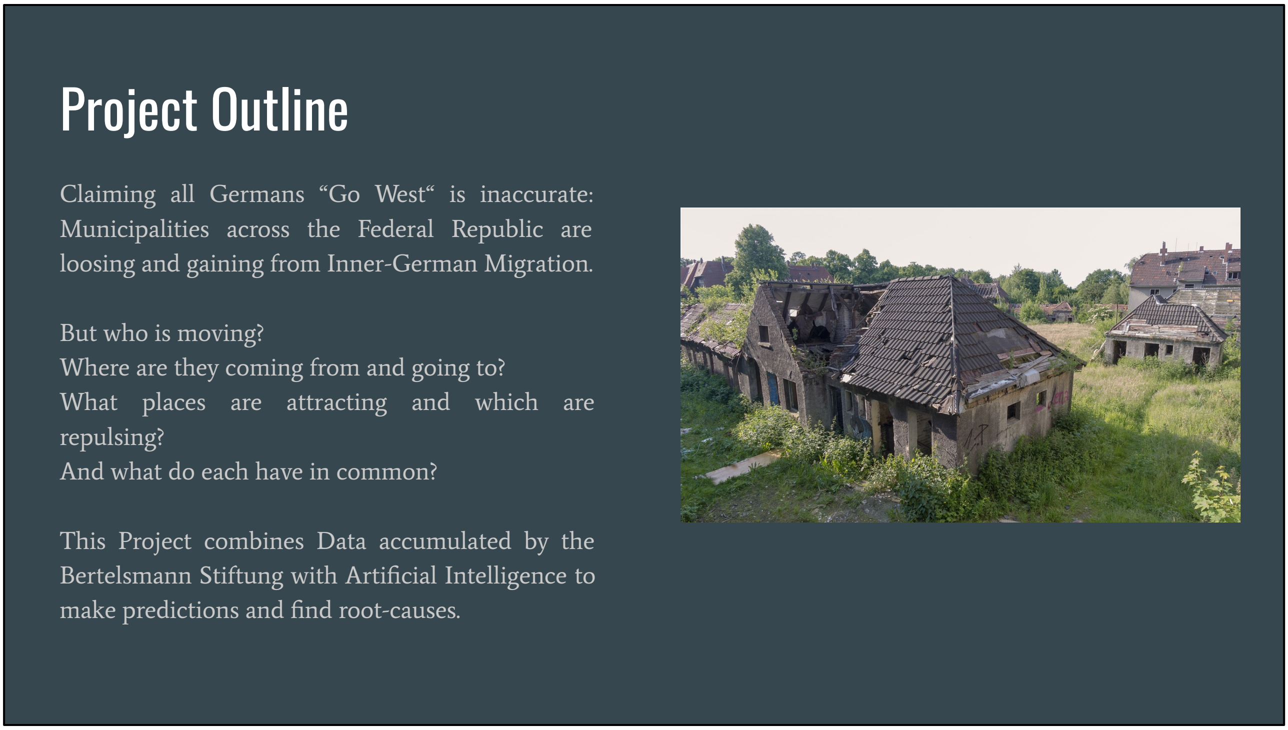
Where did the Village People go?

20th of January 2020



Project Outline

Claiming all Germans “Go West“ is inaccurate: Municipalities across the Federal Republic are loosing and gaining from Inner-German Migration.

But who is moving?
Where are they coming from and going to?
What places are attracting and which are repulsing?
And what do each have in common?

This Project combines Data accumulated by the Bertelsmann Stiftung with Artificial Intelligence to make predictions and find root-causes.

A quick outline what this Project is about.

The Structure

The Project

The Domain

The Analysis

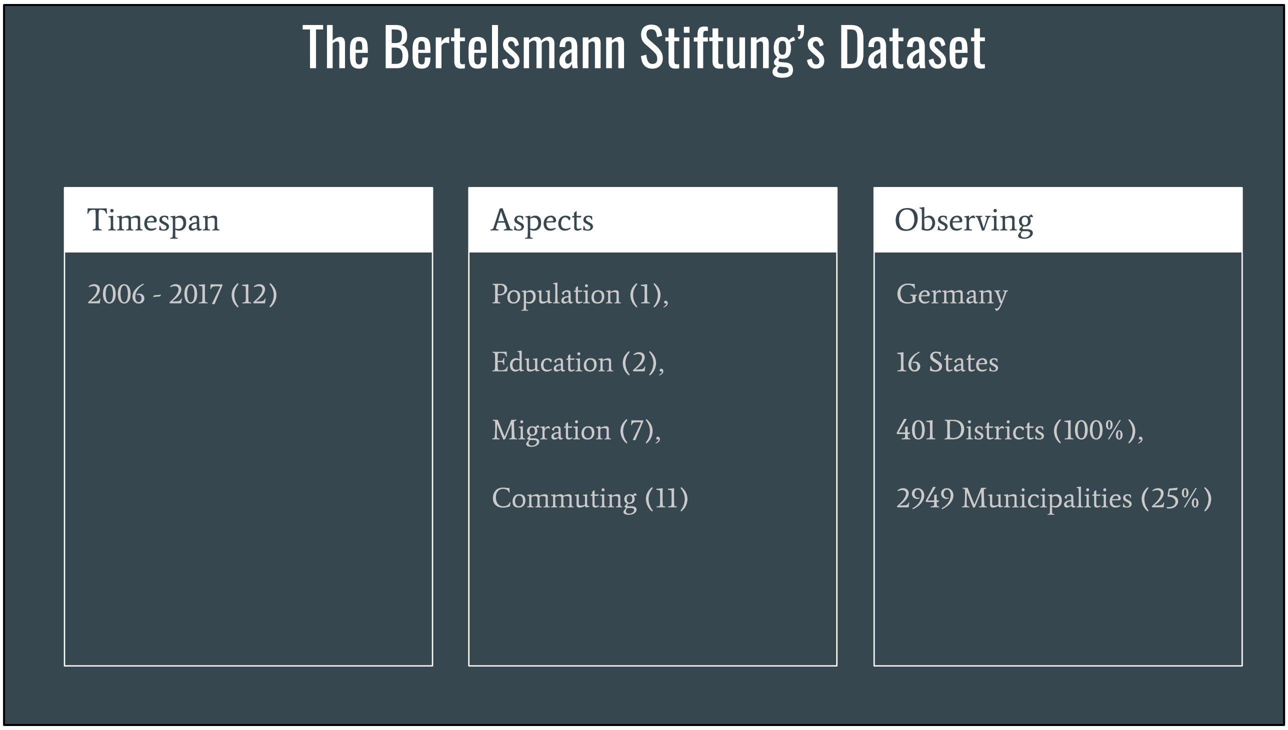
The Predictive Modelling

The Conclusion

How this Presentation is structured.

The Project

Let me begin by give you the general information on the project.



The Bertelsmann Stiftung’s Dataset

Timespan	Aspects	Observing
2006 - 2017 (12)	Population (1), Education (2), Migration (7), Commuting (11)	Germany 16 States 401 Districts (100%), 2949 Municipalities (25%)

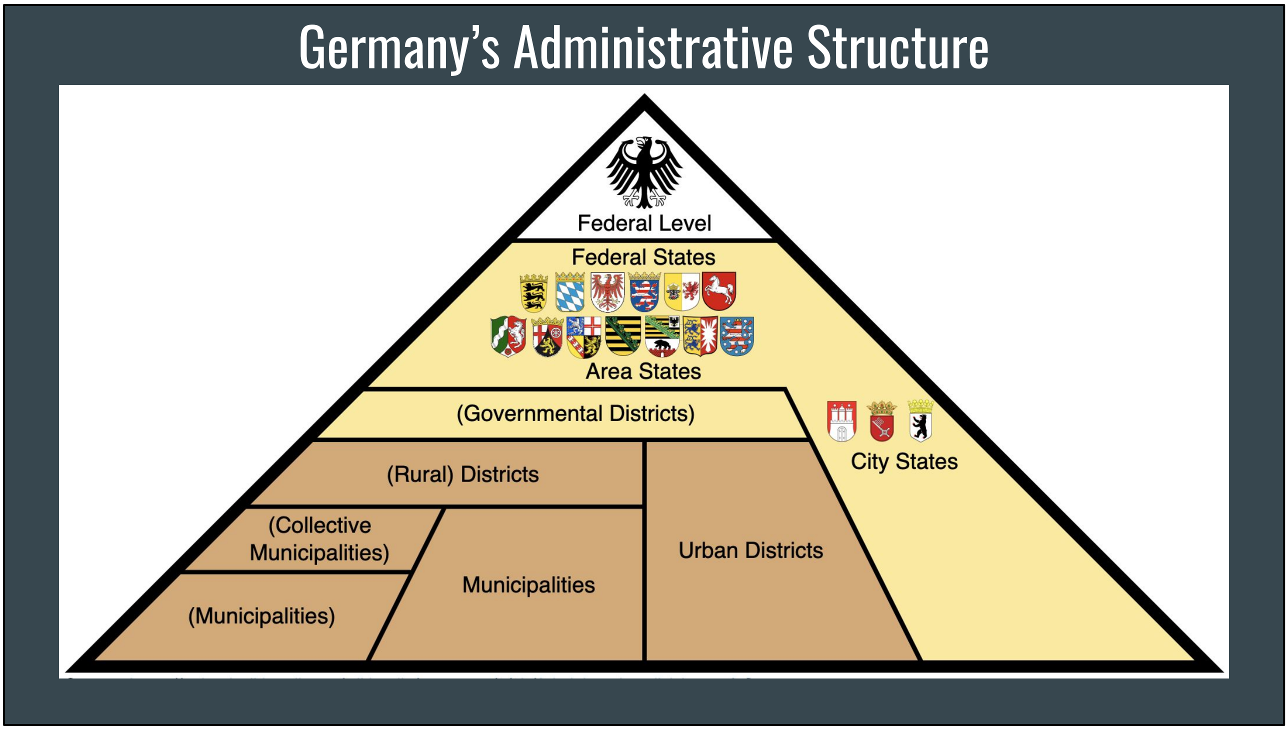
The DataSet I was provided by the Bertelsmann Foundation consists of Demographic Data on a National, States, District and Municipality Level. Only Data on Municipalities with more than 5000 inhabitants was collected.
The Data consists of the Number of Inhabitants, The Level of Eductaion, Information on the Migration, Commuting and spans the years 2006 - 2017.
(Details are in the Data Dictionary)

