Geo-Prediction

Advanced Software-Engineering
Dr. Harald Stein
Prof. Dr. Selcan Ipek-Ugay
WS 2024/25

Agenda

- Introduction
- Binary Logistic Regression
- Multinomial Logistic Regression (MLR)
- Finding new customers, approach 1: MLR
- Similarity analysis with cosine distance
- Finding new customers, approach 2: Cosine Similarities

Introduction

Targeted marketing for existing and new customers in particular sociodemographic segments and geolocations

- Benefits of Targeting New Customers Analytically
- Existing vs. New Customers
- Existing customer clusters
- Clustering as part of machine learning

Benefits of Targeting New Customers Analytically

… by predicting potential customers' interests and preferences based on existing customer clusters/segments

Enhanced Precision in Marketing

- Utilizes data-driven insights to accurately identify and reach the ideal customer profile.
- Creates personalized experiences based on customer data, increasing engagement and loyalty.

Competitive Advantage

- Provides insights into market trends and customer behavior, offering a strategic edge over competitors.
- Conversion rates: Analytics enable prediction and targeting of those most likely to convert, improving overall marketing effectiveness.
- Reduces wasted resources by focusing marketing efforts where they are most likely to yield results.

Existing vs. New Customers

Focus: Sociodemographic segments of new customers

Focus

Existing Customers

Understanding the Audience

- Analyse historical data understanding segments, preferences

Communication Strategy

- Focus on customers or segments with high turnover, loyalty

Customized Offers

- Tailor promotions based on past behavior and preferences.

Feedback Loop

- Gather and act on feedback for continuous improvement.

New Customers

- Use market research, predictions to identify potential needs.
- Emphasize awareness and brand introduction.
- Provide introductory offers to attract trial and engagement.
- Use initial interaction to understand preferences, expectations.

Effect Analysis Table: Key metrics

Interpreting Mean Values, Eta Squared and Z-Transformations for all segmentation features

Mean Values

- Central measure indicating average outcomes within each group
- Fundamental for identifying scales and differences

Eta Squared (η^2)

- Measures proportion of total variance in dependent variable that is attributable to a factor,
- Key for understanding effect size in ANOVA tests

Z-Transformed Values

- Standardization of data points based on mean, standard deviation
- Useful for comparing scores from different distributions and identifying outliers

Index	description	1	2	3	4	Total mear	Eta squared	z_1	z_2	z_3	z_4
0	Population ratio: school diploma	23.5387	2.54316	11.5277	5.43047	4.89768	0.886345	3.8453	-0.485695	1.36765	0.109903
1	Population ratio: college degree	35.4876	4.19433	16.1253	7.77358	7.38703	0.874354	4.01073	-0.455686	1.24719	0.0551722
2	Population in the house	74.7547	10.8867	37.1235	18.1483	17.6136	0.851693	3.85619	-0.453966	1.31664	0.0360881
3	Population ratio: university degree	17.2787	2.46361	9.34252	4.64443	4.21684	0.774792	3.46458	-0.465035	1.35956	0.113415
4	Tendency of having children	7.3713	3.36733	7.5142	6.21051	4.564	0.730642	1.44134	-0.6144	1.51471	0.845363
5	Interest: social media	6.84566	4.0987	6.96723	6.85288	5.08522	0.566425	0.990565	-0.555095	1.05897	0.994629
6	House size: number of flats	2.37237	1.31871	3.11911	2.56002	1.81605	0.564792	0.601959	-0.538126	1.40994	0.804996
7	Interest: green energy	8.29343	5.7935	8.80948	8.966	6.87341	0.506247	0.688072	-0.523269	0.938124	1.01397
8	Main purchase criterion: low price	7.78607	5.10585	7.99666	8.19596	6.16072	0.439162	0.752138	-0.488142	0.84959	0.941815
9	Population density (households per qkm)	8.00399	5.56313	10.599	9.90372	7.11272	0.438949	0.275552	-0.479079	1.07785	0.862881
10	Interest: online shopping	5.44948	4.56909	6.50069	6.69782	5.2555	0.300827	0.111423	-0.39428	0.715252	0.828487
11	Status of Location	2.80903	2.05673	3.00239	2.73242	2.32961	0.177814	0.536042	-0.305113	0.752239	0.450386
12	Interest: politics	6.8313	6.29871	7.65781	8.1999	6.8604	0.174146	-0.0151308	-0.292022	0.414574	0.696403
13	Interest: football	3.33846	2.91265	4.27357	3.76197	3.25778	0.168634	0.0662808	-0.283511	0.834444	0.414179
14	Interest: travel	6.13116	6.33569	6.71202	7.83844	6.67687	0.085086	-0.265093	-0.165737	0.0170743	0.564263
15	Tendency of having a photovoltaic system	3.3949	3.48795	3.35546	4.39615	3.65331	0.0582588	-0.165109	-0.105654	-0.190305	0.474621
16	Interest: economics	5.91435	5.16395	5.28867	5.81669	5.33658	0.0507714	0.465909	-0.139203	-0.038631	0.387153
17	Main purchase criterion: brand/quality	4.32879	3.90469	3.96371	4.4824	4.04322	0.0280004	0.20326	-0.0986019	-0.0565912	0.312599
18	Interest: alternative energy	3.77087	3.87738	3.87853	4.48164	3.99626	0.0240819	-0.142215	-0.0750053	-0.0742799	0.306257
19	PackageName_Cruise Getaway	0.203716	0.201072	0.192521	0.191537	0.198236	0.0165496	0.160826	0.0832289	-0.167756	-0.196628
20	PackageName_Mountain Adventure	0.190032	0.193817	0.201803	0.200199	0.195909	0.00849139	-0.157375	-0.0560108	0.157819	0.114883
21	PackageName_City Break	0.192885	0.202445	0.209366	0.205772	0.203595	0.00821105	-0.311798	-0.0334729	0.168006	0.0633862
22	PackageName_Beach Escape	0.207413	0.201423	0.198233	0.196831	0.200328	0.00449757	0.202274	0.0312598	-0.0598113	-0.0998352
23	PackageName_Safari Expedition	0.205954	0.201242	0.198077	0.20566	0.201931	0.00444456	0.115062	-0.0197062	-0.110236	0.106634

Interpreting cluster descriptions

… using most extreme, i.e. most positive or negative z-transformed values is useful tool for targeted marketing among existing customers

Cluster 1:
Educated urban families

- People with higher education degree
- Families with kids
- Living in larger buildings, urban
- Being interested in social media
- Travel preference: Beach escape

Cluster 3:
Upper-class Families interested in sports

- People with college degree
- Families with kids
- Living in small houses of higher status
- Interested in sports
- Travel preference: City Break

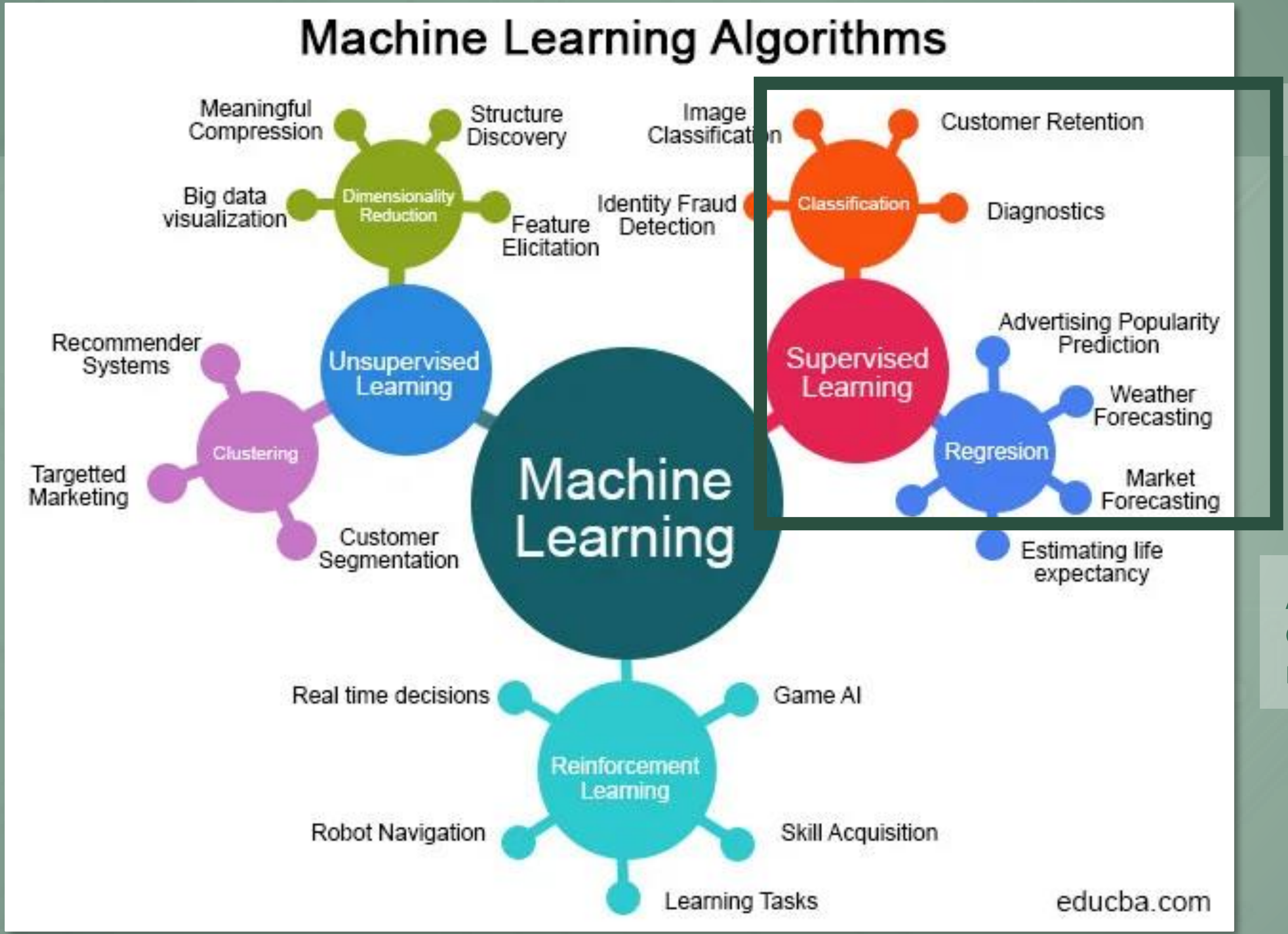
Cluster 2:
Brand aware working class

- Working class
- Interested in photovoltaics, travelling
- Brand and quality aware
- Travel preference: Cruise gateway

Cluster 4:
Low-income suburbian families

- Low population density, i.e. suburbian
- Families with kids
- Low price oriented, rather poor
- Travel preference: Mountain adventure

Classification/Prediction as part of Machine Learning



Focus

Additionally:
Cosine similarity approach is not machine learning

Binary Logistic Regression

… for data generalization of type classification with two outcome possibilities

- Aim of binary logistic regression
- Centerpiece of logistic regression
- Backpropagation
- Fit the model with L-BFGS
- Validation of results

Aim of binary logistic regression

… is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable) and a set of independent variables.

