UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

Master's thesis
(Magistrsko delo)

Predicting the gullibility of users from their online behaviour

(Napovedovanje naivnosti uporabnikov na podlagi njihovega obnašanja na spletu)

Ime in priimek: *Mateja Jovanović*
Študijski program: *Podatkovna znanost, 2. stopnja*
Mentor: *izr. prof. dr. Marko Tkalčič*
Somentor: *doc. dr. Vida Groznik*

Koper, avgust 2022

Ključna dokumentacijska informacija

Ime in PRIIMEK: Mateja JOVANOVIĆ
Naslov magistrskega dela: Napovedovanje naivnosti uporabnikov na podlagi njihovega obnašanja na spletu

Kraj: Koper
Leto: 2022

Število listov: 89 Število slik: 24 Število tabel: 13
Število prilog: 4 Število strani prilog: 25
Število referenc: 27
Mentor: izr. prof. dr. Marko Tkalčič

Somentor: doc. dr. Vida Groznik

UDK: 316.472.4:004.738.5(043.2)
Ključne besede: naivnost, strojno učenje, twitter, napovedno modeliranje
Izvleček:

Namen te raziskave je bil raziskati, ali je naivnost uporabnikov na družbenih medijev mogoče predvideti na podlagi njihovega vedenja na spletu. Pregledali smo sorodno delo v psihologiji in našli lestvico naivnosti, ki smo jo kasneje uporabili v našem eksperimentu. Raziskavo smo izvedli na 159 uporabnikih Twitterja (81F, 72M, 6O), pri čemer je večina udeležencev (103) spadala v starostno skupino od 21 do 40 let. Anketa je vsebovala vprašanja o lahkovernosti, občutku lastne vrednosti, čustvenosti, kognitivni refleksiji, finančne pismenosti, zaupanju in demografskih podatkih. Iz lestvice naivnosti smo oblikovali oceno naivnosti, ki smo jo poskušali napovedati z modeli strojnega učenja. Podatke o družbenih medijih smo zbrali prek vmesnika Twitter API. Podatke smo dodatno očistili in obdelali s standardnimi tehnikami NLP (tokenizacija, odstranjevanje stopic in lematizacija). Pri oblikovanju funkcij smo uporabili programa LIWC in fastText. Podatke smo razdelili s tehniko gnezdenega navzkrižnega preverjanja in se problema lotili z uporabo klasifikacijskih in regresijskih modelov. Obe skupini modelov sta dosegli boljše rezultate od ustreznih osnovnih modelov. Naši rezultati kažejo, da je naivnost mogoče napovedati na podlagi spletnega vedenja in da ima prihodnje delo na tem področju velik potencial.

Key document information

Name and SURNAME: Mateja JOVANOVIĆ
Title of the thesis: Predicting the gullibility of users from their online behaviour

Place: Koper
Year: 2022

Number of pages: 89 Number of figures: 24 Number of tables: 13
Number of appendices: 4 Number of appendix pages:25
Number of references:27
Mentor: Assoc. Prof. Marko Tkalčič, PhD

Co-Mentor: Assist. Prof. Vida Groznik, PhD

UDC: 316.472.4:004.738.5(043.2)
Keywords: gullibility, machine learning, twitter, predictive modeling

Abstract:

The aim of this research was to explore if the gullibility of social media users is predictable from their online behavior. We reviewed the related work in psychology and found a self-report gullibility scale that we later used in our experiment. We conducted a survey on 159 Twitter users (81F, 72M, 6O) with the majority of participants (103) belonging to the age group of 21-40 years. The survey contained questions about gullibility, sense of self, emotionality, financial literacy, cognitive reflection, trust and demographics. From the gullibility scale, we created a gullibility score, which we tried to predict using machine learning models. We collected the social media data through the Twitter API. The data was further cleaned and processed using standard NLP techniques (tokenization, stop word removal and lemmatization). For the feature engineering process, we used LIWC and fastText. We have split the data using a nested cross-validation technique and approached the problem using both classification and regression models. Both groups of models achieved better results than the respective baseline models. Our results indicate that gullibility can be predicted from online behavior, and that future work in this field has great potential.

List of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

e.g.	for example
etc.	and the rest
et al.	and others
AUC	area under the receiver operating characteristic curve
CSS	computational social science
ID	intellectual disability
KNN	k nearest neighbors
MAE	mean absolute error
MLP	multi-layer perceptron
NLP	natural language processing
PLS	partial least squares
RMSE	root mean square error
SGD	stochastic gradient descent
SVC	support vector classifier
SVR	support vector regressor

Acknowledgments

I would like to thank my supervisor Assoc. Prof. Marko Tkalčič for giving me the opportunity to work with him on this exciting topic, for being patient with me, encouraging me, and giving me advice along the way. He managed to transfer the love for this craft to me, and I think that it is priceless. I truly believe that I could not pick a better mentor than him!

I would like to also thank my co-supervisor Assist. Prof. Vida Groznik for her guidance and help throughout the process of writing my master thesis. She taught me to pay attention to the details and address critical obstacles in my experiment's pipeline.

I want to express my gratitude to the Faculty of Mathematics, Natural Sciences, and Information Technologies and their amazing staff. To PhD students Jordan Aiko Deja and Nuwan T. Attygalle for helping me in the initial steps of this journey. To my friends and colleagues Lazar Komosar, Filip Golubović, Aleksandar Tadić, Jessica Dartana, Urska Gerič, and Uros Sergaš for giving me support when I needed it the most. To my girlfriend Sladjana Novaković for being there for me through all my ups and downs.

Above all, I want to take this opportunity to express my deepest gratitude to my mother Mirjana Djurić, and my father Slobodan Jovanović. For encouraging me to continue my education in Slovenia and for pushing me forward when I was not able to do it myself. You are my biggest support and I will never have matching words to describe how grateful for having you by my side.

Lastly, I dedicate this thesis to my grandmother Milka and my pet Boža whom I lost during this journey.

1 INTRODUCTION

1.1 MOTIVATION

We live in a world of constantly changing environment with an inflow of new information being greater ever before. Filtering out what is true, useful and important has never been of such importance as it is now. Society is facing numerous problems such as spread of misinformation, fake news, media manipulation, political exploitation, as well as financial and romance scams. Understandably, it is not cost-effective to hire humans to detect and remove all the false information that is circulating online. Thus, this creates a higher demand in the development of prevention systems, fact checkers and other machine-driven solutions.

In addition to this, psychologists have tried to investigate this issue by answering the question of which human traits or a combination of traits are the key factors when it comes to people falling for scams and false information [9, 11, 15, 25, 27]. Cambridge dictionary has defined such behaviour as gullible, or precisely the gullibility is “the quality of being easily deceived or tricked, and too willing to believe everything that other people say.”[1]. Similarly, the definition of gullibility found on Wikipedia says that ”gullibility is a failure of social intelligence in which a person is easily tricked or manipulated into an ill-advised course of action”[2]. So far, there have been many different definitions of this personal trait however, all of the authors agree that there is a need for further research on measuring and describing gullibility.

Studies have shown that classes of people that are especially vulnerable to exploitation due to gullibility include children, the elderly, and the developmentally disabled [16]. Besides financial damage, scam victims face other problems such as trust issues and long-term trauma due to being a scam victim [12]. Protective organizations, banks, and insurance companies are constantly trying to inform people about threats on the internet and provide prevention systems to reduce the possibility of scams. Unfortunately, scammers are becoming much more creative and sophisticated with their ideas of tricking people and making a profit. Moreover, compared to the period before the 2016 US presidential elections, there has been an increasing number of fake news. According to Google Trends, people searched for the term “fake news” notably more

[1] https://dictionary.cambridge.org/dictionary/english/gullibility

[2] https://en.wikipedia.org/wiki/Gullibility

often than before the elections[3]. In 2016, the Oxford dictionary had declared that we are living in the "post-truth" age, and that term soon became the word of the year[4].

The impact of fake news and online scams is vast and has the potential to cause significant damage in the future. Unfortunately, catching scammers and removing the source of misinformation is often very challenging. However, scientists are constantly finding new ways to utilize the information that social media users reveal with their profiles daily. To enhance users' experience, social media companies have encouraged various studies regarding their platform and prediction of different human characteristics. So far, they have managed to successfully predict personality, depression, substance abuse, political orientation, and many other things about their users just from their social media traces. Having that in mind, we believe that the gullibility of users can also be predicted in a similar manner using social media. Moreover, we believe that identifying highly gullible individuals and applying the proper treatment to them could help us fight the uprising problems. Therefore, this work aimed to provide a tool for detecting users' gullibility. More specifically, we devised a tool that predicts gullibility from social media behaviour on Twitter.