The Technology

I used the shown Libraries / Technologies for this project.
HTML was used in the Demonstrator.
GeoPandas and Folium do not have a fancy Logo yet.

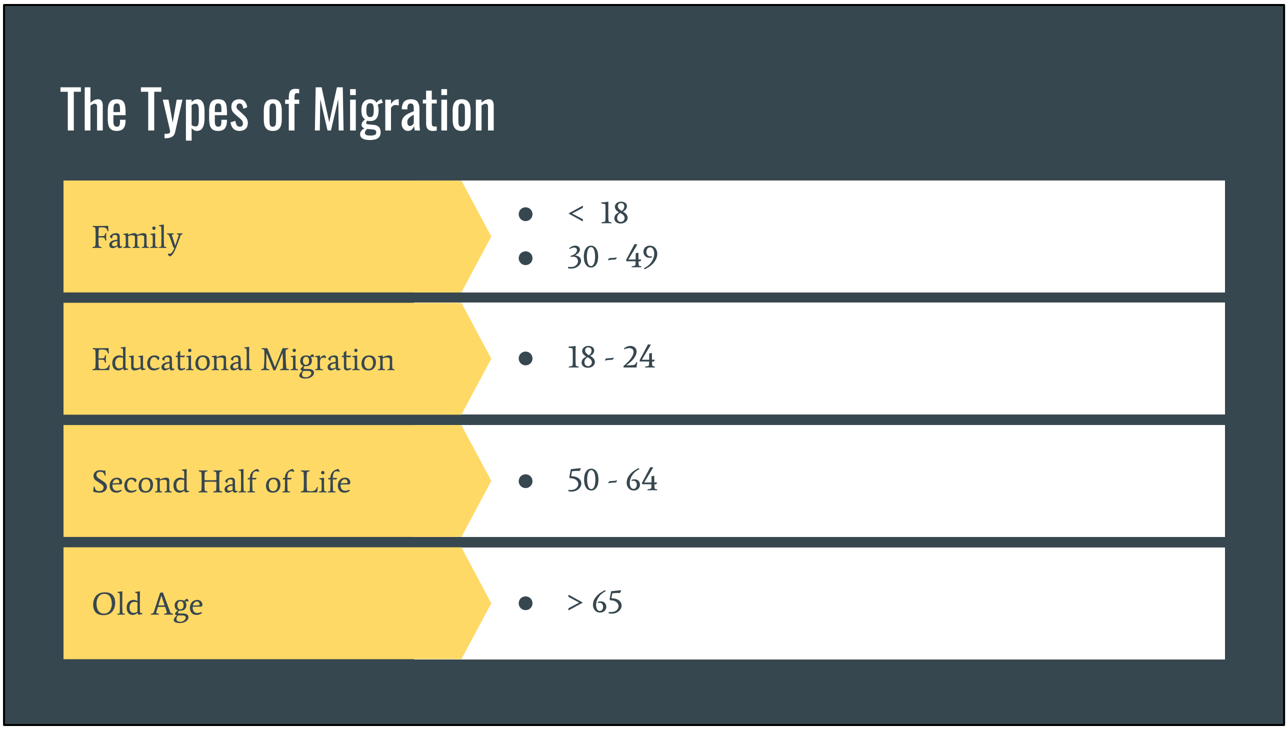
The Domain

Let me now give you some information on the topic.



Germany’s Administrative Structure

The German Federal Republic consists of 13 States and 3 City States.
Some of the States are further divided into Governmental Districts, they are not considered in this project.
All States have districts though. Even the City State of Bremen consists of two Urban Districts (Bremen and Bremerhaven).
Urban Districts are Cities with no further subdivisions, while Rural Districts consist of multiple Municipalities.
Some Municipalities decided to work togehter administratively and formed Collective Municipalities.



The Types of Migration

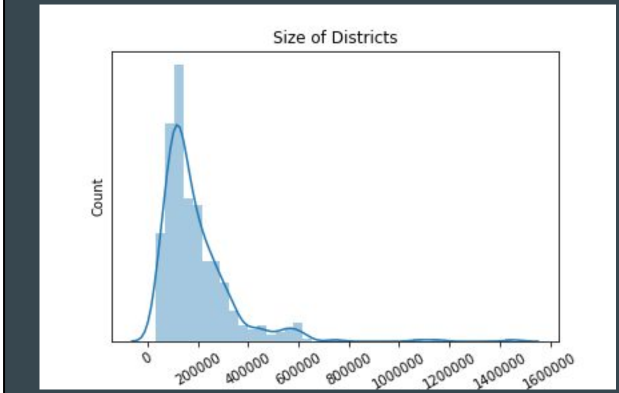
Family	● < 18 ● 30 - 49
Educational Migration	● 18 - 24
Second Half of Life	● 50 - 64
Old Age	● > 65

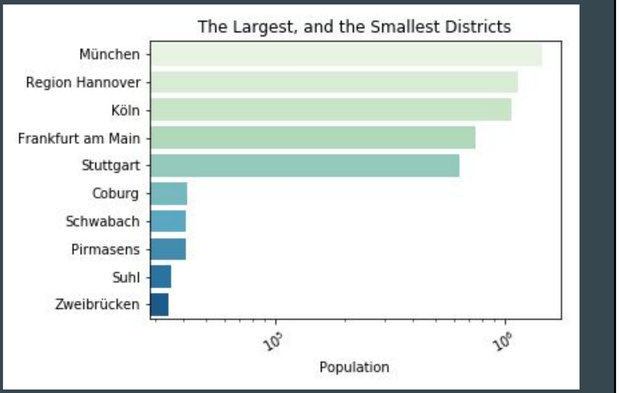
When discussing Migration the following ages are grouped.

The Analysis

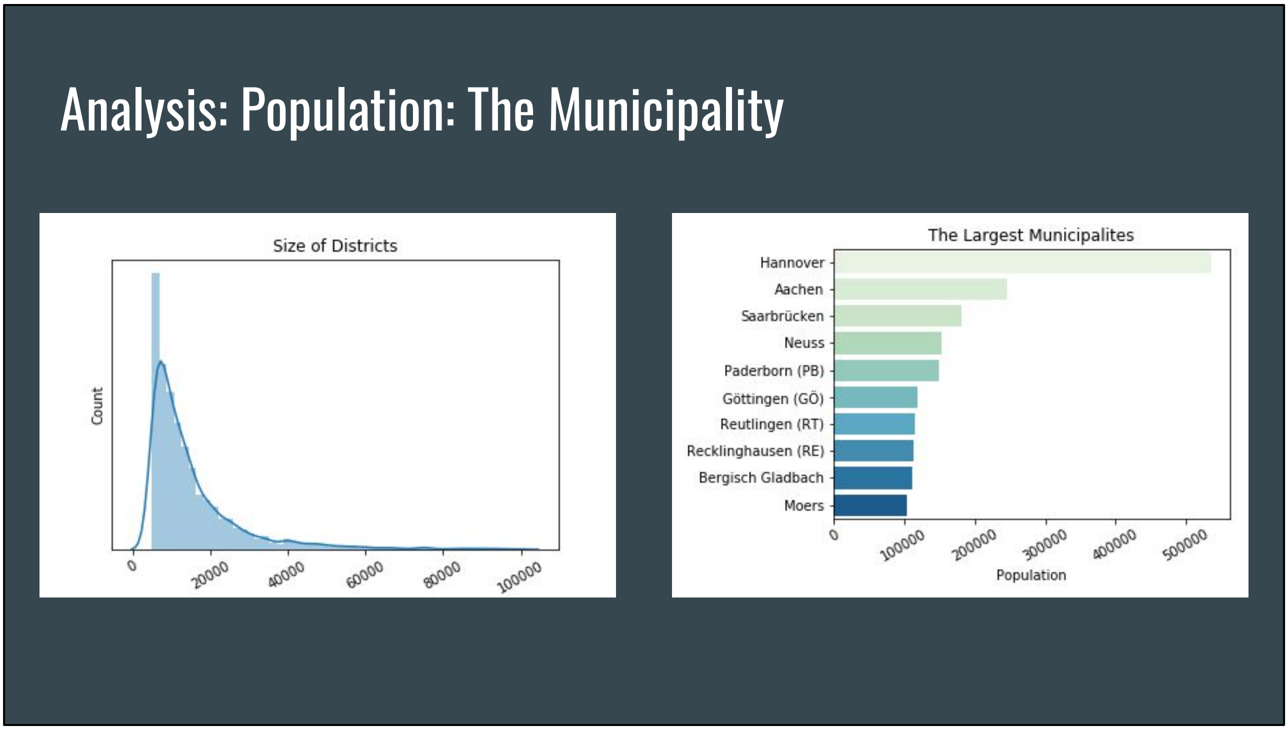
Let me now show you some of the results from analysing the Dataset. If not said otherwise, Data of the year 2017 is used.

