SELECTION OF STORE LOCATIONS

APPLIED DATA SCIENCE CAPSTONE PROJECT

MARKET CONDITIONS AND FUTURE CONSIDERATIONS

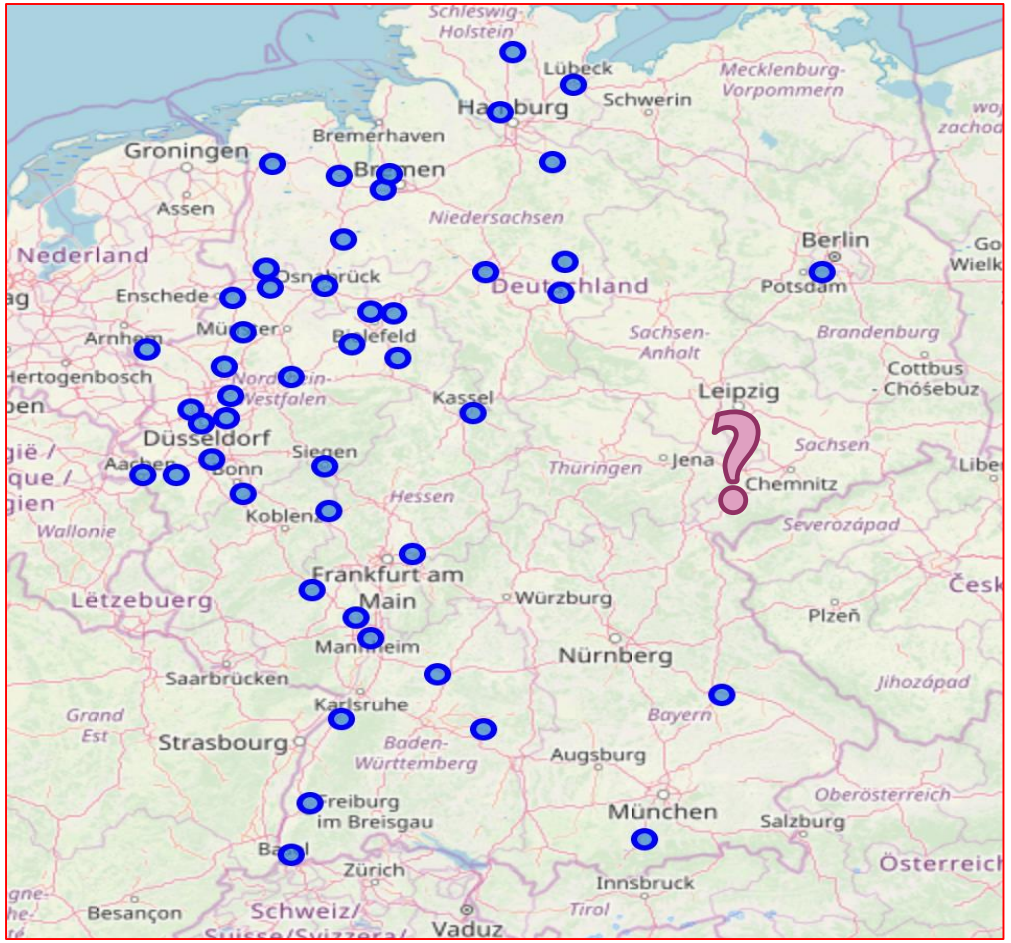
- Horse riding is very popular in Germany
- 3.9 Mio. riders, 1.25 Mio. practice riding intensely
- Market volume 4.1bn €
- Retail is very fragmented, big opportunity for growth

- Currently 46 stores opened
- Additional 5-10 stores planned for next year
- Major increase in revenue and profit possible
- BUT: only 16 of 46 stores perform as expected or exceed expectations!
- Investment in new stores should be allocated to best suited locations to optimize results