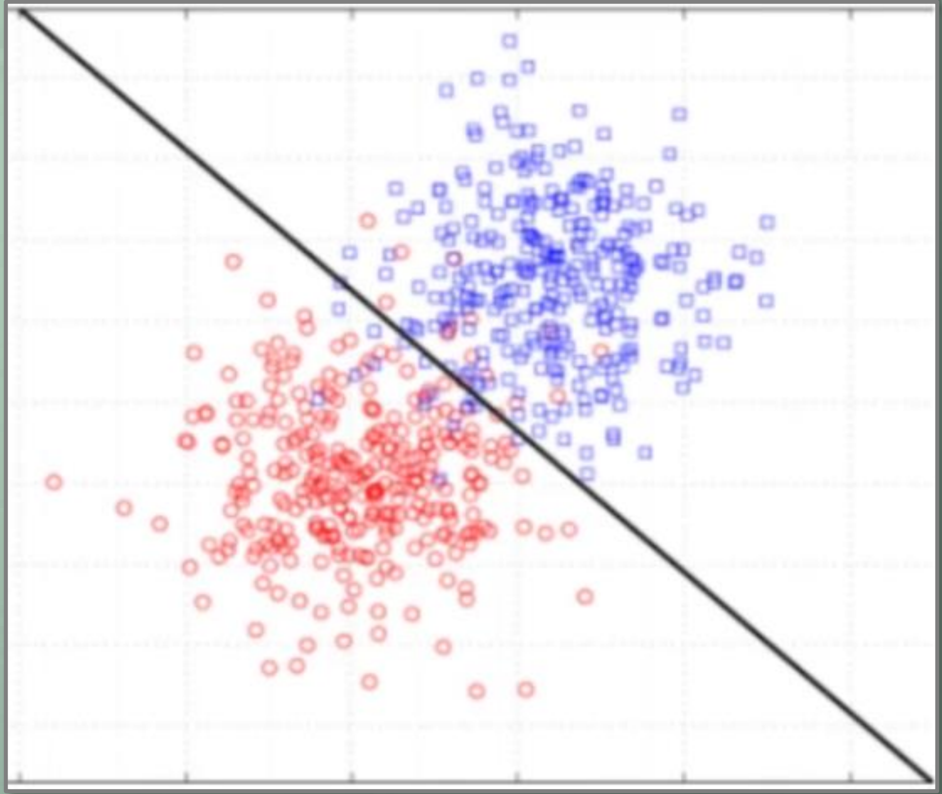
Probability Estimation

- Logistic Regression predicts the probability of occurrence of event by fitting data to logistic function.
- Trained model estimates probability that given instance falls into one of these two categories.
- Linear discriminant is trained in a way that errors, i.e. squared distances to each data point are minimized

Model Application

- Credit scoring
- measuring disease prevalence
- Any generalization of data that has binary target feature

Separation of data points with 2 features using linear discriminant



Centerpiece of logistic regression

… is linear combination of features (as in linear regression) projected on inverse logit curve

Linear combination of features

$$b_0 + b_1 x_1 + \cdots + b_n x_n = \omega^T x$$

Logit (Logged-odds) equation

$$\ln\left(\frac{P}{1-P}\right) = \omega^T x$$

Inverse of logit equation by P

$$P(y=1|x) = \frac{1}{1+exp(-\omega^T x)}$$

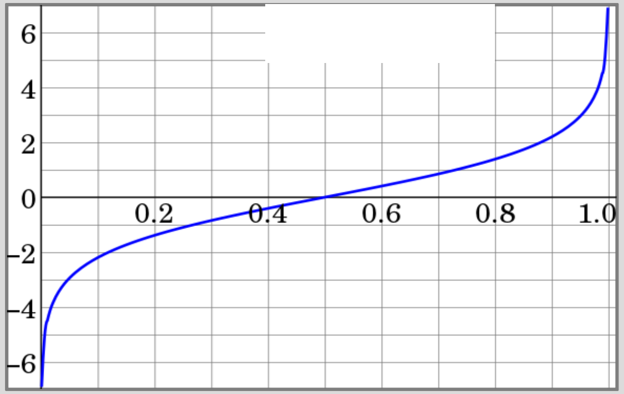
Abbreviated notation as sigmoid function

$$P(y=1|x) = \sigma(\omega^T x)$$

Last step for binary classification: Rounding to 0 or 1

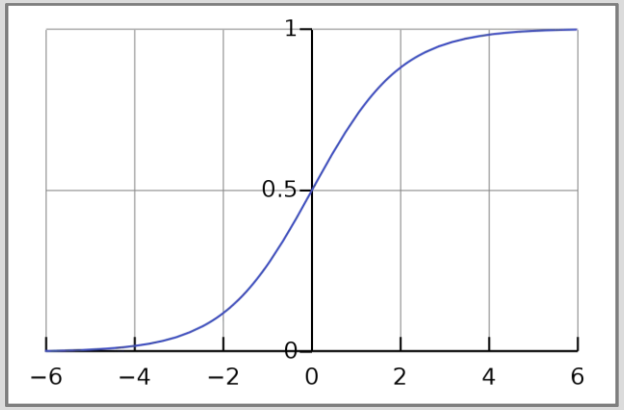
Logit equation

$$\ln\left(\frac{P}{1-P}\right) = \omega^T x$$



Sigmoid function: Inverse of logit equation

$$P(y=1|x) = \sigma(\omega^T x)$$



Backpropagation (of errors)

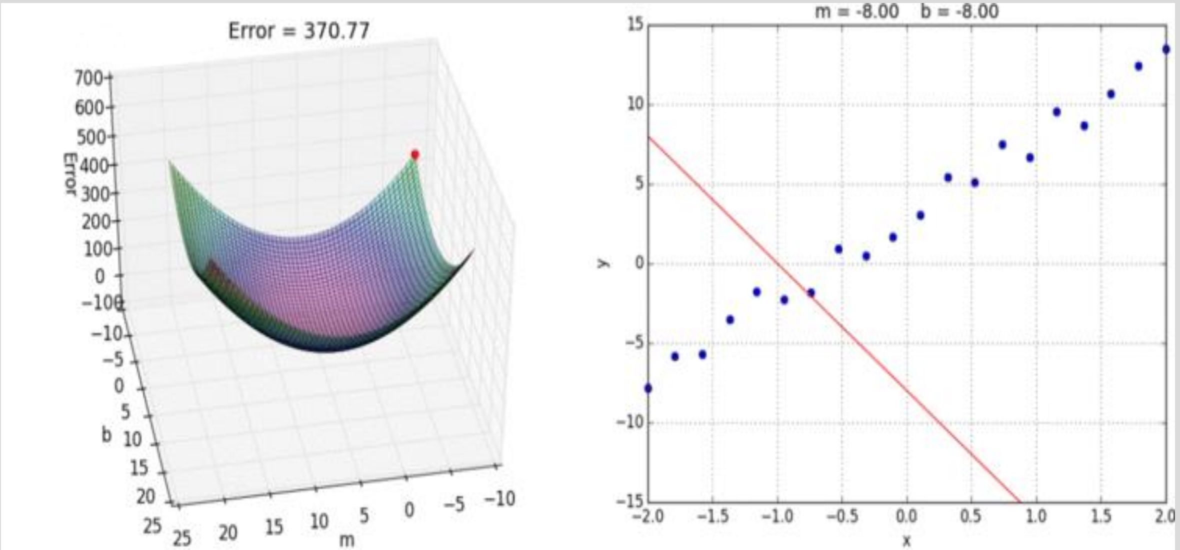
… is general optimization approach for minimizing error in supervised or reinforcement learning.
It calculates gradient of error function with respect to each weight by using chain rule.

Process

1. Feedforward Pass:
   Input a dataset and compute the predicted output using current weights.
2. Compute Loss:
   Calculate error, i.e. difference: predicted output ⇔ actual target.
3. Backward Pass:
   Propagate loss backward through network, calculating gradients for each weight.
4. Weight Update (using Gradient Descent):
   Once computed gradients are used to update weights of network.
   Aim: to adjust weights in direction that minimizes error.
5. Iterate 1–4

Challenges

- Can get stuck in local minima
  (mitigated by variants like stochastic gradient descent).
- Sensitive to hyperparameters and initial weight values.



Fit the model with L-BFGS in Sklearn

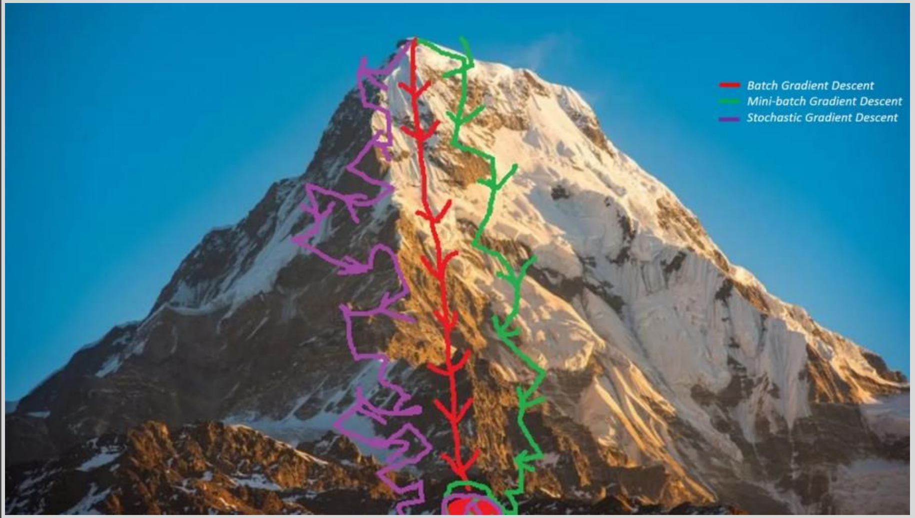
It looks for best parameters by navigating the cost function landscape, aiming to find the lowest point (global minimum)

Overview

- Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm.
- Iterative method for solving large-scale nonlinear optimization problems.
- Standard solver for logistic regression models in Python / Sklearn

How L-BFGS Works

- Approximation:
  It approximates Broyden-Fletcher-Goldfarb-Shanno (BFGS) Hessian matrix (particular second derivation matrix) using limited amount of memory.
- Updates:
  Utilizes previous gradient evaluations to construct approximation, which is then used to determine direction of next step.
- Iteration:
  Each iteration consists of line search followed by calculation of next approximation.



