Charge Task Auto-Creation AB Test Readout

Ben Laufer
Last Updated 2020-06-25

TLDR

- Charge task auto-creation increases **gross profit +3.82%*** ** over battery thresholds for scooters assigned. Region-wide we expect gross profit to increase by **+1.9%** with autocreation.
- Auto-creation creates **22.65%*** more tasks** (leading to an expected region-wide juicer cost increase of **+7.9%**) but the additional collections help us capture more demands and improve deploys.
- By-region, we see statistical significant positive results in 5 regions and negative results in 1 region, Seoul. Where regions show negative gross profit, we will use RVD information to configure the multiplier for further testing [doc].
    - In Seoul, we are creating too few tasks. Gross RVD is negative because scooters are left on the street without enough battery to capture demand. We suggest lowering the multiplier to create more tasks, and re-testing.

Experiment Overview

Product: Juicer charge task auto-creation
Product Change: ROI-based calculation is made on the scooter level in hourly intervals to determine whether a charge task is created. Charge tasks get created based on battery level, location and time instead of only battery level (in BT model).
Test Regions: Berlin, Bucharest, Budapest, Busan, Christchurch, Cologne, Copenhagen, Dunedin, Dusseldorf, Frankfurt, Hamburg, Hamilton, Hannover, Lyon, Madrid, Malaga, Malmo, Munich, Oslo, Paris, Poznan, Rimini, Ruhrpott, Seoul, Stockholm, Stuttgart, Tel Aviv, Turin, Ulsan, Vienna, Warsaw, Wroclaw
Test Observation Period: 2020-06-05 to 2020-06-26 (with rolling experiment start dates)
Test randomization level: Randomized by scooter by day (daily re-randomization)
Split: 50/50

	Number Observations
Control	233743
Test	234485

***Paris PM Collection Window Issue**: There was a change for three weekend days in Paris where the start time for PM collection window was moved to noon from 9pm. This allowed automated charge tasks to be created much earlier in the day, which prevented the scooters from capturing afternoon demand. This was an unintended change and so the days were omitted from the study.

Global Results

Metric	Control Baseline	Test Baseline	Difference	Percent Difference	P-Value	Significance
DSRVD	4.637	4.956	0.319	6.87%	0.0000	***
Gross Profit	3.845	3.992	0.147	3.82%	0.0000	***
Zero Rate	0.437	0.432	-0.005	-1.04%	0.0017	**
Task Creation Rate	0.356	0.437	0.081	22.65%	0.0000	***
Charging Rate	0.214	0.260	0.046	21.47%	0.0000	***
Juicer Cost	0.792	0.963	0.172	21.70%	0.0000	***

By-Region Breakdown

Region	Number of Observations	Gross Profit	DSRVD	Zero Rate
Berlin	21953	3.53%	2.57%	-1.98%
Bucharest	24547	6.66%**	18.33%**	-9.78%**
Budapest	12737	2.01%	4.82%`	-2.25%
Busan	18631	-2.57%	-6.81%**	2.74%
Christchurch	5751	30.41%**	-11.06%**	-1.78%

Cologne	29737	-3.41%	-1.53%	0.0217
Copenhagen	18654	-2.32%	-2.73%	-0.11%
Dunedin	1972	4.76%	-16.95%`	0.0584
Dusseldorf	11304	0.45%	3.30%	1.55%
Frankfurt	27241	-0.0383	1.75%	-1.62%
Hamburg	12735	0.0489	9.64%**	-0.54%
Hamilton	4379	4.40%	-4.78%	-0.12%
Hannover	4532	-3.97%	-4.92%	1.32%
Lyon	24815	-1.23%	-0.15%	2.10%
Madrid	10298	-5.21%	-11.69%**	5.15%`
Malaga	5592	0.1088	-0.0845	-1.25%
Malmo	4669	8.68%	9.30%	-1.35%
Munich	15084	-1.59%	-0.09%	0.17%
Oslo	25521	2.52%	8.58%**	-4.66%*
Paris	34458	15.53%**	20.91%**	-7.28%*
Poznan	6891	-0.66%	1.68%	-3.25%
Rimini	1268	15.60%	10.39%	-2.29%
Ruhrpott	4461	0.21%	0.73%	-0.89%
Seoul	31536	-5.91%**	-3.74%*	1.68%
Stockholm	24268	-2.20%	-3.18%	1.01%
Stuttgart	9719	-0.0948	5.71%	-2.80%
Tel Aviv	33092	3.82%*	8.44%**	-5.94%**
Turin	3713	7.64%	-4.94%	4.57%
Ulsan	1085	0.2279	2.64%	-6.21%
Vienna	5494	19.42%**	41.90%**	-17.22%**
Warsaw	27186	2.67%	6.80%**	-1.71%
Wroclaw	4903	5.64%	10.19%*	-4.70%

