Hertie School
of Governance

Privacy and Democracy in a Cambridge Analytica World: Predicting Party Choice from Browsing History

Master's Thesis authored by:
Felix Martens – MPP – Class of 2019
Hannah Miles – MIA – Class of 2019

Thesis Supervisor: Dr. Simon Munzert

Table of Contents

List of Tables

List of Figures

Appendices

Executive Summary

The use of online microtargeting by the Brexit and Trump campaigns has fuelled the debate about the impact of the increasing digitalisation of society on privacy and democracy. The ability to predict political preferences based on digital traces is a precondition for successful online microtargeting. It is however not well understood what type of traces can be used to that end. While research suggests that data gathered from social networks is suitable for the prediction of political preferences, there is a gap on another prevalent type of online traces we leave online every day: our browsing history. Insights into the predictive value of browsing history are particularly relevant as this data is widely available to numerous data-driven businesses who use it for advertising, but also resell it to third parties.

Our study contributes to filling this gap. We use machine learning techniques on a dataset of German citizens that contains their complete online history in the four months around the 2017 federal election to assess whether it is possible to predict party choice based on online traces. We focus on Support Vector Machines and Random Forests in particular, because they are highly effective at prediction based on sparse high-dimensional data. All results are assessed against theoretically driven baseline models, in part constructed from socio-demographic data.

This study finds that all but the simplest technique we employed – a Naïve Bayes classifier – were able to pick up on signals in our data that predict party choice. However, the models did not perform consistently across all versions of our data. Of the models we were able to test extensively, none achieved an accuracy that would allow their application for microtargeting operations. Computationally intensive models that we were only able to run once on one version of our data did however perform significantly better and provide a promising avenue for future research.

The importance of microtargeting and the use of big data to predict individual characteristics stands to become even more important in the future. With increasing applications in the political arena and the looming danger of targeted disinformation

campaigns, further research into the predictive value of different types of online traces is needed. Only then can policy makers be adequately prepared for the challenges of the future.

1 Introduction

The surprising success of the 2016 Brexit and Trump campaigns sent shockwaves through the community of political commentators, pundits and pollsters alike. A common theme in the explanation of these electoral upsets has been the crucial role of digital technology, big data and targeted political advertising in making them happen. Both campaigns hired the data analytics consultancy Cambridge Analytica and have been viewed as examples of the integration of new techniques of data collection and analysis into the political process (Paterson & McDonagh, 2018; Gonzales, 2017). Cambridge Analytica's efforts at so called micro-targeting – the tailored provision of political messages based on group or individual characteristics – have been described as a "principal component" of the Trump campaign (Persily, 2017) and have received global public attention in the journalistic account of the 2016 US presidential election.

The technological precondition for this type of campaign is the increasing availability of individual-level online data. Indeed, the Trump campaign made use of the rich datasets of social media, outspending the Clinton campaign threefold on targeted social media advertising (Balswin-Philippi, 2017). Some experts expect that the ongoing digitalisation of our lives will lead to an "algorithmic turn" of democratic electoral processes (Gurumurthy & Bharthur 2018), while others predict that the time of the "individual-centred campaign" has come (Magin et al., 2017).

The data we amass is however only relevant for micro-targeting individuals if it can be used to predict political preferences. Only then can campaign messages be adjusted individually to mobilise potential followers or reduce voter turnout for the opposing site through attack ads - the strategy pursuit by the Trump campaign in 2016 (Balswin-Philippi, 2017).

It remains unclear, however, which types of digital traces that we leave allow for the prediction of political affiliation and which do not. Claims by Cambridge Analytica that the company was able to conduct a psychometric analysis of online traces based on the insights of Kosinski's 2013 study have subsequently been debunked (Balswin-Philippi, 2017). Instead, the Trump campaign mainly used Facebook's built-in targeting tools for its own micro-targeting efforts (Balswin-Philippi, 2017). Researchers have successfully proven that the concept of microtargeting based on social media profiles can be applied to German election campaigns (Papakyriakopoulos et al., 2018). The traces users leave on social media platforms, such as the ones from Facebook that were leveraged by the Trump campaign, are however only one prevalent type of individual-level online data.

The second prevalent type of online traces that advertising and data analytics companies collect is data on the online history of users – primarily the domains they visited. This information stems from online tracking through so-called cookies, fingerprinting or a combination of these methods (Sanchez-Rola et al., 2016). Cookies are small files that are installed on a user's device by a website he or she has visited. Cookies save information about a user's online history that is then retrieved once the user visits another website containing an element – an advertisement, a map, or other content – by the original provider of the cookie. While the effect of cookies can be mitigated, for example by using the "private" mode to avoid the tracking of one's browser history, fingerprinting hardly can. In fingerprinting, an individual browser is identified based on technical settings like the scripts installed for one's browser or the resolution of the screen it is displayed on (Sanchez-Rola et al., 2016). The collectors of this data are big platforms such as Google or Facebook, but also a multitude of other advertising networks. Empirical tests have shown that some 50% of websites use cookies, while 5.5% employ fingerprinting (Acar et al., 2014). If users do not take special precautions to protect their privacy, hundreds of different providers are able to reconstruct more than 40% of a user's total online history (Acar et al., 2014).

It is questionable however to what extent this second type of data – the online history of users – is or could be used to model political preferences through machine learning as a preparatory step for microtargeting. Given the vast amounts of browsing data that are amassed and the rapid development of machine learning technology, this lack of understanding constitutes a problematic blind spot in the research on micro-targeting for political purposes. This thesis therefore investigates whether it is possible to predict political preferences on the basis of domain level data. To do so, we use machine learning techniques, focusing on Support Vector Machines and Random Forests, on a dataset that contains the complete online history of 1162 Germans in the four months around the 2017 federal election. We evaluate our findings against baseline thresholds such as a majority class guess, and a baseline model constructed from socio-economic and political data and compare the model performance. In a concluding chapter, we discuss the implications of our findings, peculiarities of the German case and share the way forward for social science research and online traces.

2 Literature Review

To place our own research in context, we first briefly outline current themes of social science research on the implications of the increasing availability of digital traces (2.1), existing attempts at using them for predictive tasks (2.2 and 2.3) and on why browsing history could contain a signal for the prediction of voting behaviour (2.3).

2.1 Digital Traces, Privacy and Democracy

The ability to model political preferences based on online traces through machine learning is not just a neutral expansion of the political toolkit. It has wider societal implications – both for the individual as well as the collective – that go well beyond the practical considerations of political campaigning (Strandburg, 2014).

At the individual level, the modelling of political preferences challenges our right to privacy. Simply put, the right to privacy protects the individual's sovereignty over its personal information. Subsequently, privacy law rests on the principle that information can only be used by third parties if it was willfully revealed (Strandburg 2014). However, given that online traces can be aggregated over time and given that the tools to make predictions based on that data become ever more sophisticated, most users do not have the capacity to foresee how their traces are used downstream and what this information reveals about them (Zwart et al., 2014). Even worse, the digital nature of one's profile and the sharing of information between different service providers makes it difficult to erase information once it has been stored (Bossewith & Sinnreich, 2012).

On the collective level, observers fear that the increasing capability of machine learning undermines the ability of democracies to hold free and fair elections (Polonski, 2017). If successfully used by political entrepreneurs, the modelling of political preferences could enable them to game the political system. This could for example be achieved through the ever more precise drawing of electoral districts (so-called gerrymandering), or the undermining of the existence of a shared democratic "public" (Gurumurthy & Bharthur, 2018). By transforming a formerly overt political discourse into covert individualized campaign scripts, political entrepreneurs could use the speed and short-lived nature of online discourses to further their goals (Gurumurthy & Bharthur, 2018).

Since we are analyzing the online traces of German citizens, we want to highlight the German context in particular. In Germany, both the individual privacy risk as well as the collective risks emanating from the potential distortion of the political system are relatively low. This has primarily two reasons. First, the current campaigning practice and culture of German political parties preclude the use of microtargeting and the necessary individual level modelling of political preference. Although some parties in Germany – the CDU SPD in particular – have started limited efforts at collecting individual-level data, they currently deliberately only target down to the street level, not the individual level (Kruschinski & Haller, 2017). This is the case due to the second reason mitigating risks in Germany: European privacy law. Article 9 of the European Union General Data Protection Regulation (EU GDPR) sets restrictive limits on the processing of data revealing political orientation (Benett, 2016).

Even so, systemic risks emanating from the use of online data to predict party preference remain for Germany. As several recent elections have demonstrated, election meddling through online channels by external political actors is real and on the rise. During the 2017 French election, for example, messages posted by bots on Twitter reached almost 23 million citizens per hour during the height of the so called "Macron leaks" (Ferrara, 2017). Access to data that allows analysts to model users' political preference could therefore be used to increase the effectiveness of microtargeted disinformation campaigns employed by foreign powers.

2.2 Existing work in social media measurement

Due to the wealth of data the internet offers, there have been numerous studies using digital data to make predictions. As a reflection of human behaviour, internet search volume can be used to 'predict the present', reflecting a correlation with current macro-level events, such as house prices and unemployment levels (Goel, Hofman, Lahaie, Pennock, & Watts, 2010), flu rates (Ginsberg et al., 2009) or the evolution of pandemics

(Lampos & Cristianini, 2010). Such data has also proven highly predictive for future collective behaviour, such as box office revenue (Asur & Huberman, 2013), song rankings and sales rates (Goel et al., 2010).

Social media presents an entirely new trove of information. The ability to like and follow content of interest offers a far more detailed image of individual users, leading to numerous studies using social media data for prediction purposes. Such studies have utilised a range of statistical models, including supervised machine learning, using data such as the volume of social media mentions as a predictor, a sentiment analysis of the content of such mentions and other aspects of online behaviour and activity.

One of the earliest contributions specifically for election prediction was the study by (Tumasjan, Sprenger, Sandner, & Welpe, 2010)) concerning the 2009 German federal election. The study asks whether online Twitter messages validly mirror offline political sentiment and can thus be predictive of vote outcome in the aggregate (Tumasjan et al., 2010). Despite claiming positive results - that the ranking of tweet volume (i.e. number of tweets) mentioning a political party corresponds to the ranking by share of vote in the election results (Tumasjan et al., 2010) – the study has been the focus of criticism by Chung and Mustafaraj (2011), Jungherr, Jürgens and Schoen (2012), and Metaxas, Mustafaraj and Gayo-Avello (2011). Jungherr et al. (2012) focus on systematic problems with the methodology and the difficulties this poses for replication, for example the seemingly arbitrary choice of parties and time frame. Indeed, had the study included the Pirate Party in their analysis, it would have been predicted as the overall winner of the German election (Jungherr et al., 2012), demonstrating the overrepresentation of small parties in socially generated data (the Pirates in fact received just 2.0% of the overall vote).

Count metrics have been the focus of criticism beyond Tumasjan et al.'s study. Studies using Twitter-based count metrics rely on the assumption that tweet volume accurately reflects the current thinking of a large proportion of the population (Lui, Metaxas, & Mustafaraj, 2011). Jungherr, Schoen, Posegga, & Jürgens (2017) provide an alternate interpretation in the face of their own unsuccessful attempt to predict the 2013 German federal election using Twitter data. They argue that, rather than indicating voting intention, Twitter-based count metrics reflect users' attention to politics (Jungherr et al., 2017). Thus, political attention may be a covariate of political support in some cases, but not in all (Jungherr et al., 2017).

Even if such metrics were reflective of voting intention in the aggregate, they would only reflect the intention of the Twitter population, which is demographically very different to the general population (Metaxas et al., 2011; Yasseri & Bright, 2016; Lui et al., 2011). Because of this demographic difference, there is little theoretical reason to expect tweet volume to be proportionate to the overall vote share of a party, resulting in potentially spurious correlations (Yasseri & Bright, 2016). Selection bias is also significant here, as those who are politically active produce the most tweets about politics (Gayo-Avello, 2013, Mustafaraj, Finn, Whitlock, & Metaxas, 2011). Mustafaraj et al. (2011) refer to this as the 'vocal minority' versus the 'silent majority'. Our study avoids this bias in two ways: we are focussing on individual level prediction, meaning that our data does not need to represent the general population, and we take a broader look at an individual's entire online presence, thus removing bias inherent to individual platforms.

Applying a range of more complex machine learning algorithms to similar data, Beauchamp looks at US-State level polls in the 2012 Presidential campaign and attempts to predict polling outcomes with political tweets using time series data (Beauchamp, 2017). The study tests an elastic-net linear model (using a combination of lasso and ridge regularisation), a support vector machine, and a random forest (Beauchamp, 2017). The study finds its elastic-net linear model to be most successful in producing similar results to polls, though none of the methods were more accurate than simply extrapolating from state polling alone. Nonetheless, this study shows that more complex machine learning algorithms can provide an advantage over simple count metrics.

Rather than relying simply on volume, studies have used sentiment analysis to produce predictions using Twitter data. O'Connor, Balasubramanyan, Routledge, & Smith (2010) analysed the sentiment of Twitter messages (positive or negative), ultimately finding no correlation between electoral polls and Twitter sentiment. Ceron, Curini, Iacus, & Porro

(2014) also employed sentiment analysis on tweets in analysing online popularity of Italian political leaders in 2011 and voting intention of French internet users in 2012. The authors found a strong correlation with both traditional mass surveys and electoral results, building on research using sentiment analysis of tweets to predict the results of the 2011 Dutch senate election (Sang & Bos, 2012). Bermingham & Smeaton (2011) attempted to predict vote share using both volume and sentiment measures of tweets surrounding the 2011 Irish General Election. The study found volume to be a stronger predictor than sentiment, although the authors' predictions were still less accurate than traditional polling methods (Bermingham & Smeaton, 2011).

As can be seen, the findings on sentiment analysis remain inconclusive. This could be due to the fact that sentiment analysis struggles to pick up the nuances of political language in particular, which often include irony or sarcasm, and thus faces reliability issues (Yasseri & Bright, 2016). Indeed, with regard to analysing social media data, Metaxas et al. (2011) warn that positive outcomes are likely down to chance.

In sum, studies using social media data to make predictions face reliability issues due to the complexities of political discourse in sentiment analysis, demographic bias and self-selection bias (Gayo-Avello, 2013). Furthermore, comparisons to sensible baseline models are often missing (Gayo-Avello, 2013). We will address this issue directly in Section 4.2 . Finally, it is worth noting that all studies using social media for prediction are post-hoc studies. Currently, there are no studies using social media correctly and consistently to predict the results of elections before they occur. This represents a major weakness in social media prediction literature thus far, indicating that using social media for prediction is not yet as advanced as some of the more positive results imply.

2.3 Individual Level Prediction

While there are many studies seeking to predict overall election outcome, fewer studies attempt to predict individual voting behaviour. Rao, Yarowsky, Shreevats, and Gupta (2010) successfully used Twitter data, including simple sociolinguistic features, number of followers and rate of retweets to detect latent attributes, such as gender, age and political orientation using a stacked-SVM classification algorithm. Pennacchiotti and Popescu (2011) attempt to predict similar characteristics, also focussing on linguistic content, behaviour and network structure of the user's Twitter feed. The study found that their models performed best on classifying political orientation (Democrat vs Republican) with an overall accuracy of 80% (Pennacchiotti & Popescu, 2011).

Indeed, even where studies undertake individual-level prediction, few attempt to predict the party an individual will vote for in multi-party elections. Kristensen et al. (2017) is one such study, which attempts to predict individual political party preference from Facebook likes of political parties and figures. However, this study only looks at political likes on Facebook, and by extension only includes those individuals who 'like' political parties on Facebook, selecting on the dependent variable and facing selection bias issues.

We are not aware of any study looking specifically at predicting individual vote choices using comprehensive online data. In particular, little previous work has focused on multi-party electoral systems. This may be because it is much harder to predict party choice in multi-party systems such as Germany with a high level of accuracy compared to simple binary outcomes (Kristensen et al., 2017). Thus, our study provides important insights into the relationship between comprehensive, domain level data and individual party choice.

2.4 Why might online data be predictive? A theoretical basis

Above we have given some examples of prediction tasks conducted on the basis of online traces, including the prediction of voting behaviour. Unlike for example Kristensen et al. (2017), which relies on Facebook likes of political parties and individuals, we do not solely rely on online behaviour that can be viewed as direct expression of political affiliation. Yasseri & Bright (2016) and Metaxas et al. (2011) caution against prediction without having any theoretical understanding of why it could work. The former pair makes the point that this would heighten the chance of misinterpreting results that are the outcome of mere chance. It is therefore necessary to answer the question: why should general online behaviour be predictive of individual voting behaviour? In addition to online behaviour that is a direct expression of political affiliation, other online behaviour must capture individual properties that are also predictive of party affiliation for our attempt to be reasonable. Indeed, the literature gives us reason to expect that this is the case, as online behaviour correlates with other variables such as personality traits, gender, age, social status and ideology, which are also predictive of party preference.

2.4.1 Personality Traits

Youyou, Kosinski, & Stillwell (2015) show that general Facebook likes are predictive of the so-called 'Big Five' personality traits. Kosinski, Bachrach, Kohli, Stillwell, and Graepel (2014) use both Facebook likes for websites and an online behaviour questionnaire to show that it is possible to identify website audience personality profiles. That means that websites differ regarding the mean score of their users on the dimensions of the Big Five personality traits and that therefore online behaviour can be predictive of personality. This is highly relevant for the prediction of voting behaviour based on online traces, because personality traits have been shown to be indicative of political behaviour in general and specifically party affiliation.

Personality is an enduring psychological structure that influences patterns of behaviour (Mondak, 2010). There have been numerous taxonomies of personalities with differing dimensions throughout the 20th century. However, since its first description in the late 1950s and subsequent refinement in the 1980s the Big Five personality dimensions ("Big Five") have become the dominating paradigm. They have been described as a human universal applying cross-culturally (McCrae & Costa, 1997) and have been shown to be very time stable (Soldz & Vaillant, 1999). As stated above, the Big Five have been used to explore political behaviour – mostly in the past 25 years or so – regarding patterns of interest in politics and access to political information (e.g. Gerber, Huber, Doherty, Dowling, & Ha, 2010), political participation (Mondak, Hibbing, Canache, Seligson, & Anderson, 2010) and political orientation (e.g. Capara, Barbaranelli, & Zimbardo, 1999). With respect to political orientation, the research confirmed the effect of two traits in particular: the positive relationship of conscientiousness with a right orientation and openness with a left orientation.

Personality traits have also been matched with party choice, for example for Germany by Mondak, Hibbing, Canache, Seligson, & Anderson, 2010.

2.4.2 Personal and Socioeconomic Factors

In addition to personality traits, online traces pick up on personal and socioeconomic factors such as age, gender or social status, because internet activity can be seen as a reflection of activities and relations that exist offline (Tirado-Morueta, Aguaded-Gómez, & Hernando-Gómez, 2018). All of the factors are likewise linked to party choice. They may therefore well reinforce or leave signals in our data that can be picked up through machine learning techniques.

With respect to age, so called life-cycle perspectives on voting behaviour predict that voters tend to become more conservative as years go by. This relationship has been shown to hold for West Germany (Goerres, 2008). Age is likewise a significant factor for patterns of online use, both with respect to the type of websites that are accessed, and the time spent online (Tirado-Morueta et al., 2018; Mellon & Prosser, 2017).

Like age, gender is another important personal characteristic related to party choice and online behaviour. Women in western democracies, including in Germany, tend to vote for left wing parties (Abendschon & Steinmetz, 2014). With respect to online behaviour, research indicates that gender affects the type of content accessed, with the main difference being greater male proclivity to use the internet for entertainment purposes such as gaming or watching movies (Bujała, 2012; Fallows, 2005).

Social status – conceptualized as a combination of education, income, occupational prestige and lifestyle characteristics – has been shown to influence internet usage with respect to the type of content sought out online for German users (Zillien & Hargittai, 2009). High status individuals have for example a higher propensity to access domains on economic or political news, travel and use e-mail services as opposed to chats. Subjective social status also has an influence on voting behaviour, with lower social status increasing the likelihood that an individual vote for a right-wing populist party (Gidron

& Hall, 2017). For Germany specifically, education and social class have also been shown to be significant for voting behaviour (Schoen & Schumann, 2007).

2.4.3 Ideology

Social psychology has empirically observed that individuals prefer information that confirms their own decisions and worldviews and has developed selective exposure theory to explain the phenomenon (Kastenmüller, Greitemeyer, Jonas, Fischer, & Frey, 2010). The theory is based on the idea that individuals, after having made a decision, seek to avoid the negative emotional state of cognitive dissonance that arises out of exposure to the negative aspects of their choices or the good aspects of alternatives not chosen. This results in information searching behaviour that is inherently biased towards prior held believes and ideas – confirmation bias (Jonas, Schulz-Hardt, Frey, & Thelen, 2001). The rationale of confirmation bias has been applied to the theoretical understanding of news markets (Mullainathan & Shleifer, 2005) and has been empirically confirmed for television stations in the United States (Iyengar & Hahn, 2009). The idea that the internet will reinforce these tendencies, because it allows individuals to finetune their information consumption to a greater degree than any other medium is almost twenty years old (Sunstein, 2001) and has been affirmed early on for the blogger sphere (Glance & Adamic, 2005). Empirical research from the United States based on browsing history and a segregation index has shown that general online news consumption is also biased based on political belief – and that the bias is more pronounced than with respect to television news (Gentzkow & Shapiro, 2011). This is relevant for the prediction of party choice based on browsing data, because ideology – usually conceptualized as left-right self-placement – has been shown to be predictive of voting behaviour (Devine, 2015; Walczak, 2010). For the German case ideology has empirically been shown to be significant for voting behaviour, albeit to different degrees depending on the political context and polarization at a given point in time (Mader & Schoen, 2017).

3 The Data

3.1 Survey Design

The data for this project stems from the German YouGov Pulse panel survey “Paying Attention to Attention: Media Exposure and Opinion Formation in an Age of Information Overload”. In addition to participating in a five-wave survey between July and October 2017, sampled panellists (N = 1500) installed the web-tracking program “Wakoopa” on their computers and mobile devices, which comprehensively recorded each individual’s web usage. This study is based on domain level data (up to the first slash) for all individuals in the survey. In addition to the web data, the survey contained a range of socio-demographic and political questions, including voting intention pre-election, and a question on actual vote post-election. The survey launched on July 13, 2017, about ten weeks before the election date.

The 1,500 participants sampled for the web-tracking study were selected from the Pulse panel to match the marginals of the German internet population on three variables: age, gender and education level, as identified by the Best for Planning study (Best for Planning, 2017). Although the sample contains voters for all six major political parties, the vote share in our sample is not the same as the outcome of the federal election; we see an overrepresentation of smaller parties (see Figure 12). However, considering the small sample size, it is important that we have a sufficient number of representatives of smaller parties. Furthermore, since we are predicting individual level vote and not the outcome of the election, it is not important for the sample to reflect the real result exactly. Thus, the sample provides substantive variation in internet use according to age, gender and education, which is an important input for our predictive model.

3.2 Outcome measure

We are interested in predicting which party an individual will vote for. In German federal elections, each voter casts two votes: one for a candidate from their constituency, and the second for a party. The second vote is our variable of interest. This variable contains the following levels:

1. CDU / CSU

2. SPD
3. FDP
4. Green
5. Linke
6. AfD
7. Other party
8. Don’t know (also coded 977 in some waves)

Voting decision is asked in multiple waves of the study. We created a new variable, “secondvote”, which includes the data from survey wave 5 (post-election, i.e. who the individual actually voted for), or if this was not available, the voting intention from the 4th wave. 1307 out of 1516 individuals responded either in the 4th or 5th wave. If neither was available, the observation was dropped since adding any earlier waves may be unreliable due to the distance in time from the election. All observations coded 8 or 977 (“Don’t know”) were removed.

3.3 Pre-processing

It was necessary to clean the domain-level data before beginning the analysis. Firstly, we sorted the data by domain and removed any domains which had only been visited once, or had a total viewing time of less than 60 seconds. Domains visited only once, and therefore by only one person, will have no explanatory power and could therefore be safely discarded. Similarly, domains visited for a total of less than 60 seconds are likely to be incidental data artifacts or unintentional clicks and therefore not reflective of party choice.

Secondly, we removed any domains which included commas or long strings of digits, as these are not valid URLs.

In order to be able to differentiate between sub-domains, domains and top-level domains, we split the original ‘domain’ variable into three new variables, as below:

fulldomain	subdomain	domain	toplevel
apps.facebook.com	apps	facebook	com

This allows us to aggregate summary statistics based on domain, for example grouping all Facebook visits regardless of whether they were accessed through a '.de' or '.com' top-level domain. A similar parsing method was used to change 'used_at', containing date and time of the access, into separate 'date' and 'time' variables.

To ease handling the data and training machine learning models, the data was converted to wide format, with one variable per domain. We produced three versions of the wide format data, with values for the cumulative duration in seconds, cumulative number of visits and a simple Boolean indicator for whether or not the domain was visited, presented below.

Duration (in seconds)

Personid	facebook	google	youtube
Person 1	0	2000	3900
Person 2	12000	3340	0

Number of visits

Personid	facebook	google	youtube
Person 1	0	70	15
Person 2	100	19	0

Boolean

Personid	facebook	google	youtube
Person 1	0	1	1
Person 2	1	1	0

A categorical variable indicating party choice ("secondvote") was appended to each data frame (see Section 3.2).

3.4 Dimension reduction

At this stage, the dataset still contained 99,133 unique domains, each recorded as a variable. In order to make the data more manageable and remove further irrelevant observations, domains were removed that did not have enough traffic.

Removal of single user domains

Firstly, we removed all domains which had only been visited by a single individual, as these would provide no predictive power. Figure 1 shows the number of unique visitors per domain. The largest bar represents those domains which only had a single visitor. Removing these single-visitor domains removed 51,330 variables, more than halving the original number of domains.

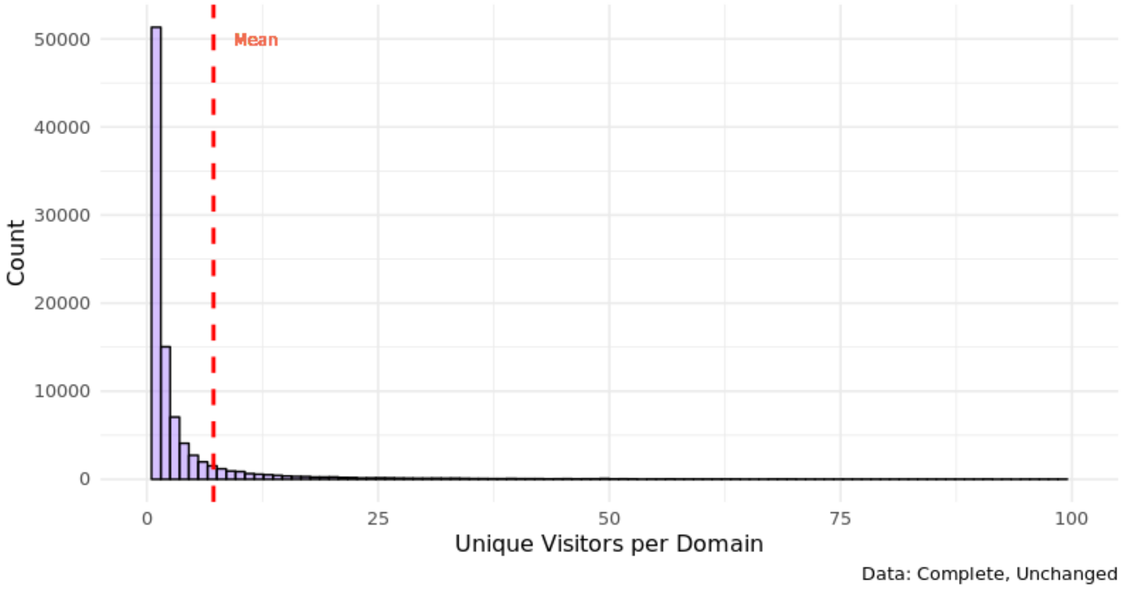


Figure 1: Histogram of Unique Visitors per Domain

Removal of rarely visited domains by duration and dealing with skew

When examining the aggregate duration of the domains, we find that many domains were visited for a very short amount of time in total. We set an arbitrary, low cut-off point at an aggregate time of 60 minutes over the entire survey period by all survey participants - any domain which did not reach this threshold was dropped. Despite being a very low bar, the data was nevertheless reduced to 9,327 unique domains. In other words, 90.6% of all domains had been removed. This alone is an interesting finding – in the aggregate, participants appear to spend the vast majority of their time online on very few domains. These domains were then selected from the duration, visits and Boolean datasets described above.

Figure 2 is a density plot of the cumulative duration of domains in hours once domains visited for one hour or less were removed. For visualisation purposes, the density plot shows up to 100 hours only, when in fact the maximum number of hours spent online was 41,564. This demonstrates that the duration metric still suffers from significant right skew. We then decided to log the data and test both logged and unlogged versions on the models. Every value was incremented by one to avoid taking the logarithm of zero. The logged distribution is presented in Figure 3.

