S-RAD Model Overview

Safety & Insurance | July 2018

Sunny Jeon | Senior Data Scientist
sjeon@uber.com

Summary

To support Uber's company-wide priority to reduce critical safety incidents, the Safety & Insurance team has developed a data-driven intervention for preventing sexual assaults. The intervention -- called **Safety Risk Assessed Dispatch (S-RAD)** -- consists of two components:

- Machine learning models that assess the safety risks associated with all potential driver-rider matches at the point of dispatch.
- A down-ranking procedure that incorporates safety risk scores when selecting the optimal driver for dispatch (subject to marketplace constraints).

Evidence suggests **this approach *may* be able to prevent up to 15% of sexual assaults in the US by down-ranking 1% of the highest risk trips.** The following sections elaborate on the intervention strategy and the performance of the v1 model:

1. Overview [back to top]

Sexual assaults[1] are one of the most tragic and infuriating forms of interpersonal conflicts that could occur on the Uber platform. To protect our users from these crimes, the Safety & Insurance team has developed a data-driven intervention that uses machine learning to score the safety risks associated with all potential driver-rider matches at the point of dispatch. The team proposes to use these scores to down-rank[2] high-risk matches subject to a set of marketplace constraints to minimize the risk of sexual assaults.

The motivating insight behind this intervention -- called **Safety Risk Assessed Dispatch (S-RAD)** -- is that many sexual assaults have correlates and precursors that may enable Uber to better anticipate them, and hence, prevent them by implementing well-designed safety products. For example, in the US, sexual misconduct and sexual assaults are disproportionately more likely to occur on:

- Trips that occur late night and on the weekends.
- Trips that originate from a bar area.
- Trips with a driver and rider of different genders.

**See descriptive statistics in Appendix A: Descriptive Statistics.*

This report provides evidence that by leveraging these types of signals, **it may be possible to predict 15% of sexual assaults that occur on the Uber platform in the US by flagging 1% of the highest risk trips** (evidence from v1 model applied to historical out-of-sample test set). Figure 1 depicts the percent of sexual assaults that can be correctly predicted (recall) for every incremental increase in the percent of trips flagged.

Prediction, however, does not equal prevention. The actions and interventions triggered by the model predictions need to change real-world outcomes. The down-ranking strategy proposed here may not have any impact in supply-constrained markets where alternative matches do not exist. And even where there is a surplus of supply, down-ranking may not be sufficient to keep predators from eventually exploiting the platform. Finally, it may not be possible to implement model-based interventions without generating prohibitive costs on the marketplace, such as

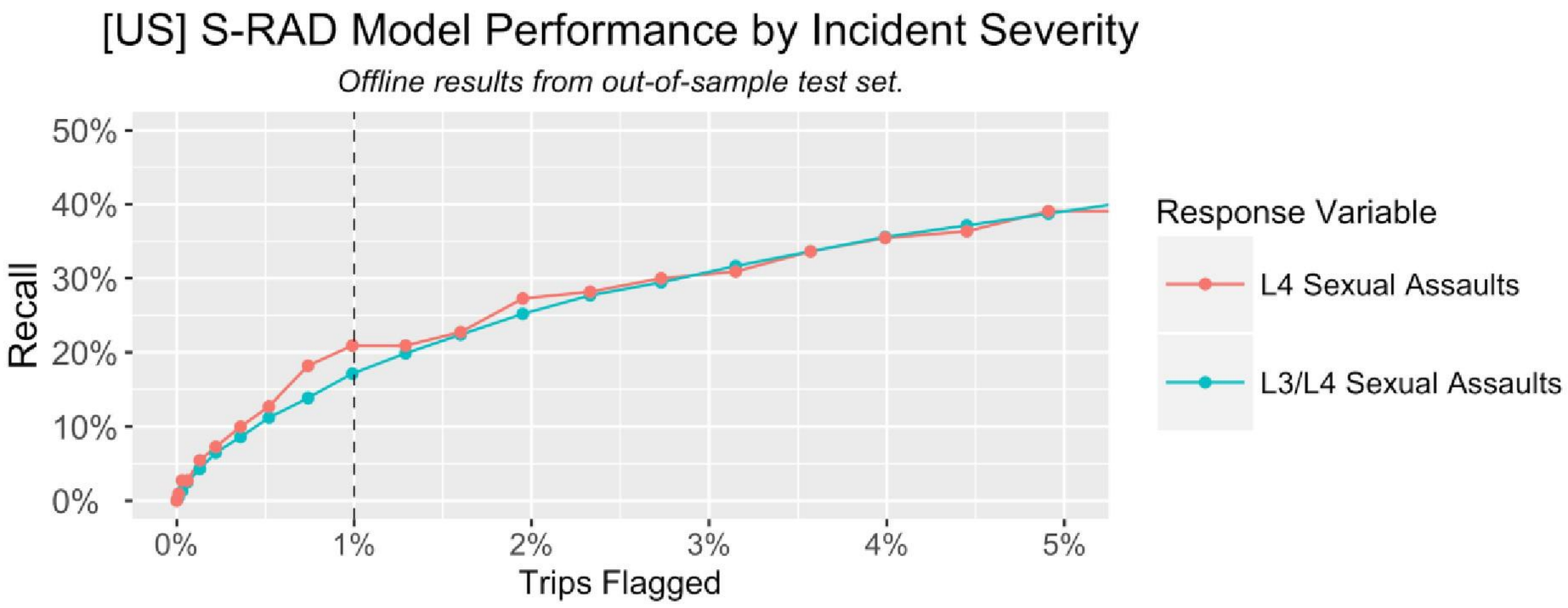
---

[1] 'Sexual assaults' are defined here as any kind of non-consensual touching or intercourse, and indecent exposure. Table 1 provides incident types according to Safety & Insurance's incident taxonomy.

[2] 'Down-ranking' is a marketplace intervention in which a particular driver-rider pair that is considered for dispatch is given an MGV (Market Generated Value) cost such that the dispatch is considered less valuable than other non-down-ranked alternative matches considered when optimizing matches across a market. Within the current dispatch framework, down-ranking is effectively blocking the pair unless there are no alternative pairs available (a.k.a., "soft filter").

undermining the ability of drivers and riders to use the platform at specific times or locations, or if they belong to a particular gender.

Figure 1



Notes: US P2P trips (including Pool) between March 1 - May 31, 2018 (100% of positives, 1% sample of negatives). Recall for all sexual assaults with both driver + rider reporters (source: JIRA). 3358 L3/L4 incidents and 110 L4 incidents in test set. Results from Female Model V1 RC 26 and Male Model V1 RC 22.

Randomized experiments will ultimately be necessary to answer these questions in a compelling way. But to demonstrate that S-RAD is viable and experiments can be conducted with minimal legal and marketplace risks, **S-RAD was deployed into shadow mode in Los Angeles on April 20, 2018**. Starting on this day, all driver-rider pairings considered for dispatch for Los Angeles UberX trip requests were scored for their safety risks in near real-time (but not actioned on). The results suggest that S-RAD is a viable strategy for prevention. More specifically, during the shadow period (April 20 - June 12):

- S-RAD correctly anticipated **15% of sexual assaults** occurring on Los Angeles UberX trips (13/85) at the 1% trip trigger rate. This included **38% of L4 sexual assaults** (3/8).

- 99% of trips flagged by the model (at the 1% trip trigger rate) had an alternative driver-rider match that passed the model check (i.e., scored under risk threshold).

- Female and male drivers are not flagged at a different rate.

**Full results available in Section 5: Shadow Mode Results.*

The sections below elaborate on the methodology, the performance of the v1 model, the most important predictors, results from shadow mode in Los Angeles, and the proposed implementation plan, which involves a series of randomized evaluations consisting of a limited pilot, a within-city switchback, and a multi-city experiment (pending discussions with Operations, Marketplace, Legal). Positive findings from these experiments -- from safety, marketplace, and legal perspectives -- are a requirement for scaling.

