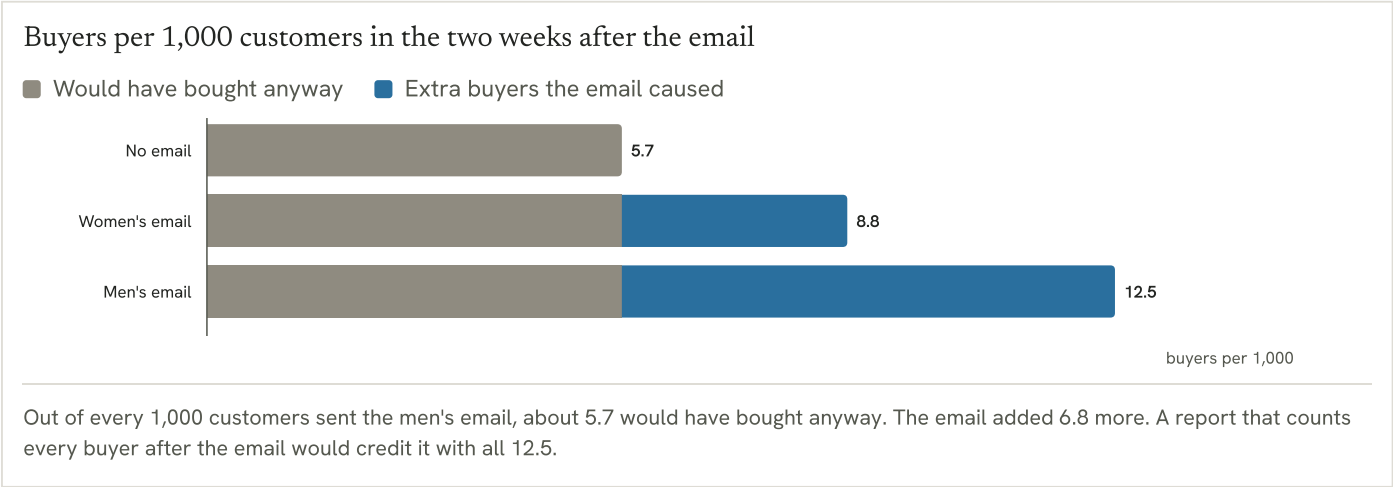


Both emails sold more. No model beat sending to everyone.

A retailer emailed 42,694 of 64,000 customers and held the rest back at random. We measured what each email really caused, then tested 6 targeting models to see if emailing fewer, better-chosen people would do better.

Independent sample using a public 2008 experiment, to show what a Dunwell Campaign Impact Check delivers. Every figure on this page comes from one results file.

Business: US online retailer, men's and women's merchandise
Data: 64,000 customers in a random email test, 2008 Outcome window: two weeks after the email
Prepared by Dunwell Insights, October 2026



Send the men's email to everyone. Don't pay for targeting yet.

6.8 Extra buyers per 1,000 men's emails range 5.0 to 8.6	46% Of men's-email buyers would have bought anyway 65% for the women's email	0 of 6 Targeting models proven better than random averaged over emailing 10% to 60%, on 19,200 customers set aside	\$0.31 Break-even cost per men's email at an assumed 40% margin
---	---	---	--

What we found

- 1

The men's email more than doubled purchases.

0.57% of customers bought with no email, 1.25% after the men's email. That is 6.8 extra buyers and \$0.77 extra spend for every customer emailed. The women's email added about half as much (3.1 buyers per 1,000).
- 2

Counting every buyer after the email would overstate it by about 1.8 times.

46% of buyers after the men's email would have bought anyway, and 65% after the women's. Only a no-email group shows this. Without one, the women's email would look 2.8 times better than it was.
- 3

The men's email sold to women's-wear buyers as well as the women's email did.

Among customers who only bought women's items last year, the men's email added +0.57 pts and the women's email +0.51 pts. We found no group where the women's email did clearly better.
- 4

The women's email did little for established customers.