Analysis: Population: The District





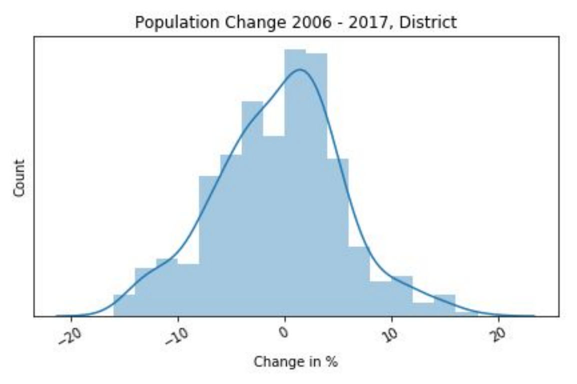
The average District size is close to 200.000 inhabitants. Districts bigger than 700k are rare.

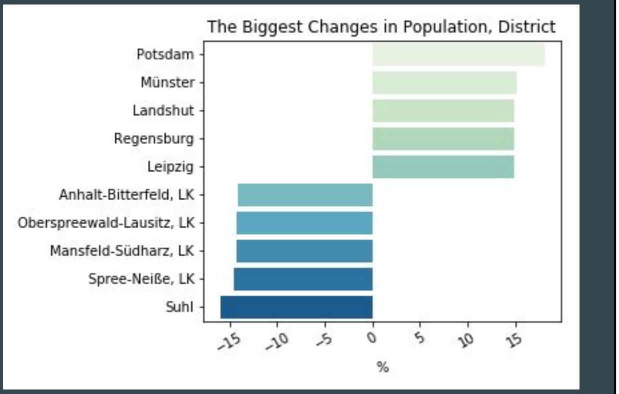


Analysis: Population: The Municipality

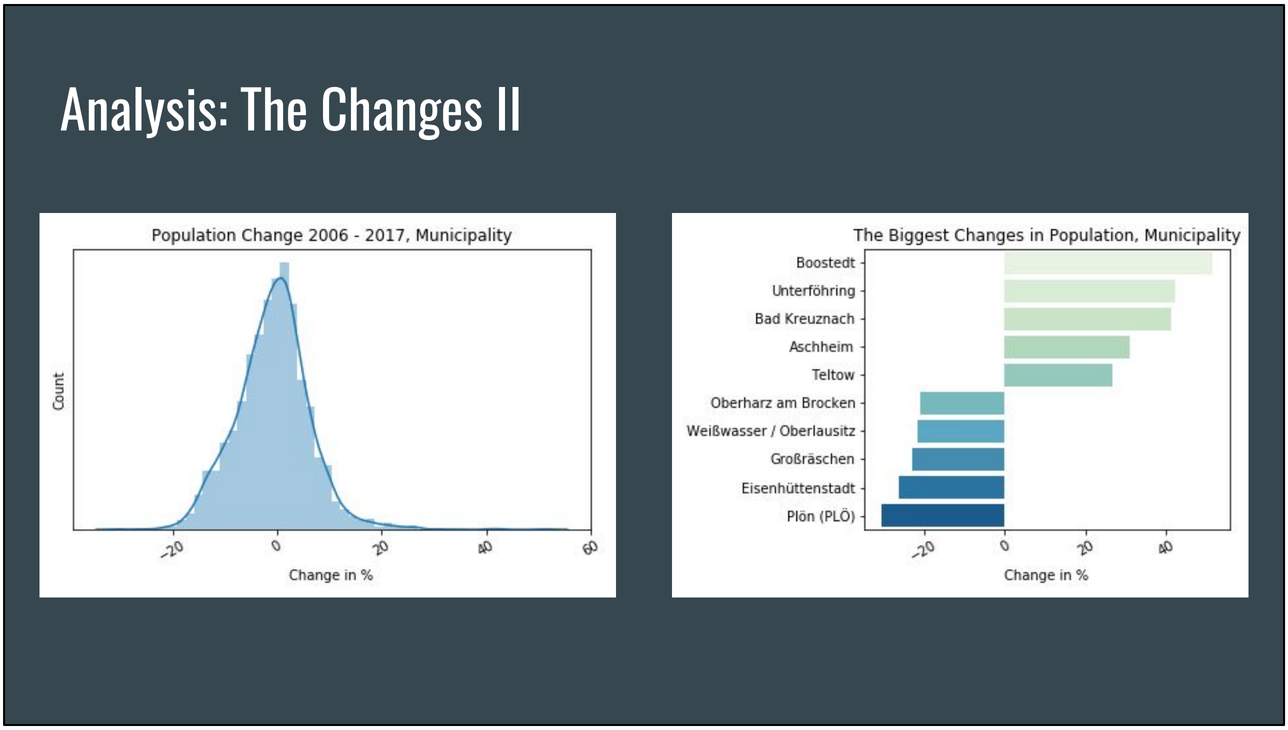
The Dataset cuts off at 5k inhabitants on the Municipality level. Therefore calculating a mean/ minimum is not representative.

Analysis: The Changes I





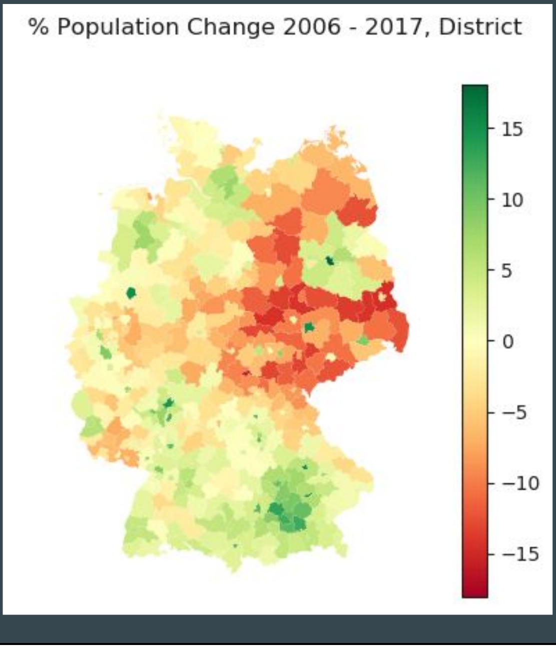
Generally Speaking, the changes in the 10 years were not drastic - however there are extremes.



Analysis: The Changes II

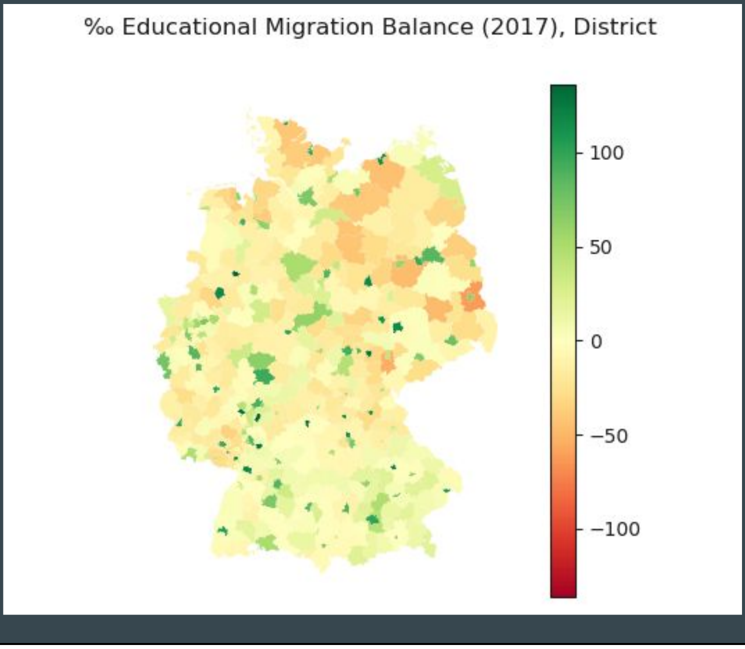
Same for Municipalities. Note that “Boostedt”, barely made the 5k cut-off in the first place.