Validation of results

… is done by comparing prediction results and test set target values

Classification and validation of results

1. Through the sigmoid function, features are projected onto the entry probability $P(y = 1|x)$ between $[0; 1]$
2. Binary classification: commercial rounding of sigmoid function results
3. Prediction results (0 or 1) are compared with test rest target values (0 or 1)
4. **Confusion matrix**: tabular depiction of correct / incorrect classification

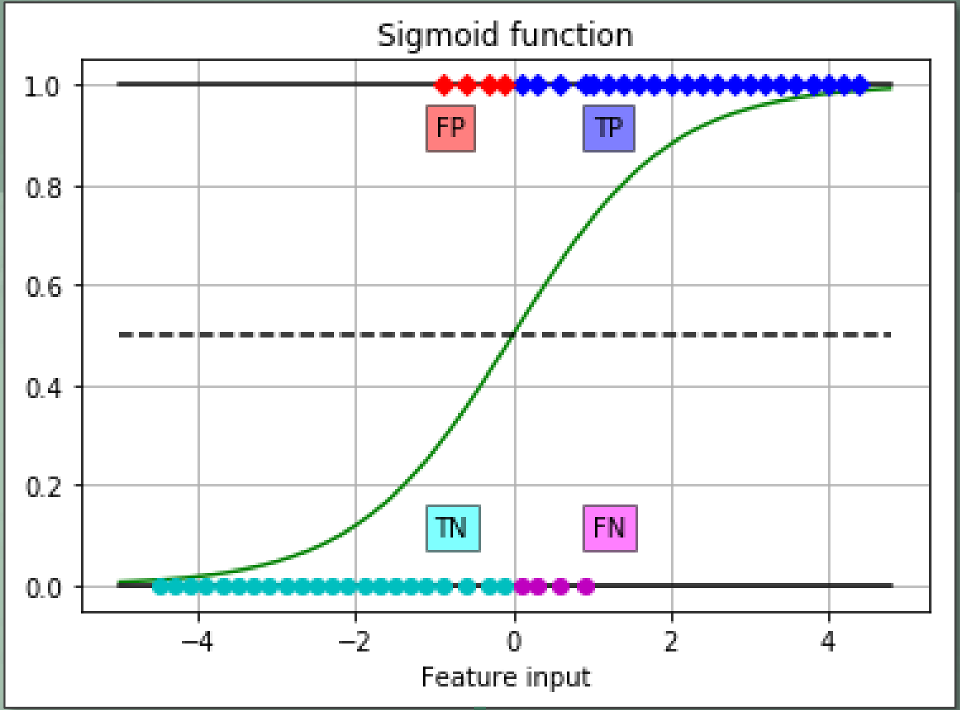
Correct vs. incorrect classification

- FP: False positives; FN: False negatives
- TP: True positives; TN: True negatives

→ Accuracy: TP + TN (Example: 80%)

Falacy of imbalanced classes

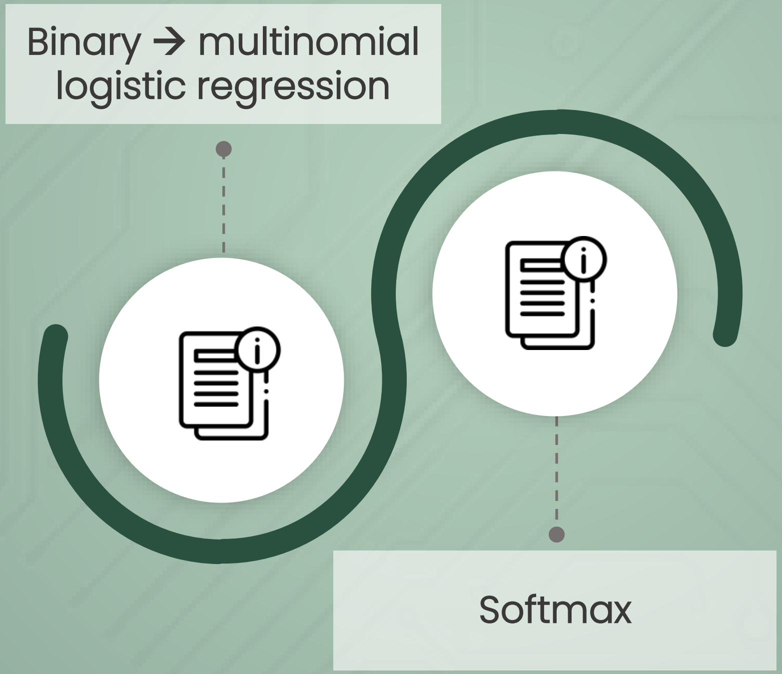
- Example: prob. of sunny weather in Greece in July → 95%
- In case outcome of 1 or very probably, "accuracy" of 95% is not good but bottom line
- Check additionally "False positives"



Example	Negative	Positive
Predicted negative	True negative (TN): 35%	False positive (FP): 15%
Predicted positive	False negative (FN): 5%	True positive (TP): 45%

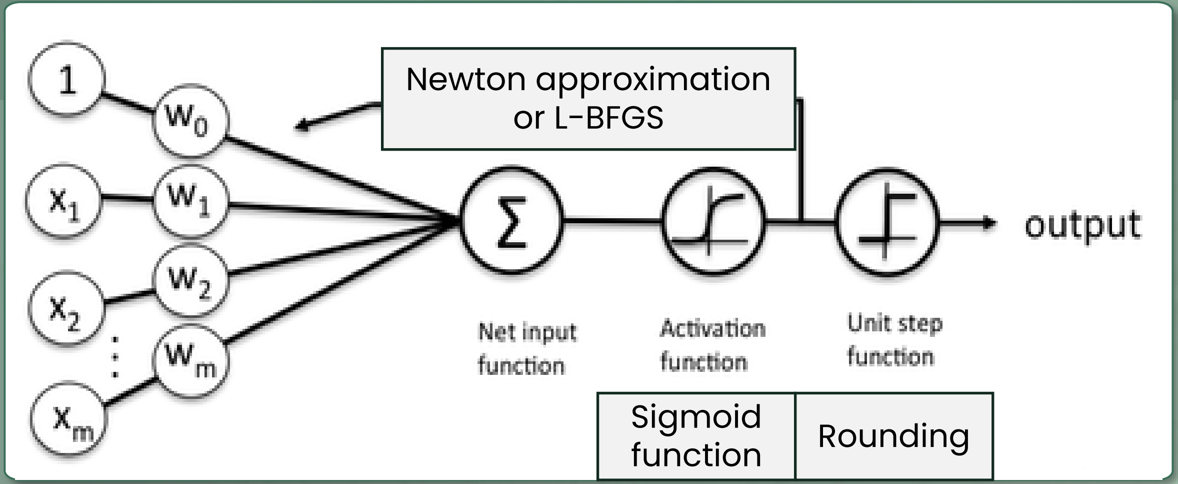
Multinomial Logistic Regression

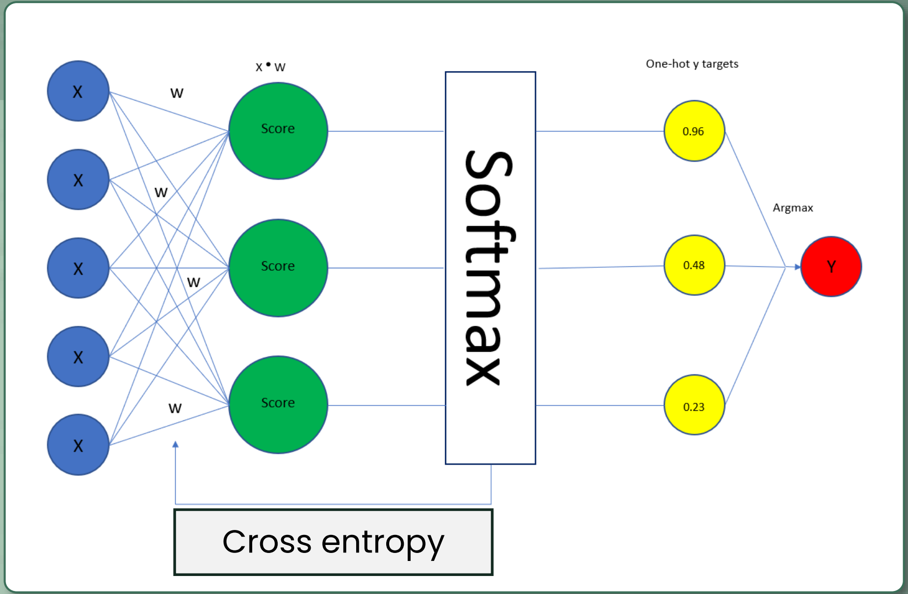
… is statistical method used for predicting, i.e. generalizing outcomes of categorical dependent variable based on multiple independent variables



From binary to multinomial logistic regression

Instead of sigmoid function with rounding softmax is used in order to match





	Binary	Multiclass	Focus
NUMBER OF OUTCOMES	2	3 or more	
PROBABILITY ESTIMATION	Sigmoid function	Softmax function	
DECISION PROCESS	Rounding probabilities	choosing the class with the highest probability	

