Notes in Statistical Learning

Viktor Zetterberg

2019/2020

Statistical learning

Background

These are my notes in the course Statistical learning given by Stanford university.

Chapter 2 - Statistical learning

Notation: $Y = f(X) + \epsilon$

Understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y.

What is a good value for $f(X)$?

$$f(x) = \mathrm{E}[Y|X = x] \tag{1.1}$$

(regression function)

Ideal/optimal predictor of Y.
$f(x) = \mathrm{E}[Y|X = x]$ is function that minimizes MSE

$$\mathrm{E}[(Y - g(X))^2|X = x],$$

over all functions g at all points $X = x$.

$\epsilon = Y - f(X)$, irreducible error

$$\mathrm{E}[(Y - \hat{f}(X))^2|X = x] = \underbrace{\left(f(x) - \hat{f}(x)\right)^2}_{\text{reducible}} - \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}} \tag{1.2}$$

Nearest neighbor good for small p, $p < 4$, and large N.
NN lousy for large p due to dimensionality.
We deal with dimensionality through structured models.

Parametric & structured models

Linear model is parametric model:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{1.3}$$

Estimate parameters by fitting model to training data.
Almost never correct but good approximation.
Splines is a sort of smoothing.
When tuning a spline so that there are no errors ϵ on training data the model is overfitted. We are overfitting the training data.

Trade-offs

Prediction accuracy vs. interpretability
- Linear models easy to interpret, thin-plate splines are not
Good fit vs. over-fit or under-fit
- When is the fit right?
Parsimony vs. black-box
- Less or more

Chapter 3 - Linear regression

Chapter 4 - Classification

Chapter 5 - Resampling

Figures

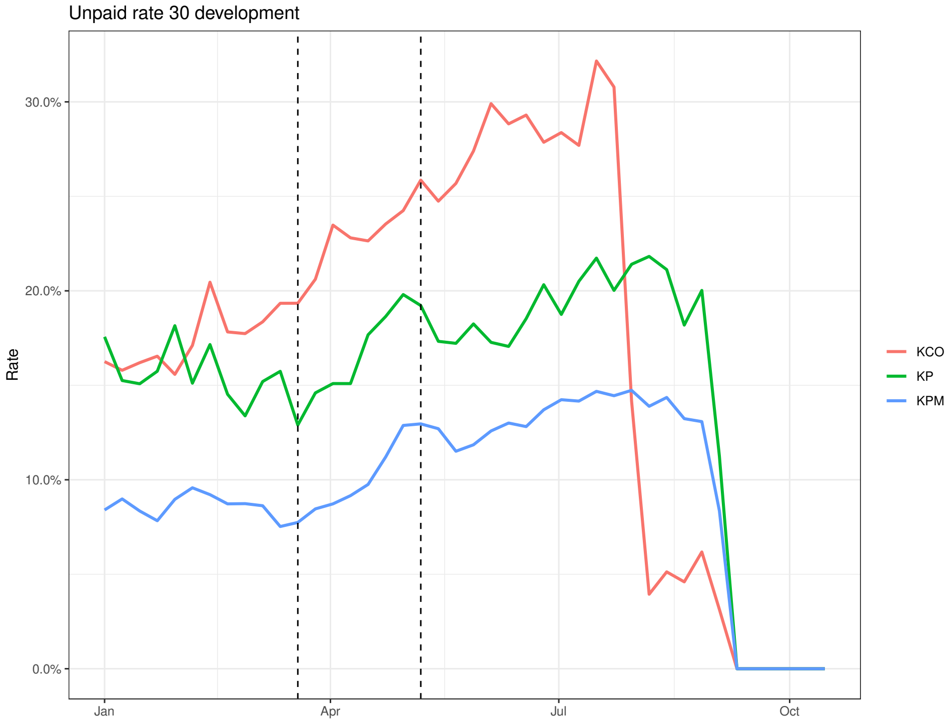

Fig. A.1. Unpaid rate 30 for the different acquiring sources. The increase is present within all acquiring sources but is more profound for KCO.