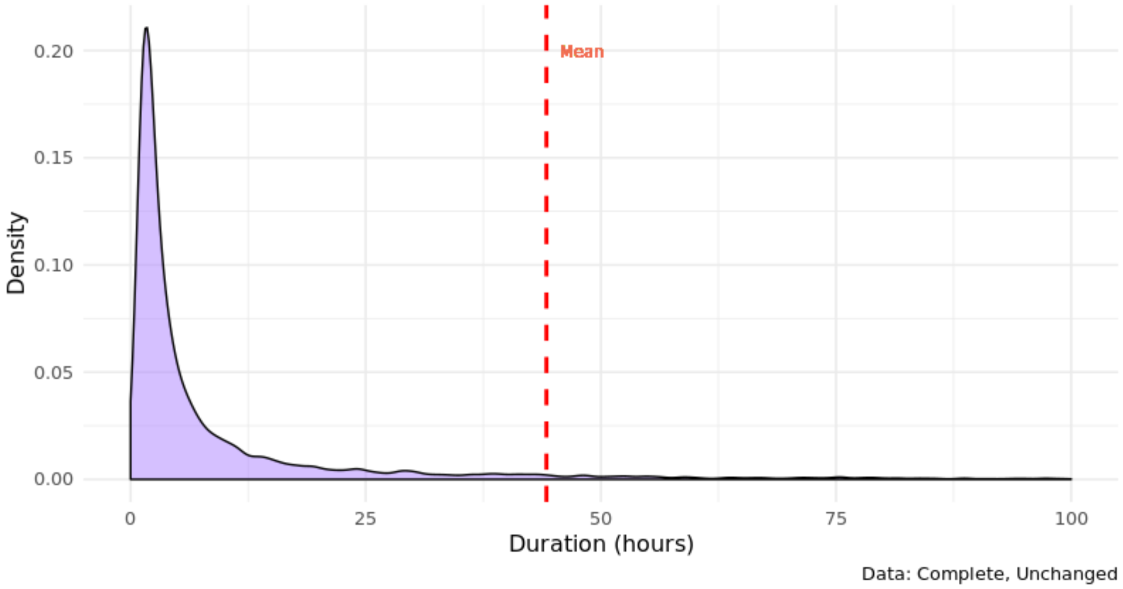


Figure 2: Density Plot of Duration of Domain Visits. (Note: x axis limited to 100, actual maximum = 41,564 hours)

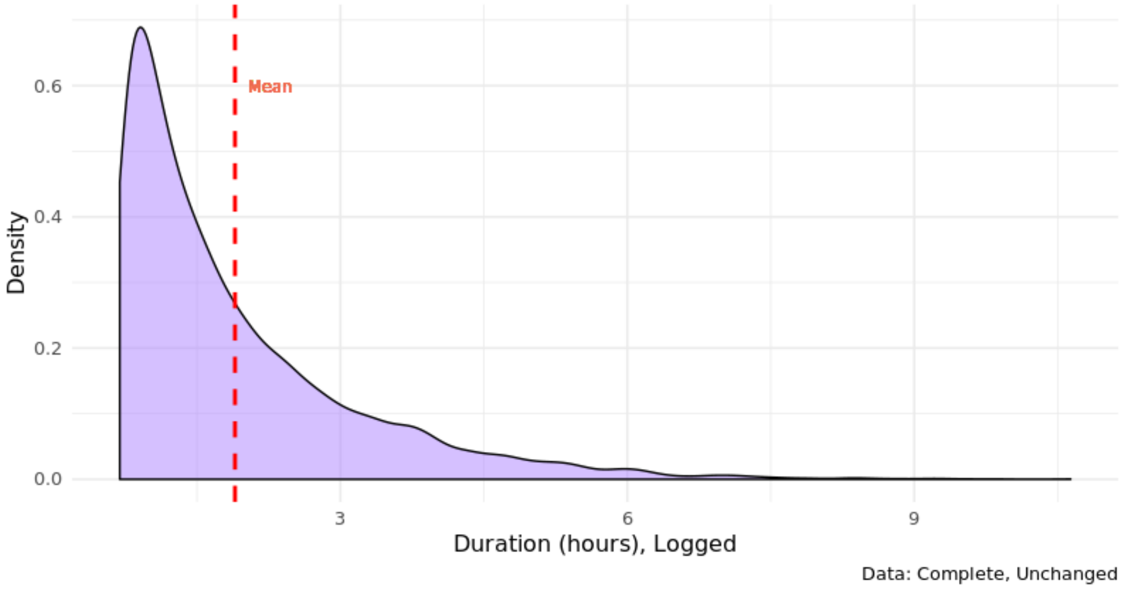


Figure 3: Density Plot of Domain Visits, Logged

Similarly, the distribution of the number of visits per domain also exhibited right-skew (Figure 4) and was subsequently logged (Figure 5).

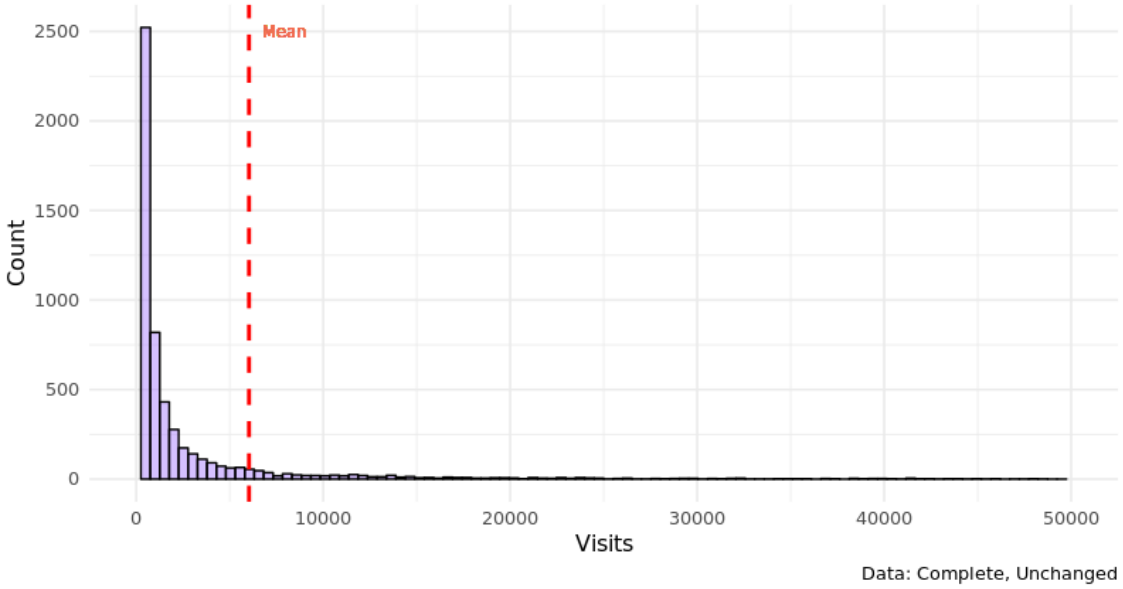


Figure 4: Histogram of Domain Visits

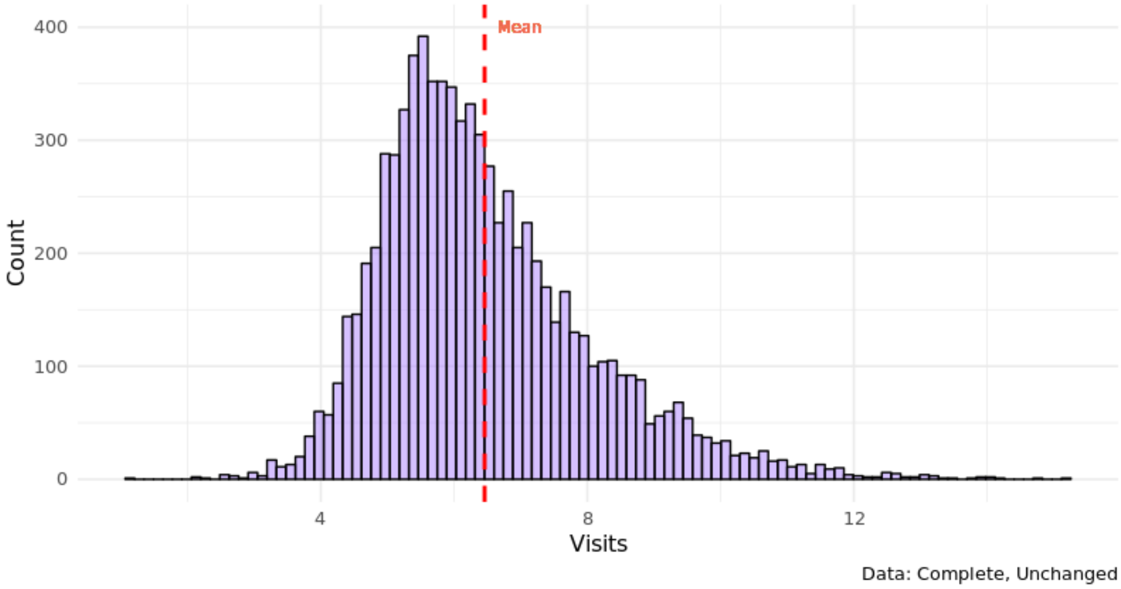


Figure 5: Histogram of Domain Visits, Logged

Finally, one larger data-frame was then created combining duration, number of visits and a Boolean indicator for those domains identified as relevant in the previous step. We hypothesise that the extra information provided by all three indicators may improve the model compared to using just one in isolation.

Removing rare users

When examining the distribution of individuals by time spent online, we observed significant right skew, meaning many participants were online for a comparatively short amounts of time over the survey period, with a few outliers who spent longer online. Figure 6 presents a density plot of the duration in hours that individuals spent online over the survey period.

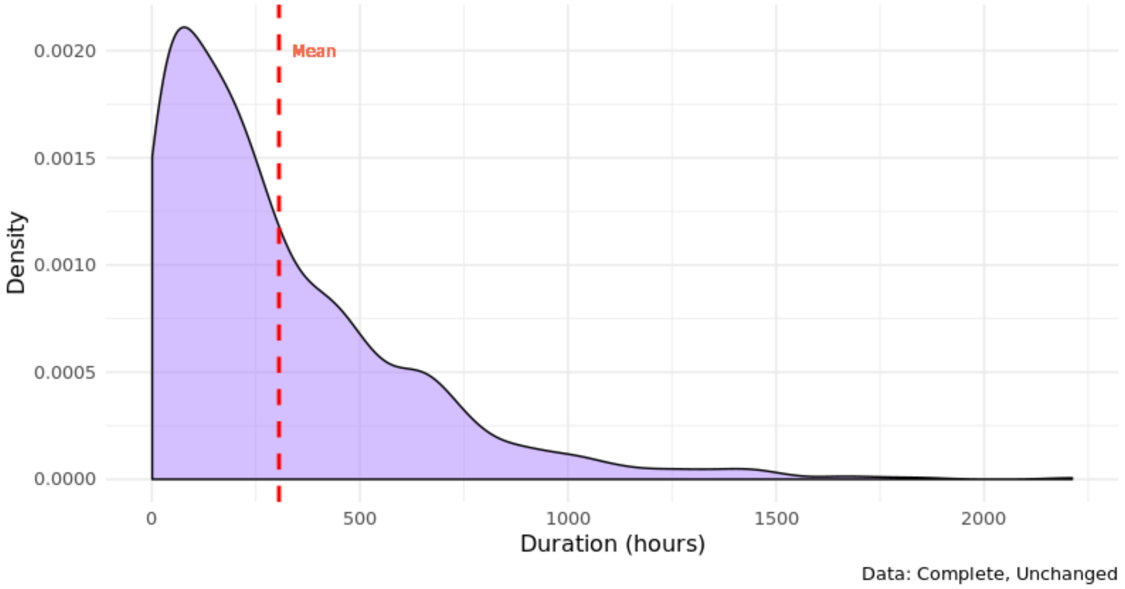


Figure 6: Density Plot of Individual Online Duration

We hypothesised that individuals who spent too little time online were unlikely to have produced enough online data to predict our outcome of interest. Therefore, we defined an arbitrary, low cut-off point of one hour online over the whole survey period. Any individual who did not meet this threshold was removed. This reduces our sample from 1199 to 1162 individuals.

However, recognising that this is a low bar and that many individuals who are not very active online remain in the sample, we created another version of our datasets in which we removed all people who were online for less than 16 hours - 1 hour per week for the duration of the study. In the model testing phase, it was tested whether removing these individuals increased model accuracy.

3.5 Descriptive Stats on Data

Our core dataset resulting from the pre-processing described above displays the following characteristics.

We have complete data on a total of 1162 individuals. With respect to our outcome variable party choice, there is mild class imbalance with the share of "SPD" voters (22.1 %) as largest class some three and a half times bigger than the smallest class "Other Parties" voters (6.4 %). Since there is most information on the SPD, CDU/CSU and Linke in our dataset, we expect our models to pick up a particularly strong signal with respect to these classes.

Table 1: Summary Statistics Online Behaviour

Party	*Cases*	*Share*	*Mean Hours Online*	*SD Mean Hours Online*	*Domains Visited*	*SD Domains Visited*	*Share of Bild Readers*
CDU/ CSU	228	19.6 %	304	309	380	256	58.8 %
SPD	257	22.1 %	274	266	358	260	53.7 %
FDP	113	9.7 %	292	283	361	243	55.8 %
Green	105	9.0 %	283	270	353	248	45.7 %
Linke	219	18.8 %	331	315	374	229	53.4 %
AfD	166	14.3 %	360	348	423	265	65.1 %
Other Parties	74	6.4 %	275	267	374	248	62.2 %

The table above shows the online behaviour of voters by party. It illustrates that there are tangible differences in online behaviour between the voter groups, for example with respect to the time spent online and the content accessed. AfD voters spent on average 360 hours online over the four months of data collection, while SPD voters only used the internet for 274 hours.

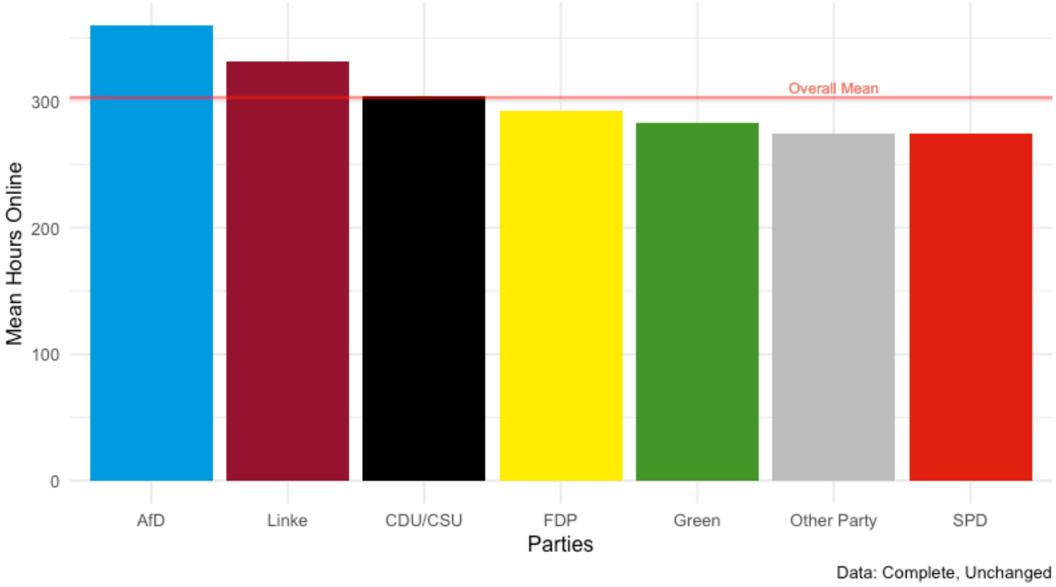


Figure 7: Mean Hours Online by Party

AfD voters also on average accessed the largest number of different domains (423), while Green voters displayed the lowest diversity in domains accessed (353). However, the standard deviations provided for both the time spent online and domains accessed shows that there is great variability even within party voter groups with respect to these measures.

As we will discuss further below, it is likely that these metrics are at least in part driven by intensive users in the respective groups. We also provide the share of party voters who accessed the domain of Germany's most successful tabloid Bild as an example for the variance in content sought by the members of different parties. Again, the largest share of Bild readers are AfD voters (65.1 %), while only party with a minority of Bild readers and the lowest score overall are the Greens (45.7 %).

In addition, the plot above visualises an example for the variation in online behaviour between the voters of all different parties on one domain. It displays the variation in time spent on Ebay, with FDP and AfD voters using Ebay most intensively. The plot below shows that the voter group that uses a domain most intensely is not merely a function of the total time that group spends online. While Green voters are third to last in terms of mean duration online, they have the highest density of intensive users of the social media platform Instagram.

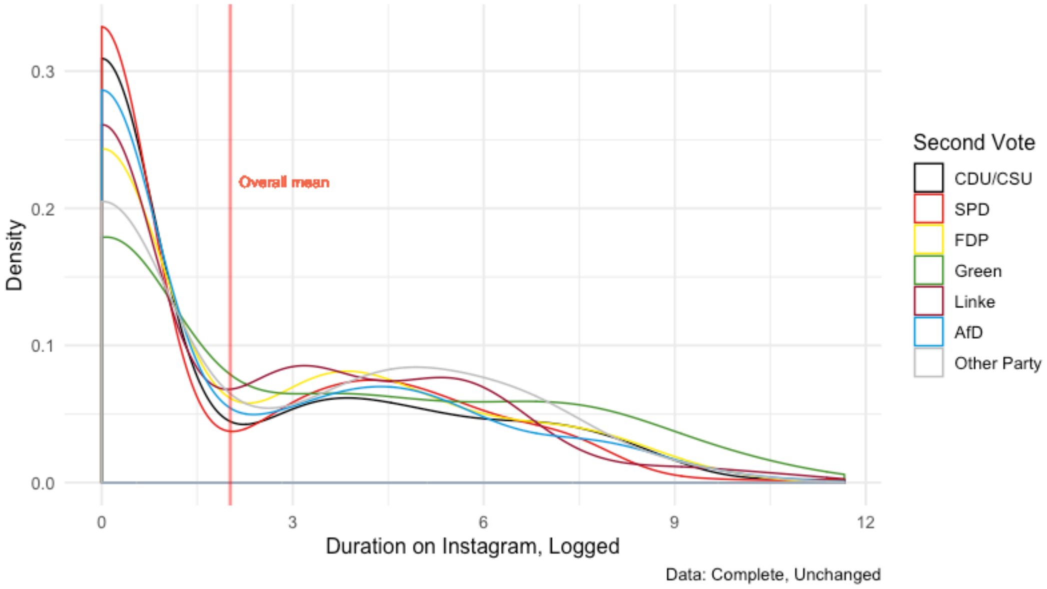


Figure 8: Density of Duration on Instagram

These findings suggest that there is indeed variance with respect to online behaviour and party choice in our data that we can hope to exploit in our attempt to predict party choice through machine learning. It also shows however that despite the differences, the curves follow quite similar trajectories.

Table 2: Summary Statistics Socio-Demographic Variables

Party	*Cases*	*Share*	*Female Share*	*Mean Ideology*	*Mean Age*	*High Education*
CDU/ CSU	228	19.6 %	44.7 %	6.4	49	42.5 %
SPD	257	22.1 %	45.1 %	5.2	48	40.1 %
FDP	113	9.7 %	36.3 %	6.4	48	46.9 %
Green	105	9.0 %	52.4 %	5.0	42	63.8 %
Linke	219	18.8 %	52.5 %	4.2	48	48.4 %
AfD	166	14.3 %	45.8 %	7.2	48	34.9 %
Other Parties	74	6.4 %	45.9 %	5.8	40	41.9 %

In section 2.4 above we have discussed other characteristics predictive of party choice that might leave traces in the online data collected: inter alia gender, age, social status as a function of education and ideological leaning as left-right self-placement. The table above shows that there is variation with respect to our outcome variable across all of these dimensions.

Gender as share of female voters for example is lowest for the FDP (36.3 %) and almost equally high for both Die Linke (52.5 %) and the Greens (52.4 %). The variation in age is limited with most parties displaying a mean of 48. The CDU/CSU has a slightly older voter base though (mean 49 years) and both the voters of the Greens (mean 42 years) and other parties (mean 40 years) are on average considerably younger. With respect to education, the Greens and the AfD again constitute the extreme cases in our data. We conceptualised the share of highly educated as the share of voters who attained a secondary-school diploma that qualifies them to enter college in Germany (either Abitur or Fachabitur). 63.8 % of Green voters fulfil this criterion, while only 34.9 % of AfD voters do. Unsurprisingly the party voters also vary by ideological self-placement on a one to eleven scale. The bar-chart below illustrates that AfD voters place themselves furthest to the right (mean score 7.2) while voters of Die Linke place themselves furthest to the left (mean score 4.2). Interestingly, those individuals who vote for “Other Parties” are closest to the overall mean score.

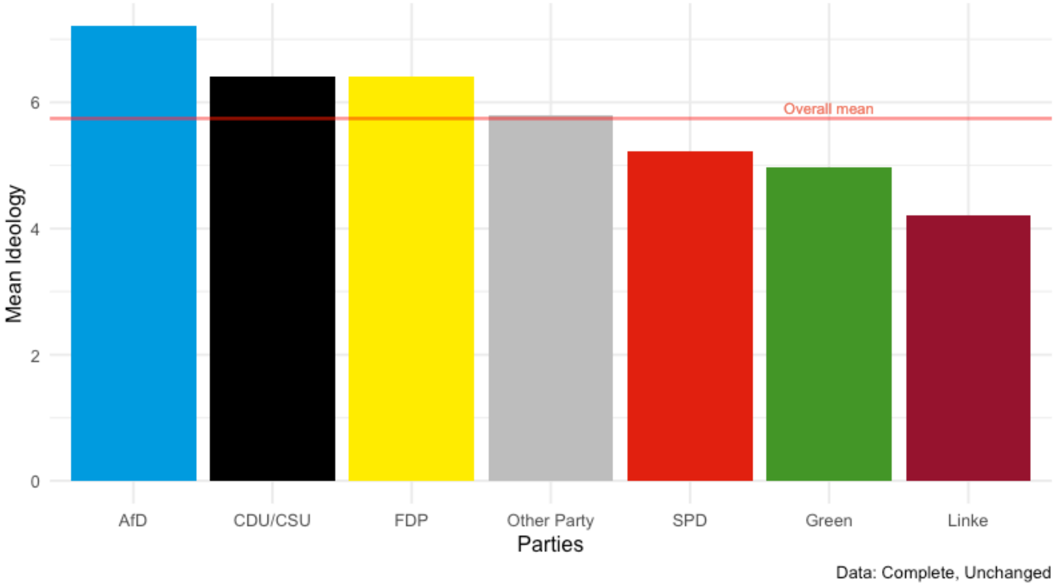


Figure 9: Mean Ideology by Party

We also explored the effects of some of those variables that we suspected to leave traces in our browsing history data in addition to also affecting party choice. We hope that these will allow the algorithms we employ to pick up on political preferences by enriching data points that do not constitute direct expressions of political preference. The plot below visualises the correlations of age to time spent on some exemplary websites such as YouTube, the website of Deutsche Bahn and the German broadsheet Süddeutsche Zeitung. While age is negatively correlated with the first two examples, it is positively correlated latter.

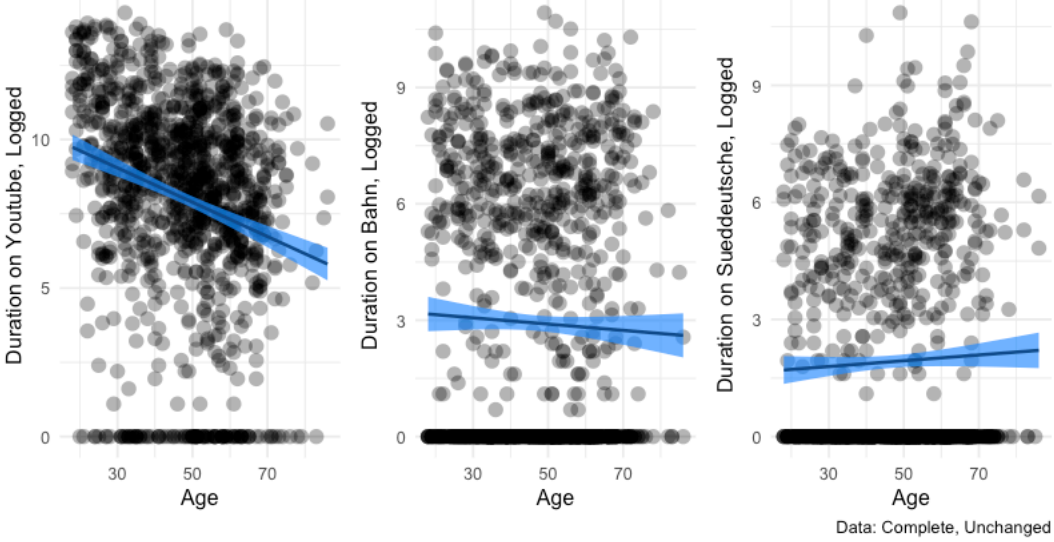


Figure 10: Scatterplots Age on Duration Online

The user profile of domains also differed with respect to attributes such as gender, high education or ideology. To illustrate our observations, we provide comparative bar charts for some exemplary domains below. The first chart reaffirms Bujala's finding that men tended to seek out entertainment, while women were more likely to be seeking information on health. Two-thirds of the users of the domain "Gesundheits-Fakten" (health facts) are female, while only 20.8 % of the users of the gaming site "Gameforge" were female.

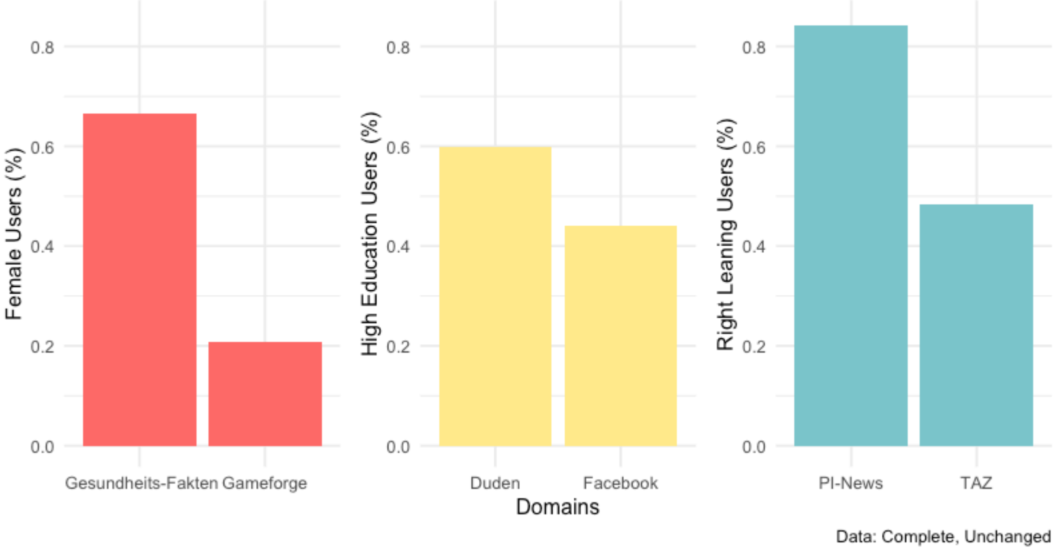


Figure 11: Variance in Socio-Demographic Variables by Domain

Among the users of "Duden" - the German equivalent of the Oxford Dictionary – 59.9 % had were highly educated (had the qualification to enter college), while only 43.9 % of Facebook users held the same qualification. This is in line with the findings of Zillien & Hargittai that high status individuals tend to use the internet more for self-enhancement and less for activities such as chatting (Zillien & Hargittai, 2009). We also find confirmatory examples for the hypotheses related to news consumption, ideology and confirmation bias discussed above (see Chapter 2.4). For this, we looked into the share of right leaning users of a domain, conceptualised as users who have a self-placement score higher than 5.5 on the one to eleven scale. The share of right leaning users of PI-News, a right-wing online outlet, is 84.2 %. The users of TAZ on the other hand, a newspaper with a left-wing reputation, is 48.4 %.

4 Methodology

The goal of this research project is to establish whether browser history can be used to predict individual political preferences. In order to do so, we use machine learning techniques suited to the high dimensionality of our data (see Section 4.3). We train these models on different versions of our datasets (see section 4.1). This gives us more than one measurement of accuracy per model and allows us to compare the strength of the signal in different partitions of our data. To assess our models, we compare their accuracies with each other and - more importantly - evaluate their performance against baseline thresholds (see section 4.2), including ones derived from models trained on socio-demographic data of the individuals in our dataset.

4.1 Datasets and Versions

In order to assess the performance of our models, we decided to apply them repeatedly to our data and not rely on merely one application as a measure of model effectiveness. We use three datasets in addition to the complete dataset described above (see chapter 2). This was motivated by the high dimensionality of our complete dataset – there are far more features (p) in our cleaned dataset (some 8,000 domains) than observations n (1,162 individuals). Where p >> n, certain statistical techniques can become less accurate due to noise (James, Witten, Hastie, & Tibshirani, 2013). We therefore pursued two commonly recommended strategies (James et al., 2013):

- select a subset of features which might be particularly related to the variable of interest;
- use regularisation to reduce the amount of noise produced by negligible features;

We employed regularisation techniques on two of the models that will be discussed in greater detail below (see section 4.3). Second, we selected subsets of features for our three additional datasets in a theoretically driven way.

We created one dataset that contains only the 500 most popular domains by duration (Top500). Our assumption behind this decision was that focusing on the top 500 websites would allow us to pick up on enough variation between the different classes, while at the same time eliminating mere idiosyncratic behaviour. We also created one dataset containing only domains related to news. We made this decision based on the theoretical and empirical insights from selective exposure theory and biased news media (see Section 2.4). Third, we constructed a dataset that only contains data from September 2017, the month of the federal election. Our motivation behind this step was to see whether online behaviour in close temporal connection to the election would reduce noise and improve accuracy. Indeed, Yasseri and Bright (2016) hypothesise that there are differences in how and why people seek information online at election time. We will test our models on these three datasets, in addition to the full dataset containing all features.

