Personalized Email Marketing Analysis

Lin Chen, Gaurav Choudhary, Yuke Liu

Background and problem

Email marketing is a useful strategy to promote related products to existing users and to increase the number of potential customers landing on the product page. The success of such email marketing campaigns is usually measured by the click through rate on the emails. The quality and the timing of emails plays a key role in improving this click through rate. Another effective strategy is personalizing the emails for each potential customer. In this project we aim to measure the benefit of personalizing the content and time of marketing emails by comparing the success of such emails with that of a generic email sent at random times during the day.

Personalized marketing, also known as one-to-one marketing or individual marketing, is the practice of using data to deliver brand messages targeted to an individual prospect. This method differs from traditional marketing, which mostly relied on casting a wide net to earn a small number of customers.

There are several benefits of using personalized marketing:

- Effectively target specific audiences:The main benefit of personalized marketing is the ability it gives the companies to reach specific audiences. By collecting user data from list segments, surveys, or studies companies can create more effective email campaigns targeting audiences based on their interests or buying habits.
- Improving brand value: Personalization also helps companies stand out from the crowd by creating better and unique content that customers are likely to remember when they come across the product in an online or offline retail setting. This helps companies in improving their brand value.

- Build deeper relationships with customers: Personalizing marketing will also help build stronger and more personal relationships with customers. Companies can show how much they care about every one of their customers by showing their gratitude, by sending an email wishing for their birthday, or sending a thank you email on the anniversary of joining their email list.
- Boost sales and conversions: Personalized marketing is not just about connecting with the audience. It's also a great way to help customers and also grow the sales at the same time. A simple recommendation or a suggestion can help bring better results for the company in terms of higher revenue and has the potential of providing a superior user experience for the customer.

In our project, we firstly try to analyze the important features that affect the click-through-rate such as timing of the email, country, past purchase history and length of the marketing email sent. Then we segment users into several groups and create group-specific emails to different groups. We bring up the idea of personalization and find out that the estimated click-through-rate for personalized email is double that for a generic email.

We downloaded the email dataset from kaggle and built several models that helped us decide the best email strategy for each user with the goal of maximizing the click through rate.

Data Summary and Exploratory Analysis

Dataset

The dataset that we worked on had details on different users and their response to the marketing email. There were 8 columns in the dataset:

1. Email_id: Denoted the email id of the user. This is the unique identifier for each user to whom a marketing email was sent.
2. Email_text: This denotes whether the email sent to the user was long or short.

3. Email_version: This denotes whether the email sent to the user was generic or personalized.
4. Hour: This denotes the hour of the day when the email was sent.
5. Weekday: This denotes the day of the week when email was sent.
6. User_country: This denotes the country where the user is from.
7. User_past_purchases: This denotes the number of purchases the user has made in the past.
8. Clicked: This denotes whether the user clicked on the email or not.

Exploratory Data Analysis

We used Python to perform Exploratory Data Analysis on the dataset. The key findings from the EDA are mentioned below:

1. There are 99950 rows in the dataset and no missing values

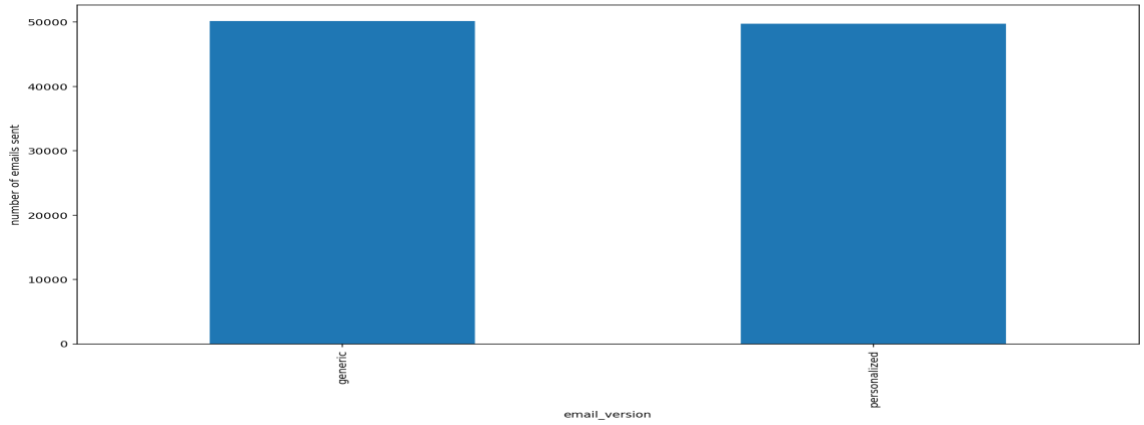
```
email_id              99950 non-null int64
email_text            99950 non-null object
email_version         99950 non-null object
hour                  99950 non-null int64
weekday               99950 non-null object
user_country          99950 non-null object
user_past_purchases   99950 non-null int64
clicked               99950 non-null int64
```

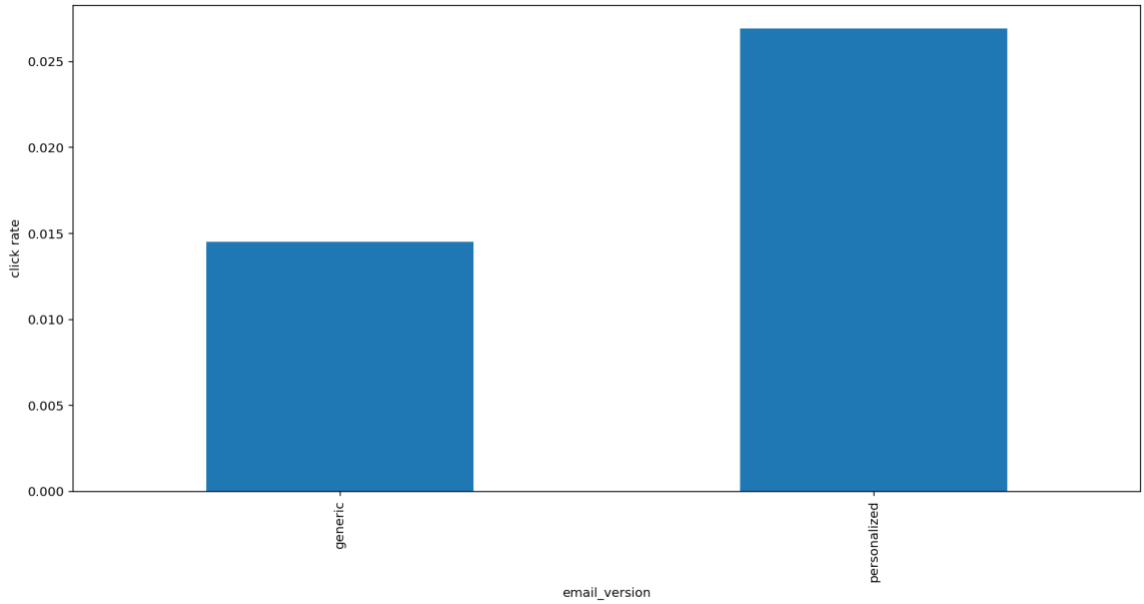
2. We observed that only 2.07% of users clicked in the dataset. This should serve as the baseline for any classifier built for this dataset.

```
df['clicked'].value_counts(normalize=True)
```

