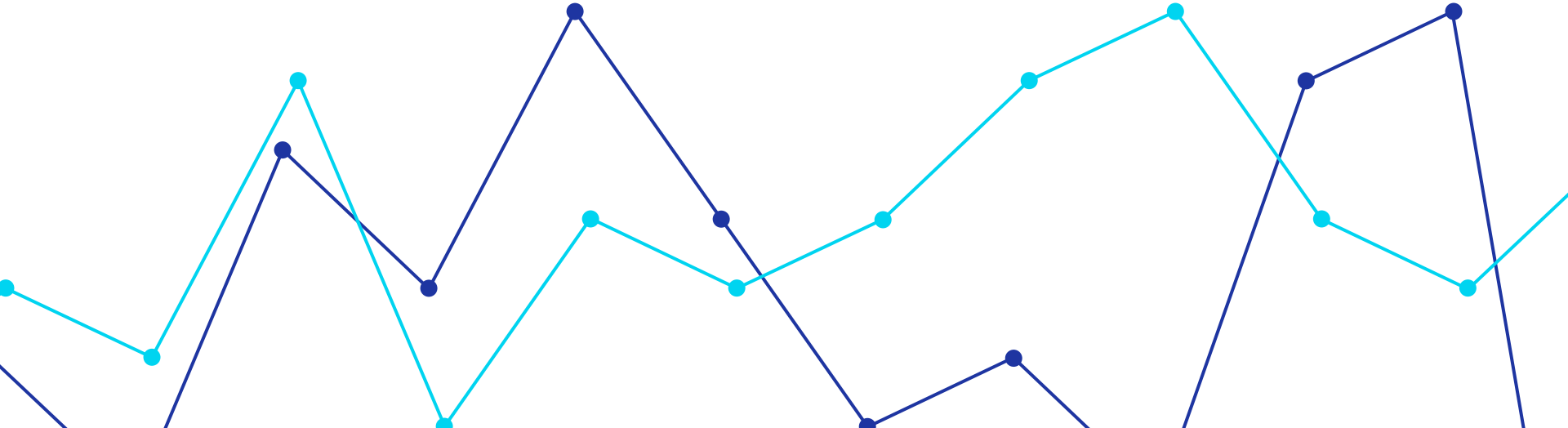
Understanding people in Germany through numbers

Group F (Jenny Schönfeld, Jimi Kim, Felix Daubner)
11.01.2024



Content

1. Introduction and Motivation
2. Classification
3. Regression
4. Conclusions

Introduction

Data

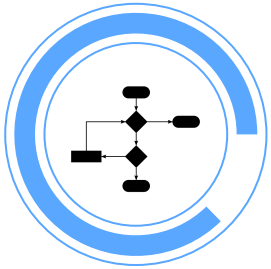
"European Social Survey (ESS)"
- across 28 countries
- paper-based survey

Motivation

Understanding the interconnections
- Social & Political & Personal Well-Being & Economical aspects

Research question

Understanding the interconnections
- Social & Political & Personal Well-Being & Economical aspects



Hypothese

Regression: H1 - H4
Classification: H5 - H10

Motivation and Research Questions

Data

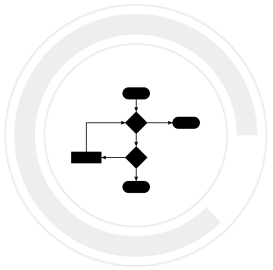
"European Social Survey (ESS)"
- across 28 countries
- paper-based survey

Motivation

Understanding the interconnections
- Social & Political & Personal Well-Being & Economical aspects

Research question

Understanding the interconnections
- Social & Political & Personal Well-Being & Economical aspects



Hypothese

Regression: H1 - H4
Classification: H5 - H10

Research Questions and Objectives

Data

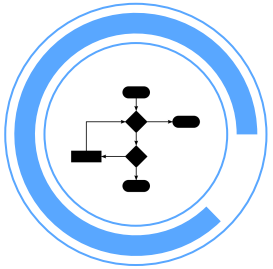
"European Social Survey (ESS)"
- across 28 countries
- paper-based survey

Motivation

Understanding the interconnections
- Social & Political & Personal Well-Being & Economical aspects

Research

Understanding the interconnections
- Social & Political & Personal Well-Being & Economical aspects



