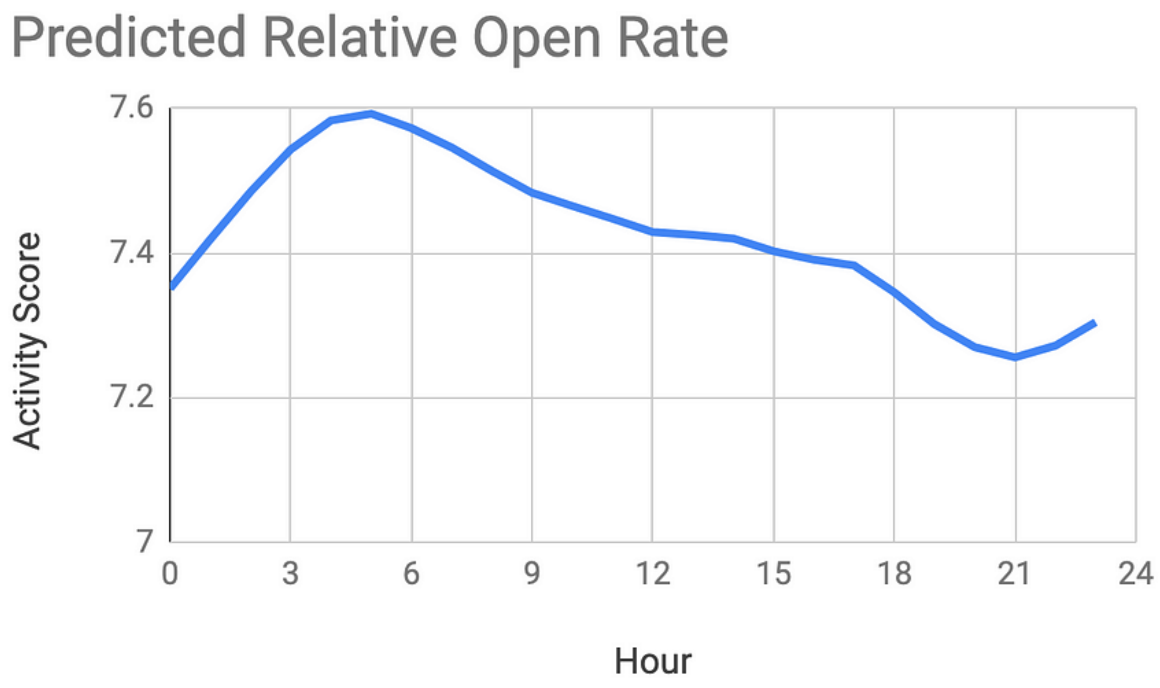
Anonymized historical campaign from April 2018. 2 p.m. has a much higher open rate than 10 a.m.

We saw that there were open rate and click rate differences for sends at different times of the day. This told us that there were better and worse times to send an email. While we didnâ€™t know what caused some hours of the day to do better than others, it showed that brands could see lift by sending at their best time. It was enough to move forward with a thorough investigation of how we could maximize recipient engagement by sending at the best time.

Can we learn about optimal times based on past email recipient behavior?

Klaviyo has a ton of data about how and when recipients interact with email. We hoped that we could use this data to build a model of what the best times would be best to send, so we started by predicting the best hour of day to send to an entire list.

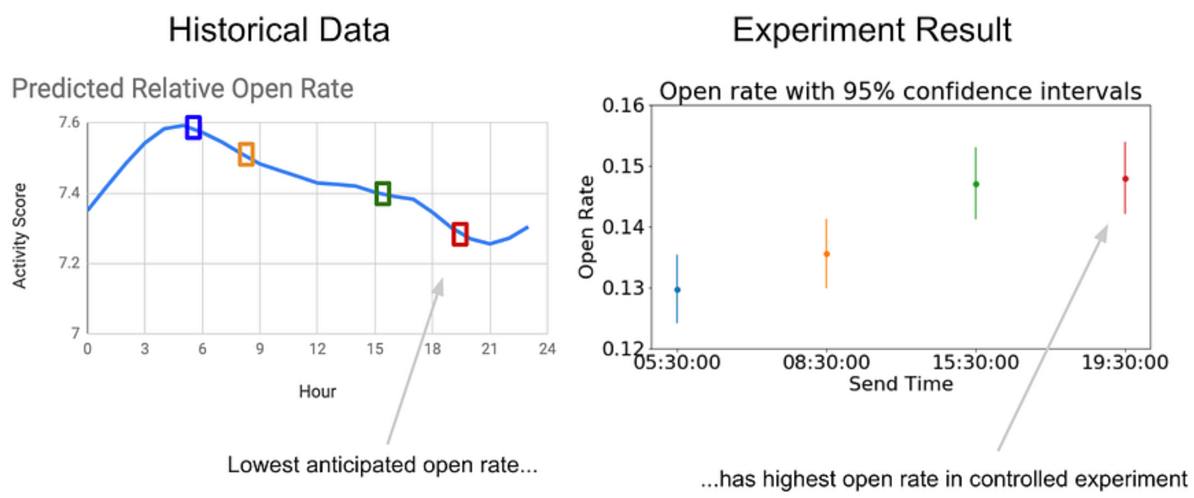
Our first model was a basic multivariable regression model where we predicted the percentage of email opens in an hour based on the hour of the day the email was originally sent and the number of hours it had been since the send. We controlled for the frequency of sends at different times of the day. Things were looking really good. We were able to plot how recipients interacted with email by time of day.