1.2 STRUCTURE OF THE THESIS

In the second chapter, we talk about gullibility as a psychological trait, how other authors have approached it, and their findings. Besides this, we also discuss a prediction of user characteristics from social media traces, what it is, and why it is interesting for our work. This helps us understand the reasoning behind our approach to the experiment.

The third chapter is dedicated to the experiment's methodology and the pipeline's development. Here we go into the exact details of each step, from pre-study and designing of the survey to the feature engineering and model evaluation.

After this, in chapters four and five, we go through the experiment's results and their interpretation. Then, we discuss the possible limitations of the experiment and what can be improved in further research. Finally, we take a short retrospective of the whole thesis.

[3] https://trends.google.com/trends/explore?date=all&q=fake\%20news

[4] https://languages.oup.com/word-of-the-year/2016/

2 RELATED WORK

2.1 GULLIBILITY

So far many researchers have tackled the topic of gullibility [11, 15, 17, 19, 22, 25, 27]. It is expected that there are different definitions of this human trait coming from many authors. However, there are more similarities than differences that we will try to cover and clarify in this chapter. Going through the related work of other researchers will help us answer what gullibility is and what we know about it.

Before we go into detail, it is essential to emphasize that in several occurrences, researchers have pointed out that gullibility should not be observed separately as a general and common trait. It is a very delicate topic dependent on the context and other factors, which we will discuss in the further text [11, 15, 17, 25].

2.1.1 Gullibility model

In their research, Greenspan et al. [17] laid down a complex model of five main factors contributing to the gullible outcome. It gives a detailed explanation and shows how multifaceted this topic is. However, it should be noted that the authors focused their work on the gullibility of people with intellectual disabilities (ID). Although this model can describe the gullible outcome of any person, regardless of their cognitive capabilities, it is important to understand that people with ID have cognitive impairments that contribute significantly to their gullible actions. Nevertheless, anybody can act gullibly, but the frequency of gullible actions is greater in the people with ID [15]. This model will help us understand why victims' default gullibility level is just one side of the problem, and the exploiter's persuasion, interrogation, and manipulative techniques the other. The model displayed in figure 1 is divided into five main parts, each representing one set of factors: environmental, intellectual, physical, communicative, and motivational. The different combinations of these factors cause gullible or non-gullible action [17].

For the environmental factors, the authors have given an excellent example of a micro-situation where a policeman interrogates the subject. In this setting, two subcomponents either increase or decrease the chances of gullible outcomes. Respectively policeman's good interrogating techniques (increasing) and the presence of the subject's lawyer (decreasing).

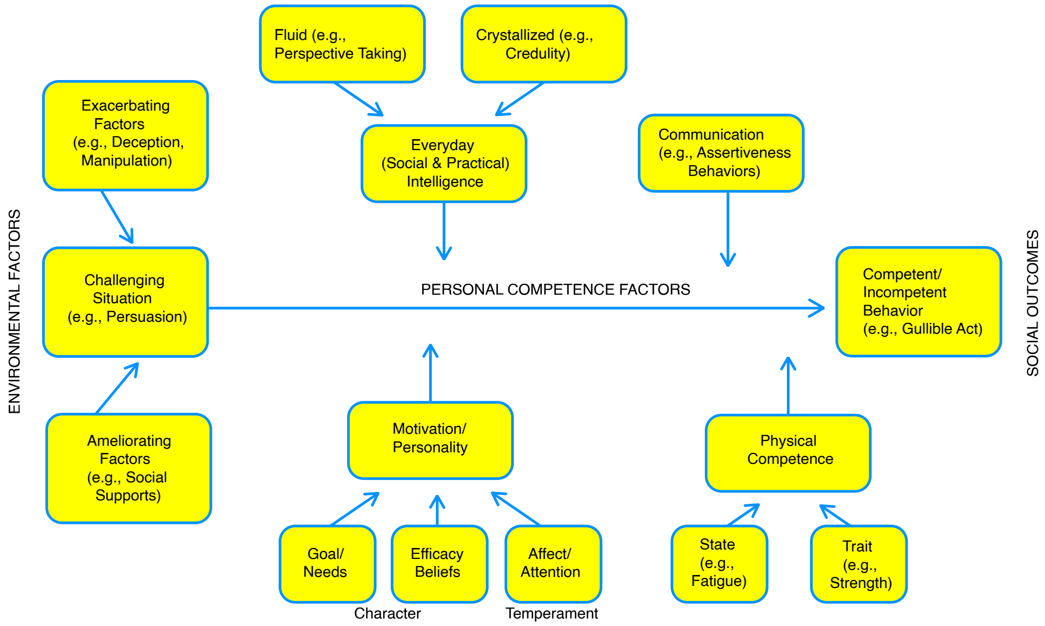

Fluid (e.g., Perspective Taking)
Crystallized (e.g., Credulity)
Exacerbating Factors (e.g., Deception, Manipulation)
Everyday (Social & Practical) Intelligence
Communication (e.g., Assertiveness Behaviors)
ENVIRONMENTAL FACTORS
Challenging Situation (e.g., Persuasion)
PERSONAL COMPETENCE FACTORS
Competent/ Incompetent Behavior (e.g., Gullible Act)
SOCIAL OUTCOMES
Ameliorating Factors (e.g., Social Supports)
Motivation/ Personality
Physical Competence
Goal/ Needs
Efficacy Beliefs
Affect/ Attention
Character
Temperament
State (e.g., Fatigue)
Trait (e.g., Strength)

Figure 1: Gullibility model proposed by Greenspan et al. [17], showing different factors that contribute to the gullible act

The following set of factors, the intellectual, are more complex and demand a greater explanation. It is not unusual that people confuse credulity, a cognitive tendency that contributes to the gullible outcome, and gullibility [17]. Credulity is described as a tendency to believe the unlikely propositions without having supporting evidence for them or, in short, a tendency to believe unbelievable [15]. As it can be seen from the definitions and the illustration of the model in Fig. 1, credulity plays a vital role in the creation process of the gullible action, but it is not comparable to gullibility. Gullibility requires an action that has a cause-effect relationship to the credulous belief [15]. Practically speaking, believing the unbelievable would be credulous, but falling for a financial scam because of that credulous belief would be gullibility.

In the model, the authors treated credulity as an instance of crystal intelligence and part of everyday intelligence. Besides crystal intelligence, there is complementary fluid intelligence. To understand the difference between the two, we should know how they are based. Crystal intelligence is experience-based, whereas fluid intelligence could be seen as a heuristic mechanism we use to cope with new social situations. Other synonyms for crystal and fluid intelligence, used by Greenspan et al. [17] are respectfully social insight and social sensitivity. Someone with high social sensitivity should be able to accurately identify the meaning of a personal state or situation. One example of the social sensitivity task is perspective-taking. In this process, people with ID have significant limitations and poorer performance, leading to difficulty recognizing manipulative intentions in social situations [17].

The third and fourth groups of factors are physical and communicative types. For example, hearing loss or poor eyesight could be a physical impairment contributing to the gullible outcome. Similarly, communication can also be an impairment (although not physical). To elaborate, people who practice their social communicative skills and have developed a strong repertoire and fluency will have an advantage when dealing with scammers or interrogators. On the contrary, if a person lacks these social skills (also referred to as soft skills), he or she is more likely to give up and easily comply instead of resisting and diverting. In people with ID, this is often linked to the feeling of hopelessness.

The last group of factors in this model are motivational factors. They are further subdivided into goals/needs, efficacy beliefs, and affect/attention. In their work, Greenspan et al. [17] said that people with ID tend to seek help and guidance from others, especially in social situations. They also model other people's behavior in social interactions since they have limited experiences in this area due to their condition [17].

While explaining the subsidiary intended for goals/needs Greenspan et al. [17] pointed out that when we are found in a particular social situation, we tend to give it some label. This helps us bring up the appropriate cooperation response (behavior) to the given situation. However, people with ID are known to have trouble with correctly labeling dangerous or coercive situations where they can be exploited. Because of this mismatch in labeling, they often respond in a gullible manner. The cause of this wrong situation labeling is often linked to the victim's seeking for social approval [17].

Regarding efficacy beliefs or self-efficacy, it is known that people with ID think that they cannot influence others; in other words, they have low social self-efficacy beliefs. This is why they often take the agreeable (submissive) approach in social situations and let others (people of influence) have what they want. It is believed that these beliefs contribute to the feeling of hopelessness [17].