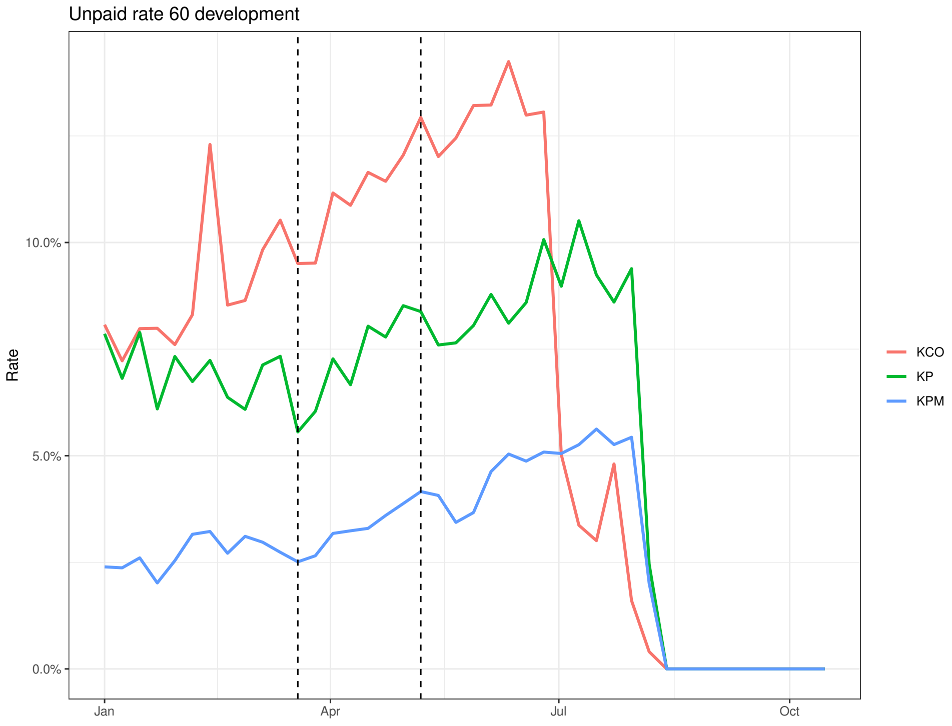

Fig. A.2. Unpaid rate 60 for the different acquiring sources. Same conclusions as for the unpaid rate 30 development.

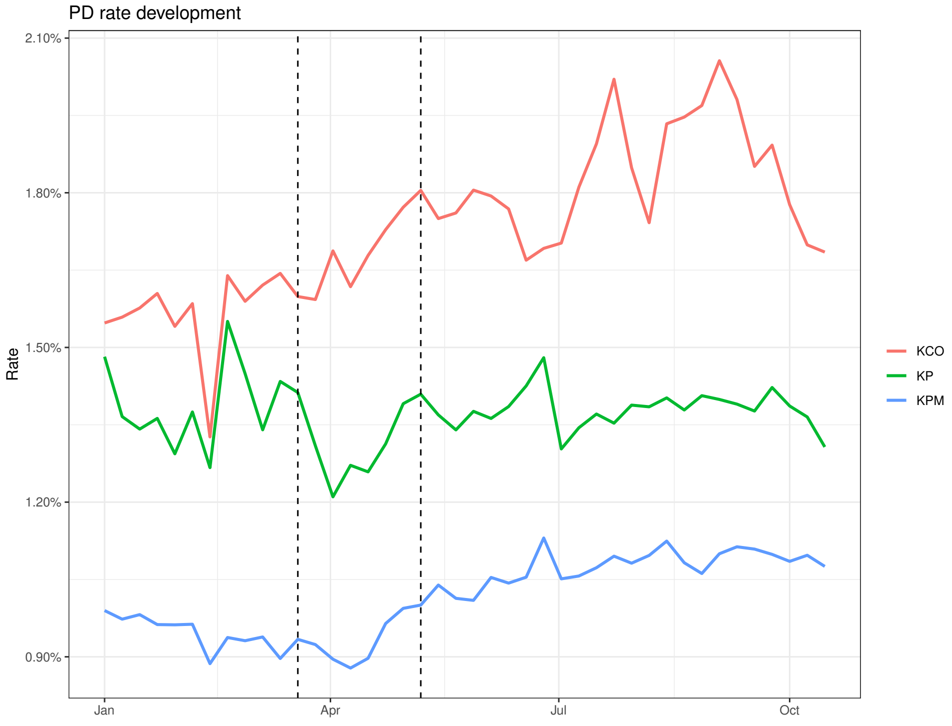

Fig. A.3. PD development. The increase is mainly present for KCO and KPM. KP seems to have a stable PD rate but the volume for KP is quite small compared to KPM and KCO.

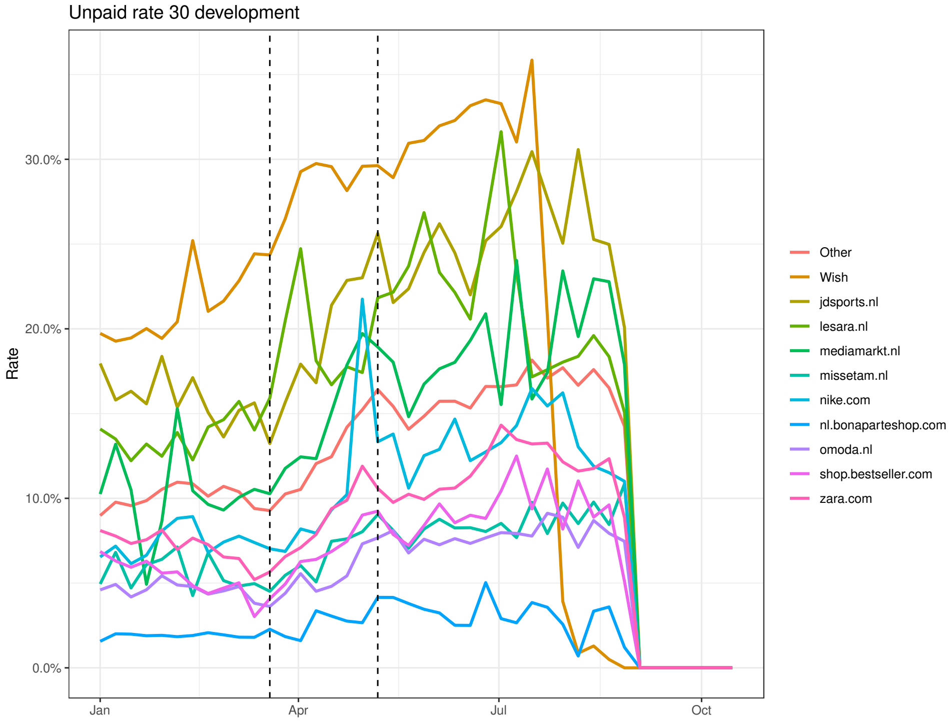

Fig. A.4. Unpaid rate 30 for merchants. It is evident that the increase is present within more or less all merchants. Hence it is not possible to draw any conclusions about the increase just looking at specific merchants/groups/segments.