ATTORNEY-CLIENT PRIVILEGED

2. Methodology [back to top]

The S-RAD model consists of two gender-specific models trained to predict sexual assaults at the driver-rider dyad level. More details:

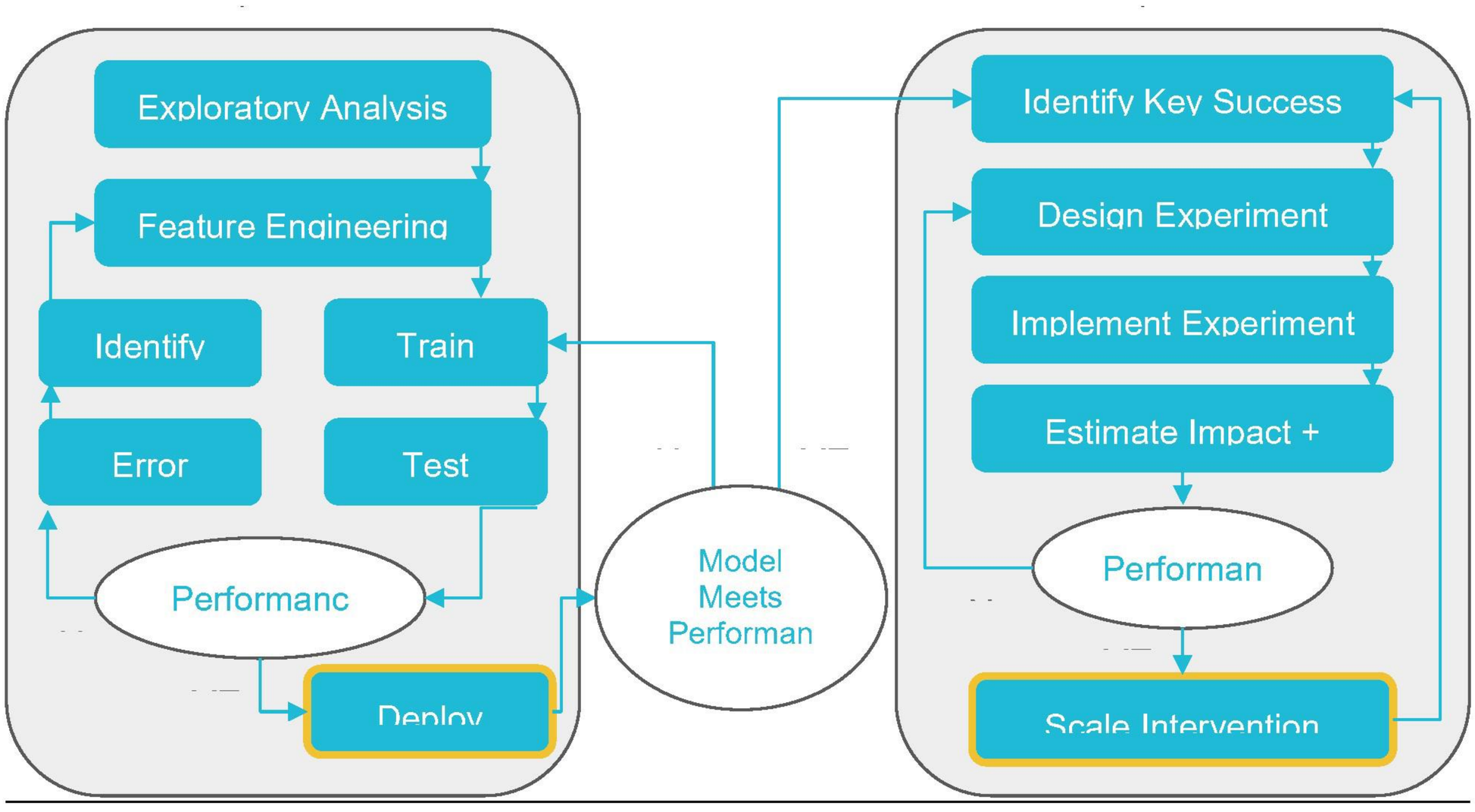
- **Gender-specific models:** separate models are trained for female drivers and male drivers because female drivers face unique risks that are better modeled using a tailored feature set.
- **Binary response variable**: 0, 1 if trip ends in a sexual assault for all models. In the training process, multiple operationalizations are considered for labels -- e.g., train on just L3/L4 sexual assaults versus all levels versus sexual assaults in addition to sexual misconduct. Types of sexual assaults and sexual misconduct under the safety ticket taxonomy is provided in Table 1.
- Use **Gradient Boosted Decision Trees** to generate predictions since the algorithm has strong predictive power, provides the ability to set a wide range of hyperparameters, and because boosting may work better (than bagging) at learning rare events.
- To deal with extreme class imbalance (i.e., incidents are rare events), use **down-sampling** to help model learn to discriminate between positive and negative cases.
- Use **out-of-sample testing protocol** to measure predictive performance (e.g., recall, precision, trigger rate computed on out-of-sample test set).
    - Optimal tuning parameters (number of iterations, interaction depth, learning rate/shrinkage, and minimum number of observations per node) identified using grid search, selecting those that maximize **recall at the 1% trip trigger rate**.
- **Curated feature set** based on past research on the correlates and causes of sexual assaults. Table 2 describes the types of features considered for the v1 model. The full S-RAD variable list + definitions is available here.
    - The final feature set is identified through an iterative process of hypothesis generation and testing in which unimportant predictors are eliminated and more powerful predictors introduced. Figure 2 depicts the model development, testing, and iteration process.

- Over 200 features considered. Final models include <40 features, which have been tailored to optimize predictive performance by driver gender.

Table 1: Sexual Assaults and Sexual Misconduct Incident Types

Sexual Assault	Sexual Misconduct	
Physical - Attempted / Accidental Touching	Staring or Leering	Verbal - Non-explicit Inappropriate Remark / Gesture
Physical - Forced or Sexual Touching	Verbal - Explicit Inappropriate Remark / Gesture	Verbal - Non-explicit Inappropriate Remark / Gesture - comments on appearance
Physical - Masturbation / Indecent Exposure	Verbal - Explicit Inappropriate Remark / Gesture - comments on appearance	Verbal - Non-explicit Inappropriate Remark / Gesture - flirting
Physical - Non-Consensual Sexual Intercourse	Verbal - Explicit Inappropriate Remark / Gesture - flirting	Verbal - Non-explicit Inappropriate Remark / Gesture - personal questions
	Verbal - Explicit Inappropriate Remark / Gesture - personal questions	Verbal - Non-explicit Inappropriate Remark / Gesture - soliciting more contact
	Verbal - Explicit Inappropriate Remark / Gesture - soliciting more contact	Verbal - Threat of Sexual Assault

Figure 2: Framework for Model Development, Testing, and Iteration



Attorney-Client Privileged and Confidential

Table 2: Predictors of Sexual Assaults

Trip-level	Driver-level	Rider-level
Businesses at Request Point (e.g., number of bars, restaurants, hotels, etc. in request point)	Driver Feedback from Riders (e.g., rider feedback describing creepy + aggressive behavior)	Device Attributes (e.g., model, device verification status, shared users, banned device, WIFI connected)
Day / Time (e.g., day of week, day of month weekend/weekday, late night v. day time)	Gender	Fraud Scores (e.g., risk scores representing probability of arrears, bans, chargebacks)
Fraud Attributes of Request Point (e.g., banned trips and unsettled fares by request geohash)	Previous Safety Incidents (e.g., incident rate, driving v. interpersonal conflicts)	Gender
Holidays (e.g., binary variable indicating national holiday or day of major celebration)	Promotions (e.g., sign-up with promos, percent of trips with promotions)	Previous Safety Incidents (e.g., incident rate, driving v. interpersonal conflicts)
Gender Difference (e.g., driver + rider different genders)	Ratings (e.g., average rating, percent 1 star ratings, ratings by opposite gender)	Promotions (e.g., cash sign-up, percent of trips with promotions)
Product Type, Fare + Surge (e.g., UberX v. Black v. SUV v. Eats)	Signup Attributes (e.g., channel, method, time, location, app language, email, referral, payment profile, card type)	Ratings (e.g., average rating, percent 1 star ratings, ratings by opposite gender)
Safety Attributes of Request + Destination Points (e.g., safety incident rate by request geohashes, destination attributes like trip counts and cancellation rates)	Telematics (e.g., speed, hard brakes + accelerations, phone handling, jerking)	Signup Attributes (e.g., channel, method, time, location, app language, email, referral, payment profile, card type)
Trips + Cancellations at Request Point	Tenure (e.g., days active)	Tenure (e.g., days active)
Weather	Trips (e.g., trips on P2P v. Black v. Eats, trips at night + weekends, cancellations)	Trips (e.g., trips on P2P v. Black v. Eats, trips at night + weekends, cancellations)

Notes: All predictors are computed using only data that is available before driver dispatch. All predictors except those in shaded cells have been tested for the v1 model described in this document. The full S-RAD Variable List + Definitions available here.

3. Model Performance [back to top]

The accuracy of the S-RAD model is assessed on two primary metrics:

- *Recall*: percent of sexual assaults correctly anticipated.
- *Trip Trigger Rate*: percent of completed trips flagged as high-risk.