It added +0.53 pts for customers new in the last year but only +0.09 pts for the rest. This showed up in training and pointed the same way on the hold-out, but the hold-out alone was not conclusive. Treat it as a strong lead to test.
- 5

No targeting model was proven better than emailing at random.

We tested 6 models, from logistic regression to a causal forest and a neural network. Averaged over emailing 10% to 60% of the list, every rule's range of likely gain on the hold-out included zero, and the model that won in training came last. At single shares, 5 of 48 rule comparisons did clear zero, more than chance alone would give. So ranking customers probably helps a little, but which rule wins keeps changing. With 191 buyers in the hold-out, that is as far as the data goes.

What we'd do

Send the men's email to the whole list. At an assumed 40% margin it pays for itself up to \$0.31 per email, far above what sending costs. On a list of 100,000, it should bring about 681 extra buyers and \$76,983 extra spend in two weeks.

Keep a random 10% back from every send, so you can keep measuring what the email really adds.

Don't pay for targeting on this evidence. If you must email fewer people, a simple "most likely to buy" order did as well as any model. To settle targeting, run a bigger test (section 6).

A clean random test: three equal groups that looked the same before the email

Each of the 64,000 customers had bought in the last twelve months. A third were picked at random for the men's email, a third for the women's email, and a third got nothing. Purchases were tracked for two weeks after the email. Before looking at any result we checked that the random split worked. The group sizes are as close to equal thirds as chance allows ($p = 0.90$), and no customer detail differs between groups by more than 0.016 standard deviations.

Customers in each group



We also set aside 30% of each group (19,200 customers) before any modelling. Targeting rules were judged only on them.

The groups matched on everything known before the email

Standardised difference from the no-email group, every customer detail. Under 0.1 is the usual sign of a fair split.



0 of 36 comparisons cross 0.1. The largest gap is 0.016. The file has no missing values. 6,562 rows match another row exactly, which is expected with coarse columns and no customer ID, so we kept them.

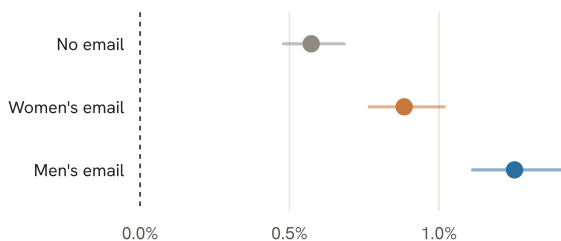
2 · DID THE EMAILS WORK?

Yes. The men's email added 6.8 buyers per 1,000, the women's email 3.1

Purchase was the main outcome, fixed before the analysis. Both emails beat no email after allowing for two comparisons (Holm-adjusted p under 0.001 and under 0.001). The men's email also beat the women's email head to head, by +0.37 pts (range +0.17 pts to +0.56 pts). Put another way, it took about 147 men's emails, or 321 women's emails, to create one extra buyer.

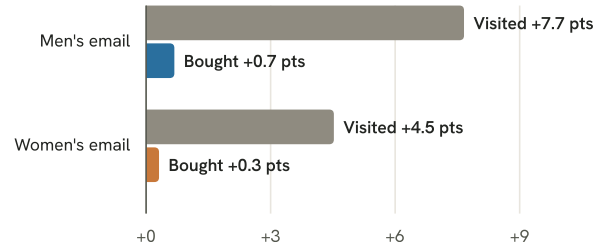
Share who bought, with its likely range

95% interval for each group



The email mostly brings visits. Few extra visitors buy.

Change caused by each email, percentage points



The men's email brought 11 extra site visits for every extra buyer. Most of the email's pull ends at the website. That points at the product page and checkout, not the email, as the next place to look.

Every result, email against no email

Change caused by the email	Men's email	Women's email
Purchase rate (main outcome)	+0.68 pts +0.50 pts to +0.86 pts	+0.31 pts +0.15 pts to +0.47 pts
Spend per customer emailed	\$0.77 \$0.49 to \$1.06	\$0.42 \$0.16 to \$0.68
Visit rate	+7.7 pts +7.0 pts to +8.3 pts	+4.5 pts +3.9 pts to +5.2 pts
No-email level	0.57% bought · \$0.65 per customer · 10.6% visited	
Spend per buyer (context only)	median \$66.98	median \$94.71 no email: \$88.37

Spend per customer counts everyone sent to, with non-buyers at \$0. That is the figure that matters for the campaign. Spend per buyer compares different groups of people, so it's shown for context only.