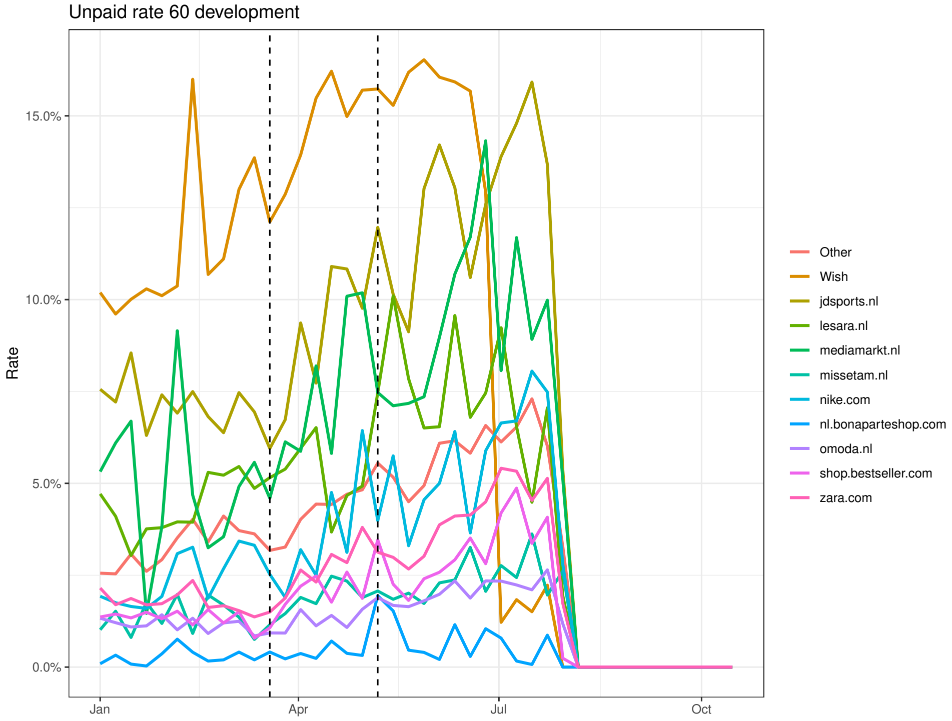

Fig. A.5. Unpaid rate 60 for merchants. Similar development as for unpaid rates 30. Same reasoning. The increase is more present for some merchants than for others but on a whole the increase is within all merchants/groups/segments.

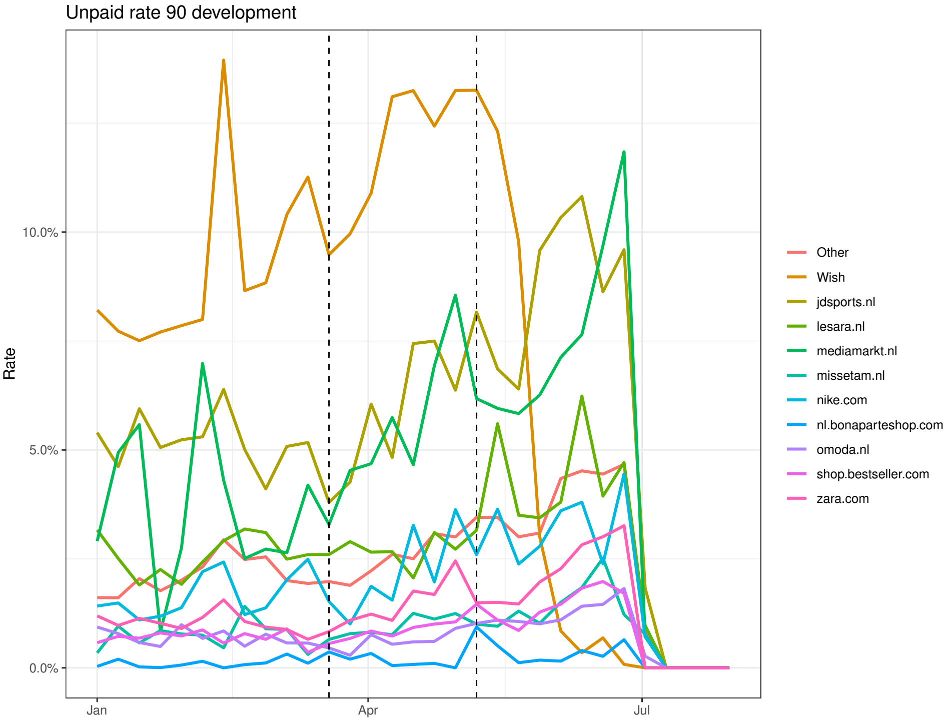

Fig. A.6. Unpaid rate 90 for merchants. Similar development as for unpaid rates 30 and 60. Same reasoning. The increase is also here present within (more or less) all merchants/groups/segments.

