Kwantitatieve validatie voor pilot Alpha

Directie Wonen
Werkplaats DAO

Huidige voorstel in pilotplan Alpha

"Bij zaken met een risico indicatie voor illegaal vakantieverhuur van 50% of hoger, wordt op het moment dat het stadium "rapportage naar handhaving (RAP NAAR HAN)" wordt opgevoerd in BWV, beschouwd als een bevestiging dat het algoritme Alpha een juiste risico indicatie heeft gegeven"

Opmerkingen

1. Deze meting is niet heel sterk, omdat het model dan bijvoorbeeld áltijd een uitkomst van 55% of hoger kan geven en dan dus altijd succesvol is.
2. Er kan geen vergelijking met bijvoorbeeld het jaar 2019 plaatsvinden doordat de verhuursituatie in Amsterdam door corona geheel veranderd is.

2

Dia 2

Vd1 Ik denk dat het de leesbaarheid van dit document ten goede komt als we een introductieslide maken wordt uitgelegd wat een kwantitatieve validatie is en waarom we dit doen.
[redacted]; 6-10-2021

Nieuwe voorstel vanuit de data science voor pilot Alpha

Zaken waar het stadium "rapportage naar handhaving (RAP NAAR HAN)" opgevoerd wordt in BWV worden gecategoriseerd als "zaken met woonfraude". Zaken die wel onderzocht worden, maar zonder het stadium "rapportage naar handhaving" afgesloten worden, worden beschouwd als "zaken zonder woonfraude". Zaken die niet zijn onderzocht worden voor het succescriteria buiten beschouwing gelaten.

We gaan dus een nulmeting doen en een meting tijdens/na de pilot op 2 kengetallen.

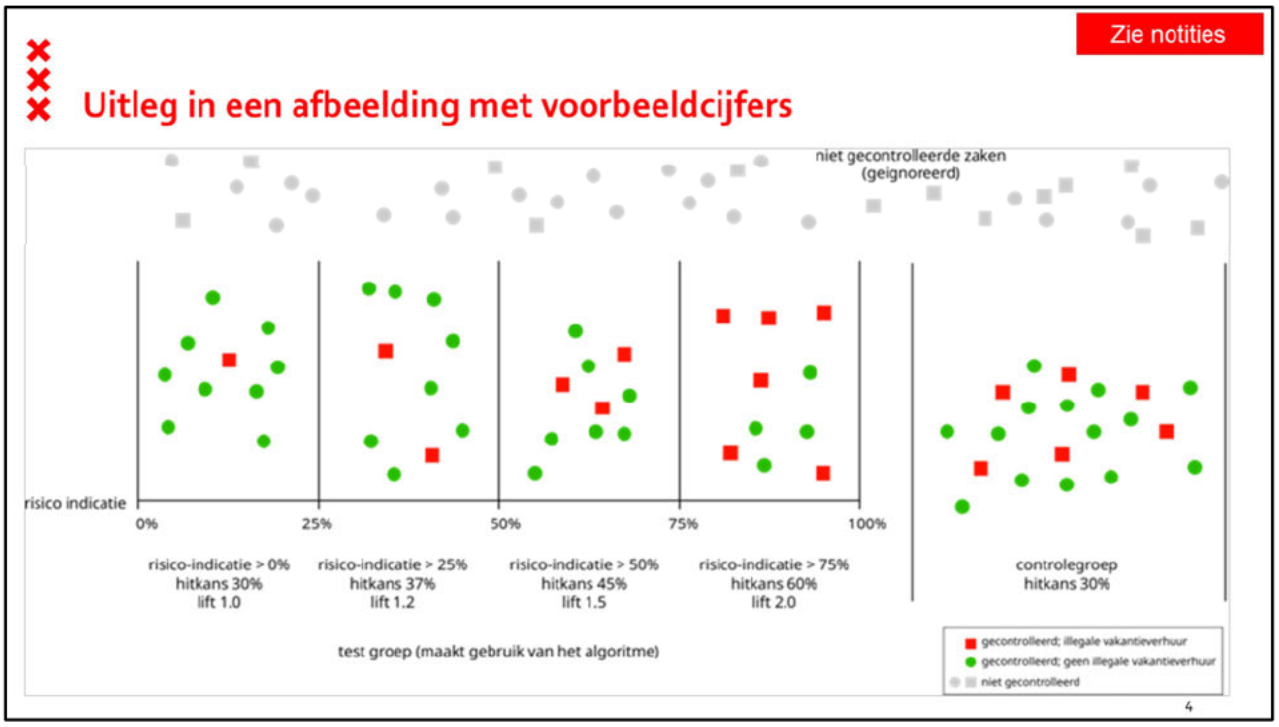
1. Zaken met woonfraude = alle zaken die als RAP naar HAN worden opgevoerd in BWV
2. Zaken zonder woonfraude = alle zaken die wel onderzocht zijn, maar zonder stadium RAP naar HAN

De verwachting is dat het relatieve aandeel van "zaken met woonfraude" onder de onderzochte zaken hoger wordt naarmate het model een hoge hitkansvoorspelling geeft. Bijvoorbeeld: model geeft bij 10 signalen een percentage van 80% en bij 10 signalen een percentage van 60%. Dan is de verwachting dat het er bij de gegeven percentage van 80% vaker "zaken met woonfraude" zijn opgevoerd dan bij de signalen waarbij 60% kans was gegeven.