The final motivational sub-factors are affect and attention. They are one of the reasons why people in general (not just ones with ID) act gullibly in social situations. We can look at those sub-factors as persistence, control of emotions, and attention while being in coercive or forced situations. However, this is another aspect that exploiters use to lure victims into acting gullibly. They purposely tap some affective schema (vulnerable spot) of the victim, for example, someone's greediness or fear. On top of that, the impaired will was also mentioned as one of the potential contributors [17].

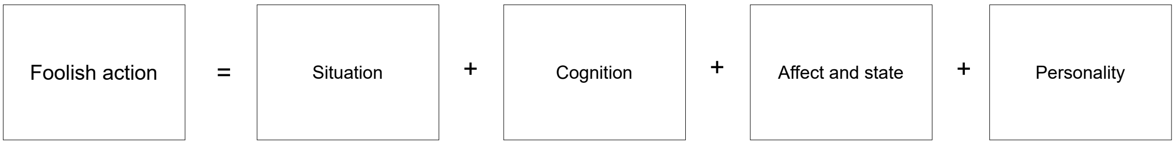


Figure 2: Four factor model of foolish action proposed by Greenspan [15]

2.1.2 Four factor model

Greenspan [15] has later tried to broaden his view on gullibility by expanding it to the concept of foolish behaviors. He made clear that every foolish action can be classified as either practical or social and that both practical and social foolish actions can be subdivided into induced and non-induced. Since he was trying to connect the concept of gullibility to the foolish action, he carefully isolated gullibility from the other foolish behavior by identifying it as a socially induced foolish action [15]. Furthermore, he proposed the four-factor explanatory model of foolish action, concentrating on gullibility as a sub-type of socially foolish behavior. The model is displayed in figure 2.

Looking back at the findings and his previous work on gullibility that we discussed in section 2.1.1 we can draw a parallel between the four-factor model and the gullibility model displayed in figure 1. Four factor model of foolish behavior consists of the following factors: situational (time pressure, social pressure, novelty, and ambiguity), cognition, affect and state, and personality. If we put these two models side by side, we can see that they overlap in many areas, although the four-factor model is much simpler and does not go into detail.

However, in the four-factor model, we notice that novelty and ambiguity are emphasized as separate entities regarding situational factors. We can see that Greenspan [15] puts stress on situational factors as he considers them a compulsory component in the creation of every foolish action. He believes that every foolish act originates from a failure to solve a situational problem. Besides novelty and ambiguity, time and social pressure are other essential parts that make up situational factors. While there is not much to be said about time pressure, we found exciting notes regarding social pressure. People with ID have increased susceptibility to social influence caused by social neediness. In social situations, they tend to search for clues in others and model others' behavior to not appear foolish. This technique is generally effective, but it can be purposely used against the victim [15].

Cognition, as a contributing factor, is described similarly to intellectual factors from the gullibility model. Nevertheless, the author drew attention to the rise in risk outcomes such as sexual and financial exploitation in people with ID. Greenspan [15] suggested that those outcomes result from deficits in practical and social intelligence.

Affect and state were described with more clarity and detail, giving us a better understanding of the relationship between the two. It is mentioned that affective disbalance engages the subject into acting foolishly but state disbalance blocks him from course correcting from a foolish action. Additionally, it is said that adults with ID are generally (as a group) more emotionally reactive, which intuitively raises the chances of acting foolishly if the emotion is activated [15].

The last contributing factor in this model is personality; it refers to human needs, traits, and tendencies that describe one individual. Two subsidiaries of personality are character (degree of one's moral strength) and temperament (manner of reacting to stimuli). Regarding foolish actions, character strength is seen as an essential aspect of personality. The main reason is that it shows how much a person can maintain moral autonomy while facing temptations that lead to foolish (gullible) acts. This comes in hand with another personality facet, and that is willpower. It is argued that willpower is tied to emotional intelligence, implying that people with weak willpower rely on a hot emotional system while making decisions instead of a cold cognitive system. We could derive from the model that people who are more impulsive, emotionally reactive, have weaker willpower, and have poor self-efficacy beliefs also have higher chances of acting foolishly and gullibly [15].

2.1.3 Gullibility and trust

There is a popular narrative that makes people confuse gullibility with default (general) trust. Meaning if someone is by default a trusting person, that automatically means that they are naive, easily fooled, or simply gullible. This might sound intuitive, but research suggests that it is quite the opposite [22, 27].

In their research, Yamagishi et al. [27] investigated this popular belief about default trust, credulousness, gullibility, and social intelligence. They defined general trust as the default expectation of other people's trustworthiness and have further clarified that it has nothing to do with credulousness. According to them, the difference between general trust and credulousness is in the presence of supportive information about someone. If such information does not exist and person A trusts person B, we say that person A has a high default trust. But suppose the supportive information is present, for example, and it carries the sentiment that person B should not be trusted. Then, if person A ignores it and does not affect his trust towards person B, it is considered credulousness [27]. Because of this confusion, people associate a generally trusting person with a credulous and gullible one. However, default trust and credulousness are not the same nor positively correlated. This has been proven in earlier studies done by Rotter [22]. Moreover, in their study Yamagishi et al. [27] tried to prove that high trusters are in fact less gullible than low trusters. They constructed a set of experiments, one of which investigated the difference in sensitivity to additional information about the subject's trustworthiness between high and low trusters.

Before the experiment, participants were given a survey examining their initial trust levels. They were classified into two groups, high and low trusters, and were given 15 scenarios each. Every scenario could come with one of the following information conditions: one positive information, two positive information, neutral, one negative information, and two negative information. Everything was assigned randomly and without any specific order. Results showed that both high and low trusters decreased the likelihood of the subject being trustworthy when negative information was present. However, high trusters were more vigilant since they were faster in changing their opinion once the negative information about the subject was provided [27]. We will not explain other experiments in this research, but we will note that they all had a similar conclusion; high trusters are more vigilant, sensitive to new information, and less gullible than low trusters. Yamagishi et al. [27] explained this paradox, saying that social intelligence is the accounting factor for such results. They mentioned that general trust is supported by social intelligence, and individuals with high social intelligence are better at understanding their own and other people's internal states, which they use in social relations. This advantage lets people with high social intelligence maintain their high levels of trust, although making others with low social intelligence rely on their mistrust. Besides this Yamagishi et al. [27] conclude that high trusters, who have a higher level of social intelligence, are less gullible than low trusters per social interaction. But, the total number of situations in which high trusters acted gullibly is higher than with low trusters. The reasoning is simple; high trusters are more willing to enter into risky, or high-risk, high-reward, social interaction. This could explain the belief that high trusters are more gullible [27].

Interestingly results from two research Yamagishi et al. [27] and Greenspan [15] seem to share the view in one aspect which might not be so straightforward to notice. In their research, Yamagishi et al. [27] argue that gullibility is not related to the high initial trust of the participants and show that it is quite the opposite. High trusters had been better at detecting untrustworthiness cues and more sensitive to new supporting information about the subject. They explained this by saying that high trusters have developed the cognitive ability to detect untrustworthiness cues simply because they were much more exposed to the situations where they had used this skill [27]. Similarly, Greenspan [15] said that people with ID belong to the group of people with higher gullibility levels and that this group has trouble with perspective-taking (fluid intelligence) as well as some other experience-based (crystal intelligence) social tasks due to the limited opportunities for a broad social experience. For example, people with ID are credulous and gullible because they lack the practical knowledge to recognize a lie. That is understandable since they haven't dealt with many liars in their life because society is usually trying to protect this group of people. Both researches found that the lack of experience in social situations is a big reason behind gullible actions [17, 27].

2.1.4 Measuring gullibility

Until recently, there have not been many attempts to make such a scale for measuring gullibility. But, the fast-paced environment of constantly circulating false information has sped up that process. Therefore in one recent study, Teunisse et al. [25] have conducted a series of 5 thorough experiments to address this matter. Through the experiments, researchers tested different questionnaires, and scales, analyzed the accountable factors, and as a result, produced a self-report gullibility scale consisting of 12 questions. Each question from this scale was answered using a 7-point Likert scale ranging from strongly agree to strongly disagree. Teunisse et al. [25] have defined gullibility as an ”individual’s propensity to accept a false premise in the presence of untrustworthiness cues.” According to them the two contributing factors to gullibility are persudability and insensitivity (to untrustworthiness cues) [25]. The scale was validated in one of the experiments, which compared the controlled group with groups of two assumed extremes, members of the skeptics society and scam victims. The results showed that the group of skeptics scored significantly lower than the controlled group on the gullibility scale. On the other hand, scam victims scored significantly higher than the control group, which confirmed the validity of this scale. Moreover, in another experiment, authors found that participants with higher gullibility scores were more likely to report that they would respond to the scam emails. Apart from this, they also found those emails significantly more persuasive than participants with low gullibility scores [25].