PROBLEM: SELECT STORE LOCATIONS

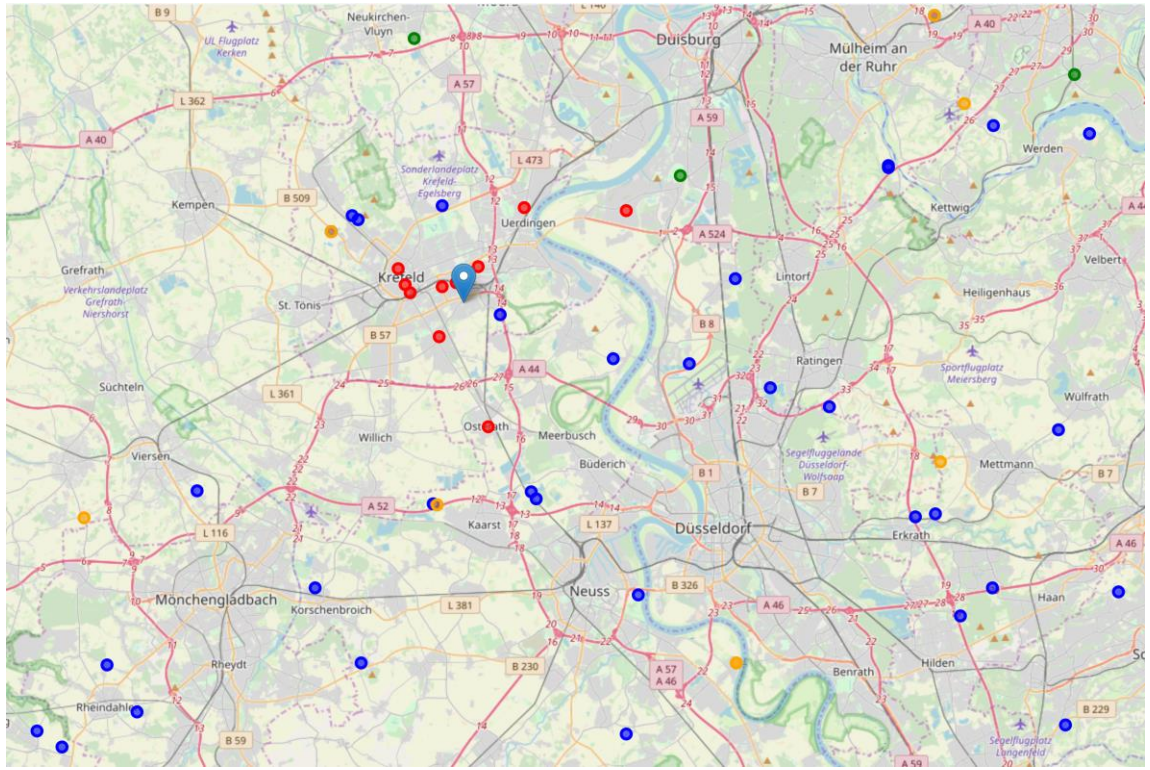
- Selecting a new store location is an opportunity to increase revenue, but also a risk
- Where should additional stores be opened?

- Idea:
  - Analyze existing stores to identify venues in the surrounding area, that favour or interfere store success
  - Hypothesis is, that success is fostered by a lot of customers in proximity and hindered by extensive competition nearby
  - Competition means other/competing stores
  - Sign of customers is existence of amenities used by riders (e.g. stables, riding schools, riding clubs)