Hypothese

Regression: H1 - H4
Classification: H5 - H8

Key variables

Response Variables:

- grspaya - Usual gross pay in euro, before deductions for tax and insurance
- happy - How happy are you

Predictor Variables:

- emplnof - Number of employees father had
- emplmom - Number of employees mother had
- dsgrmnya - How often disagree with husband/wife/partner about money
- gincdif - Government should reduce differences in income levels

Financial well-being

- polintr - How interested in politics
- mmbprty - Member of political party
- lrscale - Placement on left right scale

Political beliefs

- tvtot - TV watching, total time on average weekday
- tvpol - TV watching, news/politics/current affairs on average weekday
- nwsptot - Newspaper reading, total time on average weekday
- netuse - Personal use of internet/e-mail/www
- impfun - Important to seek fun and things that give pleasure
- ipgdtim - Important to have a good time
- iprspot - Important to get respect from others
- wrywprb - Worry about work problems when not working, how often

Social preferences

- edude1-3 - Highest level of education, Germany: höchster allgemeinbildender Schulabschluss / höchster Studienabschluss / höchster Ausbildungsabschluss
- edufde1-3 - Father's highest level of education, Germany: höchster allgemeinbildender Schulabschluss / höchster Studienabschluss / höchster Ausbildungsabschluss
- edumde1-3 - Mother's highest level of education, Germany: höchster allgemeinbildender Schulabschluss / höchster Studienabschluss / höchster Ausbildungsabschluss

Education

Hypothesis - Linear Regression

H1

Media Impact: The level of media engagement (internet use and TV watching) correlates with an individual's gross-pay in Germany.

H2

Political Opinion: Political factors (interest in politics and placement on the left-right scale) play a substantial role in predicting an individual's gross-pay.

H3

Family Impact: Variables related to family upbringing, (number of employees parents had and the highest level of education for both the individual and their parents) significantly predict an individual's gross pay.

H4

Education Impact: Educational factors (the highest general educational qualification, the highest degree obtained, and the highest vocational qualification) significantly correlate with gross pay.

Hypothesis - Logistic Regression

H5

Internet Use, Media Consumption, and Happiness: The reported level of happiness is influenced by an individual's personal use of the internet, as well as the total time spent on newspaper reading on an average weekday.

H6

Political Interest: The reported Happiness-Score is predicted by their level of interest in politics.

H7

Education Impact: The reported Happiness-Score is predicted by the highest level of education.

H8

Personal Values: The reported Happiness-Score is influenced by personal values, including the importance placed on seeking fun and getting respect from others.

Classification

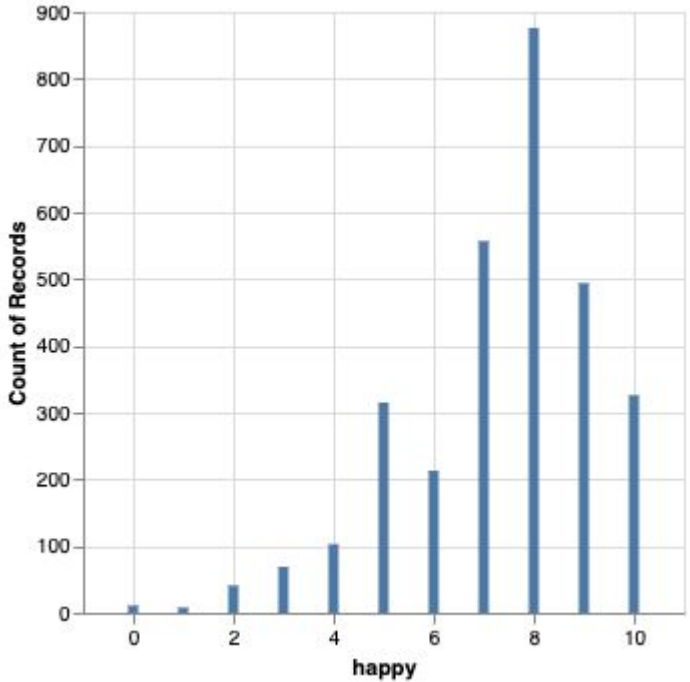
Response Variable

happy: How happy are you?