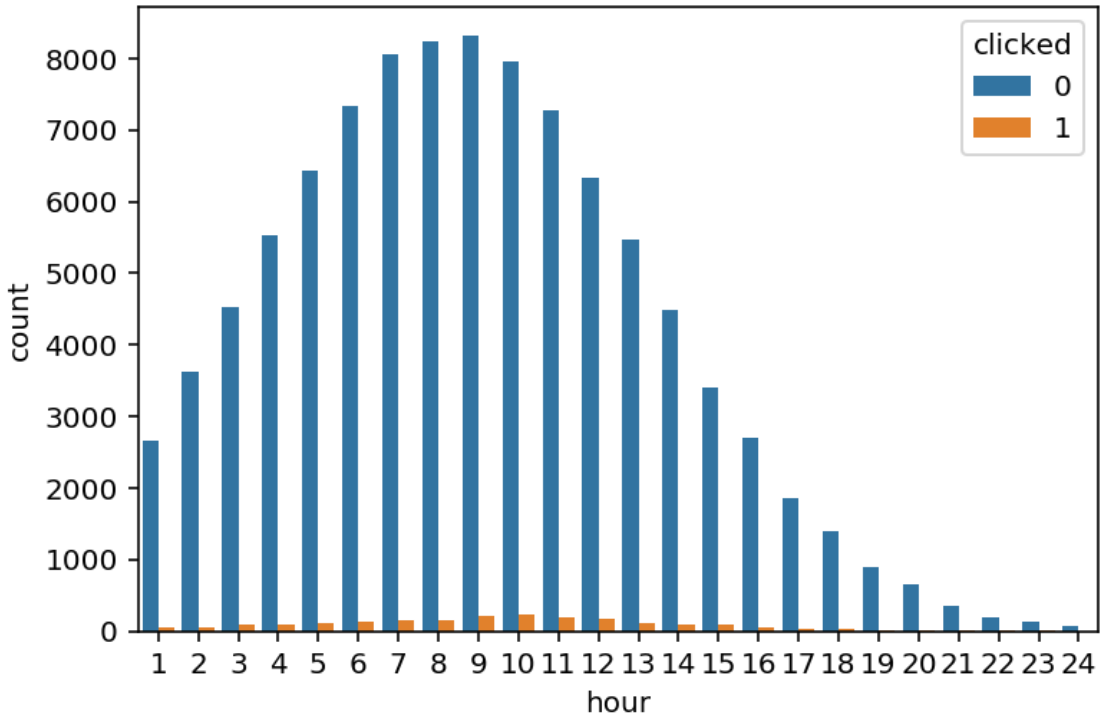
```
0    0.9793
1    0.0207
Name: clicked, dtype: float64
```

3. Almost an equal number of emails were generic and personalized. However, the response rate was double for personalized emails.