Regional Configuration Updates

On 07/06/2020, we launched regional-configured multiplier changes based on the results above. We used the following logic:

- We say a region has a negative impact if:
    - If gross profit is negative with p-value < 0.1
    - If gross profit is negative not sig but gross rev is negative with p-value < 0.05
- For negative regions, we change the multiplier as follows:
    - If gross rev is positive (means charge too many); increase the multiplier to 2
    - If gross rev is negative (means charge too few); decrease the multiplier to 1.5
- Launched Monday 07/06/2020 (based on updated results):
    - **Seoul, Busan, Madrid**: decrease the multiplier to 1.5
    - **Stuttgartt**: increase the multiplier to 2

In-Depth Explanation of Model

Documentation

For more in-depth descriptions of the model and implementation, see the following documentation:

- Motivation and Modelling doc
- Experiment Design and Plan doc
- Hotspot matching model: deployed code (not in use)
- No hotspot matching model: deployed code (tested)
- Shadowing dashboard
- Testing (assignment + results) dashboard
- Shadowing results readout doc
- Charge & rebalance task organized documentation reference doc

Motivation

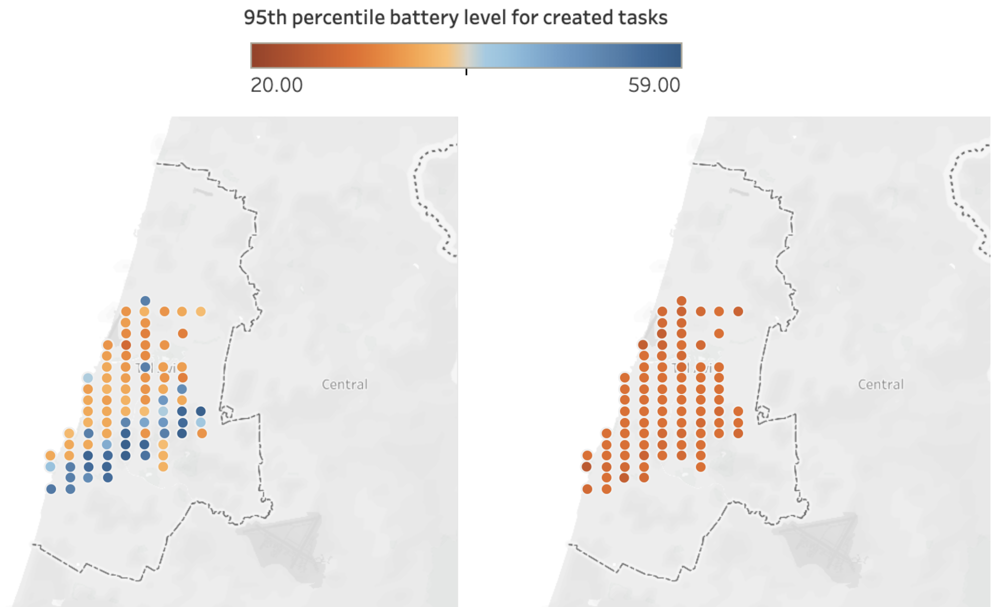
The broad purpose of the model is to move beyond the battery threshold feature, which uses only the current battery level of a scooter to determine if a charge task is created.

Since we know that demand is different depending on the location of a scooter, we know some scooters can capture more revenue if left on the streets. Others would benefit from being picked up and moved, even if the battery level is higher.

Decision: A battery threshold for every geohash-day

In-depth discussion of the model can come with jargon and confusion. But, on a basic level, we do something simple: we compute a battery threshold for every geohash-day (instead of region). Thus, instead of using the demand of an entire region to guess the battery threshold, we use the demand of a particular geohash, on a particular day.