The other findings indicated that gullibility is related to low social intelligence, specifically social information processing (e.g., I can predict other people’s behavior). Teunisse et al. [25] have suggested that this is connected to the inability to detect untrustworthiness cues. However, we would like to note that social information processing has another interpretation that we have already seen in 2.1.1, and that is perspective-taking. Additionally, results suggest that gullibility is associated with high agreeableness, high social vulnerability, and a tendency to have paranormal beliefs. However, it is not related to Machiavellianism and interpersonal trust. For future studies, authors have proposed a further validation of the scale besides other variables that could investigate the connection between gullibility and the tendency to fall to fake news (misinformation) or different scams (financial, romance, etc.) [25].

Thankfully, another research done by George et al. [11] used this gullibility scale and made an experiment that behaviourally tests and validates it. In this study, participants were given examples of phishing emails (which they rated) and a survey that included HEXACO personality factors, questions measuring the need for cognition, need for closure, sense of self, and gullibility scale. Six weeks after the experiment, participants received the simulated phishing emails. Results have shown that participants who clicked the link in the simulated email also scored significantly higher on the gullibility scale. Furthermore, gullibility was associated with higher ratings of the example phishing emails (likelihood of responding), higher emotionality, and a weaker sense of self. On the other hand, neither need for cognition nor closure was related to the gullibility [11].

2.2 PREDICTION OF USER CHARACTERISTICS FROM SOCIAL MEDIA TRACES

With machine learning and big data development, new possibilities for other scientific disciplines appeared. In addition, researchers realized the power of the data and algorithms and have started combining them with traditional techniques. This significantly improved many domains, including healthcare, quality assurance, linguistics, the automotive industry, and many more. However, it also created a new branch of social science that is of special interest for our work on gullibility, a computational social science (CSS).

Social surveys were and still are a good tool for measuring different aspects of human nature, but there are certain limitations to it, such as slow and costly data collection, controlled environment, and scalability of the survey. One of the ways that CSS addresses those limitations is through the collection of data from social media. The usual approach to the problem is the combination of observing behavior through social media activity (big data) and asking questions through the survey, as described by Salganik [23]. In his work, he explains why we should not use only big data. He emphasizes that big data has its flaws, and no matter how big the dataset we manage to get, it always draws us to ask more questions [23].

2.2.1 Research question

Much research has been done in taking social media traces of users and predicting various characteristics, such as personality, gender, political orientation, depression, substance use (tobacco, alcohol, drugs), and others [4, 5, 6, 7, 13, 14, 18, 24, 26].

Inspired by the work of other researchers, we constructed an experiment that investigates whether the gullibility of Twitter users can be predicted from their social media traces. Reviewing the literature on gullibility, we acknowledged the lack of research focused on unobtrusively measuring gullibility in an uncontrolled environment. On top of that, we wanted to address the contextual aspect of gullibility by testing if there is a negative correlation between gullibility and financial literacy. The motivation for such a hypothesis comes from an increasing number of financial scam victims believed to be exploited mainly because of their high gullibility.

To achieve our goal, we combined the data from the social survey and big data from the Twitter profiles of the participants. We used Twitter API and natural language processing (NLP) techniques to obtain information from Twitter profiles. Specifically, we used Linguistic Inquiry and Word Count (LIWC) text analysis tool since other authors used it in their studies to get more features from the tweets [5, 13, 21, 26]. Moreover, we wanted to test whether obtaining word vectors using the fastText library could help in our experiment, as it was successfully used for predicting the text's sentiment in the past [2]. For the prediction of participants' gullibility, we used a gullibility scale described by Teunisse et al. [25] and various machine learning models. Because this is one of the first exploratory research on this topic, we decided to approach the problem of predicting gullibility both as a classification and regression problem. To explore what set of features contributes the most when it comes to predicting the gullibility of Twitter users, we tested different combinations of features on each model.

3 METHODOLOGY

The related work shows different facets of gullibility, potential causes of gullible behavior, and its relation to other personality traits. Moreover, we have seen how other researchers predicted different personality traits using data from participants' social media profiles. In this chapter, we will build a pipeline for predicting the gullibility score of Twitter users using their social media activity. In order to devise a method for detecting gullibility from users' social media traces, we used the methodology depicted in Fig 3. We first performed a pre-study to validate the questionnaire, then collected the data in the main study. We then proceeded with data pre-processing and feature engineering; finally, we evaluated the predictive model.

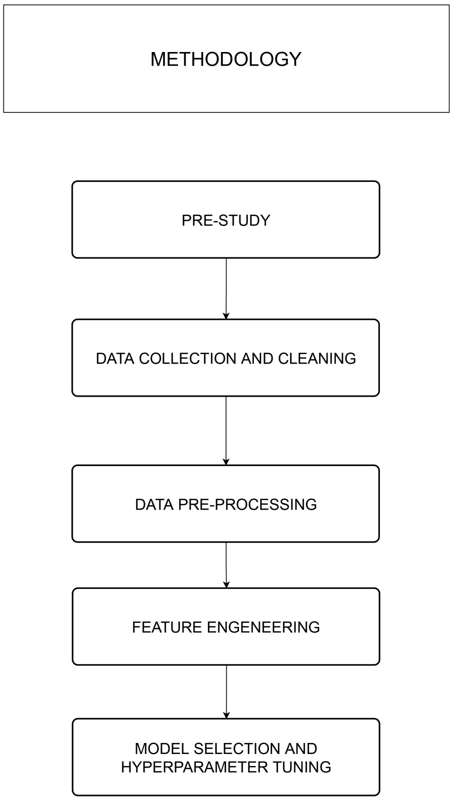

METHODOLOGY
PRE-STUDY
DATA COLLECTION AND CLEANING
DATA PRE-PROCESSING
FEATURE ENGENEERING
MODEL SELECTION AND HYPERPARAMETER TUNING

Figure 3: Methodlogy pipeline

3.1 PRE-STUDY

We have constructed a pre-study as we needed a well-rounded survey that would reproduce the findings mentioned in Chapter 2. such as a high correlation between gullibility and emotionality and a weak sense of self. But also the one that introduces new questions and helps us investigate other aspects of gullibility, such as it being highly contextual and dependent on micro-situation. The goal of the pre-study was to check the feasibility of the survey, get insights on what should be adjusted, and produce an optimal questionnaire as an output. We focused on reducing completion time and maximizing information gain to make the survey optimal. Long completion time causes people to quit the survey, leaving us with a high number of uncompleted questionnaires and poor data quality.

The pre-study survey included the gullibility scale mentioned in 2.1.4 and some other personality-related questions. Knowing that Teunisse et al. [25] and George et al. [11] suggested further validation of the gullibility scale alongside other measures, we added the following set of questions:

- Gullibility scale, 12 items, scale range 1-7 [25],
- Cognitive reflection test, 6 items [10],
- General trust question, 1 item,
- Emotionality questions extracted from HEXACO-60 personality factors, 10 items, scale range 1-5 [1],
- Sense of self scale, 12 items, scale range 1-4 [8].
- Financial literacy questionnaire, 19 items [20],
- Demographic questions, 4 items.

We included questions on emotionality, sense of self, general trust, and cognitive reflection as we wanted to replicate the results from the reviewed studies [11, 25]. In addition, we wanted to test out if the financial literacy questionnaire could reveal any new connection between gullibility. We hypothesize that financial literacy could be negatively related to gullibility, given that individuals with more practical and field-focused knowledge of finances will be less prone to fall for financial scams, the most prominent real-world example of gullible behavior. The financial literacy questionnaire we used in the experiment consists of three groups of questions about financial knowledge, financial skills, financial attitude, and two individual questions regarding subjective financial satisfaction and knowledge. We acknowledge that this will not help us explore general predictors of gullibility; however, it will help bring up more clarity in this specific use case (financial scams), which accounts for a significant share of total gullible outcomes. Lastly, we included the demographic questions related to age, gender, country of residence, and level of education. Alongside these questions, we have put three attention check questions to prevent poor quality answers and speed up the data cleaning process. The survey was hosted on www.1ka.si and all participants were recruited through their personal social networks. Access to the study was granted through the shareable link.