The goal is to maximize **recall** (predictive performance) subject to a constraint or "budget", which is defined here as the **trip trigger rate** (*threshold to be selected with stakeholders*). All recall and trigger rate metrics are computed on an out-of-sample test set that is never analyzed before generating model predictions.

Over 50 models were tested to select the v1 S-RAD model presented in this report, with each model differing on at least one of the following dimensions: feature set, hyperparameters, training labels, down-sampling ratio, and training set time period.

The best performing v1 model is able to achieve **>15% recall on sexual assaults at the 1% trip trigger rate** in out-of-sample test sets. The v1 model has the following attributes:

- *Model:* Gradient Boosted Decision Trees (Binary Classification)
- *Sample:* All US sexual assaults and random sample of 3.5M US P2P trips
- *Training Data:* December 1, 2016 - February 28, 2018
- *Test Data:* March 1 - May 31, 2018
- *Partitions:* Separate models for female drivers and male drivers
- *Female Driver Model:*
    - *Number of Positive Labels: 4,519 in training set, 880 in test set.*
    - *Number of Features:* 34
    - *Hyperparameters:* depth 5, bin 25, minimum rows 100, learning rate 0.01, balance 1, 250 iterations
    - Link to Michelangelo Model
- *Male Driver Model:*
    - *Number of Positive Labels: 11,865 in training set, 2,588 in test set.*
    - *Number of Features:* 36
    - *Hyperparameters:* depth 5, bin 25, minimum rows 100, learning rate 0.01, balance 1, 250 iterations
    - Link to Michelangelo Model

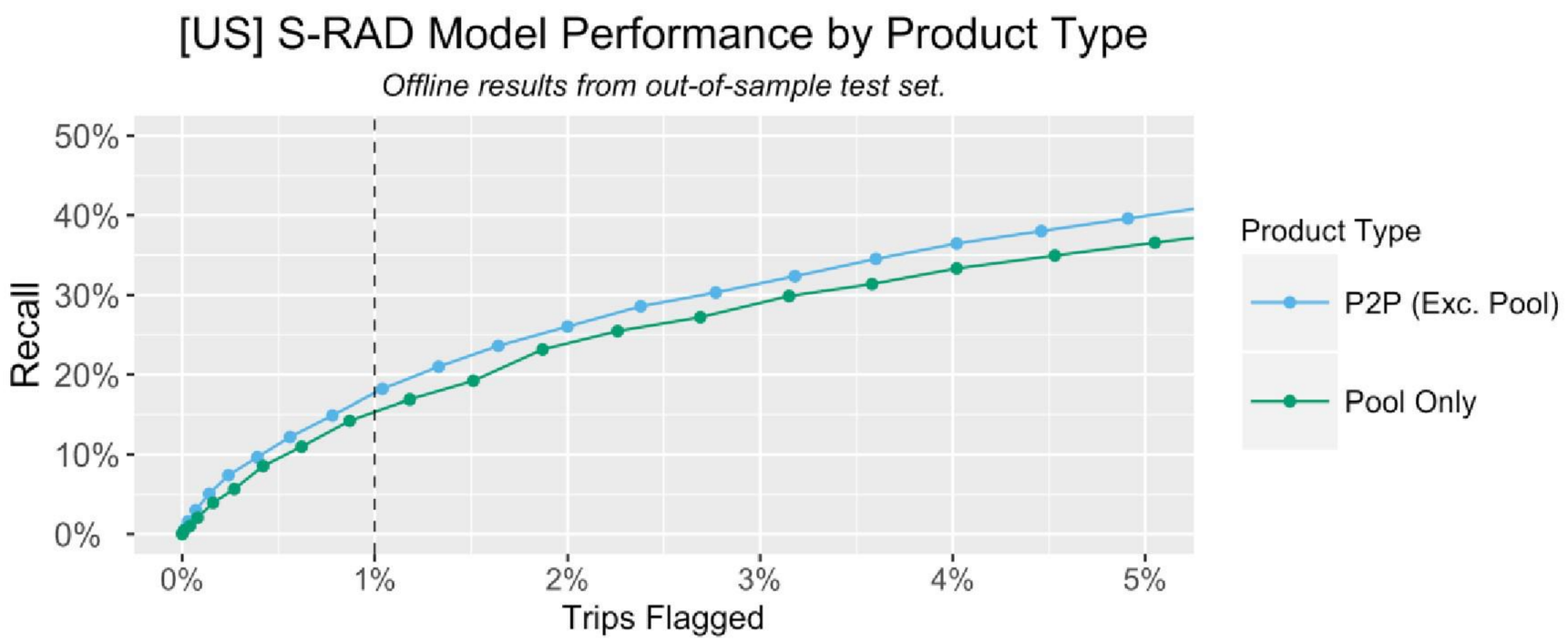
One of the advantages of this machine learning-based approach is the ability to tune the model to catch more incidents by flagging more trips (but at the cost of more false positives), or to reduce business impact by flagging only the trips where there is high confidence it is high risk.

Figure 3 shows this relationship by plotting recall as a function of trip trigger rate. The figure further breaks-down performance by P2P excluding Pool versus Pool only to shed more light on model performance. In some respects, incidents on Pool may be more difficult to anticipate because the model is trained to assess risk at the driver-rider level and not the rider-rider level. Indeed, results indicate that the model is better at predicting incidents on non-Pool P2P trips, but the difference is marginal (18% recall v. 15% recall at 1% trip trigger rate in out-of-sample test set).

Results also reveal that predictive performance is stronger for sexual assaults involving female drivers compared to male drivers (see Figure 4). **In the majority of cases involving female drivers, the female driver is the alleged victim and the alleged offender is the rider.** The recall-trigger rate curve in Figure 5 shows that the female driver model is able to achieve >20% recall on these types of incidents with a rider alleged offender (534 cases on P2P excluding Pool, 83 on Pool in test set). The model performs substantially worse for incidents involving a female driver alleged offender (likely due to small sample size; 34 cases on P2P excluding Pool, 9 on Pool in test set).

For sexual assaults involving male drivers, the male driver is more often than not the alleged offender. Figure 6 shows the model is marginally better at predicting these incidents where the alleged offender is the male driver (522 cases on P2P excluding Pool, 273 on Pool) rather than the rider (951 cases on P2P excluding Pool, 368 on Pool), but the difference is small.

Figure 3



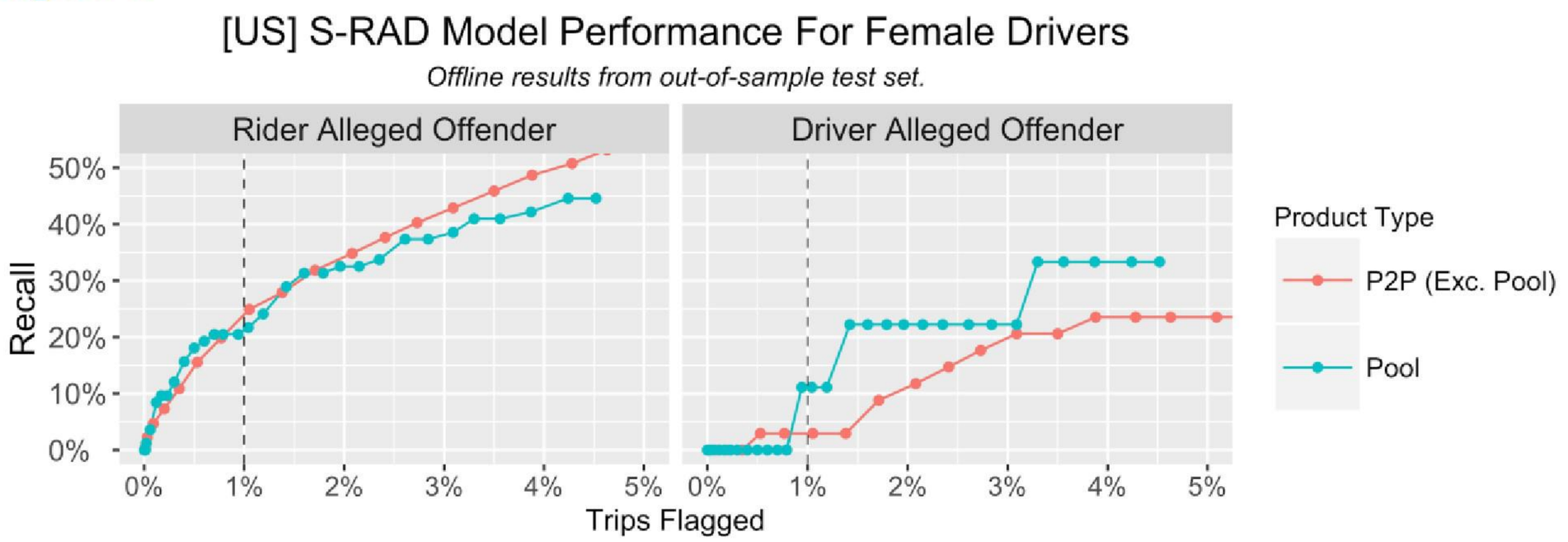
Notes: US P2P trips between March 1 - May 31, 2018 (100% of positives, 1% sample of negatives). Recall for all sexual assaults with both driver + rider reporters (source: JIRA). 2482 incidents on P2P (exc. Pool), 864 incidents on Pool. Results from Female Model V1 RC 26 and Male Model V1 RC 22.