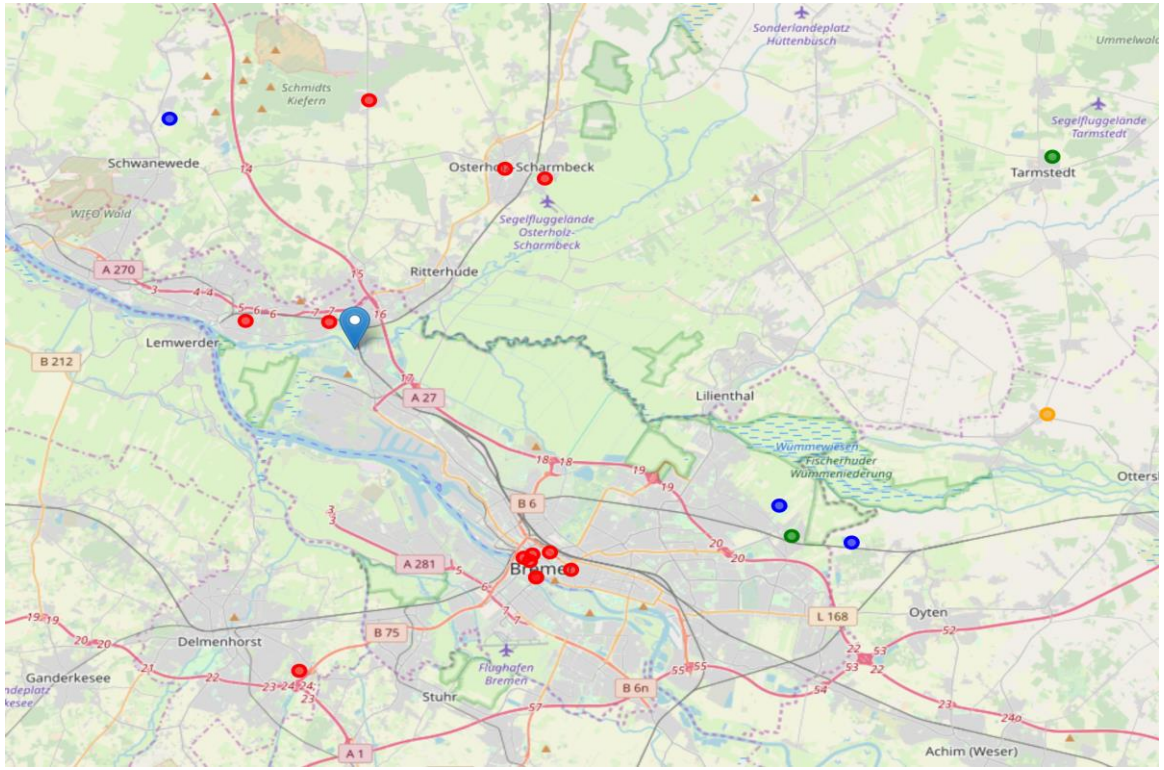


STORE COMPARISION: SUCCESS DRIVEN BY SURROUNDING AREA?

Store „great performance“
lot of riders amenities, some competitors



Store „poor performance“
few riders amenities, many competitors



Marker: Store --- Points: Red: Competition – Blue: Stables/Orange: Riding Schools/Green: Riding Clubs

DATA AND DATA SOURCES

- Store Data:
  - Number
  - Address
  - ZIP Code
  - City
  - Category (1 = exceeds expectations, 0 = below expectations)
- Geospatial Data:
  - longitude and latitude of stores
- Competition:
  - Equitation stores in surrounding area for all stores
- Indicators for customers:
  - Horse Stables
  - Riding schools
  - Riding clubs

Company internal data source, Category calculated based on revenue & profit compared to budget

Nominatim used to gain information from OpenStreetMap

Foursquare dataFoursquare Developers Access to venue data: https://foursquare.com/

STEPS OF ANALYSIS AND METHODOLOGY

1. load the data file
2. assign longitude und 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

DATA LOAD & PREPARATION

- load data file to 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

- create new address field to fit the needs of Nomatim

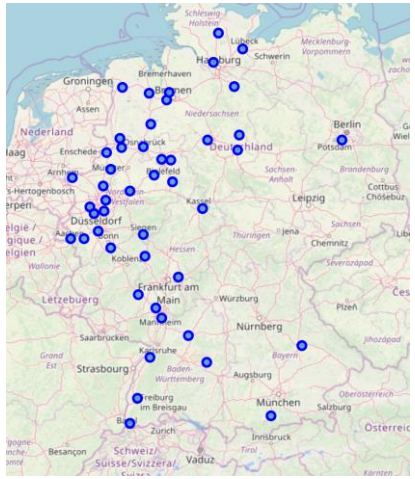
	Store	Address	ZIP	City	Group	StoreAddress
0	F101	Elberfelder Str. 86	40822	Mettmann	1	Elberfelder Str. 86, Mettmann
1	F102	Frankfurter Str. 243C	38122	Braunschweig	0	Frankfurter Str. 243C, Braunschweig
2	F103	Fallerslebener Str. 2	38518	Gifhorn	0	Fallerslebener Str. 2, Gifhorn
3	F104	Hafelsstr. 237	47809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