Map of Tel Aviv battery threshold using new model (left) vs battery threshold (right)



Modelling geohash demand: “Average hourly revenue”

It is very hard to determine where a scooter will end up after 1, 2 or 3 trips. Therefore, we use a very basic heuristic approach to model downstream revenue for a scooter at a given geohash and time of day. Using three weeks of historical data, we compute the average (hourly-normalized) revenue that a scooter at that geohash generate. Instead of looking downstream at where scooters end up in a city (introducing more noise), we simply average hourly activity over every hour in the day for the geohash in question. We effectively **assume** the scooter stays in place and generates revenue commensurate with its current geohash.

We use this historical geohash revenue as the expected “demand” for a scooter at a given hotspot. We **assume** that demand is independent of the scooter’s current battery level, and that instead battery level *constrains* the amount of demand that a scooter can capture.

To account for the battery level, we constrain our downstream expected revenue by a “maximum-revenue” given the current battery level of the scooter. In computing this maximum, we **assume** that the entire region has one “revenue per battery percentage” rate, and constrain the expected downstream revenue by the amount of revenue that would completely deplete a scooter at its current battery.

More concisely, our 24h revenue prediction takes the ***minimum*** of the following two values:

1. 24 hour downstream revenue: (24h)*(avg hourly rev for geohash,day)

2. Battery-constrained max rev: (current battery level - 20)*(region avg rev/battery %)

The "ROI" calculation

To decide for a given scooter (geohash, battery, day of week) whether to create a charge task, we model the cost and revenue associated with completing that charge task. To determine whether to create the task, we compare revenue to cost. An ROI positive task will have:

$$next\ day\ revenue > opportunity\ cost + juicer\ cost$$

See each value explained below:

Opportunity Cost (24h revenue)

We consider the opportunity cost of completing a charge task to be the 24h predicted revenue from its current location, battery level, and day of week. Our model uses a 3-week lookback historical average, computes the hourly average revenue, and constrains the prediction with battery level, as explained above.

Juicer Cost

We use the region-wide base cost for charging tasks.

Revenue (next day)

For the revenue prediction, we simply use the regional average RVD.

(It is very difficult to predict where a scooter will end up the next day… we have tried matching scooters with the supply of hotspots and its creates too much modelling complexity.)

"The Multiplier"

Since our model uses a significant amount of heuristics and assumptions to reduce complexity, we expect to see a high degree of prediction error. Because of this, and because we want to have a high degree of certainty that the tasks we create are profitable, we use a "multiplier". Instead of simply requiring tasks to be ROI-positive, we require the RVD to be a certain margin *greater* than the expected costs.

