The Battle of Coffee Shops

Maksim Agafonov, 23 Aug 2020

Introduction: Problem and background

In this project we are going to propose the optimal choice of coffee shop type for a given location. This study can be used by stakeholders who are interested in **coffee shop market in Warsaw**.

There are 2 ways to enter coffee shop market: you either buy franchise from a large chain or open an original coffeehouse.

- With franchise you'll get a fine tuned business model, brand and supply chain with prices negotiated for the large volumes
- With original coffeehouse you'll get an opportunity to make it from scratch

Interesting observation was made that a single location and small chain coffee shops co-exist with big chains. They didn't disappear due to competition powered by the economy of scale.

Let's not talk about countries with ages of coffee culture (Italy) or top coffee drinkers (Finland). We'll focus on Warsaw(Poland) where 2 big chains (Costa Coffee and Starbucks) are available.

There's the study showing that people are most likely looking for different experience when going to a big chain coffeehouse compared to a single location/small chain.

Customer experience from the place is significantly defined by it's spatial context.

We'll be using data science tools to determine **how similar or dissimilar are spatial contexts for different types of coffee shops**. And **what type of coffee shop is the optimal choice for a given location represented by it's spatial context**.

Data

Data sources

According to Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things". Following this idea will be doing study on spatial context of coffee shops in Warsaw.

Foursquare City Guide API will be used

- to obtain coffee shops locations in Warsaw
- to get spatial context for coffee shops - nearby venues

OpenStreetMap (OSM) data (obtained via Overpass API) will be used to enrich spatial context data by bringing more map features to our dataset

Geocoder Python library will help to get coordinates for Warsaw districts

Getting and cleaning the Data

To get all coffee shops in Warsaw available via Foursquare we first got Warsaw neighbourhoods (from Wikipedia) and geocoded them with OSM Geocoder help.

Then we used neighbourhoods coordinates as center point and searched Foursquare with 2000m radius limited to only Coffee Shop venue category (4bf58dd8d48988d1e0931735).

As 2000m radius search created duplicates – we cleaned them using unique venue ID (standard venue attribute).

Similar search but with 500m radius and without venue category limitation was done to obtain **nearby venues for coffee shops.**

Obtained venues data weren't required cleaning as we'll be using venue category only and this attribute is assumed to be correct.

OSM Overpass API was used to obtain documented map features (OSM tags) representing more **general features of spatial context.**

Some cleaning was applied to remove features(tags) not bringing much general spatial context (names, websites, emails, etc…) and remove to unique features (<50 occurrences in whole set).

