Coursera: Applied Data Science Capstone

Selection of Store Locations

Author: Roland Schubert

May 17, 2019

1. Introduction

1.1 Background

In Germany, about 14 Million people are interested in equitation, 3.9 million classify themselves as “riders”, 1.25 million practice this sport intensely, 78% of them are women. 900,000 people own horses. So it is no surprise, that this hobby (or better call it passion) is also economically significant.
The turnover of the German horse industry is estimated at 6.7 billion euros. This includes 39% (€ 2.6bn) of expenditure on horse keeping, and 61% (€ 4.1bn) on retail and services. Retail is widely fragmented, with lots of individual enterprises, and a few chain stores. Within the last decade, these chains have expanded and this trend is still alive.

1.2 Problem

A retail company runs about 50 equestrian shops in different regions and is planning to expand its business. 5-10 new stores shall be opened within the next year and the expansion department already has identified possible locations. As opening a new store is a considerable investment and in addition, rental agreements are concluded on terms of at least 3 years, it is essential to perform an in-depth analysis of the respective locations and the specific conditions, and to predict the chances to have success at this new location, before a contract is signed and a store is opened, especially as existing stores differ extremely in terms of revenue and profit.
This project is intended to perform an analysis of store success based on specific market conditions of the surrounding area; especially customer situation, competition and additional surrounding factors will be analyzed.

1.3 Interest

Obviously, the management of the retail company is interested to make the right decisions on expansion strategy. Especially the managers in sales and marketing departments have to ensure success of new stores.

Of course, the companies’ shareholders also are interested in this decision. Based on currently 50 stores, the expansion plans mean increasing the company by up to 20%, which is a great opportunity, but in case of failure also a significant risk.

2. Overview of Data and Data Sources

2.1 Introduction

The intention of this project is to predict store success based on various information on the surrounding area.
Step 1: Identify various amenities in the surrounding area for existing stores; data on store success, geospatial data and amenities in surrounding area is needed
Step 2: Use classification models to predict store success; amenities in surrounding area is needed

While existing store addresses and data regarding success of existing stores are only available in company internal databases, is geospatial data and data referring venues in the surrounding area available from cloud resources.

2.2 Business Data

Extracts from the company database will be used:

- Addresses existing stores
- Success data for existing stores, Year 2018 (stores are classified as “0” (success below expectation) and “1” (success meets or exceeds expectations)

Data will be provided as “.csv” files

2.3 Google Maps Geocoding API

To convert addresses to geospatial information, OpenStreetMaps Geocoding API will be used.

Data will be accessed by Python using a geocoder library.

2.4 Foursquare location data

To determine the proximity of various amenities (e.g. riding stables, riding school, riding club, competition stores), Foursquare location data is used. Data will be accessed by Foursquare API calls.

3. Data Analysis and Modeling

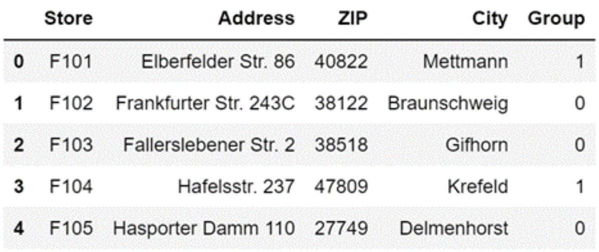
3.1 Introduction

Our analysis is conducted in 7 steps:

1. Load the data file
2. Assign longitude and latitude to each record (= store)
3. Add a count for competition, horse stables, riding clubs and riding school to each record (= store)
4. Split data to “training” and “test” data set
5. Build classification models with target “store success” and features “count of competitors”, “count of stables”, “count of clubs” and “count of schools” in surrounding area using
    a) K-Nearest Neighbours
    b) Decision Tree
    c) Support Vector Machine
    d) Logistic Regression
6. Perform a comparison (confusion matrix, metrics) and select the “best” model
7. Determine whether a model should be used to support decisions on store locations

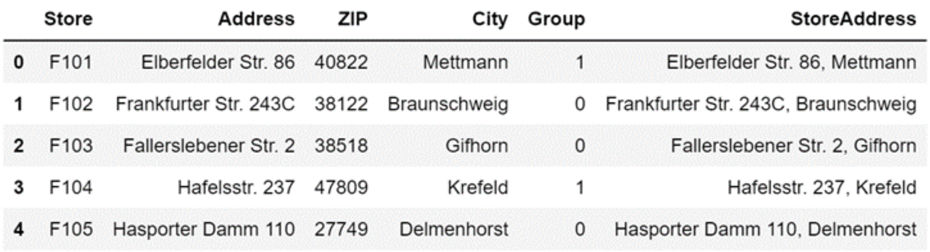
3.2 Load and Prepare Data

As the first step, we load the provided list of stores including store number, address, zip code, city and store category to a dataframe:



	Store	Address	ZIP	City	Group
0	F101	Elberfelder Str. 86	40822	Mettmann	1
1	F102	Frankfurter Str. 243C	38122	Braunschweig	0
2	F103	Fallerslebener Str. 2	38518	Gifhorn	0
3	F104	Hafelsstr. 237	47809	Krefeld	1
4	F105	Hasporter Damm 110	27749	Delmenhorst	0

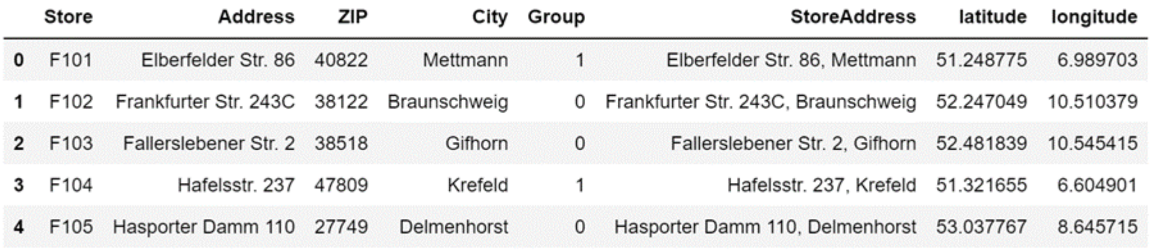
Unfortunately, the address format is not useable for the Nominatim API, so we have to create an additional column as a combination of address and city



	Store	Address	ZIP	City	Group	StoreAddress
0	F101	Elberfelder Str. 86	40822	Mettmann	1	Elberfelder Str. 86, Mettmann
1	F102	Frankfurter Str. 243C	38122	Braunschweig	0	Frankfurter Str. 243C, Braunschweig
2	F103	Fallerslebener Str. 2	38518	Gifhorn	0	Fallerslebener Str. 2, Gifhorn
3	F104	Hafelsstr. 237	47809	Krefeld	1	Hafelsstr. 237, Krefeld
4	F105	Hasporter Damm 110	27749	Delmenhorst	0	Hasporter Damm 110, Delmenhorst

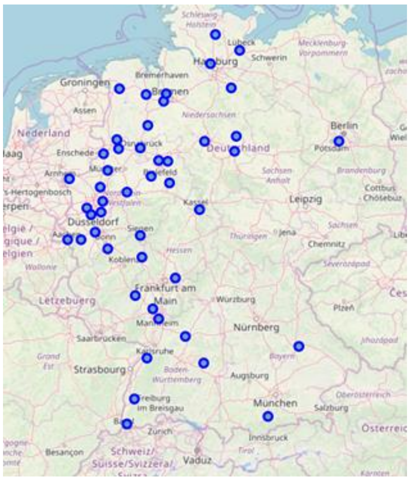
3.3 Add geospatial information

In the next step, we add geospatial information to the records. Using Nominatim to access the OpenStreetMap API, we retrieve latitude and longitude and add this information to the store data.



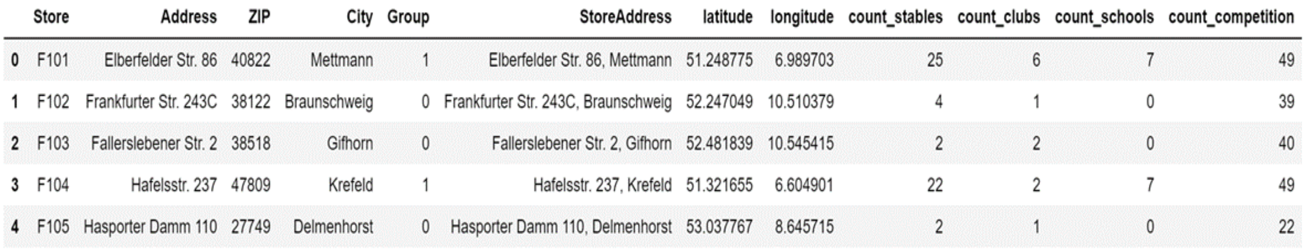
	Store	Address	ZIP	City	Group	StoreAddress	latitude	longitude
0	F101	Elberfelder Str. 86	40822	Mettmann	1	Elberfelder Str. 86, Mettmann	51.248775	6.989703
1	F102	Frankfurter Str. 243C	38122	Braunschweig	0	Frankfurter Str. 243C, Braunschweig	52.247049	10.510379
2	F103	Fallerslebener Str. 2	38518	Gifhorn	0	Fallerslebener Str. 2, Gifhorn	52.481839	10.545415
3	F104	Hafelsstr. 237	47809	Krefeld	1	Hafelsstr. 237, Krefeld	51.321655	6.604901
4	F105	Hasporter Damm 110	27749	Delmenhorst	0	Hasporter Damm 110, Delmenhorst	53.037767	8.645715