Attorney-Client Privileged and Confidential

Figure 4

Notes: US P2P trips (exc. Pool) between March 1 - May 31, 2018 (100% of positives, 1% sample of negatives). Recall for all sexual assaults with both driver + rider reporters (source: JIRA). Female Drivers: 733 incidents on P2P (Exc. Pool); Male Drivers: 1749 incidents on P2P (Exc. Pool). Results from Female Model V1 RC 26 and Male Model V1 RC 22.

Figure 5



Notes: US P2P trips between March 1 - May 31, 2018 (100% of positives, 1% sample of negatives). Recall for all sexual assaults with both driver + rider reporters (source: JIRA). Rider Alleged Offender: 534 incidents on P2P (Exc. Pool); 83 incidents on Pool. Driver Alleged Offender 34 incidents on P2P (Exc. Pool); 9 incidents on Pool. Results from Female Model V1 RC 26 and Male Model V1 RC 22.

Figure 6

Notes: US P2P trips between March 1 - May 31, 2018 (100% of positives, 1% sample of negatives). Recall for all sexual assaults with both driver + rider reporters (source: JIRA). Rider Alleged Offender: 522 incidents on P2P (Exc. Pool); 273 incidents on Pool. Driver Alleged Offender 951 incidents on P2P (Exc. Pool); 368 incidents on Pool. Results from Female Model V1 RC 26 and Male Model V1 RC 22.

4. Most Important Predictors [back to top]

The S-RAD model leverages a variety of trip, driver, and rider-level predictors to detect driver-rider matches with elevated risk of sexual assaults (full variable list and definitions available here). Different feature sets are used for the female driver model and the male driver model to tailor the predictors to the different types of risks female drivers and male drivers may face.

To identify the most important predictors, variable importance ranks are computed for each feature for each model. These ranks are then aggregated by "feature bundles" -- groups of predictors representing conceptually similar variables. Table 3 provides variable importance ranks by feature bundle (ordered by harmonic mean rank). Figure 7 plots the variable importance ranks of each predictor.

Insights on critical predictors:

- **Spatial** predictors are the most important bundle for both the female and male driver models. These spatial variables capture key environmental risk signals, such as the number of bars and the historical safety incident rate at a trip request point.

- **Driver Prior Feedback** variables are the second most important bundle for both the female and male driver models. These variables operationalize different types of feedback that may be correlated with unsafe behavior, such as the number of 1-star ratings and safety complaints they received historically.

- **Client Trips** features are the third most important bundle for both the female and male driver models. These variables capture client trip patterns that may be associated with an elevated risk of sexual assaults, such as a high proportion of trips late night on the weekends, or a high historical average ATA, which could indicate the client frequently travels to and from an area that is remote and far from where supply is normally positioned.

Comparisons of variable importance ranks across the gender-specific models also generate interesting insights on the different risks that female drivers and male drivers may face, and how they might be better modeled in the next iterations. For example:

- Driver often use cancellations to avoid dangerous or undesirable parts of the city, and this **historical driver cancellation rate near a trip request point** (*Spatial - Geohash6 Driver Cancellation Rate 4week*) ranks as one of the most important predictors for the female driver model, but not the male driver model (in fact, the variable was not selected into the final v1 male driver model because of a lack of predictive power). One

of the implications here is that female drivers are more susceptible to these types of environmental risks distributed across space.

- Because female drivers are usually not the alleged offender, the **number of sexual assault / sexual misconduct complaints a driver received historically** (*Driver - Sexual Incidents Driver Alleged Offender*) is not important for predicting incidents involving female drivers. However, this variable ranks highly in importance for the male driver model.

- **Clients that receive low ratings from the opposite gender** (*Client - 1-Star Rate from Opposite Gender*) are a greater risk to female drivers compared to male drivers. By contrast, a **driver's historical rating from the opposite gender** (*Driver - 1-Star Rate from Opposite Gender*) is only important for the male driver model. The is consistent with the observation that sexual assaults involving female drivers usually involve a male client offender, whereas sexual assaults involving male drivers usually involve a male driver offender.

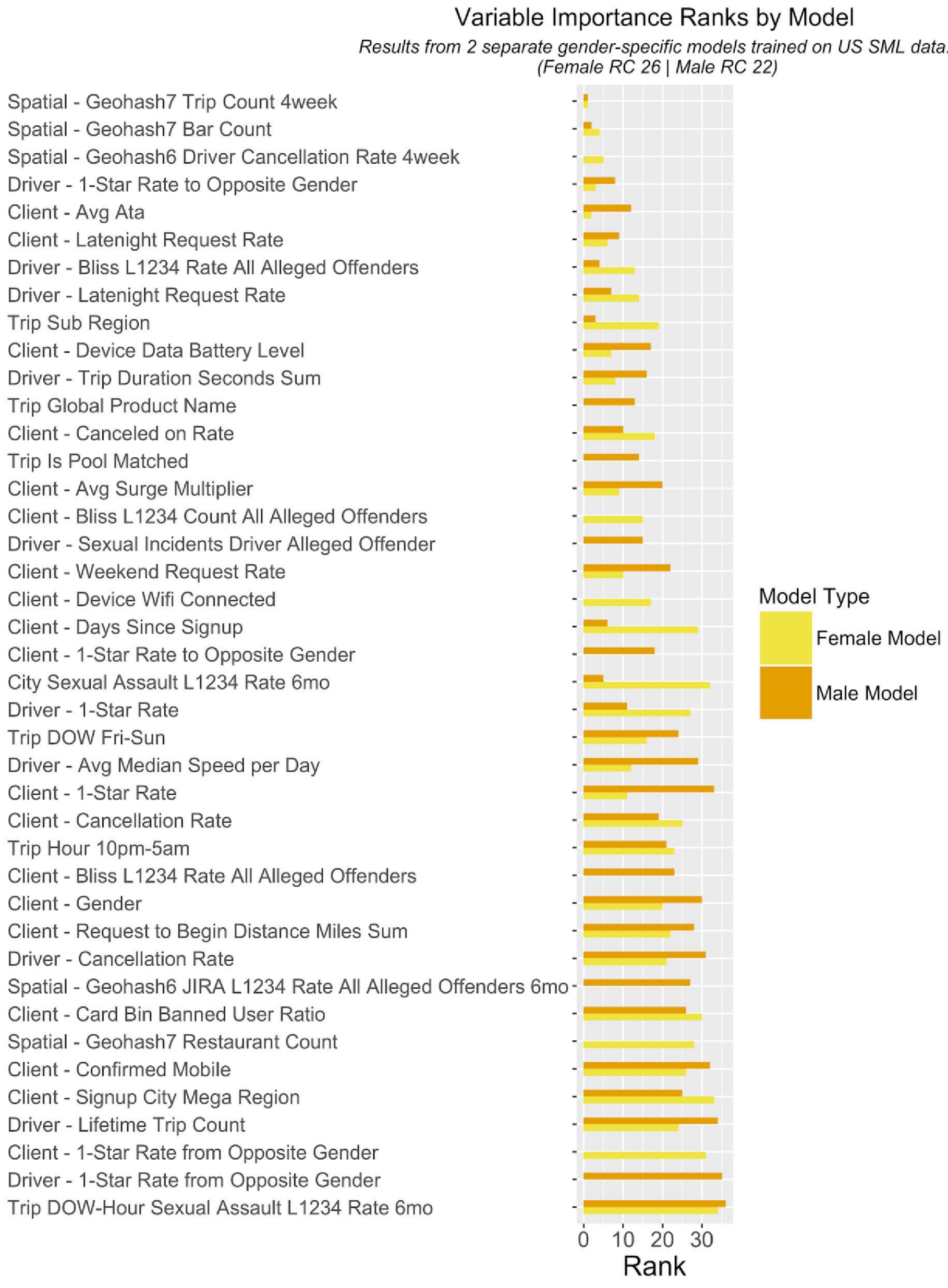
Table 3: Variable Importance Ranks by Feature Bundle and Model Type

	Feature Bundle	
Rank	Female Driver Model	Male Driver Model
1	Spatial	Spatial
2	Driver Prior Feedback	Driver Prior Feedback
3	Client Trips	Client Trips
4	Driver Trips	Product
5	Client Prior Feedback	Driver Trips
6	Client Account	Client Prior Feedback
7	Gender	Client Account
8	Time	Time
9	--	Gender

Notes: Features falling under each "feature bundle" are described in the S-RAD Variable List + Definitions spreadsheet. Bundle ranks computed by taking the harmonic mean of variable importance ranks of individual features in each feature bundle.