In sum, this leaves us with:

- The full dataset, as detailed in section 3 (Complete);
- A dataset containing only the top 500 most visited domains by duration (Top500);
- A dataset containing only news domains (News);
- A dataset containing only domains visited in September 2017, the month of the election (September).

Finally, as discussed in Chapter 3, we will investigate whether accuracy increases for each of the four datasets outlined above under the following conditions:

- The dataset as it is (Unchanged)

- A logged version (Logged)
- A version removing those individuals who were not online for at least one hour per week for the duration of the study (Minus Rare Users)
- A logged version, with rare users removed (Logged Minus Rare Users)

In total, this leaves us with sixteen iterations of our data that we apply our models to.

It is important to reiterate that the Top500, News and September datasets contain three different variables per domain: one capturing the duration spent there, one the number of visits to a domain and a Boolean indicator, showing whether or not the domain was accessed at all. However, the Complete dataset only contains one variable per website - on the duration spent there. This dataset would have included all domains in all three measurement metrics, totalling 24,101 variables, however using a dataset of this size proved computationally impossible for the scope of this project.

4.2 Model Evaluation

4.2.1 Assessing model accuracy

Before running any models, it was necessary to define how we would measure the success or failure of a model. In classification tasks, the most commonly used metrics are overall accuracy (the percentage of correct predictions across the entire model). However, this can be misleading in cases where there is a strong class imbalance: if one class dominates, a model will achieve high accuracy by simply predicting this class for all observations but provide little informational value. The classes in our data are distributed as follows:

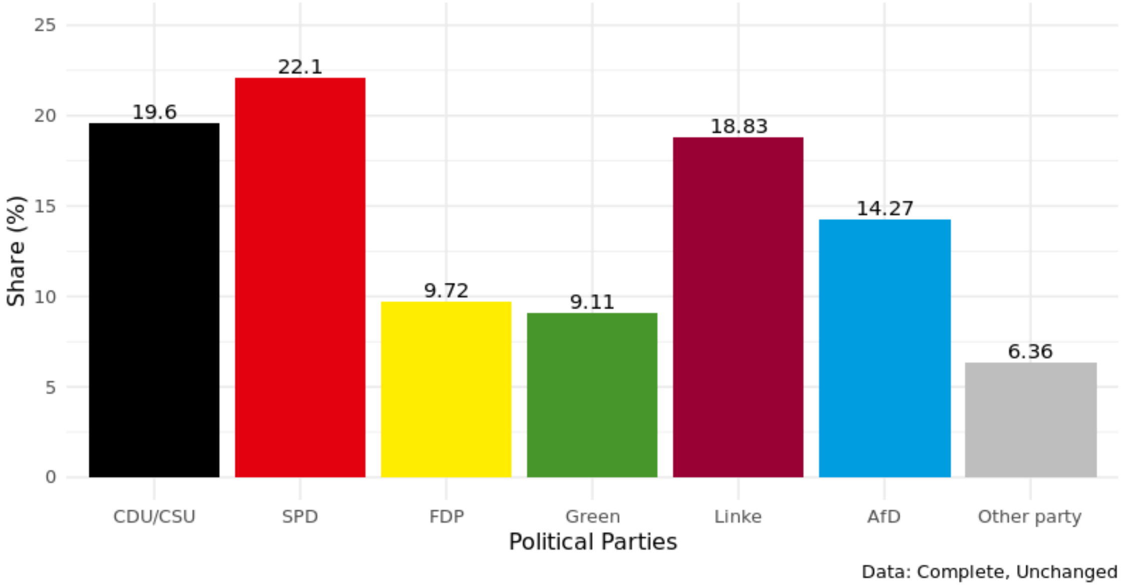


Figure 12: Party Share in Complete Dataset

As can be seen above, we have a mild class imbalance. There is no one party that dominates the data to such an extent that predicting this party for all observations would produce a high accuracy. For this reason, we will rely on accuracy to compare the performance of the different models. However, as CDU, SPD and Linke are significantly larger than other parties in our data, we will also report precision, recall and F1 statistics for the classes within the model. This will give us a more detailed view of the usefulness of a particular model.

Precision is the fraction of correct predictions of a certain class out of all predictions for that class (the ability of the model to identify only the relevant data points). Recall is the fraction of true observations in a certain class that were correctly predicted (the ability of the model to identify all the relevant data points). There is an inherent trade-off between the two, particularly with unevenly distributed classes, and it is not possible to maximise one without reducing the other. The F1 statistic deals with this trade-off by providing a harmonic mean between precision and recall, punishing extreme values of either and creating an optimal balance.

4.2.2 Comparing to a baseline model

Metaxas et al. (2011) point out that many studies present accuracy statistics of models - for example, percentage of correctly guessed electoral races - without qualifying this in comparison to traditional methods of prediction, such as polling, or even more trivial measures, such as incumbency. Many studies are guilty of comparing predictions simply to vote rates, including Bermingham & Smeaton (2011), Ceron et al. (2014) and Tumasjan et al. (2010), without comparing to a suitable baseline model (Gayo-Avello, 2013).

Cranmer & Desmarais (2017) define three features of a sensible baseline predictive model: it must be one which can be specified without the proposed theory; make predictions purely out-of-sample and that it take into account the distributional features of the predicted variable. In this study, the baseline model must rely on variables other than domains in the prediction, must predict based on a test set separate to that used for training purposes, and take into account the relative distribution of the outcome variable, ‘second vote’. While Metaxas et al. (2011) and Gayo-Avello (2013) look solely at binary predictions, namely US elections, and thus set a high bar for accuracy since “the incumbent candidate gets re-elected 9 out of 10 times” (Metaxas et al., 2011), we will be predicting seven classes and thus expect much lower accuracy ratings as the task of differentiating these is more difficult. There are multiple options for baseline models and a successful model should at the very least be better than chance. Our models will be compared against the following baselines.

Majority-class metric

When one class dominates a dataset, predicting this class for all cases may produce a high overall accuracy. In this dataset, the most common party is 2 (SPD) and predicting this for all classes would produce an accuracy of 22%.

Predictive classifier: Socio-demographic classifier

A number of studies attempting to predict vote choice using socially produced internet data have compared their models to baseline models comprised of socio-demographic attributes of individuals, for example in studies of the effect of personality on voting outcomes (see Giebler & Regel, 2018) and how online data predicts voting outcomes (see Kristensen et al., 2017). Such attributes include age and gender, but also more fine-grained data on issue attitudes and party preferences[1].

[1] See Appendix 1

Many of the socio-demographic factors included in our socio-demographic baseline model can or could be deduced from online history (see Do Viet Phuong & Tu Minh Phuong, 2014; M. Kosinski, Stillwell, & Graepel, 2013, PrakashSwami, Bhalchandra Tarte, Kisan Rakshe, Maroti Raut, & Faiz Shaikh, 2015; Youyou et al., 2015). For example, age, gender, political leaning, relationship status and many more can be found on Facebook, or predicted from the content of websites visited. Even more sensitive information can be gleaned from domains visited, such as if a woman is pregnant, looking for a new job, or approximate wealth levels. These factors in turn have an effect on which party a person votes for.

We compute four multinomial logit models, three based on different socio-demographic or political indicators, and the final model including all variables from the first three models. The models represent different types of information that might be available. The first model is based on socio-demographic variables, the second model on ideology, the third examines issue-related voting, and the fourth combines all the variables from the previous models. The results and specific variables are detailed in the Table 3. Further details of the variables and regression tables can be found in Appendix B.

Table 3: Performance of Predictive Baseline Models

Focus of Model	Variables	Accuracy
Socio-demographic	Gender, education, marital status, employment status, household income	21%
Ideology	Self-reported left-right scale, rating of the government, rating of the following parties: CDU, CSU, SPD, Green, FDP, Linke, AfD	42%
Issue attitudes	Attitudes towards the following issues: migrants, coal, migrants' culture, taxes, EU cooperation, tax reduction, terrorism, Brexit, fake news, international crises, investment, limit on refugees	29%
Complete model	All of the above	38%

As can be seen from the above accuracy levels, even with comprehensive information about an individual's opinion of the political parties, the model is still only correct in 42%

of cases. Furthermore, the models indicate that socio-demographic data alone is not sufficiently predictive, performing worse than the majority-class metric. This corresponds to findings that support for the AfD, for example, does not come from socially marginalised groups (i.e. those with low income, the unemployed etc.) but those who experience feelings of 'cultural threat' from immigration (Baron, 2018).

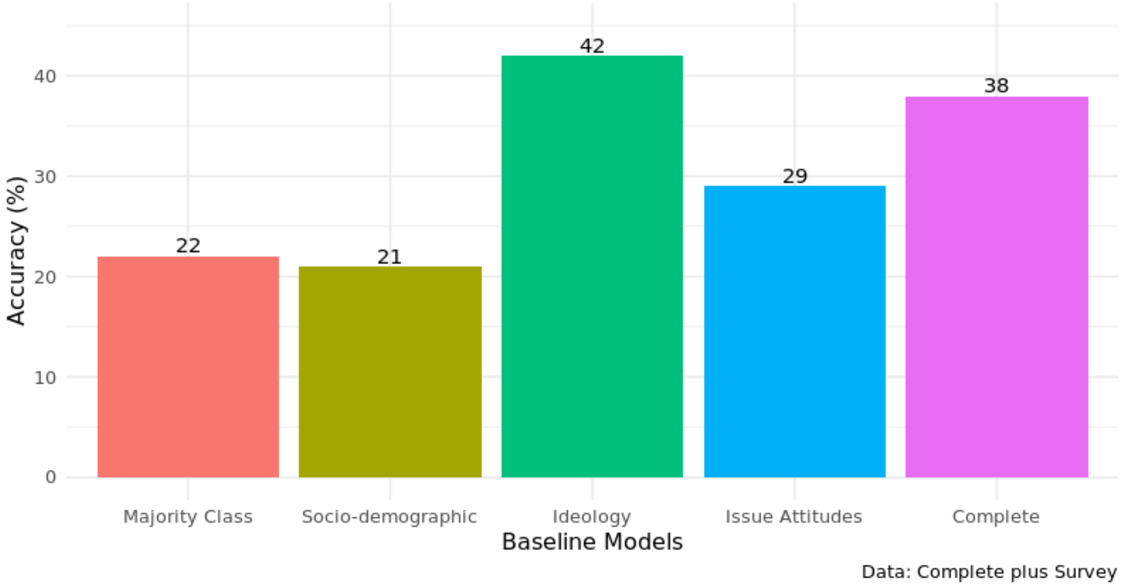


Figure 13: Accuracy of Baseline Models

The ideology model represents an extremely high bar, and such detailed information about party preferences is rarely available to companies. Indeed, it represents they very information we are hoping to predict. Issue attitudes therefore represents a more reasonable baseline model, as such information could likely be inferred based on the kind of websites that individuals visit. We will evaluate our models based on both the issue attitudes baseline, and the majority class metric in order to have a lower and a higher bar of accuracy for comparison purposes.

Finally, these results caution against expecting very high accuracy levels for our models. Even a model based on variables rating political parties is only correct 42% of the time.

4.3 Machine Learning Models

Since we are interested in the performance of contemporary machine learning techniques on these datasets, we focussed in particular on two classes of algorithms that are known for their outstanding performance in dealing with high-dimensional data: Random Forests and Support Vector Machines (Caruana, Karampatziakis, & Yessenalina, 2008). In addition, we wanted to assess how well a regularised multinomial logit classifier would compare to these more sophisticated machine learning techniques and had also hoped that it would provide us with further leads regarding highly predictive features in our data. However, we started out with an even more simplistic classification approach in order to see whether even it would be capable of producing tangible results: a Naïve Bayes classifier. Before we present our findings, we will provide brief summaries on the four models selected.

4.3.1 Naive Bayes

Bayesian Classifiers are types of probabilistic classifiers - that is, they calculate the likelihood of an observation being in a certain class given a certain set of parameters. Naive Bayes Classifiers are so called because they rely on the strong (naive) assumption that, given a certain class, the variables in the model are conditionally independent of (i.e. unrelated to) each other. In reality, this is rarely the case. However, even if this assumption is not completely met, Naive Bayes often outperforms more complex methods. The variables in our data will likely be highly correlated with one another; very frequently visited domains, such as Google and Facebook, will be correlated with one another, as will domains relating to a specific topic, such as news or shopping. Despite expectations that this model would not perform well on our data, we were interested to see what levels of accuracy a rudimentary model such as this would produce for comparison purposes.

4.3.2 Regularization and Multinomial Logit Classification

Regularization techniques, such as the widely used LASSO (L1) or ridge (L2) methods, aim to reduce the variance of a regression of classification (Kuhn & Johnson, 2013). This prevents an algorithm from overfitting on the training data, i.e. becoming very precise in predicting the data it is trained on, but increasingly bad at predicting unseen data. In the variance bias trade-off of a predictor, regularization therefore reduces variance at the risk of introducing bias. This trade-off however has been shown to be particularly beneficial in a setting with a high number of predictors and a relatively small number of observations (Urda et al., 2017). Both approaches, L1 and L2 regularization introduce a penalty term that shrinks the relative importance of features that are not highly predictive (Kuhn & Johnson, 2013). The difference between the two methods is that due to its mathematical formulation L1 regularization can lead to the exclusion of unpredictive features in the final model, while L2 regularization only reduces the relative importance of less predictive features.

In this project we decided to employ L1 regularization on a multinomial logit classifier in order to identify highly predictive features early on before applying other machine learning methods to the dataset. The multinomial logit estimates the probability that an observation belongs to a specific class based on a maximum likelihood function. It produces coefficients that indicate the increase of the log-odds (or logit) of an observation to belong to class a if its value for feature x_i increases by one unit (James et al., 2013). Since we are employing L1 regularization in combination with the multinomial logit, we tuned the penalty parameter λ using repeated ten-fold cross validation.

In addition to the L1 regularization applied to the multinomial, we used L2 regularization in combination with a linear kernel Support Vector machine, a type of machine learning method that will be discussed in greater detail below.

4.3.3 SVM

A Support Vector Machine or SVM (Cortes & Vapnik, 1995) is a supervised machine learning technique that can be used for both regression and classification. We decided to investigate the performance of SVMs on the domain level data because it is particularly suited to high-dimensional cases (Caruana et al., 2008).

In a classification application – as done in this paper, by assigning party choice to individuals based on their browser history – SVMs rely on identifying the linear maximum margin between the classes of interest. This separation is drawn as a (p-1)-dimensional hyperplane in the p-dimensional space generated by the p features contained in the dataset for classification (Hastie, Tibshirani, & Friedman, 2017). To determine the location of the separating hyperplane or decision boundary that creates the maximum margin between the classes, only the observations closest to the decision boundary are decisive. These observations are called the support vectors. The SVM may therefore be particularly prone to changes in performance when observations are removed from the training data, as is the case in our datasets that only contain individuals that spent more than one hour per week online, as opposed to the full population.

Since it is often impossible to draw a perfect linear boundary between data points to separate them, SVMs allow for a slack variable that tolerates some misclassification during the training of the algorithm (Hastie et al., 2017). In our application, we determine the optimal values of the slack variables of the different SVMs used through repeated cross validation.

In addition to the slack that SVMs allow for – their soft margins – SVMs are furthermore capable of employing the so-called "kernel trick" to classify data that is not separable linearly. A kernel is a mathematical transformation that simulates a higher dimensional space ($q > p$ dimensions) than our p features would allow for (Hofmann, Schölkopf, & Smola, 2008). It is often possible to draw a linear boundary in the simulated higher dimensional space based on the transformed data. When the separation has taken place and the data is "transformed back" the linear separation in the higher dimensional space translates to non-linear boundary in the original feature space. In this project we experimented with polynomial and radial kernels (creating a polynomial or radial separation in the p-dimensional space) and tuned their parameters such as the order of the polynomial using repeated cross validation.

In total we worked with four different types of SVM. Two SVMs with a linear kernel, one of them employing an L2 regularization method to prevent overfitting described above, one SVM with a polynomial and one with a radial kernel. Since SVMs are computationally intensive we ran parallelized versions of all SVMs and could only implement the computationally least taxing linear SVM to all versions of our data.

The metric used to train the models is Kappa. This has two reasons. First, there is a mild imbalance in our dataset with respect to CDU (19.60 %), SPD (22.10 %) and Die Linke (18.83 %). These three parties are overrepresented as compared to FDP (9.72 %), B90 (9.11 %), minorly AfD (14.27 %) and particularly Other Parties (6.36 %). While accuracy could be high by chance in tuning, just because some classes are overrepresented compared to others, Kappa takes agreement by chance into consideration and therefore corrects for the mild class imbalance (Kuhn & Johnson, 2013). Second, experimenting with Kappa and accuracy as tuning parameters showed that models tuned with Kappa had an equal or better test set accuracy, even when training set accuracy was better for the model trained with accuracy as metric.

4.3.4 Random Forest

Classification trees (classifying decision trees) are a simple machine learning algorithm making use of binary or multivariate trees to classify data. Classification trees tend to enjoy low bias and work well on the training data, but are prone to overfitting and high variance, producing large errors on previously unseen test data (Hastie et al., 2017). Random Forests (Breiman, 2001) are an ensemble learning method that aims to reduce the variance inherent to classification trees. By maintaining their low bias, but reducing overfitting of the training data, random forests work more effectively with previously unseen data.

The variance of classification trees can be reduced by *bagging* (bootstrap aggregation), whereby we subset the dataset, apply the model and then average the output. This increases bias slightly, but allows for a large reduction in variance and avoids overfitting. However, with bagged classification trees, certain strong predictors will appear in every tree. This results in a large number of very similar trees which do little to reduce the variance when averaged (James et al., 2013). A random forest solves this issue by reducing the correlation between the trees generated from different bootstrapped samples from the training data. This is done by selecting a random number of variables before each split/node (Hastie et al., 2017). Averaging such decorrelated trees generally leads to a considerable reduction of variance.

Random forests also benefit from the fact that, on average, one third of observations do not appear in the bootstrapped dataset. This unused (and unseen) training data can be "recycled" to test the model. We nonetheless chose to split the data into a training and test set, mainly to ensure comparability between the random forest and other models on the same data points.

Parameter tuning

Random forests have one tuning parameter, namely the number of variables selected at random for splitting at each node (Mtry). Tuning this parameter can involve trying random values or defining a list of values for the model to try. The default value of Mtry for classification tasks is generally taken to be the square root of the number of predictors (p). We defined six trial values of Mtry for each dataset, always including the square root of p as well as five other values evenly distributed between zero and the number of predictors.

Random Forests calculate the most beneficial split based on a given metric, often the Gini impurity or reduction in entropy (information gain) (Hastie et al., 2017). The Gini impurity quantitatively evaluates how good a split is by assigning a probability between 0 and 1 to the chance of classifying a randomly chosen data point incorrectly. Given a number of classes C where the probability of choosing the i^{th} observation is *p(i),* the Gini impurity is calculated as follows:

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

The best split is chosen by maximizing the Gini Gain, which is calculated by subtracting the weighted impurities of the branches from the original impurity of the dataset.

"Extratrees" stands for 'extremely randomised trees' and is a further variant of the random forest algorithm (Geurts, Ernst, & Wehenkel, 2006). In such trees, the next split is the best split among random uniform splits in the selected variables for the current tree. Extratrees is often computationally faster and can be more accurate than random forests, although can perform worse if there is a high number of noisy features, as in high dimensional datasets (Geurts et al., 2006). For this reason, we will calculate both random forests using the Gini split and an ‘Extratrees’ model, and use the variant most accurate on our data.

Finally, it is also possible to specify the number of trees (ntree) in the forest. Once all parameters are fixed, the model’s loss (the summation of errors made by the model) stochastically decreases as the number of trees increases, however with diminishing marginal returns. Increasing the number of trees cannot overfit the model, therefore it is sensible to choose a sufficiently large value for ntree. Growing more trees requires more computational power and considering our resource limits we decided to keep ntree at the library default of 500. This will therefore give us a conservative estimate of accuracy that would increase if more trees are grown in subsequent studies.

5 Results and Discussion

5.1 Naive Bayes

As expected, the Naive Bayes classifier does not perform well on our data. This is due to high-dimensionality of the dataset and the correlation between many of the variables, which violates the strong independence assumption that Naive Bayes relies on. The model performed worst on the complete dataset and the September domains, but better on the top 500 domains and the news domains. The latter two datasets are the smallest ones, containing 1502 and 1652 features respectively. That the model performs better on datasets with lower dimensionality indicates that the model is overfitting on the larger datasets, reducing accuracy on unseen test data.

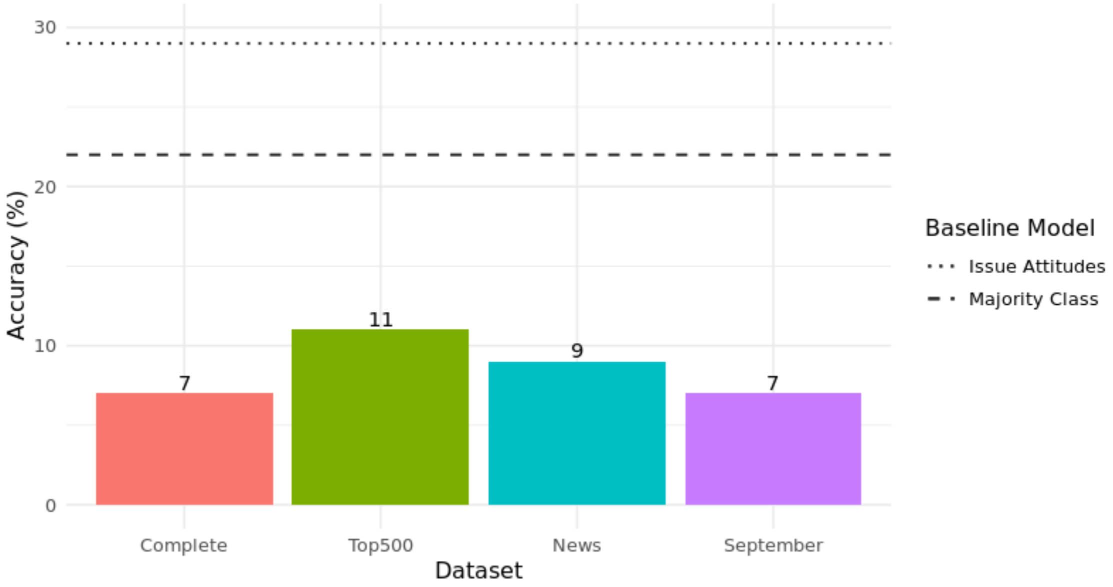


Figure 14: Accuracy of Naive Bayes by Dataset

None of the Naive Bayes models meets either of the baseline model thresholds. Even the majority class metric, i.e. predicting SPD for every individual, produces an accuracy of 22%, twice that of the most accurate Naive Bayes model. Since the model performs so poorly, we did not continue our analysis with the logged versions or removing rare users, as we expected this to only produce minor gains in accuracy which would not push the model over the threshold of the baseline models.

The most accurate Naïve Bayes model, using the Top500 domains, returns the following values for precision, recall and F1.

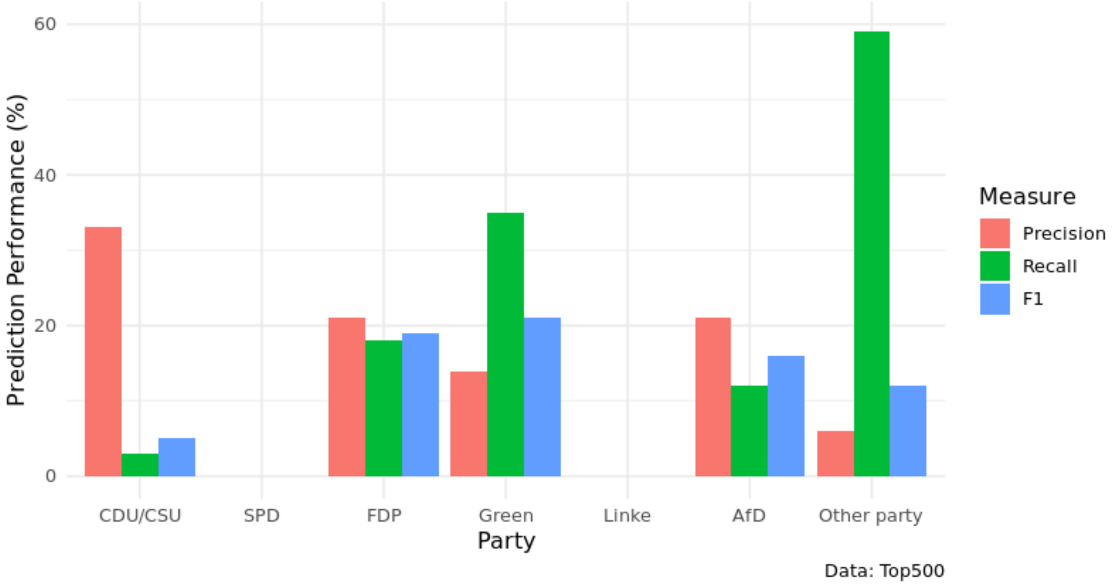


Figure 15: Precision, Recall, F1 by Party for Highest Accuracy Naive Bayes

The figure shows that the naïve Bayes model over-predicts 'other party', with a very low level of precision. The model also failed to correctly predict any votes for the SPD or Linke. This further calls into question the usefulness of a measure that fails to correctly predict the largest and third largest parties in our sample.

5.2 LASSO Multinomial Logit

The L1 regularised multinomial logit model achieved some of the highest accuracy in predicting party choice. It nonetheless failed to outperform the issue attitudes baseline, as can be seen in Figure 16 below.

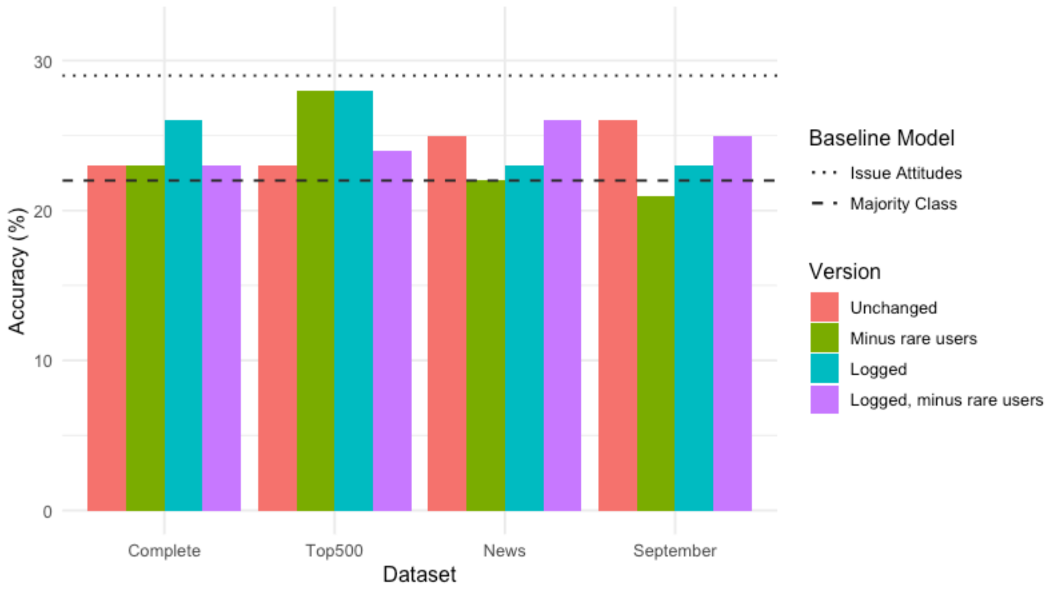


Figure 16: Accuracy of LASSO by Dataset and Version

It did however outperform a majority class guess for 14 out of 16 dataset versions. It achieved its highest accuracy of 28 % correctly predicted class labels on the Top500 dataset's logged version, and the version which we removed rare users from.

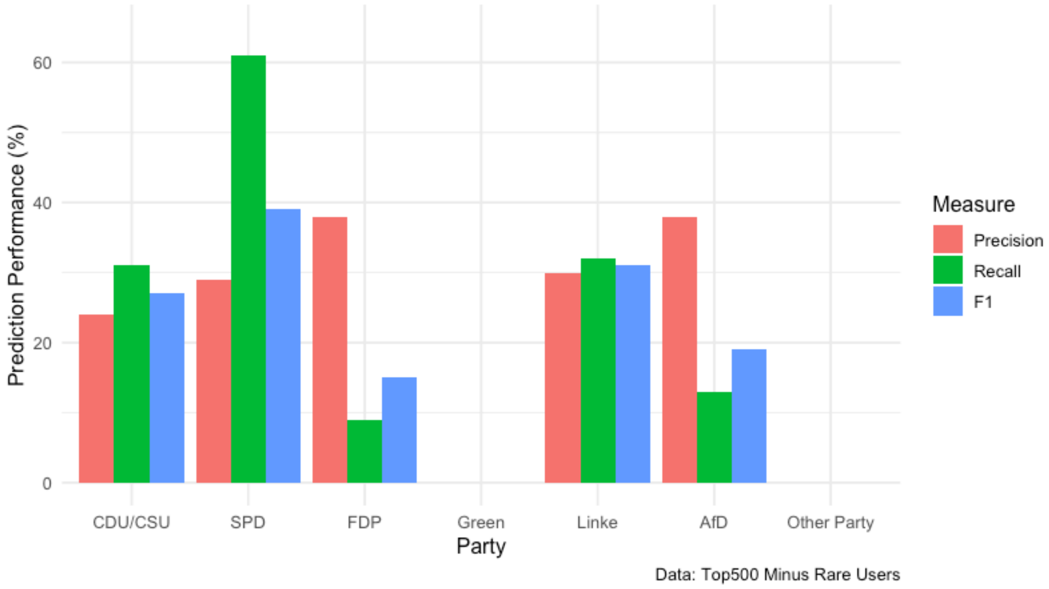
The Top500 subsequently yielded the best average accuracy of any dataset of 26 %. The differences between all datasets are minimal however and could be due to the performance of our models on the particular training and test set split we used. The same goes for the different versions of our data. None massively outperformed the others and each version is once the most accurate on the different datasets. There is simply neither a dataset nor a version of data that has consistently and substantively outperformed all others.