Om de verwachting te controleren kan er met een controlegroep getest worden. In de controlegroep zal het algoritme niet gebruikt worden om zaken te kiezen, en de handhavers hebben tijdens de controle geen inzicht in de risico-indicatie of de verklaringen van het algoritme. Dit kan bijvoorbeeld ook door het algoritme twee weken wel te gebruiken en dan weer twee weken niet etc. Er moet worden verkend of dit technisch mogelijk is in de TOP applicatie.

3

Nieuwe zaaksysteem: zit daar wel Rap naar Han in? Hoe borgen we dezelfde meting als het zaaksysteem verandert?
Nulmeting: waarop doen we een nulmeting? Is het nodig? Als we zonder controlegroep werken is het wel goed een nulmeting te doen, zodat we de hitkans kunnen vergelijken.
A/B: opletten met teams vergelijken, sommige teams presteren beter dan andere. De tijdsdeling A/B test heeft wel praktische beperkingen: de toezichthouders moeten dan steeds switchen.
Belangrijk: waar wordt de risicoindicatie per zaak opgeslagen? Dit is essentieel voor de meting. > Dit wordt opgeslagen in the TOP applicatie



Als bijvoorbeeld de gemiddelde hitkans onder alle onderzochte zaken 10% is, maar de hitkans voor onderzochte zaken met een risico-indicatie van groter dan 70% bij 20% ligt, is de *lift* 20%/10% = 2.0. Met andere worden is de hitkans bij de controle van zaken met een risico-indicatie van 70% twee keer zo hoog als bij de controle van een toevallig gekozene zaak -- onder de veronderstelling dat het model *niet* gebruikt wordt om de zaken te kiezen en de looprouten te plannen.

De lift wordt berekend voor risico-indicaties van 10% tot 90% in stappen van 10%. Deze lifts zalen geanalyseerd worden, bijvoorbeeld of eer voor een bepaalde risico-indicatie-waarde een voldoende hoge lift bestaat *en* het een voldoende aantal zaken met een hogere risico-indicatie bestaan.

In addition to Lift analysis: A/B testing

1. Based on team
    1. Disadvantage: there is a lot of difference between prestations of teams
2. Based on period
    1. Disadvantage: Seasons
    2. Disadvantage: covid changed the periods a lot, not really ideal for comparison
    3. Disadvanatge: practical constraints: alternate weeks, this is not usability friendly for users
        1. > we should ask [redacted]

3. **>>>> Based on the technichal side**
    1. TOP application asks for all the cases the percentage, but show for some of the cases (randomly assigned)
       To do TOP: Show only for randomly assigned outcomes of the model, but store all outcomes, store which was showed and which not, front-end show versus don't show
    2. Disadvantage: What if a user sees some percentages and some not: you might focus on the adresses with percentages instead of all
    3. Possible solution: fix this in the work instructions: everyone has to pick 50% with results and 50% without.
    4. Hoe controleren we of er daadwerkelijk 50-50 is aangehouden? > dit moet gedurende de pilot worden gevolgd door de teamleiders

Question: hoeveel zaken hebben we nodig voor een goede validatie ? Wanneer start de pilot, bij 50% drukte van 2019?
> We hebben in totaal 700 zaken nodig. We kunnen de pilot 6 maanden kunnen laten duren. Hoeveel worden er momenteel gelopen per week door vv?

Je kunt het percentage van een op de looplijst geplaatste adres zonder percentage weer inzien ALS dit adres is bezocht. De adressen die worden bezocht worden opgeslagen in de met hun op dat moment bijbehorende resultaat vanuit het model.

Het zou in de "looppijstitem" geregisteerd kunnen worden dat 50%50% moet worden getoond en de andere niet. Voorwaarde: er moet[5]

Technichal refinement: requirements for AB testing

1. Storing of the results and outcomes of the model
    1. We need to know if a visited adress was fraud or not fraud
    2. We need to know if a visited adress was shown with or without the outcome of the model
    3. We need to know the score predicted by the model (independently if it was showed or not)

2. Showing results in TOP
    1. The A group and the B group should be randomly assigned and +- even size
    2. If a case was signed towards the A group, it should stay in the A group (same goes for the B group)
    3. The A Group (adresses with result of the model) should be adresses with different scores, high and low, randomly assigned

    Estimated time for TOP developers to implement this: 3 days in total by [redacted]

    •Note: the data of these results are in the TOP back-end. We need to check how to receive an export from TOP.

6

Vervolgstappen

1. Dit plan vertalen naar het pilotplan
2. Ammar de nulmeting en analyse uitvoeren met ondersteuning van [redacted]
3. AB test technisch realiseren tussen model en TOP: [redacted] ontwikkelaar TOP
    1. Hoe en waar worden de zaken in TOP opgeslagen om te kunnen analyseren of de zaken met of zonder percentage het beter doen?
4. AB instrueren: werkinstructies maken voor de toezichthouders om met de AB test te werken [redacted]
5. AB monitoren: Zorgen voor een overzicht van de 50-50 verhoudingen tussen belopen adressen met percentage en zonder (dashboard voor AB testing) (data analisten Wonen)

7