Attorney-Client Privileged and Confidential

Figure 7


Variable Importance Ranks by Model
Results from 2 separate gender-specific models trained on US SML data.
(Female RC 26 | Male RC 22)
Spatial - Geohash7 Trip Count 4week
Spatial - Geohash7 Bar Count
Spatial - Geohash6 Driver Cancellation Rate 4week
Driver - 1-Star Rate to Opposite Gender
Client - Avg Ata
Client - Latenight Request Rate
Driver - Bliss L1234 Rate All Alleged Offenders
Driver - Latenight Request Rate
Trip Sub Region
Client - Device Data Battery Level
Driver - Trip Duration Seconds Sum
Trip Global Product Name
Client - Canceled on Rate
Trip Is Pool Matched
Client - Avg Surge Multiplier
Client - Bliss L1234 Count All Alleged Offenders
Driver - Sexual Incidents Driver Alleged Offender
Client - Weekend Request Rate
Client - Device Wifi Connected
Client - Days Since Signup
Client - 1-Star Rate to Opposite Gender
City Sexual Assault L1234 Rate 6mo
Driver - 1-Star Rate
Trip DOW Fri-Sun
Driver - Avg Median Speed per Day
Client - 1-Star Rate
Client - Cancellation Rate
Trip Hour 10pm-5am
Client - Bliss L1234 Rate All Alleged Offenders
Client - Gender
Client - Request to Begin Distance Miles Sum
Driver - Cancellation Rate
Spatial - Geohash6 JIRA L1234 Rate All Alleged Offenders 6mo
Client - Card Bin Banned User Ratio
Spatial - Geohash7 Restaurant Count
Client - Confirmed Mobile
Client - Signup City Mega Region
Driver - Lifetime Trip Count
Client - 1-Star Rate from Opposite Gender
Driver - 1-Star Rate from Opposite Gender
Trip DOW-Hour Sexual Assault L1234 Rate 6mo
0
10
20
30
Rank
Model Type
Female Model
Male Model

5. Shadow Model Results (Los Angeles) [back to top]

As a first step towards demonstrating that S-RAD can prevent sexual assaults within marketplace constraints, S-RAD was deployed in shadow mode on April 20, 2018. Starting on this day, all driver-rider pairings considered for dispatch for Los Angeles UberX trip requests were scored for their safety risks in near-real time (but not actioned on). On average, there were 175k jobs per day and 8M driver-rider pairings considered for dispatch per day.

Results provide evidence that S-RAD is a viable strategy for prevention. More specifically, during the shadow period (April 20 - June 12):

- S-RAD correctly anticipated **15% of sexual assaults** occurring on Los Angeles UberX trips (13/85) at the 1% trip trigger rate. This included **38% of L4 sexual assaults** (3/8).
- 99% of trips flagged by the model (at the 1% trip trigger rate) had an alternative driver-rider match that passed the model check (i.e., scored under risk threshold).
- Female and male drivers are not flagged at a different rate.

Predictive Performance

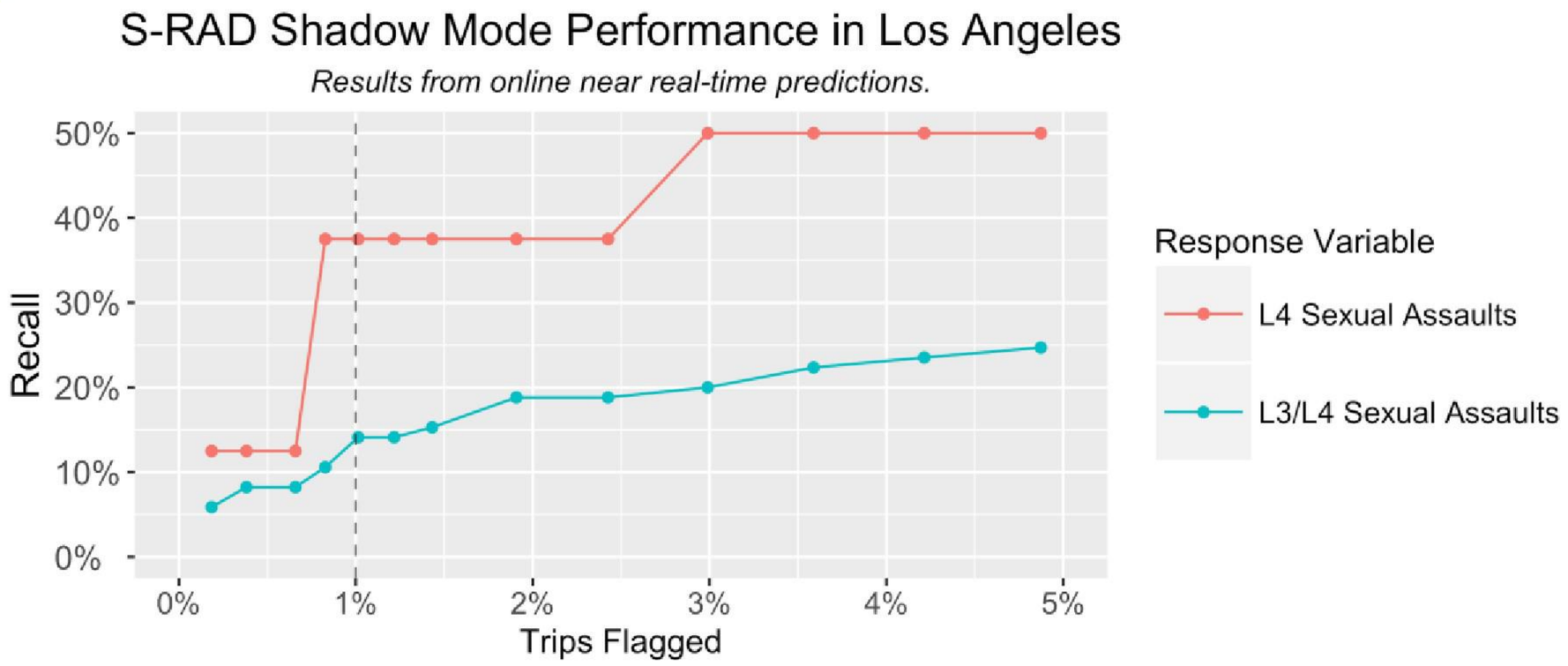
Figure 8 depicts the recall-trigger rate curve for shadow mode predictions for Los Angeles UberX trips. At the 1% trip trigger rate, the model predicts 15% of sexual assaults, including 38% of L4 sexual assaults. If the model were calibrated at the 0.5% trip trigger rate, the model would have flagged ~8% of L3/L4 sexual assaults.

Importantly, however, prediction does not equal prevention. Because prevention requires down-ranking to effectively block driver-rider matches with elevated risk, there needs to exist alternative driver-rider matches for the flagged pair that are scored as low risk. This may not be possible when there is a shortage of supply, or if the model tends to flag all potential dispatches for a rider's request as risky (e.g., if rider features in the model dominate predictions).

Randomized experiments will ultimately be necessary to address these concerns in a compelling way, but as a first step we can examine if trips that were flagged in shadow (positives) had alternative driver-rider pairings that were considered for dispatch and passed the model check (e.g., had a risk score under the threshold for down-ranking). Furthermore, we can examine what the potential change in ETA and risk score would have been by dispatching the next best driver.

Attorney-Client Privileged and Confidential

Figure 8



Notes: Los Angeles UberX trips between April 20 - June 12, 2018. Recall for all sexual assaults with both driver + rider reporters (source: JIRA). 85 L3/L4 incidents and 8 L4 incidents. Results from V0 Combined Gender Model.

Table 4: Next Best Unflagged Match for Flagged Trips (LA Shadow; UberX Only)

	Flagged Trips with Incidents	Flagged Trips
Trips	13	5000
Trips with Alternative Match	92% (12/13)	>99% (4986/5000)
Median Change in ETA* (mean in parentheses)	+12 secs (+72 secs)	+32 secs (+52 secs)
Median Change in Model Risk Score* (mean in parentheses)	-0.11 (-0.20)	-0.13 (-0.20)

Notes: Model calibrated at the 1% trip trigger rate. Sample consists of all 13 sexual assaults correctly flagged (true positives) during April 20 - June 12, and a random sample of 5k flagged trips (positives) during June 1 - 7.

