Picnic: where to deliver next?

Data Science Capstone Assignment **11-10-2020**

Luc Keizers

In this presentation

1) Business case
2) Data
3) Methodology
4) Results
5) Discussion
6) Conclusion

1)Business case

- Online grocery shopping becomes more popular over years

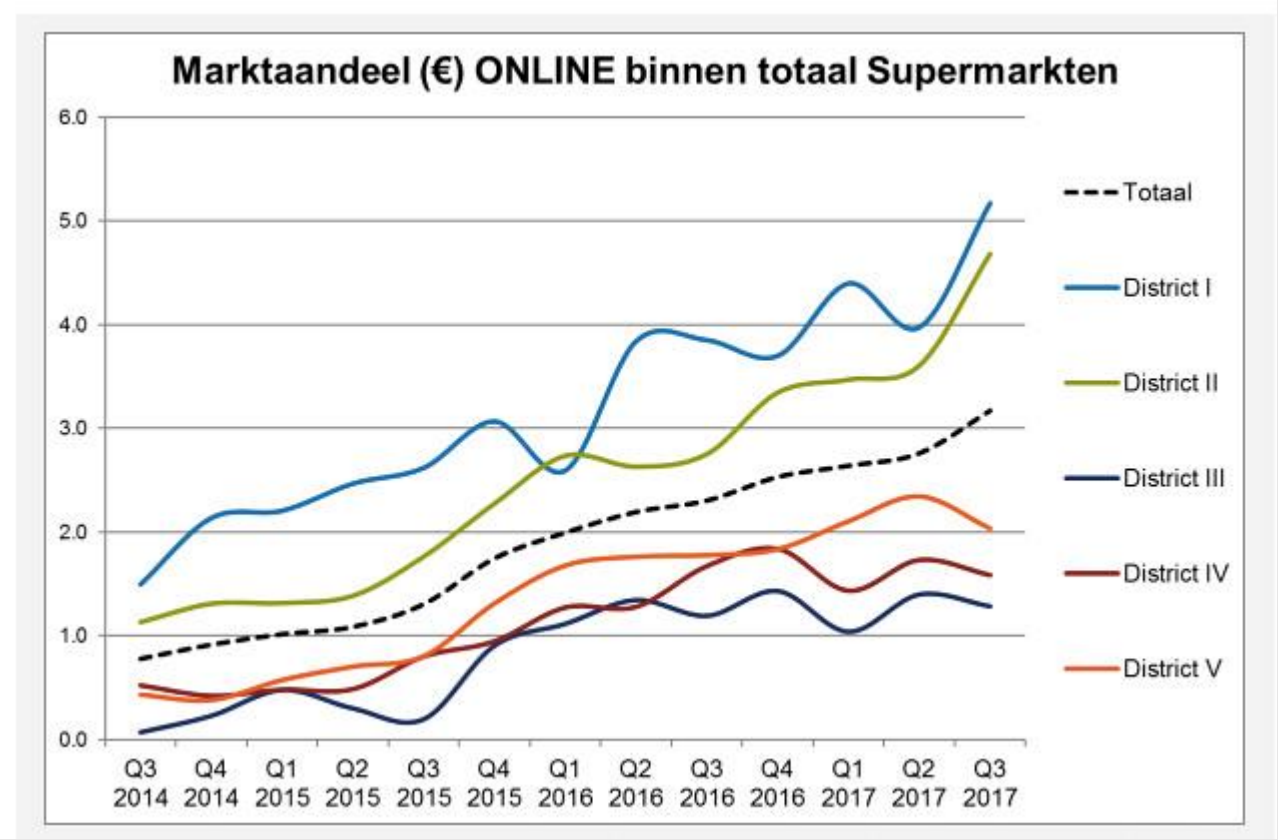

Fig. 1 Increase of market share in online grocery shopping in five different districts in the Netherlands *[1]*

1) Business case

- Picnic is the first online-only grocery store
- Delivery using electric vehicles: not only more easy, but also more eco-friendly!
- Already in >70 places!

- Keep expanding: but where to start delivering next?