Morning hours were predicted to have the highest open rate for this Home + Gardens brand

Our predictions looked great when we held back data and tested our models. Excited, we asked some customers to be part of beta tests for our new models of send time optimization. We set up A/B tests of the different hours of the day and worked with five customers to test our research.

And for all five cases, our predictions failed. And not just by a little. They were completely backwards. The times we had predicted to be the best had the lowest open rates and the times we expected to do worst were most engaging.



Our predictions did not hold up in an A/B test of send times. Open rates are plotted with 95% confidence intervals.

While communicating results like these to customers was embarrassing, failures like these are actually my favorite types of experiments. A lot of life advice says you learn most from your failures, and in science and engineering, we have a name for that process: root cause analysis. Root cause analysis involves looking at your model, experiment, or solution and digging into the specifics of what went wrong with it to find the underlying causes of the observed failures.

We had five experiments all fail in similar ways. In all five, the historic data was a poor predictor of how an A/B test of send times would perform. We also saw a general trend that morning send times were performing poorly and afternoon and evening send times were doing better. Somehow, morning opens were being weighted too heavily in our model. But why was the data biased? Why were morning opens overrepresented?

What is survivor bias?

Survivor bias is a form of selection bias where only the things or events that make it past some sort of threshold are considered in the data set. Before we talk about how this relates to email opening behavior, letâ€™s talk about one of the most well known stories about survivorship bias, Abraham Wald and the Case of the Missing Bullet HolesÂ[1].

The year is 1942. The armed forces are deeply involved in WWII and looking for ways to operate with the most efficiency. As part of the war effort, the Statistical Research Group, an assembly of statisticians, was formed to work on difficult operational problems. One problem they identified was that planes return from battle full of bullet holes. How could they armor planes to reduce damage? Armor was heavy and slowed planes down â€” so what parts of the plane are most important to protect?

The armed forces collected data on where bullet holes were located on planes returning from combat. Different parts of airplanes had different densities of bullet holes. So where is the most optimal location for armor?

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel System	1.55
Other	1.80

The â€˜otherâ€™ parts of the plane, the wings and tail, have the most bullet holes. So, if you have a model that says more bullet holes = more damage = higher cost to repair, youâ€™d be inclined to armor the most damaged parts of the plane.

But youâ€™re missing something. Youâ€™re only measuring returning airplanes, the survivors. What about the planes that do not return? How can you measure the most important bullets, the ones that bring down an airplane? The answer is you canâ€™t. Those planes arenâ€™t part of the data set. Wald concluded that the most important part of the plane to armor were the areas with the least damage since those were the bullet holes missing from the data set.

So how does this apply to emails? Emails are another survivor scenario. Most emails are opened as soon as they go out, meaning the email open data we collect is localized around the send time. Also, emails are usually only read once, which creates an environment where every hour passing means fewer and fewer emails survive. Together, these factors mean itâ€™s very difficult to tell how a later hour performs based on an earlier hour. If an email is sent at 10 a.m., it tells us very little about how that email would perform at 5 p.m. The only way to tell is to compare directly: send it at 5 p.m., then compare the results with an A/B test. Going back to our bullet holes, weâ€™re missing bullet holes from recipients at 5 p.m. because they were already recorded at 10 a.m. when they received the email.

While Klaviyo is sitting on a treasure trove of data about when recipients have interacted with email in the past, we were unable to use this to make optimal send time predictions because of the survivor bias in the data. We tried all kinds of bias detection and correction techniques. They didnâ€™t help. We had tons of data at our disposal, but it was less than useless, it was leading us in the wrong direction.