Table 4: Average Accuracy of LASSO by Dataset and Version

Dataset	*Mean Accuracy*	*Version*	*Mean Accuracy*
Complete	24	Unchanged	24
Top500	26	Minus Rare Users	24
News	24	Logged	25
September	24	Logged Minus Rare Users	25

To investigate the LASSO model's class specific precision, recall and F1 we looked at its best-performing iteration, the one on the Top500 dataset minus rare users. Since the model does not predict Green or Other Party at all, there are no outputs for these parties in the graph below. Precision is comparatively high for both the AfD and the FDP – both reach 38 % – but both recall and consequently F1 are the two lowest values achieved on these metrics. This indicates that both parties were underpredicted. This is surprising since it could be hypothesised that AfD voters as affiliates of a party that is considered to be one of the two ideological poles in the Bundestag give of a particularly strong signal. CDU/CSU and Linke show more balanced metrics.

Table 5: Precision, Recall, F1 LASSO

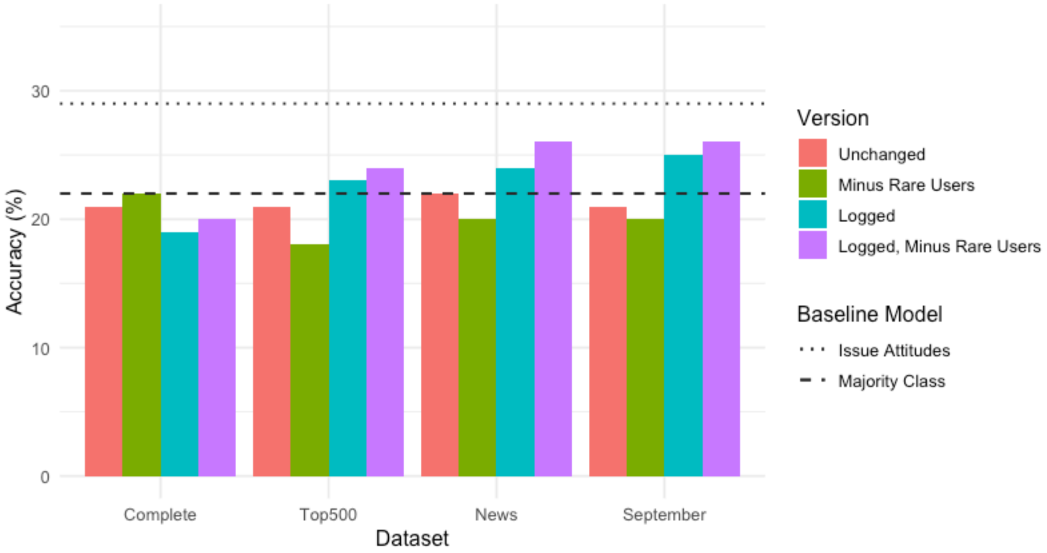


The big outlier is SPD – the largest class – with a recall of 61 %. It is the only class for which a majority of cases was correctly predicted. Since the precision is however relatively low at 29 %, the high recall is the effect of the model over predicting SPD, despite the fact that Kappa was used to train the model.

5.3 SVM

We chose to implement an SVM on our data because this type of algorithm is known to perform well on high-dimensional data. It is however also known to be computationally taxing and thus time intensive. Given an ideal computational set-up, we would have computed four different types of SVMs on all versions of our data: a linear SVM, a radial SVM, an L2 regularised linear SVM and a polynomial SVM. However, due to a bottleneck in processing capacity, even when parallelising the SVMs on six processor cores, the net computing time for the linear SVMs alone amounted to several days. Thus, we decided to compute the linear kernel across all dataset versions and run the computationally more taxing models on the unchanged Top500 dataset. This provides us with an indication of whether these kernels should be a focus of future research.

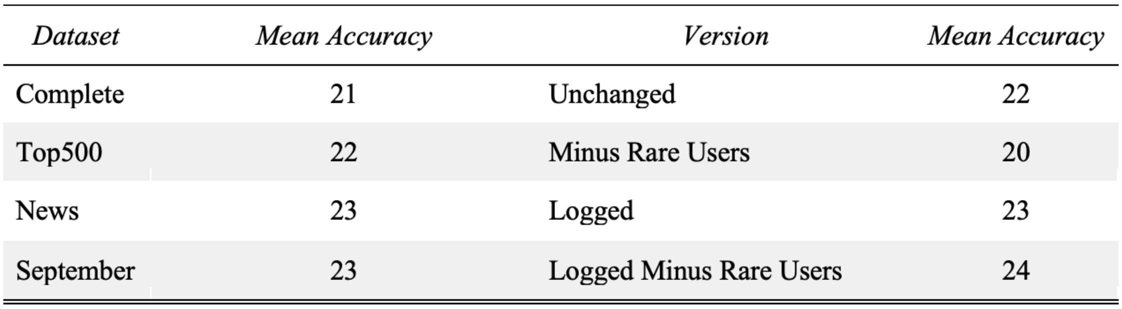
Table 6: Accuracy SVM by Dataset and Version



The linear SVM outperformed the majority class guess baseline on six of the sixteen different versions of our data. It did not however reach higher accuracies than the issue attitudes baseline model. It performed best on those versions of the logged and logged minus rare user versions of the News and September datasets. These versions of the data allowed the SVM to outperform the majority class guess baseline on three out of our four datasets, suggesting that the reduction in variation achieved through logging the data and removing infrequent users was particularly beneficial for the SVM. Only on the complete dataset did none of the versions produce a higher accuracy than the majority-class baseline. This is the case despite the fact that both Complete and September contain a large number of variables (p>8000). This suggests that the data retrieved in the last month around the election indeed contains a stronger signal with respect to party choice – at least for the linear SVM.

Looking at averaged accuracy scores by dataset and version in Table 7, the linear SVM reached an average accuracy of 23 % on the September and News datasets, but only a 21 % average on Complete. The average values of accuracy also show that the logged version of our data outperform the non-logged versions. There is no consistent effect of removing rare users on accuracy. While the logged minus rare user datasets have the highest average accuracy of 24 % across versions, the minus rare user datasets have the lowest with 20 %.

Table 7: Average Accuracy SVM by Dataset and Version



Dataset	*Mean Accuracy*	*Version*	*Mean Accuracy*
Complete	21	Unchanged	22
Top500	22	Minus Rare Users	20
News	23	Logged	23
September	23	Logged Minus Rare Users	24

To investigate the linear SVM model's class specific precision, recall and F1 we again looked at its most predictive iteration – here the September logged minus rare users. Since the model does not predict FDP, Green or Other Party, there are no corresponding outputs for these parties in the graph below. Precision is comparatively high for both the AfD and the Linke. The recall and F1 for AfD are however the lowest values achieved on these metrics. This indicates that the model underpredicted AfD. The Linke and SPD show more balanced metrics and are the models with the highest F1 score, with the SPD predictions again being most accurate.

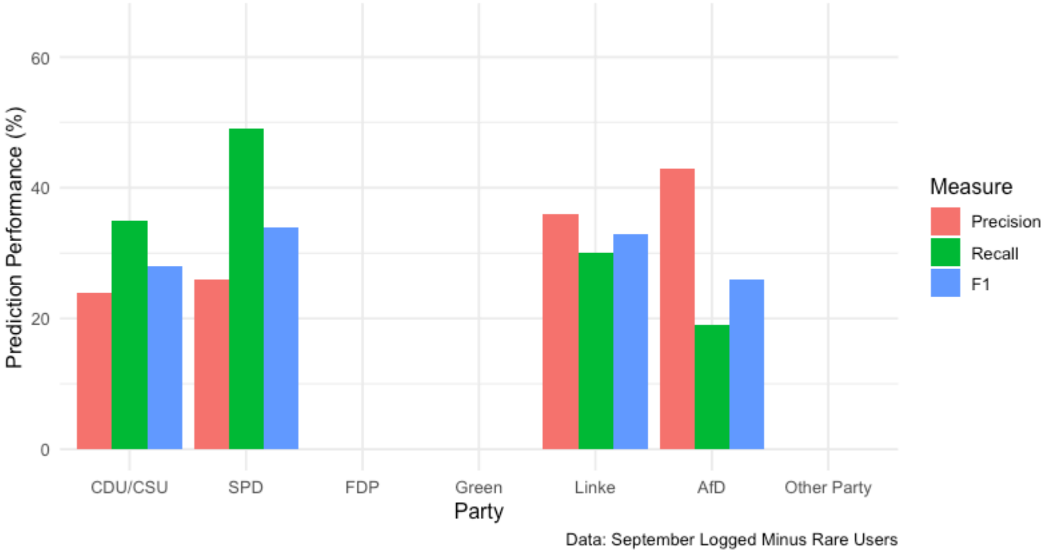


Figure 17: Precision, Recall, F1 SVM

Two of the three more complex iterations of our SVM performed outstandingly well on the unchanged Top500 dataset. The L2 regularized linear SVM achieved an accuracy of 39 % and the polynomial SVM even 77 %, outperforming all baseline models comfortably.

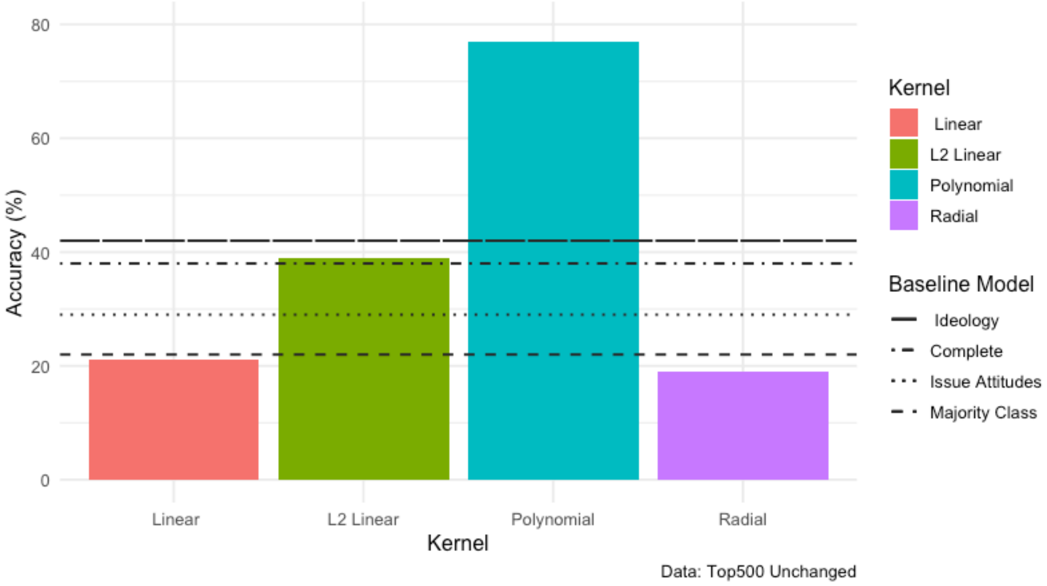


Figure 18: Accuracy SVM by Kernel

However, caution is advised. While these results strongly suggest that there is a signal in our data that allows for the prediction of political affiliation, our computational limitations did not allow us to run these models several times and on different versions of our data. The better performance is likely a result of the reduction of noise in the case of the regularized SVM. The polynomial SVM on the other hand is able to parse the data in a more flexible way than the linear kernel and might therefore have been able detect relevant relationships between our features not accessible to the linear SVM.

5.4 Random Forest

The random forest models were tuned to find the level of optimum value of Mtry (the number of variables considered at each node of the tree) which maximised accuracy. Accuracy levels were calculated for regular random forests using the Gini Index to split the trees and for 'Extratrees' (as explained in Section 4.3.4), with the one that maximised accuracy being chosen. The model fit plot in Figure 19 shows the changes in accuracy for models using both Gini and 'Extratrees' according to the number of variables taken into account at each split (Mtry) for the top 500 domains.

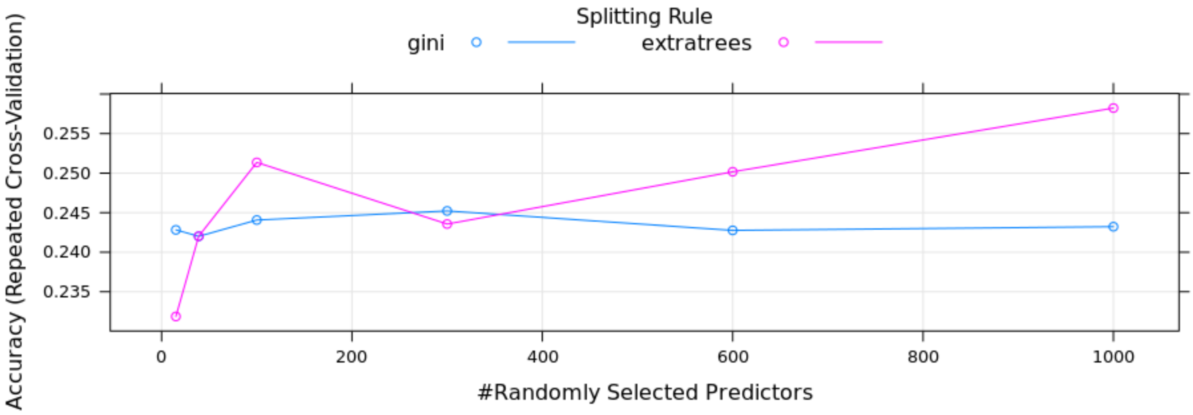


Figure 19: Tuning the Paramter Mtry by Gini and Extratrees

We see that there is an optimum level of Mtry which maximises accuracy. Model fit plots for all dataset versions can be found in Appendix D. Neither 'Gini' nor 'Extratrees' was consistently more accurate; the results depend heavily on the version of the dataset. However, the difference in accuracy of different values of Mtry between a regular Random Forest or an 'Extratrees' Random Forest is minimal, generally ranging only 2 to 3 percent difference.

The accuracy of all random forest models are summarised in Figure 20 below:

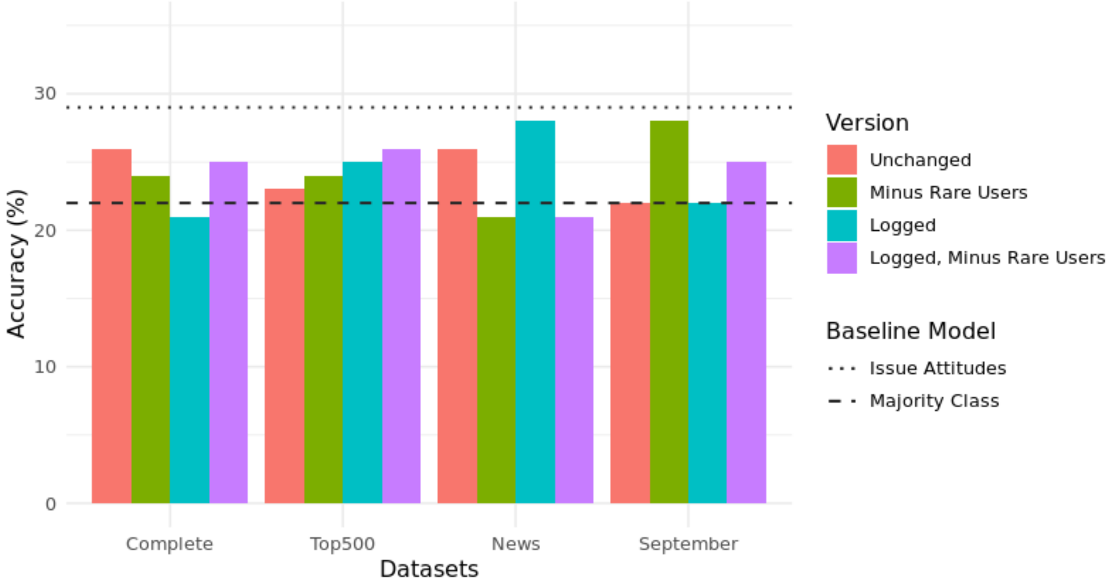


Figure 20: Accuracy of Random Forest by Dataset and Version

12 out of the 16 models were at least as or more accurate than predicting the majority class for all observations, and the 4 remaining models were only 1 percentage point less accurate. While all models were more accurate than the socio-demographic classifier, no model was more accurate than using issue attitudes. This indicates that domain level data does not provide as detailed a picture of an individual's opinion towards certain political issues as the survey did.

The highest accuracy was achieved using the news domains, logged, and September domains excluding rare users. This may indicate that that accuracy improves when focussing on the month of the election, or on inherently more political content like news. These models achieve an accuracy closest to the issue attitudes classifier. However, the other dataset versions in the September and news groups produced mixed results, and indeed some versions failed to reach the majority-class baseline. This suggests that the positive results may have been produced by chance. Furthermore, there is no pattern in the results indicating that removing rare users, logging the data, or a combination of both has any consistent effect on the accuracy of the model. This is further confirmed when averaging the accuracy scores by dataset and by version, producing the results in Table 8. Here we see no distinguishable difference in datasets or versions.

Table 8: Averaged Accuracy of Random Forest Models by Version and Dataset

Accuracy by Version		**Accuracy by Dataset**	
Version	*Average*	*Dataset*	*Average*
Unchanged	24	Complete	24
Minus Rare Users	24	Top500	25
Logged	24	News	24
Logged, Minus Rare Users	24	September	24

Variable Importance

A variable importance plot (see **Error! Reference source not found.**) was produced using one of the models with the highest accuracy (September, Minus Rare Users). Variable importance plots indicate how effective a feature is at reducing uncertainty (Parr, Turgutlu, Csiszar, & Howard, 2018). The importance measure is a sum of weighted impurity decreases for all nodes at which a predictor is used, averaged across all trees in the forest (Charpentier, 2015), and is scaled between 0 and 100, with the most important variable obtaining a score of 100, and thus allowing us to compare relative variable importance.

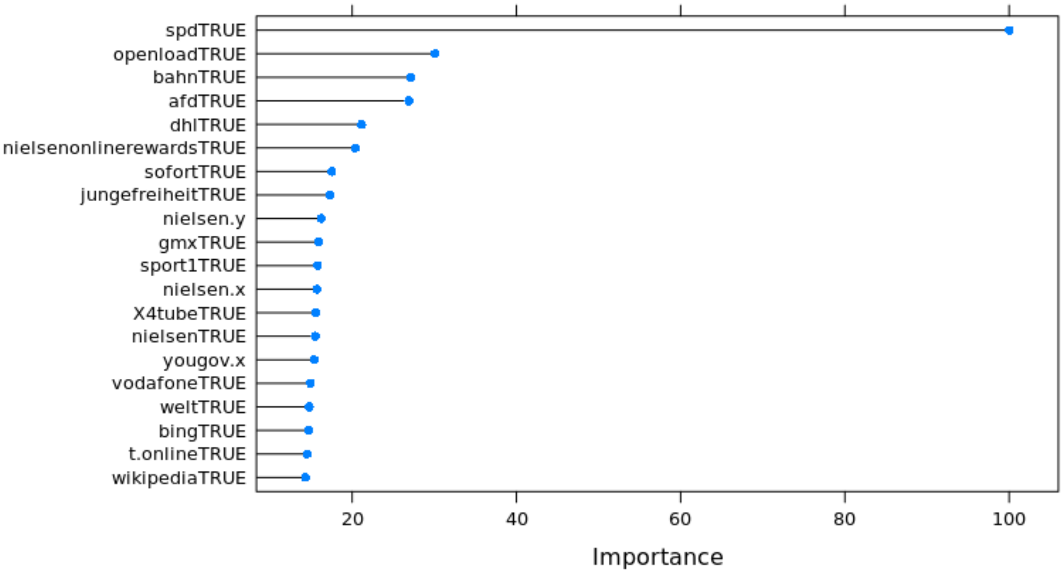
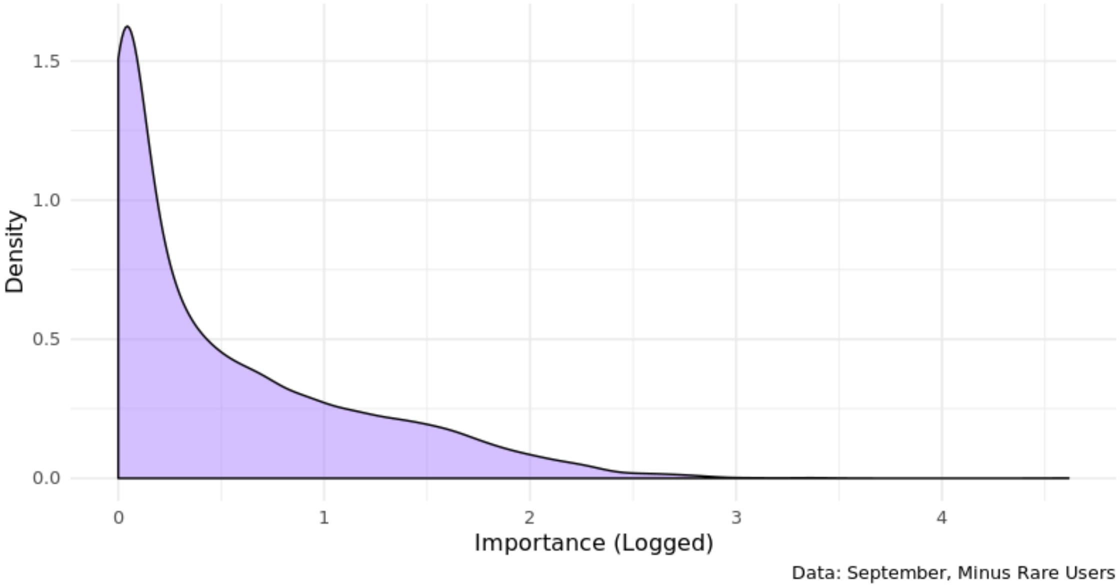


Figure 21: Variable Importance Plot, September, Minus Rare Users.

Note: TRUE indicates Boolean variable, x indicates duration and y visits.

In Figure 21, we see that the most important variable was the website of the SPD. Comparing this figure to the variable importance plot of the Top500 domains (see Appendix E) which covers the whole survey period, we see that the most important domains in the month of September included many more political and news websites, such as the websites of the SPD and the AfD, potentially reflecting that people inform themselves about politics online near election time. Other politically relevant websites, such as Junge Freiheit, a conservative and borderline right-wing news website, also appear among the most important.

However, we should not overemphasise the presence of political websites. As this conforms to our expectations of which domains will be more predictive of vote choice in the time leading to the election, we risk a form of confirmation bias. Indeed, there are also many non-political websites among the most important in this plot: 'bahn', the German national rail website, is more important than the AfD website, and the logistics service 'DHL' more important than 'Junge Freiheit'. If we look at the distribution of importance for all variables in September, Minus Rare Users, we see that the vast majority of domains had a near zero importance score (see Figure 22). Out of 8219 domains, 8216 had an importance score of less than one quarter of that of the most important variable, SPD. 8157 were not one tenth as important. This shows that the vast majority of variables are not important in determining the fit of the model.



Data: September, Minus Rare Users

Figure 22: Density Plot of Variable Importance (Logged)

Such plots can help us evaluate the reliability of the model by allowing us to assess the weight given to individual variables (Túlio Ribeiro, Singh, & Guestrin, 2016). For example, the different random forest models produce very different importance plots (see Appendix E). Only few domains, such as *Politically Incorrect*, a right-wing news blog, appear in multiple models with high importance. While this indicates that this website is particularly important in identifying party affiliation, the fact that the plots vary so widely indicates that the results, and therefore also the accuracy of the models, depend strongly on which variables were randomly selected at each node. The plots also include variables for which we have no theoretical reason to believe are predictive of party choice (e.g. DHL). It can be that such theoretically irrelevant variables are highly correlated with the response variable in both the training and the test set by chance (Kaufman, Rosset, & Perlich, 2011). Such correlations are impossible to identify by looking at accuracy metrics alone, demonstrating the importance of examining variable importance plots (Túlio Ribeiro et al., 2016). The prominence of seemingly irrelevant variables in the importance plots casts doubt on how well our model would apply to new data and if the same patterns would persist.

5.5 Binary response variable

Considering the weak performance of our models in comparison to the majority-class metric and issue attitudes classifiers, we decided to explore whether this resulted from the inherent difficulty of accurate multi-class classification by testing our models on a binary response variable.

In doing so, we created a binary variable, named 'left', coded 1 for left-wing parties, and 0 for right-wing parties. We coded CDU, FDP and AfD as right-wing, and SPD, Green and Linke as left-wing, removing 'Other Party' as it could not be determined where to place these observations.

We will evaluate the binary models based on the majority-class metric (53% for left wing). Any model must be at least this accurate to be better than guessing the majority class for all observations.

The results of binomial logit models on all versions of the four datasets are presented in Figure 23.

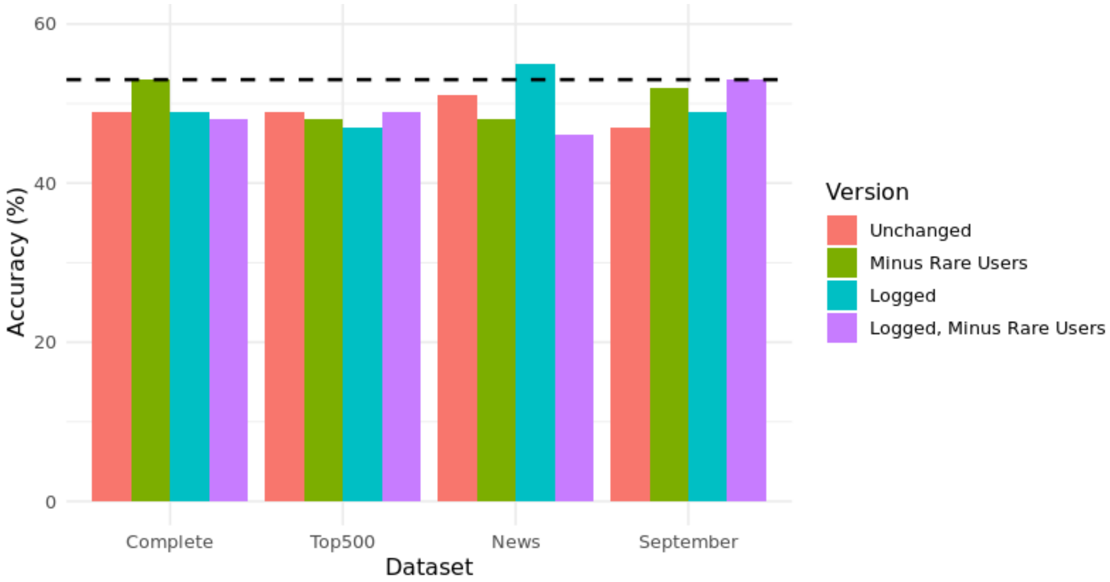


Figure 23: Accuracy of Binomial Logit by Dataset and Version

Only the news domains (logged) achieved an accuracy higher than the 53% baseline, with 55%. However, the results varied widely depending on which version of the dataset was used. While news domains (logged) performed best overall, the news domains also had the poorest performer, namely news domains (logged, minus rare users), with an accuracy of 46%. The variability within each dataset rules out drawing general conclusions about which dataset is most predictive or most accurate.

In addition to the binomial logit, we trained a random forest model on the binary outcome variable, using the news domains (logged) dataset, as this performed best in the multiclass random forest models and in the binomial logit. Since the difference in accuracy between the multinomial random forest models is limited and inconsistent across versions, we tested on one dataset only to explore whether a random forest on a binomial variable would in general be an effective approach. Further research considering a binomial response variable would ideally test all variations of all datasets.

The random forest returned an accuracy of 59%. We also see that the lower bound of the 95% confidence interval is at 53%, thus the accuracy of the random forest prediction is greater than that of the majority-class metric to a statistically significant level. Table 8 shows the precision, recall and F1 score for this random forest model. Despite an accuracy of 59%, the model only identified 41% of those who voted left wing. Of those it did classify as left, 58% were correct, producing a balanced accuracy of only 48%.

Table 9: Random Forest: Precision, Recall, F1 for Binary Response Variable

	Precision	*Recall*	*F1*
News, logged	0.58	0.41	0.48

Despite simplifying the classification task from seven classes to two, the model still fails to achieve high levels of accuracy. This indicates that the difficulty in achieving high accuracy in the multinomial random forest is not purely due to the general difficulties of multi-class classification, as a binomial variant of the model was only 6 percentage points more accurate than random guessing and classified leaving 41% of individuals incorrectly. Rather, this result further implies that domain-level data is not sufficiently predictive of political preference.

5.6 Cross-Model Comparisons

The goal of our project was to assess whether domain-level data could be used to predict individual party choice through machine learning methods. Our results however are not consistent enough to allow us to make a general recommendation. Of the models which we were able to run on all versions of the datasets, totalling 52 iterations over different versions of our data, none were more accurate than the issue attitudes baseline. The highest accuracy was achieved by the L1 regularized multinomial logit model and the Random Forest with 28 %. This is not an accuracy suitable for the implementation of a broad microtargeting campaign that relies on the identification of individual level political preference.