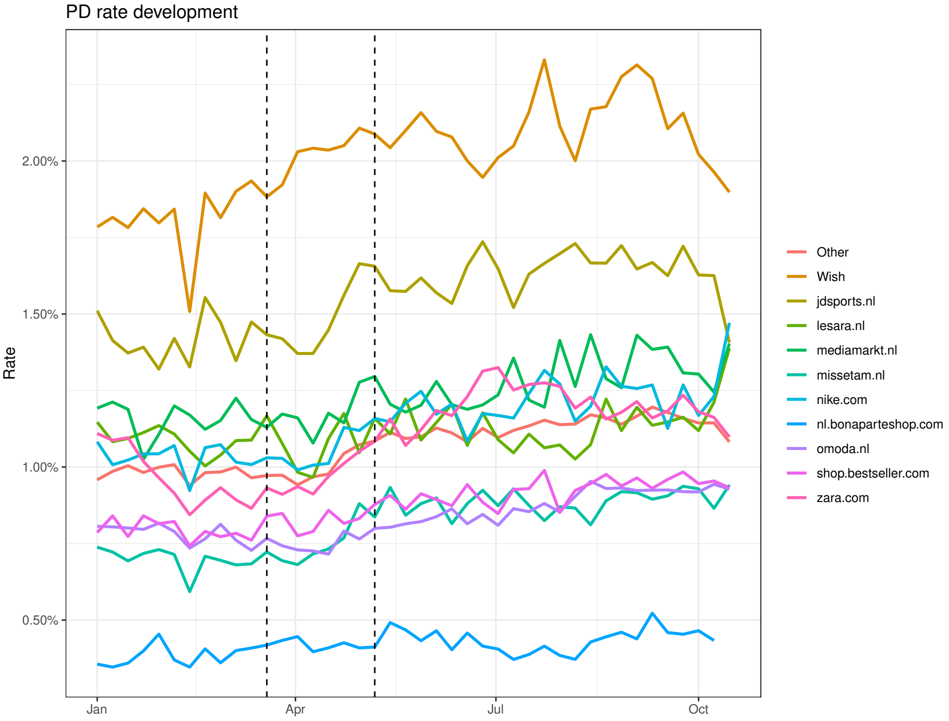

Fig. A.7. PD rate for merchants. Even though not at evident as for the other unpaid rates, the PD rates are increasing as well. This suggest that the population entering the market entail a higher risk than previous.

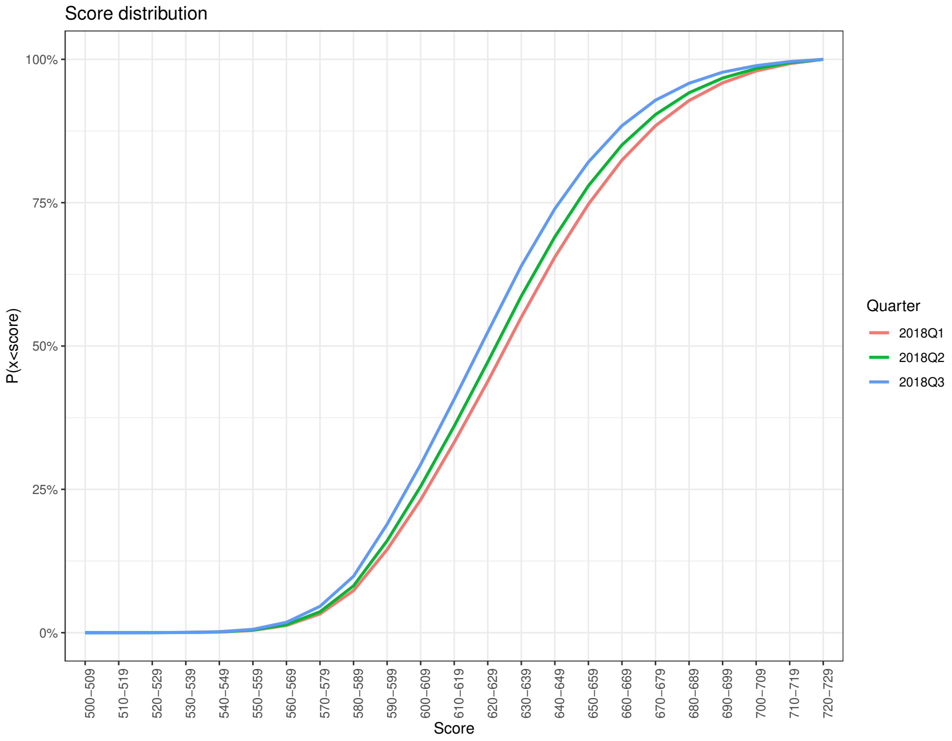

Fig. A.8. Score distribution each quarter of 2018. It is present that the score distribution is successively skewed to the left each quarter, this implies that the population is entailing higher risk and is a driver of the increase.