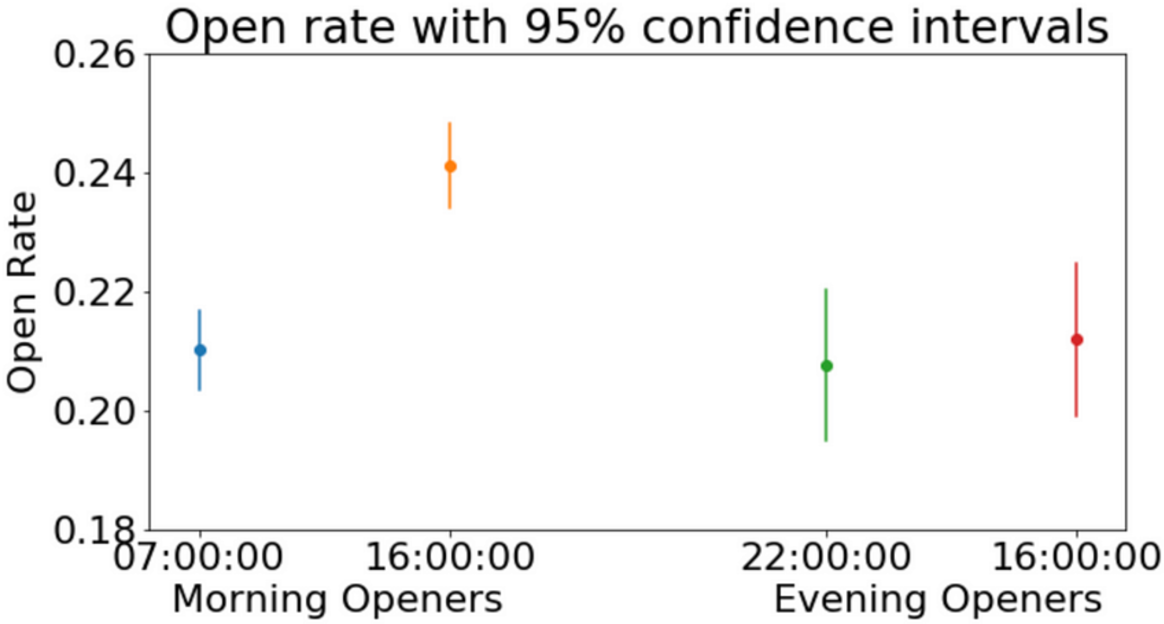
Can we create lift using personalized send time methods?

While we were waiting for the experimental results from our list wide send time experiments, we started a series of personalized send time experiments. We took past engagement behavior and fed it into various clustering methods to assign personalized best send times to recipient. Then, we sent to recipients at their personalized best times as well as control times throughout the day. We worked with 6 brands to test how personalized strategies did compared to sending at their previous business practice time.

Company	Open rate at previous business practice	Open rate for personalized time	Lift from using personalization
Fashion + Apparel	18.0%	18.3%	+ 2%
Animal + Pet Care	31.6%	31.8%	+ 1%
Fashion + Apparel	27.2%	28.0%	+ 3%
Beauty + Cosmetics	17.0%	18.0%	+ 6%
Adult Products	16.5%	17.2%	+ 4%
Fashion + Apparel	20.1%	19.3%	- 4%
			Median: +2.5%, Mean +2% std = 3.4%

After a series of experiments, we had measured a 2.5% lift in open rates. This was a lot better than our list wide model that was predicting backwards! But, we still needed to understand our results. And we had to explain why we had one company who had worse open rates at their personalized send times while others were successful.