- Logistic Regression: binary prediction
- Current spectrum: 0 to 10

0	1	2	3	4	5	6	7	8	9	10

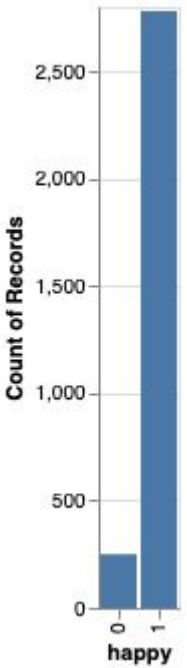
- Categorize “happy” into two values



Classification - Data Corrections

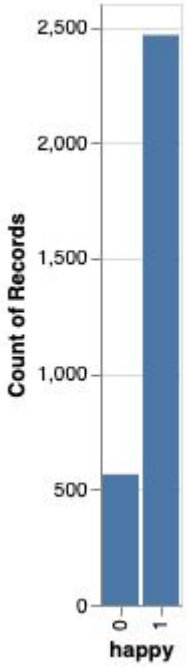
0	1	2	3	4	5	6	7	8	9	10

happy
1 2781
0 250



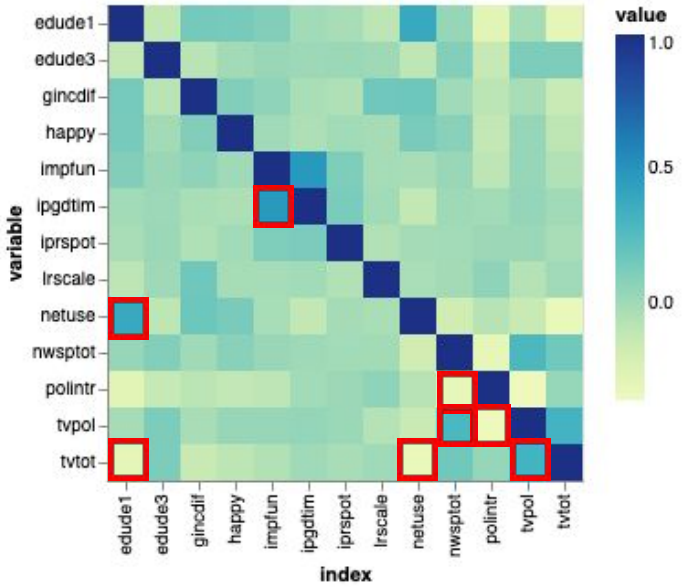
0	1	2	3	4	5	6	7	8	9	10

happy
1 2466
0 565



Classification - Feature Selection

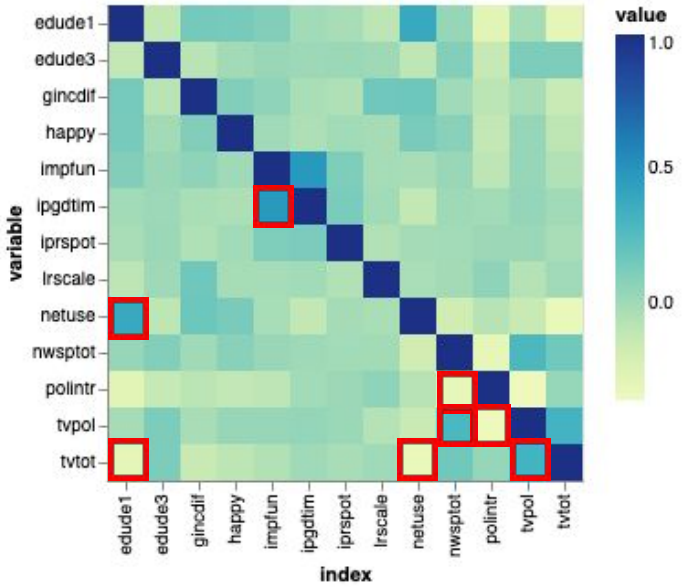
Correlation Matrix



feature	VIF
lrscale	7.586451
nwsptot	2.870115
netuse	4.493601
tvtot	6.384523
tvpol	4.692509
impfun	9.723894
ipgdtim	8.024356
polintr	7.536917
iprspot	7.418480
gincdif	5.606190
edude1	8.702929
edude3	2.518125

Classification - Feature Selection

Correlation Matrix



feature	VIF
~~lrscale~~	7.586451
nwsptot	2.870115
netuse	4.493601
~~tvtot~~	6.384523
tvpol	4.692509
~~impfun~~	9.723894
ipgdtim	8.024356
polintr	7.536917
~~iprspot~~	7.418480
gincdif	5.606190
~~edude1~~	8.702929
edude3	2.518125

Classification - Model

Logistic Regression Model

- 7 predictor variables
- Response variable: “happy”
- 1 = “happy”
- 0 = “unhappy”
- Train-test-split: 70:30

Trained on unbalanced dataset

Trained on oversampled dataset

Classification - Evaluation

What do we want to achieve?