Softmax

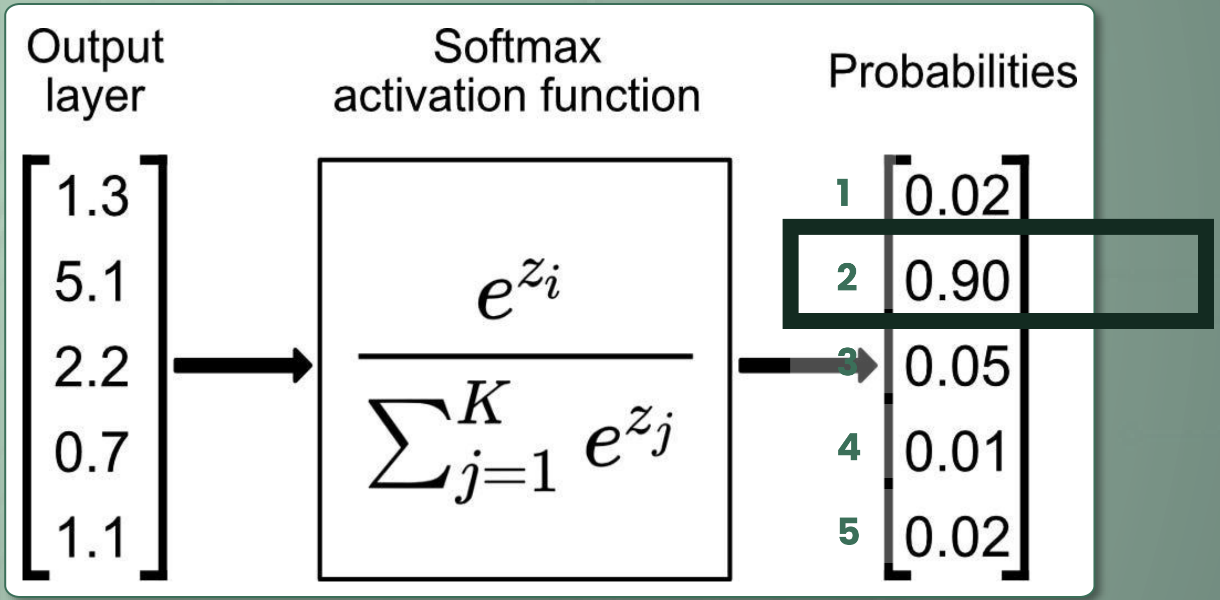
… is activation function that turns numerical output of n levels into probabilities that sum to one. Decision approach in multinomial logistic regression

Purpose:

- Transforms vector of K real numbers into probability distribution over K possible outcomes.
- Class / level with highest probability is selected

Why exponentiation: e^{z_i}

- Normalization:
    - Each value in vector is normalized
    - Negative values become positive
    - Division by sum make that probabilities sum up to 1
- Particular class / level becomes dominant compared with others



Selected level due to highest probability:
2

Finding new customers, approach 1: Multinomial Logistic Regression

Original and enlarged predicted clusters are compared

Prediction of new customers per cluster

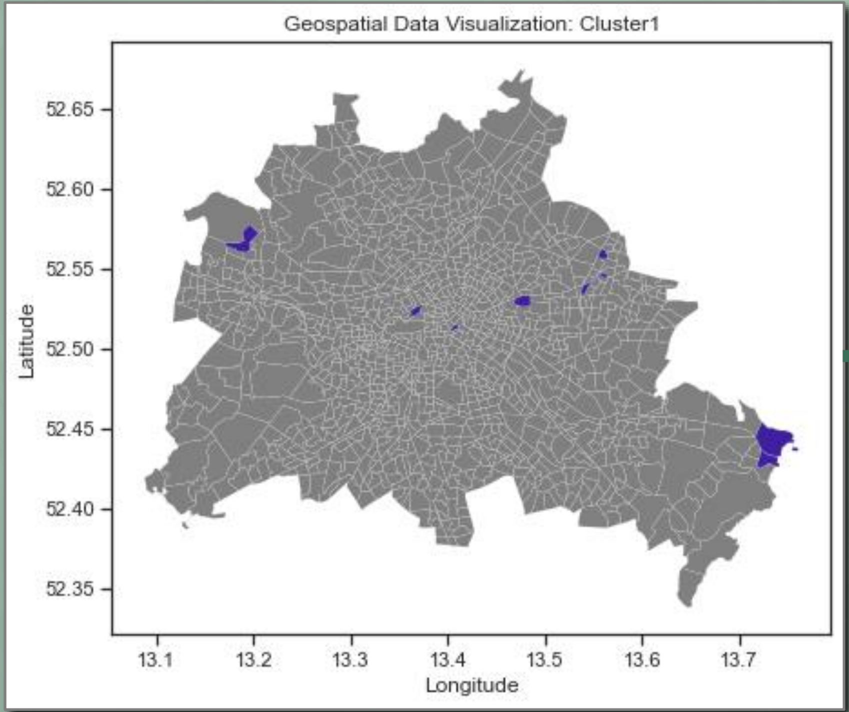
Cluster 0 as residual cluster

Prediction of new customers in cluster "educated urban families"

Obviously geo-visualization depicts implausibility both of existing and of predicted customers, i.e. it shows that underlying data is artificial & arbitrary

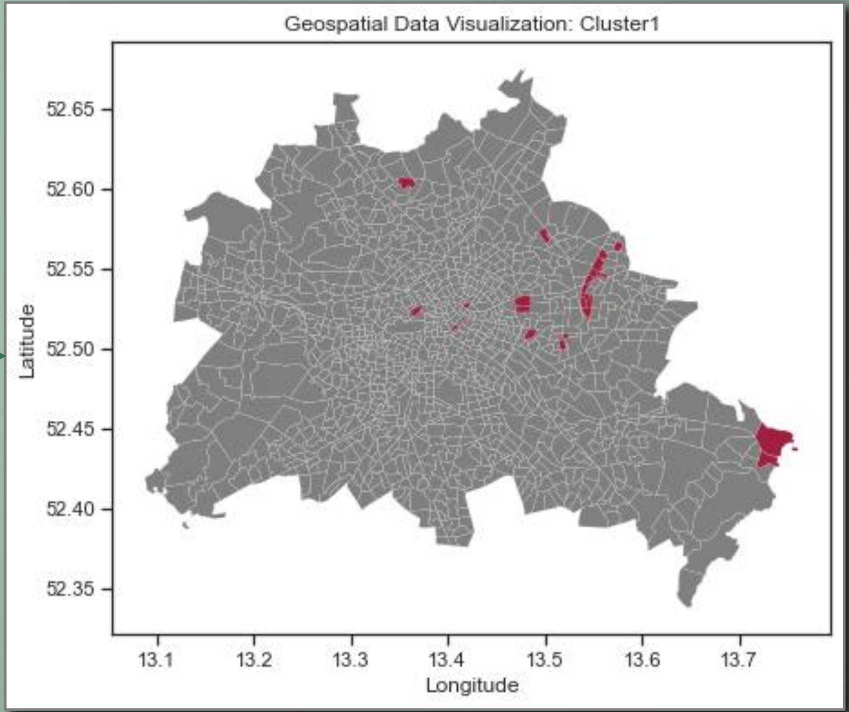
Cluster 1:
Educated urban families

- People with higher education degree
- Families with kids
- Living in larger buildings, urban
- Being interested in social media
- Travel preference: Beach escape



Existing customers

8 Polygons



Predicted customers

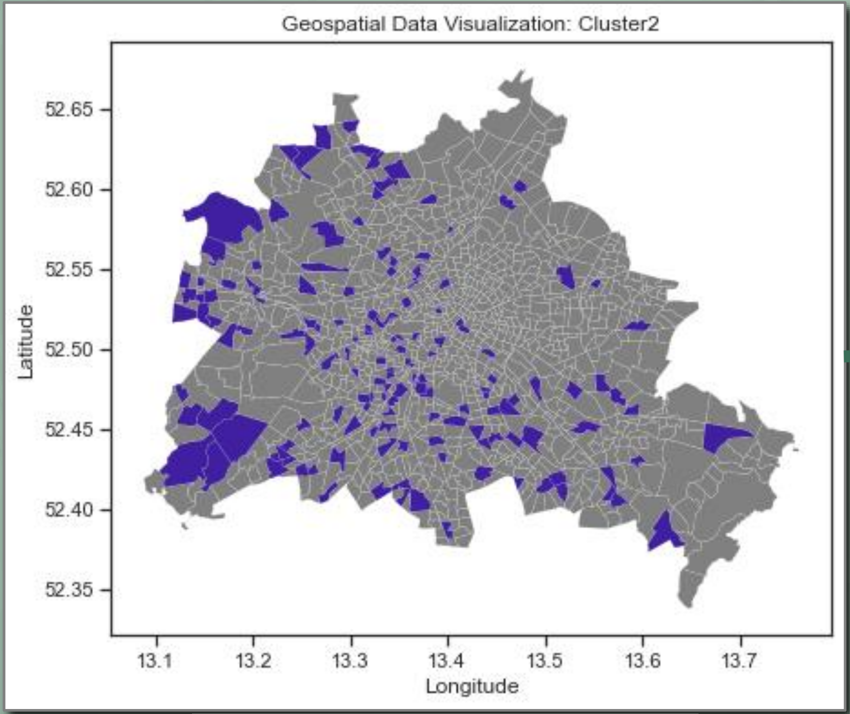
23 Polygons ↑

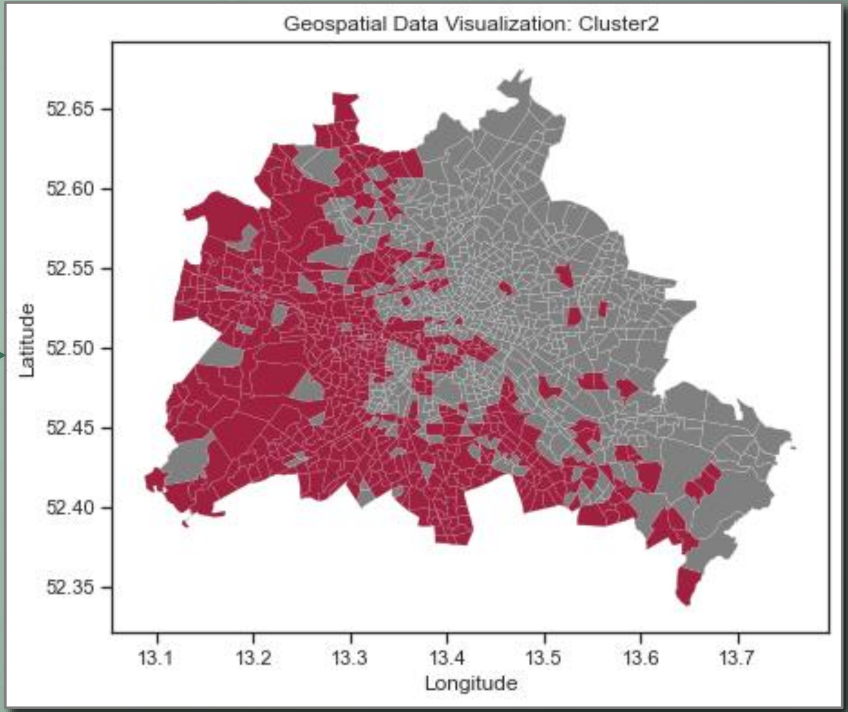
Prediction of new customers in cluster “brand aware working class”