$$next\ day\ revenue > (multiplier) * (opportunity\ cost + juicer\ cost)$$

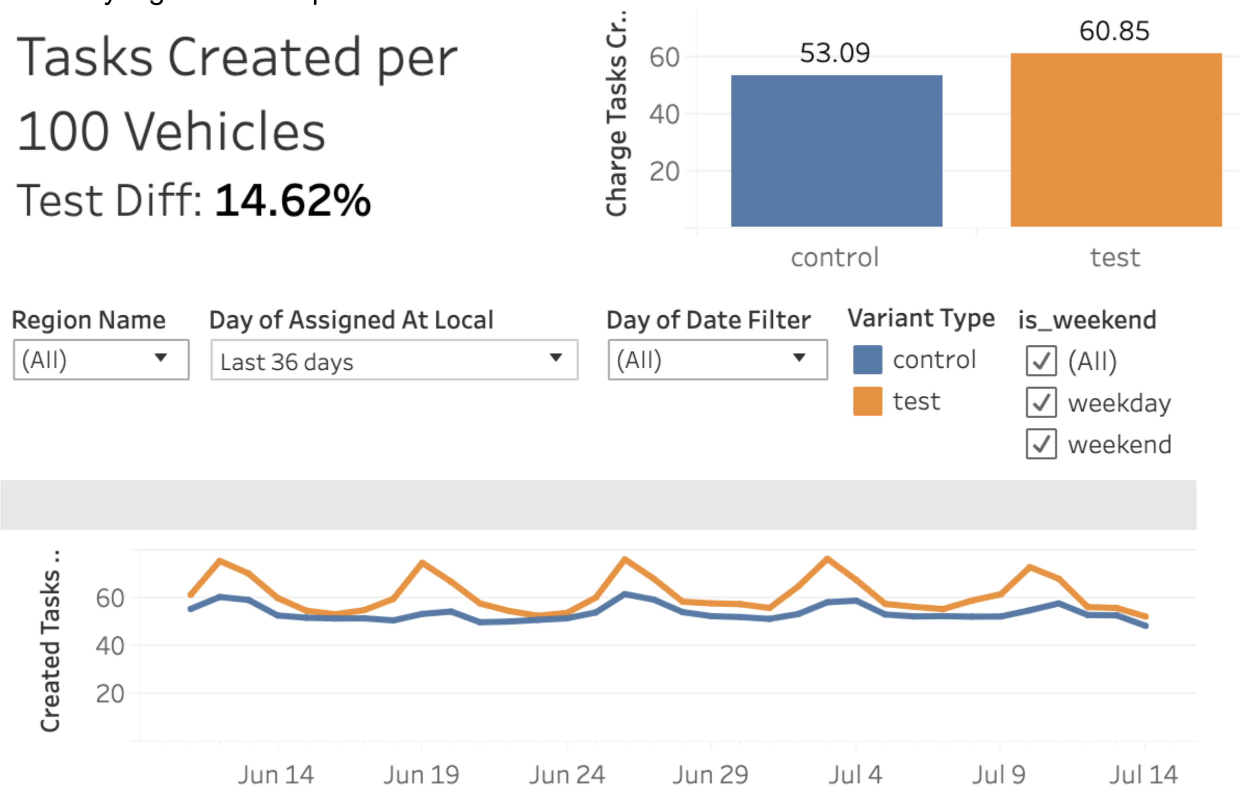
This is a very powerful parameter as it determines whether, on the most basic level, we create more tasks or fewer tasks. It is our strongest lever to pull if we see a region has negative testing results.

General Experiment Observations

Here we note a few general take-aways about the impact of our automated charge task logic.

