Modelling and Understanding Conspiracy Belief: Model Optimization and Interpretation

Capstone Sprint 3

Elliot Carter

BrainStation®

The Problem Area

- Who is most susceptible to conspiracy belief and misinformation?
- Big problem, big consequences

How Data Science Can Help

- Much research on conspiracy theories has small sample sizes
- Recent survey by Imhoff et al produced a large dataset (~100k respondents).
- My project: explore and model their data

ARTICLES
https://doi.org/10.1038/s41562-021-01258-7
nature human behaviour
Check for updates

Conspiracy mentality and political orientation across 26 countries

Roland Imhoff[1]✉, Felix Zimmer[1], Olivier Klein[2], João H. C. António[3], Maria Babinska[4], Adrian Bangerter[5], Michal Bilewicz[4], Nebojša Blanuša[6], Kosta Bovan[6], Rumena Bužarovska[7], Aleksandra Cichocka[8], Sylvain Delouvée[9], Karen M. Douglas[8], Asbjørn Dyrendal[10], Tom Etienne[11], Biljana Gjoneska[12], Sylvie Graf[13,14], Estrella Gualda[15], Gilad Hirschberger[16], Anna Kende[17], Yordan Kutiyski[11], Peter Krekó[17], Andre Krouwel[18], Silvia Mari[19], Jasna Milošević Đorđević[20], Maria Serena Panasiti[21], Myrto Pantazi[22], Ljupcho Petkovski[23], Giuseppina Porciello[21], André Rabelo[24], Raluca Nicoleta Radu[25], Florin A. Sava[26], Michael Schepisi[21], Robbie M. Sutton[8], Viren Swami[27,28], Hulda Thórisdóttir[29], Vladimir Turjačanin[30], Pascal Wagner-Egger[31], Iris Žeželj[32] and Jan-Willem van Prooijen[18]

Introduction to the Dataset

- Target: scores on *Conspiracy Mentality Questionnaire (CMQ)*
- Predictors: demographic features, political orientation measures
- Feature engineering of country-level features:
  - GDP and unemployment % from previous year
  - Economist Democracy Index