Analysis: The Changes III



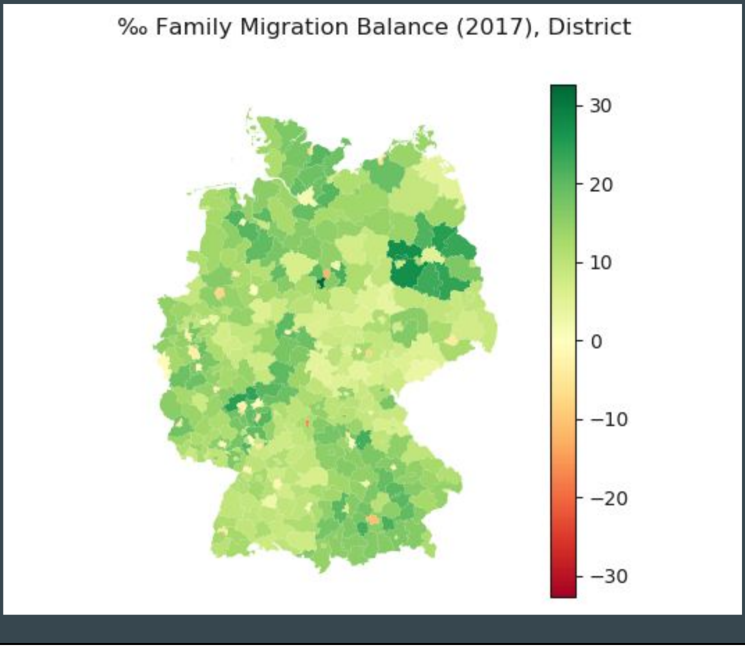
The East is loosing its population, but only on in the rural areas. Urban centers nationwide are growing. So are the belts around Germanies Top 3 cities: Berlin, Hamburg and Munich - which are also growing. Cologne's surroundings are not profiting from its growing center.

Analysis: Migration: The Scholars



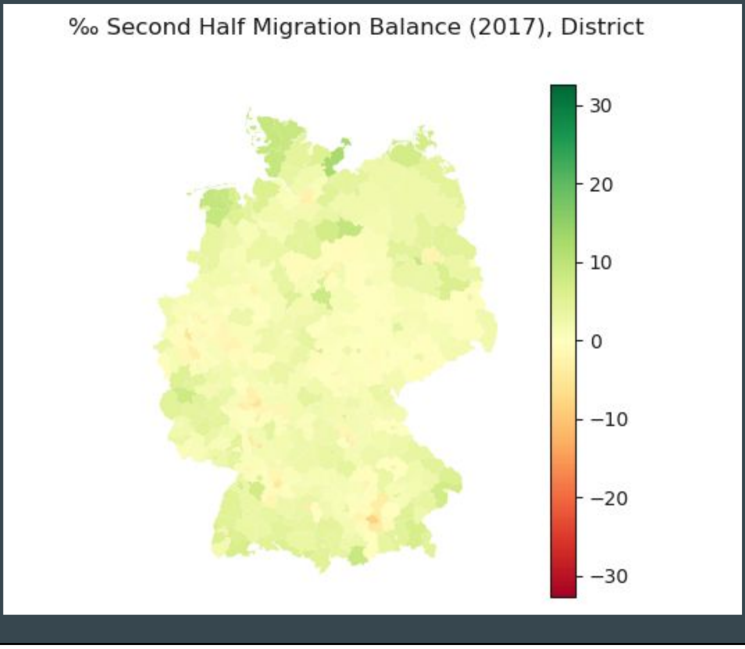
Families leave the cities for the rural areas.

Analysis: Migration: The Family



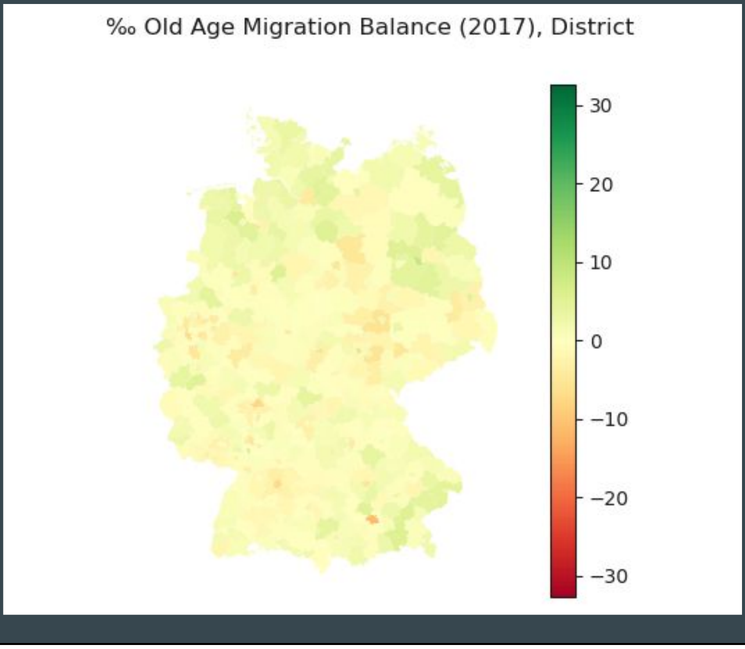
While people 18-24 leave the countryside and flock to the cities.

Analysis: Migration: The 2nd Half of Life

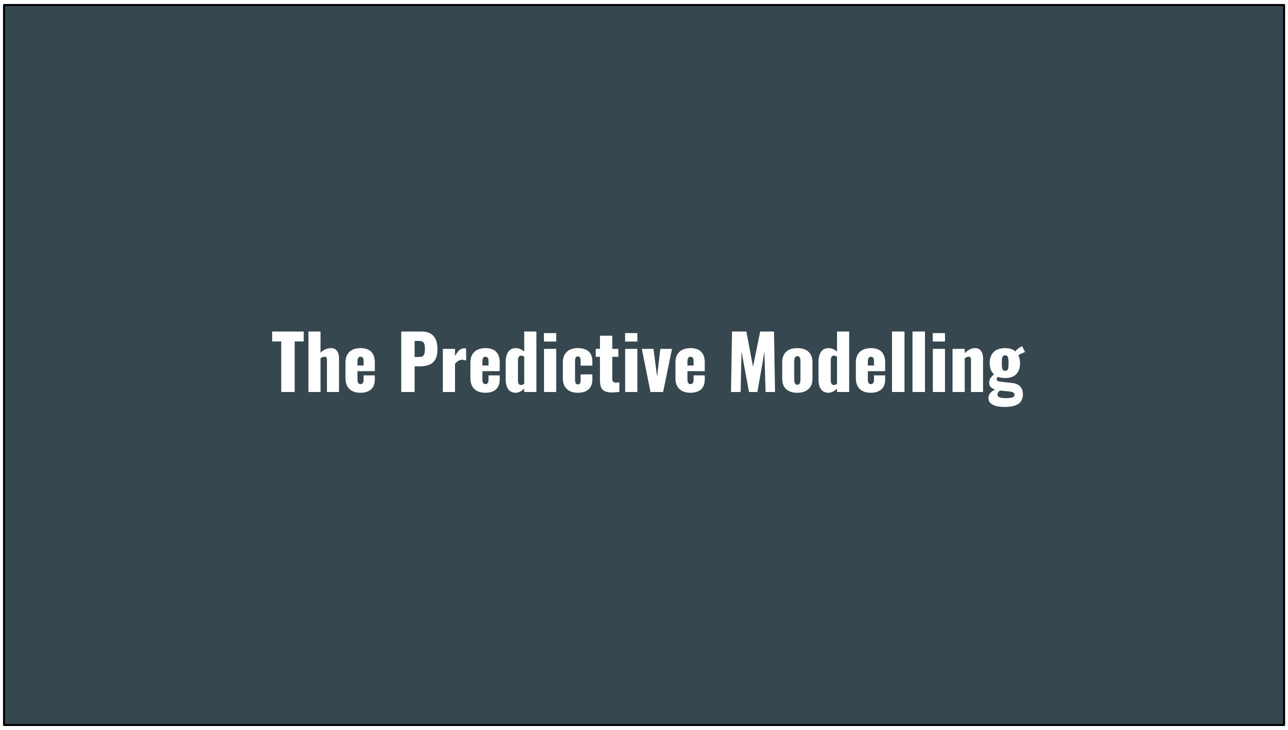


And around 50 the Germans return.