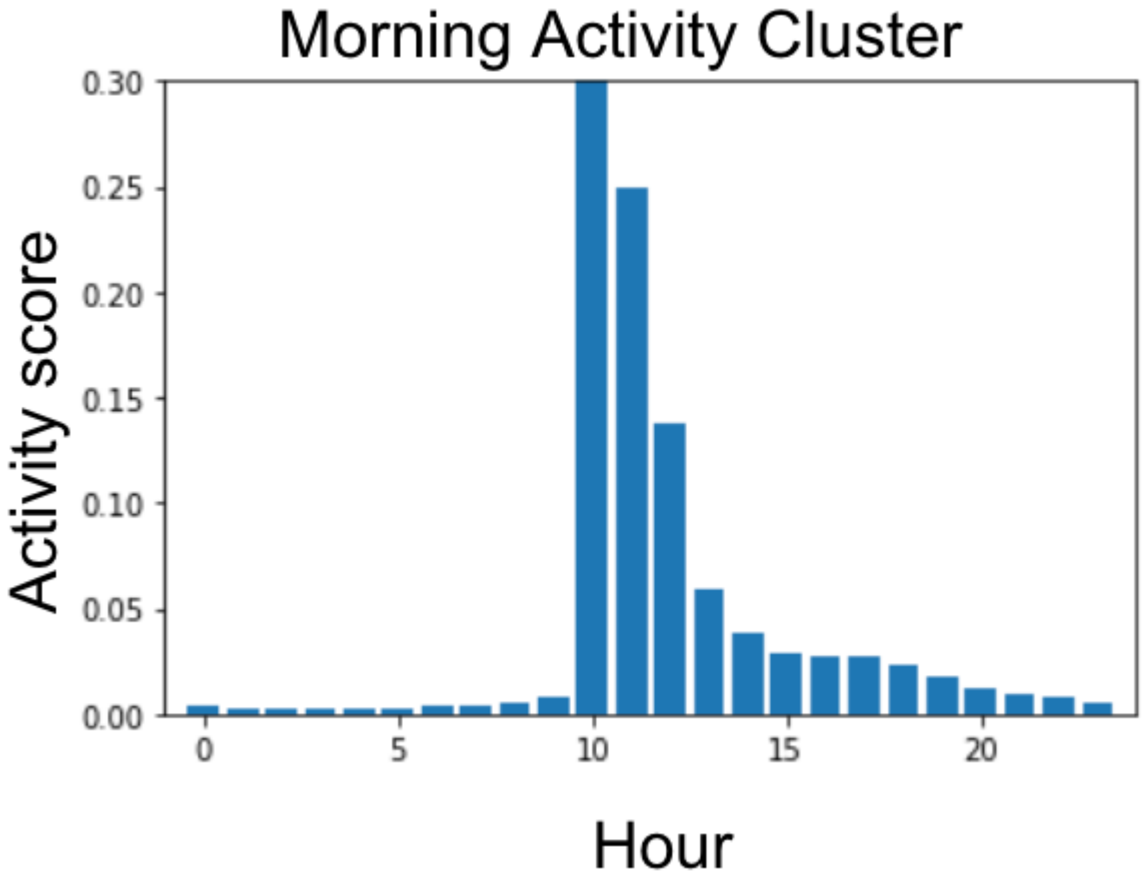
Once we started with root cause analysis on our personalized send time models, we started to see that they were problematic. We werenâ€™t getting lift because we had identified good send times for people, we were getting lift because we had mixed up the send times. For instance, in our experiment with a Fashion + Apparel company, we saw the following results:



Two clusters were personalized send times didnâ€™t have higher open rates. Open rates are plotted with 95% confidence intervals.

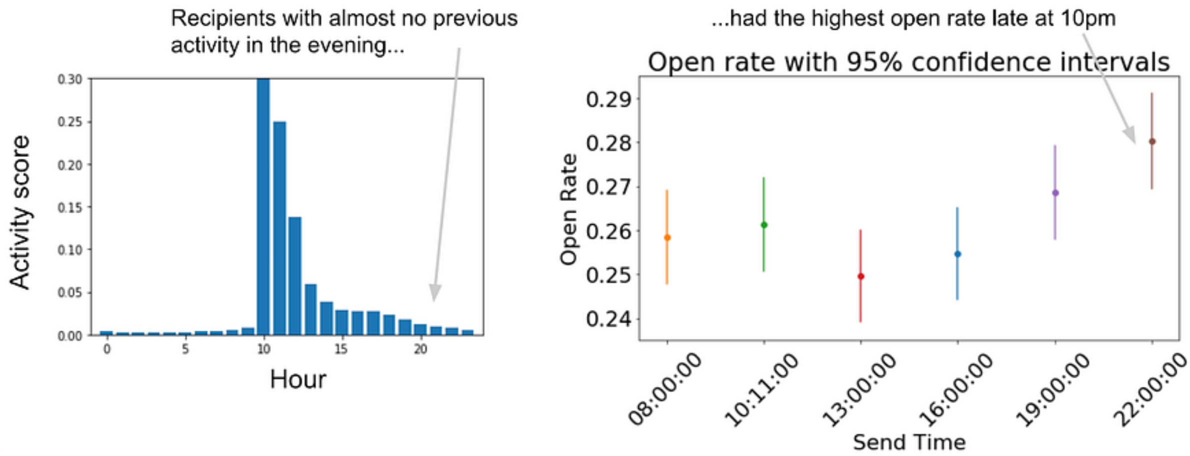
On the left we have our cluster of morning openers â€” people whose past behavior shows they are more active in the morning. We sent them an A/B tests between their 7 a.m. personalized time and a 4 p.m. control time. Which time did they prefer? 4 p.m, meaning the personalized send time has a lower open rate when we send at these recipients personalized send time. Another example on the right is our evening openers. Based on their past behavior, they should be more active in the evening. But, in the A/B test, 10 p.m. and 4 p.m. performed exactly the same. Why werenâ€™t these recipients responding better at their personalized send times?