- Model: predicts whether a person is happy or unhappy
- Application examples could be:
  - Depression prediction

Worst case

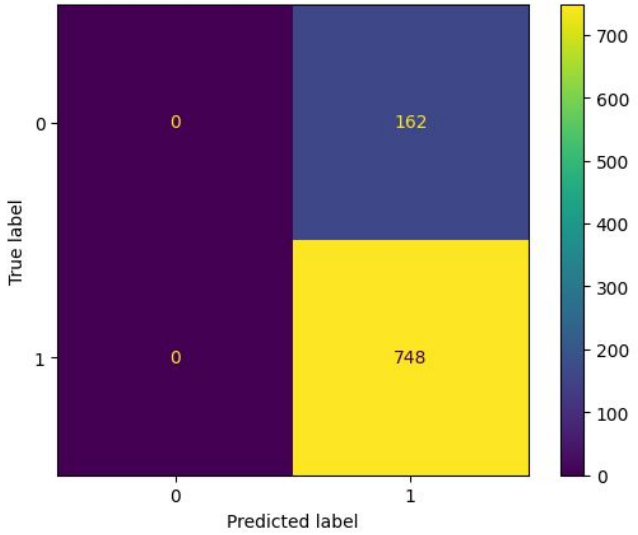
Unhappy person is predicted to be happy → “False Positive”

Classification - Evaluation

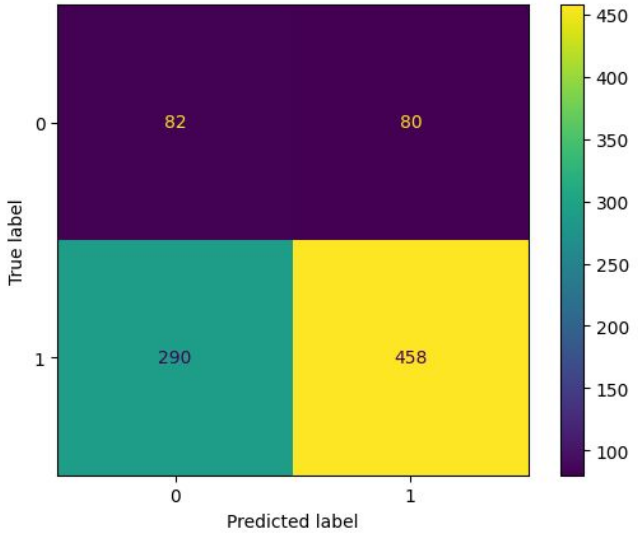
Unbalanced Data

Accuracy

Oversampled Data



Confusion Matrix



Classification - Evaluation

Unbalanced Data

precision	recall	f1-score	support
0.00	0.00	0.00	162
0.82	1.00	0.90	748
		0.82	910
0.41	0.50	0.45	910
0.68	0.82	0.74	910