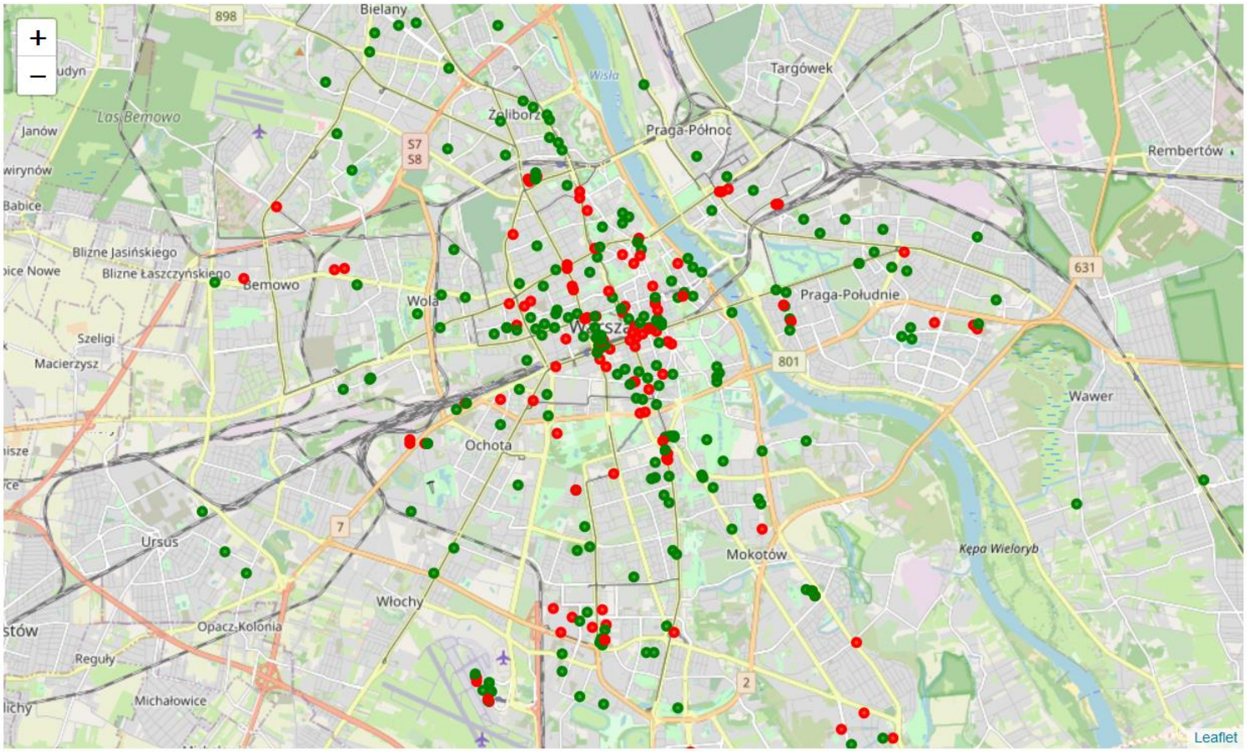
Exploratory data analysis

First look shown presence of 3 big chains in the city:

```
In [520]: coffee_shops['Coffee Shop'].value_counts().head()
```

```
Out[520]: Costa Coffee            67
          Green Caffè Nero        22
          Starbucks               21
          Tchibo                   4
          Kawiarnia Waszyngton     2
          Name: Coffee Shop, dtype: int64
```

It was decided to manually classify coffee shops to see their locations and construct ideas about suitable methodology.



Main observation: there are areas where single location/small chain coffee shops (marked green on map) co-exist with big chains (marked red). In these cases spatial context will be the same and we won't be able to separate these cases based only on map features.

Methodology

In this project we'll try to build classifier to assess how suitable is some given location for single location/small chain coffee shop in Warsaw.

We'll be doing it purely based on spatial context data obtained from Foursquare API (venues, PoI) and OSM Overpass API (OSM tags, more focused on general map features).

Raw data (both venues and OSM tags) were gathered with 500m radius around existing coffee shops locations. There are 301 unique venues categories and 612 OSM tags. We'll process the raw data to obtain occurrence frequency for venues and OSM tags (separately) - it will be used as features for our classifier.

2 features representing city density will be added - sum of all venues and OSM tags.

It was observed that there are areas where single location/small chain coffee shops (marked green on map) co-exist with big chains (marked red). In these cases spatial context will be the same and we won't be able to separate these cases based only on map features. We'll manually classify coffee shops as below:

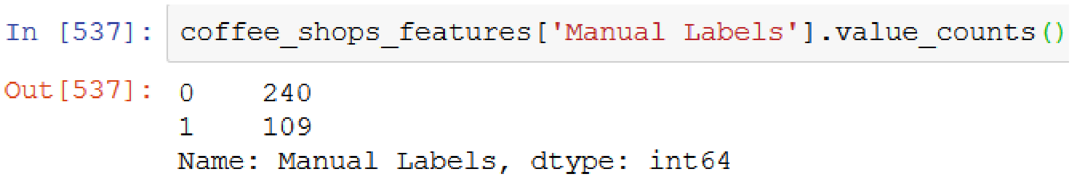
- Big chains (Costa, Starbucks, Nero) - class 0
- Single locations and small chains with big chains nearby - class 0
- All other single locations and small chains - class 1

We'll be using Logistic Regression classifier from scikit-learn - logistic regression will help to obtain actual probability values on top of predicted classes.

Finally we'll make use of our classifier to classify regular grid points in the part of Mokotów district of Warsaw. We'll visualize it using probability values on map to see final deliverable of this project.

Model development

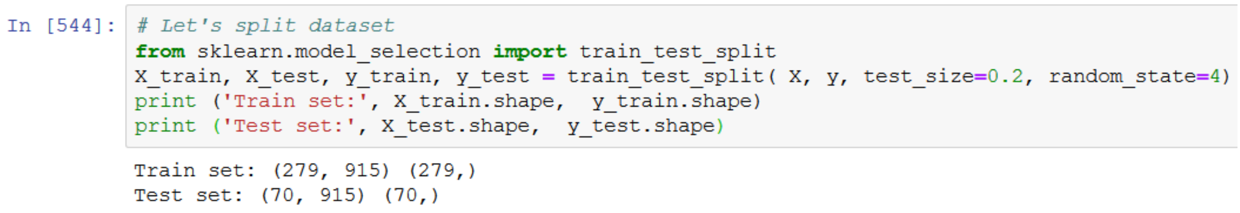
Manual classification using described above methodology gave us the following samples per class:



```
In [537]: coffee_shops_features['Manual Labels'].value_counts()

Out[537]: 0    240
          1    109
          Name: Manual Labels, dtype: int64
```