In another experiment, we focused on a Beauty + Cosmetics companyâ€™s morning openers. We found morning openers by clustering recipients by the times of days they were active and identifying people who opened emails in the morning. We show their activity score by time of the day â€” times of day with a higher activity score were times of day the recipients typically interacted with email. Our morning cluster was very active at 10 a.m. with decaying activity throughout the day.



The activity of customers in the morning activity cluster

We tested sending email at six different times spaced throughout the day to this group, expecting to see the highest open rate at 10 a.m. when they were most active. But, just like our failing list wide models, our cluster behavior prediction was wrong. 10 a.m. was not the best time for these recipients. 10 p.m. had the highest open rate, despite almost no historic activity at that time. And there were other oddities, 8 a.m. did approximately as well as 10 a.m. despite showing almost no historic activity at that time. While our aggregate results were looking ok, when we looked at how personalized models were doing for individuals, we saw they were failing.



Customers historically most active in the morning had the highest open rate in the late evening. Open rates are plotted with 95% confidence intervals.

After investigating, we found personalized send time models were hugely vulnerable to all kinds of bias. Theyâ€™re biased by the times of day when brands send their emails â€” meaning weâ€™ll see peopleâ€™s open activity cluster around the times of day emails get sent. Theyâ€™re vulnerable to the same types of survivor bias described above. They also horrifically overfit to a few data points or timestamps.

Weâ€™ve heard stories of personalized models that fail more spectacularly than the ones we created. Some use the last time a recipient was active as their optimal send time. We talked to a marketer who opened a testing email at 1 a.m. during a product launch. Thereafter, they received all their own brandâ€™s emails at their 1 a.m. personalized time even though they typically went to sleep at 9 p.m. Thatâ€™s not personalization. Itâ€™s bad models making bad inferences about people. Itâ€™s bad science and ultimately it doesnâ€™t create highly engaged experiences with recipients.

At this point, we were getting really worried about Send Time Optimization. The past data was biased. We could only get an unreliable sometimes lift out of personalized models. But we had hope. Data science isnâ€™t just about building really good models on the data we have. Itâ€™s also a field where we can design how we collect our data. The data we had was biased. But what if we could design a way to collect unbiased data? What if we changed the problem around â€” instead of trying to use past data to make predictions, what if we could create the type of data that actually measured how different send times performed?

What is the Klaviyo Smart Send Time method?