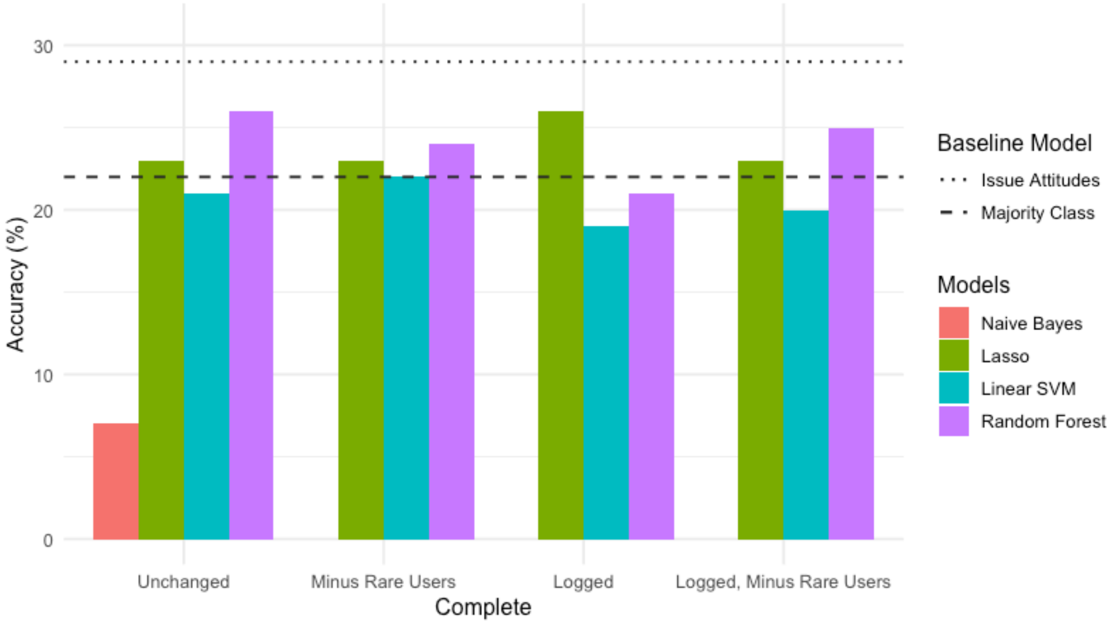


Figure 24: Model Comparison, Complete Dataset

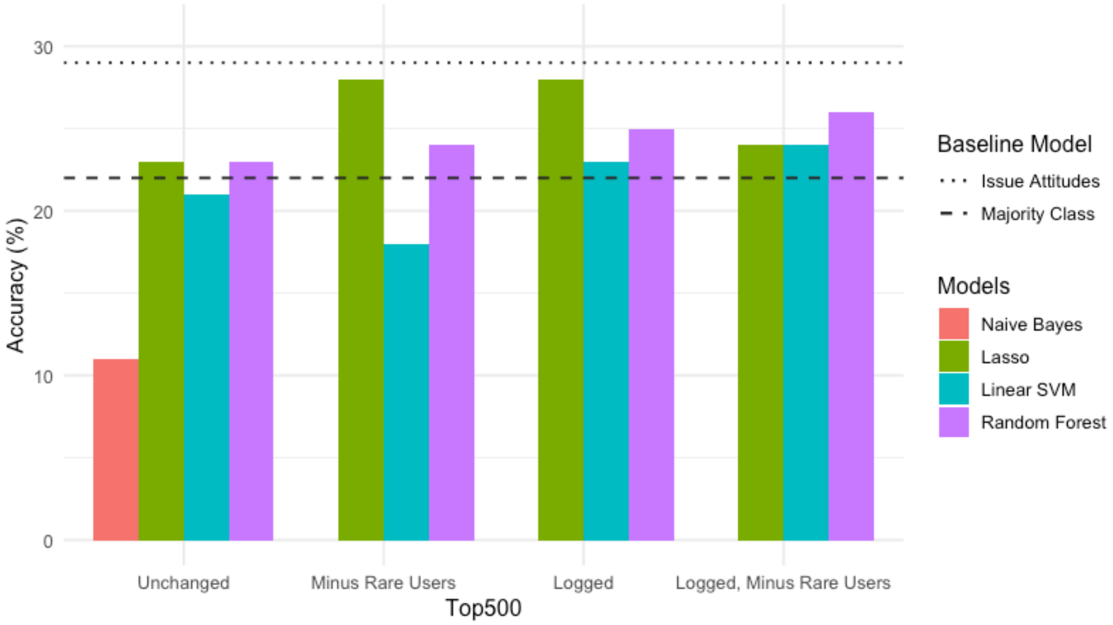


Figure 25: Model Comparison, Top 500 Dataset

At the same time however, 31 of our 52 multi-class models did outperform a majority-class guess. This clearly shows that domain-level data contains a signal that allows for the construction of models predicting individual level political affiliation. Furthermore, our study was particularly ambitious, as we attempted to predict seven different class labels simultaneously. In reality, this may not be necessary in all cases for the implementation of a microtargeting campaign. A foreign disinformation campaign meant to mobilise the voters of one particular party would only require the correct prediction of two different labels: voter and non-voter of that party. Despite this challenge, some of our models have nevertheless achieved relatively high levels of recall, particularly the LASSO (see Confusion Matrices in Appendix C). The LASSO on news domains, logged minus rare users, was able to correctly identify 74% of SPD voters.

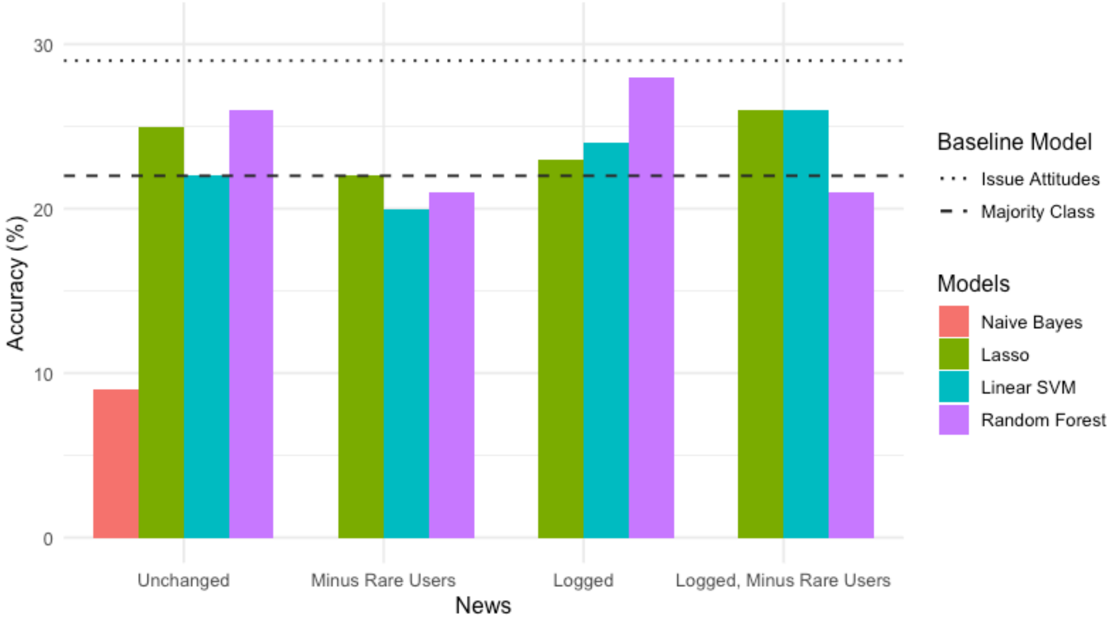


Figure 26: Model Comparison, News Dataset

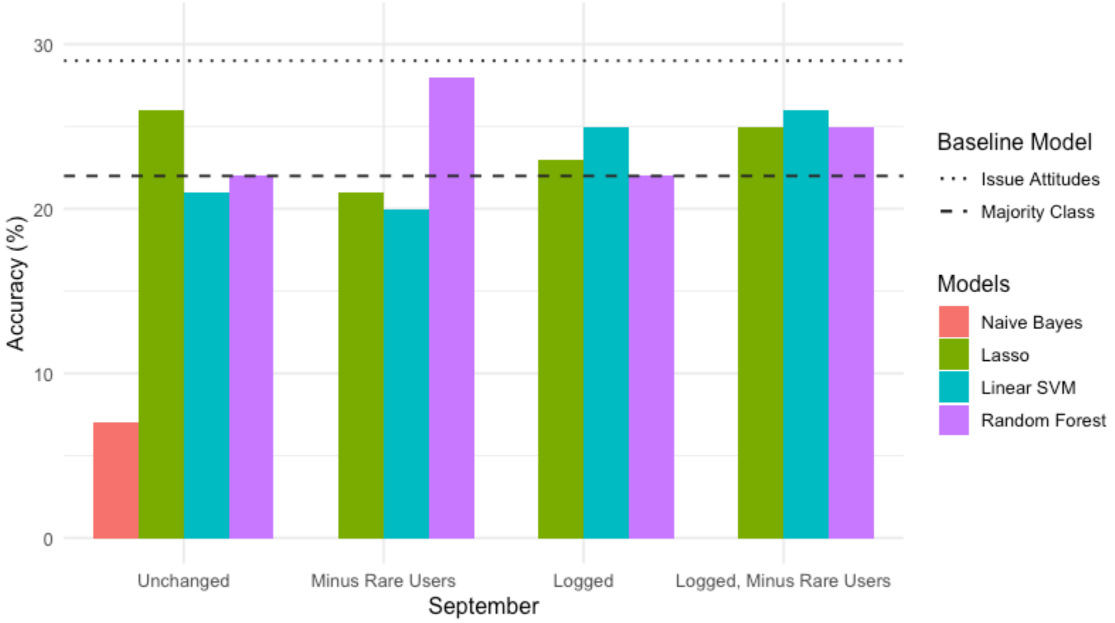


Figure 27: Model Comparison, September Dataset

The most promising models of those we were able to test on all dataset versions were the Random Forest and the LASSO. They had the highest mean accuracy across all different model versions, at 24 %. They also outperformed the majority class guess most consistently, namely 14 and 10 out of 16 times respectively.

Table 10: Model Comparison by Dataset and Version

Dataset	*Version*	*Naïve Bayes*	*LASSO Logit*	*Linear SVM*	*Random Forest*
Complete	Unchanged	7	23	21	26
Complete	Minus Rare Users	n/a	23	22	24
Complete	Logged	n/a	26	19	21
Complete	Logged Minus Rare Users	n/a	23	20	25
Top500	Unchanged	11	23	21	23
Top500	Minus Rare Users	n/a	28	18	24
Top500	Logged	n/a	28	23	25
Top500	Logged Minus Rare Users	n/a	24	24	26
News	Unchanged	9	25	22	26
News	Minus Rare Users	n/a	22	20	21
News	Logged	n/a	23	24	28
News	Logged Minus Rare Users	n/a	26	26	21
September	Unchanged	7	26	21	22
September	Minus Rare Users	n/a	21	20	28
September	Logged	n/a	23	25	22
September	Logged Minus Rare Users	n/a	25	26	25
	Overall Average	**9**	**24**	**22**	**24**

Finally, our study provides an avenue for future research with the promising results of the polynomial and L2 linear SVMs. The radial SVM performed poorly with only 19 % accuracy, however the polynomial SVM achieved an accuracy of 77 %, far surpassing any other model. We were only able to test these resource intensive SVMs on one dataset within the scope of this project, which naturally makes the insights less reliable than those of other models. The accuracies of the polynomial and regularized SVM however suggest that a more complex and computationally taxing SVM could be a method of choice for further research. Particularly the polynomial SVM suggests that there could be signal in our data we were so far not able to use to their fullest extent.

6 Conclusion

This study utilised machine learning in an attempt to predict individual voting preferences from comprehensive online tracking data. Contrary to previous studies reporting positive outcomes from using online socially generated data to make predictions, our findings indicate that online tracking data is not sufficient to make accurate vote choice predictions, at least within the context of our test. While some models did achieve higher accuracy than the majority-class baseline metric, these were inconsistent across datasets and versions. Furthermore, this is a very low bar by which to judge success: no model matched or bettered the issue attitudes baseline. From our study, we therefore conclude that online tracking data provides little, if any, accuracy advantage over other metrics of predicting vote choice. The domain-level tracking data used in this study did not sufficiently reflect individual political preferences and opinions in a useable, evaluable way, nor did it indicate that certain behavioural patterns are strongly associated with voting for a certain party. Coupled with the fact that such comprehensive online tracking data is (thankfully) not widely available, this means that this approach is unlikely to revolutionise the field of voter prediction.

6.1 Challenges and limitations

Our goal was indeed ambitious – there are few studies attempting to predict individual vote choice, and of those we found only one other example which attempted to predict vote choice in a multi-party system (see Kristensen et al., 2017). Multiclass classification tasks are a priori more difficult than simple binary tasks, which necessarily implies lower expected levels of accuracy when predicting vote choice in a multi-party system. However, as was presented in section 5.5, the inherent difficulty of the task is not the only issue at play. The poor performance of the binomial logit, performing worse than random guessing, shows that domain-level data is not as predictive as the literature may lead one to believe. Further research could investigate whether having more information, such as complete URLs or the content of websites, would improve the predictive power of the models.

The subject of this study also posed a unique set of challenges. The 2017 German federal election saw the AfD, a party previously without any parliamentary representation, become at a stroke the third largest party in the Bundestag (Arnett, 2017). The AfD gained supporters from across the political spectrum - 980,000 from the centre-right CDU, but also 400,000 from the Linke, a far left party (Bonsen, 2017). This blurring of traditional political lines may make it more difficult for our model to predict party support accurately. Future research could assess this postulation by applying similar methods to historical federal elections that did not see such voter migration. However, a model only able to predict elections that follow the anticipated trajectory contributes little to the field of election prediction.

Our study also faced limitations with respect to the data sample. Bias may have been introduced through individuals selecting into the sample. Those who choose to participate in paid online surveys such as this one may have very different internet habits than the general population. For example, in the variable importance plots of the Random Forest models, presented in Appendix E, survey sites such as 'YouGov' repeatedly appear among the top 20 most important variables. It appears unlikely that such results would hold true for the general population.

6.2 Implications

The success of the Trump and Brexit campaigns and the role of companies such as Cambridge Analytica has brought the issue of microtargeting to the forefront of public debate on elections and the use of big data (see for example, Dachwitz & Kurz, 2018). In order to anticipate the progress of machine learning and its potential role in changing the face of election campaigns worldwide, it is paramount that we have a solid understanding of the data available and the possibilities it opens for actors of various means and political affiliations. We have addressed the latter part of this question and shown that, even with comprehensive online tracking data, it is not possible to predict the party for which an individual will vote to a high level of accuracy.

This is not to say that there are no risks to democracy and privacy through the mass collection of large amounts of data. While we were not able to predict individual votes in the German context to an extent that would be necessary for a microtargeting campaign, online tracking data may be more predictive in other contexts, such as countries with a two-party system like the USA, or areas with a highly polarised political discourse. If similar prediction difficulties persist even in other contexts, this may imply that individuals' online behaviour is largely homogenous, which in turn leads to interesting implications regarding the existence of filter bubbles on the internet, or lack thereof. It may also provide an insight into general patterns of political engagement.

Finally, although domain-level data did not prove to be predictive of vote choice, this is not to say that individualised campaigning is not possible, nor that the data we produce cannot be used to ascertain information that individuals would not ordinarily divulge. Looking further into the content of visited domains could give companies an insight into individuals opinions on certain issues, which is a strong classifier for party choice as presented in Section 4.2.2. Indeed, internet advertising networks or large social media platforms already have access to much of this information. Complete URLs and even page content would also provide a fuller picture than was available to us in this study. Privacy concerns are paramount here, as the use of online tracking data deprives individuals of control over what is known about them and by whom. Not only is personal privacy at risk, but such detailed information about what the public is engaging with online leaves democracy vulnerable to foreign interference with the intent to disrupt public debate and spread misinformation.

Specifically in the German context, the implications are difficult to discern. Microtargeting may be most useful in closely fought elections to influence the vote outcome (Kind & Weide, 2017), particularly relevant in multi-party systems like Germany where an overall majority is difficult to achieve. At the same time, because of the protections of the GDPR, such microtargeting is more likely to come in the form of foreign interference. Nevertheless, the effectiveness of microtargeting in election campaigns is far from proven (Kind & Weide, 2017). This makes it difficult to predict the scale of potential foreign interference through disinformation campaigns. Research such as ours is therefore vital in planning to mitigate this risk.

6.3 Avenues for future research

Due to computational constraints resulting from the high-dimensionality of the data, we were unable to implement some more sophisticated machine learning methods which may have improved prediction accuracy. Regarding random forests, boosting methods such as AdaBoost (Freund & Schapire, 1999) and XGBoost (Chen & Guestrin, 2016) have proven effective in increasing accuracy by growing trees sequentially rather than in parallel, creating a stronger learner from an ensemble of weak learners. Similarly, ensemble LASSO methods (Urda, Franco, & Jerez, 2017) could be used to build on our LASSO model from section 5.2. Finally, smooth Support Vector Machines have proven effective in predicting outcomes using high-dimensionality data (Purnami, Andari, & Pertiwi, 2015).

Open questions exist surrounding individual-level vote prediction and the use of metadata in future political campaigning. Future research should focus on what data can tell us about individuals or groups of individuals in order to be able to anticipate its uses in the political sphere. It is critical for the public debate surrounding data-based prediction methods to keep pace with rapid technological developments. As political advertising adapts to the modern age, so too must regulations on campaign methods. The public and those charged with supporting robust democratic debate must be aware of and able to respond to disinformation and targeted campaigns, something that only possible with a good understanding of the tools and capabilities of both sides.

7 Bibliography

Abendschon, S., & Steinmetz, S. (2014). The Gender Gap in Voting Revisited: Women's Party Preferences in a European Context. *Social Politics: International Studies in Gender, State & Society*, *21*(2), 315–344. https://doi.org/10.1093/sp/jxu009

Arnett, G. (2017, September 25). *Bundestagswahl 2017: Wie Wähler die Parteien gewechselt haben - in Grafiken*. Retrieved from https://www.welt.de/politik/deutschland/article169010727/Die-Waehlerwanderung-in-Bildern.html

Asur, S., & Huberman, B. A. (2013). Predicting the Future with Social Media. *Applied Energy*, *112*, 1536–1543. https://doi.org/10.1016/j.apenergy.2013.03.027

Baron, D. (2018). *Who Identifies with the AfD? Explorative Analyses in Longitudinal Perspective* (No. 983; p. 20). Retrieved from Deutsche Institut für Wirtschaftsforschung website: https://www.diw.de/documents/publikationen/73/diw_01.c.601358.de/diw_sp0983.pdf

Beauchamp, N. (2017). Predicting and Interpolating State-Level Polls Using Twitter Textual Data: PREDICTING POLLS WITH TWITTER. *American Journal of Political Science*, *61*(2), 490–503. https://doi.org/10.1111/ajps.12274

Bermingham, A., & Smeaton, A. F. (2011, November 13). *On Using Twitter to Monitor Political Sentiment and Predict Election Results*. 9. Chiang Mai, Thailand.

Bonsen, G. (2017, September 25). Infografiken zur Bundestagswahl 2017: Wählerwanderung: Wie sich die Parteien gegenseitig die Stimmen abluchsten |

shz.de. *Shz*. Retrieved from https://www.shz.de/deutschland-welt/bundestagswahl/waehlerwanderung-wie-sich-die-parteien-gegenseitig-die-stimmen-abluchsten-id17921181.html

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bujała, A. (2012). GENDER DIFFERENCES IN INTERNET USAGE. *Social Science Quarterly*, *90*(2), 274–291.

Capara, G. V., Barbaranelli, C., & Zimbardo, P. G. (1999). Personality Profiles and Political Parties. *Political Psychology*, *20*(1), 175–197. https://doi.org/10.1111/0162-895X.00141

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 96–103. https://doi.org/10.1145/1390156.1390169

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France. *New Media & Society*, *16*(2), 340–358. https://doi.org/0.1177/1461444813480466

Charpentier, A. (2015, July 15). 'Variable Importance Plot' and Variable Selection. Retrieved 8 May 2019, from DZone Big Data website: https://dzone.com/articles/variable-importance-plot-and-variable-selection

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining - KDD '16, 785–794. https://doi.org/10.1145/2939672.2939785

Chung, J., & Mustafaraj, E. (2011). Can Collective Sentiment Expressed on Twitter Predict Political Elections? *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Presented at the AAAI, San Francisco, California, USA.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cranmer, S. J., & Desmarais, B. A. (2017). What Can We Learn from Predictive Modeling? *Political Analysis*, *25*(02), 145–166. https://doi.org/10.1017/pan.2017.3

Dachwitz, I., & Kurz, C. (2018). *Microtargeting und Manipulation: Von Cambridge Analytica zur EU-Wahl*. Retrieved from /v/35c3-10037-microtargeting_und_manipulation

Devine, C. J. (2015). Ideological Social Identity: Psychological Attachment to Ideological In-Groups as a Political Phenomenon and a Behavioral Influence. *Political Behavior*, *37*(3), 509–535. https://doi.org/10.1007/s11109-014-9280-6

Do Viet Phuong, & Tu Minh Phuong. (2014). Gender Prediction Using Browsing History. In *Advances in Intelligent Systems and Computing*: *Vol. 244. Knowledge and Systems Engineering* (pp. 271–283). https://doi.org/10.1007/978-3-319-02741-8_24

Fallows, D. (2005). *How Women and Men Use the Internet*. Retrieved from Pew Research Center website: https://www.pewresearch.org/wp-content/uploads/sites/9/2005/12/PIP_Women_and_Men_online.pdf

Freund, Y., & Schapire, R. E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*(5), 771–780.

Gayo-Avello, D. (2013). A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, *31*(6), 649–679. https://doi.org/10.1177/0894439313493979

Gentzkow, M., & Shapiro, J. M. (2011). Ideological Segregation Online and Offline. *The Quarterly Journal of Economics*, *126*(4), 1799–1839. https://doi.org/10.1093/qje/qjr044

Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M., & Ha, S. E. (2010). Personality and Political Attitudes: Relationships across Issue Domains and Political Contexts. *American Political Science Review*, *104*(1), 111–133. https://doi.org/10.1017/S0003055410000031

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gidron, N., & Hall, P. A. (2017). The politics of social status: economic and cultural roots of the populist right. *British Journal of Sociology*, 29.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Glance, N., & Adamic, L. (2005). The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Proceedings of the 3rd International Workshop on Link Discovery*, 16. Chicago, Illinois.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, *107*(41), 17486–17490. https://doi.org/10.1073/pnas.1005962107

Goerres, A. (2008). The grey vote: Determinants of older voters' party choice in Britain and West Germany. *Electoral Studies*, *27*(2), 285–304. https://doi.org/10.1016/j.electstud.2007.12.007

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning* (2nd ed.). Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, *36*(3), 1171–1220. https://doi.org/10.1214/009053607000000677

Iyengar, S., & Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, *59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: with applications in R*. New York: Springer.

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, *80*(4), 557–571. https://doi.org/10.1037/0022-3514.80.4.557

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, TO, Sander, PG, & Welpe, IM "Predicting Elections With Twitter:

What 140 Characters Reveal About Political Sentiment”. *Social Science Computer Review*, *30*(2), 229–234. https://doi.org/10.1177/0894439311404119

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support. *Social Science Computer Review*, *35*(3), 336–356. https://doi.org/10.1177/0894439316631043

Kastenmüller, A., Greitemeyer, T., Jonas, E., Fischer, P., & Frey, D. (2010). Selective exposure: The impact of collectivism and individualism. *British Journal of Social Psychology*, *49*(4), 745–763. https://doi.org/10.1348/014466609X478988

Kaufman, S., Rosset, S., & Perlich, C. (2011). Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), 556–563. https://doi.org/10.1145/2020408.2020496

Kind, S., & Weide, S. (2017). *Microtargeting: psychometrische Analyse mittels Big Data* (No. Themekurzprofil Nr. 18; p. 12). Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kosinski, Michal, Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, *95*(3), 357–380. https://doi.org/10.1007/s10994-013-5415-y

Kristensen, J. B., Albrechtsen, T., Dahl-Nielsen, E., Jensen, M., Skovrind, M., & Bornakke, T. (2017). Parsimonious data: How a single Facebook like predicts voting behavior in multiparty systems. *PLOS ONE*, *12*(9), e0184562. https://doi.org/10.1371/journal.pone.0184562

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Lampos, V., & Cristianini, N. (2010, July). *Tracking the flu pandemic by monitoring the Social Web*. Presented at the Cognitive Information Processing (CIP). https://doi.org/10.1109/CIP.2010.5604088

Lui, C. T., Metaxas, P. T., & Mustafaraj, E. (2011, March). *On the Predictability of the US Elections Through Search Volume Activity*. Presented at the IADIS International Conference on e-Society, Avila, Spain. Retrieved from https://pdfs.semanticscholar.org/5ca6/58cf819fd3ee220c2526054b3de84e34d4ab.pdf

Mader, M., & Schoen, H. (2017). Ideological voting in context: The case of Germany during the Merkel era. In *Voters and Voting in Context: Multiple Contexts and the Heterogeneous German Electorate*. Oxford: Oxford University Press.

McCrae, R. R., & Costa, P. T. (1997). Personality Trait Structure as a Human Universal. *American Psychologist*, 8.

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3), 205316801772000. https://doi.org/10.1177/2053168017720008

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October 9). *How (Not) to Predict Elections*. Presented at the PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Confernece on Social Computing (SocialCom), Boston, MA, USA. https://doi.org/10.1109/PASSAT/SocialCom.2011.98

Mondak, J. J., Hibbing, M. V., Canache, D., Seligson, M. A., & Anderson, M. R. (2010). Personality and Civic Engagement: An Integrative Framework for the Study of Trait Effects on Political Behavior. *American Political Science Review*, *104*(1), 85–110. https://doi.org/10.1017/S0003055409990359

Mullainathan, S., & Shleifer, A. (2005). The Market for News. *The American Economic Review*, *95*(4), 24.

Mustafaraj, E., Finn, S., Whitlock, C., & Metaxas, P. T. (2011). Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail. *Proceedings of PASSAT/SocialCom*, 8.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010, May 23). *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*. Presented at the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington DC, USA. Retrieved from https://www.researchgate.net/publication/221297841_From_Tweets_to_Polls_Linking_Text_Sentiment_to_Public_Opinion_Time_Series

Parr, T., Turgutlu, K., Csiszar, C., & Howard, J. (2018, March 26). Beware Default Random Forest Importances. Retrieved 8 May 2019, from Explained AI website: https://explained.ai/rf-importance/index.html#3

Pennacchiotti, M., & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter User Classification. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Presented at the International Conference on Web and Social Media, Barcelona, Spain.

PrakashSwami, C., Bhalchandra Tarte, P., Kisan Rakshe, S., Maroti Raut, S., & Faiz Shaikh, N. (2015). Detecting the Age of a Person through Web Browsing Patterns. *International Journal of Computer Applications*, *118*(12), 8–12. https://doi.org/10.5120/20795-3455

Purnami, S. W., Andari, S., & Pertiwi, Y. D. (2015). High-Dimensional Data Classification Based on Smooth Support Vector Machines. *Procedia Computer Science*, *72*, 477–484. https://doi.org/10.1016/j.procs.2015.12.129

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in Twitter. *In 2nd International Workshop on Search and Mining UserGenerated Content. ACM*.

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch Senate Election Results with Twitter. *Proceedings of the Workshop on Semantic Analysis in Social Media*, 53–60. Retrieved from http://dl.acm.org/citation.cfm?id=2389969.2389976

Schoen, H., & Schumann, S. (2007). Personality Traits, Partisan Attitudes, and Voting Behavior. Evidence from Germany. *Political Psychology*, *28*(4), 471–498. https://doi.org/10.1111/j.1467-9221.2007.00582.x

Soldz, S., & Vaillant, G. E. (1999). The Big Five Personality Traits and the Life Course: A 45-Year Longitudinal Study. *Journal of Research in Personality*, *33*(2), 208–232. https://doi.org/10.1006/jrpe.1999.2243

Sunstein, C. (2001). *Republic.com*. NJ, USA: Princeton University Press.

Tirado-Morueta, R., Aguaded-Gómez, J. I., & Hernando-Gómez, Á. (2018). The socio-demographic divide in Internet usage moderated by digital literacy support. *Technology in Society*, *55*, 47–55. https://doi.org/10.1016/j.techsoc.2018.06.001

Túlio Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Computing Research Repository*, *abs/1602.04938*. Retrieved from http://arxiv.org/abs/1602.04938

Tumasjan, A., Sprenger, T. O., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Journal Of The International Linguistic Association*. Presented at the Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM, Washington DC, USA. Retrieved from https://www.researchgate.net/publication/215776042_Predicting_Elections_with_Twitter_What_140_Characters_Reveal_about_Political_Sentiment

Urda, D., Franco, L., & Jerez, J. M. (2017). Classification of high dimensional data using LASSO ensembles. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7. https://doi.org/10.1109/SSCI.2017.8280875

Walczak, A. (2010, March). *Values, Ideology and Party Choice in Europe*. 30. Münster.