Large parts of city are predicted to be typical places for cluster 2 people. Again implausibility of prediction results from artificiality of underlying data

Cluster 2:
Brand aware working class

- Working class
- Interested in photovoltaics, travelling
- Brand and quality aware
- Travel preference: Cruise gateway





Existing customers

150 Polygons

Predicted customers

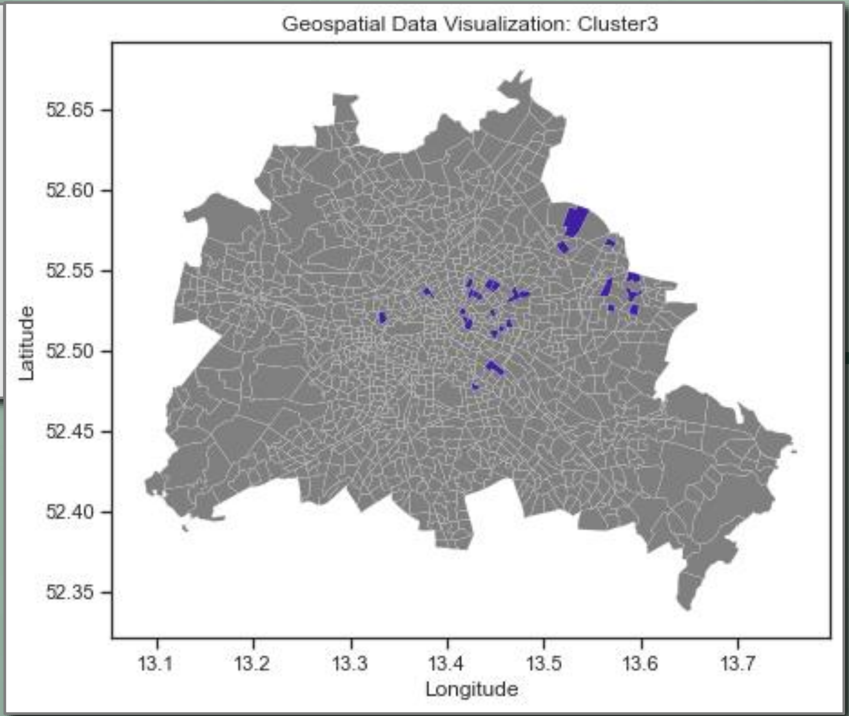
531 Polygons ↑

Prediction of new customers: “Upper-class Families interested in sports”

New cluster 3 type customers are predicted in vicinity of existing customers

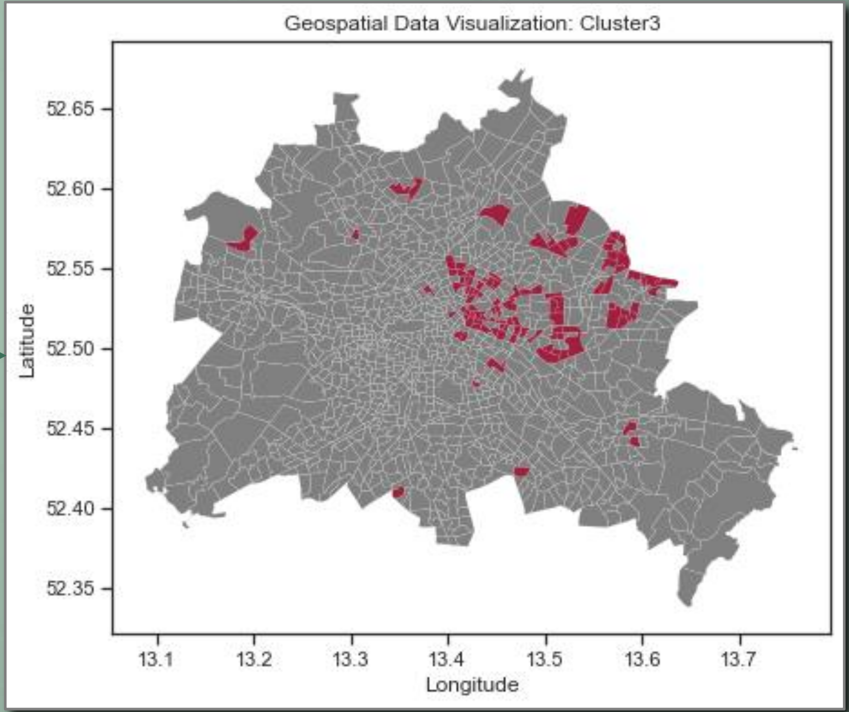
Cluster 3:
Upper-class Families interested in sports

- People with college degree
- Families with kids
- Living in small houses of higher status
- Interested in sports
- Travel preference: City Break



Existing customers

27 Polygons



Predicted customers

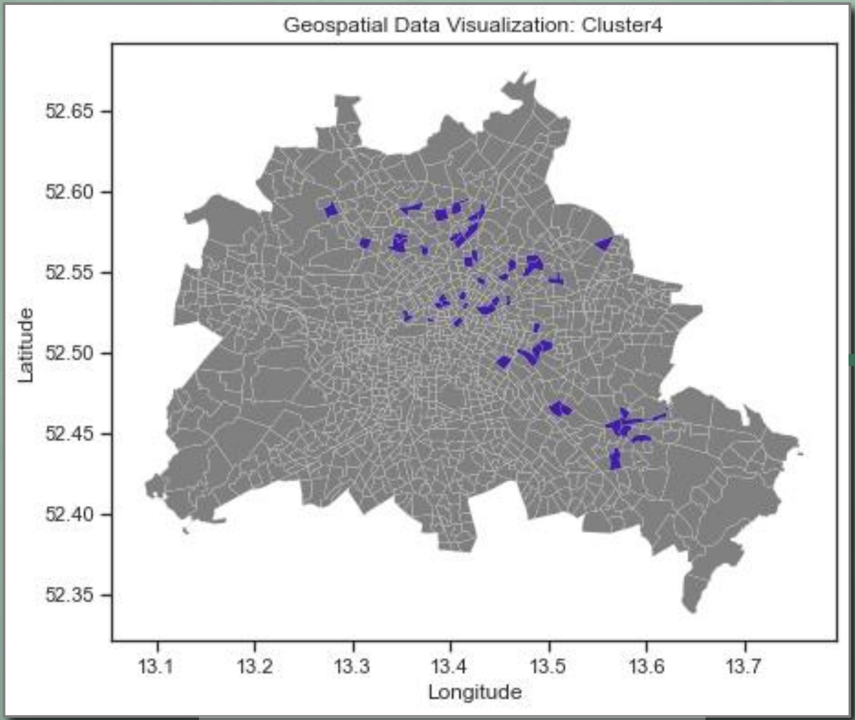
107 Polygons ↑

Prediction of new customers: “low-income suburbian families”

Predicted cluster 4 customers increase scope tremendously

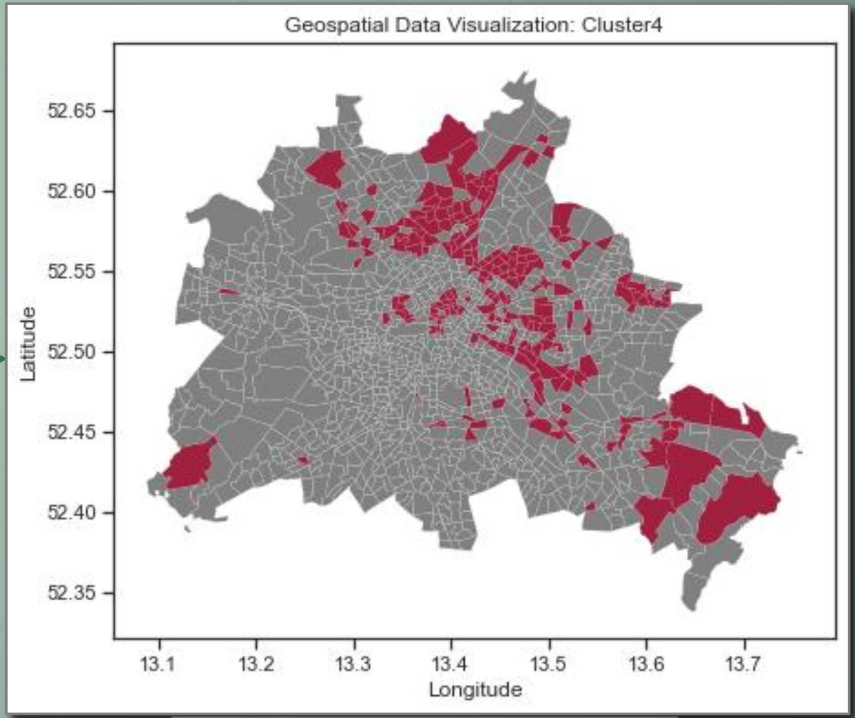
Cluster 4:
Low-income suburbian families

- Low population density, i.e. suburbian
- Families with kids
- Low price oriented, rather poor
- Travel preference: Mountain adventure



Existing customers

47 Polygons



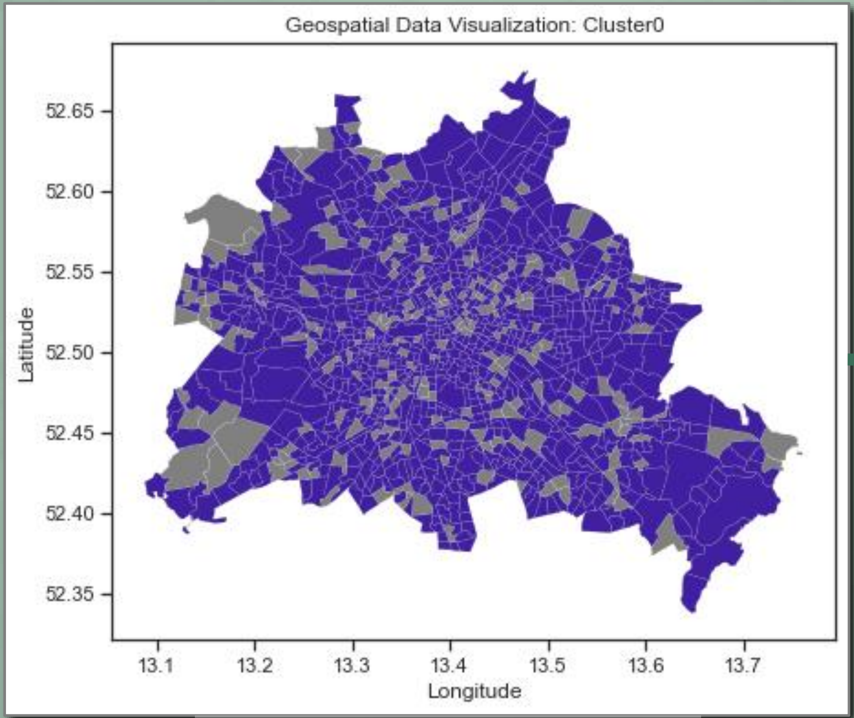
Predicted customers

221 Polygons ↑

Cluster 0 is residual cluster with so called “average Germans”