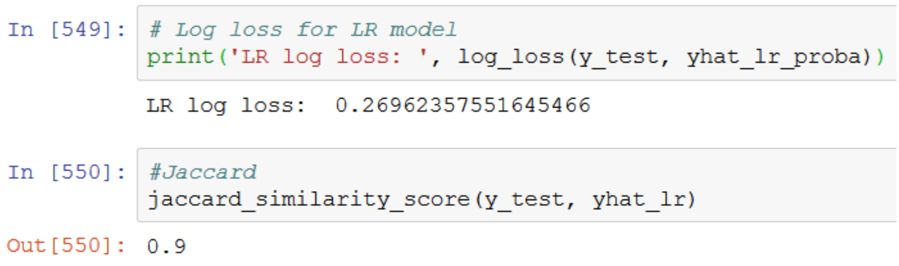
Train/test split as below:



```
In [544]: # Let's split dataset
          from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)
          print ('Train set:', X_train.shape,  y_train.shape)
          print ('Test set:', X_test.shape,  y_test.shape)

          Train set: (279, 915) (279,)
          Test set: (70, 915) (70,)
```

Quite good performance was observed:



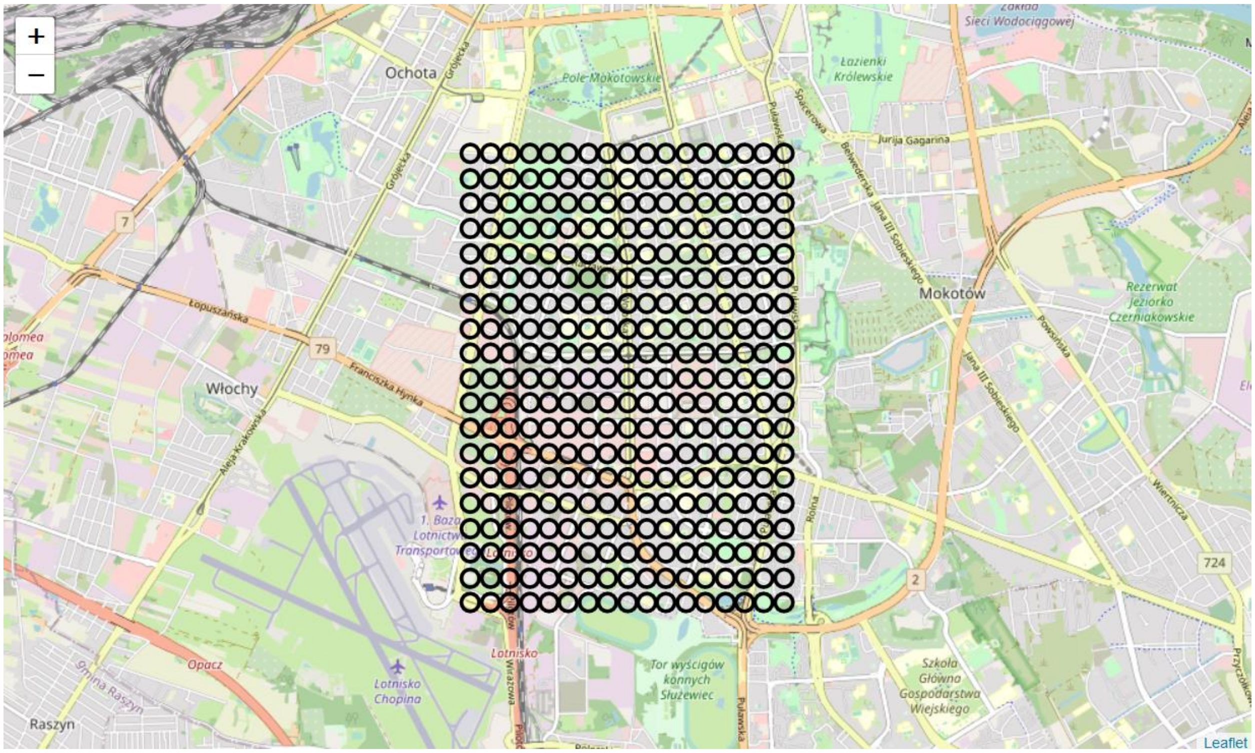
```
In [549]: # Log loss for LR model
          print('LR log loss: ', log_loss(y_test, yhat_lr_proba))

          LR log loss:  0.26962357551645466

In [550]: #Jaccard
          jaccard_similarity_score(y_test, yhat_lr)

Out[550]: 0.9
```

