The data trails of causality

Wählerstruktur bei der Bundestagswahl 2017

Dr. Ana Moya, Handelsblatt
Dr. Andreas Loos, ZEIT online

Four Types of Data Analytics

Descriptive: what's happening in my business?

- comprehensive, accurate and live data
- effective visualization

Diagnostic: Why is it happening?

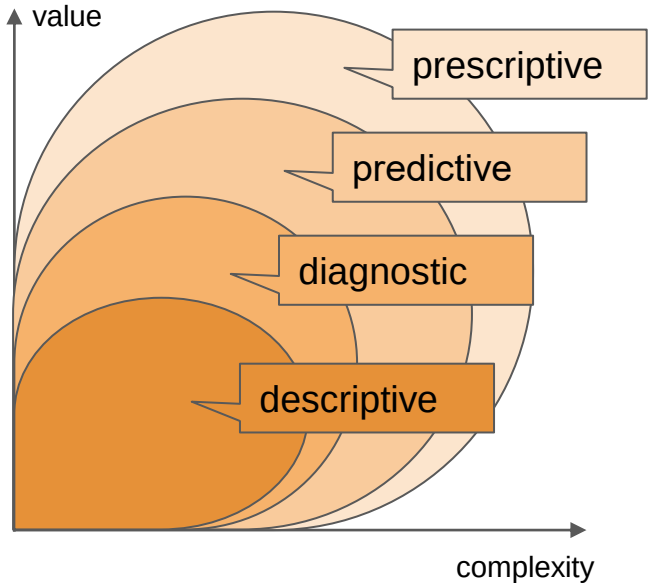
- ability to drill down to the root-cause
- ability to isolate all confounding information

Predictive: What's likely happening?

- historical patterns being used to predict specific outcomes using algorithms
- decisions are automated using algorithms and technology

Prescriptive: What do I need to do?

- applying advanced analytical techniques to make specific recommendations



The Data to be analysed

Election results in each constituency and Party der Bundestagswahl 2017 as well as Structure Date in each constituency.

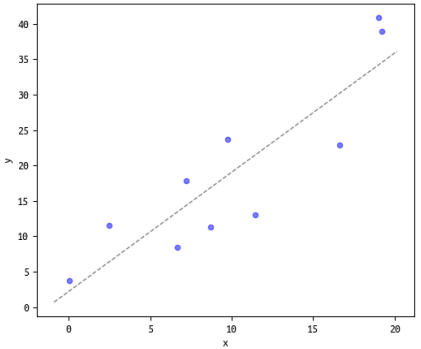
Output Variable (y): The proportion of AFD votes

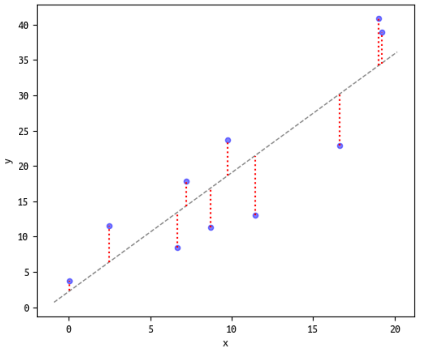
Input Variables (x): Structure Date

Source: Bundeswahlleiter.

(Some Maths Behind) Linear Regression

- (Simple) linear regression:
  - **“Find a line such that the sum of distances between this line and any of the y-values becomes minimal over all points x”**
  - maths: set of data points (measurements) D containing pairs (x, y) of real numbers; find a linear function f(x) that minimizes the sum of all squared distances: $\sum_{x,y \in D}(f(x)-y)^2$
  - quality measures:
    - p-value (is there a correlation at all?)
    - correlation coefficient and R^2 (“goodness of fit”)
- Multiple linear regression
  - **“Find a plane such that the sum of distances between this plane and the y-values becomes minimal over all points x”**
  - maths: x becomes a point in $\mathbb{R}^k$, the line becomes a k-1-dimensional hyperplane
  - don’t try to imagine this at home!