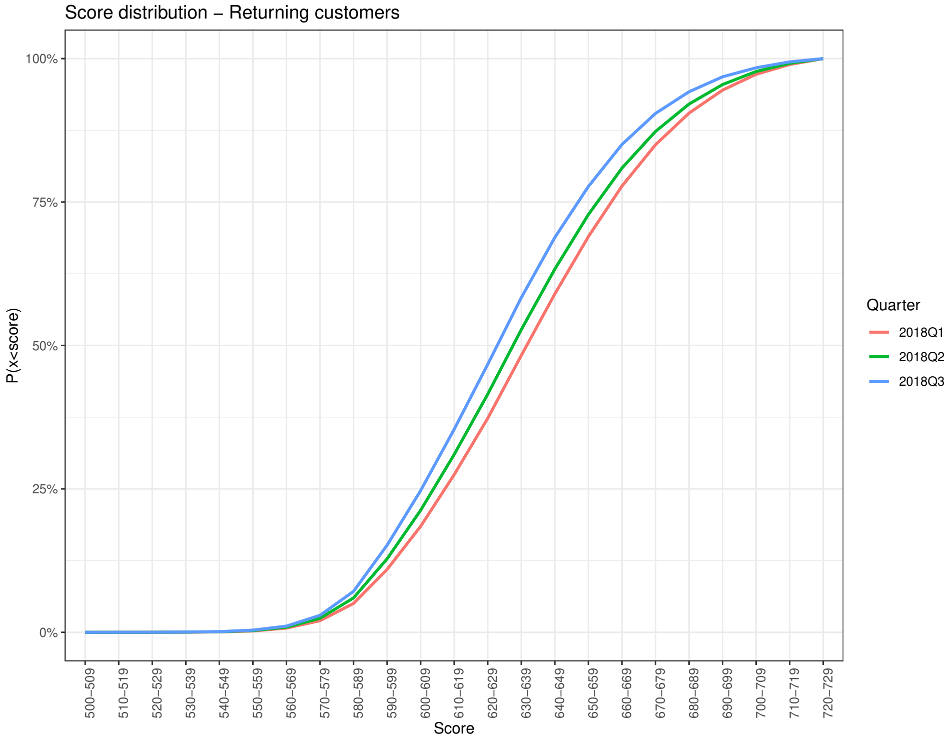

Fig. A.9. Score distribution each quarter of 2018 for returning customers. The score distribution for returning customers shows the largest change during the period. It is evident that the distribution is significantly different fore each quarter.

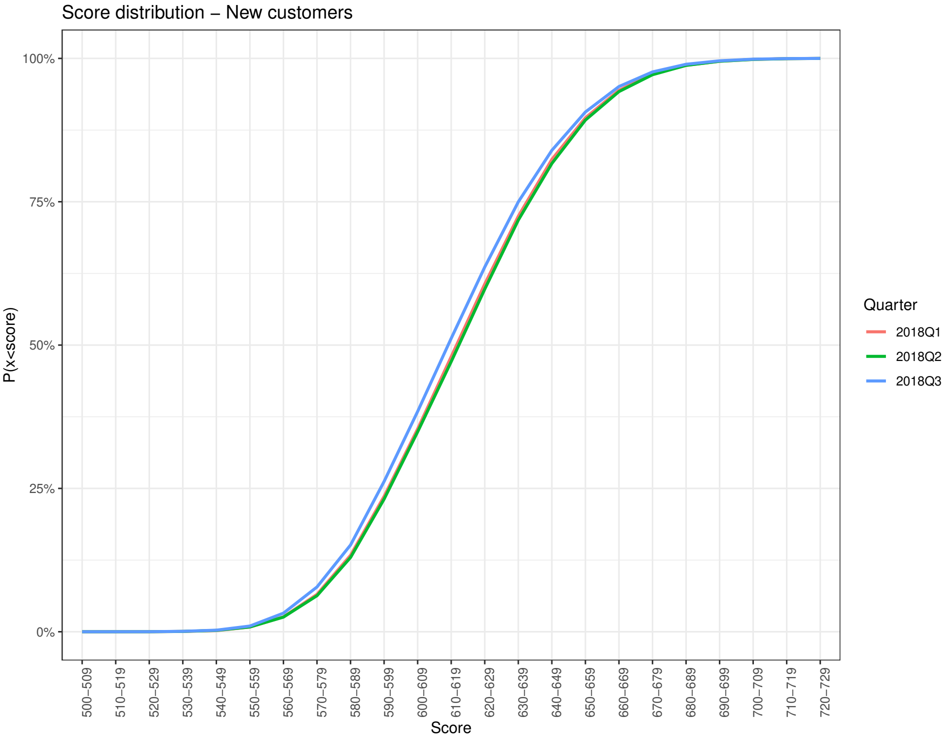

Fig. A.10. Score distribution each quarter of 2018 for new customers. It seems to be the case that the score distribution for new customers have changed just a little during the period.

