LUKAS
WEIDICH

APPLIED DATA SCIENCE CAPSTONE PROJECT

Using Foursquare and eBay to analyze computer & hardware sellers in my area

INTRODUCTION

What is it about?

MOTIVATION

There are way too many different stores and sellers to buy computers from, can I apply what I‘ve learned in the course and gain insights?

The scenario is purely constructed and not connected to a real life use case, still:

- I want to collect data from different sources and unify them
- I want to visualize all possible stores and sellers I can visit
- I want to apply machine learning algorithms to predict features

GETTING THE DATA

Using Foursquare and eBay as my sources

GETTING THE DATA

Foursquare

- As required, I will use Foursquare
- Foursquare provides geographical information about venues and places of interest
- I will use foursquare to find computer stores within a 100km radius of my area

eBay

- Additionally to Foursquare, I will also use the eBay API
- The API provides information about all active listings for a specific keyword or category
- I used the category computers and notebooks and also only considered results within a 100km radius

METHODOLOGY

Which methods were applied?

UNIFYING AND VISUALIZING ENTRIES

The two datasets from Foursquare and eBay were available in different structures, so I needed to unify them. Missing columns were added using geocoding.

Steps I took before visualizing the data

1. Retrieve the data
2. Define columns I need to work with later in the project
3. Fill missing columns
4. Merge the two datasets into one

UNIFYING AND VISUALIZING ENTRIES

Datasets before unifying

```
df_foursquare.head()
```

	Name	Latitude	Longi
0	Computer Service Dröge	52.279037	8.90
1	TM Computer e.K.	52.116941302856596	8.6726503444
2	TM Computer e.K. - Thomas Müller (Büro)	52.15775752422116	8.6342995069
3	Schormann-Computer	52.18319156583845	8.87479877468
4	4you-computer.de	52.24314880371094	8.84024906158

```
df_ebay.head()
```

	itemId	title	globalId	primaryCategory/categoryId	primaryCategory/categoryName
0	233800267763	ACER Nitro 5 (AN515-52-76YJ) schwarz Gaming-No...	EBAY-DE	177	PC Notebooks & Netbooks
1	184560017711	MacBook Pro (Retina 15 Zoll, Anfang 2013), 8 G...	EBAY-DE	111422	Apple Notebooks
2	154228405289	MacBook Pro 13" 2016 Touch Bar, generalüberholt	EBAY-DE	111422	Apple Notebooks
3	133590834366	HP Pavillion Notebook, 17", 16GB, AMD A8-6410 ...	EBAY-DE	177	PC Notebooks & Netbooks
4	383843713150	Laptop Acer Extensa 5635*15,6 Zoll*Intel Core ...	EBAY-DE	177	PC Notebooks & Netbooks

Explaining the data

- Foursquare only hat coordinates
- eBay data could only provide the city name

➔ To fill the missing columns, I used geocoding

UNIFYING AND VISUALIZING ENTRIES

Datasets after unifying

```
# add columns city, state to foursquare df
states = []
cities = []
for lat, lng in zip(df_foursquare["Latitude"], df_foursquare["Longitude"]):
    g = geocoder.osm(str(lat+","+lng)).json
    states.append(g["state"])
    if("town" in g):
        cities.append(g["town"])
    elif("city" in g):
        cities.append(g["city"])
    else:
        cities.append("-")
df_foursquare["State"] = states
df_foursquare["City"] = cities

df_foursquare.head()
```

	Name	Latitude	Longitude	Source	State	City
0	Computer Service Dröge	52.279037	8.902915	Foursquare	Nordrhein-Westfalen	Minden
1	TM Computer e.K. - Thomas Müller (Büro)	52.15775752422116	8.63429950693619	Foursquare	Nordrhein-Westfalen	Hiddenhausen
2	TM Computer e.K.	52.116941302856596	8.67265034442815	Foursquare	Nordrhein-Westfalen	Herford-Stadt
3	Schormann-Computer	52.18319156583845	8.874798774686496	Foursquare	Nordrhein-Westfalen	Vlotho
4	4you-computer.de	52.24314880371094	8.840249061584473	Foursquare	Nordrhein-Westfalen	Bad Oeynhausen

```
df_ebay_corrected.head()
```

	Name	City	State	Latitude	Longitude	Source
0	ACER Nitro 5 (AN515-52-76YJ) schwarz Gaming-No...	Ronnenberg	Niedersachsen	52.316662	9.653208	eBay
1	MacBook Pro (Retina 15 Zoll, Anfang 2013), 8 G...	Warstein	Nordrhein-Westfalen	51.445811	8.353682	eBay
2	MacBook Pro 13" 2016 Touch Bar, generalüberholt	Hannover	Niedersachsen	52.374478	9.738553	eBay
3	HP Pavillion Notebook, 17", 16GB, AMD A8-6410 ...	Paderborn	Nordrhein-Westfalen	51.717704	8.752653	eBay
4	Laptop Acer Extensa 5635*15,6 Zoll*Intel Core ...	Bückeburg	Niedersachsen	52.261104	9.048896	eBay

USING KNN TO PREDICT FEATURES

Working with geographical data

- My combined dataset was missing sales data or any other more insigtful values
- So I decided to apply KNN to predict the state based on the coordinates the entry is located at

