Picnic: where to deliver next?

L.S. Keizers

November 10, 2020

1 Business case

Picnic is a Dutch online supermarket. Ordering groceries online has become more popular over the past years: nowadays the amount of people doing online groceries has doubled compared to 2016 [1]. A company anticipating on this trend is Picnic: the first supermarket that has no physical store anymore, but only distribution centers. For the company this is convenient because it requires e.g. less rent and less staff to pay. The groceries are delivered by small electric vehicles. In a world that becomes more aware of climate change, it is more convenient to have few electric vehicles driving around for delivering groceires, than that everyone has to drive to the supermarket by themselves. Furthermore, plastic bags and bottles can be returned every time groceries are delivered. For this reason, ordering groceries online is not only attractive for the eldery, but also for younger people.

The vehicles of Picnic are found in more and more cities throughout the Netherlands and the supermarket is very popular: there are huge waiting lists for people who want to become a customer, e.g. in Enschede, there are 20.000 people waiting [2]. Due to the popularity, a logical step is to to keep expanding to other cities. However, which city is most suitable to start delivering next? In which city can Picnic gain most profit? That is something to find out using data science.

2 The data

To find out where to make the most profit, it is required to determine factors that influence this. Unfortunately, Picnic does no provide data about their profits in certain regions of the Netherlands, but assumptions can be done on important factors. A multiple regression model can be constructed and it can be checked whether the predicted regions to make profit are indeed the regions where Picnic is already active or not. If this is true, from the model it can be determined what the next region can be.

The Dutch Central Statistical Office [3] provides many public data sets with information about e.g. population, income, house prices and more. This data base can provide useful parameters for a regression model, e.g. in regions where many people live or where people are wealthy, more profit can be made compared to poorer, less populated regions. Another indicator to determine if Picnic would be successful is the amount of grocery stores in a certain area. It is expected that Picnic would be more successful in regions where there are little grocery stores per 1.000 people compared to regions where there are many grocery stores per 1.000 people. For this reason, the Foursquare API can be used to determine the grocery store-density in certain regions.

It is also useful to know where Picnic is already delivering, first of all for validation of the model and secondly to be sure that in the conclusion it will no be recommended to start delivering in a region where Picnic is already active. This information is available on the website of Picnic [4].

Unfortunately, data of the Central Statistical Office is based on municipalities, where the Picnic delivers in cities rather than in whole municipalities. To be able to do an analysis, the analysis will be performed based on municipalities. For this reason, the city names from the website of Picnic are manually transformed to the municipalities where the cities are located in.

The remainder of this section consists of preprocessing data to obtain the required data frames.

2.1 Population data frame

First of all, a data frame is required which gives the population for each municipality. However. this population information is only found linked to municipality codes. For this reason, another data set is required to link municipality codes to the municipity names. By merging these two data sets, the required data set is created:

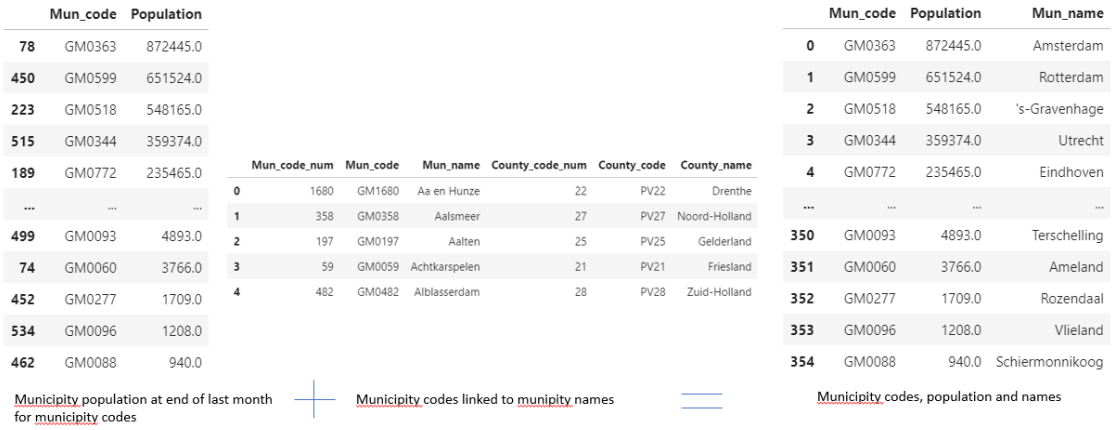

Figure 1: Population in municipalities

2.2 House prices data frame

From the CBS website, the average house price in each municipity is found:

	Mun_name	Average House Price
0	Appingedam	194.8
1	Delfzijl	155.1
2	Groningen	258.9
3	Loppersum	192.4
4	Almere	294.3
...	...	...
350	Hoeksche Waard	297.3
351	Het Hogeland	188.8
352	Westerkwartier	240.8
353	Noardeast-Fryslân	194.8
354	Molenlanden	302.0

Figure 2: Average house prices of municipalities

2.3 Grocery stores

At last information about existing grocery stores in the municipality is required. This leads to the following data frame:

	Mun_name	Mun_Latitude	Mun_Lon	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amsterdam	52.36993	4.90788	SPAR city Korstanje	52.370554	4.910020	Convenience Store
1	Amsterdam	52.36993	4.90788	Albert Heijn	52.368762	4.903477	Grocery Store
2	Amsterdam	52.36993	4.90788	SPAR city Van Kempen	52.366537	4.906142	Convenience Store
3	Rotterdam	51.91438	4.48716	SPAR city Glashaven	51.916405	4.485656	Convenience Store
4	Den Haag	52.08409	4.31732	SPAR Den Haag Centraal	52.081856	4.321592	Grocery Store
...	...	...	...	...	...	...	...
356	Eemnes	52.25336	5.25860	Lekker Gezond Groente En Fruit	52.252966	5.264246	Grocery Store
357	Terschelling	53.35839	5.21635	SPAR Berghuis West Terschelling	53.360309	5.215170	Convenience Store
358	Ameland	53.44296	5.68673	SPAR Manje	53.445197	5.686481	Convenience Store
359	Schiermonnikoog	53.47723	6.16121	SPAR Brunekreef Schiermonnikoog	53.478417	6.160780	Convenience Store
360	Schiermonnikoog	53.47723	6.16121	SPAR Brunekreef Schiermonnikoog	53.478307	6.160970	Convenience Store

Figure 3: Foursquare grocery store data frame

2.4 Final data frame

Now all features of the data frames described above are merged. The different grocery stores found using Foursquare are counted in each municipality and divided by the total population in the municipality as additional feature. An additional column is added to indicate whether Picnic is already active in the area (1) or not (0).

Using this data frame, the data analysis will be performed as discussed in the following sections.

	Mun_code	Population	Mun_name	Count	Average House Price	Population per store	Existing
0	GM0363	872445.0	Amsterdam	3.0	485.0	0.000003	1
1	GM0599	651524.0	Rotterdam	1.0	282.9	0.000002	1
2	GM0518	548165.0	Den Haag	3.0	324.8	0.000005	1
3	GM0344	359374.0	Utrecht	4.0	379.8	0.000011	1
4	GM0772	235465.0	Eindhoven	2.0	308.8	0.000008	1

Figure 4: Head of final data frame

3 Methodology

Because there is no information available about the profit in the municipalities where Picnic is already active, it is a binary problem: the model should be trained on information whether Picnic is already active or not. The features will be normalize and the data frame will be split in in a training- and test data set. Then the logistic regression model predicts based on the features of each municipality if Picnic is active in the municipality. The Jaccard score gives insight in the quality of the model.
However, the aim of the project is to find out which new municipalities might be interesting for Picnic. For this reason, for each municipality, the probability that Picnic is already active will be determined. If the probability for a municipality is high, while Picnic is not active, this is an interesting area for further expansion of the company. The log-loss metric indicates the quality of the predictions. The following section describes the results of the developed models.

4 Results

Now the results of the logistic regression models are discussed. For this purpose, the data features are normalized using the StandardScaler and the data set is divided into a training- and a test data set. The C of the logistic regression model is set to 1. The model leads to the following coefficients:

Population	House prices	Store count	Store / population
2.3	0.8	0.7	-1.0

Table 1: Logistic regression coefficients for each feature

The Jaccard score of the model is 0.58, which is bad. To evaluate the results, a confusion matrix is created as shown in figure 5.