Analysis: Migration: The Old Age

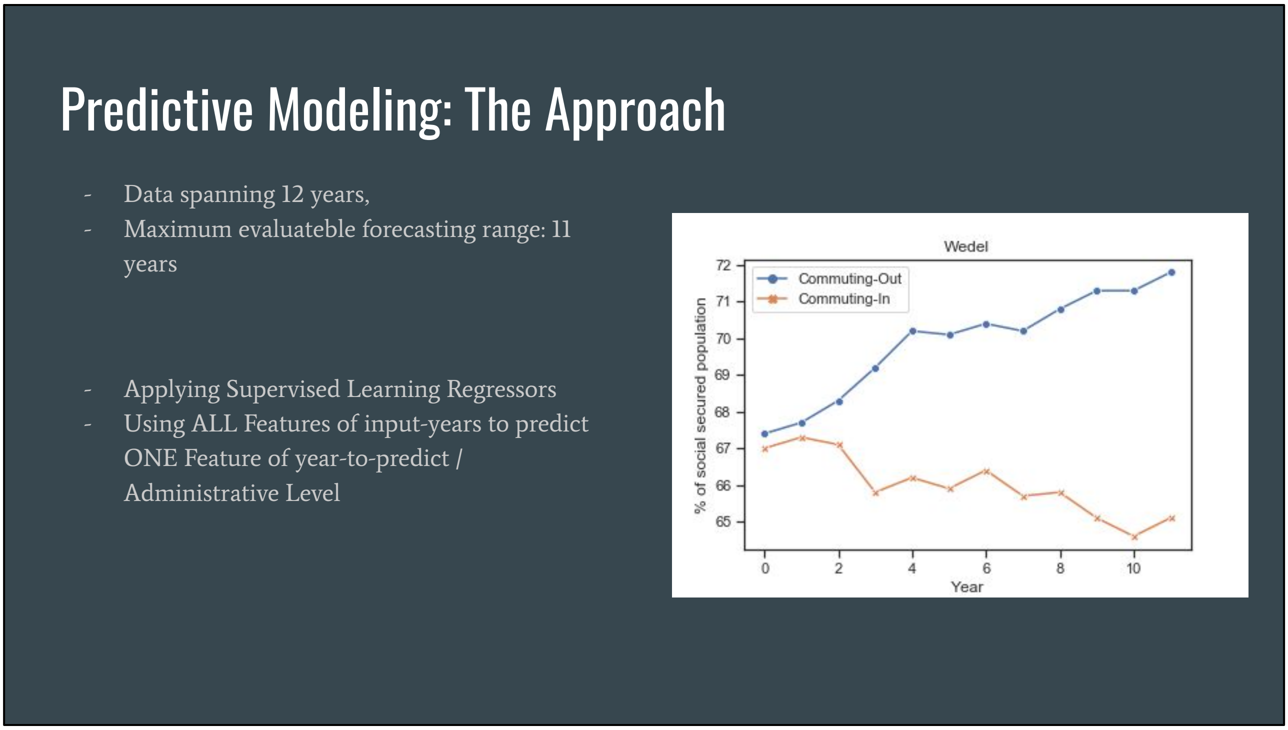


And when reaching the pension-age, people continue to leave the cities - however they also leave certain rural counties for others.



The Predictive Modelling

Let me get to the Predictive Modelling I did.

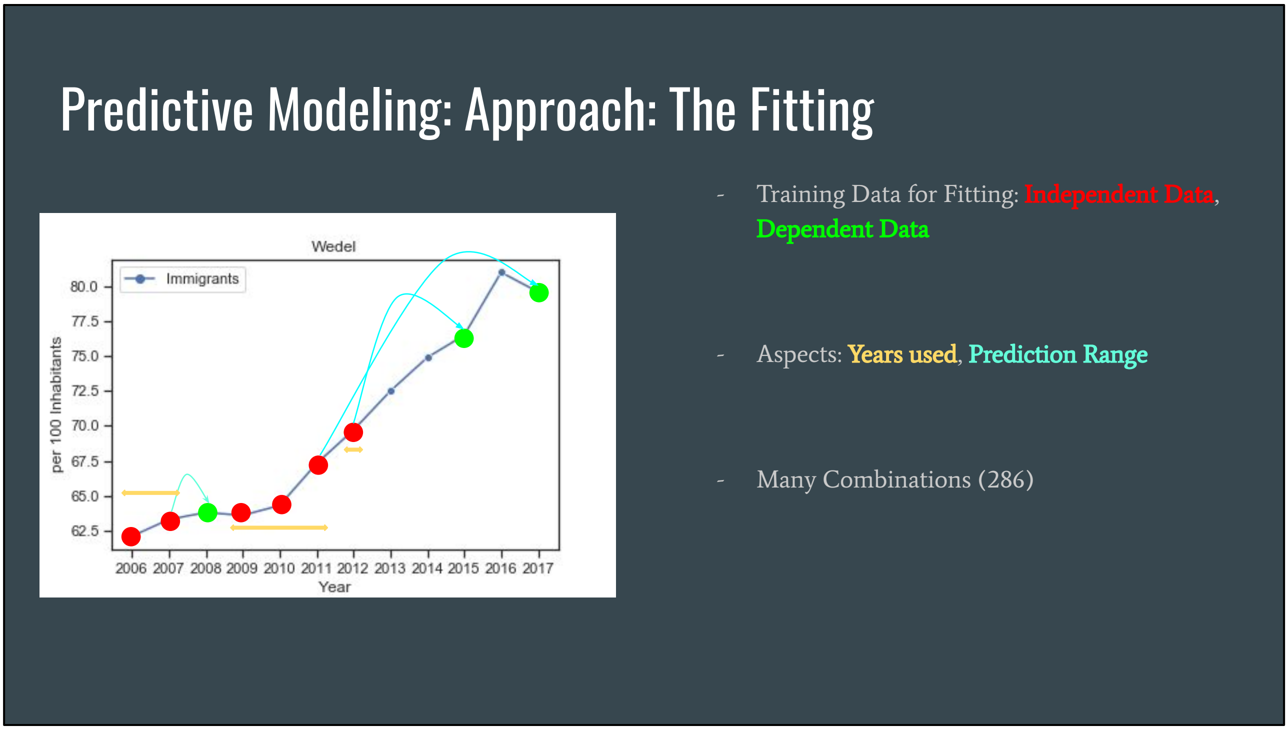


Predictive Modeling: The Approach

- Data spanning 12 years,
- Maximum evaluateble forecasting range: 11 years

- Applying Supervised Learning Regressors
- Using ALL Features of input-years to predict ONE Feature of year-to-predict / Administrative Level

Let me start by describing my approach.



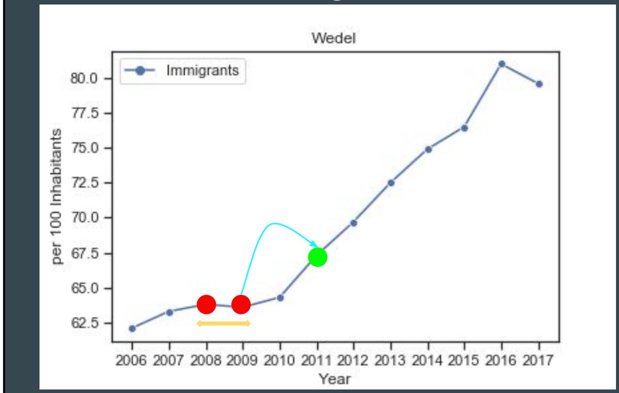
Predictive Modeling: Approach: The Fitting

- Training Data for Fitting: Independent Data, Dependent Data
- Aspects: Years used, Prediction Range
- Many Combinations (286)

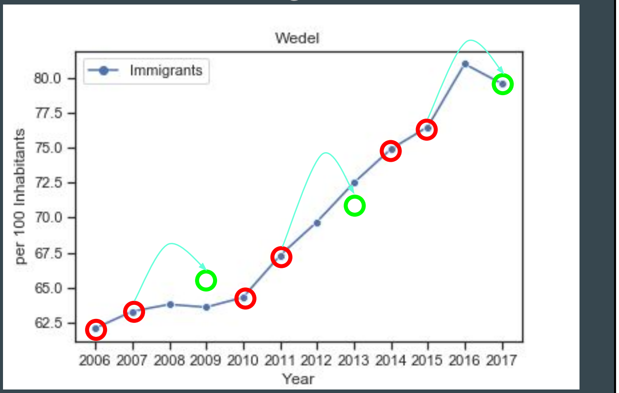
While a Data Scientist mainly sees the world in “dependent Data” and “independent Data”, I will have to set up my own terminology for this problem. I will mainly speak of “years used” (= the number of years’ which all features are used as independent data) an “prediction range” or “years looked ahead” (= how many years after the last year of the “years used” is taken as dependent data).

