UBER_JCCP_MDL_005620999

Metadata

#Author	dharmin@uber.com	SEMANTIC
#Date Modified	04/13/2021	SEMANTIC
#DateCreated	04/13/2021	SEMANTIC
#Title	[ACP] S-RAD questions from Sukumar	SEMANTIC
All Custodians	Uber Technologies, Inc.;	SEMANTIC
All Paths	Uber Technologies, Inc.: \MassTort_Category3_DRIVE\MassTort_Category3_DRIVE_3.zip	SEMANTIC
Application	Microsoft 2007 Word Document	SEMANTIC
Begin Family	UBER_JCCP_MDL_005620999	SEMANTIC
Collaborators	eboman@uber.com; srad-working-group@uber.com	SEMANTIC
Confidentiality	Confidential	SEMANTIC
Date Created	04/13/2021 3:34 am	SEMANTIC
Date Modified	04/13/2021 4:29 pm	SEMANTIC
Document Type	4/13/2021	SEMANTIC
End Family	UBER_JCCP_MDL_005621001	SEMANTIC
File Path	\MassTort_Category3_DRIVE\MassTort_Category3_DRIVE_3.zip	SEMANTIC
File Size	13190	SEMANTIC
Filename	-ACP- S-RAD questions from Sukumar_1F8_riGCNKGjp3LSm1dMuN9_NdHK8Co28r6NwHvpBP8g.docx	SEMANTIC
GoogleDocumentType	DOCUMENT	SEMANTIC
Hidden Content	Yes;	SEMANTIC
ILS All Bates	UBER_JCCP_MDL_005620999;UBER_JCCP_MDL_005621000;UBER_JCCP_MDL_005621001	SEMANTIC
ILS Prod Date	05/12/2025	SEMANTIC
ILS Prod Vol	JCCP_MDL181	SEMANTIC
LINKSOURCEBEGBATES	UBER_JCCP_MDL_003384231	SEMANTIC
Other Custodians	Uber Technologies, Inc.;	SEMANTIC
Primary Date	04/13/2021 3:34 am	DOC_TYPE_ALIAS
Production Volume	JCCP_MDL181;	SEMANTIC
Redacted	Yes	SEMANTIC
Sort Date	04/13/2021 4:29 pm	SEMANTIC
SourceHash	3c9d0df41910cdb3f3c0065e3fffbd9a	SEMANTIC

Trial Exhibit No.
P-01630

Attorney-Client Privileged & Confidential

FRANK IS RESPONDING IN E-MAIL

Commented [1]: Thank you!

Commented [2]: np. I thought about it and it has to come from me.

Q: Do we have 2 parameters P/R or 3 parameters P/R/Trigger rate we are trying to make trade-offs across?

There is a single free parameter (trigger rate / model threshold); this selection determines precision and recall.

However, the trade-off is between safety incident reduction (where P/R are defined) and marketplace impact (e.g. average/p95/p99 post-dispatch ETA, Rider cancellation rate and C/R).

Marketplace impact cannot be measured either offline or in shadow mode, so trigger rate (% of trips where an S-RAD recommended plan would be dispatched instead of best ETA plan) is our proxy for discussion prior to experimentation.

For safety incident reduction, we could have chosen either precision or recall; the convention has been to use recall but the data science team has done research into using precision instead. We can share the white paper separately.

After deliberation and consensus, the trigger rate is translated to a model threshold for implementation and a switchback experiment is performed to estimate the trade-off. These results are then presented again to business stakeholders to make a roll-out decision.

NOTE: There are myriad reasons why switchback experiments are the only reliable method for computing marketplace impact (vs. random saturation, A/B, synthetic control). We have discussed this at length, over months, with the marketplace experimentation team and have a consensus both on the experimentation method as well as the calculation of marketplace results.

Q: What is the science behind triggering at 1.25%? "input from ops" in my book is usually not a scientific way to do this. Why can't we learn that threshold/use statistics to estimate it?

Over the past two and a half years, we have created multiple models with variations on gender variables (more below). These models have been put into experimentation with proxy trigger rates between 1% and 2% to better understand the trade-off. Each iteration pre- and post-experiment has been reviewed by Ops, Legal, Policy, Comms, Marketplace, and Mac. Depending on the circumstances, other ELT members, such as Tony, Jill, Sundeep, and Dara, have been involved in the review.
The 1.25% proxy trigger rate was selected because it provided the optimal trade-off across all stakeholders.

Attorney-Client Privileged & Confidential

NOTE: The relative rarity of safety incidents increases the needed switchback experiment period to get reliable results. Ideally, the experiment needs to run for a few months before we can describe the safety impact, so we only have the ability to test a handful of proxy trigger rates.

Q: I am not sure what the variables "driver_male_rider_different" and "driver_female_rider_different" actually mean? Why not just use rider/driver gender and let the model figure out the weights etc.
Q: What the curves suggest is gender is almost as important as all the other driver attributes combined

The definitions of the variables:

- `driver_male_rider_different`: the driver is male and the rider is of a different gender
- `driver_female_rider_different`: the driver is female and the rider is of a different gender

Past models have included direct gender variables for rider/driver, no gender variables whatsoever, and the two highlighted variables.

While direct gender variables provide the strongest model, the use of gender comes with additional considerations, which we must weigh against performance on safety incidents. Many rounds of experiments and model iterations focused on gender; these are summarized below:

1. Estimating disparate earning impact - research was done through a fairly complex experiment design.
2. REDACTED - PRIVILEGED
3. (now) Gender-agnostic comparison

REDACTED - PRIVILEGED

Q: Given the data sparsity, would we consider an ensemble like approach that risk has taken where you build both precision optimized and recall optimized models and make a decision based on multiple models?

We are open to new approaches and recently collaborated with the risk team on a similar model (SDM) to test their multi-stage approach, which did not produce improved results.

Please note that, for S-RAD, the trade-off is against direct marketplace impact, which is not easy to measure, and, as above, our feature space has been under discussion.
Finally, this model will be reviewed at an upcoming Senior Science Review (audience Dir+ DS with some L6+), where the goal is to pressure-test significant models against a broad DS audience.

~~It also looks like we are making a prediction based on a <driver, rider> pair when the reality is there is correlation across <driver,*> and <*,rider> pairs. In reality the likelihood of an event perpetrated by one party has a proportion (value TBD) that is independent of who the other party is. The modelling here is complex and there are policy considerations we want to have in how we design the models~~

Commented [3]: Let's not go overboard. This is a suggestion, not a question and I'm going to focus on questions only.

My other point is any discussion of models with P/R and AUC curves always raises questions about the trade-offs we want to make. I would like to see for any/all models not just P/R and AUC but also Tp, Tn, Fp, Fn.

Gender-based

	incident	no-incident
flagged	207	1,963,170
non-flagged	735	155,106,116

Gender-agnostic

	incident	no-incident
flagged	168	1,963,209
non-flagged	774	155,106,077

Q: Do these type of models go through a review process? Who is on the reviews and what feedback is provided?

Commented [4]: @frank@uber.com - can you please help frame this?

Q: Finally have we considered DL models?