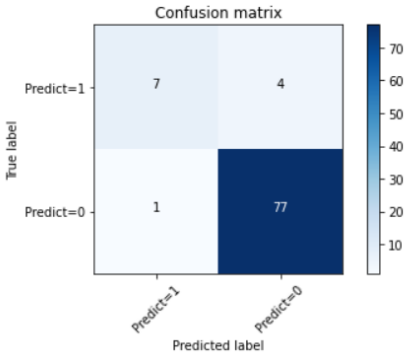

Figure 5: Confusion matrix of the logistic regression model

From the eleven locations where Picnic is already active, 7 out of the 11 are considered as good locations. However, there are also 4 municipalities classified as municipalities where Picnic is not already active. In section 5 it will be discussed how this model can be improved.

4.1 Probability logistic regression

The logistic regression classifies the municipalities in 0 or 1. If 1, the municipality is similar to the municipalities where Picnic is already active and if 0, not. But most of the most profitable municipalities are already occupied by Picnic and therefore the look is for the municipalities which thereafter will gain most profit. E.g. in the confusion matrix, the one municipality that is predicted that Picnic is active there, but where it is not, it interesting for future expanding of the company. To find out which municipalities have potential, the probabilities that a municipality belongs to class 1 are determined. Unfortunately, the log-loss metric is found to be 0.16, which is bad. The probabilities for each municipality are shown in figure 6. Red dots indicate a municipality where Picnic is not active yet. The green dots indicate that Picnic is there active already.

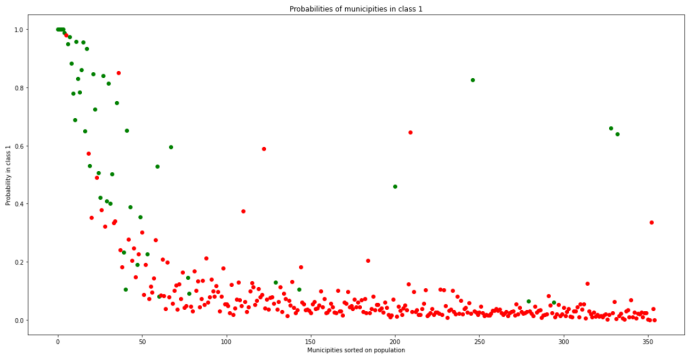

Figure 6: Scatter plot of the probabilities of class 1 for each municipality

The municipality are sorted on population, so the data points at the left are larger municipalities and many of municipalities where Picnic is active, are located at the left side of the spectrum.

However, some high probabilities are found at the right. To evaluate which municipalities these are, the data set is again divided in active- and non-active municipalities. First, the probabilities are added to the data frame. Then the data set is split. Figure 7 shows the data of the smallest municipalities in the data frame where Picnic is already active.

	Mun_code	Population	Mun_name	Count	Average House Price	Population per store	Existing	Predicted proba
77	GM1926	55441.0	Pijnacker-Nootdorp	0.0	375.9	0.000000	1	0.145476
78	GM0603	55096.0	Rijswijk	1.0	280.8	0.000018	1	0.090552
129	GM0406	41139.0	Huizen	0.0	408.5	0.000000	1	0.129256
143	GM0385	36220.0	Edam-Volendam	3.0	326.3	0.000083	1	0.104878
200	GM0397	27459.0	Heemstede	1.0	626.6	0.000036	1	0.458526
246	GM0377	23568.0	Bloemendaal	0.0	831.6	0.000000	1	0.826514
279	GM1842	19389.0	Midden-Delfland	0.0	399.6	0.000000	1	0.064697
294	GM0852	17374.0	Waterland	1.0	424.2	0.000058	1	0.061416
328	GM0376	11863.0	Blaricum	0.0	775.9	0.000000	1	0.660784
332	GM0417	11379.0	Laren	0.0	768.1	0.000000	1	0.640519

Figure 7: Data of the smallest municipalities in the data frame where Picnic is already active.

It is found that the house prices in some of these regions are very high and the amounts of grocery stores is low. This explains why although the population is small, these municipalities are interesting. However, other municipalities in this list are small, do no have high house prices and not few grocery stores. The locations of these municipalities are plotted on a map of the Netherlands to see where they the municipalities are located:

Figure 8: Areas where Picnic is active, low- and high probability values

This map shows that many of the municipities are relatively close to each other. What to do with this information, will be discussed further in the discussion section. Now first it will be determined

where Picnic should start delivering next, and where not, based on this logistic regression model:

	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
...	...	...	...	...	...	...	...	...	...
302	326	GM0765	12155.0	Pekela	1.0	161.3	0.000082	0	3.127557e-03
303	336	GM0965	10517.0	Simpelveld	2.0	227.4	0.000190	0	2.055613e-03
304	350	GM0093	4893.0	Terschelling	1.0	327.5	0.000204	0	1.863941e-03
305	351	GM0060	3766.0	Ameland	1.0	321.2	0.000266	0	6.650821e-04
306	354	GM0088	940.0	Schiermonnikoog	1.0	350.8	0.001064	0	3.952682e-09

Figure 9: Potential new areas for Picnic to start

This shows that Groningen should be the next step. Thereafter, Amstelveen is lsited, which is relatively large and has relatively high house prices. Wassenaar is listed third, which is small, but a very wealthy population.

5 Discussion

The metrics of the logistic regression model are, unfortunately, poor. Therefore, it is good to discuss how to improve the model.

5.1 Municipality - City level

First of all, because the Dutch Statistical Office provides data on municipality level rather than on city-level, the analysis is performed on municipality level. However, a municipality can consist of multiple cities from which one can be attractive for Picnic and the other is not. Furthermore, the coordinates to look for the amount of grocery stores per population can be picked in an area in a municipality where no city center is located, and because only in a certain radius grocery stores are searched using FourSquare, this can give a false representation of the grocery store density in the cities in the municipality.

For this reason, it would be interesting to have knowledge available on city-level rather than on municipality-level for more accurate prediction models.

5.2 Feature selection

Only a few features are selected: population, estimated amount of grocery stores in a radius in the municipality, grocery stores per population and house prices. However, possibly there are more factors playing a role in selecting new cities to start delivering. During this assignment, it was found that a similar analysis is performed by a professional company as well to look for potential areas for grocery delivery services, where also features as family size and age were considered [5]. From this analysis, it was actually found that income is a very good feature.

However, also the distance to already existing distribution centers or active areas can play a role in the selection of new areas. For example, it is found that Picnic is active in almost all large cities in the Netherlands, except for the number 5 city, which is Groningen, and which is located far from

the current active area of Picnic. And when looking at the map indicating where Picnic is already active, it is found that some municipalities with relatively low probabilities are close to existing areas. So distance to active areas can be a good feature to improve the model.

5.3 Clustering

Some of the municipalities where Picnic is already active have a very low probability value from the model. However, when looking at the map of the Netherlands indicating where Picnic is already active, it is found that multiple red spots (lower probability areas) are near the high probability areas. If multiple small municipalities are very close to each other, together they form a higher populated cluster, so it can be interesting to not look at individual municipalities, but cluster municipalities- or cities that are close to each other together.

5.4 Multiple Linear regression

Because no data was available about the profits of Picnic in certain areas, this analysis is limited to a classification model. However, if there is data available about the earnings of Picnic in already active areas, a multiple linear regression model can be build to predict the earnings in new areas. A training- and test data set can be constructed of already active areas to determine the accuracy of the profit predictions and if the model performs well, predictions can be done on the unknown areas. This can lead to better insights than only a classification model.

6 conclusion

Data analysis can be a powerful tool to predict where regions are with a high market potential. In this analysis, a logistic regression model is created to predict where Picnic should start delivering groceries. Unfortunately, the model did not perform very well. To improve the model, it would be interesting to start looking at city-level, to investigate additional features and to start clustering cities or municipalities. In a real scenario, it would be useful to have information about the earnings in active areas to be able to predict the market potential in new areas.

References

[1] "Online boodschappen doen wint aan populariteit." https://www.trouw.nl/nieuws/online-boodschappen-doen-wint-aan-populariteit~b232e89d/?referrer=https%3A%2F%2Fwww.google.com%2F, 2020. Retrieved on 10-11-2020.

[2] "Alweer 8.000 klanten op de wachtlijst: 'enschede is een echte picnic-stad'." https://www.tubantia.nl/enschede/alweer-8-000-klanten-op-de-wachtlijst-enschede-is-een-echte-picnic-stad~afe3c277/?referrer=https%3A%2F%2Fwww.google.com%2F, 2020. Retrieved on 10-11-2020.

[3] "Centraal bureau voor statistiek." https://www.cbs.nl/. Retrieved on 10-11-2020.

[4] "Picnic locaties." https://picnic.app/nl/locaties. Retrieved on 10-11-2020.

[5] V. R. A. Team, "Korte termijn groeipotentie van online boodschappen,"