USING KNN TO PREDICT FEATURES

Steps I took before applying KNN

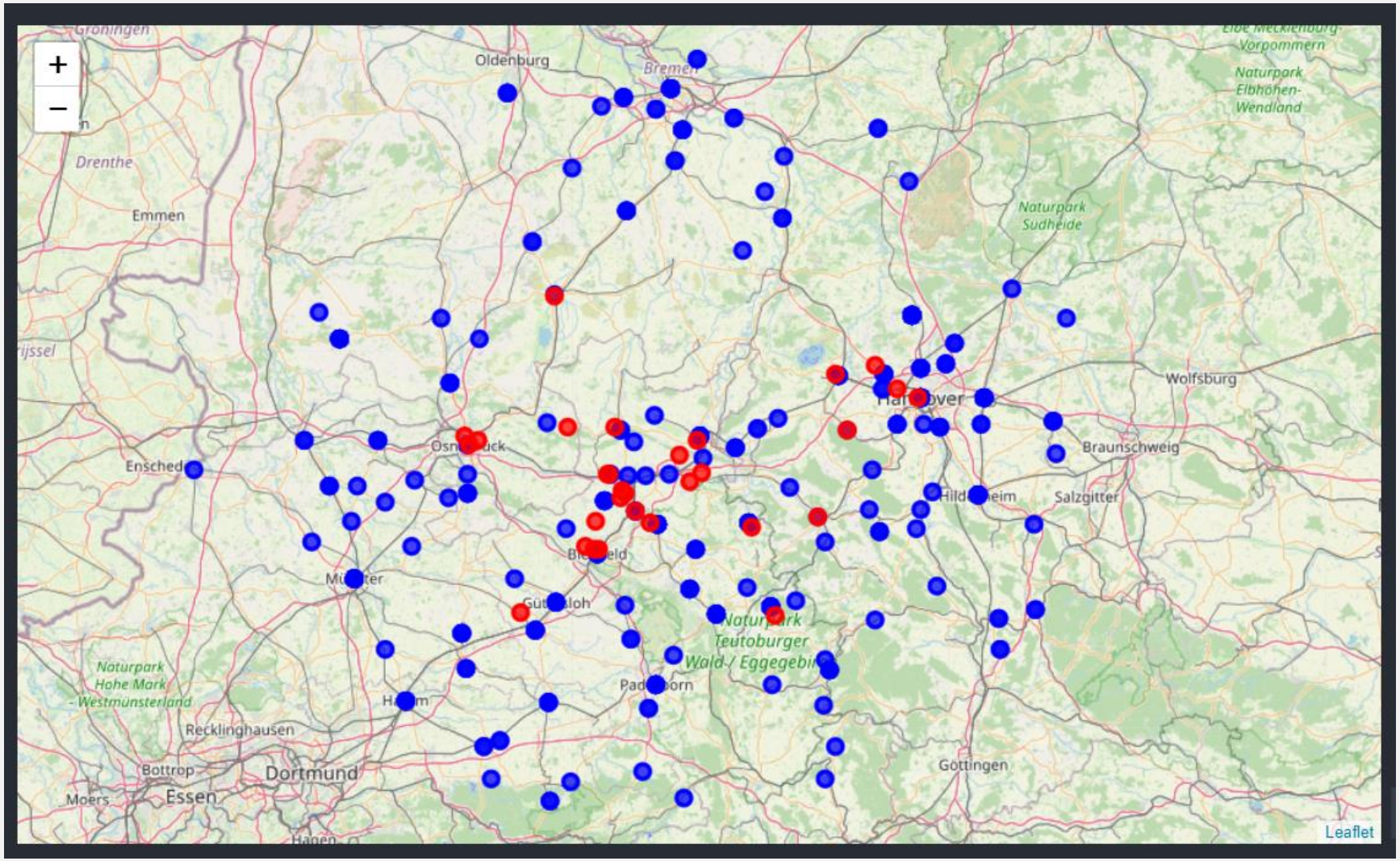
1. Select relevant columns for independent (X) and dependent (y) variables
2. Preprocess X and y
    1. Standardize numerical values (Latitude and Longitude)
    2. One Hot Encode categorical values in state
3. Define test and train set
4. Find best k

RESULTS

Insights and learnings

UNIFYING AND VISUALIZING ENTRIES

Datasets visualized using a folium map (red: Foursquare, blue: eBay)



USING KNN TO PREDICT FEATURES

Finding best k (=6) and predicting the state, then measure accuracy

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics

k = 6
neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
neigh
```

```
KNeighborsClassifier(n_neighbors=6)
```

```
yhat = neigh.predict(X_test)
yhat[0:5]
```

```
array([[0, 0, 0, 0, 0, 0, 1, 0, 0],
       [0, 1, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 1, 0, 0],
       [0, 0, 0, 0, 0, 0, 1, 0, 0],
       [0, 0, 0, 0, 0, 0, 1, 0, 0]], dtype=uint8)
```

```
from sklearn import metrics
print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh.predict(X_train)))
print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))
```

```
Train set Accuracy:  0.9682539682539683
Test set Accuracy:  0.9841269841269841
```

DISCUSSION AND CONCLUSION

Challenges I faced and what I take away

FINALLY,

I am looking back at the project and want to summarize the challenges I faced what and conclude what I have learned and take away from this short project

DISCUSSION

- I was able to collect data from computer stores in my area using eBay and Foursquare
- Results are somewhat insightful, as I could unify the two datasets, visualize the different locations and offerings within my area and apply machine learning techniques
- Some cities were mistakenly matched to a different state or country when using geocoding, but these outliers did not affect the result very much

CONCLUSION

- I really enjoyed applying the concepts and methods that were presented in the course
- Although the project is constructed and not tied to a real world business case, I still think it acts as a solid foundation for future work