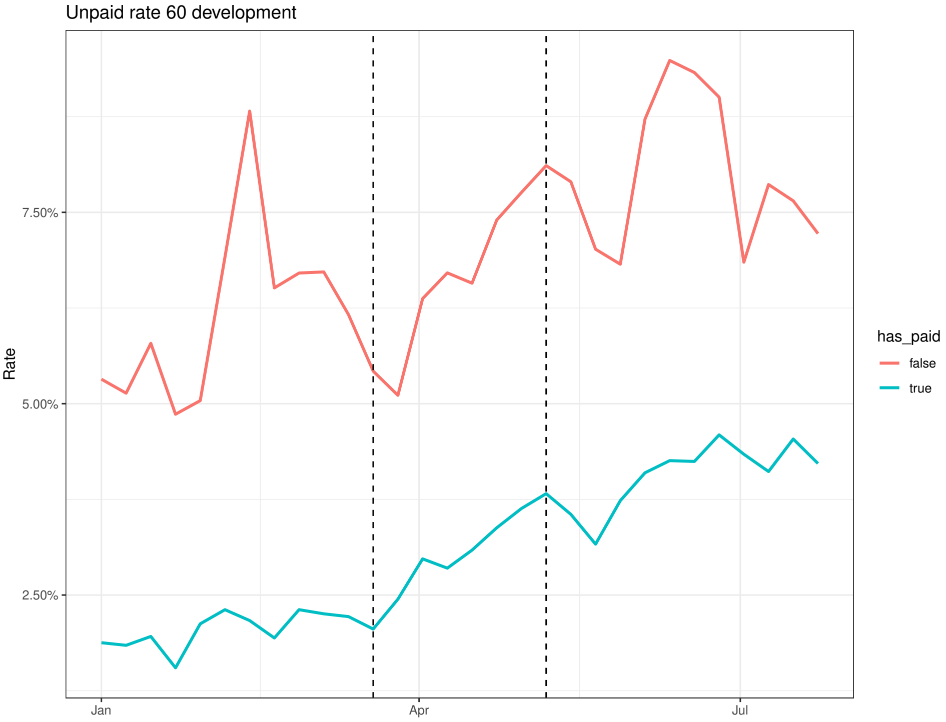

Fig. A.11. Unpaid rate 60 for new and returning customers.

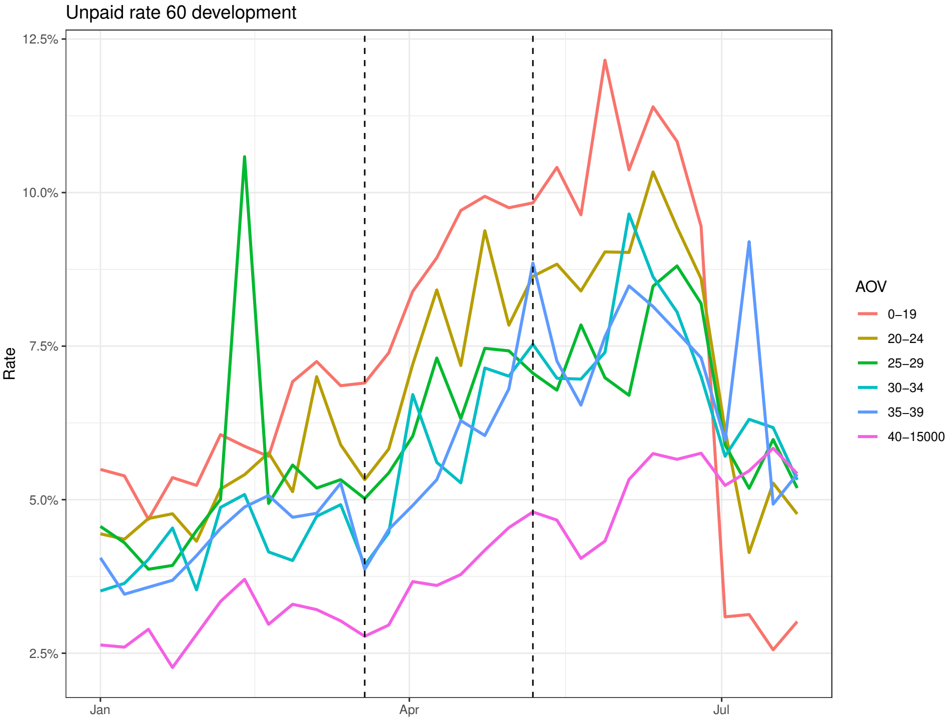

Fig. A.12. Unpaid rate 60 for different AOV bins.