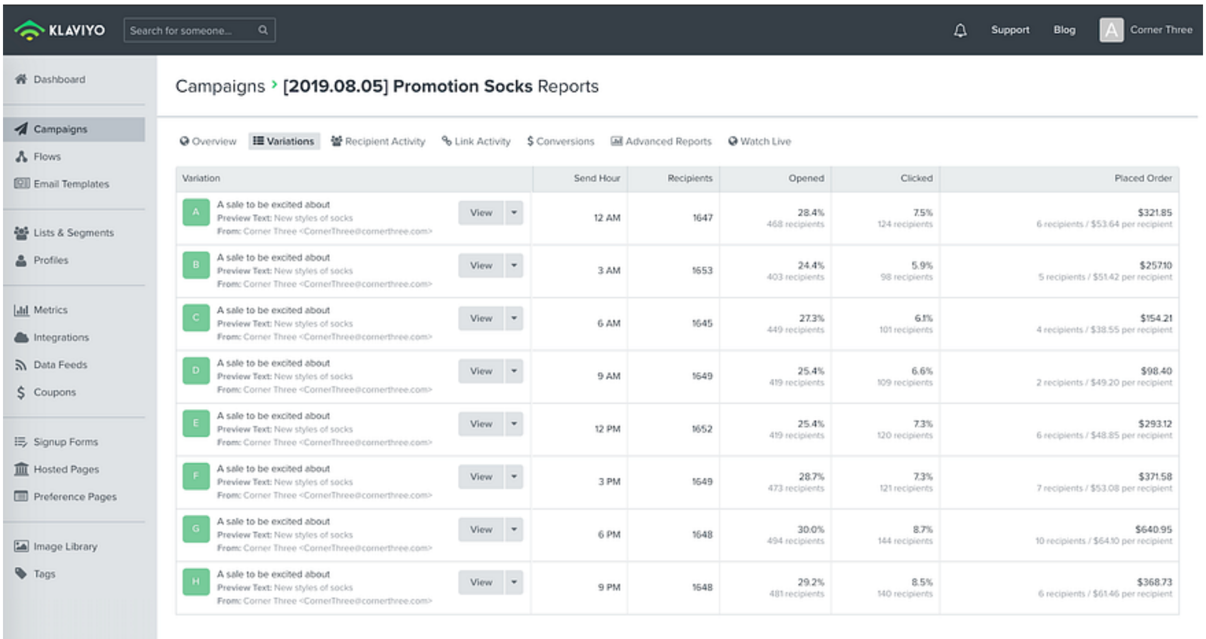
Klaviyoâ€™s Smart Send Time method automates the process of experimenting and finding the best send time for an audience. It rests on a key tenet of data science: the exploration-exploitation tradeoff. Smart Send Time first explores how recipients respond to email at different times of the day. Once we understand that, Smart Send Time exploits that information by sending at the optimal time, maximizing open rates, while also continuing to learn and explore customersâ€™ habits.

Smart Send Time learns about every hour of the day by setting up a 24-hour A/B test. By testing the entire day, we donâ€™t leave any gaps. And thereâ€™s no need to make assumptions about customer activity at different hours â€” we know how hours compare because we directly compare them! All together, this gives us complete certainty about how a brand can expect its customers to interact with email throughout the day. And once weâ€™ve explored all 24 hours to find the best one, itâ€™s just a matter of *exploiting* that information by sending at the optimal time.

For some list sizes, we only need one exploratory send to gather enough data to narrow to the optimal send time. For smaller lists, we might need several sends to discover the optimal send time. Our back end keeps track of that, and once weâ€™ve collected enough data, optimal send time becomes available as a sending option. Using this optimal send time exploits the learnings from the previous experiments. By focusing into sending at this optimal time, brands can maximize their open rate.

Smart Send Time isnâ€™t done after narrowing to the best time. It continuously explores by sending in a four hour window around the best send time to learn more about how recipients interact with email sent at different times. This continuous exploration lets us track any potential changes in the best send time so we can guarantee that weâ€™re continuously exploiting the most up to date send time. It allows us to adapt as we get more information.

Everything is done in recipient local time to make sure to reach an entire global or localized audience at their best time. And, all of the results for each campaign are available in the campaign variation tab. This allows marketers full transparency of the results from their sends.



Send time results update in real time and are available for each Smart Send Time send

What lift can we create using Smart Send Time?

Like our other send time optimization methods, we needed to put Smart Send Time to the test. We worked with 15 companies in a private beta starting in April to test out Smart Send Time. Smart Send Time consistently provides lift. Hereâ€™s the data we collected during the beta test, anonymized by our beta testerâ€™s industries.