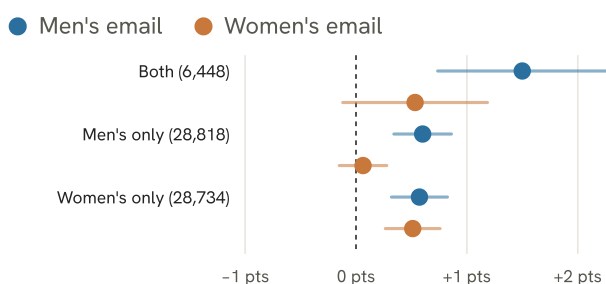
Why this matters: a report that counts every buyer after a campaign can't tell who would have bought anyway. Here it would count 12.5 buyers per 1,000 for the men's email when the email caused 6.8. For the women's email it would count 8.8 against 3.1. A small no-email group is the cheapest way to know what a campaign is really worth.

3 · WHO RESPONDS MORE?

The men's email worked across the board. The women's email looked weaker for established customers.

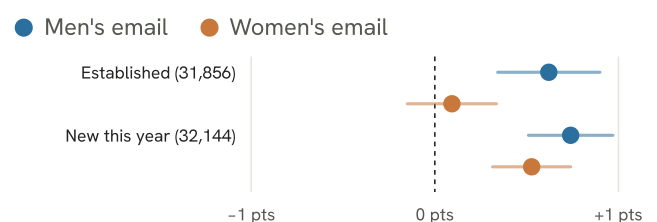
We split customers by six details known before the email and asked whether either email worked better for some groups. For the men's email, none of the six showed a real difference in training. It lifted purchases by a similar amount everywhere. For the women's email, two details stood out in training. Past spend held up on the hold-out, but its shape is odd: the effect was strongest above \$500, near zero at \$200 to \$500 and positive below that. We wouldn't act on it without a new test. Being a new customer pointed the same way on the hold-out without being conclusive.

Extra purchases by what customers bought last year



Each dot is the extra purchase rate caused by the email, with its 95% range. Customers who bought both kinds last year responded most to the men's email (+1.50 pts), but that group is small and the difference was not significant in training.

Women's email: new customers against established ones



"New" means a first purchase in the last twelve months. Customers who had been around longer barely reacted to the women's email (+0.09 pts, range -0.15 pts to +0.34 pts).

How we kept chance patterns out

Email	Customer detail	Training p	Hold-out p	Result
Women's email	Spend last year Higher: \$500+. Lower: \$200-500.	0.033	0.044	Held up
Women's email	New in the last 12 months Higher: New this year. Lower: Established.	0.013	0.325	Same direction, not conclusive

Not flagged in training (p 0.05 or more): Men's x what they bought last year, Men's x months since last order, Men's x spend last year, Men's x new in the last 12 months, Men's x where they bought, Men's x area, Women's x what they bought last year, Women's x months since last order, Women's x where they bought, Women's x area.