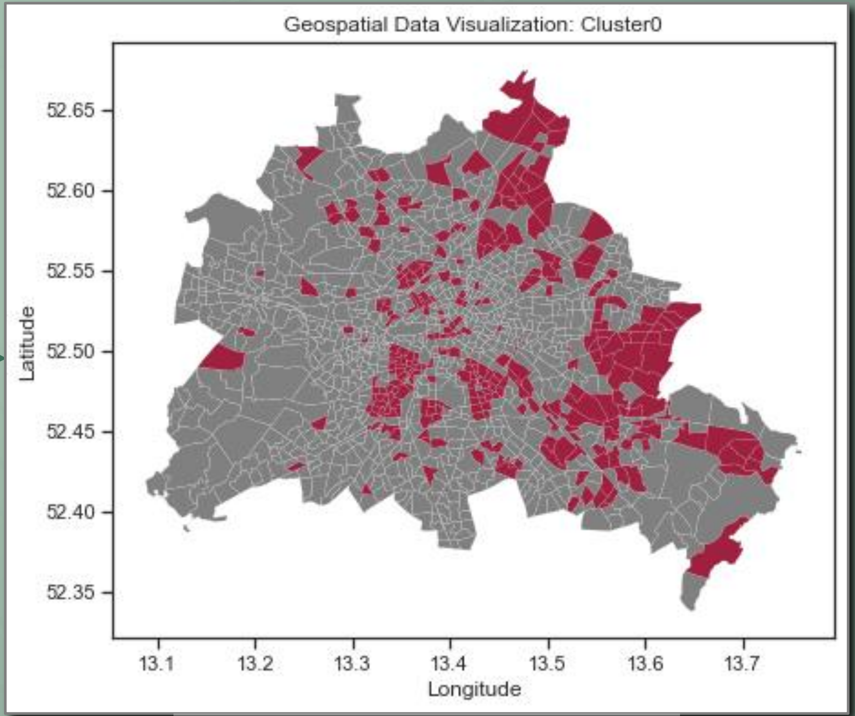
… Fields that are not assigned to any cluster after multinomial prediction are considered to be average, not having significant descriptive properties

Cluster 0 (residual cluster): Average Germans



Existing customers

950 Polygons

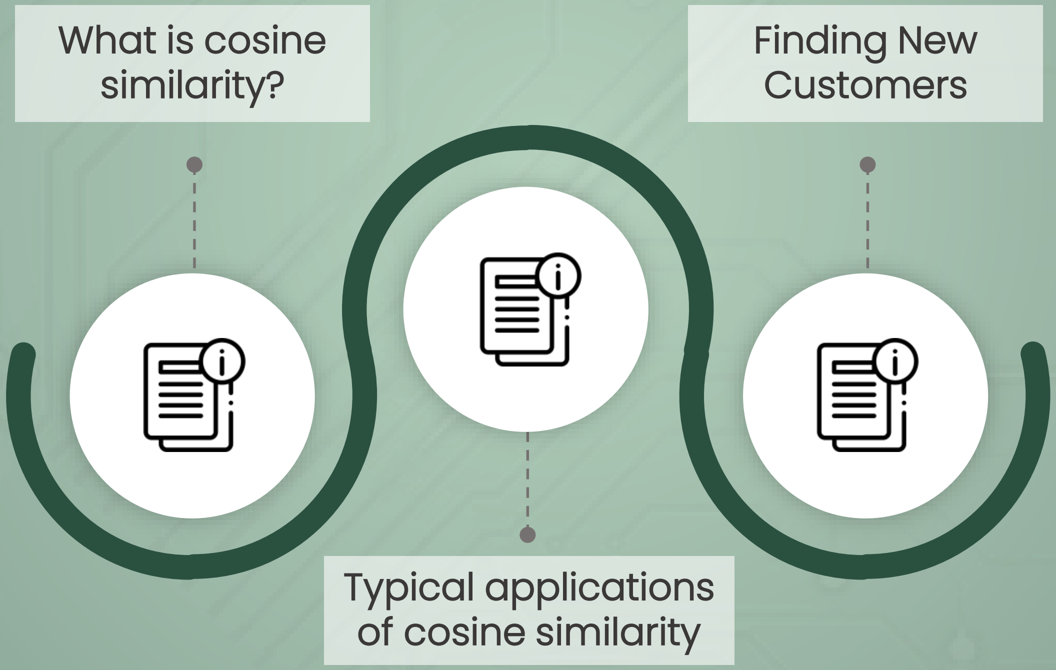


Predicted customers

300 Polygons ↓

Similarity analysis with cosine distance

… or cosine similarities are used to measure angular similarity between multidimensional vectors, commonly applied in text mining and recommender systems.



Cosine similarity

… is metric used to measure how similar two entities (documents, vectors, etc.) are irrespective of their size. It is not machine learning but frequently used together.

It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.

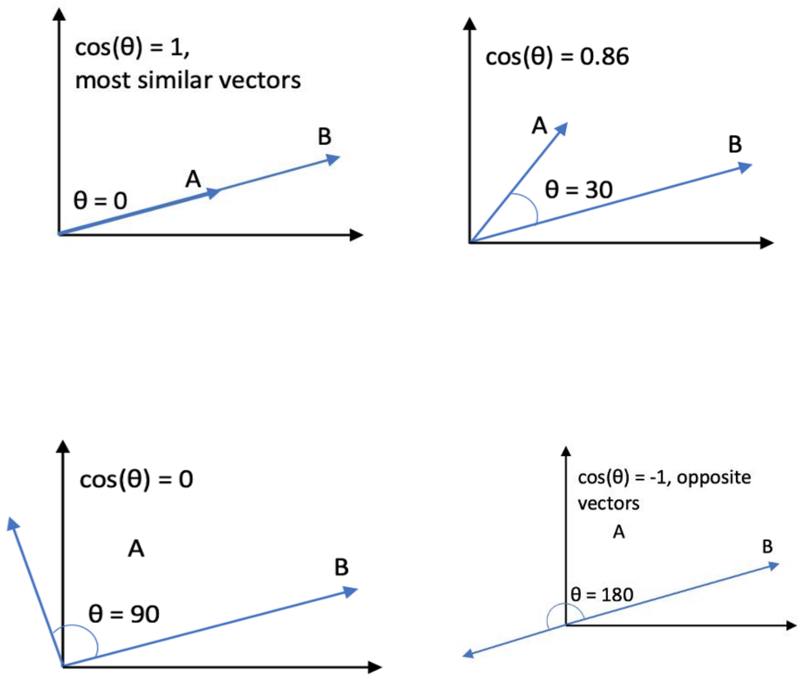
Formula

$\cos\boldsymbol{\theta} = \frac{\vec{a}\times\vec{b}}{\|\vec{a}\|\times\|\vec{b}\|}$ with $\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots a_n^2}$ and $\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + \cdots b_n^2}$

- $\vec{a}$ and $\vec{b}$ are vectors, $\vec{a} \times \vec{b}$ is their dot product
- $\|\vec{a}\|$ and $\|\vec{b}\|$: magnitudes i.e. lengths of vectors $\vec{a}$ and $\vec{b}$

Explanation

- Ranges from -1 meaning exactly opposite, to 1 meaning exactly the same
- 0 usually indicating independence
- in-between values indicating intermediate similarity or dissimilarity



Typical applications of cosine similarity

It is commonly used in text analysis for document clustering, recommender systems to match user preferences, in machine learning for feature similarity and semantic analysis.

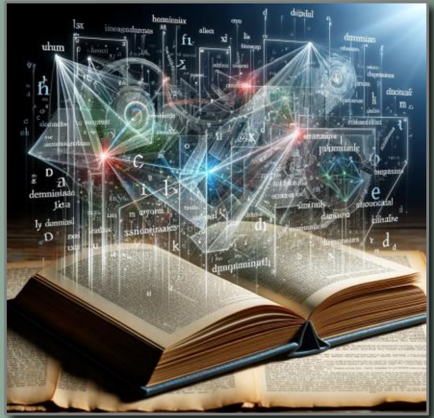
Text Analysis

- Plagiarism Detection:
  Identifying similarities between documents.
- Document Clustering:
  Grouping similar documents in data mining.

Focus

Machine Learning

- Feature Similarity:
  Measuring similarity in feature space for classification or clustering.
- Semantic Analysis:
  Understanding text data by comparing word or sentence vectors.

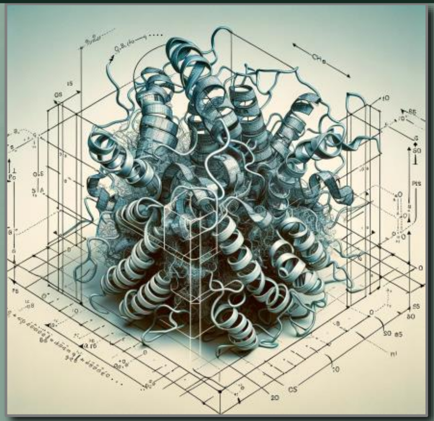


Recommender Systems

- Product Recommendations:
  Matching user preferences with product features.
- Content Recommendations:
  Suggesting similar articles, movies, or music.

Bioinformatics

- Gene Expression Analysis:
  Comparing gene expression patterns.
- Protein Structure Prediction:
  Analyzing similarity in protein structures.