Yasseri, T., & Bright, J. (2016). Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0083-3

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the*

National Academy of Sciences, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zillien, N., & Hargittai, E. (2009). Digital Distinction: Status-Specific Types of Internet Usage. *Social Science Quarterly*, *90*(2), 274–291. https://doi.org/10.1111/j.1540-6237.2009.00617.x

Appendix A Variables Selected from the Pulse Survey

Variables of YouGov Pulse Survey used in Baseline Model Multinomial Logistic Regression

Variable Name	*Coding*
Gender	Male (1) or Female (2)
Age	In years
Marital Status	1. Married and living together with spouse 2. Married and separated from spouse 3. In a civil-same-sex partnership, living together 4. In a civil-same-sex partnership, living separated 5. Single 6. Divorced
Education	1. Finished school without school leaving certificate 2. Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling ("Hauptschulabschluss, Volksschulabschluss" 3. Intermediary secondary qualification, after 10 years of schooling ("Mittlere Reife, Realschulabschluss, or Polytechnische Oberschule mit Abschluss 10. Klasse") 4. Certificate fulfilling entrance requirements to study at a polytechnical college ("Fachhochschulreife (Abschluss einer Fachoberschule etc.)") 5. Higher qualification, entitling holders to study at a university ("Abitur or Erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife)") 6. I am still in high school
Employment Status	1. in full-time employment (more than 30 hours/week) 2. in part-time employment (up to 30 hours/week) 3. in a traineeship or apprenticeship 4. high school student 5. college student 6. currently on a retraining course 7. currently unemployed 8. currently on short-time work 9. alternative community service 10. in early retirement, retirement, on a pension (formerly employed) 11. on maternity leave, parental leave 12. not in full or part-time employment (Housewife/househusband)

Household Income (monthly)	1. below 500 Euros 2. 500 up to 1000 Euros 3. 1000 up to 1500 Euros 4. 1500 up to 2000 Euros 5. 2000 up to 2500 Euros 6. 2500 up to 3000 Euros 7. 3000 up to 3500 Euros 8. 3500 up to 4000 Euros 9. 4000 up to 4500 Euros 10. 4500 up to 5000 Euros 11. 5000 up to 10000 Euros 12. 10000 Euros and more
Self-reported left-right scale	1 = left, 11 = right
Government rating	Are you rather satisfied or rather dissatisfied with the performance of the federal government made up of CDU/CSU and SPD? (scale = -5 to 5)
Party rating: CDU	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Party rating: CSU	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Party rating: SPD	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Party rating: Green	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Party rating: FDP	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Party rating: Linke	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Party rating: AfD	-5 = I have a very negative view of this party | 5 = I have a very positive view of this party
Issue attitudes: Immigrants who come to Germany due to economic reasons should be deported.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don't know
Issue attitudes: Germany should back out as quickly as possible from coal energy even if this means higher energy prices.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don't know

Issue Attitudes: Immigrants should be obliged to adapt to the German culture.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: Taxes for people with high income should be increased.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: Germany should take a stand for more cooperation within the EU.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: Overall, German tax payers have to be relieved even if this means less money for public investments.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: Germany must increase protection against terrorist attacks even if this means more surveillance of all citizens.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: The EU should make no big concessions to the UK in the Brexit negotiations even if this could harm consumer prices in Germany.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: The government must take legal action against the spread of “fake news” in social media channels.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: Germany should engage more in international crises.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know
Issue Attitudes: Germany should invest more in	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know

infrastructure (e.g., streets, schools, internet) even if this means more borrowing.	
Issue Attitudes: Germany needs an upper limit for refugees.	<1> fully disagree <2> rather disagree <3> neither/nor <4> rather agree <5> fully agree <977> don’t know

Appendix B Baseline Regression Output

Regression tables for Multinomial Logistic Regressions carried out to create the baseline models can be requested from our GitHub Repository.

Appendix C Confusion Matrices – All Models

Baseline Models

Confusion Matrix: Baseline Model, Socio-Demographic

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.21	0.24	0.09	0.15	0.24	0.13	0.12
Recall	0.24	0.36	0.03	0.06	0.35	0.06	0.05
F1	0.22	0.29	0.05	0.09	0.28	0.08	0.07

Confusion Matrix: Baseline Model, Issue Attitudes

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.27	0.25	0.19	0.22	0.39	0.35	0
Recall	0.33	0.28	0.11	0.24	0.38	0.41	0
F1	0.3	0.26	0.14	0.23	0.39	0.38	NaN

Confusion Matrix: Baseline Model, Ideology

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.41	0.45	0.38	0.33	0.59	0.51	0.09
Recall	0.41	0.36	0.5	0.39	0.53	0.51	0.12
F1	0.41	0.4	0.43	0.36	0.56	0.51	0.1

Confusion Matrix: Baseline Model, All Variables

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.34	0.48	0.31	0.29	0.5	0.46	0
Recall	0.28	0.46	0.32	0.25	0.55	0.49	0
F1	0.31	0.47	0.31	0.27	0.52	0.47	NaN

Naïve Bayes

Confusion Matrix: Naive Bayes, Complete, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	1	1	0	0	NA	NA	0.06
Recall	0.01	0.01	0	0	0	0	1
F1	0.03	0.03	NaN	NaN	NA	NA	0.12

Confusion Matrix: Naive Bayes, Top 500, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.33	0	0.21	0.14	0	0.21	0.06
Recall	0.03	0	0.18	0.35	0	0.12	0.59
F1	0.05	NaN	0.19	0.21	NaN	0.16	0.12

Confusion Matrix: Naive Bayes, September, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	NA	NA	0	0.33	NA	NA	0.06
Recall	0	0	0	0.08	0	0	0.95
F1	NA	NA	NaN	0.12	NA	NA	0.11

Confusion Matrix: Naive Bayes, News, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.25	0	0.25	0.07	0	0.5	0.07
Recall	0.05	0	0.07	0.15	0	0.05	0.83
F1	0.09	NaN	0.11	0.1	NaN	0.09	0.13

LASSO

Confusion Matrix: LASSO, Complete, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.26	0.26	0.0	0.17	0.21	0.18	0.0
Recall	0.15	0.7	0.0	0.06	0.14	0.08	0.0
F1	0.19	0.38	NaN	0.93	0.17	0.11	NaN

Confusion Matrix: LASSO, Complete, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.25	0.25	0.1	0.19	0.24	0.23	0.1
Recall	0.2	0.48	0.06	0.14	0.2	0.15	0.05
F1	0.23	0.33	0.08	0.16	0.22	0.18	0.07

Confusion Matrix: LASSO, Complete, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.19	0.28	0.13	0.25	0.35	0.19	0.33
Recall	0.19	0.58	0.06	0.08	0.29	0.13	0.06
F1	0.19	0.38	0.09	0.12	0.31	0.15	0.1

Confusion Matrix: LASSO, Complete, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.26	0.25	0.11	0.1	0.21	0.32	0.09
Recall	0.27	0.39	0.04	0.05	0.26	0.21	0.07
F1	0.26	0.30	0.06	0.06	0.24	0.25	0.08

Confusion Matrix: LASSO, Top 500, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.14	0.24	0.0	0.0	0.21	0.4	0.0
Recall	0.88	0.7	0.0	0.0	0.15	0.16	0.0

F1	0.11	0.35	NaN	NaN	0.18	0.23	NaN

Confusion Matrix: LASSO, Top 500, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.24	0.29	0.38	0.0	0.30	0.38	0.0
Recall	0.31	0.61	0.09	0.0	0.32	0.13	0.0
F1	0.27	0.39	0.15	NaN	0.31	0.19	NaN

Confusion Matrix: LASSO, Top 500, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.18	0.31	0.00	0.00	0.32	0.56	NA
Recall	0.28	0.66	0.0	0.0	0.27	0.13	0.0
F1	0.22	0.42	NaN	NaN	0.29	0.21	NaN

Confusion Matrix: LASSO, Top 500, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.24	0.3	0.0	0.12	0.23	0.25	0.0
Recall	0.28	0.48	0.0	0.08	0.24	0.18	0.00
F1	0.26	0.37	NaN	0.1	0.23	0.21	NaN

Confusion Matrix: LASSO, News, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.21	0.28	0.14	0.27	0.25	0.3	0.17
Recall	0.21	0.42	0.1	0.15	0.27	0.23	0.11
F1	0.21	0.34	0.12	0.20	0.26	0.26	0.13

Confusion Matrix: LASSO, News, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*

Precision	0.18	0.22	0.18	0.33	0.21	0.26	0.38
Recall	0.25	0.31	0.87	0.1	0.26	0.15	0.2
F1	0.21	0.26	0.12	0.15	0.23	0.19	0.26

Confusion Matrix: LASSO, News, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.22	0.27	0.25	0.00	0.25	0.26	0.0
Recall	0.31	0.45	0.1	0.0	0.21	0.15	0.0
F1	0.26	0.34	0.14	NaN	0.23	0.19	NaN

Confusion Matrix: LASSO, News, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.39	0.37	0.13	0.95	0.33	0.18	0.0
Recall	0.65	0.74	0.97	0.97	0.85	0.93	0.99
F1	0.21	0.27	0.3	0.25	0.35	0.3	0.0

Confusion Matrix: LASSO, September, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.19	0.3	0.17	0.17	0.3	0.3	0.0
Recall	0.2	0.62	0.1	0.03	0.24	0.15	0.0
F1	0.2	0.40	0.13	0.06	0.27	0.2	NaN

Confusion Matrix: LASSO, September, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.26	0.26	0.0	0.0	0.21	0.28	0.0
Recall	0.35	0.32	0.0	0.0	0.15	0.31	0.0

F1	0.3	0.29	NaN	NaN	0.17	0.29	NaN

Confusion Matrix: LASSO, September, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.21	0.25	0.13	0.0	0.26	0.32	0.0
Recall	0.22	0.54	0.07	0.0	0.2	0.17	0.0
F1	0.21	0.34	0.09	NaN	0.23	0.22	NaN

Confusion Matrix: LASSO, September, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.25	0.34	0.08	0.0	0.23	0.28	0.0
Recall	0.37	0.47	0.04	0.0	0.19	0.22	0.0
F1	0.3	0.39	0.06	NaN	0.21	0.25	NaN

SVM

Confusion Matrix: Linear SVM, Complete, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.23	0.25	0.16	0.15	0.18	0.21	0.1
Recall	0.12	0.55	0.12	0.13	0.09	0.16	0.09
F1	0.16	0.35	0.14	0.14	0.12	0.18	0.09

Confusion Matrix: Linear SVM, Complete, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.24	0.28	0.1	0.22	0.2	0.22	0.05
Recall	0.19	0.41	0.09	0.29	0.17	0.15	0.05

F1	0.21	0.33	0.1	0.25	0.18	0.18	0.05

Confusion Matrix: Linear SVM, Complete, Logged

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.15	0.21	0	0.11	0.19	0.25	0.12
Recall	0.16	0.43	0	0.04	0.16	0.15	0.06
F1	0.15	0.28	NaN	0.06	0.17	0.19	0.08

Confusion Matrix: Linear SVM, Complete, Logged Minus Rare Users

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.17	0.22	0.2	0	0.16	0.27	NA
Recall	0.16	0.53	0.04	0	0.17	0.12	0
F1	0.16	0.31	0.07	NaN	0.16	0.17	NA

Confusion Matrix: Linear SVM, Top 500, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.13	0.26	0.12	0.24	0.24	0.24	0
Recall	0.15	0.42	0.06	0.13	0.23	0.16	0
F1	0.14	0.32	0.08	0.17	0.23	0.19	NaN

Confusion Matrix: Linear SVM, Top 500, Minus Rare Users

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.22	0.22	0.11	0.1	0.24	0.1	0.12
Recall	0.27	0.23	0.12	0.07	0.18	0.09	0.16
F1	0.24	0.22	0.12	0.08	0.21	0.09	0.14

Confusion Matrix: Linear SVM, Top 500, Logged

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*

Precision	0.19	0.25	0	0.13	0.27	0.25	0
Recall	0.28	0.49	0	0.08	0.21	0.1	0
F1	0.23	0.34	NaN	0.1	0.24	0.15	NaN

Confusion Matrix: Linear SVM, Top 500, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.2	0.28	0.17	0.12	0.26	0.33	0
Recall	0.28	0.48	0.03	0.08	0.2	0.21	0
F1	0.24	0.35	0.06	0.1	0.22	0.26	NaN

Confusion Matrix: Linear SVM, News, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.23	0.25	0.19	0.06	0.27	0.2	0.07
Recall	0.24	0.49	0.1	0.04	0.18	0.1	0.06
F1	0.23	0.34	0.13	0.05	0.22	0.14	0.06

Confusion Matrix: Linear SVM, News, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.18	0.19	0.17	0	0.33	0.21	0.2
Recall	0.09	0.61	0.04	0	0.11	0.12	0.07
F1	0.12	0.29	0.07	NaN	0.16	0.15	0.1

Confusion Matrix: Linear SVM, News, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.28	0.24	0.14	0	0.33	0.14	0
Recall	0.38	0.52	0.03	0	0.16	0.08	0
F1	0.32	0.33	0.05	NaN	0.22	0.1	NaN

Confusion Matrix: Linear SVM, News, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.21	0.24	0.17	0.17	0.35	0.38	0
Recall	0.25	0.57	0.04	0.05	0.28	0.15	0
F1	0.23	0.34	0.07	0.07	0.31	0.22	NaN

Confusion Matrix: Linear SVM, September, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.28	0.25	0	0	0.27	0.44	0
Recall	0.2	0.78	0	0	0.12	0.1	0
F1	0.24	0.38	NaN	NaN	0.16	0.16	NaN

Confusion Matrix: Linear SVM, September, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.22	0.23	0.04	0.13	0.25	0.22	0.06
Recall	0.2	0.32	0.04	0.11	0.23	0.16	0.06
F1	0.21	0.27	0.04	0.12	0.24	0.18	0.06

Confusion Matrix: Linear SVM, September, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.28	0.25	0	0	0.27	0.44	0
Recall	0.2	0.78	0	0	0.12	0.1	0
F1	0.24	0.38	NaN	NaN	0.16	0.16	NaN

Confusion Matrix: Linear SVM, September, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.24	0.26	0	0	0.36	0.43	0

Recall	0.35	0.49	0	0	0.3	0.19	0
F1	0.28	0.34	NaN	NaN	0.33	0.26	NaN

Confusion Matrix: L2 Regularized Linear SVM, Top 500, Unchanged

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.48	0.43	0.55	0.17	0.57	0.31	0.6
Recall	0.4	0.51	0.18	0.45	0.4	0.33	0.27
F1	0.44	0.47	0.27	0.25	0.47	0.32	0.37

Confusion Matrix: SVM Polynomial, Top 500, Unchanged

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.76	0.67	0.86	0.88	0.8	0.78	0.89
Recall	0.79	0.82	0.76	0.71	0.72	0.78	0.73
F1	0.78	0.74	0.81	0.79	0.76	0.78	0.8

Confusion Matrix: Radial SVM, Top 500, Unchanged

Row	*CDU/CSU*	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.18	0.21	NA	0.25	0.18	0	NA
Recall	0.16	0.48	0	0.03	0.23	0	0
F1	0.17	0.29	NA	0.06	0.2	NaN	NA

Random Forest

Confusion Matrix: Random Forest, Complete, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.23	0.26	NA	0	0.18	0.62	NA
Recall	0.29	0.61	0	0	0.17	0.2	0
F1	0.26	0.36	NA	NaN	0.18	0.31	NA

Confusion Matrix: Random Forest, Complete, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.21	0.26	NA	NA	0.22	0.27	NA
Recall	0.28	0.54	0	0	0.23	0.15	0
F1	0.24	0.35	NA	NA	0.22	0.19	NA

Confusion Matrix: Random Forest, Complete, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.2	0.22	NA	0	0.24	0.21	NA
Recall	0.29	0.43	0	0	0.21	0.13	0
F1	0.23	0.29	NA	NaN	0.22	0.16	NA

Confusion Matrix: Random Forest, Complete, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.21	0.22	0.5	0	0.28	0.35	NA
Recall	0.27	0.41	0.04	0	0.37	0.21	0
F1	0.24	0.29	0.08	NaN	0.32	0.26	NA

Confusion Matrix: Random Forest, Top500, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.22	0.25	0	0.25	0.23	0.4	NA
Recall	0.29	0.61	0	0.03	0.17	0.12	0
F1	0.25	0.35	NaN	0.06	0.19	0.19	NA

Confusion Matrix: Random Forest, Top500, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.2	0.28	0	NA	0.24	0.17	NA
Recall	0.27	0.55	0	0	0.32	0.07	0
F1	0.23	0.37	NaN	NA	0.27	0.09	NA

Confusion Matrix: Random Forest, Top500, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.17	0.29	0.33	0	0.26	0.45	NA
Recall	0.22	0.58	0.03	0	0.3	0.13	0
F1	0.2	0.38	0.06	NaN	0.28	0.2	NA

Confusion Matrix: Random Forest, Top500, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.26	0.25	NA	0.25	0.26	0.33	NA
Recall	0.37	0.56	0	0.04	0.24	0.11	0
F1	0.31	0.35	NA	0.07	0.25	0.16	NA

Confusion Matrix: Random Forest, News, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.23	0.27	0	0	0.25	0.5	NA
Recall	0.29	0.51	0	0	0.34	0.18	0
F1	0.26	0.35	NaN	NaN	0.29	0.26	NA

Confusion Matrix: Random Forest, News, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.2	0.17	1	NA	0.25	0.25	0

Recall	0.32	0.29	0.09	0	0.33	0.12	0
F1	0.24	0.21	0.16	NA	0.29	0.16	NaN

Confusion Matrix: Random Forest, News, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.22	0.29	NA	0	0.29	0.44	NA
Recall	0.29	0.51	0	0	0.41	0.21	0
F1	0.25	0.37	NA	NaN	0.34	0.28	NA

Confusion Matrix: Random Forest, News, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.17	0.18	1	0	0.25	0.25	NA
Recall	0.23	0.29	0.09	0	0.39	0.15	0
F1	0.19	0.22	0.16	NaN	0.31	0.19	NA

Confusion Matrix: Random Forest, September, Unchanged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.22	0.21	NA	0	0.22	0.4	NA
Recall	0.29	0.41	0	0	0.27	0.15	0
F1	0.25	0.27	NA	NaN	0.24	0.21	NA

Confusion Matrix: Random Forest, September, Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.26	0.24	0	1	0.3	0.69	NA
Recall	0.33	0.45	0	0.05	0.38	0.28	0
F1	0.29	0.31	NaN	0.1	0.33	0.4	NA

Confusion Matrix: Random Forest, September, Logged

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.2	0.2	1	NA	0.2	0.5	NA
Recall	0.32	0.33	0.03	0	0.25	0.17	0
F1	0.25	0.25	0.06	NA	0.22	0.25	NA

Confusion Matrix: Random Forest, September, Logged Minus Rare Users

	CDU/CSU	*SPD*	*FDP*	*Green*	*Linke*	*AfD*	*Other Party*
Precision	0.2	0.23	NA	NA	0.25	0.83	NA
Recall	0.26	0.51	0	0	0.3	0.16	0
F1	0.23	0.32	NA	NA	0.27	0.26	NA

Bivariate Regression

Confusion Matrix: Binomial Logit, Various Datasets

	Precision	*Recall*	*F1*
Complete, Unchanged	0.46	0.47	0.47
Complete, Minus Rare Users	0.5	0.56	0.53
Complete, Logged	0.45	0.51	0.48
Complete, Logged Minus Rare Users	0.45	0.57	0.5
Top500, Unchanged	0.45	0.46	0.46
Top500, Minus Rare Users	0.46	0.48	0.47
Top500, Logged	0.43	0.45	0.44
Top500, Logged Minus Rare Users	0.45	0.48	0.47
News, Unchanged	0.48	0.53	0.5
News, Minus Rare Users	0.45	0.53	0.49
News, Logged	0.52	0.65	0.57
News, Logged Minus Rare Users	0.42	0.5	0.46

September, Unchanged	0.43	0.41	0.42
September, Minus Rare Users	0.47	0.45	0.46
September, Logged	0.45	0.48	0.47
September, Logged Minus Rare Users	0.49	0.54	0.51

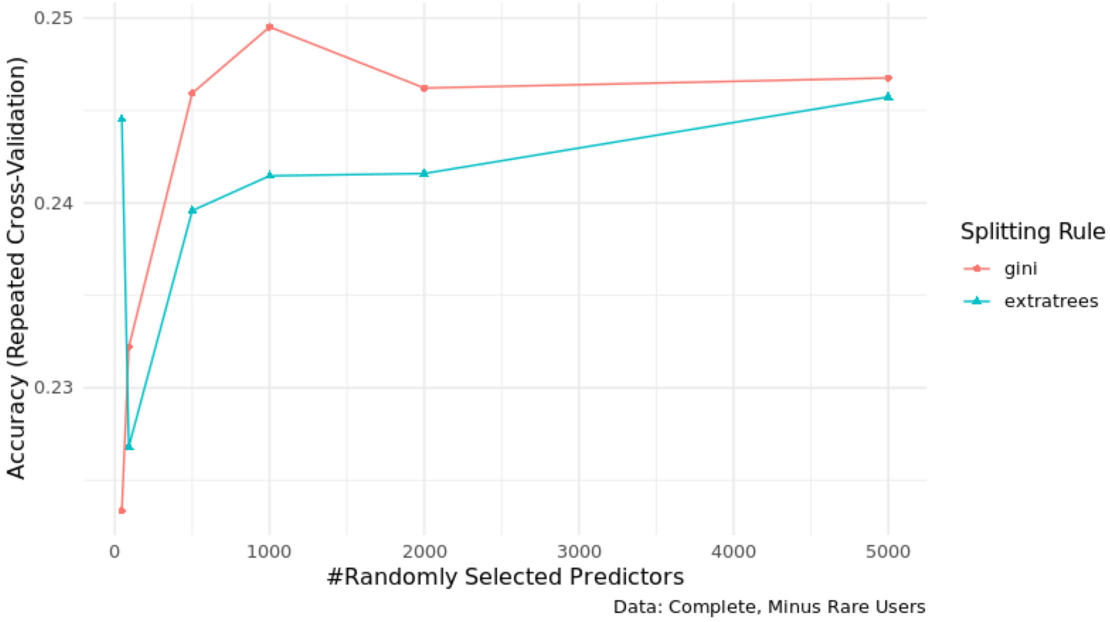
Confusion Matrix: Random Forest, News, logged

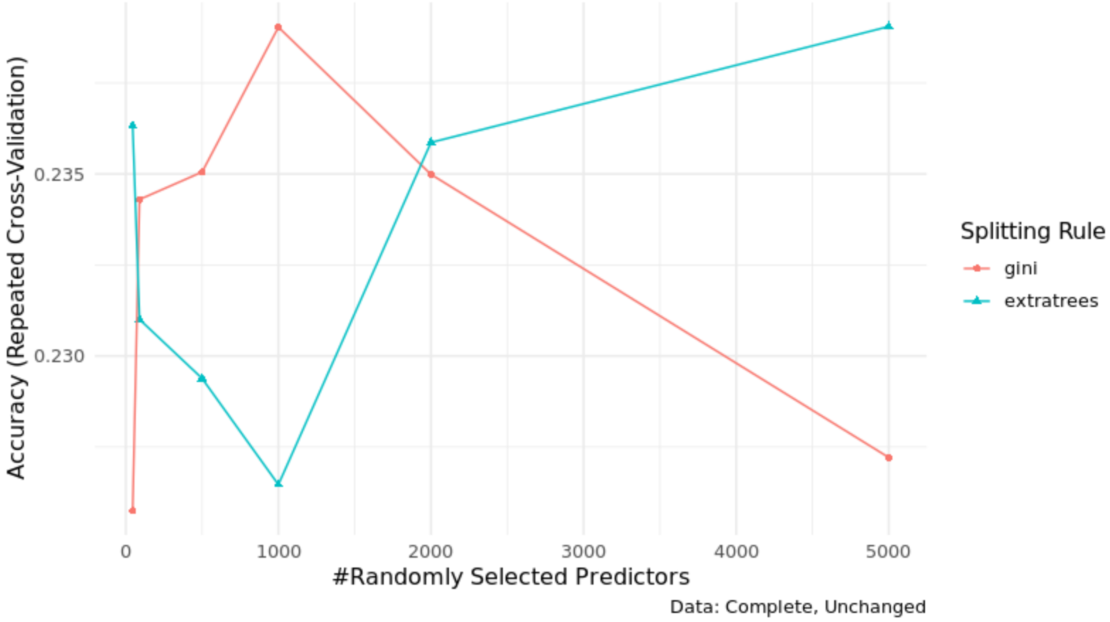
	Precision	*Recall*	*F1*
News, logged	0.58	0.41	0.48

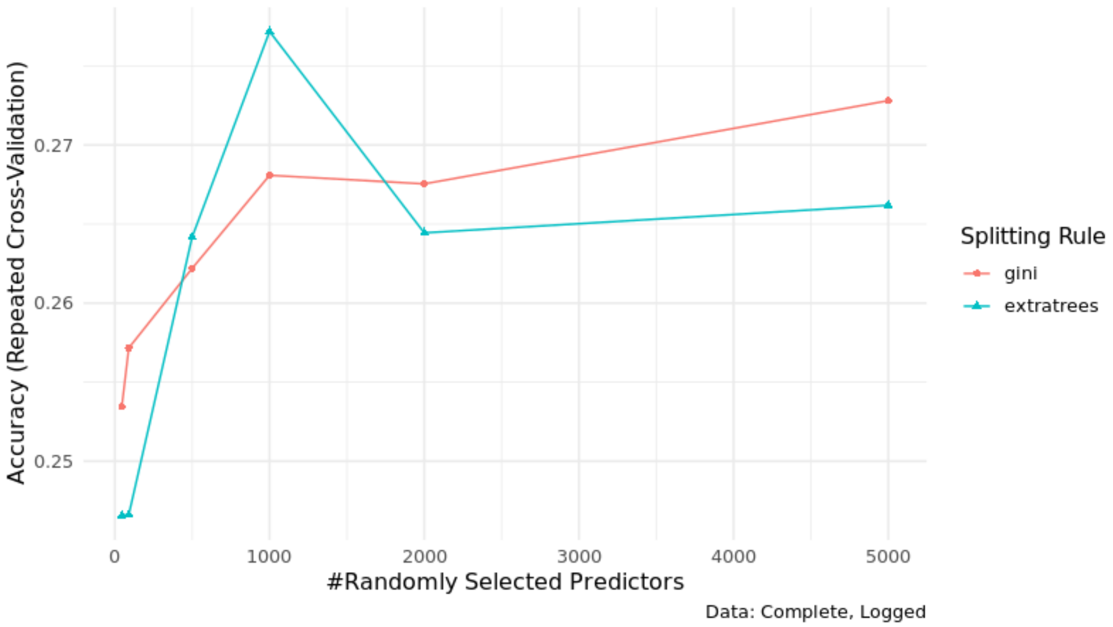
Appendix D Random Forest: Model Fit Plots

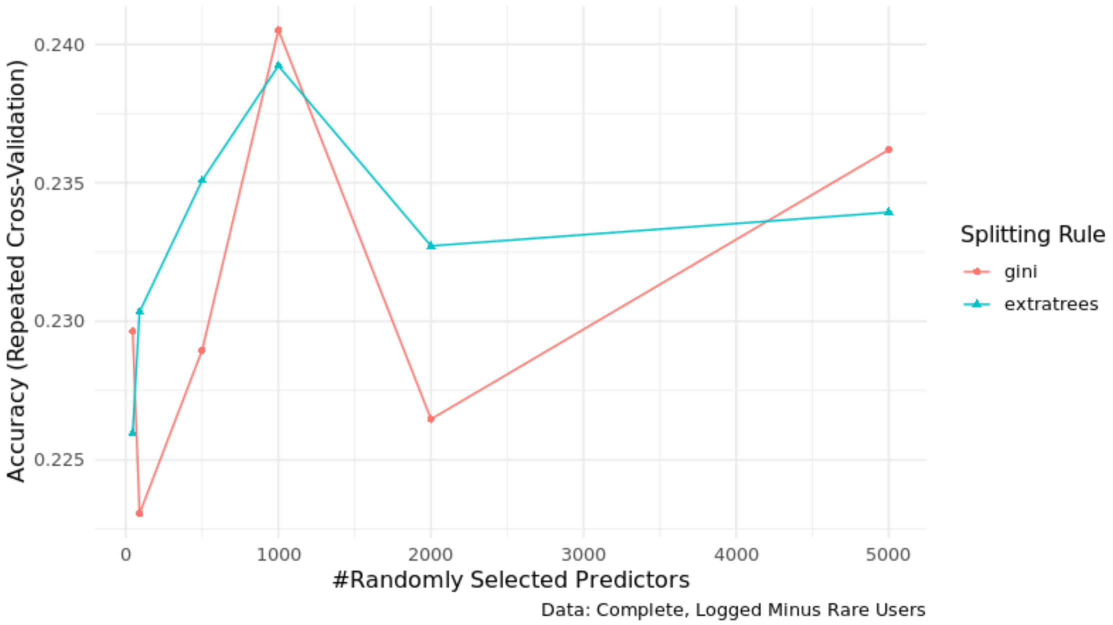
Plots showing the values of Mtry (the number of variables randomly selected at each split) which maximise accuracy in both a Random Forest model and an Extratrees model.