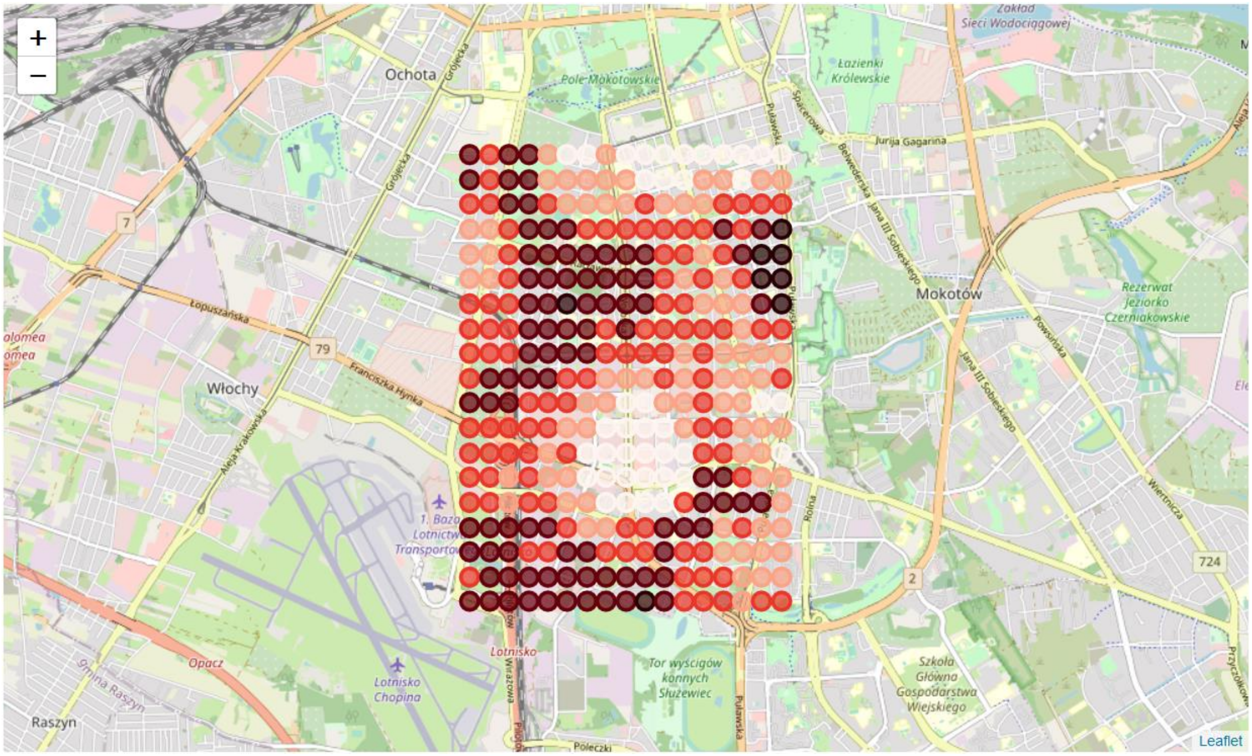
For project results regular grid was generated:



Results and Discussion

Our project resulted in working classification model helping to highlight suitable locations for new unique coffee shops in Warsaw. Example deliverable for a part of Mokotów district shows expected behavior:

- office and malls areas are not recommended as they are common location for big chain coffee shops – light points
- living areas, park areas, leisure areas are recommended ones – dark points



Areas of improvement for the same data are:

- optimization of the search radius for venues and OSM tags
- generalization of the model considering several cities of Poland

Conclusion

Obtained classifier most likely could be used as one contributing element of data science ensemble for selection proper location of unique coffee shop.

Model uses only documented OSM map features and Foursquare venues. And the main limitation is inability to separate single location coffee shops with big chains nearby.

It's expected that model could be improved with other datasets representing (for example) temporal features (people mobility, vehicles mobility).