Predictive Modeling: Approach: The Testing

Training Data



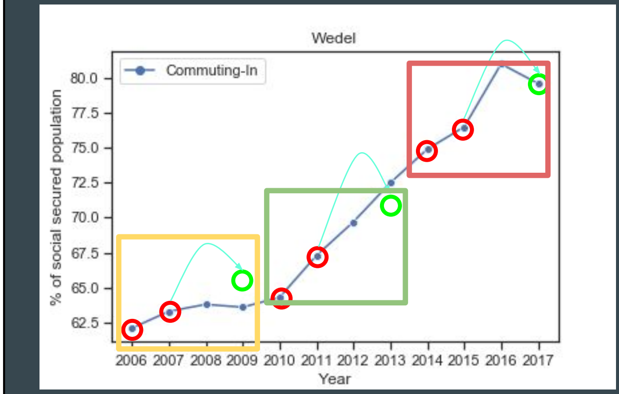
Testing Data

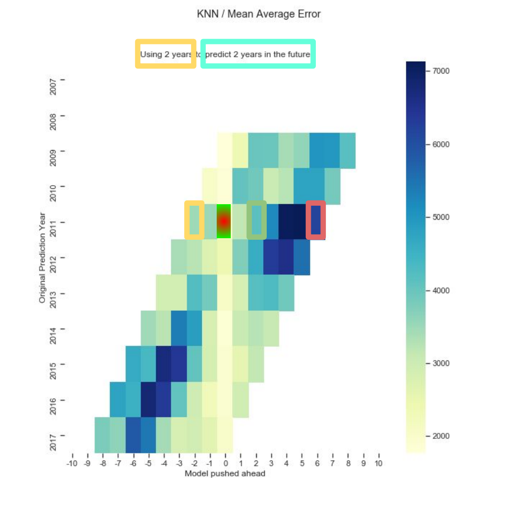


Dependent Data, Independent Data

Each combination of “years used” and “prediction range” has itself multiple possibilities to fitted into a model. Using 2 years to look 2 years into the future can be done by fitting a model with 2008 and 2009 as X and 2011 as y. This can then be tested, using the other combinations as testing data.

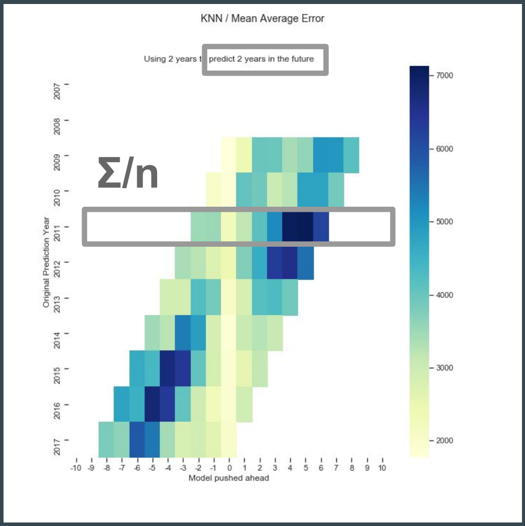
Predictive Modeling: Approach: The Scoring I

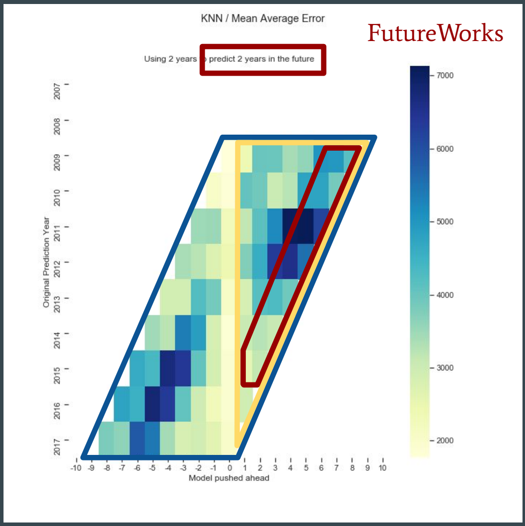




… as seen here. A heatmap is used to show which model worked well with which testing data. The idea behind this is, that some years are more representative for prediction than other, and some years are easier to predict than others.

Predictive Modeling: Approach: The Scoring II

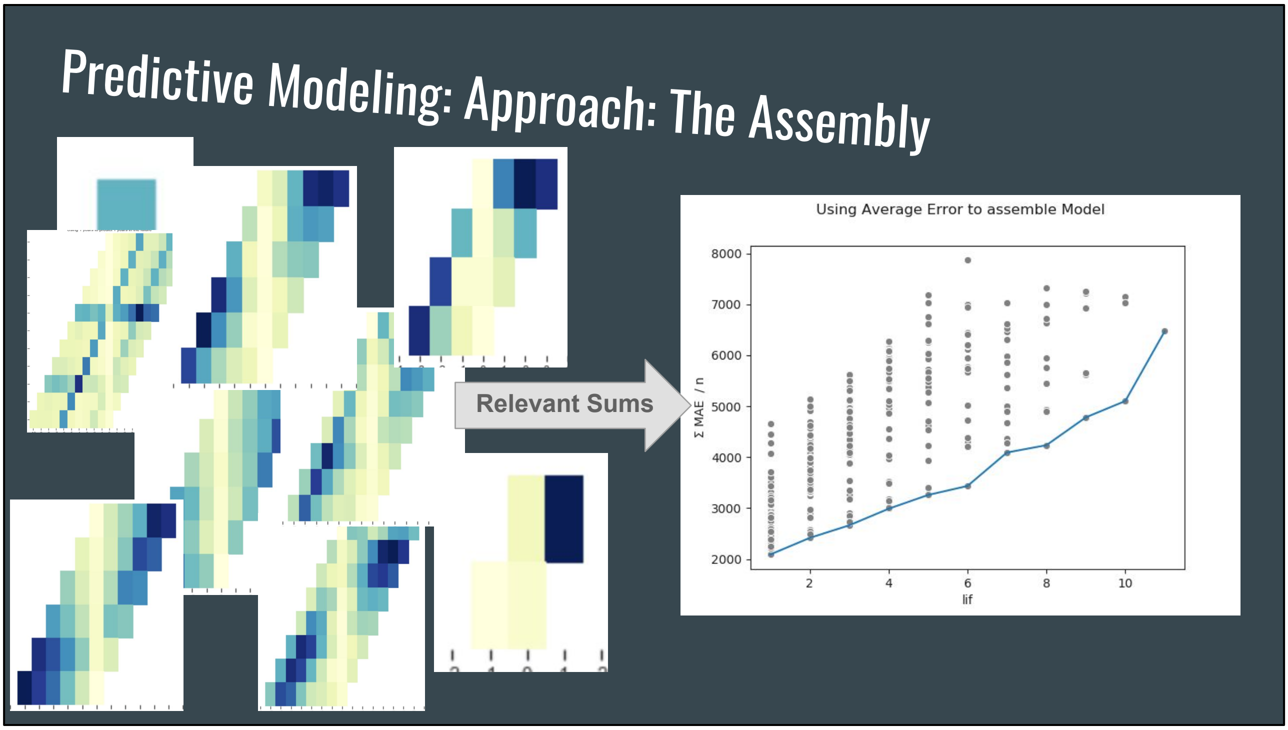




In the end, a lot of models are set up an evaluated. In order to pick the right model for each prediction range the scores are averaged. There are multiple ways to pick relevant scores:

- using the whole rhombus (blue),
- ore using only the prediction, not pastdiction (orange).