3.1.1 Pre-study results

In Tab. 1, we can see the descriptive statistics of the answers collected in the pre-study survey. In total, 26 people completed the questionnaire, and the average gullibility score was 31.69, with a standard deviation of 11.40. These gullibility scores' results are not surprising because Teunisse et al. [25] reported similar scores, across three different test groups, in their original study. Skeptics had a mean gullibility score of 27.94 and std of 9.07, scam victims 40.57 and std of 13.12, controlled group of psychology students 35.41 and std of 12.62. All three groups combined (615 participants) had a gullibility score of 33.24 and an std of 12.27, which is very close to what we managed to measure in our small sample. However, the completion time was 23 minutes with a standard deviation of 9 minutes. This alone was a good indicator that the questionnaire was too long and needed to be reduced, so we removed questions that did not contain significant information. After the revision survey had 43 instead of the initial 64 questions making it more time-efficient and suitable for the main study; for a detailed view of both the pre-study and main study questionnaire, please check appendix A and B.

Table 1: Summary statistics of the pre-study results

	mean	minimum recorded	maximum recorded	median	std
Gullibility	31.69	12	55	30.5	11.40
General Trust	3.73	1	6	4	1.56
Sense of self	25.80	18	38	25	4.56
Emotionality	36.11	27	52	35.5	6.88
Subjective Financial Knowledge	4.26	1	7	4	1.56
Subjective Financial Satisfaction	4.46	1	7	5	1.60
Financial Attitude	46.07	31	55	46.5	6.85
Financial Knowledge	2.03	0	5	2	1.48
Financial Skills	2.42	0	5	3	1.50
Cognitive reflection test	2.96	0	6	3.5	1.92
Time in minutes	22.73	9	45	21.5	9.15

3.2 MAIN STUDY

The survey in the main study was set up in the same way as the pre-study survey; however, we added one additional question dedicated to the Twitter username, resulting in a total of 43 questions. The main study questionnaire is available in appendix B.

3.2.1 Data collection and cleaning

Before taking the survey, participants gave us consent to use their data. We had let them know the purpose of the study, details on data collection, and the criteria for participation. Participants were informed that their results could be deleted, edited, and retrieved at their request. Furthermore, we assured them that their data would never be reported anywhere in an identifiable form or shared with any third party. Lastly, participants were informed that they were free to quit participating in the study at any moment.

Criteria for eligible participants were the following:

- participant has completed the whole survey
- participant has passed all of the attention check questions
- participant does not possess protected Twitter account
- participant is an active Twitter user (has tweeted more than 20 times in the past 12 months)

After filtering out the participants, only 159 were considered eligible and their data was taken to the further stages of the pipeline. Twitter data was only collected from the eligible participants. For the Twitter data collection, we used Twitter API and Tweepy python library [1] [2].

For each user, we have collected the most recent 3,250 statuses from the profile or all statuses of the user if the total number of statuses was less than 3,250. For each user, we collected a simple set of variables available directly through the Twitter API and Tweepy. These variables were the following: number of followers (people following the user), number of friends (people the user follows), number of likes (total number of given likes), number of statuses (tweets, retweets, replies), status text, location, protected account (boolean), likes received on the status (number), retweets received on the status (number), number of lists, profile's date of creation.

We have also derived the number of mentions, links, hashtags, and retweets from the "status text" since all of those features have unique interpretations in the text.

[1] https://developer.twitter.com/en/products/twitter-api

[2] https://www.tweepy.org/

For example, every retweet starts with "RT". Besides them, we created variables for the number of words, characters, full stops, commas, semi-columns, columns, question marks, exclamation marks, quotes, apostrophes, dashes, brackets ('[]'), parentheses ('()'), braces (''), emojis and scraped statuses. Further in the text, we will refer to this set of features as "basic profile metrics and punctuation."

3.2.2 Data pre-processing and feature engineering

To prepare the collected data for the machine learning models, we had to transform it into a set of meaningful aggregated variables. The first step was to correct all the reversed questions' answers. Answers to the reversed questions were treated as if the scale was turned opposite. For example, if the scale ranges from 1 to 7 and the user answers with 1 we will treat that answer as 7; the same goes for other values, as displayed in Fig. 4. The analogy is the same for other scales that have a different range of answers.

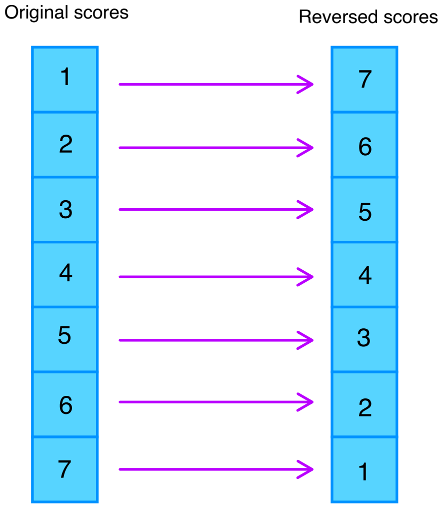


Figure 4: Process of reversing scores of the survey answers

Another type of questions we had to adjust was free-form and multiple-choice questions. We created dummy variables from the answers. If the answer was right, it was labeled "True," and if not, False. Later converted all "True" values to 1 and all "False" to 0 and summed them up. This process is displayed on the figure 5.

In the end, we summed up the rest of the scores with respect to their original scales. In such a manner that 12 gullibility scores were summed together, 4 sense of self questions were summed together, etc. This procedure reduced the dataset to 10 features instead of the initial 36 (excluding demographic and attention check questions). An example of such a procedure is displayed in Fig. 6.

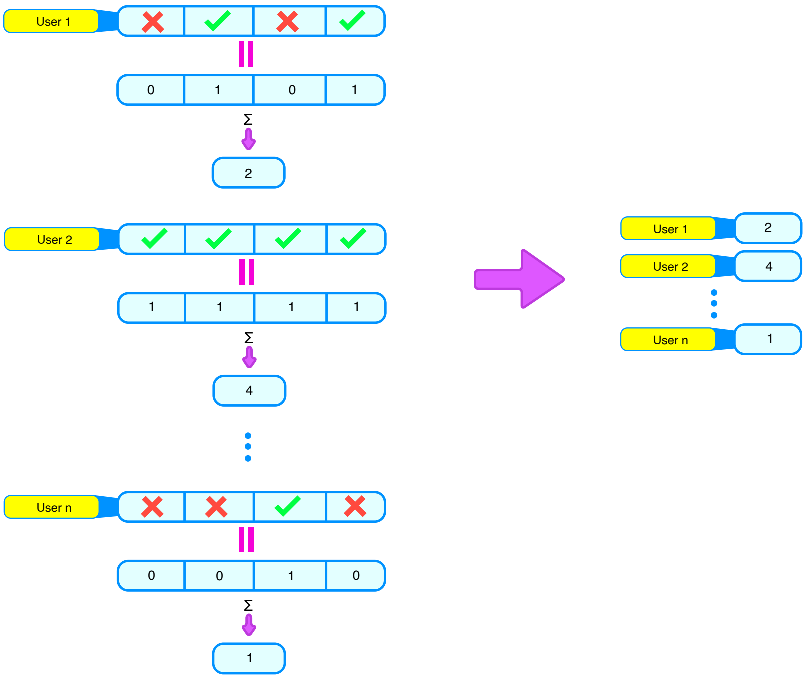


Figure 5: Aggregation of the free from and multiple choice questions

Other than social survey data, we had to pre-process the Twitter data. We have decided to split the Twitter data into three sets of features, as demonstrated in Fig.7. The details of the ”basic profile metrics and punctuation” have already been covered in the subsection 3.2.1. For the LIWC and fastText features, we merged each user's English tweets (statuses). For the LIWC features, we did not perform any additional preprocessing of the merged tweets. Instead, we passed the merged tweets as an input to the LIWC software to obtain the features from the text. In the software settings, we have marked all of the available features to be extracted from the text. In the end, we were given an output set with 117 features per user, or in other words, a dataframe of shape 117 by 159.

Similarly, we used merged tweets to create fastText sentence vector features. The idea behind the word and sentence vectors is to represent words or sentences by vectors and capture hidden information about the language, its semantics, and word analogies. It is also used to enhance the performance of the text classifiers [3]. In our example, we used pre-trained word vectors for the English language, trained on Wikipedia using fastText. The vectors had a dimension of 300 and were obtained using the skip-gram model described by Bojanowski et al. [2].