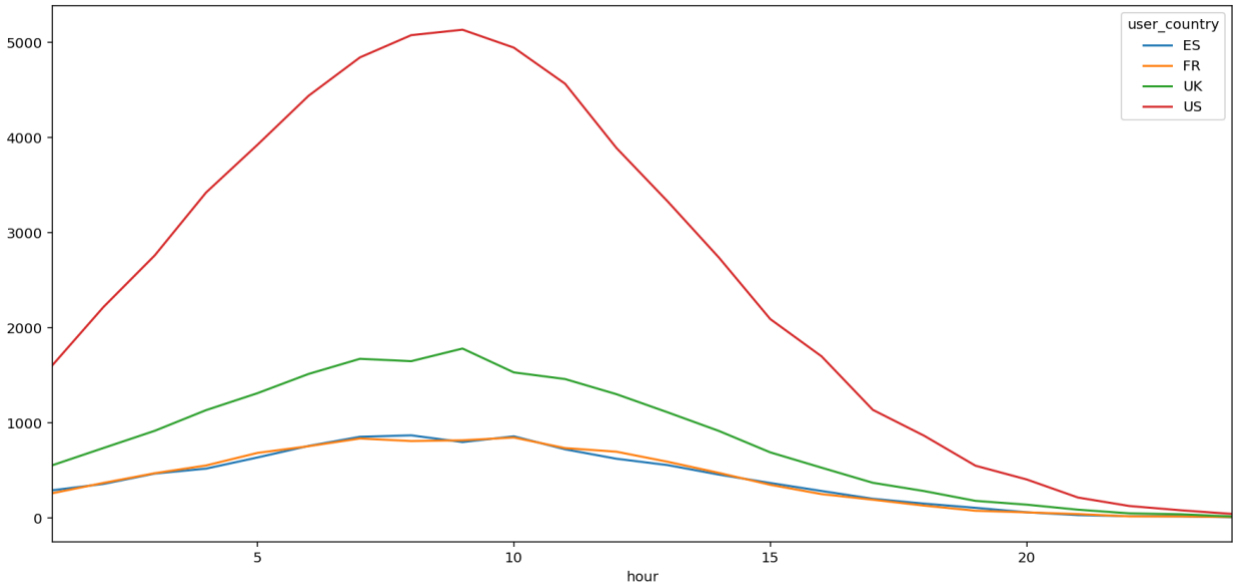




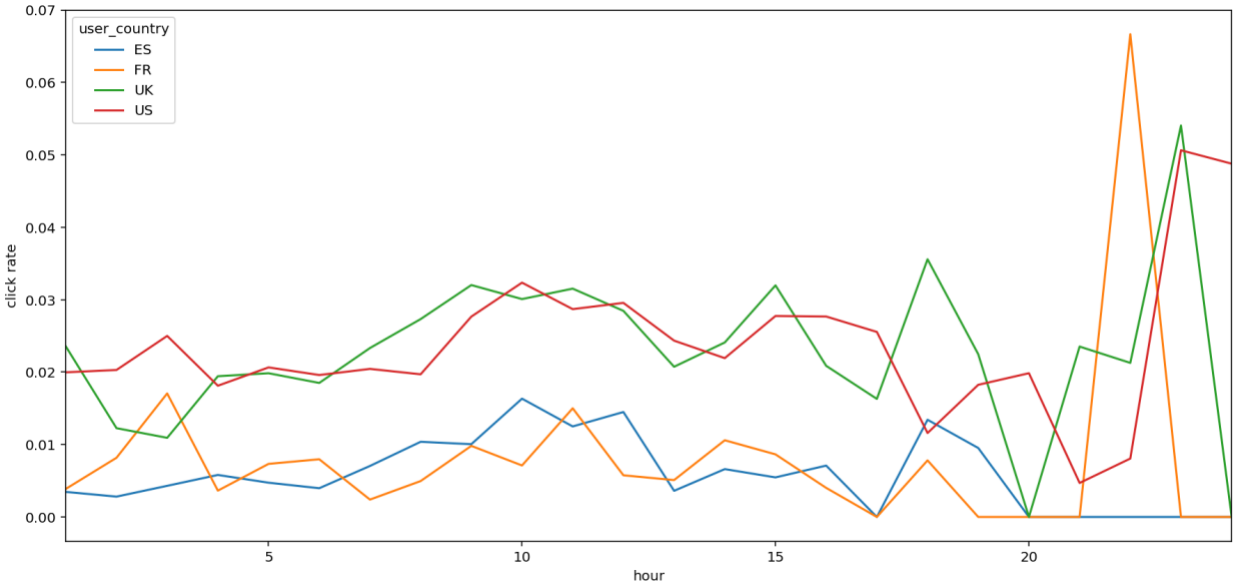
4. Most of the mails were sent early in the day with the peak around 9 AM. Correspondingly the highest number of responses were also made around morning.



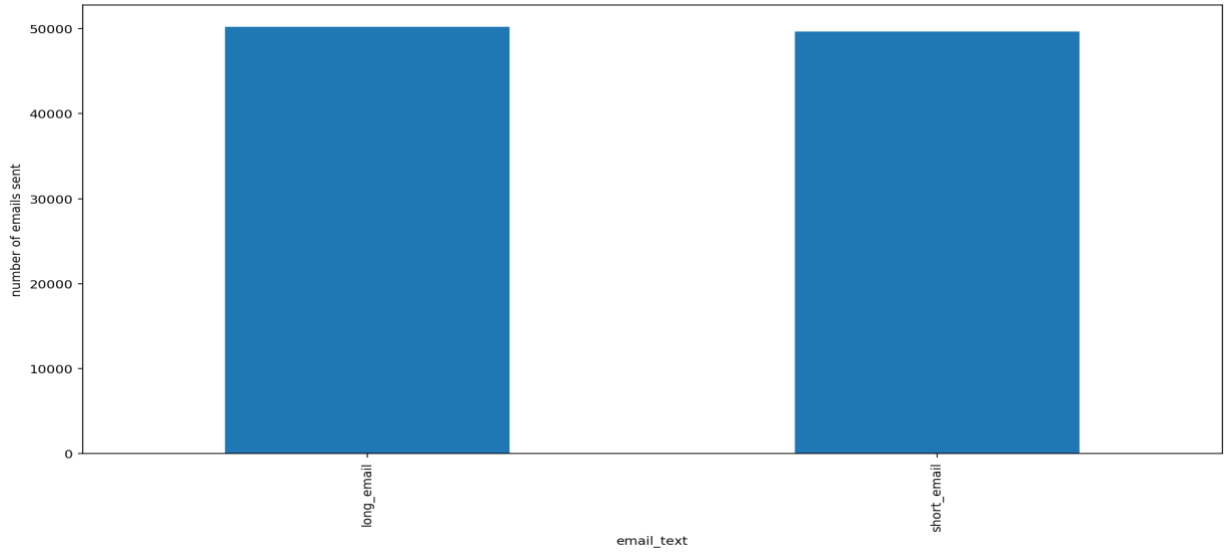
5. This trend of sending most emails early in the day was also consistent across all countries. Most emails were sent from 5 AM to 2 PM. We believe that the rationale behind sending these emails early is that it would ensure that users see the marketing email first thing in the morning, much like traditional marketing flyers sent along with newspapers.

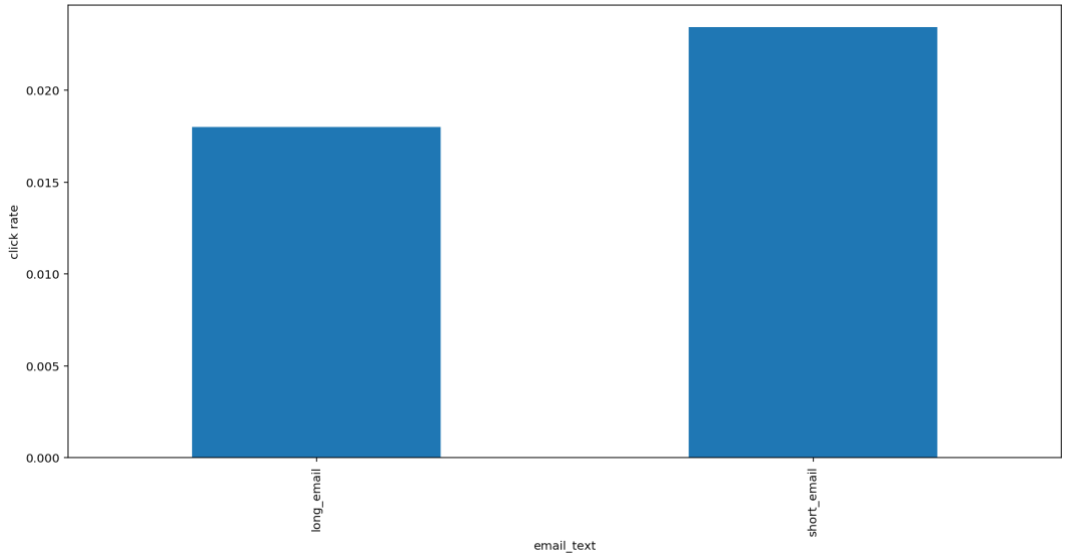


6. Although most emails were sent early in the day, these hours did not ensure a high click rate on emails. Highest click rates were observed after 6 PM, with click rates peaking around 10 PM. This trend was consistent across all countries. This could be because users are perhaps on their way back from their offices or classes and hence more likely to check any unread emails once they are less busy.