Oversampled Data

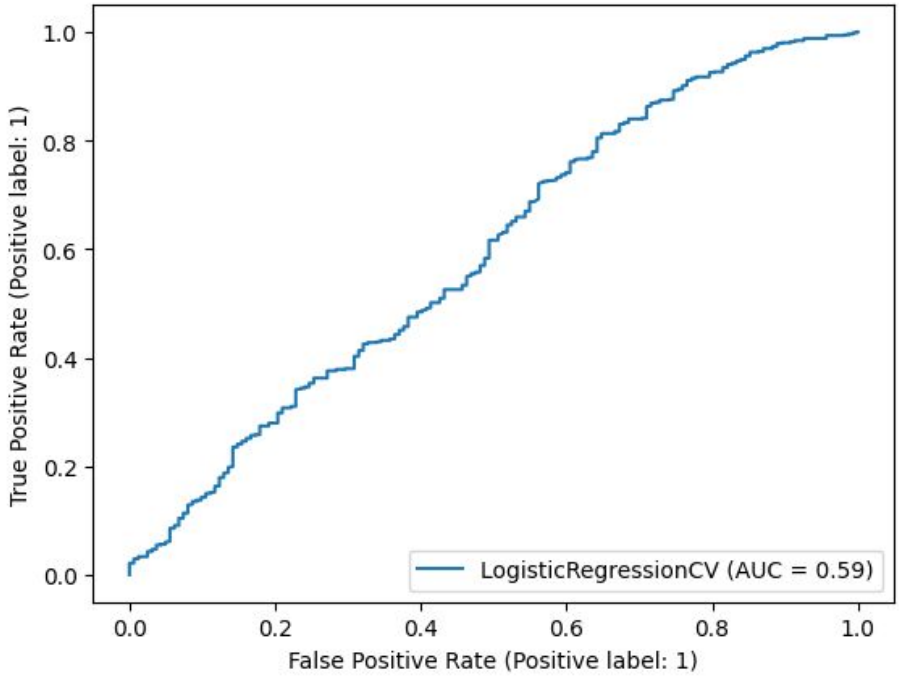
	precision	recall	f1-score	support
unhappy	0.22	0.51	0.31	162
happy	0.85	0.61	0.71	748
accuracy			0.59	910
macro avg	0.54	0.56	0.51	910
weighted avg	0.74	0.59	0.64	910

$$\textbf{specificity} = \frac{tn}{tn + fp}$$

Out of all people who are unhappy, how many got predicted to be unhappy?

Classification - Evaluation

Oversampled Data

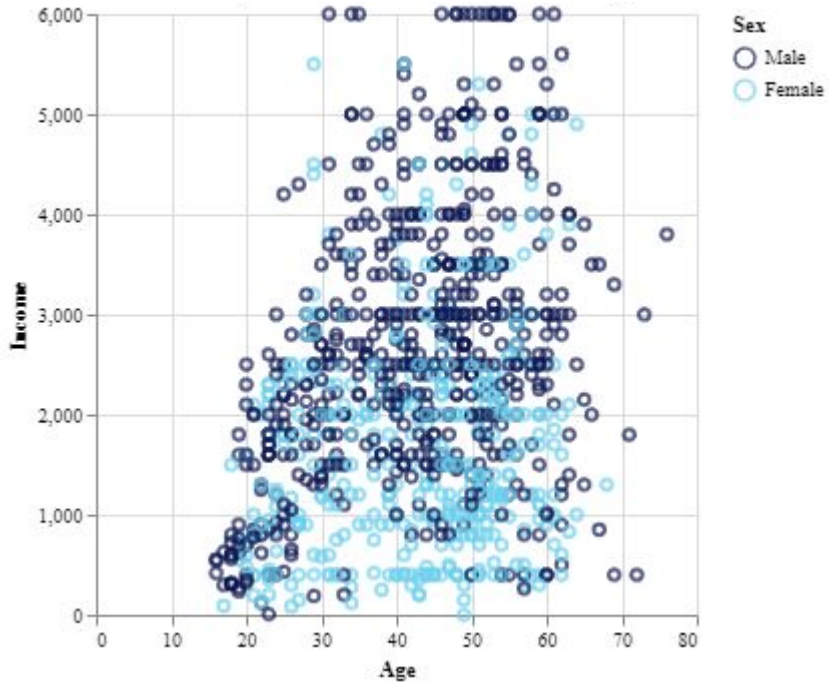


Classification - Evaluation

Oversampled Data

	Threshold: **0.4**		Threshold: **0.5**		Threshold: **0.6**	
	recall	f1-score	recall	f1-score	recall	f1-score
unhappy	0.29		0.51		0.81	
happy	0.85		0.61		0.28	
macro avg	**0.57**	0.57	**0.56**	0.51	**0.54**	0.37