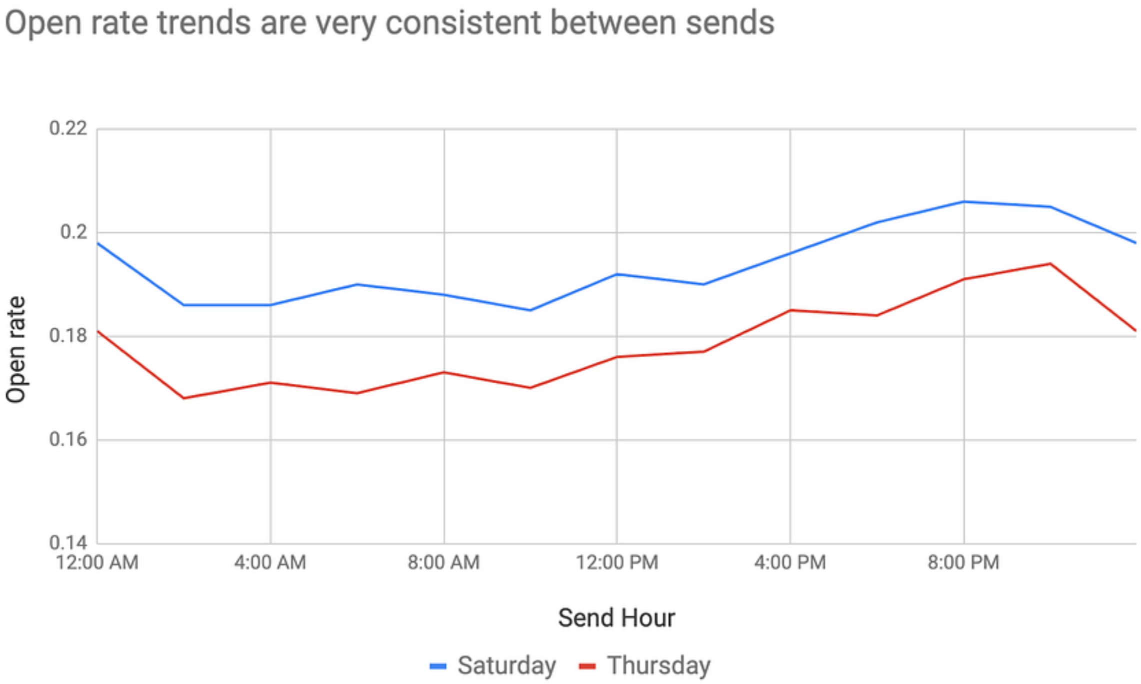
Company	Open rate at previous business practice	Open rate at optimal send time	Improvement
Home + Garden	19.3%	21.7%	+ 12.4%
Beauty + Cosmetics	13.2%	15.7%	+ 18.9%
Adult Products	6.3%	6.6%	+ 4.8%
Health + Fitness	20.3%	20.6%	+ 1.5%
Beauty + Cosmetics	14.2%	17.3%	+ 21.8%
Jewelry + Accessories	8.5%	9.1%	+ 7.1%
Fashion + Apparel	22.7%	25.5%	+ 12.3%
Home + Garden	18.3%	18.7%	+ 2.2%
Outdoor + Wilderness	6.7%	7.1%	+6.0%
Beauty + Cosmetics	15.2%	17.4%	+14.5%
Health + Fitness	6.5%	6.7%	+3.1%
Jewelry + Accessories	18.0%	19.7%	+9.4%
Jewelry + Accessories	19.9%	21.9%	+10.0%
Health + Fitness	18.4%	20.6%	+12.0%
Fashion + Apparel	3.8%	4.0%	+5.2%
			Median: +10%, Mean +10% std = 6%

If a brand is already sending close to its optimal time, theyâ€™ll see a small change in their open rates. However, if theyâ€™re far away, they can see up to a 20% increase in open rates.

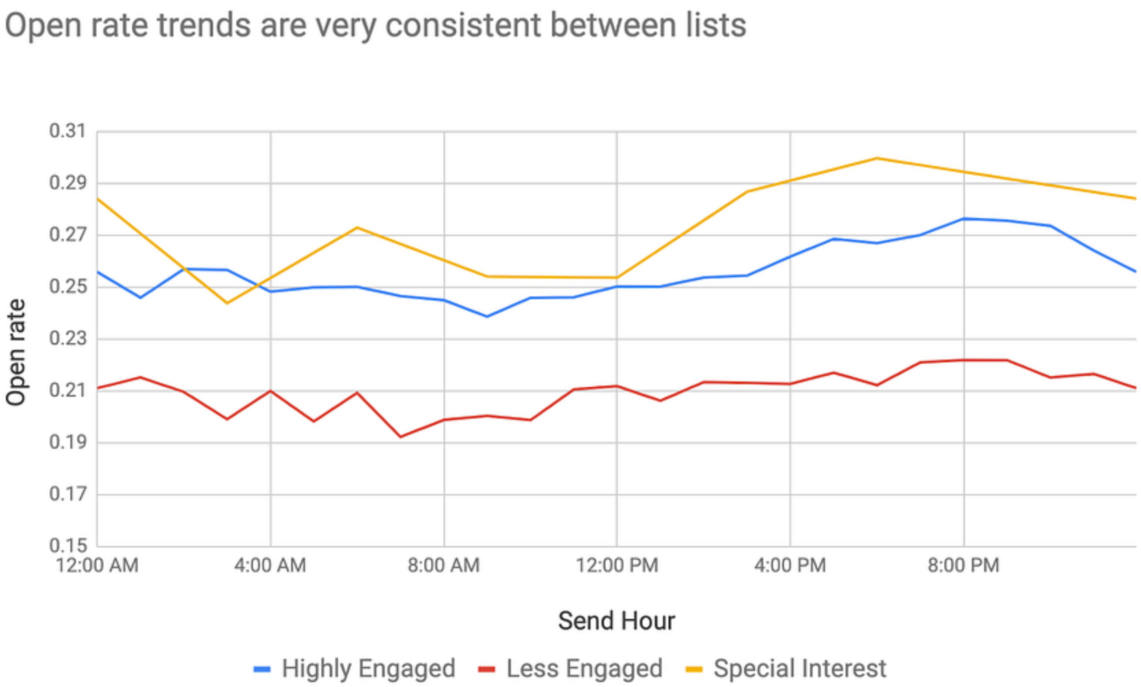
When is the best time to send an email?