However, before obtaining the vectors, we had to clean up the text by applying the following procedure. Firstly we have tokenized the text using the TweetTokenizer, which is a tokenizer from the natural language toolkit library specially designed for

[3] `https://fasttext.cc/docs/en/unsupervised-tutorial.html`

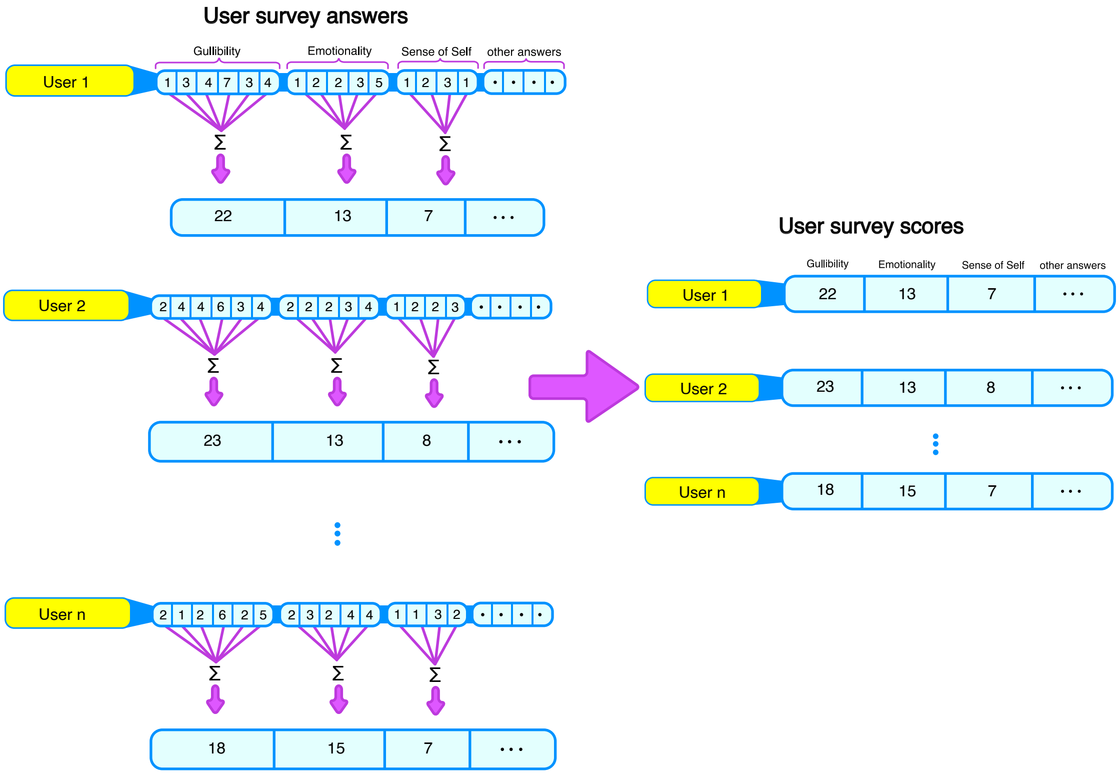


Figure 6: Aggregation of the survey answers which belong to the same scale

handling tweet-like texts [4]. Next, we removed the tokens that represented the user mentions or replies (they start with @), links, hashtags, emojis, and retweet indicator (”RT”). Tokens were then transformed to a lowercase and were filtered out by a set of stop words. Stop words in the English language are, for example, ”a”, ”is”, ”are”, ”the”, etc. All of the tokens that passed the filtering process were lemmatized. Lemmatization is the process that groups together the inflected forms of a word so they can be analyzed as a single item, also known as the word's lemma. After this, tokens were merged and saved. We obtained one sentence vector from this cleaned text for each user using the fastText function “get_sentence_vector.” We treated all (cleaned up) user tweets as one long sentence. At the end of this procedure, we got the output set consisting of sentence vectors of each user. This process is well illustrated in Fig. 8.

3.2.3 Model selection and hyperparameter tuning

The goal of our modeling phase is to predict the gullibility score of Twitter users. The gullibility score has been made by summing up answers to the 12 questions from the gullibility scale. For measuring responses to 12 gullibility questions, we used the 7-point Likert scale. The gullibility score ranges from 12 (low gullibility) to 84 (high

[4] https://www.nltk.org/howto/tokenize.html#regression-tests-tweettokenizer

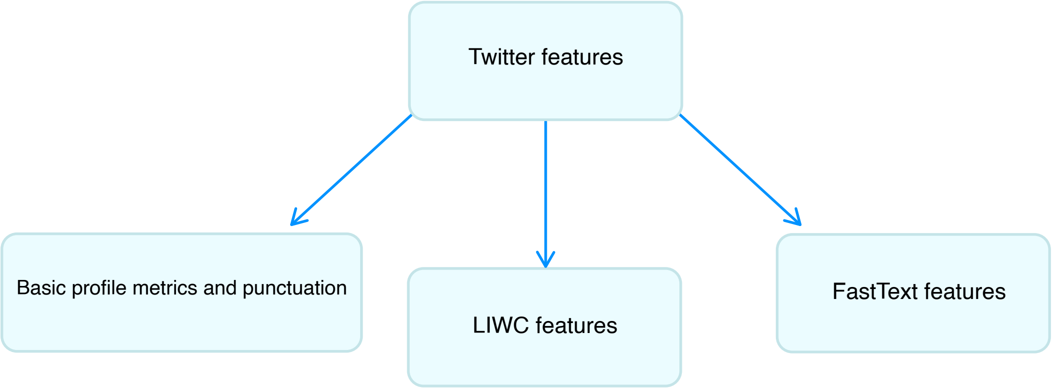


Figure 7: Three different sets of features obtained through Twitter

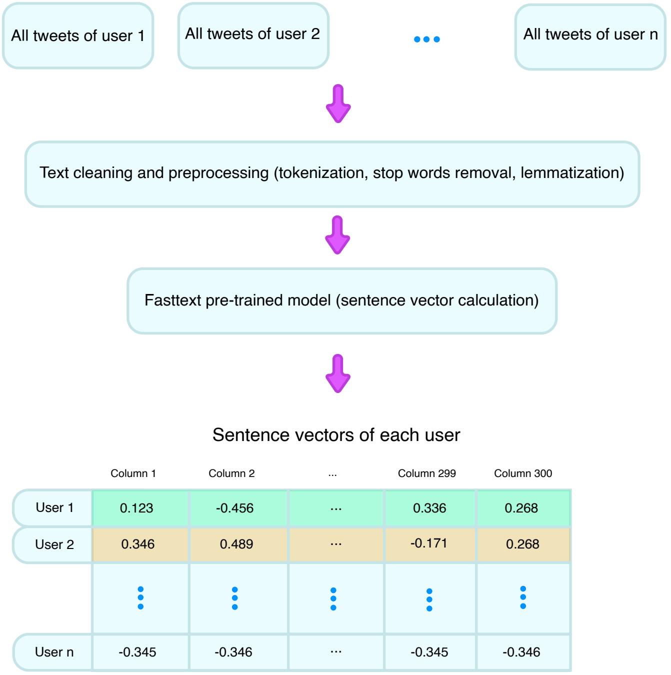
gullibility); however, we only managed to record values ranging from 12 to 60.

To predict the users' gullibility scores, we used 4 different sets of features, three Twitter sets, as mentioned in section 3.2.2. and the fourth set with survey answers. In order to obtain the best results, we have used the following combinations of the features:

1. Basic profile metrics and punctuation + Survey answers
2. Basic profile metrics and punctuation + LIWC features
3. Basic profile metrics and punctuation + Survey answers + LIWC features
4. Basic profile metrics and punctuation + Survey answers + LIWC features + fastText features

Because of the model complexity (a large number of input features), we selected the top 20 features using the "SelectKBest" function for feature selection from sklearn, with "score_func" parameter set to "f_classif"[5]. We did not apply feature selection on fastText dataset.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html


All tweets of user 1
All tweets of user 2
•••
All tweets of user n
Text cleaning and preprocessing (tokenization, stop words removal, lemmatization)
Fasttext pre-trained model (sentence vector calculation)
Sentence vectors of each user

	Column 1	Column 2	...	Column 299	Column 300
User 1	0.123	-0.456	···	0.336	0.268
User 2	0.346	0.489	···	-0.171	0.268
	⋮	⋮	⋮	⋮	⋮
User n	-0.345	-0.346	···	-0.345	-0.346

Figure 8: Process of obtaining fastText sentence vectors from user's tweets

Since we did not know if predicting the gullibility from social media traces was possible, and if it was more of a classification or regression problem, we decided to take both approaches. For the classification task, we performed the median split of the gullibility scores, meaning all values smaller than the median were labeled with 0 and bigger with 1. For the regression task, we tried to predict the exact value of the user's gullibility score.