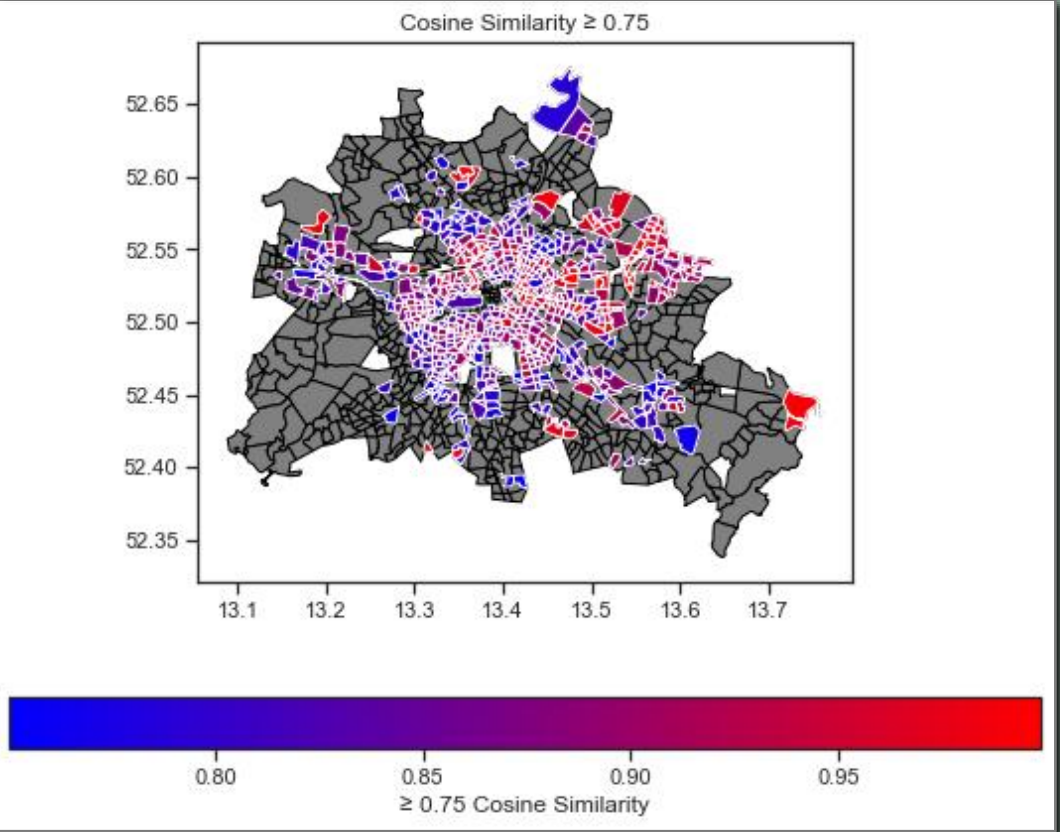
Finding New Customers using Cosine Similarities

… of sociodemographic geodata

Cosine similarity

- … helps in quantifying likeness in consumer profiles across different geographic regions or polygons
- … allows understanding and depiction of gradual differences in similarities
- … provides geo-locational focus to marketing campaigns

Cluster 1:	Educated urban families
Travel preference:	Beach escape



Index	description	z_1
0	Population ratio: school diploma	3.8453
1	Population ratio: college degree	4.01073
2	Population in the house	3.85619
3	Population ratio: university degree	3.46458
4	Tendency of having children	1.44134
5	Interest: social media	0.990565
6	House size: number of flats	0.601959
7	Interest: green energy	0.688072
8	Main purchase criterion: low price	0.752138
9	Population density (households per qkm)	0.275552
10	Interest: online shopping	0.111423
11	Status of Location	0.536042
12	Interest: politics	-0.0151308
13	Interest: football	0.0662808
14	Interest: travel	-0.265093
15	Tendency of having a photovoltaic system	-0.165109
16	Interest: economics	0.465909
17	Main purchase criterion: brand/quality	0.20326
18	Interest: alternative energy	-0.142215
19	PackageName_Cruise Getaway	0.160826
20	PackageName_Mountain Adventure	-0.157375
21	PackageName_City Break	-0.311798
22	PackageName_Beach Escape	0.202274
23	PackageName_Safari Expedition	0.115062

Finding new customers, approach 2: Cosine Similarities

Original clusters and gradual extensions are compared with alternative approach to supervised learning

Cluster solutions as basis of cosine similarity calculation

Geo-visualized comparison of similarities >95% and >75%

Cluster solutions as basis of cosine similarity calculation

Variant 1: small clusters, i.e. original clusters from segmentation

Variant 2: large clusters, i.e. predicted clusters from multinomial logistic regression (MLR)

Variant	Relationship to MLR	Residual Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	instead	950	8	150	27	47
2	Based on results, i.e. new, larger clusters	300	23	531	107	221

Here:
Small clusters used

Index	Cluster_1_simcount	Cluster_2_simcount	Cluster_3_simcount	Cluster_4_simcount
>70%	711	1178	1167	1178
>75%	611	1175	1110	1177
>80%	524	1157	884	1172
>85%	384	1132	704	1151
>90%	242	1068	587	1062
>95%	89	808	398	677

Similarities per cluster:

The lower the similarity, the more polygons

Educated urban families are among existing customers in few polygons

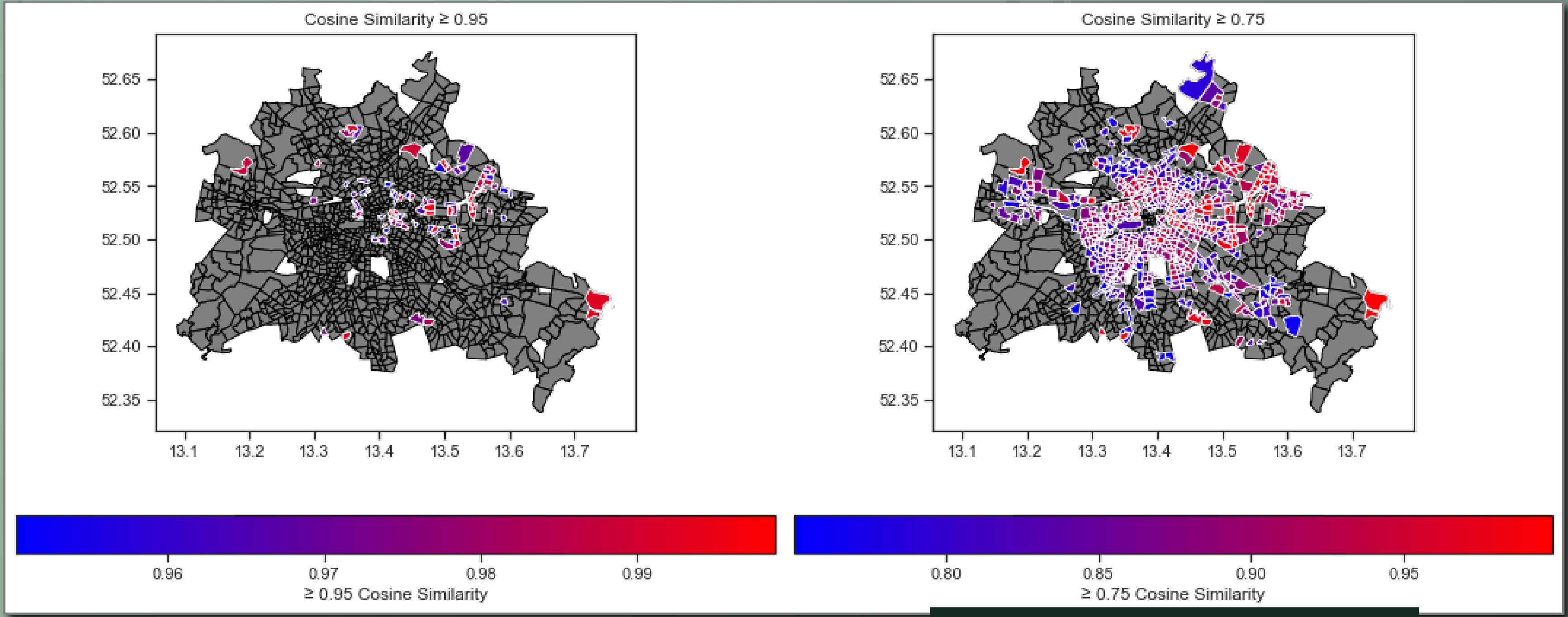
Based on cosine similarity potential customers can be found all over the city in different polygons. 75%-filter provides appropriate number of polygons to find new customers

Cluster 1:
Educated urban families

- People with higher education degree
- Families with kids
- Living in larger buildings, urban
- Being interested in social media
- Travel preference: Beach escape

Existing customers

8 Polygons



Similar customers >95%

89 Polygons

Similar customers >75%

611 Polygons

Many of existing customers are among brand aware working class

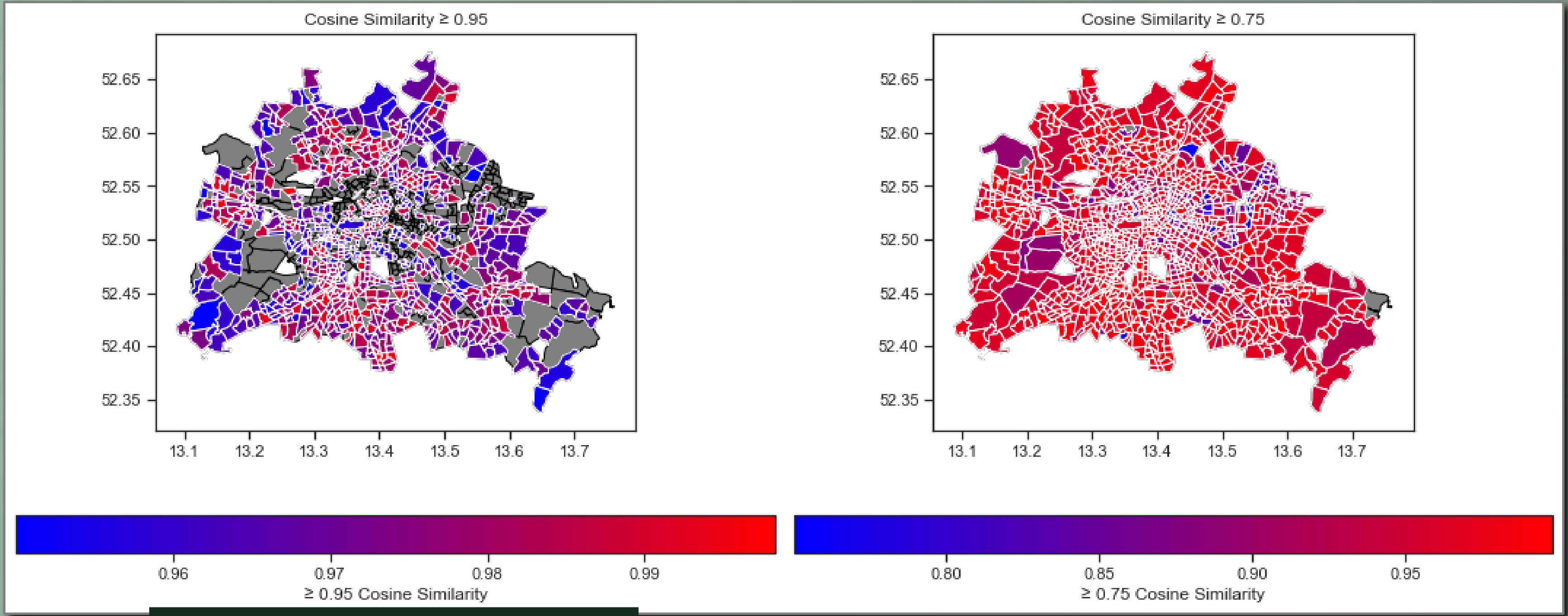
95%-filter provides already provides high number of polygons to generate leads