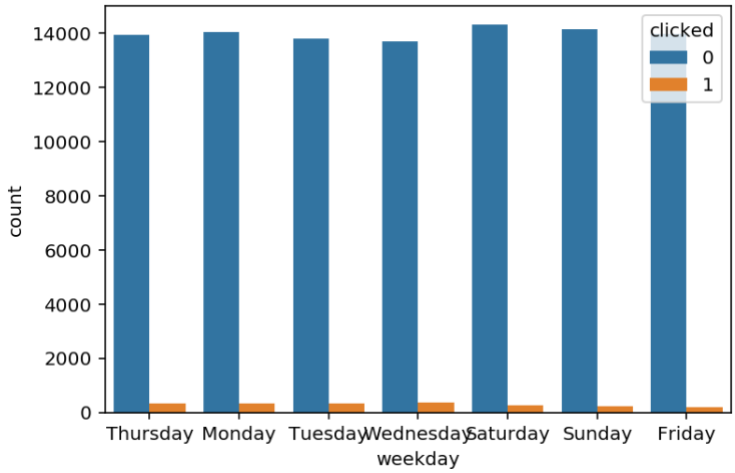


7. Similar numbers of emails were in the long and short email category. The click rate was higher for shorter emails than for longer emails. This suggests that shorter mails may be more successful in generating response from customers, perhaps because they are concise and require less attention.

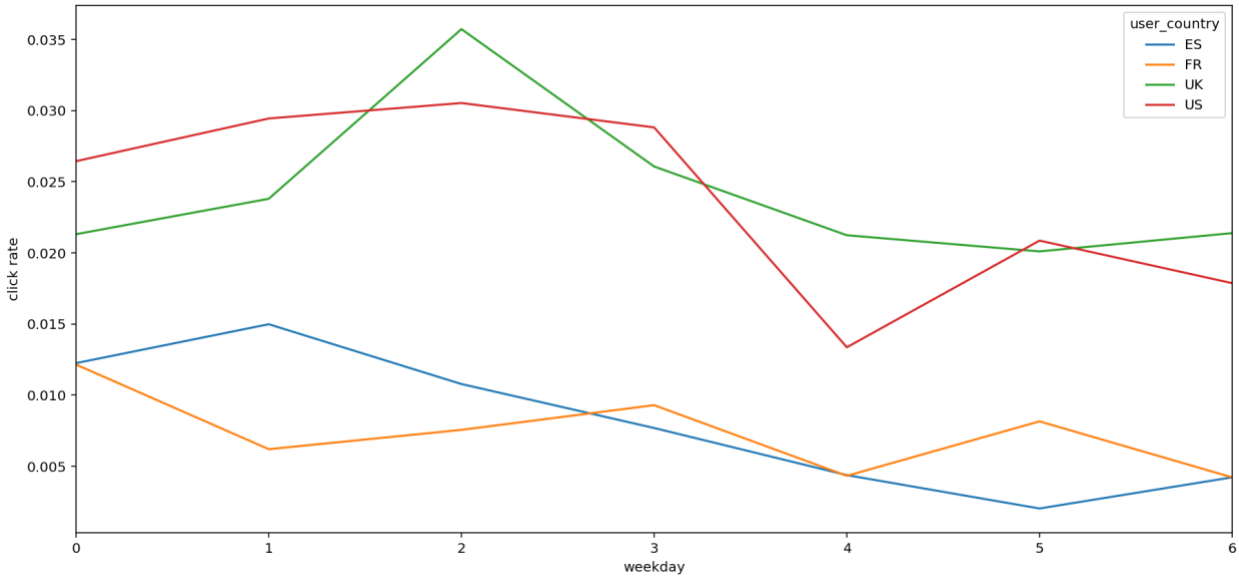




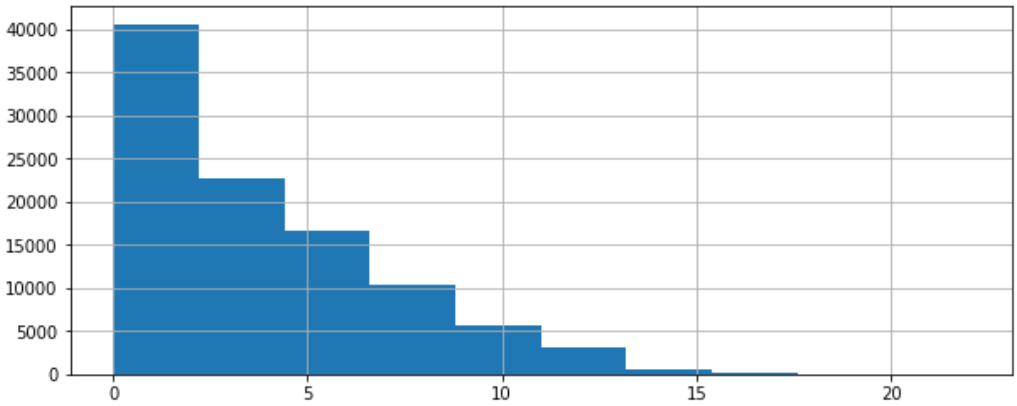
8. Highest number of emails were sent on the weekends (Saturday and Sunday) however emails sent on weekdays generated a higher click rate. This could be because users are less likely to check their mail when they don't have office or school work and hence their emails remain unopened during weekends.