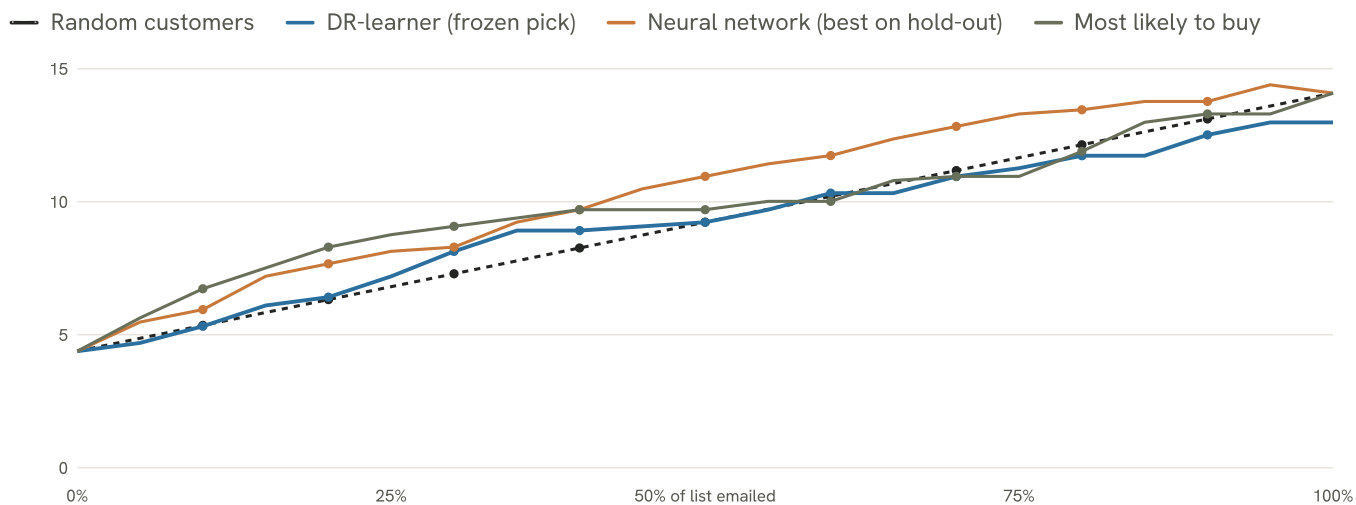
We looked for differences in the training 70% only. Anything that showed up there was checked once on the 30% we had set aside. Twelve checks were run in training, so one or two false alarms are expected. That is why the second check matters.

4 · CAN A MODEL PICK WHO TO EMAIL?

Six models, and none was proven better than picking customers at random

If you can only email part of the list, who should get it? The usual answer is "the people most likely to buy". But many of them would buy anyway. Uplift models try to find the people the email will change. We trained 6 of them on the 70% training set, picked one by cross-fitting before seeing the hold-out (DR-learner), froze it, and then scored every rule on the 19,200 customers set aside.

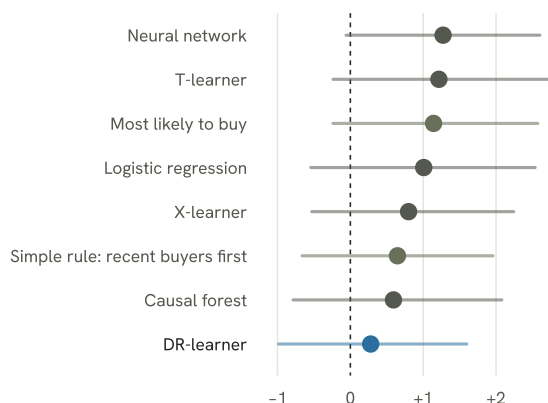
Buyers per 1,000 customers, by how many you email



Value is estimated from the random assignment: what each rule would have produced if the whole hold-out had followed it. The lines sit on top of each other. Emailing more people matters far more than choosing who.

Gain over emailing random customers

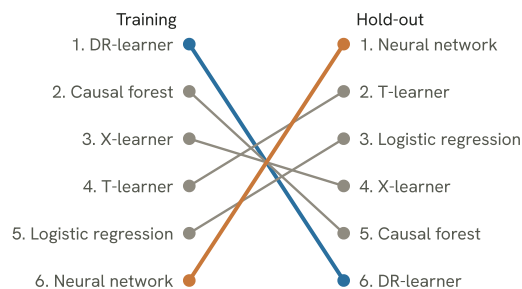
Extra buyers per 1,000 customers, averaged over emailing 10% to 60% of the list. Tested on the 30% hold-out, with 95% range.