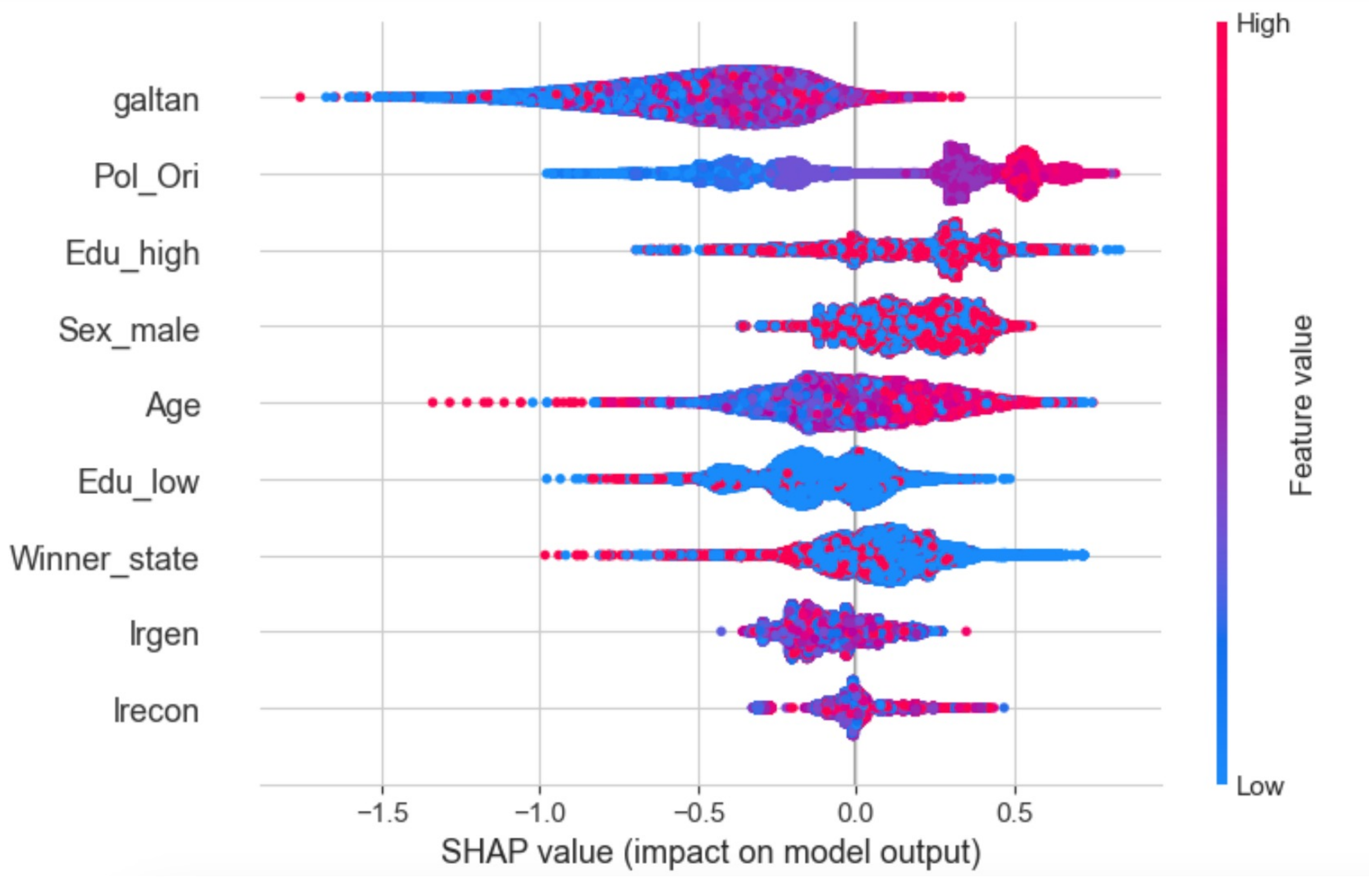
Final Modelling Results

Type of Model	Scoring	Notes
Linear Regression	R^2: 0.19	• Same R^2 with just the top 5 country variables and individual variables; drops to 0.13 with no country variables • Some issues with distribution of residuals
Gradient Boosting Regressor	R^2: 0.27	• Can achieve the same R^2 without country variables; can achieve R^2 of 0.25 with no country-level info (e.g., GDP)
Logistic Regression	Accuracy: 0.66 F1 Score: 0.67 for class 1; 0.65 for class 0	• Slight drop in accuracy using 'balanced' weight setting to address minor class imbalance
XGBoost Classifier	Accuracy: 0.69 F1 Score: 0.72 for class 1, 0.66 for class 0	• Includes all features, but can achieve practically equal performance without any country variables