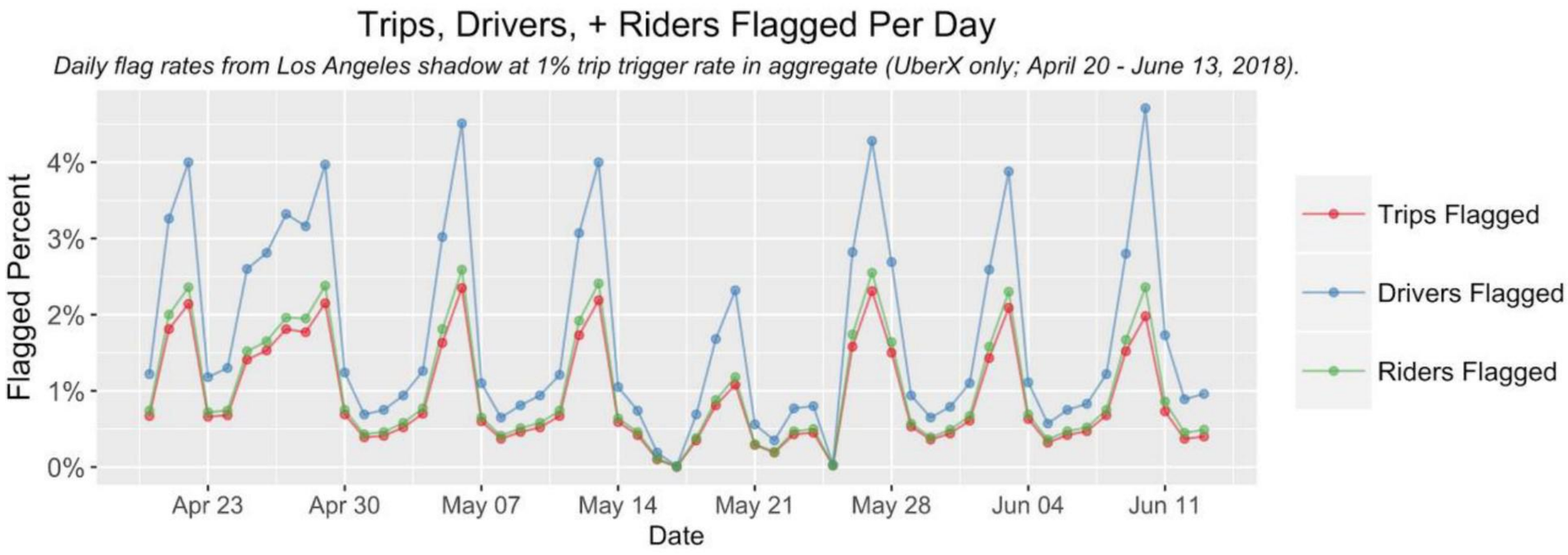
To generate these insights, two types of positives are sampled: all 13 true positives (flagged trips with incidents), 5k randomly selected positives (flagged trips). For each of these trips, we collect all driver-rider pairings considered for that trip request (according to rawdata.kafka_hp_multileg_mgv_log_nodedup), score the risks associated with the pairings, and examine the percent of trips with an alternative driver-rider pairing that passed the model check. Results are presented in Table 4, and show that 92% of true positives and 99% of positives had alternative driver-rider pairings that passed the model check. If the driver with the lowest ETA amongst all drivers passing the model check were dispatched, the median change in ETA is +12 seconds for true positives and +32 seconds for positives. **Note, however, that this analysis does not solve the global dispatch optimization problem used in production, so these high-level estimates need to be re-assessed via rigorous online experiments.**

Flag Rates at 1% Trip Trigger Rate

During the shadow mode period, there were ~500k UberX trips per day in Los Angeles. A 1% trip trigger rate thus equates to approximately ~5k down-ranked trips per day. However, because safety risks vary by time and space, flag rates can vary significantly depending on the unit of analysis (e.g., day-level, hour-level, hexagon-level).

Figure 9, for example, shows that at the 1% trip trigger rate calibration in the aggregate, the daily trip trigger rate fluctuates between 0.01% to 2.35% (<25 trips flagged per day to >15k trips flagged per day). On the average, 1.74% of drivers and 1.02% of riders taking trips are impacted each day.

Figure 9



Another way of assessing the potential marketplace costs is to examine the percent of **driver-rider pairs** flagged, where driver-rider pairs are comprised of all driver-rider pairs considered for an UberX request (not just those culminating into a trip).

At the 1% trip trigger rate, approximately 1.26% of all driver-rider pairings considered for dispatch are flagged for down-ranking each day on the average, with the highest flag rates occurring on the highest risk days (e.g., >3% on weekends). These flag rates are presented in Figure 10.

The rate at which driver-rider pairs are flagged also increase significantly on weekends between 12-3am -- the time of the day when the sexual assault incident rate is highest (see Appendix Figure A.1).

Finally, Figure 12 illustrates the distribution of flag rates across space (at the hexagon 7 level, with each hexagon 7 being approximately 5.16 km^2 in size). On the average, 1.22% of driver-rider pairs are flagged across each hexagon 7, with the highest flag rate being 6.35%.

Figure 10

Driver-Rider Pairings Flagged Per Day

Daily flag rates from Los Angeles shadow at 1% trip trigger rate in aggregate (UberX only; April 20 - June 13, 2018).