Dots show the average gain. Lines show the 95% range from 2,000 resamples. Every range crosses zero. At single shares a few rules did clear zero (Most likely to buy at 10%, Most likely to buy at 20%, X-learner at 30%, Neural network at 50%, Neural network at 60%), but the winner changes from share to share. The frozen pick (DR-learner) gained +0.28 buyers per 1,000 (range -0.99 to +1.60).

The order of the models flipped on new data

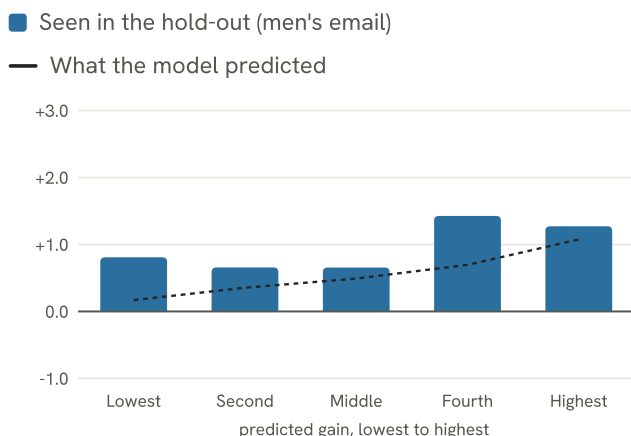
Rank in training (cross-fitted) against rank on the hold-out



DR-learner ranked first in training and last on the hold-out. Neural network went the other way. When the order reverses like this, the differences between models are noise. More complexity didn't buy anything here.

Predicted gain against what really happened

Hold-out customers split into five equal groups by the DR-learner's predicted gain from the men's email, in points



The model's top two fifths did respond more (+1.35 pts on average against +0.71 pts), so there is some signal. But the predictions were too low across the board, and each group's range is wide.

Likely buyers or persuadable buyers?

Emailing 30% of the hold-out, two ways

Who gets emailed	Bought, no email	Bought, men's email
Most likely to buy 5,760 customers · 35% new	0.85%	2.35% +1.5
DR-learner pick 5,760 customers · 60% new	0.47%	1.90% +1.4
Picked by neither 11,021 customers · 54% new	0.33%	0.91% +0.5

The uplift model did avoid people who would buy anyway: 0.47% of its picks bought with no email, against 0.85% of the likely buyers. But likely buyers also responded strongly to the email, so both picks gained about the same. 58% of the two lists were the same people.

5 · WHAT IT'S WORTH

Each men's email is worth about \$0.77 in extra sales within two weeks

Extra sales per email come straight from the test. To turn them into profit we assumed a 40% margin and \$0.01 to send each email. On those numbers, the men's email returns \$298 per 1,000 sent. It only stops paying if each email costs more than \$0.31. Real sending costs are a small fraction of that, so the case for cutting the list would have to rest on unsubscribes or email fatigue, which this test did not measure.

Email	Extra sales per email	Range	Profit per 1,000 sent	Break-even cost per email
Men's email	\$0.77	\$0.49 to \$1.06	\$298	\$0.31

Email	Extra sales per email	Range	Profit per 1,000 sent	Break-even cost per email
Women's email	\$0.42	\$0.16 to \$0.68	\$160	\$0.17

These figures use a margin and an email cost we assumed. They are not in the data. The dashboard lets you put in your own. The window is two weeks, so this is not lifetime value, and the test did not record unsubscribes.

6 · NEXT STEPS

Send to everyone, keep a hold-back, and run the targeting test properly

- 1

Send the men's email to everyone, and hold back a random 10%.

The send list marks the hold-back. Compare the two groups after two weeks. It costs a little in sales and tells you every time what the email is really worth.
- 2

Judge campaigns by the hold-back difference, not total sales after the send.

Total sales after a send include buyers who would have bought anyway.
- 3

Test the women's email on new customers only.

New customers were the one group where it came close to the men's email, and the pattern pointed the same way on the hold-out. Send it to half of new customers and the men's email to the other half.
- 4

If targeting still matters, size the test for it.