Cluster 2:
Brand aware working class

- Working class
- Interested in photovoltaics, travelling
- Brand and quality aware
- Travel preference: Cruise gateway

Existing customers

150 Polygons



Similar customers >95%

808 Polygons

Similar customers >75%

1175 Polygons

Rather few existing customers are among Upper-class families

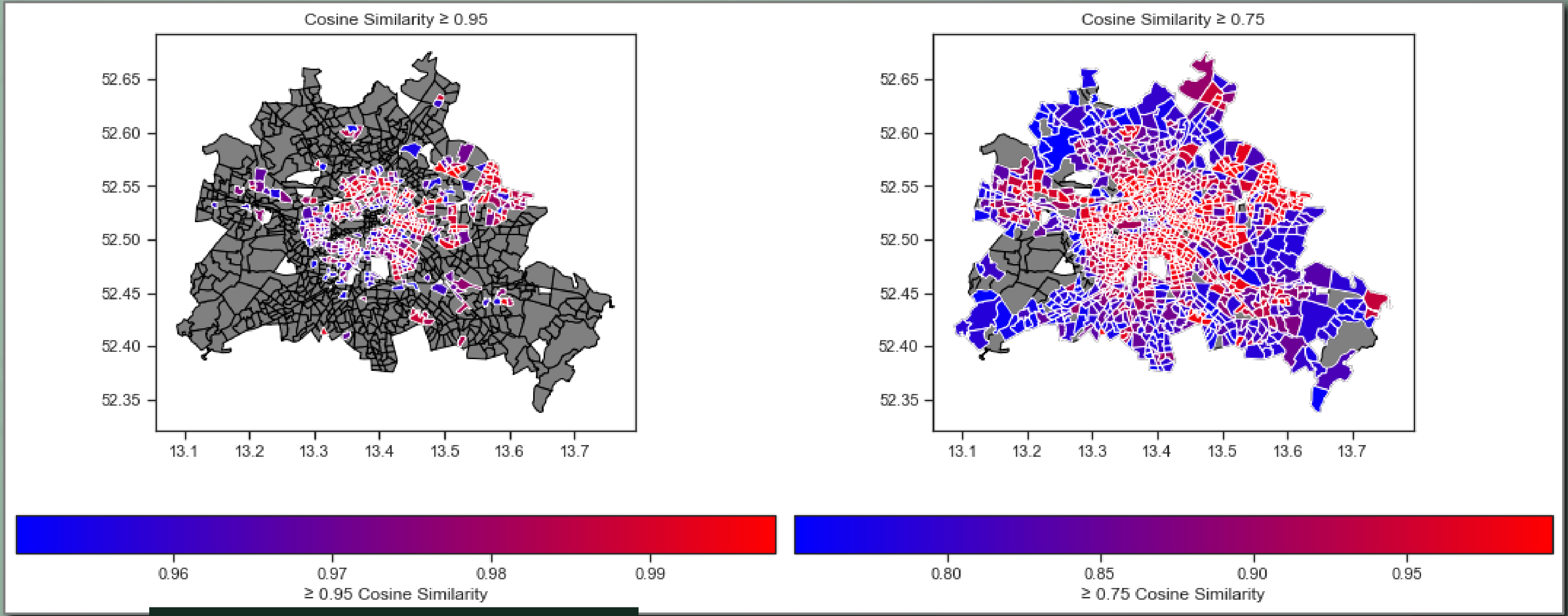
Nevertheless 95%-filter provides appropriate number of polygons to generate leads

Cluster 3:
Upper-class Families interested in sports

- People with college degree
- Families with kids
- Living in small houses of higher status
- Interested in sports
- Travel preference: City Break

Existing customers

27 Polygons



Similar customers >95%

398 Polygons

Similar customers >75%

1110 Polygons

Medium share of customers is among Low-income suburbian families

However 95%-filter already provides appropriate number of polygons to generate leads

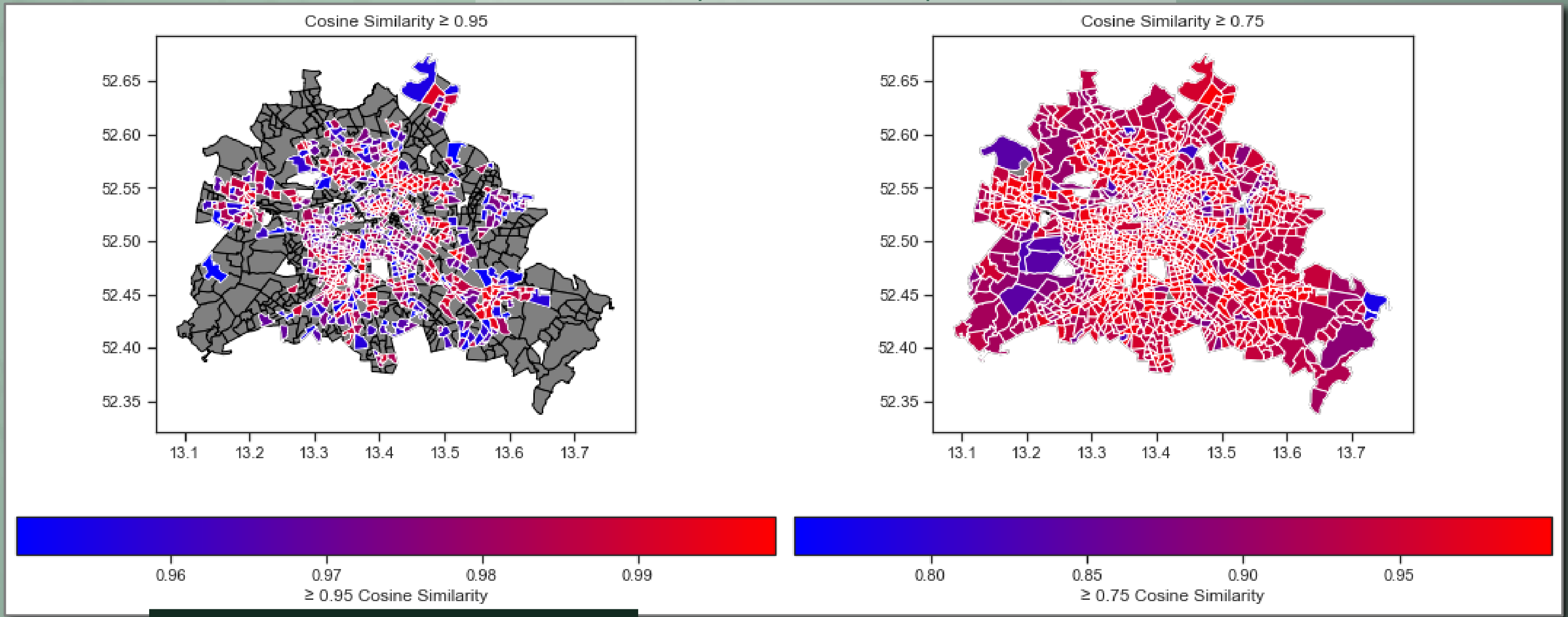
Cluster 4:
Low-income suburbian families

- Low population density, i.e. suburbian
- Families with kids
- Low price oriented, rather poor
- Travel preference: Mountain adventure

Existing customers

47 Polygons

Untypical locations for "suburbian families" are caused by artificial example data



Similar customers >95%

677 Polygons

Similar customers >75%

1177 Polygons

Links

i.e. sources for self-learning

	Title	Link
Logistic regression	Logistic Regression — Detailed Overview	https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
	SpeechandLanguageProcessing. DanielJurafsky&JamesH.Martin. Copyright © 2023. All rightsreserved. DraftofJanuary7,2023. LogisticRegression	chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://web.stanford.edu/~jurafsky/slp3/5.pdf
	Logistic Regression	https://www.sciencedirect.com/topics/computer-science/logistic-regression
	A Beginner's Guide to Logistic Regression	https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/
	Python ressource: sklearn.linear_model.LogisticRegression	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
	How to implement Logistic Regression from scratch with Python	https://www.youtube.com/watch?v=YYEJ_GUguHw
	Step by Step Tutorial on Logistic Regression in Python | sklearn |Jupyter Notebook	https://www.youtube.com/watch?v=bSXIbCZNBw0
	Logistic Regression 3 A Sentiment Example	https://www.youtube.com/watch?v=u_PiPYi4s7E

Links

i.e. sources for self-learning

	Title	Link
Multinomial logistic regression	Multinominal logistic regression, Part 1: Introduction	https://www.youtube.com/watch?v=JcCBIPqcwFo
	Week 4: Logit Model | Video 4: Multinomial Logit Model	https://www.youtube.com/watch?v=tIYjpmBtPq4
	605 Discrete Choice Conjoint Analysis with multinomial model in R	https://www.youtube.com/watch?v=ra8Y2FjRqOE
	Multinomial Logistic Regression	https://www.ibm.com/docs/en/spss-statistics/29.0.0?topic=regression-multinomial-logistic
	Multinomial Logistic Regression With Python	https://machinelearningmastery.com/multinomial-logistic-regression-with-python/

	Title	Link
Cosine Similarity	What is a cosine similarity matrix?	https://medium.com/acing-ai/what-is-cosine-similarity-matrix-f0819e674ad1
	Python: sklearn.metrics.pairwise.cosine_similarity	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
	How to Calculate Cosine Similarity in Python?	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
	How to Implement Cosine Similarity in Python	https://datastax.medium.com/how-to-implement-cosine-similarity-in-python-505e8ec1d823
	Cosine Similarity Explained Using Python	https://towardsdatascience.com/cosine-similarity-explained-using-python-machine-learning-pyshark-5c5d6b9c18fa

About me

Dr. Harald Stein

- Data Scientist ~ 8 years experience
- Algotrader ~ 4 years experience
- Ph.D. in Economics, Game Theory

- LinkedIn: https://www.linkedin.com/in/harald-stein-phd-1648b51a
- ResearchGate: https://www.researchgate.net/profile/Harald-Stein