Models used for classification tasks were logistic regression, Gaussian naive Bayes, k nearest neighbors (KNN), support vector classifier (SVC), gradient boosting, random forest, extra trees, decision tree, stochastic gradient descent (SGD), ridge, multi-layer perceptron (MLP) and majority voting. Besides them, we combined bagging and ada boost with some of the classification models to achieve better results.

Models used for regression tasks were k nearest neighbors (KNN), partial least squares (PLS), decision tree, Bayesian ridge, Huber regressor, ridge, linear regression, support vector regressor (SVR), extra trees. Similarly, we combined bagging and ada boost with some of the regression models to achieve better results.

Due to a small number of data points (159), we decided to use cross-validation for data splitting. However, because of hyper-parameter tuning, we had to be extra careful. Suppose we were to use the same data for tuning and evaluation of the model. In that case, some of the information may leak into the model, causing the overfit of the data and optimistic evaluation. To solve this problem, we used the nested cross-validation technique as described by Cawley and Talbot [3]. In our particular case we used the 5-fold cross-validation procedure for model hyper-parameter optimization and nested it inside the 5-fold cross-validation procedure for model selection. An illustration of the nested cross-validation technique we used is displayed in figure 9.

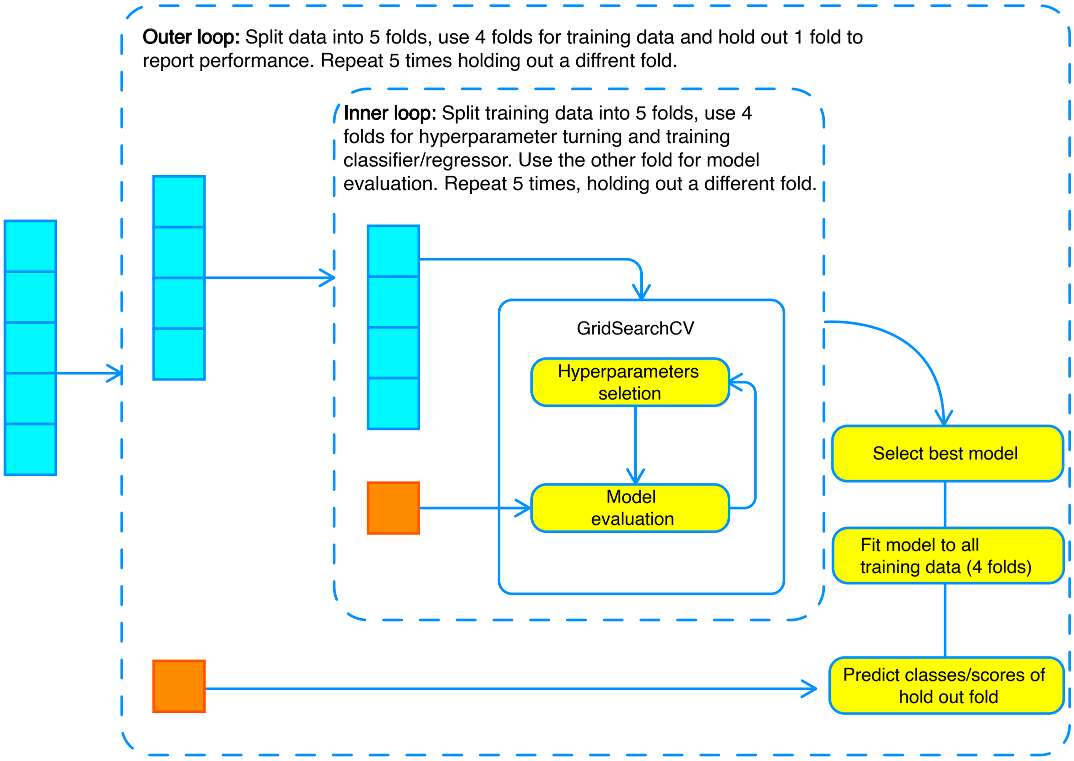

Outer loop: Split data into 5 folds, use 4 folds for training data and hold out 1 fold to report performance. Repeat 5 times holding out a diffrent fold.
Inner loop: Split training data into 5 folds, use 4 folds for hyperparameter turning and training classifier/regressor. Use the other fold for model evaluation. Repeat 5 times, holding out a different fold.
GridSearchCV
Hyperparameters selection
Model evaluation
Select best model
Fit model to all training data (4 folds)
Predict classes/scores of hold out fold

Figure 9: Nested cross-validation consisted of two loops with five folds each. The outer loop's training set is an input of the inner loop, which we used to perform the hyper-parameter tuning and model evaluation. The test set of the outer loop is used for reporting the results

To make our results replaceable, we set the random state of each model and each data splitter to 1. Every model produced 5 results, one for each combination of the folds; from those results, we reported the average score and its standard deviation. The baseline model used for the classification task was set to always predict the most frequent label in the training set. The baseline model used for the regression task was set to always predict the mean value of the training set. Hyper-parameter optimization was performed using "GridSearchCV" [6]. For classification models, we optimized hyper-parameters for accuracy, and for regression models for root mean square error (RMSE). A set of hyper-parameters and their respective models that we used across all combinations of features are displayed in the appendix C.

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

4 RESULTS

4.1 EVALUATION CRITERIA

Since we are researching a pretty undiscovered topic and using a nontraditional method for measuring gullibility, we decided to address the problem using both classification and regression models with their respective metrics.

To evaluate the proposed classification models, we used the following metrics: accuracy, precision, recall, F1, and area under the receiver operating characteristic curve (AUC). We optimized the models for accuracy but included other metrics simply because the accuracy alone does not offer much information about the model's performance. For example, a highly imbalanced dataset model could always predict the majority class and get very high accuracy. Still, the truth is, it fails to predict the other classes, which might be crucial. In our case, we are not dealing with an imbalanced dataset regarding the classification task because we did the median split of the target variable. However, in reality, we want to have a good balance between precision and recall or, in other words, a high f1 score and area under the curve. Formulas for calculating accuracy, precision, recall, and f1 that we used in the classification task are the following:

$$\text{Accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

where y is the vector of observed values and ŷ is the vector of predicted values

$$Precision = \text{true positives}/(\text{true positives} + \text{false positives})$$

$$Recall = \text{true positives}/(\text{true positives} + \text{false negatives})$$

$$F1 = 2 * (precision * recall)/(precision + recall)$$

We used the root mean square error (RMSE) and mean absolute error (MAE) for regression models. Both express average model prediction error in units of the variable of interest and are negatively-oriented scores. But with RMSE, the errors are squared before they are averaged, meaning it penalizes large errors. It does not describe average error alone and that is why we also use MAE as a second metric for regression tasks. Mentioned metrics were calculated according to the following formulas:

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}.$$

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \left| y_i - \hat{y}_i \right|.$$

where y is the vector of observed values and ŷ is the vector of predicted values

4.2 SURVEY RESULTS

In Fig. 10 we can see the distribution of gullibility score frequencies. The minimal recorded score is 12, and the maximum is 60. The median score is 28.0, and the mean is 29.74; the whole distribution is shifted to the left side, which is expected with the answers coming from a self-report scale.

In Fig. 11 we can see the participants' distribution of age and gender. In the study participated, 81 females, 72 males, and 6 others. The gender is not evenly distributed across different age groups. The majority of the participants are people between 21 and 40 years of age.

Figure 12 shows the correlations between different features collected in the social survey. We observe a high negative correlation between subjective financial knowledge and gullibility score and financial attitude and gullibility score. Also, we observe a high positive correlation between subjective financial knowledge and subjective financial satisfaction, weak sense of self and gullibility score, emotionality and gullibility score, and financial attitude and financial satisfaction. Features correlated the most to gullibility regardless of the direction of correlation were subjective financial knowledge, weak sense of self, emotionality, and financial attitude. On the contrary, the least correlated features were general trust, education, age group, and cognitive reflection score.

4.3 CLASSIFICATION

In Tabs. 2, 3, 4, and 5 we summarized the results of the classification tasks. We classified each user as being either gullible or not based on their gullibility score. The baseline algorithm was predicting the most frequent class (majority classifier) in the training set. Each table presents the model's performances with a different set of features. We reported the average accuracy score from the 5 splits. For the detailed tables that include other metrics, please look at the appendix D.