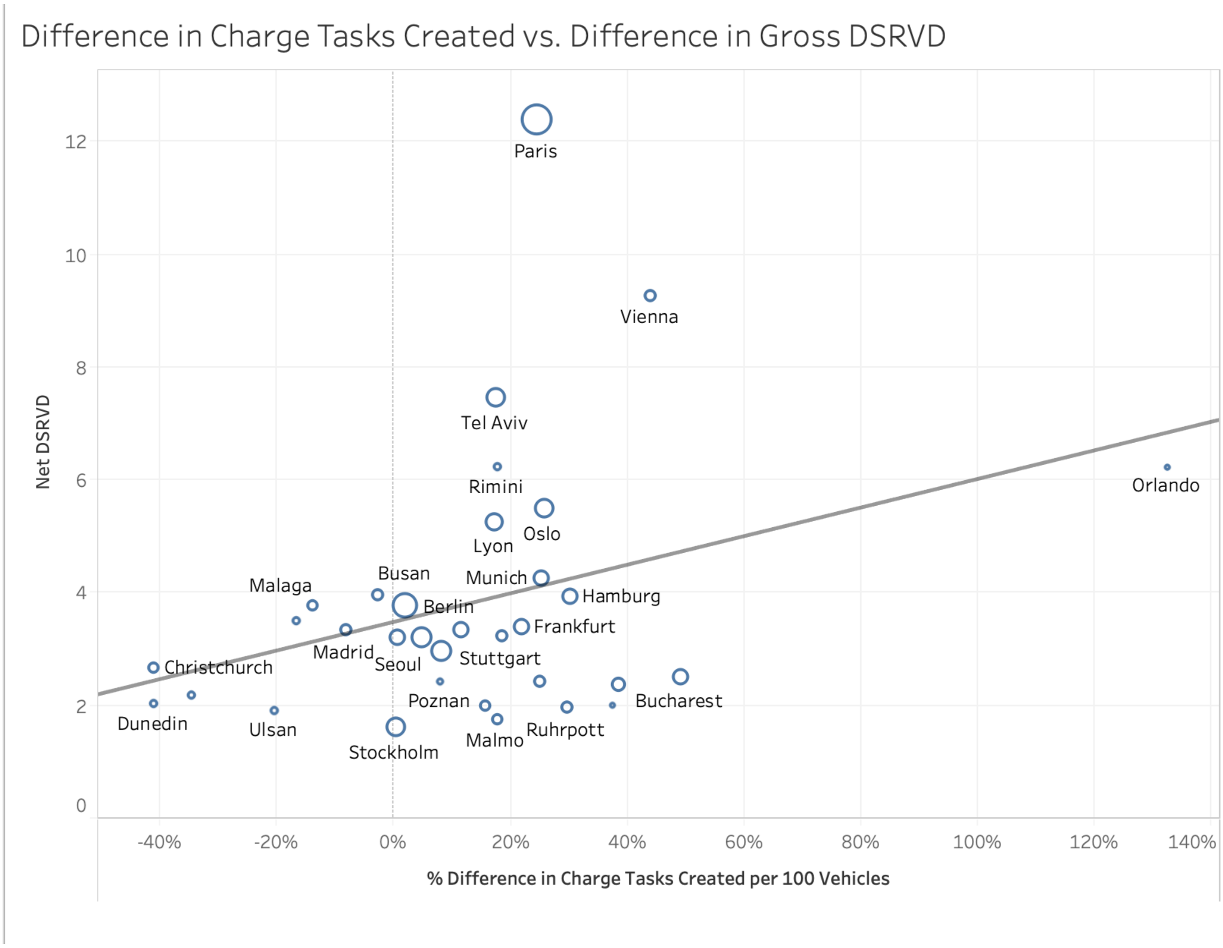
1. We create more tasks

In general, we create 14.6% more charge tasks in our new model, compared to the battery threshold model (as of 2020-07-16). This lift is especially pronounced on Thursday, Friday and Saturday nights in anticipation of weekend demand.



2. When we create more tasks, the gross revenue increases

This is an important relationship for configuring the multiplier. Gross revenue directional change will indicate whether we are creating too many or too few tasks.



At Regional Level… Understanding outcome in relation to regional metrics:

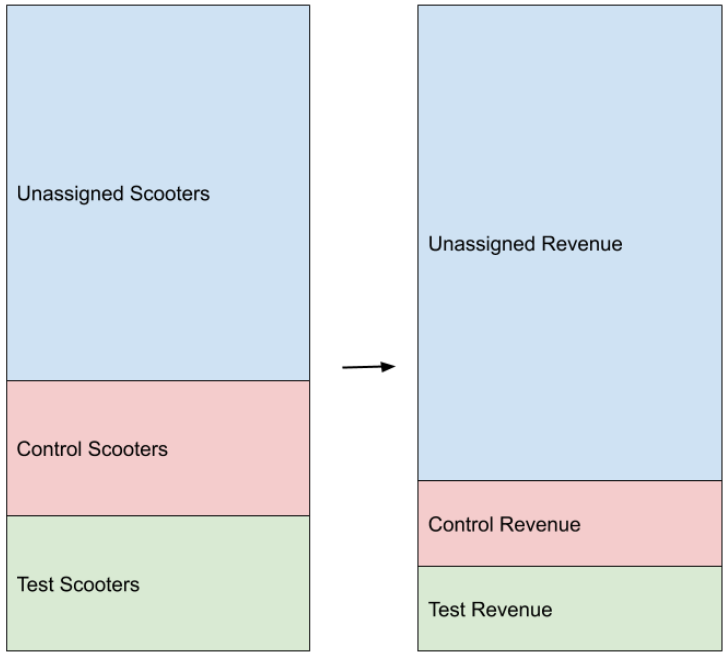
Variant Type	All Regions	Region Name	Avg. RVD_WEEKLY	Avg. OPERATIONAL_ VEHICLES_WEEKLY	% Difference in Gross Profit	% Difference in Avg. DSRVD	% Difference in Juicer Charging Cost per Scooter fr..
test	All Regions	Berlin	$7.11	5,093	-0.02%	-0.90%	-6.22%
		Bucharest	$4.69	2,533	8.35%	17.92%	84.18%
		Budapest	$6.01	1,905	4.27%	5.50%	12.16%
		Busan	$7.12	1,337	-2.13%	-4.86%	-17.63%
		Christchurch	$4.26	868	8.64%	-15.34%	-68.66%
		Cologne	$5.26	3,943	-0.10%	1.32%	8.46%
		Copenhagen	$6.18	1,758	0.77%	0.30%	-0.98%
		Dunedin	$2.67	404	-0.15%	-4.68%	-62.57%
		Dusseldorf	$4.25	1,404	1.23%	3.10%	36.02%
		Frankfurt	$6.04	2,779	0.29%	3.76%	28.89%
		Hamburg	$7.94	1,750	1.51%	6.27%	50.40%
		Hamilton	$3.76	500	8.56%	-11.91%	-50.65%
		Hannover	$5.50	598	4.94%	3.80%	-7.35%
		Lyon	$9.79	2,945	9.15%	10.13%	13.23%
		Madrid	$5.13	1,216	-4.03%	-8.62%	-29.06%
		Malaga	$5.59	852	-1.41%	-6.28%	-30.88%
		Malmo	$3.02	816	-1.85%	1.77%	20.13%
		Munich	$6.68	2,029	8.05%	11.66%	37.90%
		Orlando	$4.88	219	47.66%	58.57%	253.67%
		Oslo	$10.58	2,404	4.05%	10.99%	52.81%
		Paris	$19.32	7,156	9.43%	16.82%	64.11%
		Poznan	$4.01	785	5.77%	6.91%	14.53%
		Rimini	$10.73	366	1.52%	3.17%	11.30%
		Ruhrpott	$3.54	1,215	4.11%	4.16%	4.49%
		Seoul	$7.83	2,691	-1.01%	-1.06%	-1.22%
		Stockholm	$3.49	2,687	-4.31%	-3.87%	-2.37%
		Stuttgart	$5.80	997	-2.81%	5.21%	53.45%
		Tel Aviv	$11.62	2,706	5.41%	9.13%	29.86%
		Turin	$5.21	449	9.04%	-2.98%	-40.49%
		Ulsan	$3.75	396	1.08%	-9.69%	-38.45%
		Vienna	$13.82	1,086	19.42%	33.58%	198.52%
		Warsaw	$4.46	3,016	2.30%	6.81%	30.76%
		Wroclaw	$3.72	473	10.54%	17.55%	44.29%