Linear Regression



Response Variable

grspaya: Usual gross pay before deductions for tax and insurance

Predictor Variables

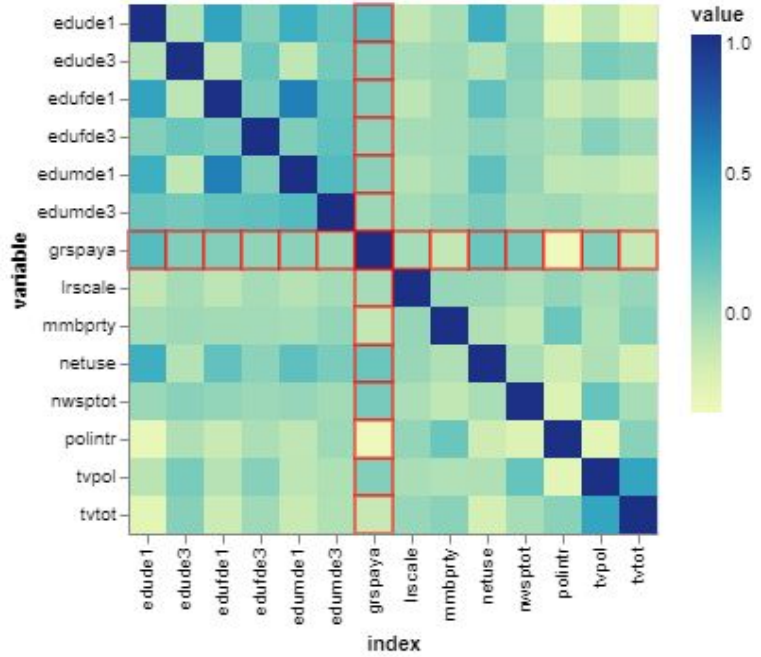
Correlation between income and various social, political and economic factors

Data Corrections

Removing missing values and outlier of response variable, dropping predictor variables, data imputation

Regression - Feature Selection

Correlation matrix

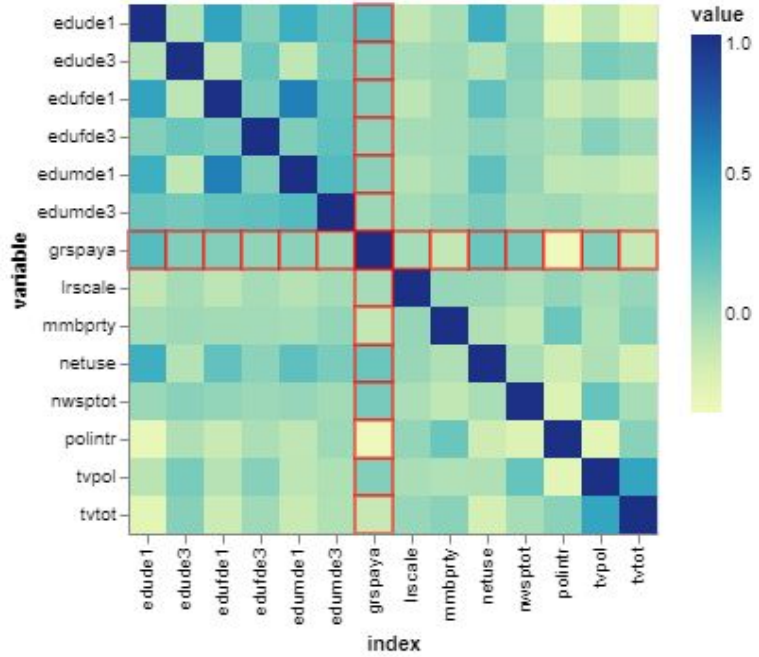


feature	VIF
tvpol	5.35
lrscale	7.96
netuse	7.53
tvtot	6.79
nwsptot	2.73
polintr	11.27
mmbprty	49.26
edude1	14.42
edude3	4.09
edufde1	10.05
edufde3	5.62
edumde1	10.81
edumde3	2.46

Regression - Feature Selection

Correlation > 0.05

Correlation matrix



feature	VIF
tvpol	5.35
lrscale	7.96
netuse	7.53
tvtot	6.79
nwsptot	2.73
~~polintr~~	11.27
~~mmbprty~~	49.26
~~edude1~~	14.42
edude3	4.09
~~edufde1~~	10.05
edufde3	5.62
~~edumde1~~	10.81
edumde3	2.46