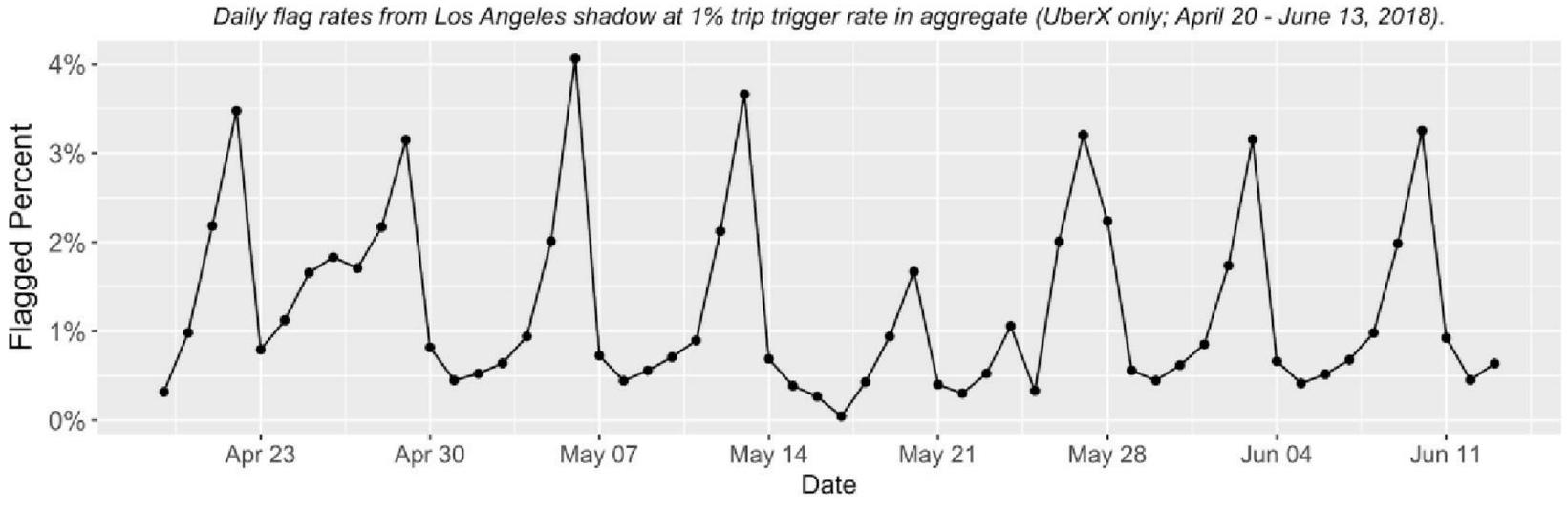
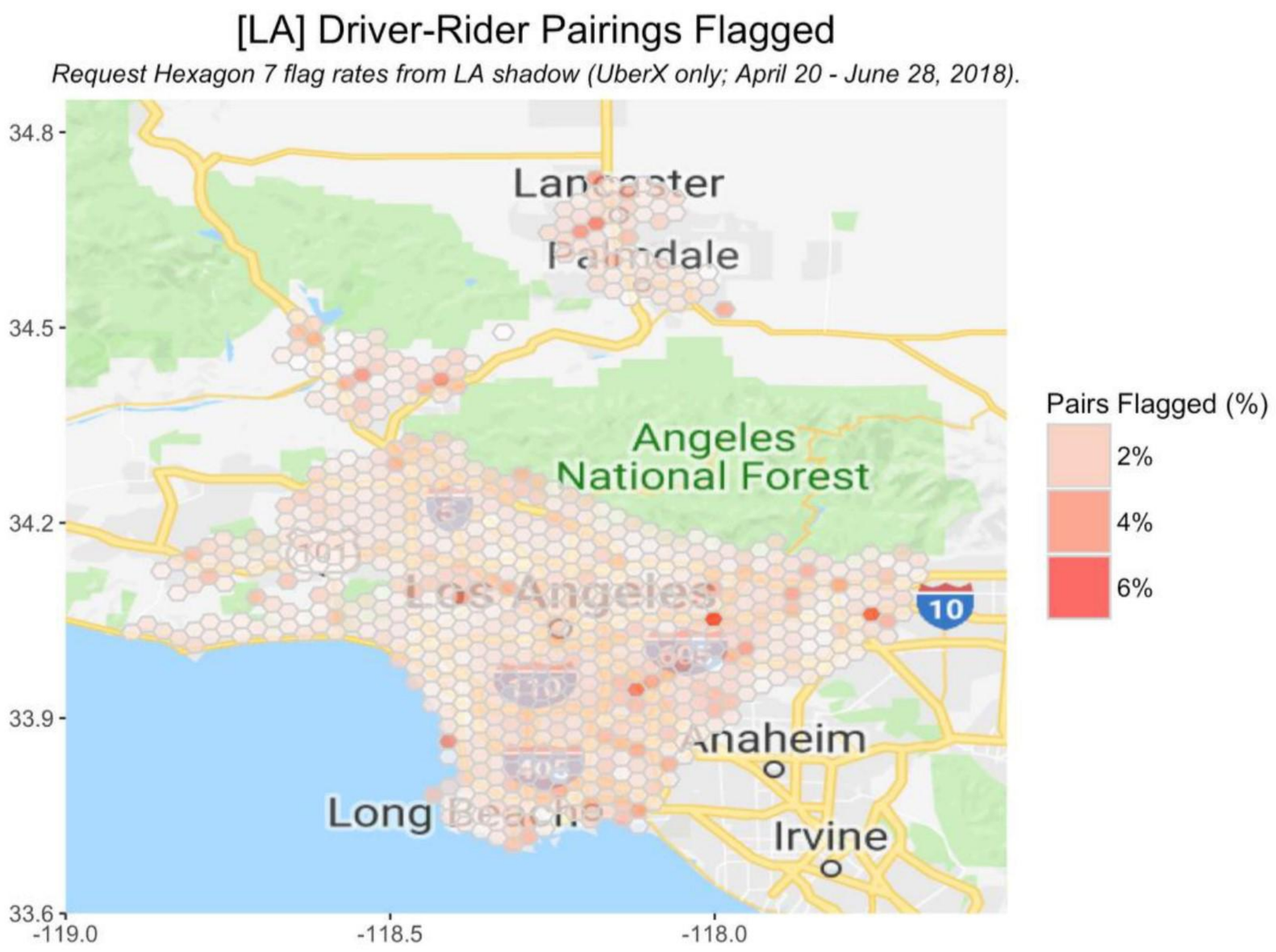


Figure 12



Notes: Only showing hexagons with at least 100 jobs. Models calibrated at the 1% trip trigger rate in aggregate. Combined Gender Model v0.

6. Implementation Plan [back to top]

Because S-RAD is likely to have a meaningful impact on both safety and the marketplace, a rigorous evaluation needs to be conducted in stages. The evaluation must demonstrate that:

- Sexual assaults can be better anticipated (prediction).
- Down-ranking based on model predictions can reduce the incident rate (prevention).
- Down-ranking can be achieved within marketplace constraints (business).
- Down-ranking does not have a disproportionate impact on driver earnings or ATAs by gender (legal).

To answer these questions, the *proposed* evaluation plan consists of a limited pilot, a within-city switchback, and a multi-city experiment (pending discussions with Operations, Marketplace, Legal). The full details of the evaluation plan is available in this document: S-RAD Evaluation Plan. A high-level roadmap is provided in Table 5 below.

Table 5. Proposed 2018 Roadmap (Pending Stakeholder Reviews)

	JUL	AUG	SEP	OCT	NOV	DEC
Shadow Deployments Select US Cities						
Evaluation #1: Blue-Red Experiment						
Review #1 with Stakeholders RGM/GMs, Operations, Safety Leadership, Legal, Marketplace.						
Evaluation #2: Within-City Switchback (Los Angeles)						
Review #2 with Stakeholders RGM/GMs, Operations, Safety Leadership, Legal, Marketplace.						
Evaluation #3: City-Level Randomized Experiment (DID) N US cities (to be selected with ops + marketplace)						
Review #3 with Stakeholders RGM/GMs, Operations, Safety Leadership, Legal, Marketplace.						
Scale (pending evaluation results)						

****Precise timelines depend on experiment results (meet acceptance criteria) and stakeholder reviews.*

7. Team [back to top]

- *Data Science:* Sunny Jeon, Katy McDonald, David Purdy, Frank Chang
- *Engineering:* Peng Sun, Misha Bosin
- *Legal:* Daniel Kolta, Scott Binnings
- *Product:* Akankshu Dhawan
- *Product Operations:* Jose Sandi
- *Safety Operations:* Eric Schroeder

Appendix A: Descriptive Statistics [back to top]