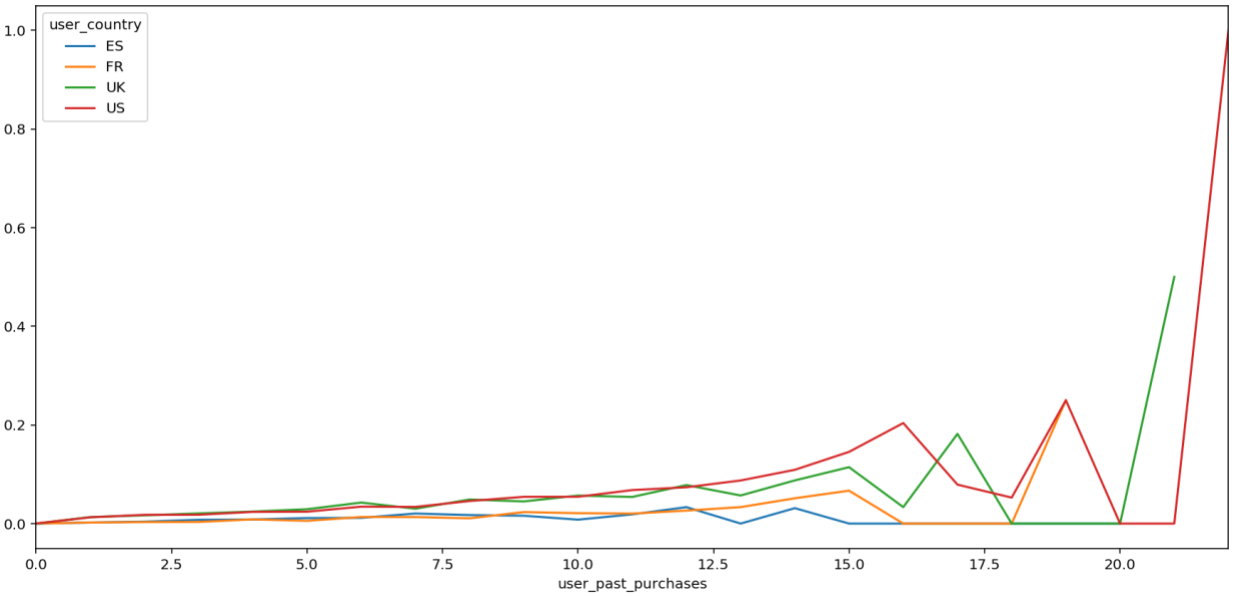
Click rate peaked on Tuesday or Wednesday in most countries and declined on the weekends.



9. Most users had made between 0 and 5 past purchases. This seems logical as marketing emails are sent to either new users or those who haven't ordered a lot. Sending marketing emails to users who are already good users might prove to be a waste of marketing expense.

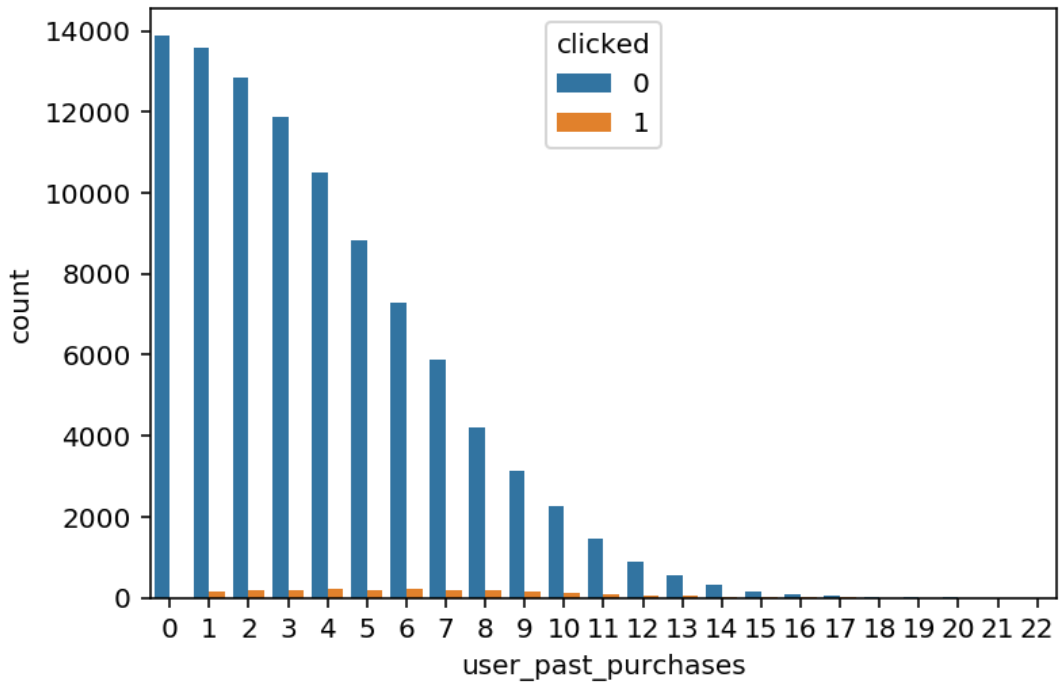


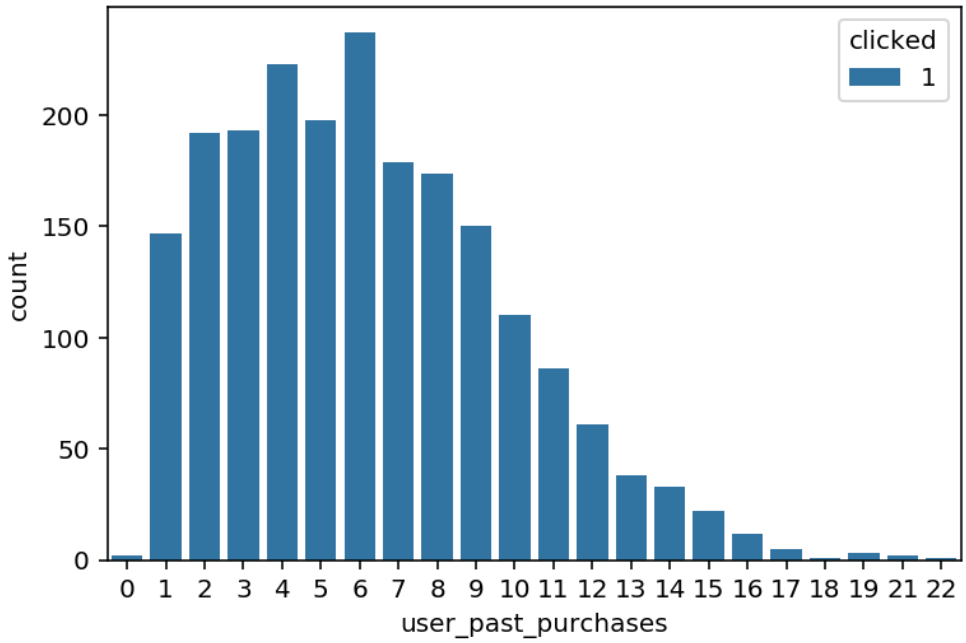
10. Despite getting the lowest number of marketing emails, the highest response rate was from users who had made more than 15 past purchases. These are the most enthusiastic customers for this company and are hence more likely to read communication from their favorite company regarding new offers and deals. However, sending offers and discounts to customers who would anyway have made a purchase might lead to lower revenue.