To tell a gain of 1 extra buyer per 1,000 from noise, the set-aside group needs about 105,000 customers, roughly 5.5 times the one here. Run the frozen rule against random selection on that many people before paying for it.
- 5

Look at the website, not just the email.

The email brought 11 extra visits for each extra buyer. A better product page could be worth more than a better email.

7 · HOW WE DID THIS

The plan was written down before any result was seen

The rules for this study were locked before we compared any outcomes. The targeting rule was frozen, with a fingerprint of its scores, before the hold-out was opened. If something had changed after that, the run would have stopped. The whole study reruns from the raw file with one command.

1 Check the data Hash, counts, missing values, the random split	2 Lock the plan Outcomes, comparisons and the hold-out fixed in writing	3 Readout Each email against no email, all customers	4 Patterns Found in training, checked once on the hold-out
5 Train six models 5-fold cross-fitting on the training 70%	6 Freeze Pick by cross-fitted value, fingerprint the scores	7 Test Every rule on the hold-out, 2,000 resamples	8 One results file Report, dashboard and send list read from it

The six models

Model	Training rank	Hold-out rank
DR-learner Frozen pick Combines outcome models with the known one-in-three assignment, so errors in one are corrected by the other	1	6
Causal forest A forest built to split customers where the email's effect differs, with honest splitting	2	5
X-learner Learns each email's gain directly from the other group's predictions	3	4
T-learner A separate gradient-boosted model for each group; gain is the difference	4	2
Logistic regression One model with email-by-customer terms, kept simple with a penalty	5	3
Neural network A small neural network: a shared layer learns the customer, one output per group. Five networks averaged	6	1

Checks we ran

- Raw file fingerprint matches before every run (SHA-256 bab6578f60db...).
- Group sizes and published purchase rates match the original challenge.
- Every purchase has spend above \$0 and a site visit; no spend without a purchase.
- Model inputs are only details known before the email. Visits, purchases and spend are never inputs.
- The hold-out's scores were saved and fingerprinted before any hold-out result was read. The test step refuses to run if they changed.
- Rule values were checked two ways (weighting and doubly robust). They track each other closely (correlation 0.99). In 4 of 48 comparisons they disagree on the sign, always within 0.27 buyers per 1,000 of zero.
- Every figure on this page is read from one results file. None is typed in.

Decisions log

Decision	Why
Purchase is the main outcome; spend per customer second	Fixed before analysis. Spend among buyers only would compare different people.
Keep the two emails separate	They had different effects. Pooling would hide that.
Holm correction across the two main comparisons	Two emails means two chances of a false alarm.
Keep the 6,562 identical rows	No customer ID, coarse columns. They are most likely different people.
Set aside 30% before any modelling	A rule must be judged on customers it never saw.
Six models instead of one (logged before any outcome was seen)	Asked for a deeper look. The pick was still made on training data only.
Pick the model by average value at 10% to 60% contact	That is the range where targeting would matter in practice.
Compare rules at the same number of emails	Otherwise a rule wins just by emailing more people.
Recommend the simple option when no model clearly wins	Complexity has to earn its place.

What this data can't tell us

- It is one email test from 2008 at one US retailer. It says nothing direct about discounts, paid ads, text messages or shops today.
- Results cover two weeks after the email. Longer effects, repeat orders and lifetime value were not measured.
- There is no margin, email cost or unsubscribe data. Every money figure rests on stated assumptions.
- No individual's response is ever observed under both emails. Model scores are estimates for groups of similar customers, not facts about a person.
- "No model beat random" means this test was too small to show a gain. It does not prove targeting can never work.
- The data has no customer ID, so the send list uses a row reference, not real people.

This analysis is based on the data described here and the methods shown. It supports decisions. It is not a promise of any result, and it is not financial, legal or tax advice.

Every job. Done well.

Dunwell Insights · info@dunwellstudio.com

Data: Kevin Hillstrom, MineThatData E-Mail Analytics and Data Mining Challenge (2008). blog.minethatdata.com. Used for analysis only; the raw file is not redistributed. Kevin Hillstrom and MineThatData do not endorse Dunwell.