Now it is possible to display a map of all existing stores:



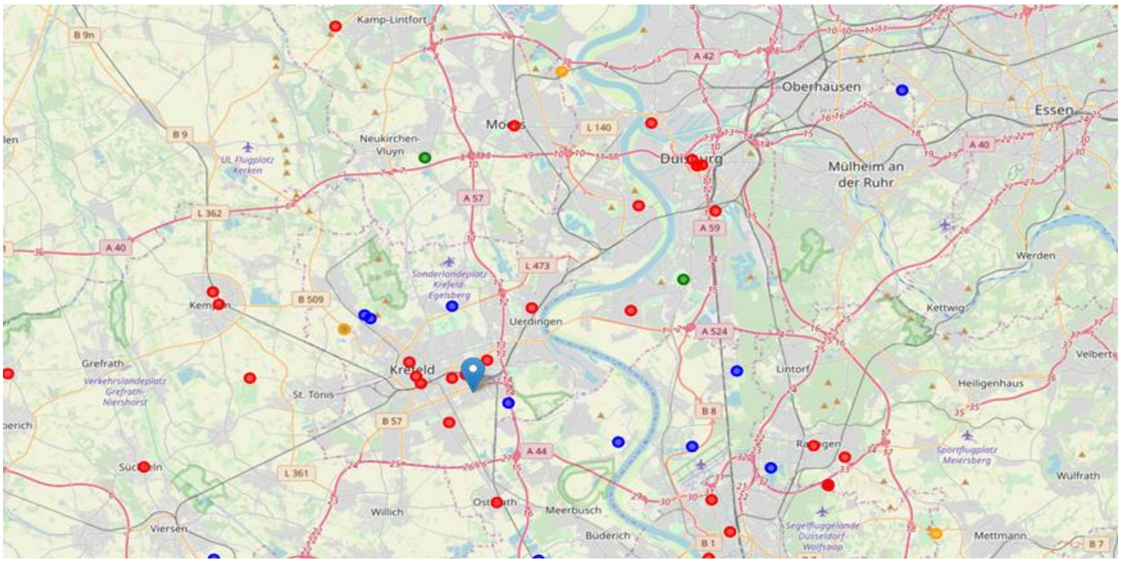
3.4 Identify rider amenities in surrounding areas

Now it is possible to retrieve information related to rider amenities and competition in the surrounding area. Using the FOURSQUARE developer API, we can these venues by a series of request for each record. For each record (store) four requests were generated (stables, riding clubs, riding schools and equitation stores). The result is transformed to a dataframe and result records were counted. Only the count for each category is added to the records.



	Store	Address	ZIP	City	Group	StoreAddress	latitude	longitude	count_stables	count_clubs	count_schools	count_competition
0	F101	Elberfelder Str. 86	40822	Mettmann	1	Elberfelder Str. 86, Mettmann	51.248775	6.989703	25	6	7	49
1	F102	Frankfurter Str. 243C	38122	Braunschweig	0	Frankfurter Str. 243C, Braunschweig	52.247049	10.510379	4	1	0	39
2	F103	Fallerslebener Str. 2	38518	Gifhorn	0	Fallerslebener Str. 2, Gifhorn	52.481839	10.545415	2	2	0	40
3	F104	Hafelsstr. 237	47809	Krefeld	1	Hafelsstr. 237, Krefeld	51.321655	6.604901	22	2	7	49
4	F105	Hasporter Damm 110	27749	Delmenhorst	0	Hasporter Damm 110, Delmenhorst	53.037767	8.645715	2	1	0	22

To have a more detailed look, it would be possible to display the single venues as shown here for a sample store. The marker indicates the store, red points indicate the competition, blue points are stables, green points are riding clubs and orange points riding clubs. For a specific analysis of stores this would be a feasible approach.



3.5 Train and test predictive model

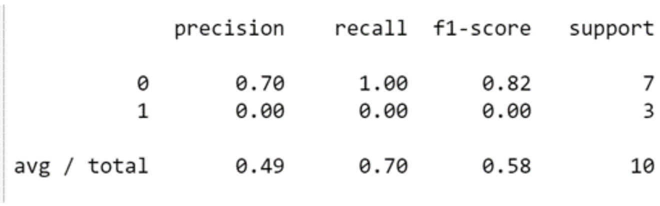
The dataset we created up to now is now split to training and test data. 80% is used as training data, the remaining 20% as test data.