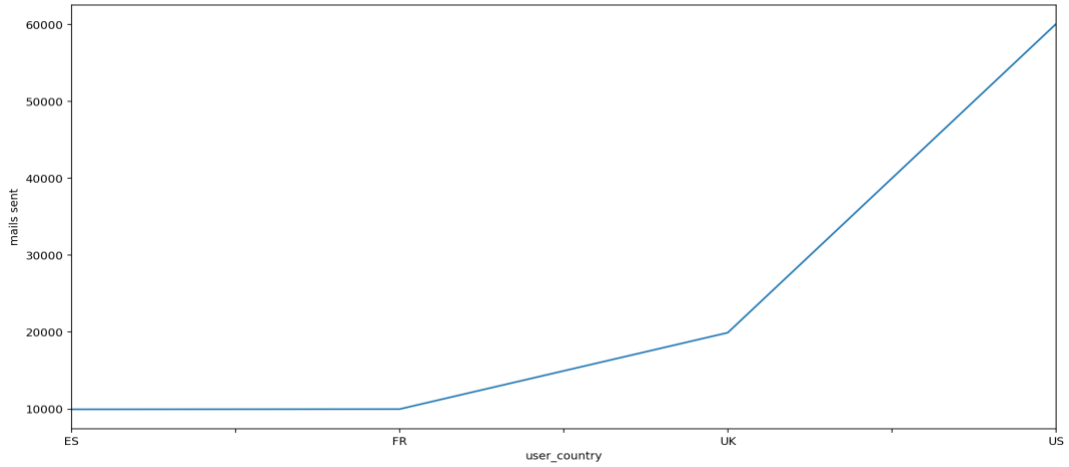
11. Although most users who received the email had made between 0 and 5 purchases, the highest number of clicks came from users who had made

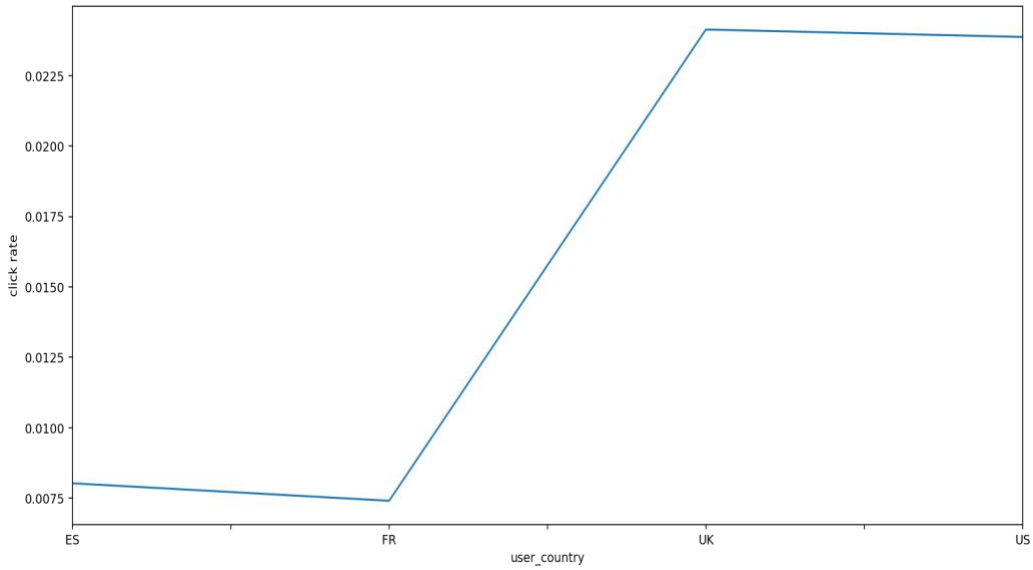
between 2 and 8 purchases. This is because they received enough emails and had a high enough click rate





12. Most of the emails were sent to users in the United States. Response rates were also highest for users in the US and UK.





Data analyses, key findings and conclusions

- **Extracting insights from user characteristics and product characteristics**
  At the beginning of the analysis, we want to dig into the user characteristics and product characteristics to extract insights about their effect on the click through

rate. For the binary classification, we'd like to utilize logistic regression and probit regression.

From the logistic regression, we can summarize some interesting findings.

- The summary shows that short email is better than long email, increasing the log odds ratio of being clicked by 0.279.
- The large coefficient is the email version and it shows that personalized email type is way more better than generic email type.
- In terms of hour, the coefficient of hour is significantly positive and it means that the latter the email sent to the customer in a day, the larger probability it would be clicked.
- Regarding weekdays, taking Friday as a benchmark, we can see that other weekdays all perform better than Friday. Within the weekday, although sending emails on Saturday and Sunday is better than on Friday, sending emails on Tuesday, Wednesday and Thursday is the best choice.
- As EDA has mentioned before, there are 4 countries(US, UK, FR and ES). Taking ES(Spain) as the benchmark, users in the US(United States) and UK(United Kingdom) are most likely to click the email. The country effect is the same for the US and UK. Although, the coefficient of FR(France) is negative but it isn't significant. users in FR and ES have no difference in clicking the email.
- Having Last purchase is also important and plays a positive role. The more purchases the customer makes before, the more likely the customer will click the email.
- We also post the regression result using the probit link function. There isn't any difference between sign and significance except for the value of coefficients.

```
Call:
glm(formula = clicked ~ email_text + email_version + hour + weekday +
    user_country + user_past_purchases, family = binomial(link = "logit"),
    data = email)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1654  -0.2218  -0.1669  -0.1248   3.5427

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -6.899891   0.151305 -45.603  < 2e-16 ***
email_textshort_email     0.279245   0.045304   6.164 7.10e-10 ***
email_versionpersonalized 0.638679   0.046914  13.614  < 2e-16 ***
hour                      0.016717   0.005006   3.340 0.000839 ***
weekdayMonday             0.540978   0.093408   5.792 6.97e-09 ***
weekdaySaturday           0.282735   0.097774   2.892 0.003832 **
weekdaySunday             0.183420   0.100117   1.832 0.066942 .
weekdayThursday           0.625210   0.092338   6.771 1.28e-11 ***
weekdayTuesday            0.615972   0.092369   6.669 2.58e-11 ***
weekdayWednesday          0.755416   0.090843   8.316  < 2e-16 ***
user_countryFR           -0.078671   0.162571  -0.484 0.628443
user_countryUK            1.155266   0.122047   9.466  < 2e-16 ***
user_countryUS            1.141360   0.115948   9.844  < 2e-16 ***
user_past_purchases       0.187795   0.005726  32.800  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20140  on 99949  degrees of freedom
Residual deviance: 18545  on 99936  degrees of freedom
AIC: 18573

Number of Fisher Scoring iterations: 7
```