Complete

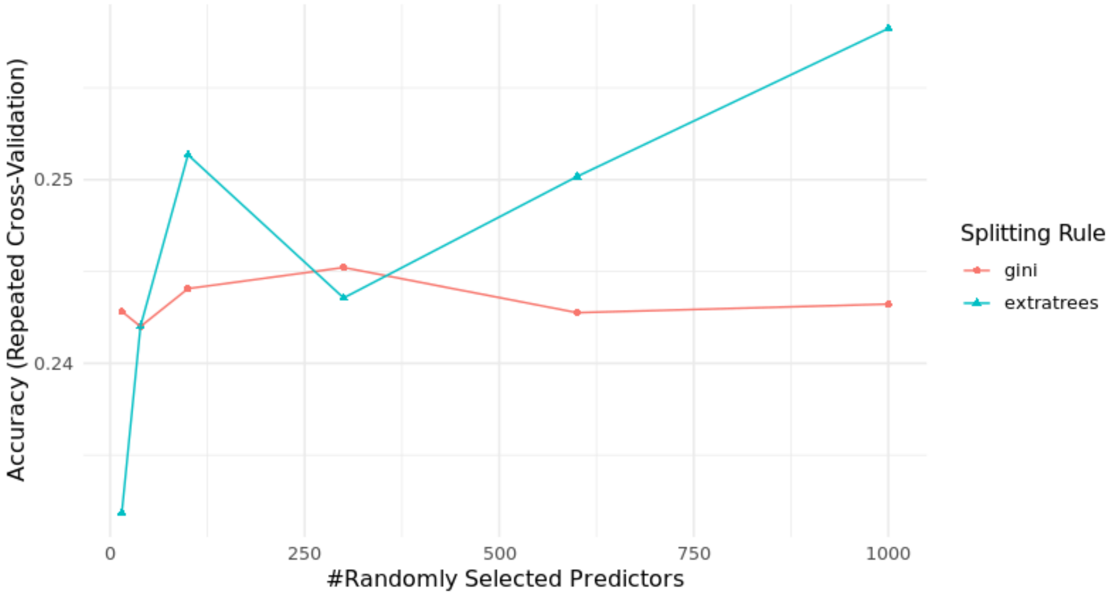




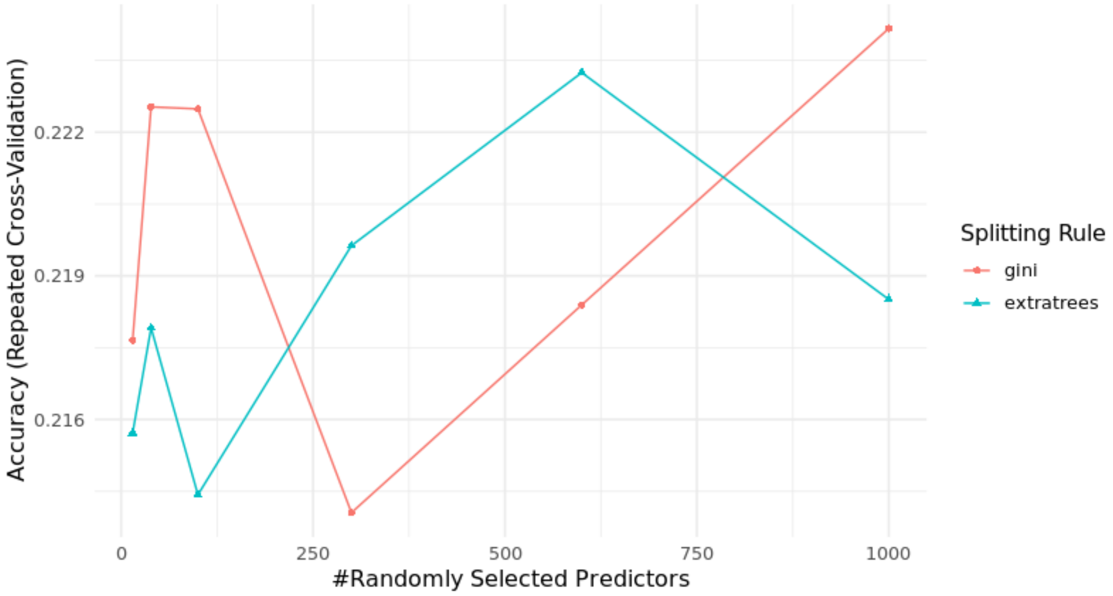

Accuracy (Repeated Cross-Validation)
0.25
0.26
0.27
0
1000
2000
3000
4000
5000
#Randomly Selected Predictors
Splitting Rule
gini
extratrees
Data: Complete, Logged


Accuracy (Repeated Cross-Validation)
0.225
0.230
0.235
0.240
0
1000
2000
3000
4000
5000
#Randomly Selected Predictors
Splitting Rule
gini
extratrees
Data: Complete, Logged Minus Rare Users

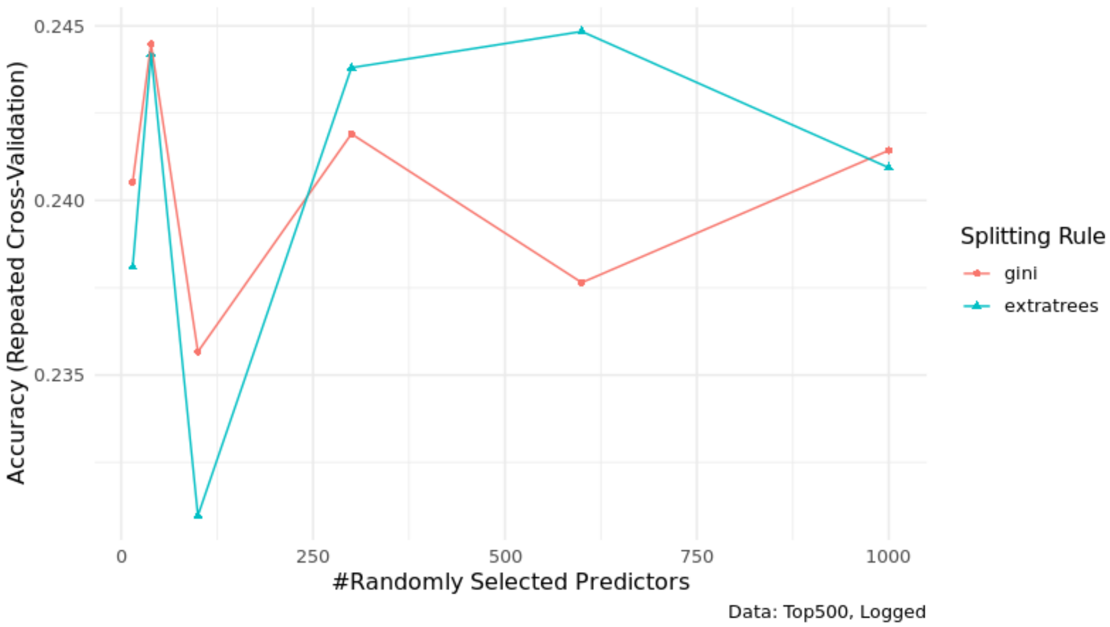
Top500



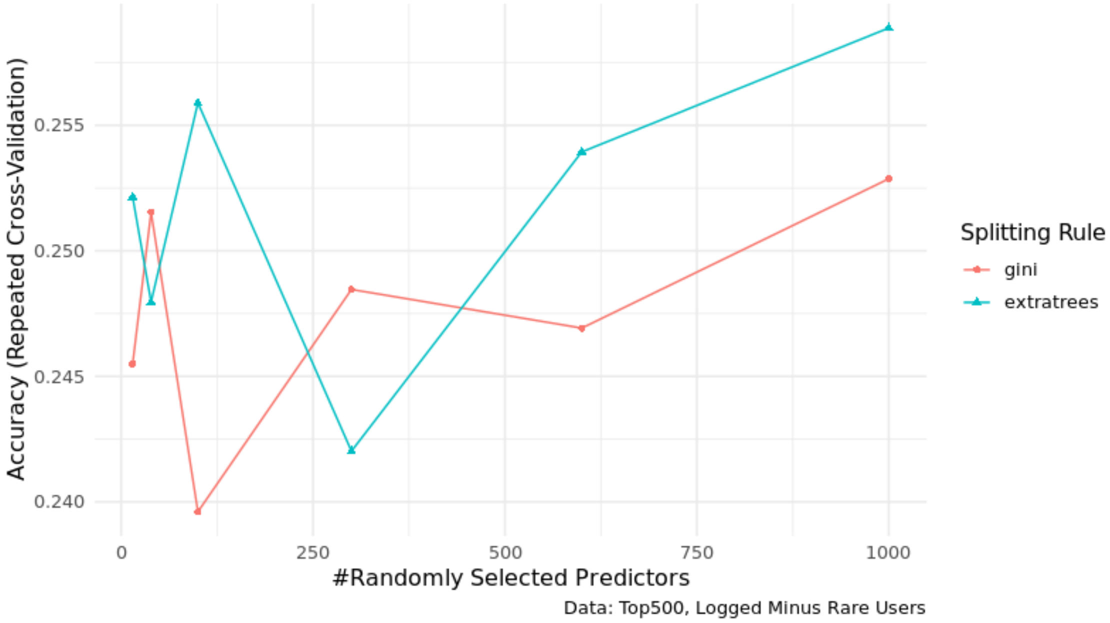
Data: Top500, Unchanged



Data: Top500, Minus Rare Users

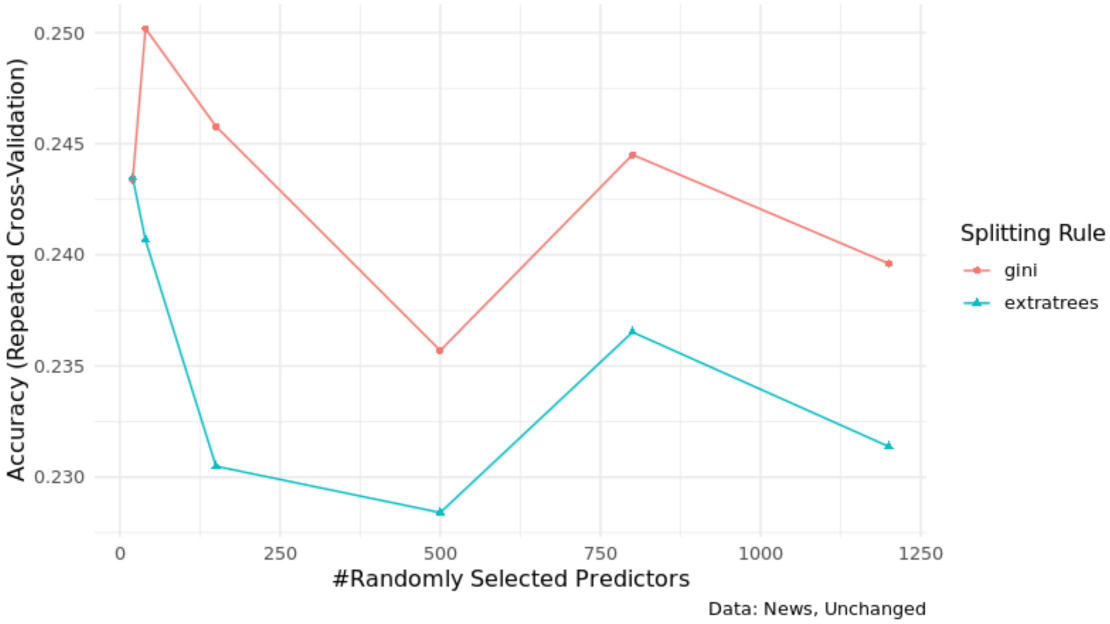


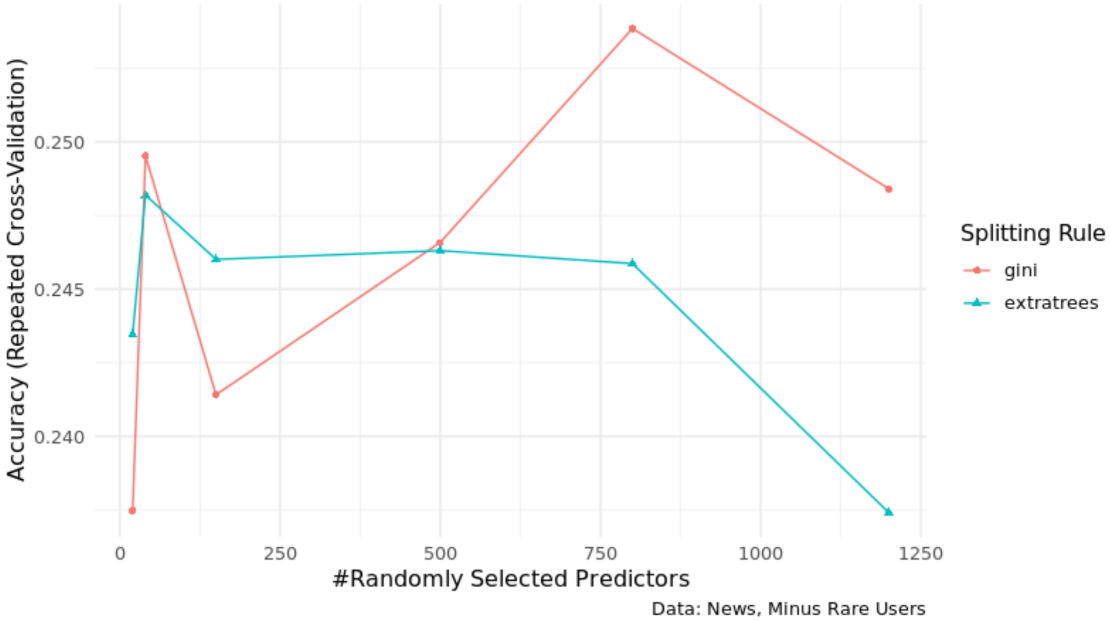
Data: Top500, Logged

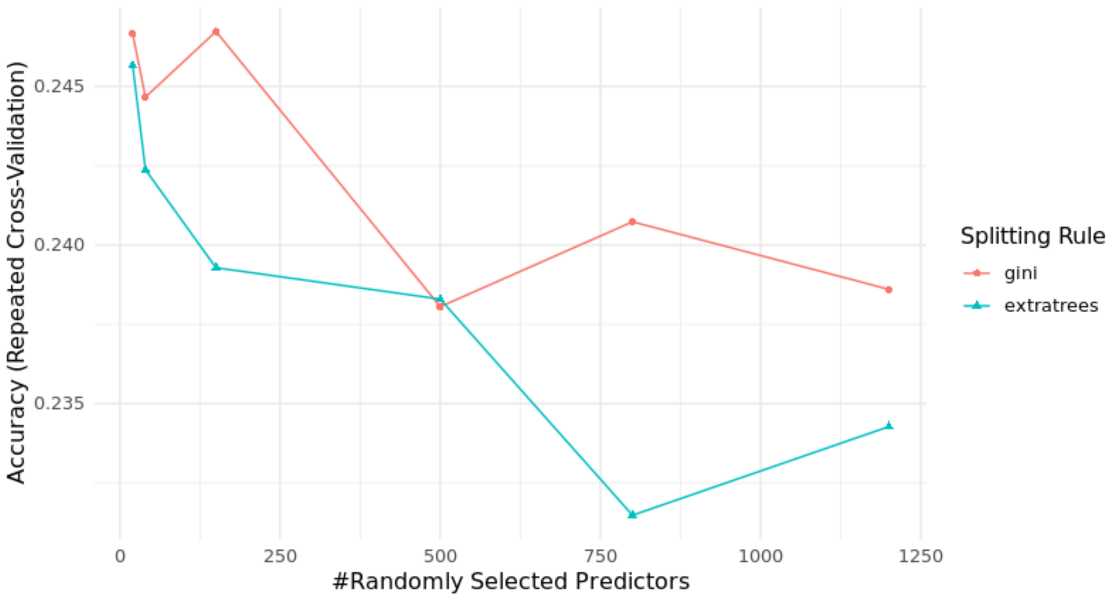


Data: Top500, Logged Minus Rare Users

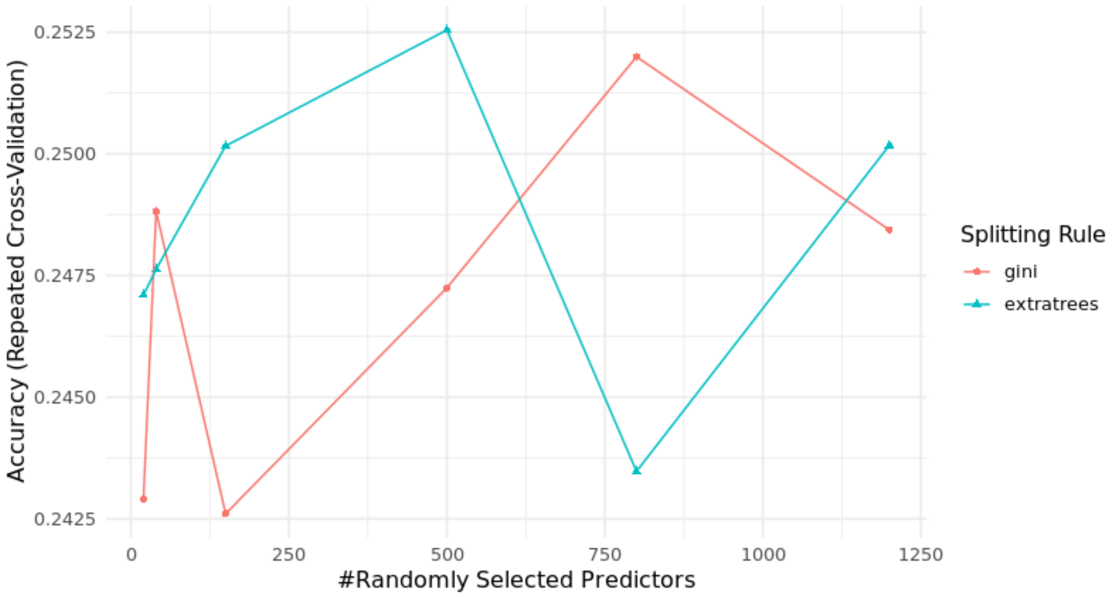
News


0.250
0.245
0.240
0.235
0.230
Accuracy (Repeated Cross-Validation)
0
250
500
750
1000
1250
#Randomly Selected Predictors
Splitting Rule
gini
extratrees
Data: News, Unchanged


0.250
0.245
0.240
Accuracy (Repeated Cross-Validation)
0
250
500
750
1000
1250
#Randomly Selected Predictors
Splitting Rule
gini
extratrees
Data: News, Minus Rare Users

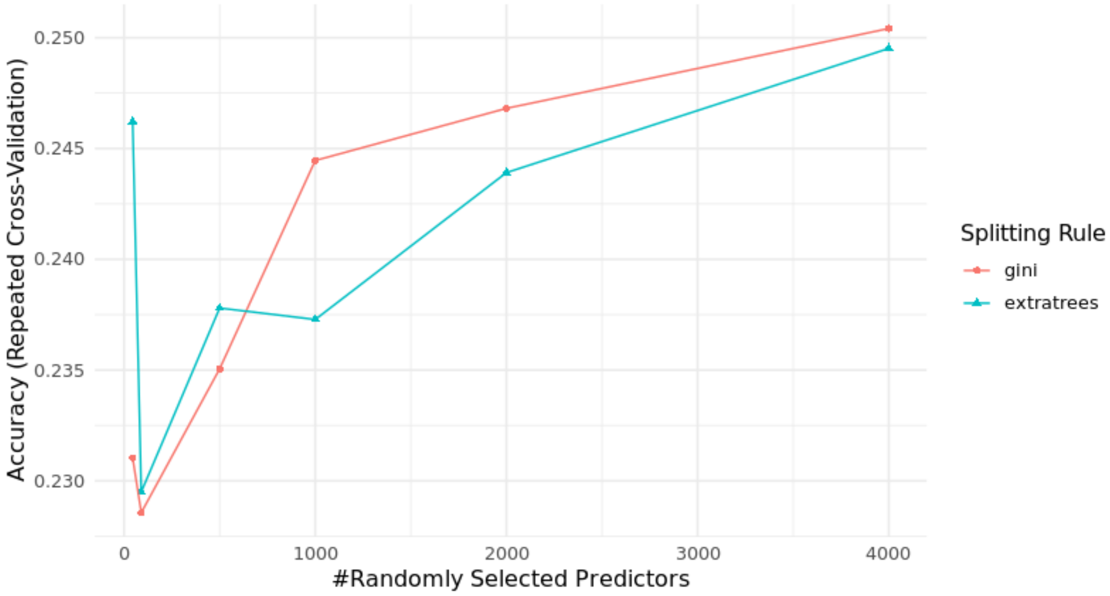


Data: News, Logged

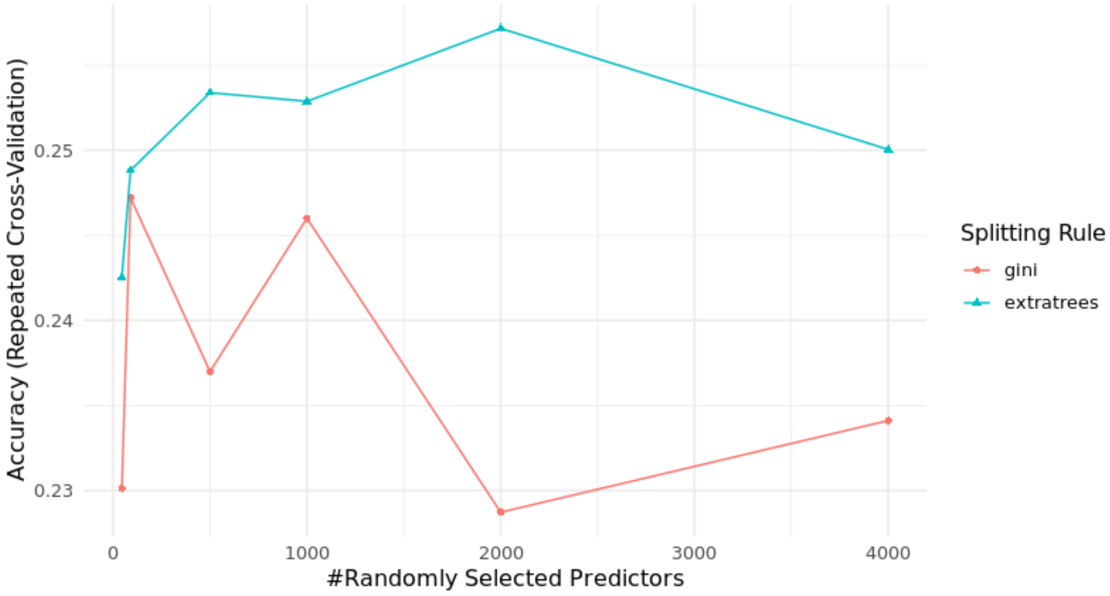


Data: News, Logged Minus Rare Users

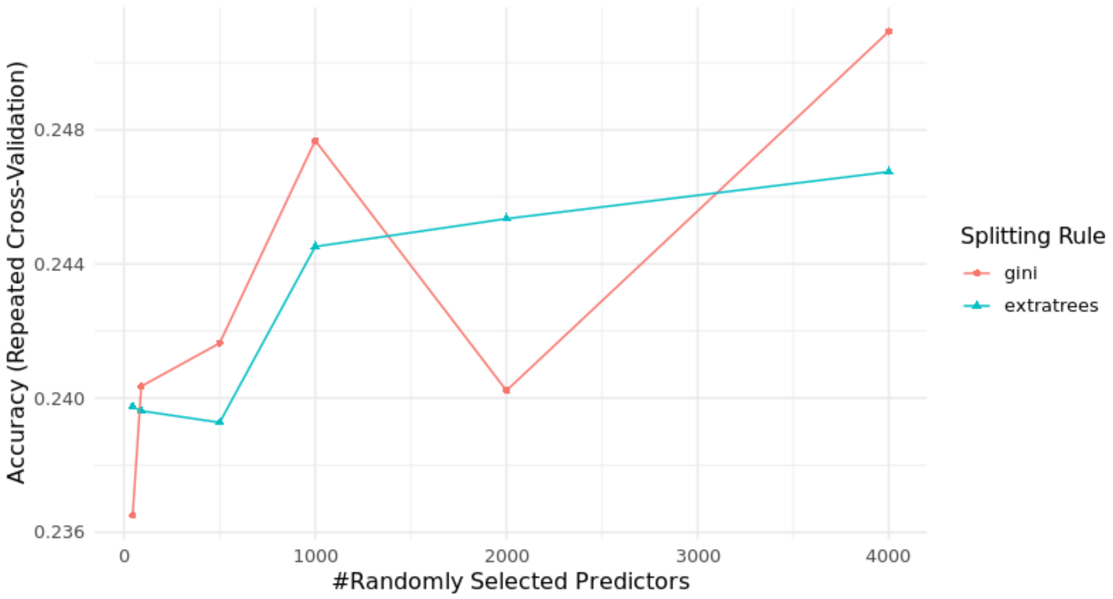
September



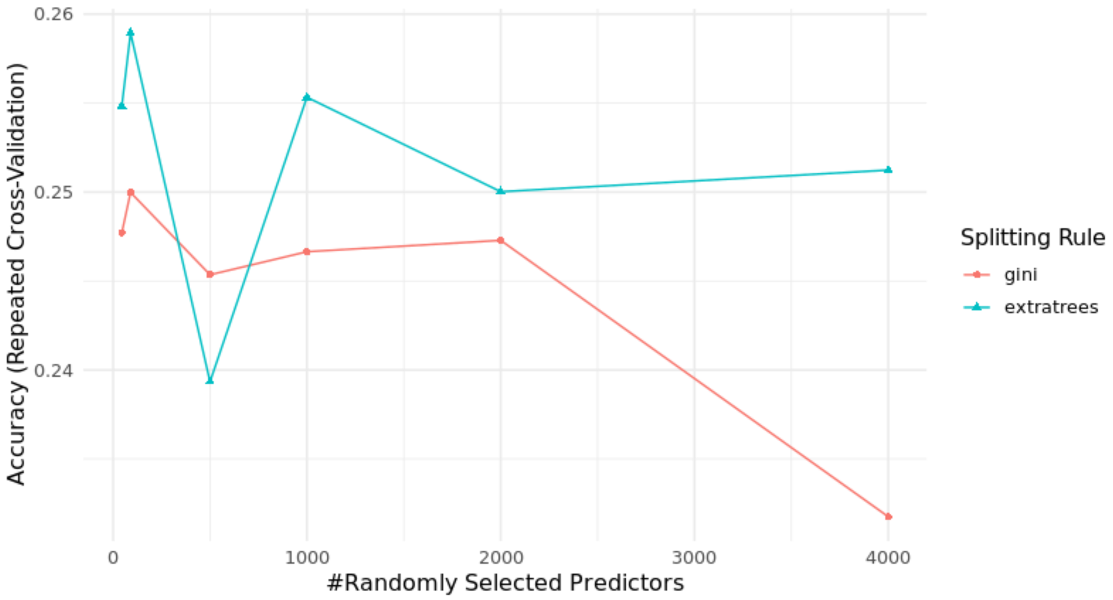
Data: September, Unchanged



Data: September, Minus Rare Users



Data: September, Logged



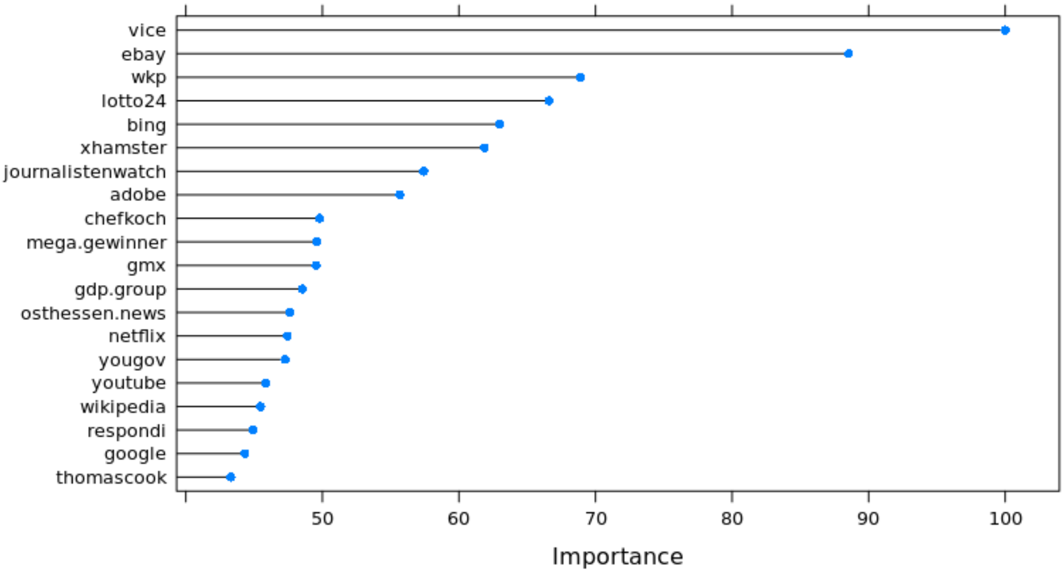
Data: September, Logged Minus Rare Users

Appendix E Random Forest: Variable Importance Plots

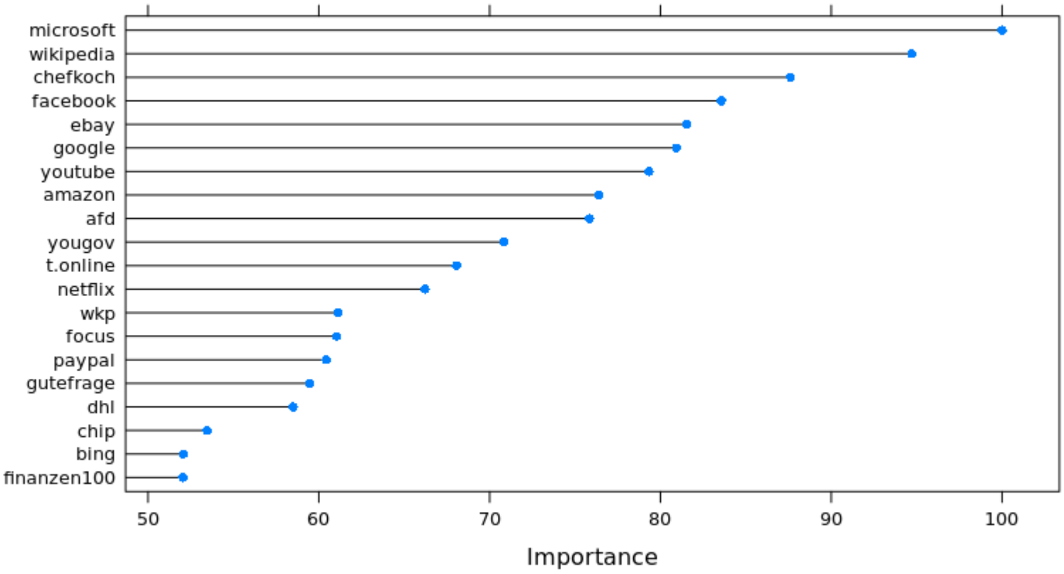
Variable importance plots showing the twenty most important variables in each Random Forest model, by dataset.