We will use this for manual changes to multiplier

Estimate of Global Impact

This experiment is a conditional AB-test, meaning there are operational scooters that are not assigned. Only scooters that reach below 60% battery are assigned. This is a non-random selection, meaning we cannot simply use the population size to estimate the market-wide treatment effect. Instead, we use the *revenue-share* of assigned scooters to estimate the overall impact on revenue.

Revenue Impact



Instead of weighting based on number of scooters, we weight the treatment effect based on revenue-share for a non-random-sample assignment experiment.

Say the treatment effect found for charging is α. Assuming the topline treatment effect only impacts the rebalanced scooters (i.e. no network effect), then we can extrapolate the total revenue lift as:

$$Overall\ Revenue\ Impact \ = \frac{R_{control}*(N_{assigned}/N_{control})}{R_{total,0}}*\alpha$$

We compute the revenue share for assigned and unassigned scooters using the week of 06/22/2020 to 06/28/2020. We use our experiment treatment effect on DSRVD, +6.87%.

$$\frac{R_{control}*(N_{assigned}/N_{control})}{R_{total,0}}*\alpha = 0.4877*(0.0687) = 0.0335$$

Thus we expect this feature improves gross revenue by **+3.35%**.

Cost Impact

$$Overall\ Cost\ Impact\ =\ \frac{C_{control}*(N_{assigned}/N_{control})}{C_{total,0}}\left(\alpha_{cost}\right) = (0.36235)(0.2170) = 0.0786$$

Thus we expect this feature increases juicer costs by **+7.9%**.

Gross Profit Impact

$$Overall\ Net\ Revenue\ Impact\ =\frac{(R-C)_{control}*(N_{assigned}/N_{control})}{(R-C)_{total,0}}\left(\alpha_{net\ revenue}\right) = (0.4966947)(0.0382) = 0.01897$$

Thus we expect the gross profit to increase by **+1.9%**.

Next Steps

- Increase to 80/20 split in all significant positive and insignificant regions.
- Configure multiplier in regions with statistically negative results. [doc]
- Check region-day level features (weather, RVD, etc) and compare to model performance