Logistic regression

```
Call:
glm(formula = clicked ~ email_text + email_version + hour + weekday +
    user_country + user_past_purchases, family = binomial(link = "probit"),
    data = email)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9603  -0.2261  -0.1659  -0.1189   3.6775

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.317494   0.059838 -55.441  < 2e-16 ***
email_textshort_email     0.120446   0.019151   6.289 3.19e-10 ***
email_versionpersonalized 0.272964   0.019560  13.955  < 2e-16 ***
hour                      0.007433   0.002128   3.493 0.000477 ***
weekdayMonday             0.229140   0.038423   5.964 2.47e-09 ***
weekdaySaturday           0.120060   0.039828   3.014 0.002575 **
weekdaySunday             0.076705   0.040626   1.888 0.059017 .
weekdayThursday           0.265164   0.038072   6.965 3.29e-12 ***
weekdayTuesday            0.259625   0.038122   6.810 9.74e-12 ***
weekdayWednesday          0.319069   0.037622   8.481  < 2e-16 ***
user_countryFR           -0.040090   0.061450  -0.652 0.514142
user_countryUK            0.468141   0.047046   9.951  < 2e-16 ***
user_countryUS            0.456565   0.044138  10.344  < 2e-16 ***
user_past_purchases       0.084062   0.002595  32.400  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20140  on 99949  degrees of freedom
Residual deviance: 18509  on 99936  degrees of freedom
AIC: 18537

Number of Fisher Scoring iterations: 7
```

Probit regression

Using hour feature as a continuous variable from 1 to 24 doesn't make sense so we bin the hour into morning, afternoon and evening and rerun the logistic and probit models using new hour_binned feature. From the new result, we can see that the probability of being clicked doesn't increase monotonically with time. Using the afternoon as the benchmark, sending emails in the morning is better than in the afternoon while sending emails in the afternoon is worse than in the night in the new logistic regression.

```
Call:
glm(formula = clicked ~ email_text + email_version + hour_binned +
    weekday + user_country + user_past_purchases, family = binomial(link = "logit"),
    data = email)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1615  -0.2219  -0.1669  -0.1240   3.5652

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -6.790026   0.152994 -44.381  < 2e-16 ***
email_textshort_email     0.277554   0.045305   6.126 9.00e-10 ***
email_versionpersonalized 0.637464   0.046918  13.587  < 2e-16 ***
hour_binnedmorning        0.119324   0.063648   1.875   0.0608 .
hour_binnednight         -0.138252   0.076189  -1.815   0.0696 .
weekdayMonday             0.541449   0.093411   5.796 6.78e-09 ***
weekdaySaturday           0.283793   0.097779   2.902   0.0037 **
weekdaySunday             0.181336   0.100131   1.811   0.0701 .
weekdayThursday           0.621743   0.092350   6.732 1.67e-11 ***
weekdayTuesday            0.614863   0.092380   6.656 2.82e-11 ***
weekdayWednesday          0.758043   0.090849   8.344  < 2e-16 ***
user_countryFR           -0.078867   0.162572  -0.485   0.6276
user_countryUK            1.155922   0.122053   9.471  < 2e-16 ***
user_countryUS            1.142221   0.115953   9.851  < 2e-16 ***
user_past_purchases       0.187970   0.005728  32.817  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20140  on 99949  degrees of freedom
Residual deviance: 18535  on 99935  degrees of freedom
AIC: 18565

Number of Fisher Scoring iterations: 7
```

Logistic regression

- **Personalization**
    - **Build the model**

Using simple logistic regression and probit regression above helped us extract insights and understand how each variable impacted the output. For instance, we found out that personalized as well as short emails are better, we should send emails on weekdays, especially on Tuesday, Wednesday and Thursday, etc. However, the fact that on average short emails are better, doesn’t imply that short emails are better for every user we have.

Next, we'd like to approach the personalization problem. The goal of personalization is to take insights one step further and find the best email characteristics for each user. So a given user will receive a long email, another one a short one, one will receive it in the night, and one in the morning, etc.

In this problem, we'd like to use Random Forest which is a powerful classification algorithm to deal with a binary classification problem. There's another point to

make before we build the model. As EDA graph shows before, user past purchases feature has a long-tail distribution and we also bin the past purchases like hour to make Random Forest more efficient and easier to run. We use the out of bag parameter on the train data and get the train accuracy.

Let's define the class 0 and class 1 error for the test data. Class 0 error rate is the number of being 0 labeled as 1 divided by the number of being 0. So as the class 1 error rate.

Random Forest Accuracy

```
        OOB estimate of  error rate: 11.54%
Confusion matrix:
      0    1 class.error
0 57973 6647   0.1028629
1   963  384   0.7149220
               Test set error rate: 11.48%
Confusion matrix:
      0    1 class.error
0 29848 3413   0.1026127
1   487  235   0.6745152
```

OOB and test error are very similar, so we are confident we are not overfitting. And overall the model is working pretty well. We only had 2% of clicks, but despite that the model is not predicting all events as class 0, we actually manage to correctly predict ~1/3 of clicks (changing weights helped). And class 0 error didn't go up that much either.