Fig. 2: Areas where Picnic is available *[2]*

2) Data

- Features to include in deciding area of interest:
- Many people -> Much profit
- Rich people -> Much profit
- Few competition -> Much profit

2) Data

- CBS database *[3]:* Municipality population, house prices
- FourSquare API: Grocery stores in municipalities

	Mun_code	Population	Mun_name	Count	Average House Price	Population per store
0	GM0363	872445.0	Amsterdam	3.0	485.0	0.000003
1	GM0599	651524.0	Rotterdam	1.0	282.9	0.000002
2	GM0518	548165.0	Den Haag	3.0	324.8	0.000005
3	GM0344	359374.0	Utrecht	4.0	379.8	0.000011
4	GM0772	235465.0	Eindhoven	2.0	308.8	0.000008
5	GM0014	233266.0	Groningen	2.0	258.9	0.000009
6	GM0855	220589.0	Tilburg	1.0	265.3	0.000005
7	GM0034	214133.0	Almere	2.0	294.3	0.000009
8	GM0758	184306.0	Breda	0.0	350.0	0.000000
9	GM0268	177525.0	Nijmegen	0.0	293.8	0.000000

Table 1: data for analysis

[3] http://cbs.nl/

3) Methodology

1) Logistic regression to see if model is able to predict in which areas Picnic is already active (binary)
2) Logistic regression to determine probability that Picnic is active in area: if high probability and Picnic not active, good potential

4) Results

4.1) binary

- Jaccard score: 0.58
- Confusion matrix

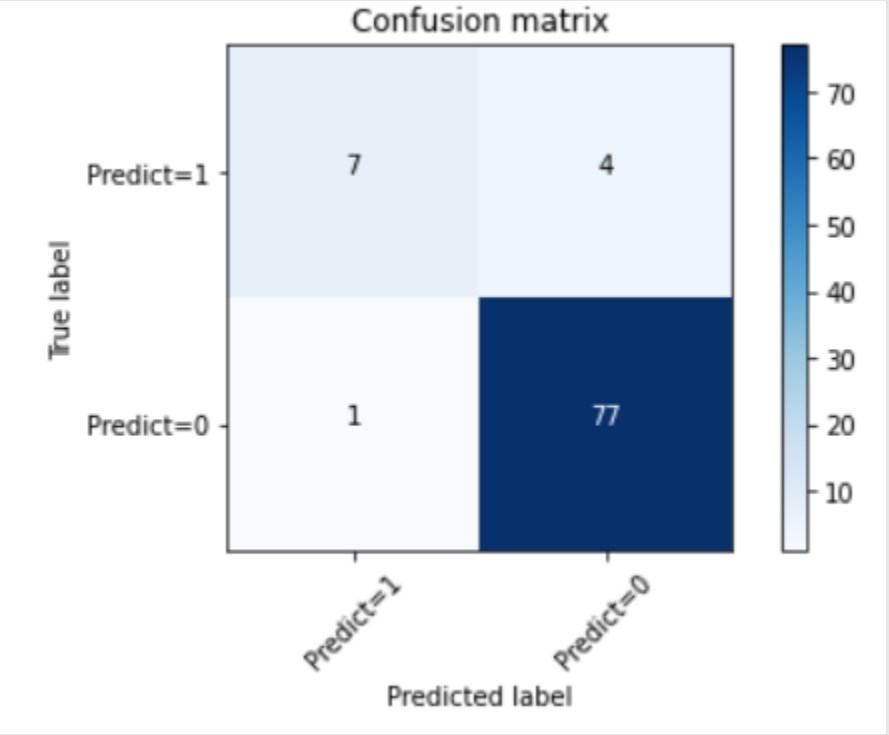

Fig. 3: Confusion matrix of logistic regression model

- Not very well..

4) Results

4.2) Probability

- Green dots:

  Picnic active
- Red dots:
  Not active yet

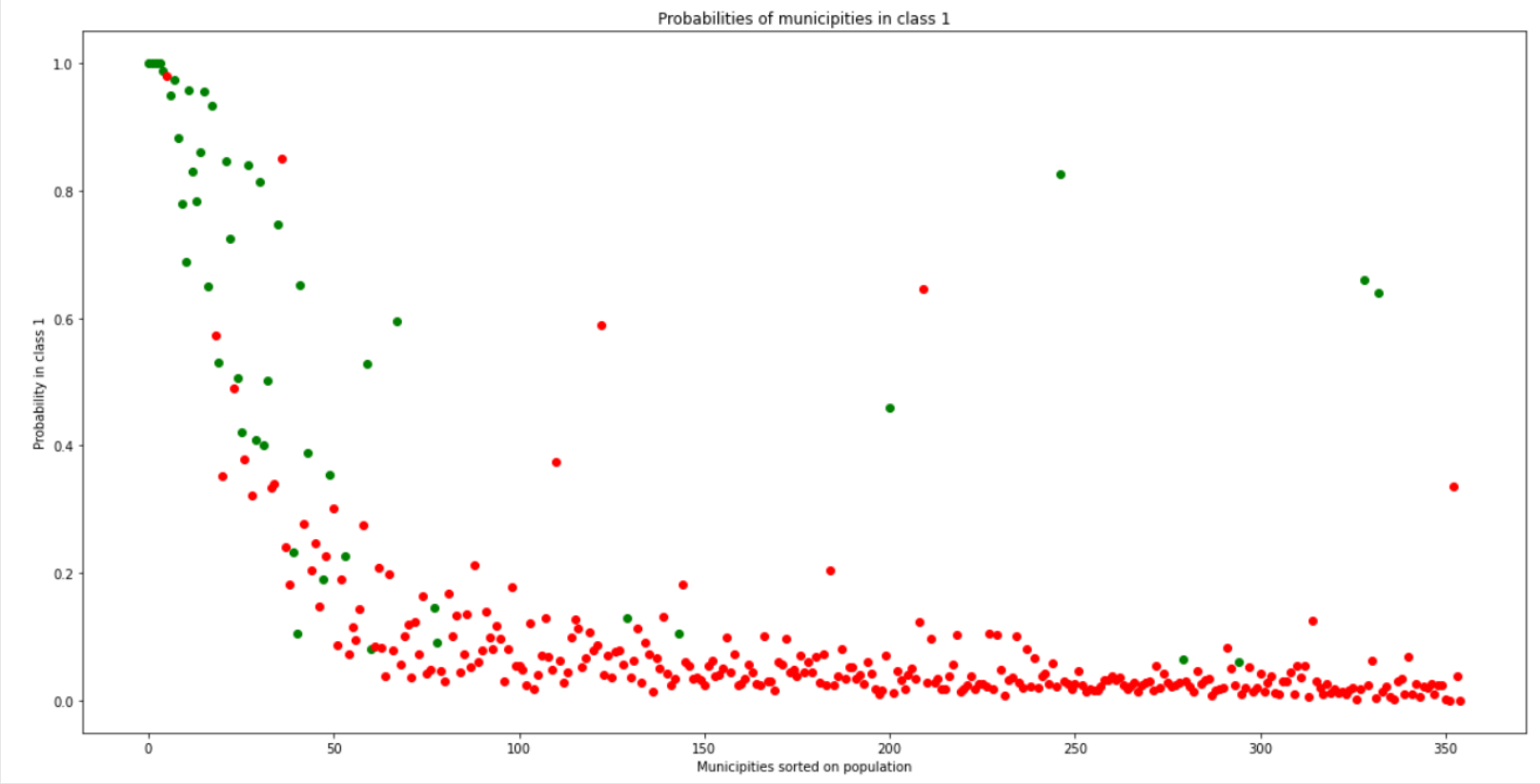

Fig 4: Probabilities of each municipality to Picnic activity

- High probabilities are interesting, lower not. Why are some green dots low and some red dots high?

4) Results

- Green dots: Picnic active, high probability value
- Red dots: Picnic active, low probability value

- Seems like distance to other active areas should also be included as a feature.

4) Results

- Based on this model, the top 5 to start delivering next is as follows:

	index	Mun_code	Population	Mun_name	Count	Average House Price	Population per store	Existing	Predicted proba
0	5	GM0014	233266.0	Groningen	2.0	258.9	0.000009	0	9.791669e-01
1	36	GM0362	91061.0	Amstelveen	2.0	525.2	0.000022	0	8.498615e-01
2	209	GM0629	26681.0	Wassenaar	1.0	711.3	0.000037	0	6.450573e-01
3	122	GM0310	43328.0	De Bilt	5.0	481.4	0.000115	0	5.895832e-01
4	18	GM0193	129465.0	Zwolle	1.0	293.1	0.000008	0	5.723832e-01

Discussion

- Look at city-level instead of municipality level
- Add relevant features, improve current features
- Cluster municipalities/cities that are close to each other
- If profit per active area is known, use this in a linear regression model!

Conclusion

- Data analysis can be a powerful tool to determine where to deliver next
- Model is poor and needs improvement
- Linear regression would be more useful than logistic regression, but requires additional profit data