Four different models were created:

- K-Nearest Neighbours
- Decision Tree
- Support Vector Machine
- Logistic Regression

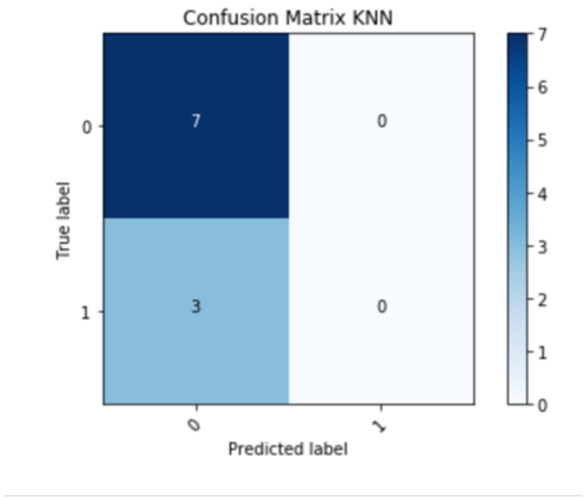
3.6 Compare models

The results were very similar for three of them; K-Nearest Neighbours, Decision Tree and Support Vector Machine perform really bad on prediction of “success stores” (in fact, no store is predicted to be a success ….):

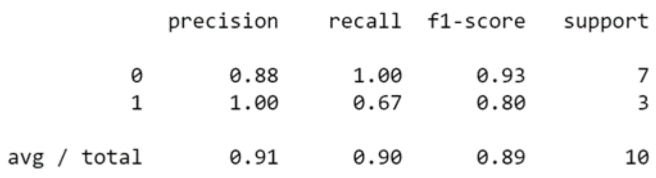


```
             precision    recall  f1-score   support

          0       0.70      1.00      0.82         7
          1       0.00      0.00      0.00         3

avg / total       0.49      0.70      0.58        10
```

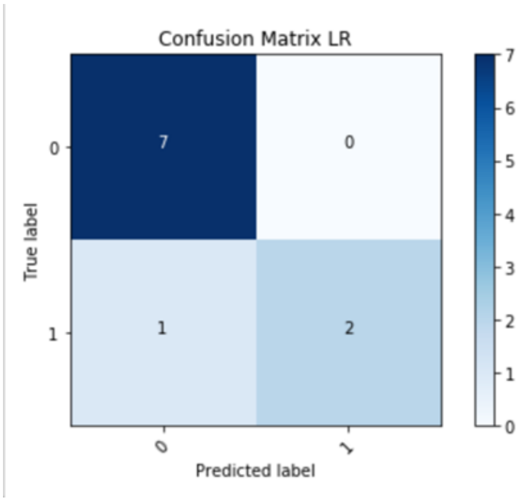
The confusion matrix reflects this:



The Logistic Regression model does better:



```
             precision    recall  f1-score   support

          0       0.88      1.00      0.93         7
          1       1.00      0.67      0.80         3

avg / total       0.91      0.90      0.89        10
```



While the prediction of “fails” works fine in all models, only Logistic Regression provides useable insights on “success stores”.

3.7 Model judgement

The only model we can use is Logistic Regression; all other models fail on “positives”.

4. Result, Discussion and Recommendations

Based on the results of Logistic Regression model, 2 out of 3 locations in the test data set are correctly predicted to be „successful“ (66% correctly predicted), while all “fails” were predicted correctly (100%).

If our intention is to avoid risks given by investments in stores, that do not achieve the expected revenue and profit, the model (especially Logistic Regression) provide useful hints – if we want to avoid to open stores at locations, that promise only moderate success, the model is useful.

But it may be, that opportunities are missed due to the “pessimistic” results; stores, that would be successful, may be predicted as fails and not opened.

As a first guideline, the model should be feasible, but not the only decision criteria.

5. Conclusion

Logistic Regression can provide some hints to explain or predict store success, but improvement seems to be possible.

Especially the retrieval of rider amenities and competitor stores is not very specific.

For all stores, a surrounding area of 20 km has been defined. I think, a differentiation between urban and rural areas can be a major improvement. In urban areas people expect shopping opportunities to be close, while in rural areas driving distance is widely accepted, so different area definitions (e.g. area of 40 km for rural, 10 km for urban) should be applied.

In general, if amenities for riders exist, it can be interpreted as an indication, that there are potential customers, but there may be overlaps (a rider makes use of a stable, visits a riding school and is member of a riding club), so simply “counting” may not be the right approach, it tends to be inaccurate. More detailed information on the venues is needed (e.g. the size of a stable, number of members in a club) to create a more dependable data set.

Count of competitors is also vague, there are stores only selling forage and litter, others only carry clothes or saddles, so more detail is needed on this to judge competition accordingly.

Further investigation is required, to make a final decision – but I think, the approach should be seen as a promising first step.