- **Predict click-through-rate for each segment**

The second step is to create a new dataset with all unique combinations of our variables. We will then feed this dataset into the random forest model and, for each unique combination, we will get a prediction. The model prediction represents click rate and, therefore, this step is meant to estimate probability of clicking for each unique combination of country, # of purchases, email text, weekday, etc.

	user_country <fctr>	purchase_binned <fctr>	email_text <fctr>	email_version <fctr>	weekday <fctr>	hour_binned <fctr>	prediction <dbl>
1	US	Low	short_email	generic	Thursday	morning	0.00
2	US	None	long_email	personalized	Monday	morning	0.00
3	US	Low	short_email	generic	Tuesday	afternoon	0.00
4	US	High	long_email	personalized	Thursday	morning	0.92
5	UK	Low	short_email	generic	Wednesday	morning	0.06
6	US	None	long_email	generic	Wednesday	morning	0.00
7	US	None	short_email	generic	Saturday	night	0.00
8	FR	Medium	long_email	generic	Thursday	afternoon	0.00
9	ES	None	long_email	personalized	Thursday	afternoon	0.00
10	US	Low	short_email	personalized	Wednesday	night	0.24

1–10 of 10 rows

So, looking at the table output, if we send a short email, generic, in the morning, on Thursday, to US customers with few(low) purchases, our model predicts no clicks. And so on for each row. For each unique segment, we have got the probability of clicking.

- **Predict click-through-rate for each segment**

The third step is to identify the variables that can be personalized. This means separating user characteristics from product characteristics, and focusing on the second ones. The reason behind it is that we only can control product characteristics. For instance, you can choose when to send the email or its message, but you can't realistically move a customer from Spain to the UK.

Then, we group by unique combinations of user characteristics and find the product characteristics with the highest probability of clicking. So, for instance, one group will be US customers with 0 purchases (these are user characteristics). And then we will look for the combination of all the other variables that maximize probability of clicking. And that's it. That combination will tell us how our product should be for those users and we will send emails accordingly.

user_country <fctr>	purchase_binned <fctr>	email_text <fctr>	email_version <fctr>	weekday <fctr>	hour_binned <fctr>	prediction <dbl>
US	High	long_email	personalized	Thursday	morning	0.92
UK	High	short_email	personalized	Sunday	morning	0.90
UK	Medium	short_email	personalized	Wednesday	morning	0.78
US	Medium	short_email	personalized	Thursday	morning	0.64
ES	High	short_email	personalized	Tuesday	morning	0.62
ES	Medium	short_email	personalized	Tuesday	afternoon	0.54
FR	High	short_email	personalized	Thursday	night	0.52
FR	Medium	short_email	personalized	Monday	night	0.48
US	Low	short_email	personalized	Tuesday	afternoon	0.34
UK	Low	short_email	personalized	Saturday	afternoon	0.30

1–10 of 16 rows

user_country <fctr>	purchase_binned <fctr>	email_text <fctr>	email_version <fctr>	weekday <fctr>	hour_binned <fctr>	prediction <dbl>
FR	Low	short_email	generic	Friday	night	0.14
ES	Low	long_email	personalized	Monday	afternoon	0.10
UK	None	short_email	personalized	Sunday	morning	0.06
US	None	long_email	personalized	Wednesday	night	0.04
ES	None	short_email	personalized	Tuesday	morning	0.02
FR	None	short_email	personalized	Monday	night	0.02

So now we have a model that returns the best email strategy for each user and that's how we should be sending email to maximize overall click-through-rate. Btw note how even the best email strategy has super low model predictions for users with no purchases, regardless of the country. Once again, it won't be able to win those people just by tweaking the email.

- **Test the personalized strategy**

Now that we have come up with a personalized strategy to send emails, the last step is to test it. In order to test, we would run our personalized algorithm on a randomized fraction of users and compare its results with the current email strategy. Since we know the predicted probability for each group, we can just estimate the weighted average to guess the final overall click rate.

Our model isn't a perfect one and has pretty high class one error so we need to adjust the predicted probabilities after taking into account the model expected error.

We can do it as fellow:
Assume, for instance, that our model output is 0.8, so it is predicting 80% clicks. Based on the confusion matrix when we built the model, that when our model predicts class 1 is right 5% of the times and when it predicts class 0 is wrong 2% of the times. So if our model is predicting 80% class 1 (and therefore 20% class 0), am actually expecting: 0.8 * 0.05 + 0.2*0.02 = 0.044 = 4.4%. It still isn't a big click rate but would still be a huge improvement because our starting point is 2.07% in EDA.

Using this calculation, we find out that this personalized algorithm does improve the click rate a lot. It would be great to apply this strategy.

predicted_click_rate <dbl>	old_click_rate <dbl>
0.03352934	0.02070035

Comparison between personalized strategy and random strategy

Marketing strategy Conclusion, Limitation and Recommendation

Conclusion

- In the first part, we identified several features that are critical to email click-through rate. For example, short emails are generally better than long emails and sending emails during Tuesday and Thursday than other weekdays. Country is also important as customers in the US and UK are more likely to click the email.
- In our second part, our model precision was ~5% for the personalized email strategy. Matter of fact, that was actually really good. Our starting click-through-rate is 2% when using the randomized email strategy. A 5% precision means we have found a region of space with 2.5X probability of clicking. That's huge when dealing with real data. And, at the end, our overall predicted click-through-rate went from 2% to 3.7%. That's also huge. It would mean almost doubling up whatever revenue is coming from emails.
- It is interesting to note that you cannot tweak the click-through rate for customers with no purchases. You cannot change their behavior only by tweaking the email.

Limitation

- Personalization sounds to be a sweet option to increase the conversion rate. But there are a lot of limitations to do it.
- Costs of personalization. Personalized advertising is typically more expensive than other types of online advertising that is less targeted. When people consider

the hundreds of applications needed to gather, manage, store, and secure personal data on every potential customer who's ever interacted with content or ads on a company's site, social channels, and ad networks, it's staggering. And it's not just the data. It's the content. You can have all the data analytics you need to understand what content to deliver to each user. But you still need to build, publish, and maintain the content.

- Difficult in data collection. It takes time and money to track more detailed customers' information.

Recommendation

- Gather as Much Customer Data as Possible. The more variables you have about your users, the more granular will be the groups, and the more specific will be the personalization. With the right technology in place, marketers can begin collecting customer data at every single touchpoint across the buying journey. When marketers know exactly what a specific customer is searching for, or is continuously clicking on, it becomes easy to target this customer with relevant content, incentives, and even recommendations.
- Continuously Update and Refine Personalization Processes. Just like any other process, personalized marketing practices require continuous review and optimization to ensure 100% effectiveness. A best practice is to let campaigns run for a couple weeks, and then analyze the results. This way, the team will have an accurate impression of actual progress over time, rather than a quick snapshot into how the personalization strategy is progressing. It's also important to remember to double-check all of the technological integrations, to make sure you have access to a holistic picture of data. During the refinement stage, marketers can take a step back and evaluate certain strategies or elements of their personalized marketing campaign.