ASSIGN GEOSPATIAL INFORMATION TO STORES

- Retrieve latitude and longiture for all records in dataframe

	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

- Review current store locations

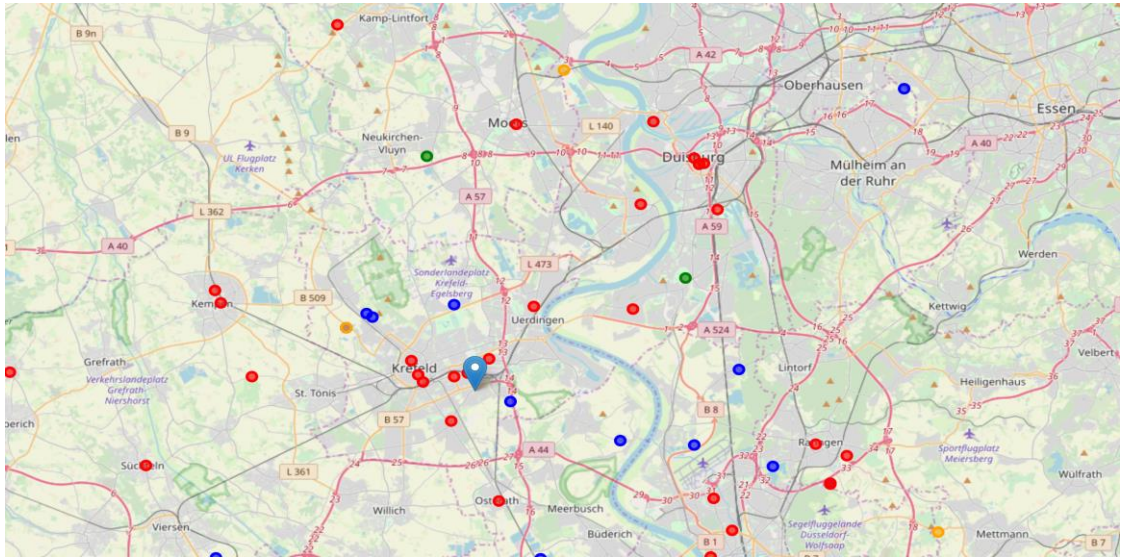


IDENTIFY RIDERS AMENITIES IN SURRONDING AREAS

- using FOURSQUARE developer API, identify riders amenities and competitors in the surrounding area, and assign the count to the records in the dataframe

- visualize for sample stores – instead of the count this time the locations (red = competition, blue/orange/green = stable/school/club)

	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

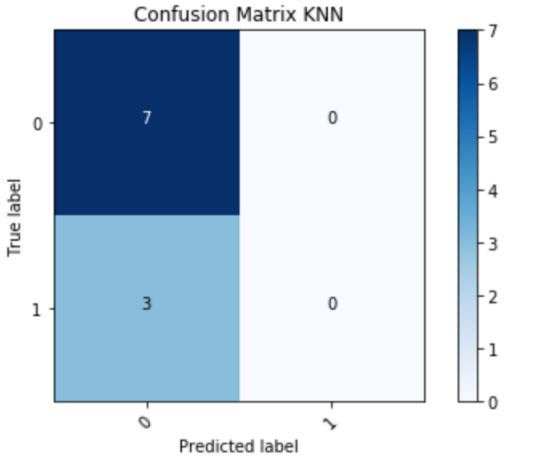


CREATE, TRAIN AND TEST CLASSIFICATION MODELS

- Split dataset to training and test data (80 % training data, 20 % test data)
- Train model on training data set
- Create prediction on test data set
- Calculate metrics

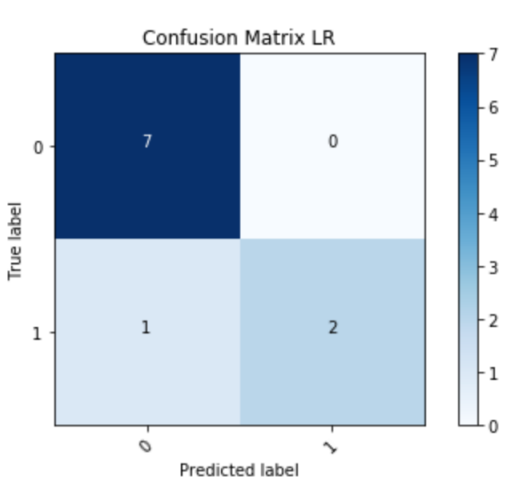
COMPARE MODELS

- Confusion Matrix for K-Nearest Neighbours, Decision Tree and Support Vector Machine looks very similar, all other metrics also differ very slightl



	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

- Logistic Regression does better on both Confusion Matrix and other metrics



	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

RESULTS

- Based on the results of Logistic Regression model, 2 out of 3 locations in the test data set are correctly predicted to be „successful“
- All other models predict 0 out of 3!
- Prediction of „possible fails“ works very well, so for a risk averse decision, 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
- Predictions of „successful stores“ are not as reliable

CONCLUSION

- Logistic Regression can provide some hints to explain or predict store success, but needs improvement

- Possible enhancements:
  - Different area definitions for rural and urban environments (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)
  - Simply counting the amenities is very inaccurate, more detailed information on the venues is needed (e.g. the size of a stable, number of members in a club)
  - Count of competitors is also vague, there a store 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!