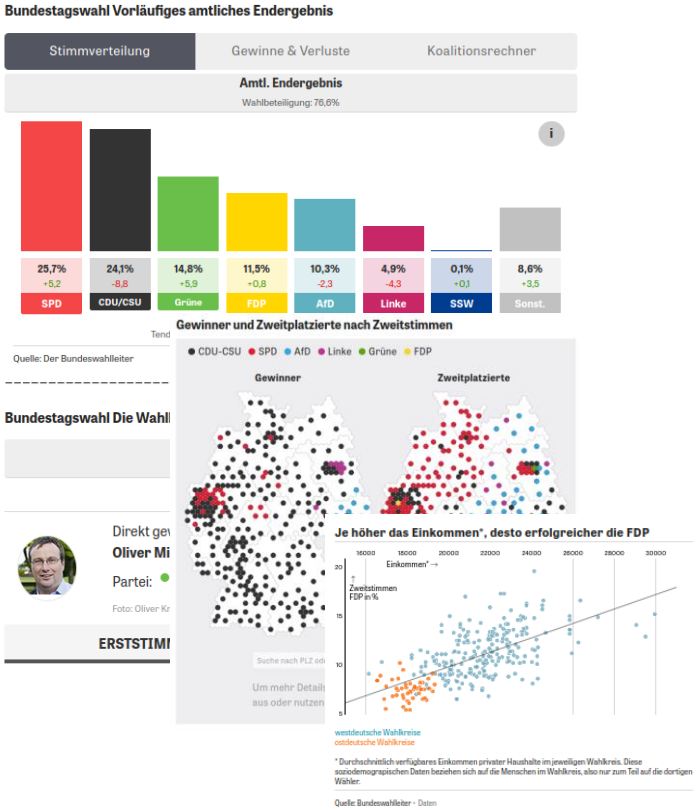
ZEIT online 2017: Voting Analysis

Analysis on the Results of the Federal Elections 2017: *Merkel-Enttäuschte und Nichtwähler machen die AfD stark*

- team of authors: Paul Blickle, Andreas Loos, Fabian Mohr, Julia Speckmeier, Julian Stahnke, Sascha Venohr und Veronika Völlinger
- article was permanently updated during the night after the voting

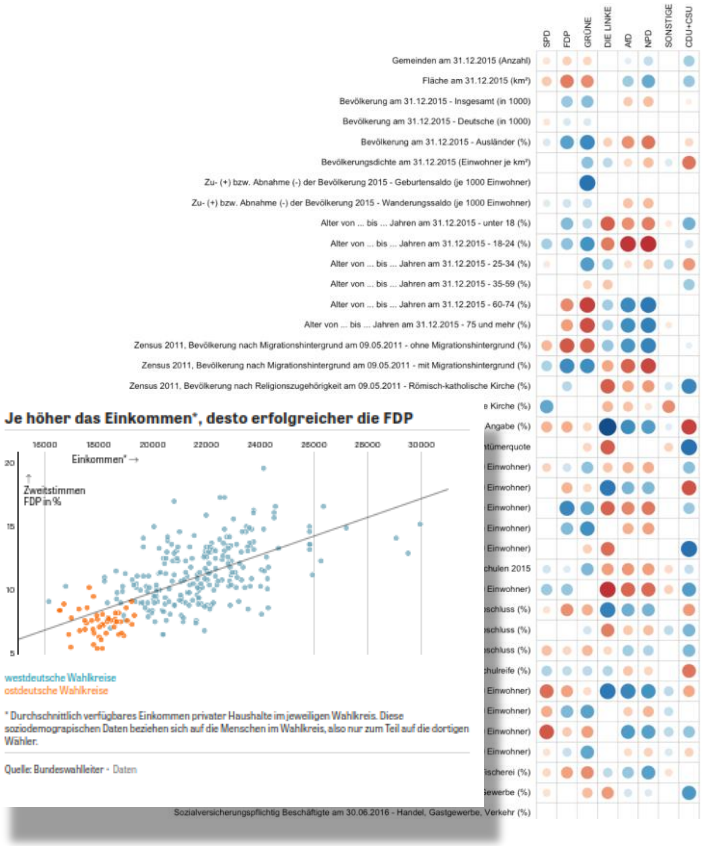
Many covered aspects:

- winning parties and candidates by first and second vote for each electoral district
- voting migration: how did voters vote now and in the election before
- voting by gender, age, education and occupation (workers, freelancers,…)
- voting among those who decided late what party to vote
- *regression: sociodemographic data and votes by electoral districts*



ZEIT online 2017: Linear Regression

- correlate any of 50 variables (cars per 1000 residents, voter turnout rate 2017 and 213, mean income, rate of residents with migration background,…) in the electoral districts with the result for any of the main parties
- compute correlation coefficient, p-value, and R^2 to measure how strong and trustable the correlation is
- show eastern and western Germany in the plots separately (to avoid Simpson's paradox here)
- all "interesting" correlation plots are analyzed by hand
- if still interesting: include them in the article
- **not implying that there is in fact a *linear correlation*, simply testing, how far we are away from this assumption**
- **NO CAUSAL INFERENCE!**



Re-Analysis: Evaluating Scenarios

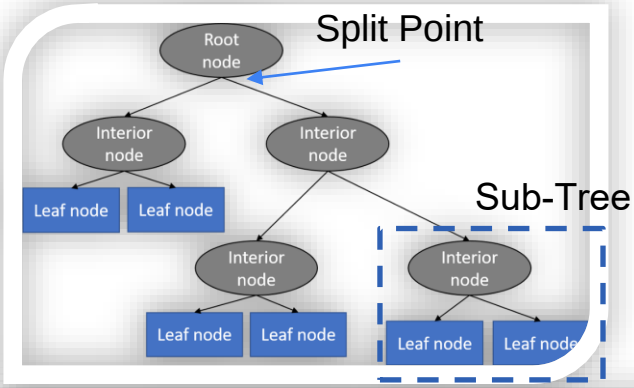
	Complexity	
Method	Interpretabilty	Elaboration
Regression Tree	low	low
Random Forest	med	low
Logistic Regression (Link: Logit)	med	med
Multi Level Analysis	high	high

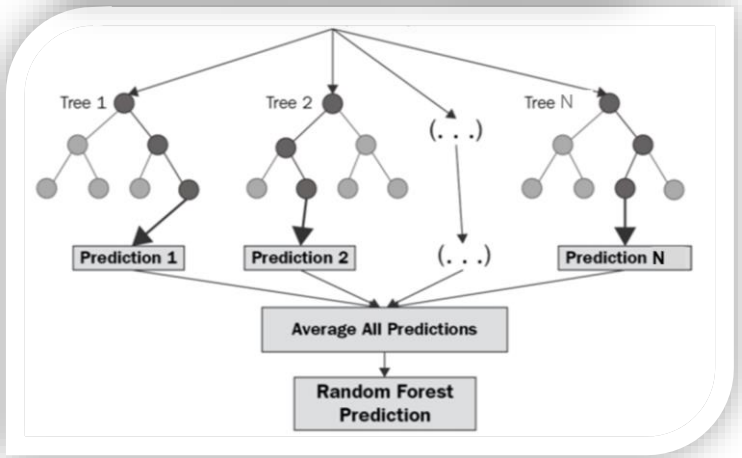
Decision Criteria: Mean Squared of error in the Test-Dataset.

- > Best Performance Random Forest followed by Regression Tree followed, therefore a combination was implemented