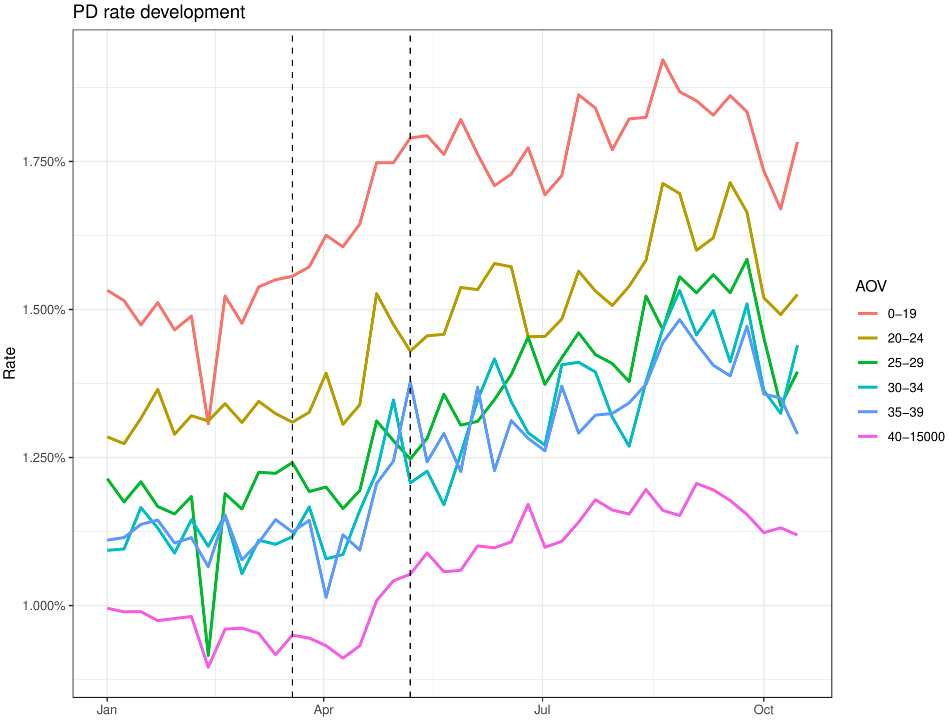

Fig. A.13. PD rate for different AOV bins.

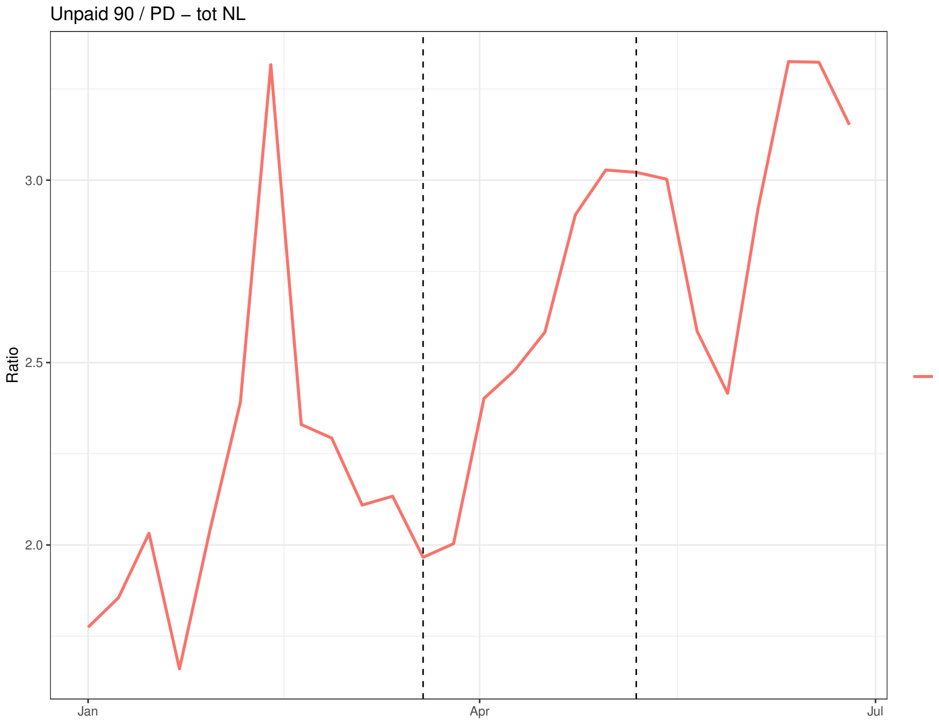

Fig. A.14. Unpaid rate 90 vs. PD rate. The ratio is increasing during the period. Since we know that PD is increasing the conclusion is that unpaid rate 90 is increasing at a faster rate than the PD rate. This is not surprising since from looking into the score development we saw that PD is growing but it did not explain the total increase.

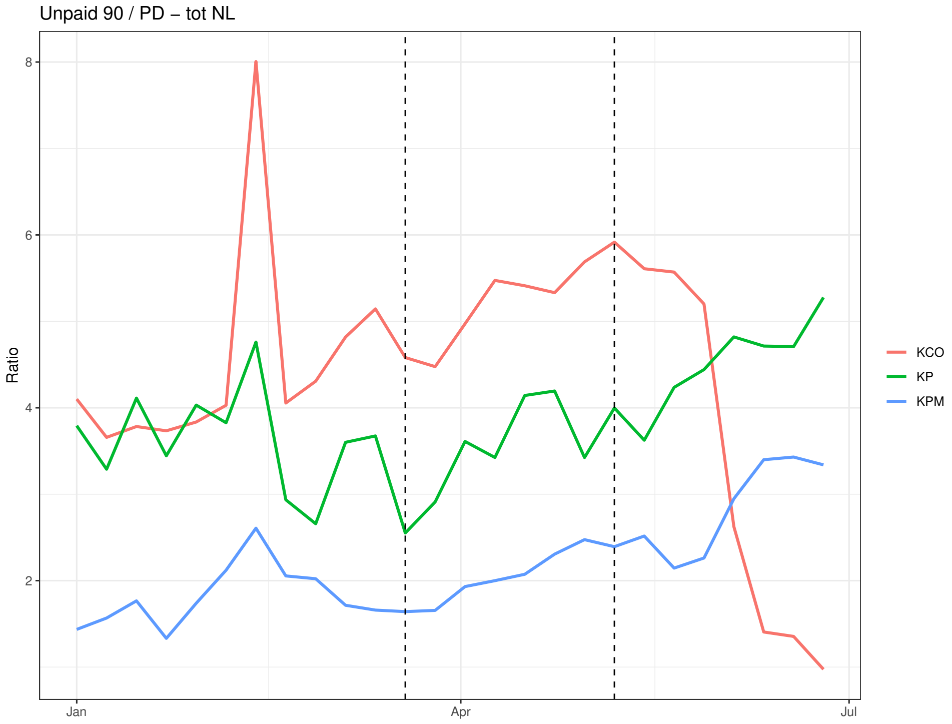

Fig. A.15. Unpaid rate 90 vs. PD rate for the different acquiring sources.