Interpretation

Shapley values for Gradient Boosting Regressor

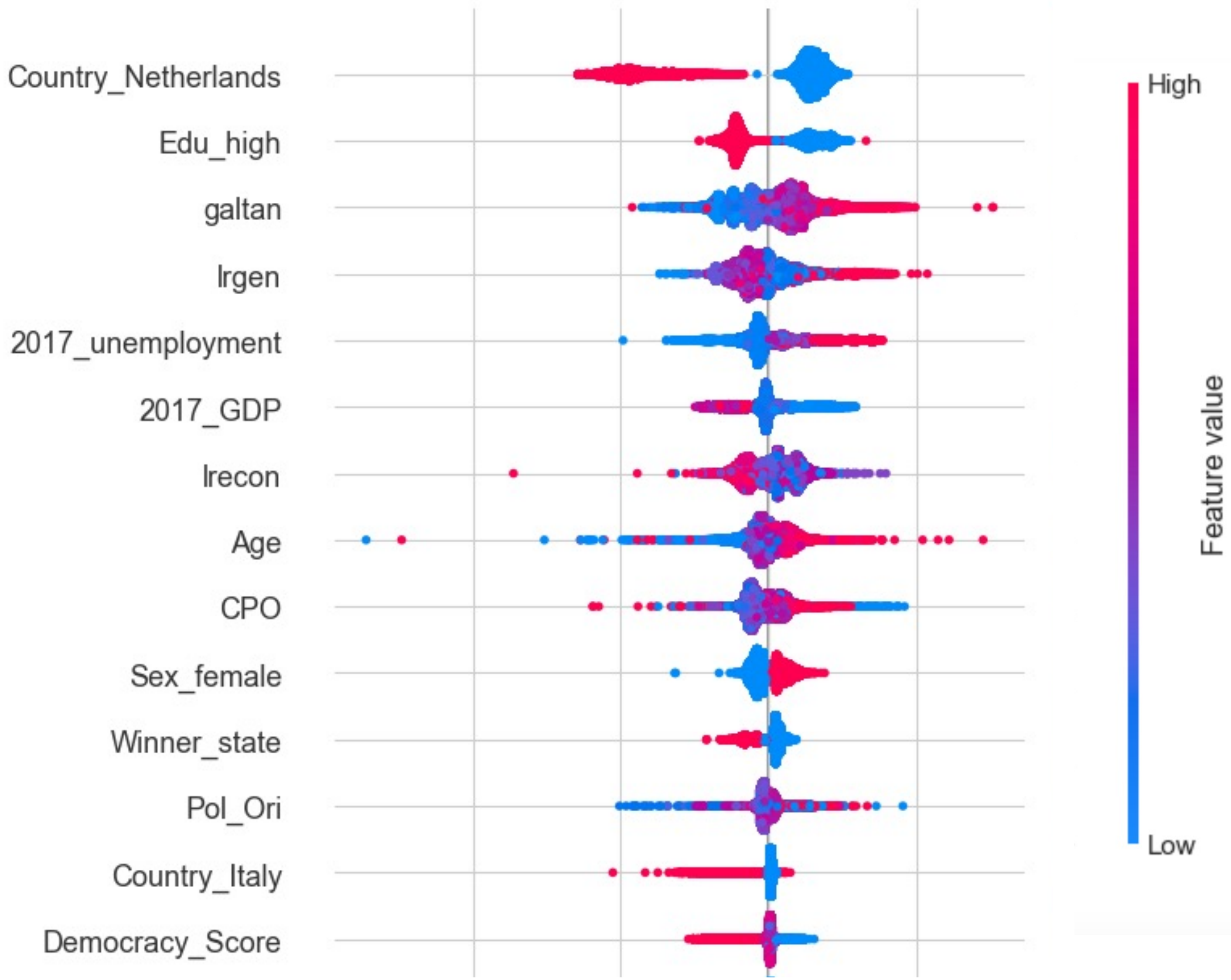
- Individual features only
- R^2: 0.25



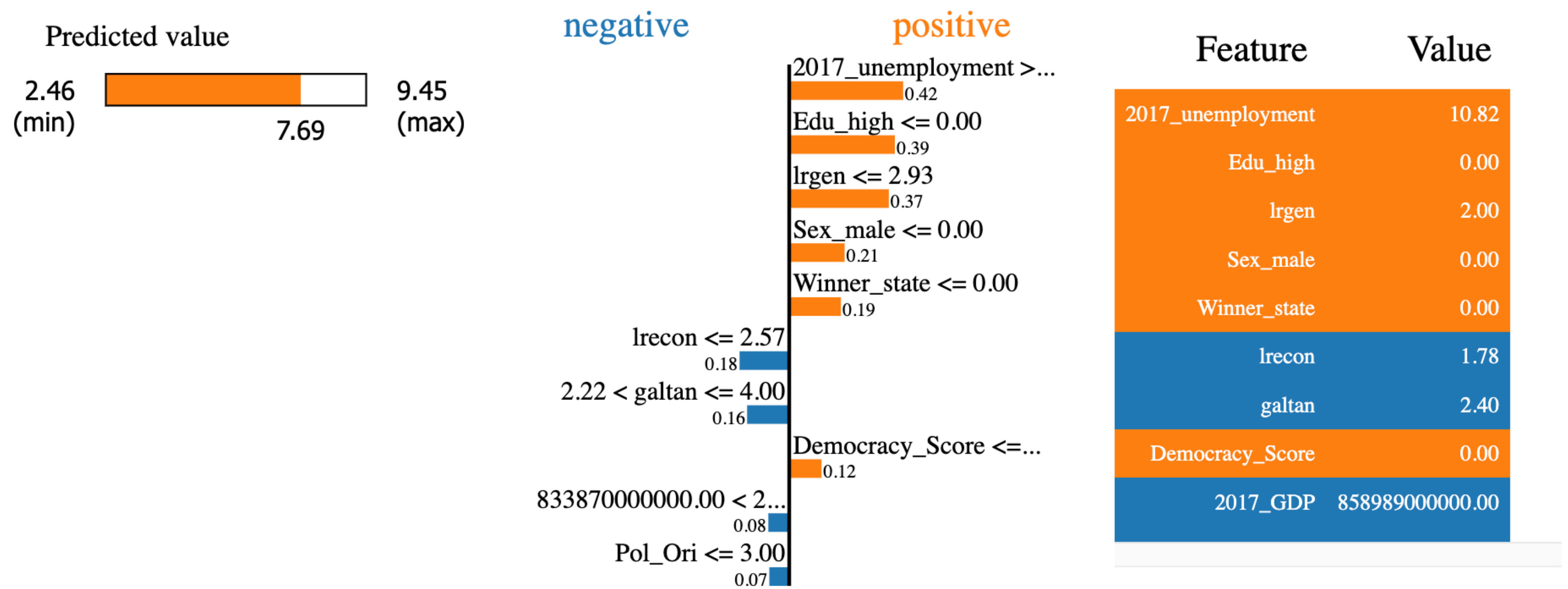
Interpretation

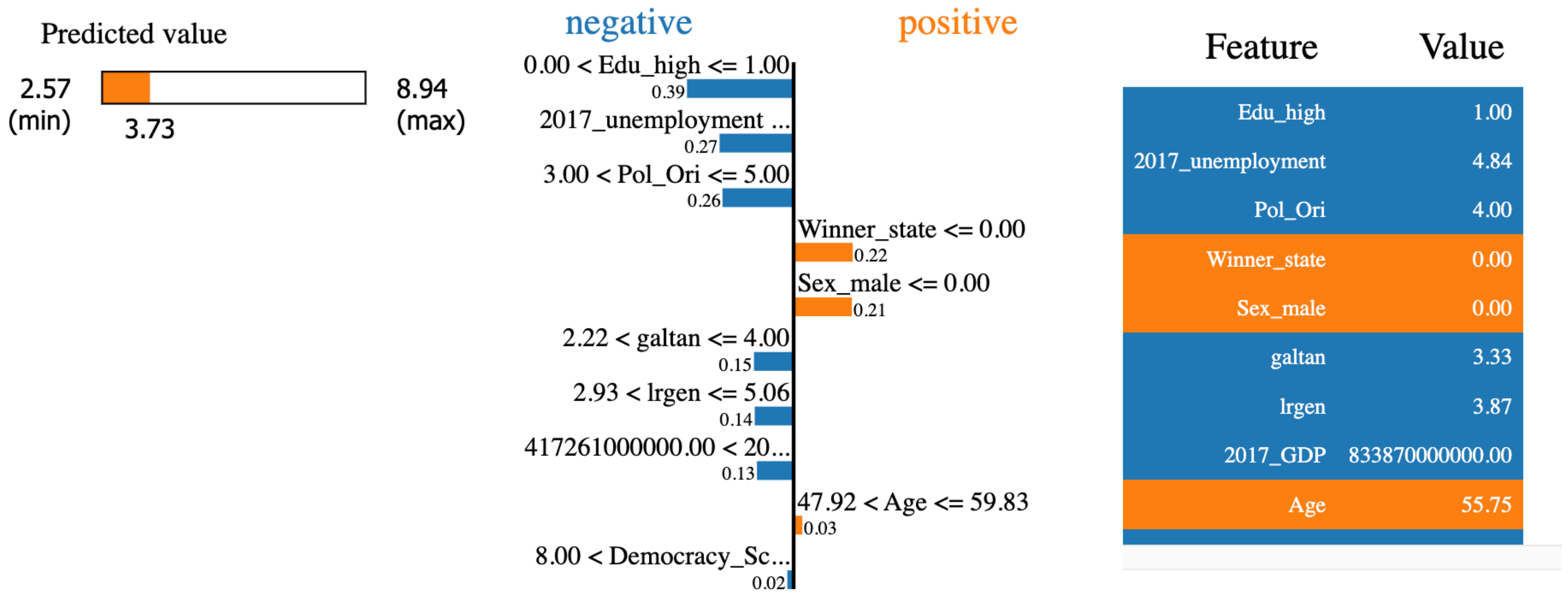
Shapley values for Gradient Boosting Regressor

- All features included
- R^2: 0.27



Interpretation: LIME Results





Conclusions

- Country information largely extraneous (except the Netherlands!)
- Strongest individual-level predictors: *social* political tilt and higher education status
- Strongest country-level predictor: unemployment rate
- Other measures of political orientation show stronger effects in non-linear models

Conclusions

- Without country-level information, model performance stable but features' impacts on predictions become highly non-linear
- Big picture: the social world is complicated!
- Modelling ≠ causal analysis…
- …but seems to support links between CM and social conservatism, political polarization, and economic environment