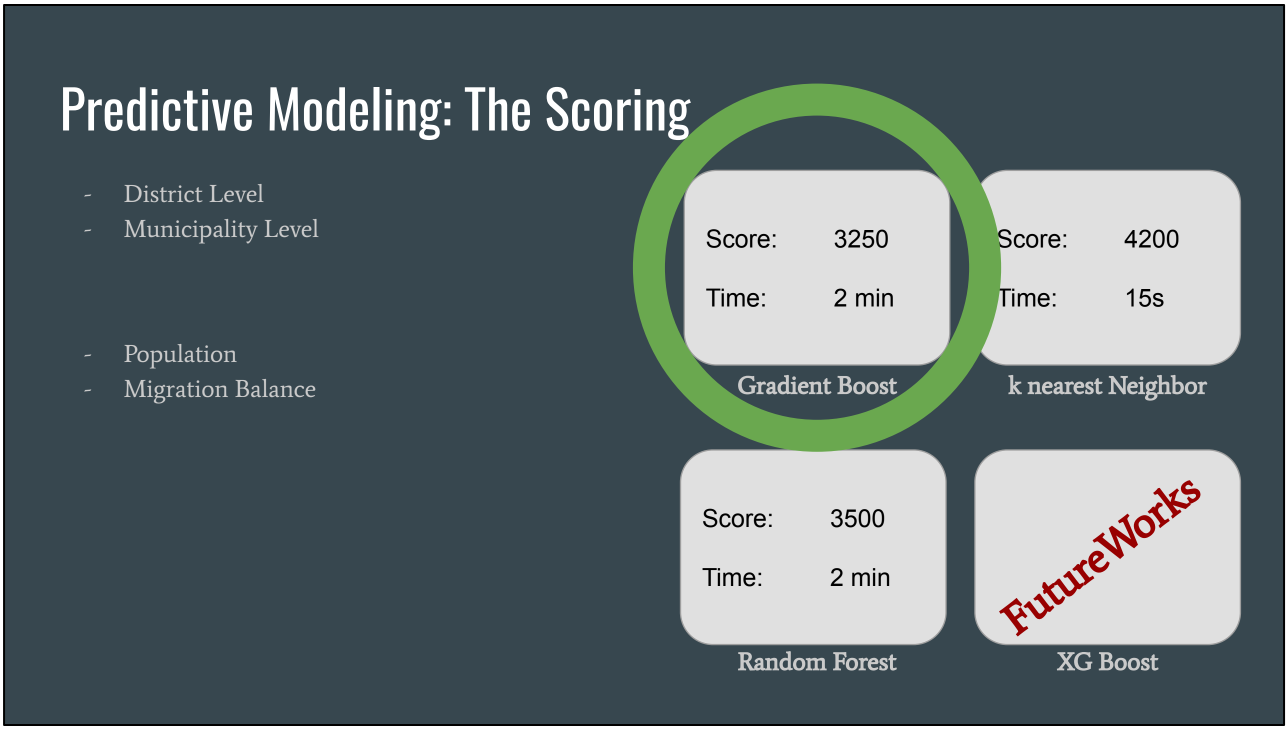
Overall, using the orange “triangle” delivered the most plausible results. In the future I would like to try to limit the selection even more.



Predictive Modeling: Approach: The Assembly

Relevant Sums

In the end, a lot of models are fitted and evaluated, and the best for each prediction range is chosen. Having a multiple options gave me the choice to either make a short, but accurate, or a long-range but inaccurate prediction. As you can see here most of the time an additional year looked ahead will bring additional inaccuracies. This was not always the case though. But in general, the farther the look the worse the prediction.



Predictive Modeling: The Scoring

- District Level
- Municipality Level

- Population
- Migration Balance

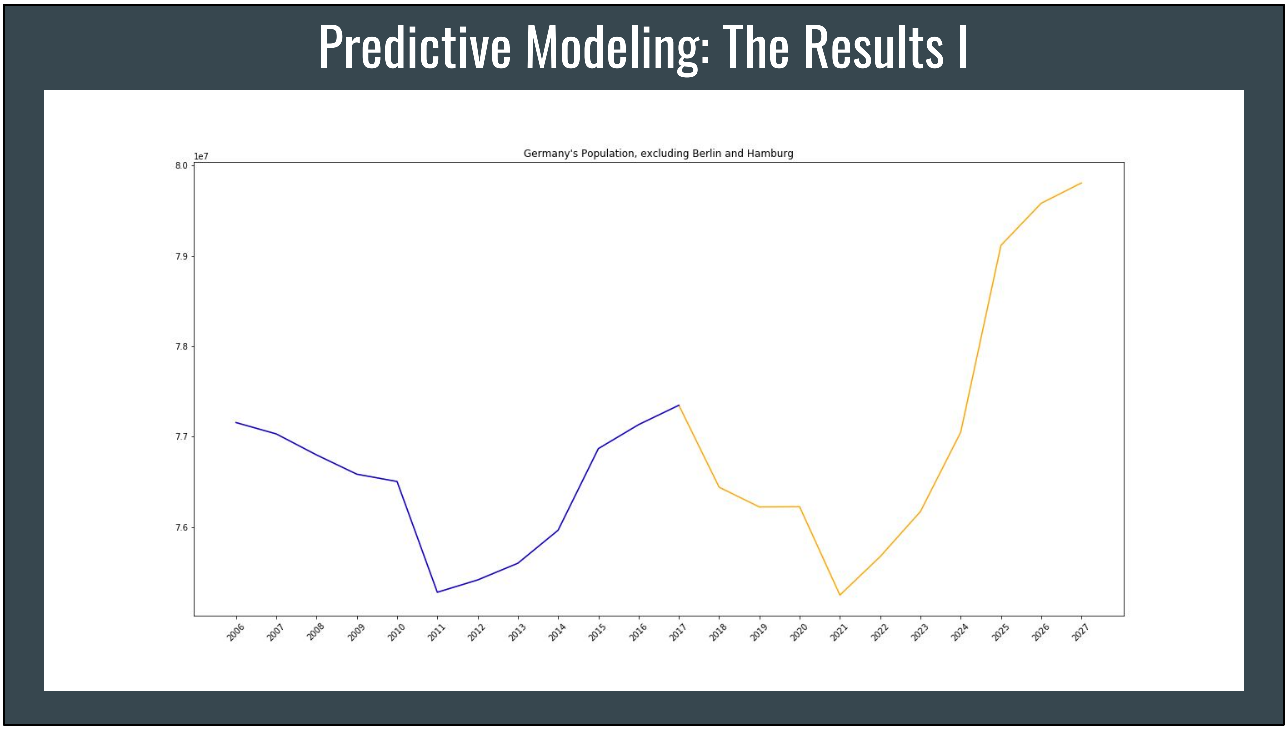
Score: 3250
Time: 2 min
Gradient Boost

Score: 4200
Time: 15s
k nearest Neighbor

Score: 3500
Time: 2 min
Random Forest

FutureWorks
XG Boost

I tried various regressors and judged them by thier score (Mean Absolute Error) and the time it took to set up the model (using my approach). In the end I chose Gradient Boost and chose the number of estimators so the fitting lasts around 1h - 30min.

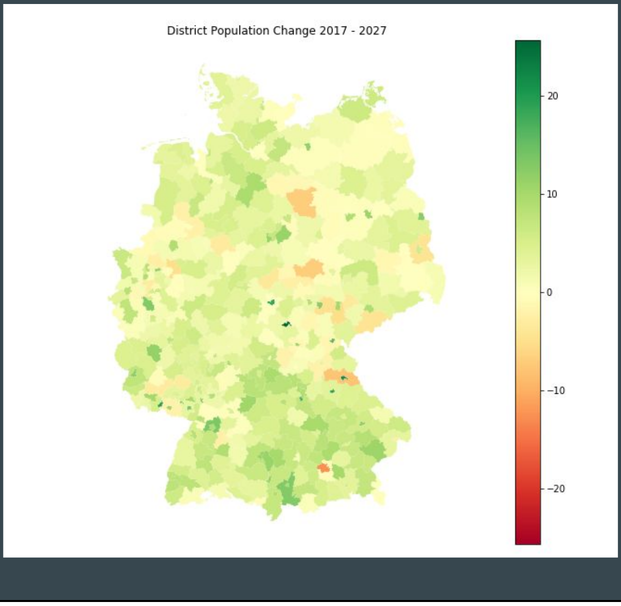


Predictive Modeling: The Results I

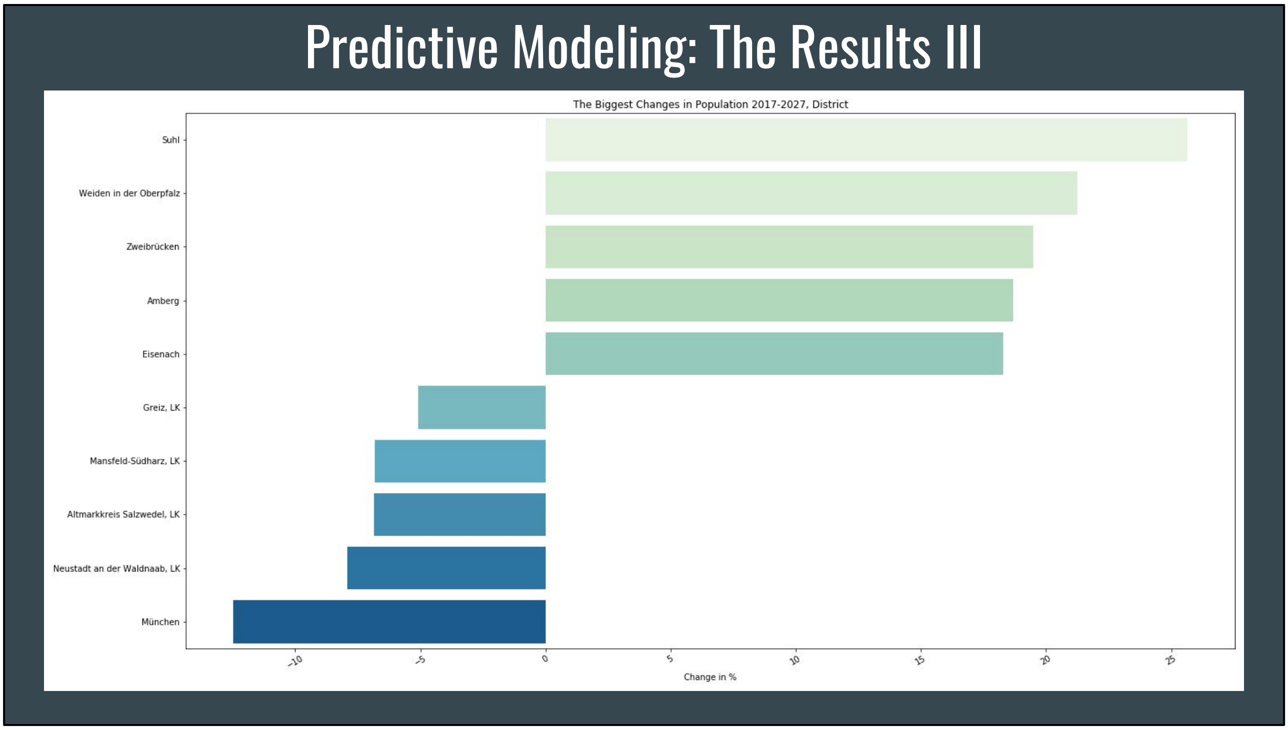
Germany's Population, excluding Berlin and Hamburg

The German Population will slump until 2021 but will recover in the following years.