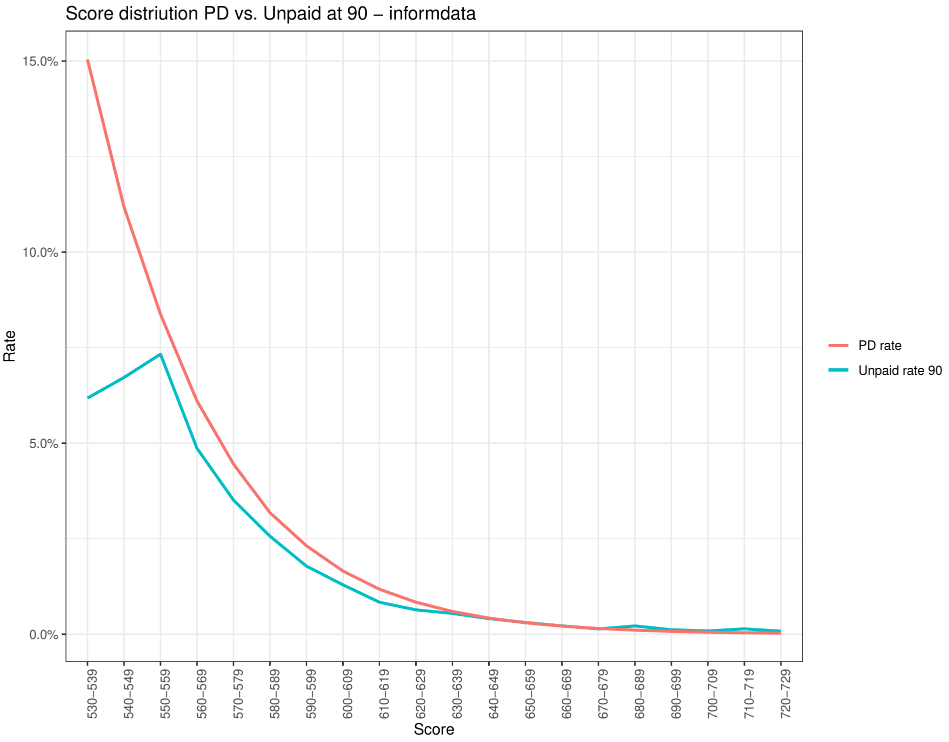

Fig. A.16. PD rate vs. unpaid at 90 rate for scorecard informdata. Largest scorecard.

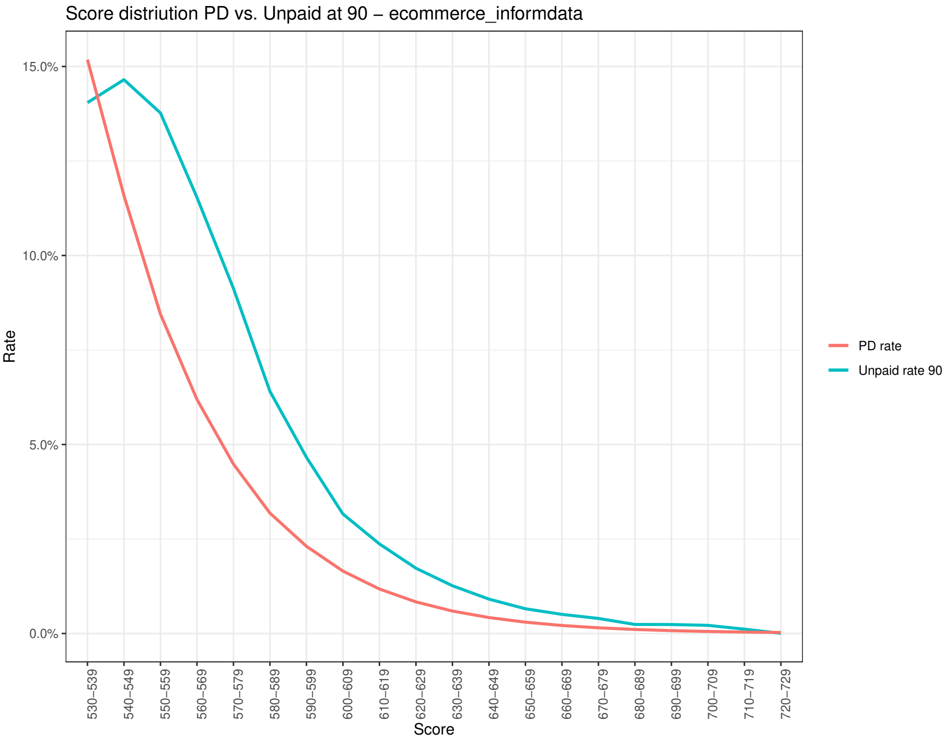

Fig. A.17. PD rate vs. unpaid at 90 rate for scorecard ecomerce_informdata. Second largest scorecard.