Complete

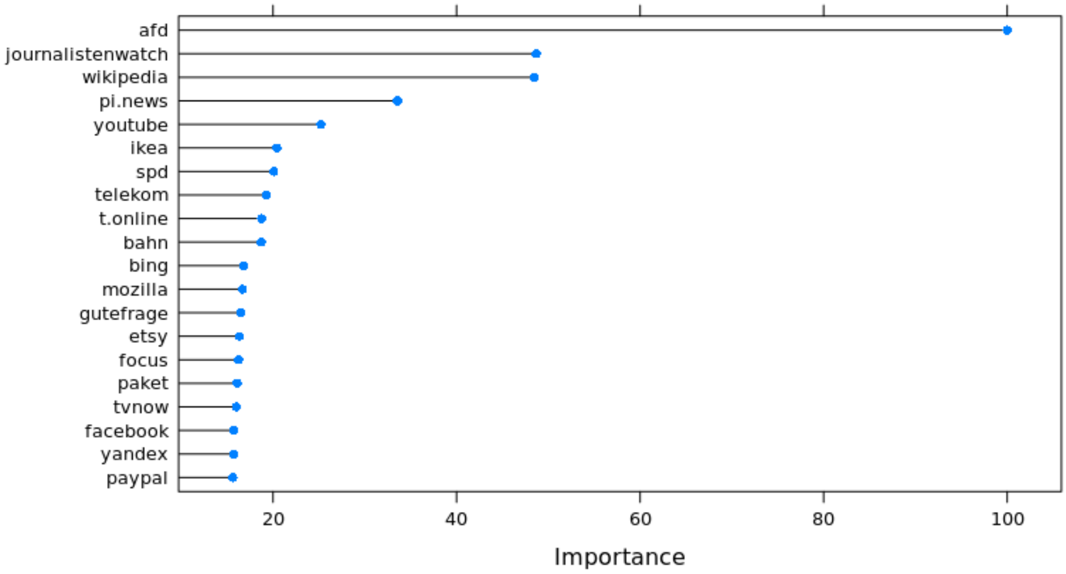
Unchanged



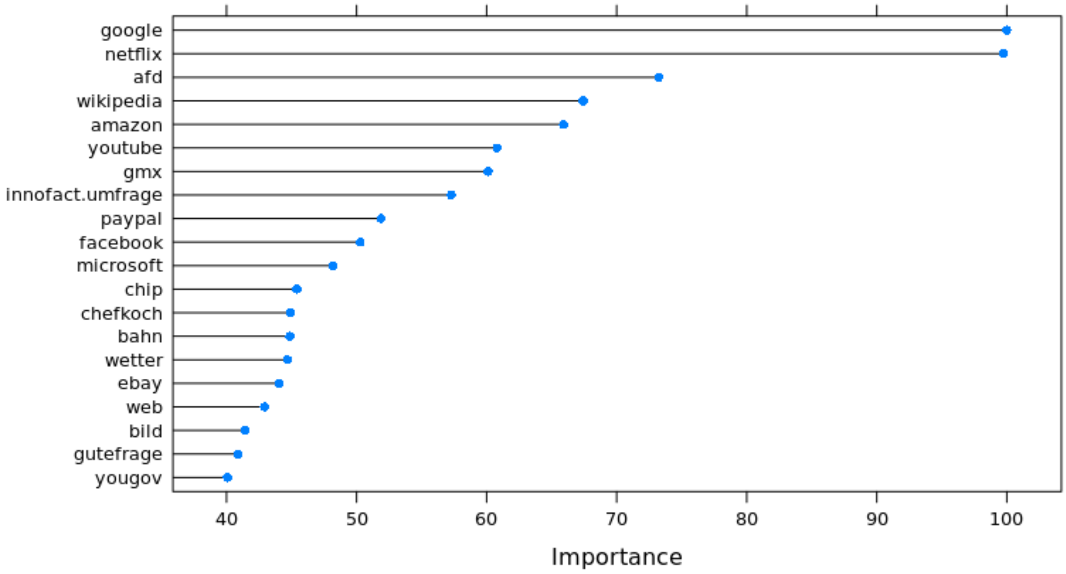
Minus Rare Users



Logged

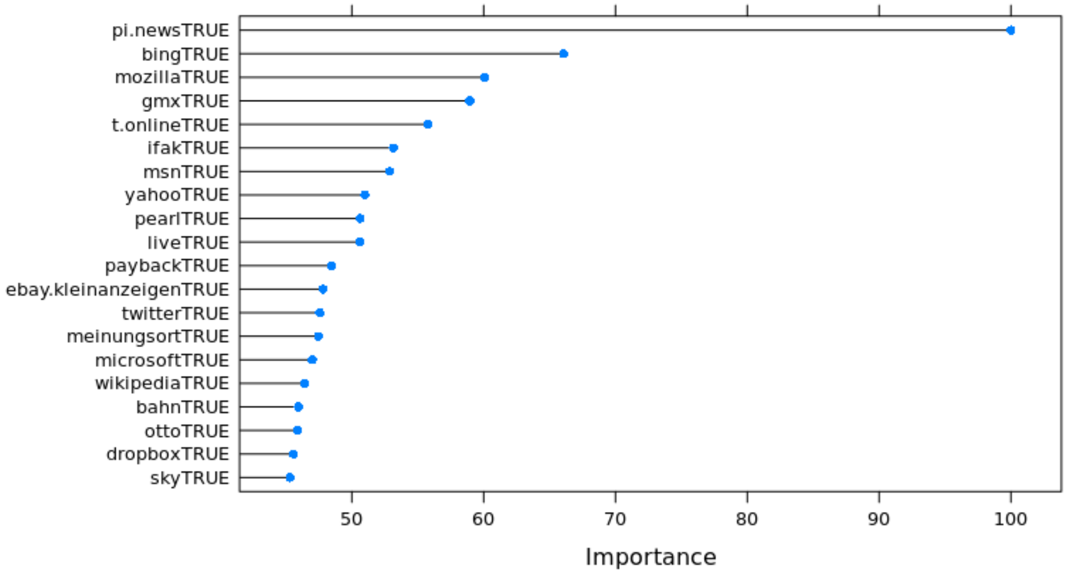


Logged Minus Rare Users

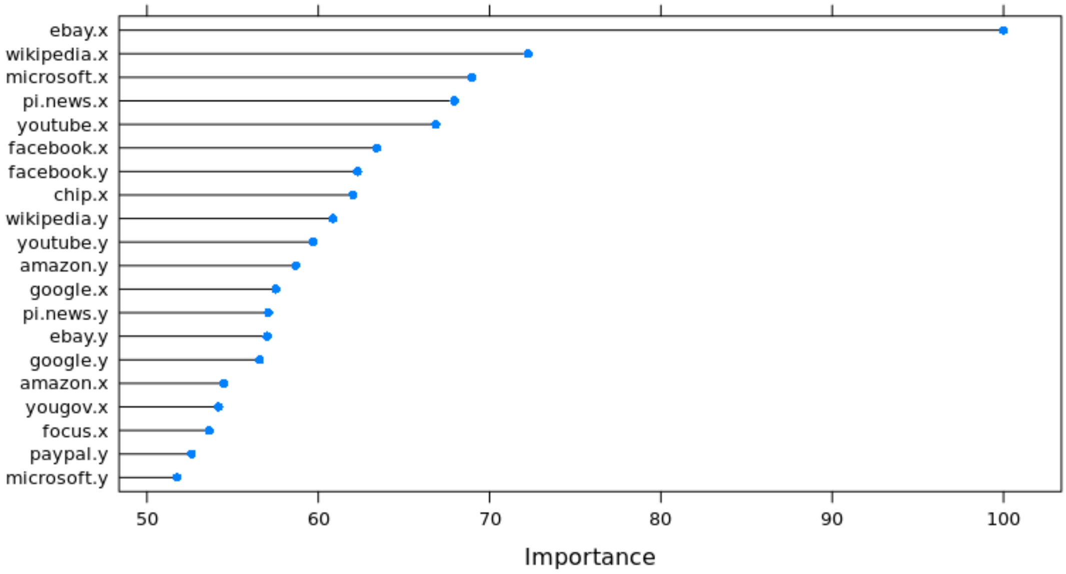


Top500

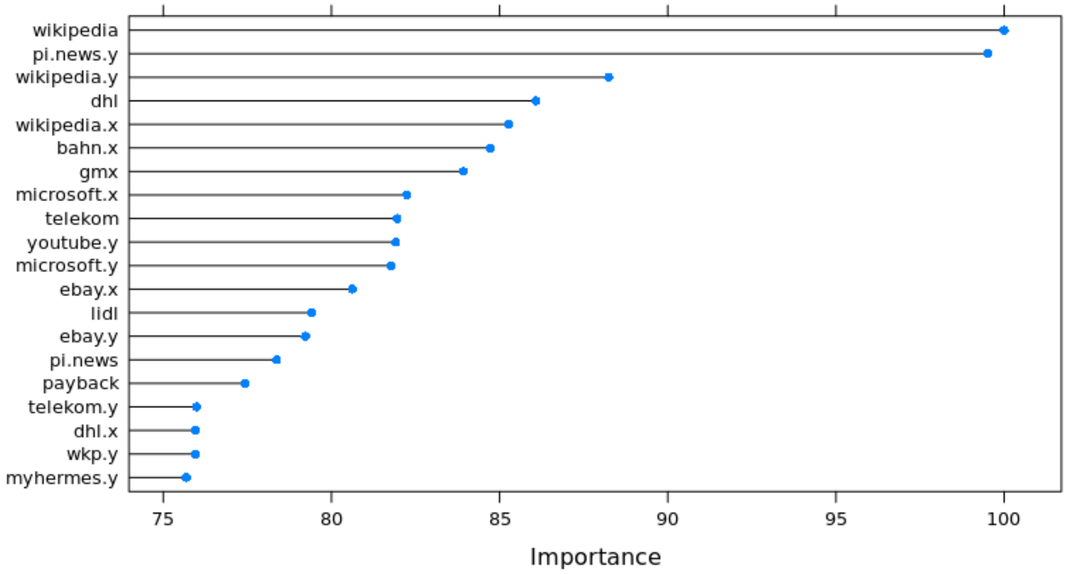
Unchanged



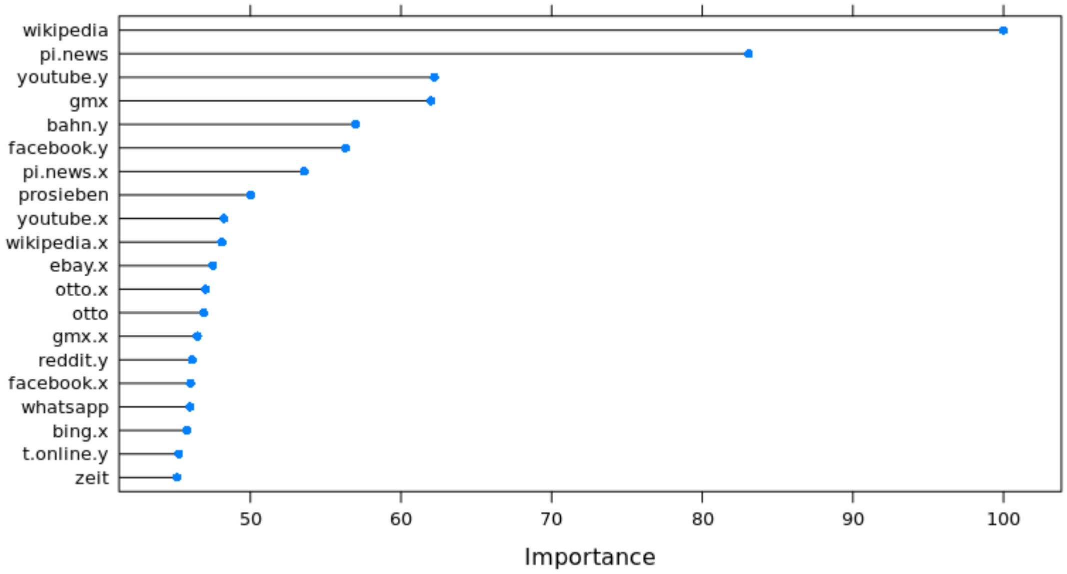
Minus Rare Users



Logged

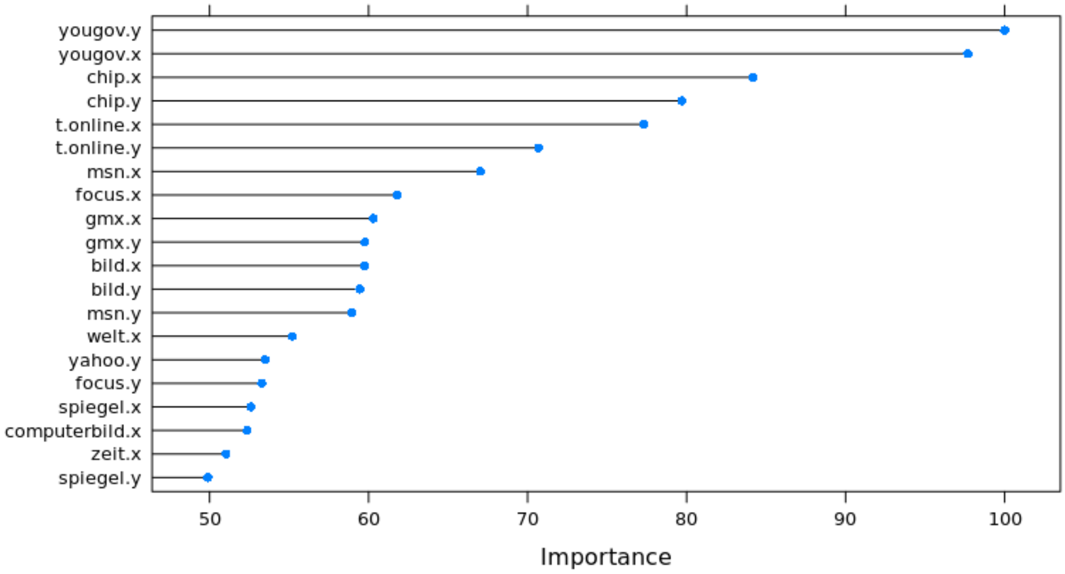


Logged Minus Rare Users

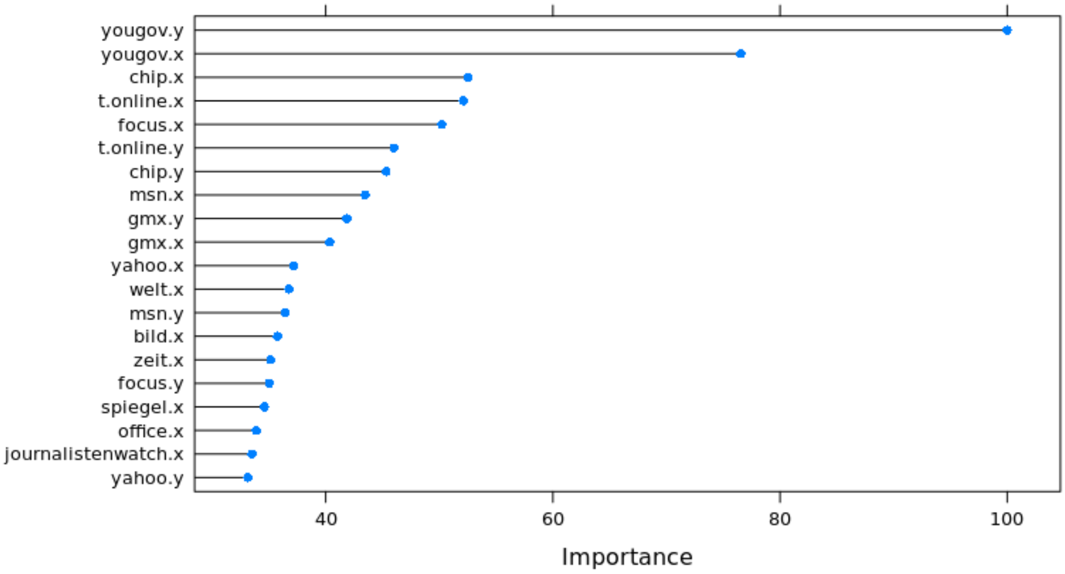


News

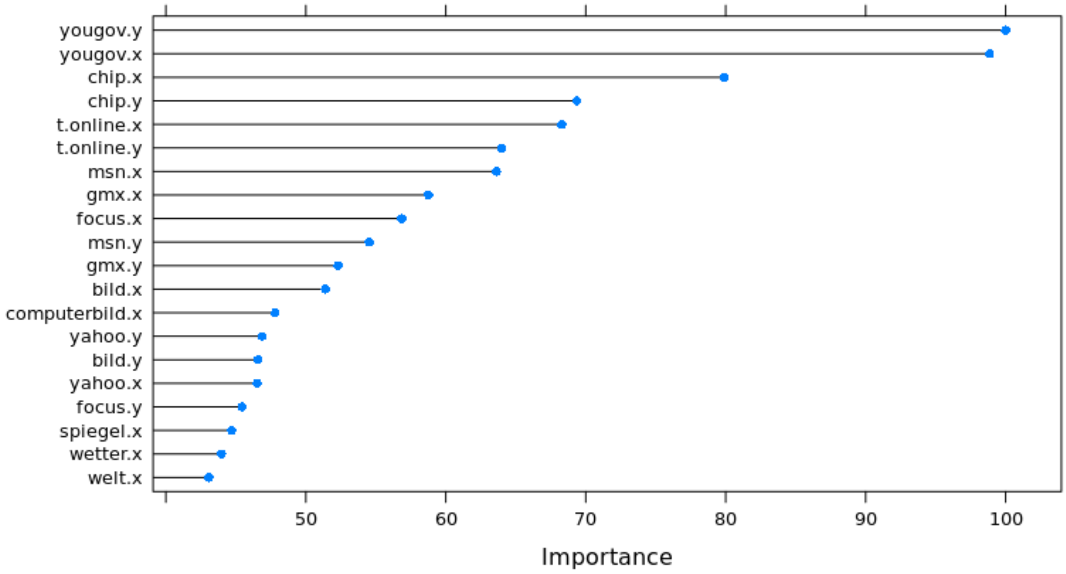
Unchanged



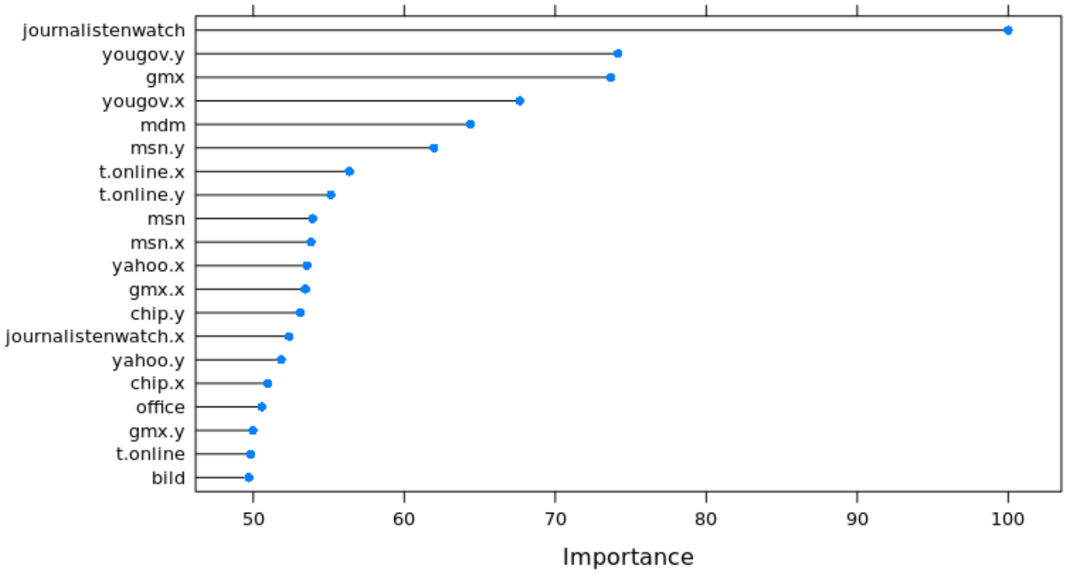
Minus Rare Users



Logged

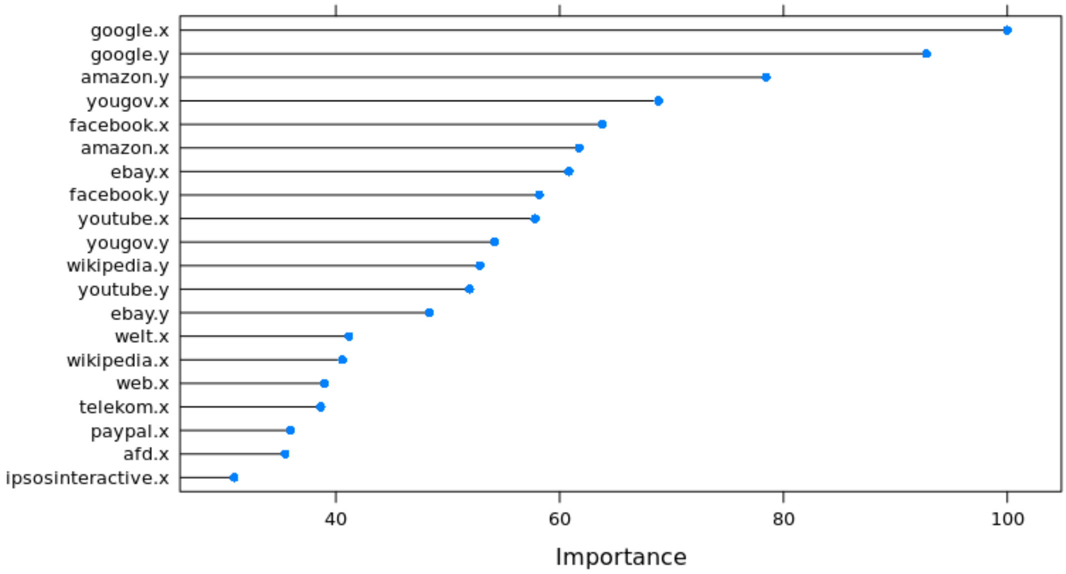


Logged Minus Rare Users

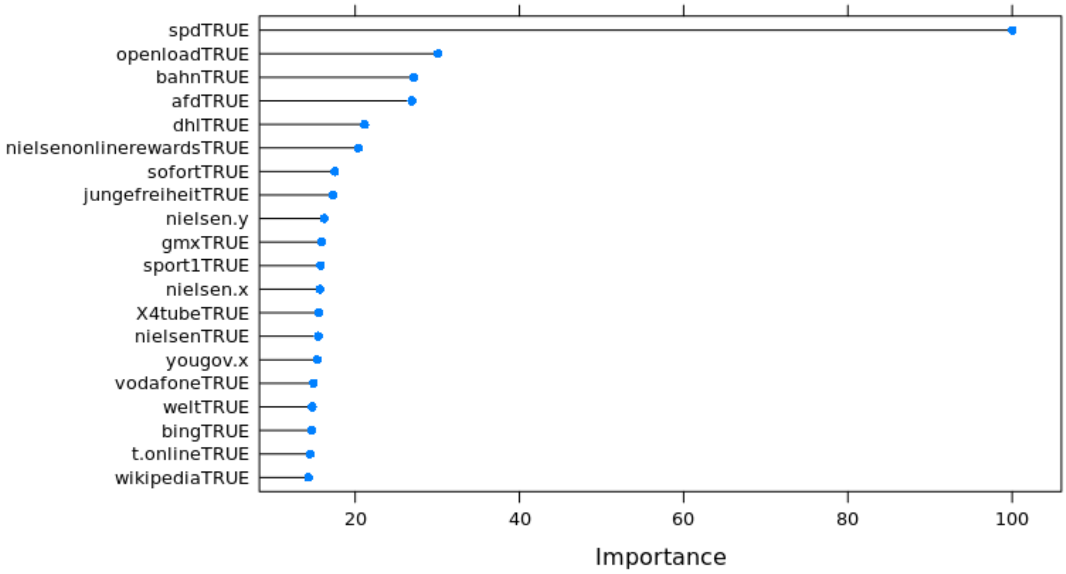


September

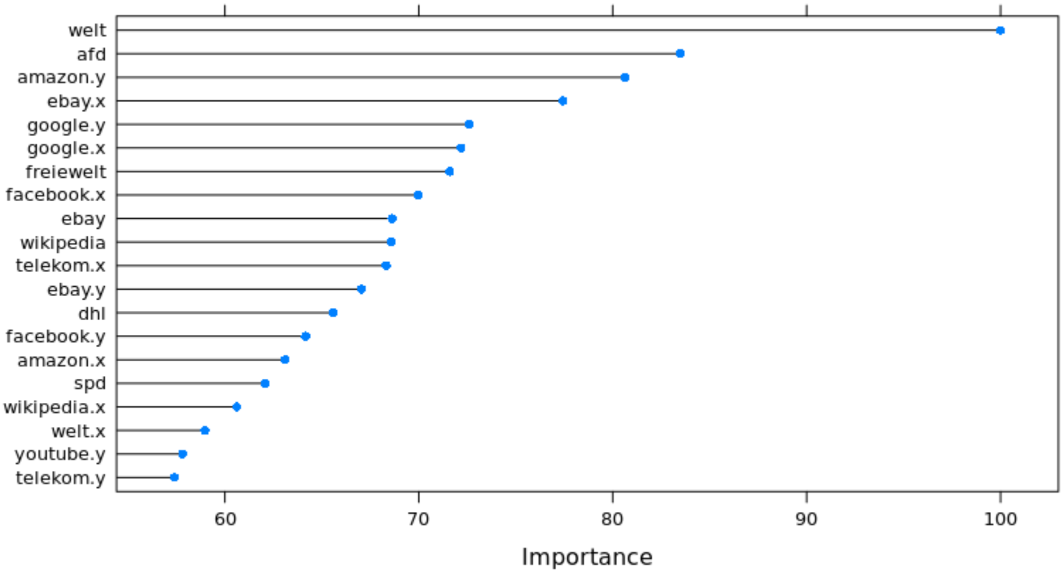
Unchanged



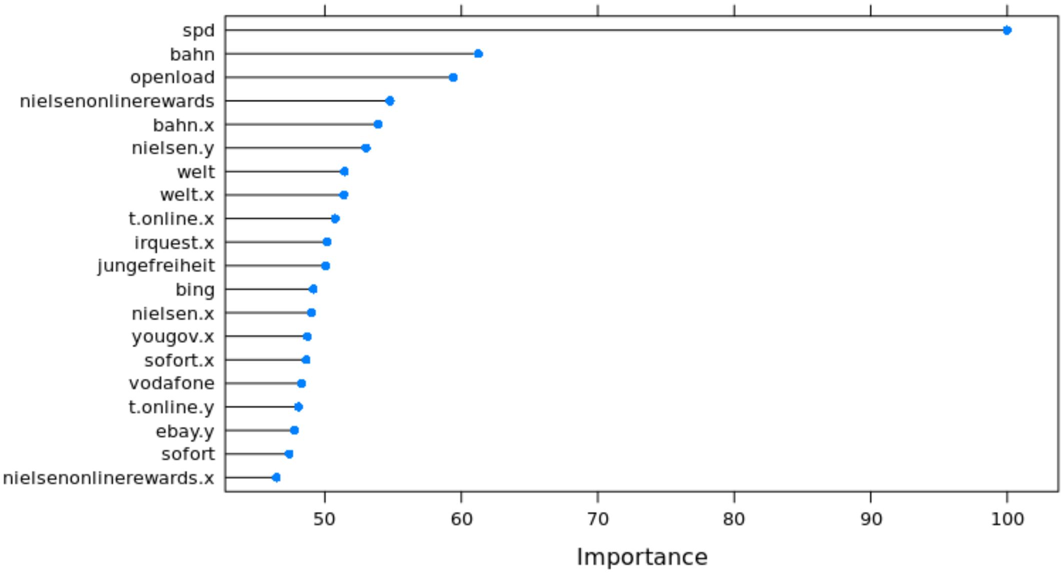
Minus Rare Users



Logged



Logged Minus Rare Users


spd
bahn
openload
nielsenonlinerewards
bahn.x
nielsen.y
welt
welt.x
t.online.x
irquest.x
jungefreiheit
bing
nielsen.x
yougov.x
sofort.x
vodafone
t.online.y
ebay.y
sofort
nielsenonlinerewards.x
Importance