Figure A.1

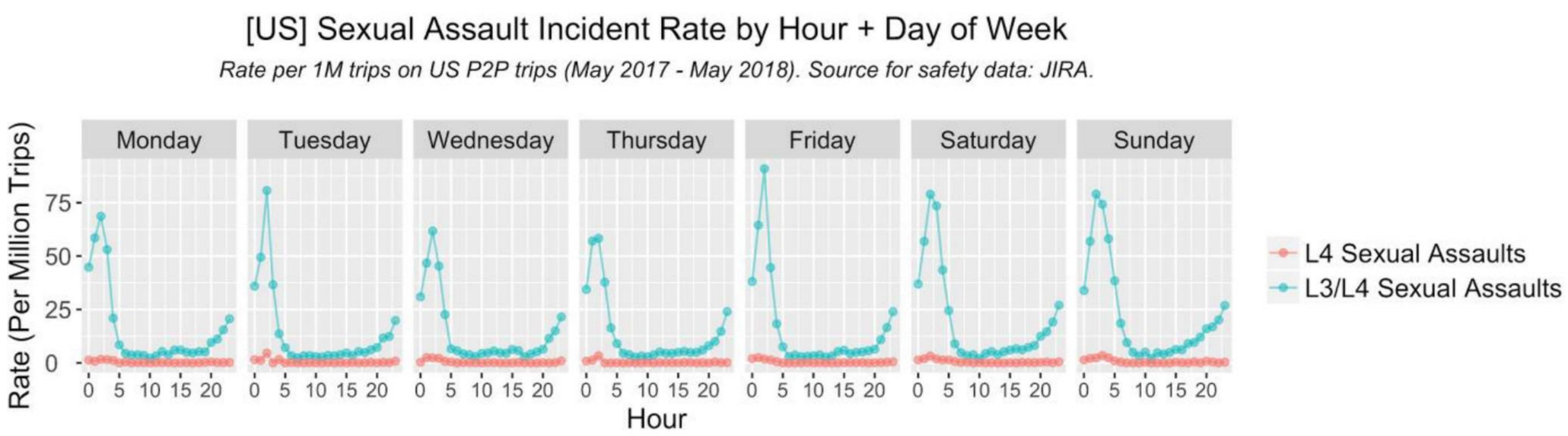
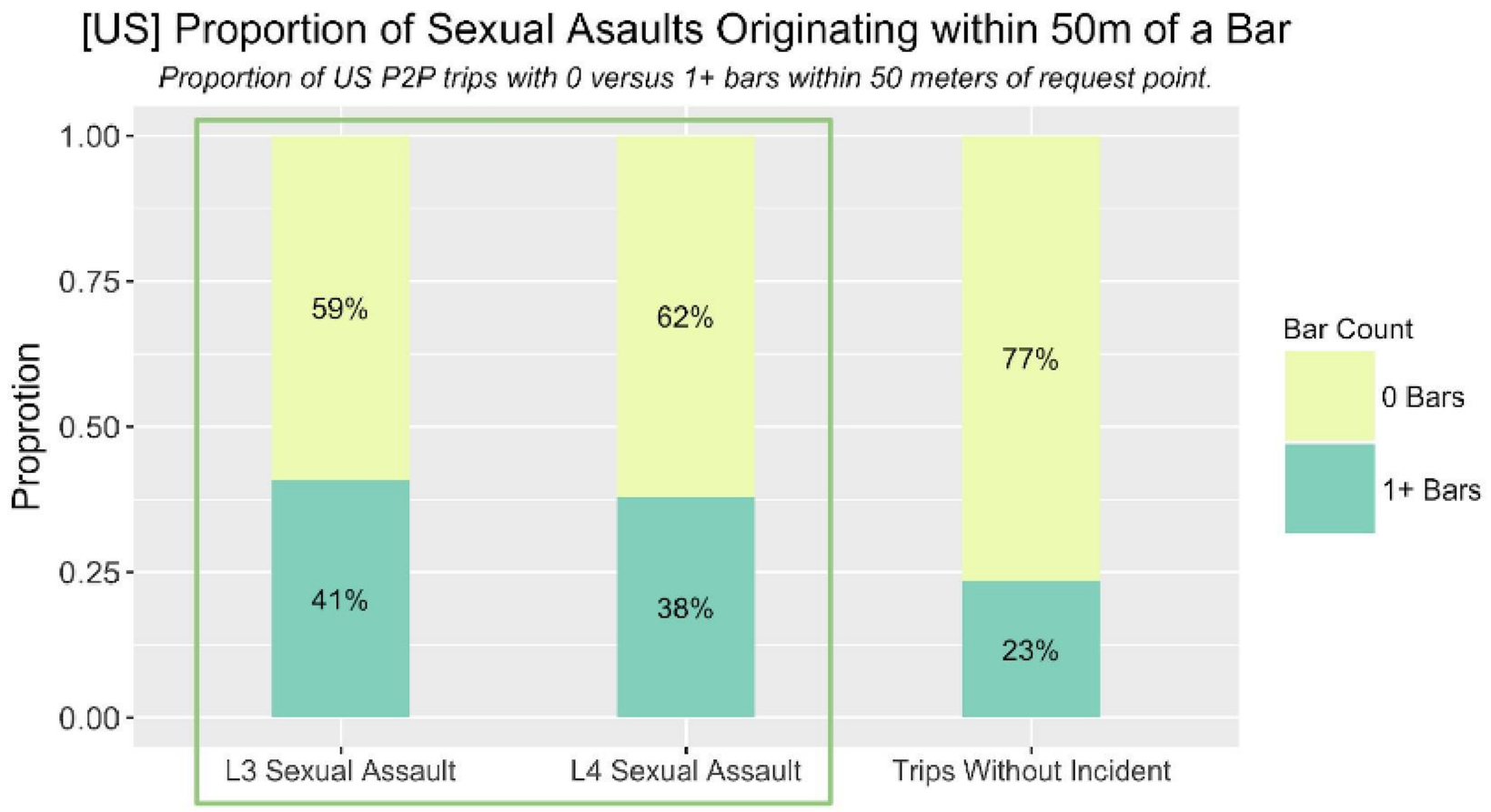


Figure A.2



Notes: 100% sample of sexual assaults (13555 L3 sexual assaults, 382 L4 sexual assaults) and 250k randomly selected trips without sexual assaults (source: JIRA). US P2P occurring March 1 - May 14, 2018.

Figure A.3

Problem for Females **and** Males

Female Users

Top 5 Most Common Safety Incident Types

Incident Category	Proportion Of Cases
1. Accident or Claim	52%
2. Sexual Assault	**22%**
3. Physical Altercation	10%
4. Sexual Misconduct	**6%**
5. Substance Abuse	2%

Male Users

Top 5 Most Common Safety Incident Types

Incident Category	Proportion Of Cases
1. Accident or Claim	50%
2. Physical Altercation	25%
3. Sexual Assault	**12%**
4. Verbal Altercation	3%
5. Health / Self-Harm	3%

Notes: Data divided by ***gender of alleged victim*** (inferred gender based on first name). JIRA data only.
Incidents from US P2P trips only (May 2017 - May 2018). Excluding Law Enforcement/Regulatory.

Figure A.4

Sexual Assault Incident Rate by Gender of Driver + Rider

US P2P incidents occurring May 2017 - May 2018

	Female Drivers		Male Drivers	
	+ Female Rider	+ Male Rider	+ Female Rider	+ Male Rider
Rate of Sexual Assaults (Per 1M Trips)	**8.36**	**38.98**	**13.10**	**7.93**
Num of Sexual Assaults	549	3114	5476	4014
Num of Trips	66M	80M	418M	506M

Notes: Only US P2P trips and incidents occurring May 2017 - May 2018 with inferred gender data (97% coverage in sample). Safety data source: JIRA.

Figure A.5

Gender of Alleged Offender of Sexual Assaults

US P2P incidents occurring May 2017 - May 2018

	Gender of **Alleged Offender**	
	Female	Male
L3 Sexual Assaults (13,342 cases)	14%	**86%**
L4 Sexual Assaults (371 cases)	4%	**96%**

Notes: JIRA sexual assault tickets from US P2P trips occurring May 2017 - May 2018. Includes only sexual assaults with inferred gender data (97% of incidents) and those where the alleged offender is the driver or rider (73% of incidents).

Table A.1: Sexual Assault Incident Rate in 2017 by US Sub-Region

US Sub-Region	Sexual Assault Incident Rate (Per 1M Trips)	Sexual Assault Incident Count	Trip Count
South	**19.60**	1723	87890588
Southwest	**15.24**	2695	176791541
Midwest	**14.30**	1779	124389720
Southeast	**13.43**	3319	247075343
Pacific Northwest	**12.69**	365	28760375
NorCal	**10.17**	944	92785270
Caribbean & Panama	**9.64**	24	2489858
New England	**9.12**	525	57588019
TRIPAD	**8.48**	1761	207577926

Table A.2: Top 10 Cities with Highest Sexual Assault Incident Rate in 2017 (Amongst Large Cities w >20M Trips)

City	Sexual Assault Incident Rate (Per 1M Trips)	Sexual Assault Incident Count	Trip Count
Atlanta	**13.56**	420	30983618
Los Angeles	**13.32**	1251	93932481
Miami	**11.34**	638	56256087
Philadelphia	**9.97**	278	27892521
Chicago	**8.92**	631	70724834
San Francisco	**8.85**	753	85055360
New Jersey	**8.59**	356	41442850
Boston	**7.97**	393	49285208
Washington D.C.	**7.79**	494	63401431
New York City	**6.16**	671	108851902

Table A.3: Top 10 Cities with Highest Sexual Assault Incident Rate in 2017 (Amongst Medium-Sized Cities w <=20M Trips but >1M Trips)

City	Sexual Assault Incident Rate (Per 1M Trips)	Sexual Assault Incident Count	Trip Count
Albuquerque	**45.18**	51	1128833
Pensacola, FL	**41.63**	42	1008853
El Paso	**37.25**	38	1020003
Wilmington, NC	**36.67**	39	1063595
Oklahoma City	**32.39**	54	1667435
Omaha	**32.35**	36	1112946
Fresno	**31.52**	41	1300872
Fort Myers-Naples	**29.03**	54	1859865
Central Atlantic Coast, FL	**28.66**	43	1500494
Kansas City	**28.08**	105	3738970

Table A.4: Top 10 Cities with Highest Sexual Assault Incident Rate in 2017 (Amongst Small Cities w <=1M Trips)

City	Sexual Assault Incident Rate (Per 1M Trips)	Sexual Assault Incident Count	Trip Count
Juneau	**103.52**	1	9660
Jonesboro	**100.21**	1	9979
Bowling Green, KY	**92.68**	7	75531
Topeka	**91.43**	6	65626
Tri-Cities	**80.00**	4	50000
Sioux City	**73.74**	2	27121
Green Bay	**66.42**	28	421590
Stillwater	**64.92**	7	107823
Tri-Cities, MI	**63.19**	3	47476
Huntsville, AL	**61.04**	10	163817

Appendix B: Variable List + Definitions [back to top]

https://docs.google.com/spreadsheets/d/1gXFrZubkGWme3d6xjgT6sScaV_bBARgt0aftdGsq3xw/edit#gid=0

Appendix C: Evaluation Plan [back to top]

https://docs.google.com/document/d/1whZdg5k4GY2q1IZHJJKUMoirEybB_jFdrCdpjf3RbV0/edit#heading=h.sk3a5ndlu6ae