Linear Regression

Forward Selection

schrittweises Regressionsverfahren

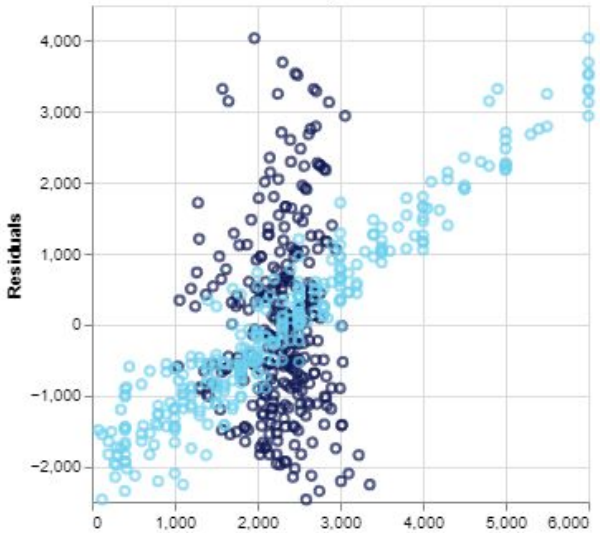
Lasso Regression

Form der regulierten Regression

Linear Regression - Performance

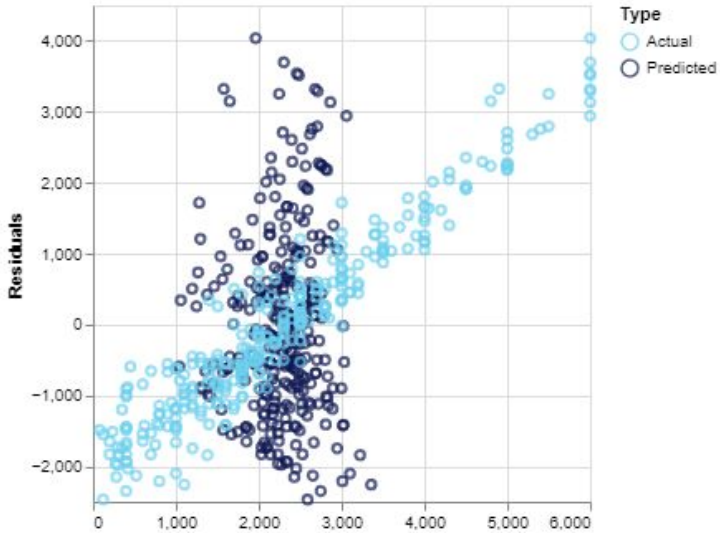
Linear Regression

MSE: 1665350.34
RMSE: 1290.48
R2 : 0.1



Lasso Regression

MSE: 167499.0
RMSE: 1292.48
R2 : 0.1



Discussion

- Impacts of the predictor variables (Media, Education, Politic Interest on the Economical Status and the Happiness → both models not robust enough to predict the relationships

- Challenges
  - Manual variable selection resulted
  - Issues such as multicollinearity and data inconsistencies
  - Log. Regression: subjective data (e.g. “happy/unhappy”)
  - Lin. Regression: inconsistency of data, e.g. Income unit (week/month/year?)
  - Ambiguities: inefficient categorization (e.g. Newspaper reading 0,5hr / 1hr / 1,5hr / 2hr / 2,5hr / 3hr / …)

Conclusion

- Our study aimed to shed light on the socio-economic structure of Germany, focusing especially on the influence of media consumption, education, political opinion for individual's happiness and the income.

- Future Research Approaches:
  - Need for more extensive data collection and improved analytical techniques
  - A wider range of relevant factors and more precise analysis methods for clearer insights.

- Potential: longitudinal Studies and structured models will provide insights into evolving societal trends and relationships in Germany.

References

- Smith, A., Johnson, R., & Brown, C. (2019). Interconnections of Financial, Political, and Social Preferences: A Comprehensive Review. Journal of Social Dynamics, 15(2), 245-267.
- Jones, M., & Brown, S. (2020). Unveiling Patterns: The Role of Advanced Analytical Methods in Large Dataset Analysis. Journal of Quantitative Research, 25(4), 511-530.
- Garcia, E., Patel, K., & Lee, J. (2021). Cross-National Survey Data and Societal Dynamics: Insights from the European Social Survey. International Journal of Social Research, 30(3), 321-340.
- Anderson, R., & Smith, B. (2020). Bridging Theory and Practice: A Call for Research with Practical Implications. Journal of Applied Social Science, 18(2), 211-228.