Predictive Modeling: The Results II

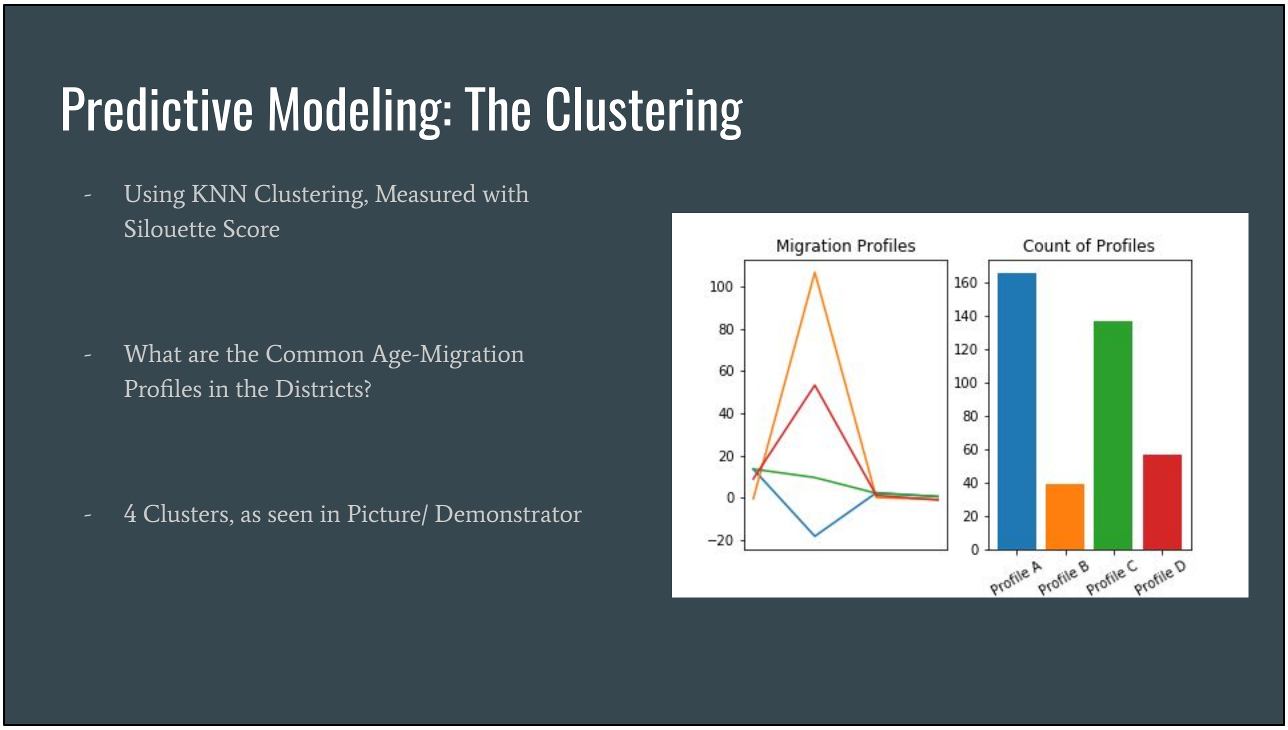


Besides some exceptions, the East will either stabalize or grow.



Predictive Modeling: The Results III

Here are the fastest growing, and shrinking districts.



Predictive Modeling: The Clustering

- Using KNN Clustering, Measured with Silouette Score

- What are the Common Age-Migration Profiles in the Districts?

- 4 Clusters, as seen in Picture/ Demonstrator

Usinge Clustering Methods I determined the 4 most prevailant types of Ageprofiles when it comes to migration. In the end, the educational migration Age group is the determining factor. More can be seen in the demonstrator.

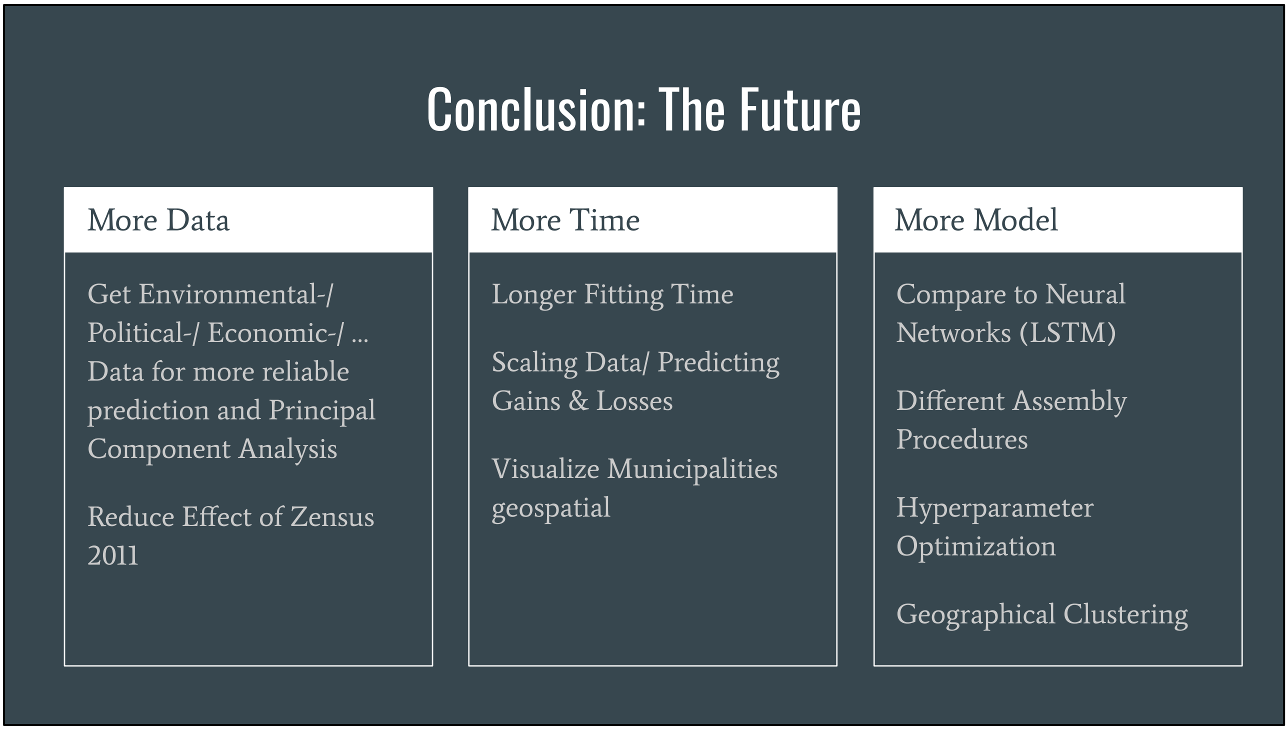
The Conclusion

The Conclusion

Some of the results are conclusive, while others seem outright unrealistic. In the end, more data is needed, data that truely holds the possibility to determine the attractiveness of a municipality or district.

In general the approach is feasable, it needs to be altered. Right now it is a purely "Data-Science" approach: Population is an infinite resource. In a future iteration the total population needs to be predicted, followed by a division into mobile, and immobile inhabitants. Then, the mobile ones are distributed among the districts/ municiplaities, based on a predicted attractiveness-factor.

In the end, the lines between Data-Science and Simulation need to be blurred.



Conclusion: The Future

More Data

Get Environmental-/ Political-/ Economic-/ ... Data for more reliable prediction and Principal Component Analysis

Reduce Effect of Zensus 2011

More Time

Longer Fitting Time

Scaling Data/ Predicting Gains & Losses

Visualize Municipalities geospatial

More Model

Compare to Neural Networks (LSTM)

Different Assembly Procedures

Hyperparameter Optimization

Geographical Clustering

This is what I would do if I had more Data/ Time/ ...

Thank you for your attention.

Do you have an questions?