The best time is when A/B tests of send time set up through Smart Send Time show the highest open rates. For most brands, this was an afternoon or evening send time. And generally, morning send times were among the worst time to send an email. So, any brands sending email in the morning is potentially losing engagement and revenue if their recipients prefer evening communication.

We saw that time of day trends were consistent between different days of the week. A Health + Fitness brand we worked with wanted to confirm they could use the same send time for their Weekday and Weekend campaigns, so they sent one exploratory send on a Thursday and one on a Saturday. Their open rate trends were the same on both days: low in the morning, increasing in the afternoon, and highest from 8â€“10 p.m.



Time of day trends were also consistent across engagement level. A Fashion + Apparel brand sent their exploratory send to three mutually exclusive lists â€” a special interest group, their highly engaged list, and their less engaged list. Again, the time of day trends were the same for each list.



Why canâ€™t we use a personalized A/B test to find personal best send time?

Smart Send Time automates collecting the best send time for a list. Why canâ€™t we do that at a personalized level? Why didnâ€™t we build a feature to continuously learn and test how each recipient responds to different send times?

To test individual send time, we need to test all times of day for each recipient. If we divide the day into three hour blocks, weâ€™ll need to test eight different send times. Then, we need to send multiple emails at each time to measure open rate for each time. How many do we need to send?

Weâ€™ll cover this topic in depth in a blog post next week, but letâ€™s briefly look at a best case scenario. Our recipient opens 20% of emails sent at non optimal time and doubles their open rate to 40% at their optimal time. We test by sending ten emails to each three hour block, a total of 80 emails. Then, we see if theyâ€™ve opened the most emails at their most optimal time. In addition, weâ€™ll assume we have perfect correction for different subject lines during these experiments. Working through the math, we find that they have the highest open rate only 46% of the time. So, after 80 emails, weâ€™re at just less than a coin flip for getting their optimal send time. It would take 440 emails to correctly guess the best send time at a 95% accuracy rate. And thatâ€™s the best case scenario.

Letâ€™s talk about what we saw in our data. We saw a +10% lift by sending at optimal send time, not a doubling in open rate. So, weâ€™re moving people from a 25% open rate to a 28% open rate. How many emails do we need to test between those two different times? We canâ€™t measure the difference with 80 emails. In fact, after sending 3,600 emails, 450 to each of the 3 hour blocks, weâ€™ve only can guess the best send time correctly 50% of the time.

Personal A/B tests would theoretically give the best personalized send time, but the math simply doesnâ€™t work out on any reasonable timescale. This is why personalized models fail. They canâ€™t collect enough data to accurately predict which times of day a recipient is most active.

This is why weâ€™ve implemented data requirements to narrow to a best send time. So we donâ€™t overfit and lead a brand in the wrong direction, we require a list of at least 12,000 recipients to start exploratory sending. We donâ€™t want to provide a feature that overfits and starts sending at a time that isnâ€™t optimal.

Does an increased open rate imply an increased conversion rate?

This is another topic weâ€™ll do a deep dive on. The short answer is yes! For both Smart Sending campaigns and campaigns in general, a higher open conversion is correlated with a higher open rate.

Smart Send Time is part of Klaviyoâ€™s owned marketing mission

Klaviyo wants to give brands the power to send to their recipients at the best hour possible. We want to give our customers the tools to automatically test and act on their results, so they can own every step of the way. We donâ€™t want to hide data and learnings inside an algorithm. Smart Send Time is fully transparent about what times weâ€™re sending to recipients, meaning brands can own that interaction.

How do we know the optimal time weâ€™ve chosen is best? Because we tested it. We performed an experiment. We saw the results. We can find the optimal send time for any audience.