(Basic Explanation) Decision Trees/Random Forests

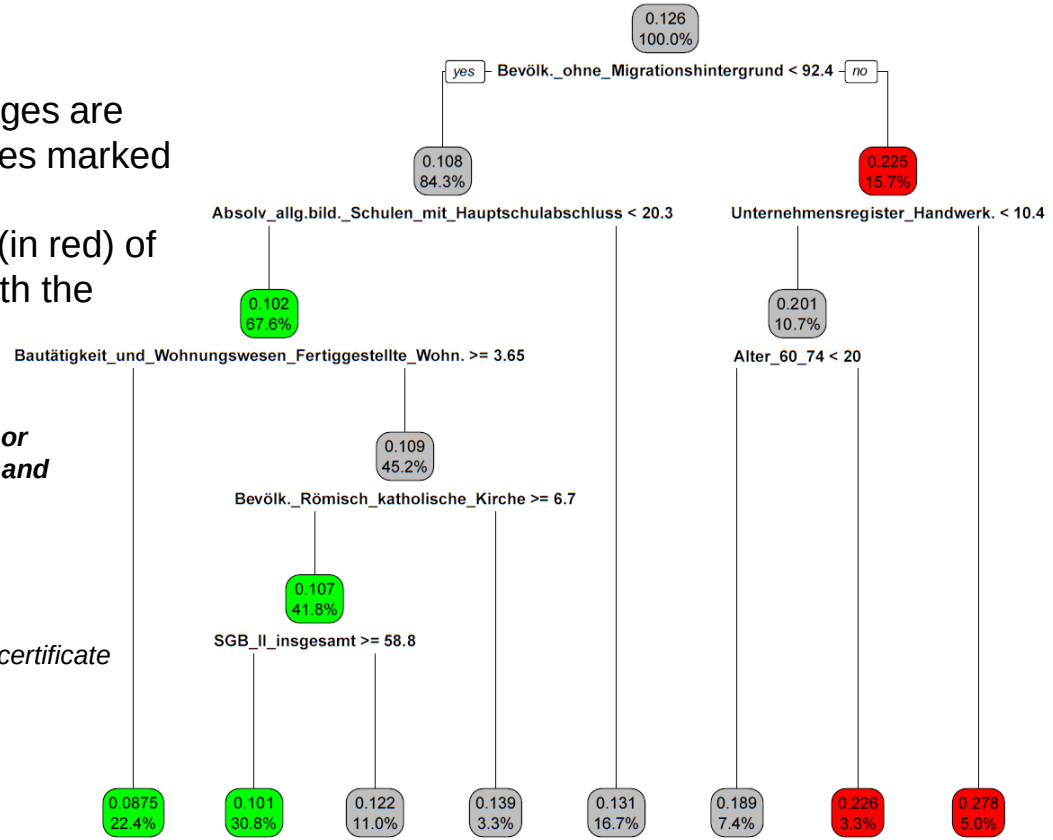
- (Regression) Decision Trees:
  - **"A Decision Tree generates a set of rules that follow a "IF Variable A is X THEN…" pattern and can be used for classification and regression".**
  - the goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data feature
- Random Forest - Ensemble Learning
  - **"A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees."**
  - Within a random forest, there is no interaction between the individual trees.
  - Variable importance is obtained as result of the random forest.





Re-Analysis: Results

- the nodes with the lowest modeled percentages are green, the nodes with the highest percentages marked in red.
- a constituency, which has a high proportion (in red) of AfD second votes (> 20%), are described with the following Properties:
  - *Proportion of population without a migration background*
  - *(according to the 2011 census) ≥ 92:4%* ***and***
  - *Business Register 2014 - Crafts (per 1000 inhabitants) ≥ 10.4* ***or***
  - *Business Register 2014 - Crafts (per 1000 inhabitants) < 10.4* ***and***
  - *Proportion of the population between 60 and 74 years ≥ 20%*
- in the first splits of the tree (highest variable importance from Random Forest):
  - *Proportion of population without a migration background*
  - *Absolv_allg.bild.Schools with lower secondary school leaving certificate*
  - *Business Register 2014*



Discussion

- simple linear regression *is* simple: no parameters to tweak, easy to access/comprehend
- so: we can keep the focus on the story (since no complicated math explanations necessary)
- – especially since the regression was only one small aspect of ZEIT online voting analysis
- usually performs ok in prediction (but that's not our goal)
- But: simple linear regression can deal only with "one fact – one result"; no combinations of "independent" variables
- But: dependent variable is a percentage. In general, a linear regression will fail here when predicting (already, because linearity is most probably not given).

- Decision trees and random forests *do* combine facts/variables to predict results
- they usually perform very good in prediction
- are also not built upon various assumptions, such as normal distribution
- But: prediction possible only in the range of the training data values, therefore it is important to consider a representative training data set.
- But: several parameters to chose (forest size, bag size, methods how to measure feature importance,…) with influence on feature importance (and thus they have to be explained)