In Fig. 13 we can see the performance of the classification models used with the first set of features. Out of 19 models, 14 of them were better than the baseline. The top 3

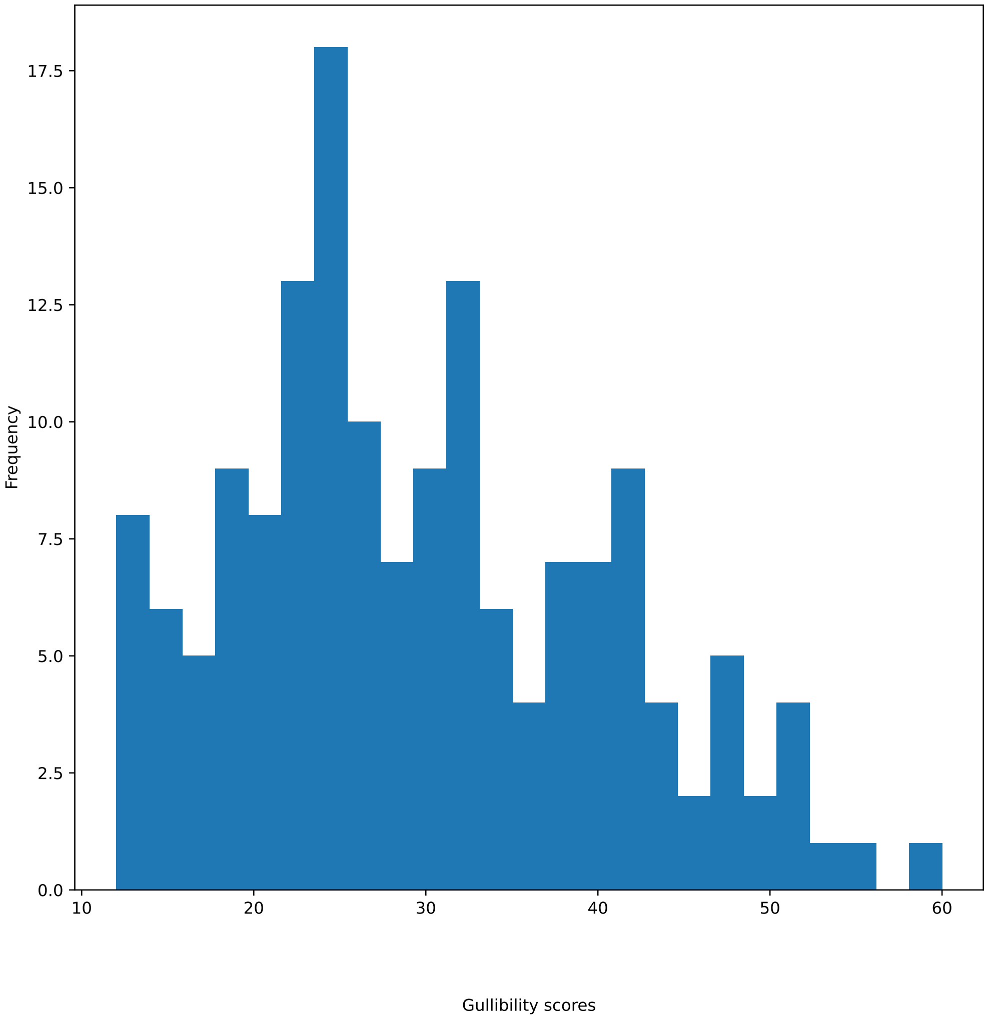


Figure 10: Frequency distribution of gullibility scores of the participants

models were majority voting (0.591), ada boost + ridge (0.585) and SVC (0.579). The bottom 3 models were KNN (0.460), decision tree (0.471), and SGD classifier (0.491).

In Fig. 14 we can see the performance of the classification models used with the second set of features. Out of 19 models, 17 of them were better than the baseline. The top 3 models were extra trees (0.636), random forest (0.604), and Gradient boosting classifier (0.603). The bottom 2 models were MLP (0.515) and KNN classifier (0.503).

In Fig. 15 we can see the performance of the classification models used with the third set of features. Out of 19 models, 17 of them were better than the baseline. The top 3 models were bagging + logistic regression (0.667), logistic regression (0.642) and ada boost + logistic regression (0.635). The bottom 2 models were KNN (0.510) and MLP classifier (0.497).

In Fig. 16 we can see the performance of the classification models used with the

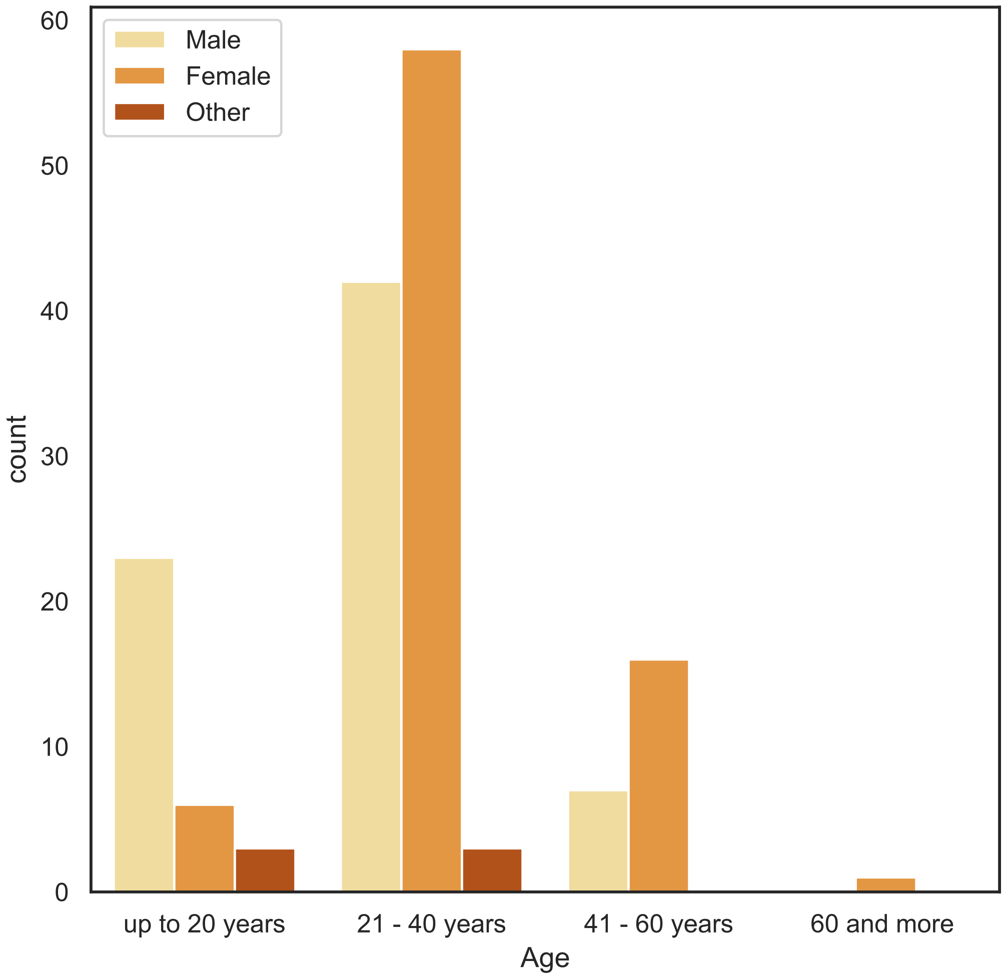


Figure 11: Age-gender distribution of participants

fourth set of features. Out of 19 models, 13 of them were better than the baseline. The top 3 models were bagging + logistic regression (0.598), majority voting (0.591), and Gaussian naive Bayes (0.585). Bottom 3 models were ada boost + random forest (0.504), KNN (0.503), and random forest classifier (0.491).

In Fig. 17 we can see the performance of top 15 classification models across all sets of features. All 15 models performed better than the baseline model. The top 3 models were bagging + logistic regression with dataset 3 (0.667), logistic regression with dataset 3 (0.642) and extra trees classifier with dataset 2 (0.636). The bottom 3 models were bagging + logistic regression with dataset 4 (0.598), ada boost + decision tree classifier with dataset 2 (0.597) and gradient boosting classifier with dataset 3 (0.592).

In Fig. 18 we show how each classification model preforms across different feature

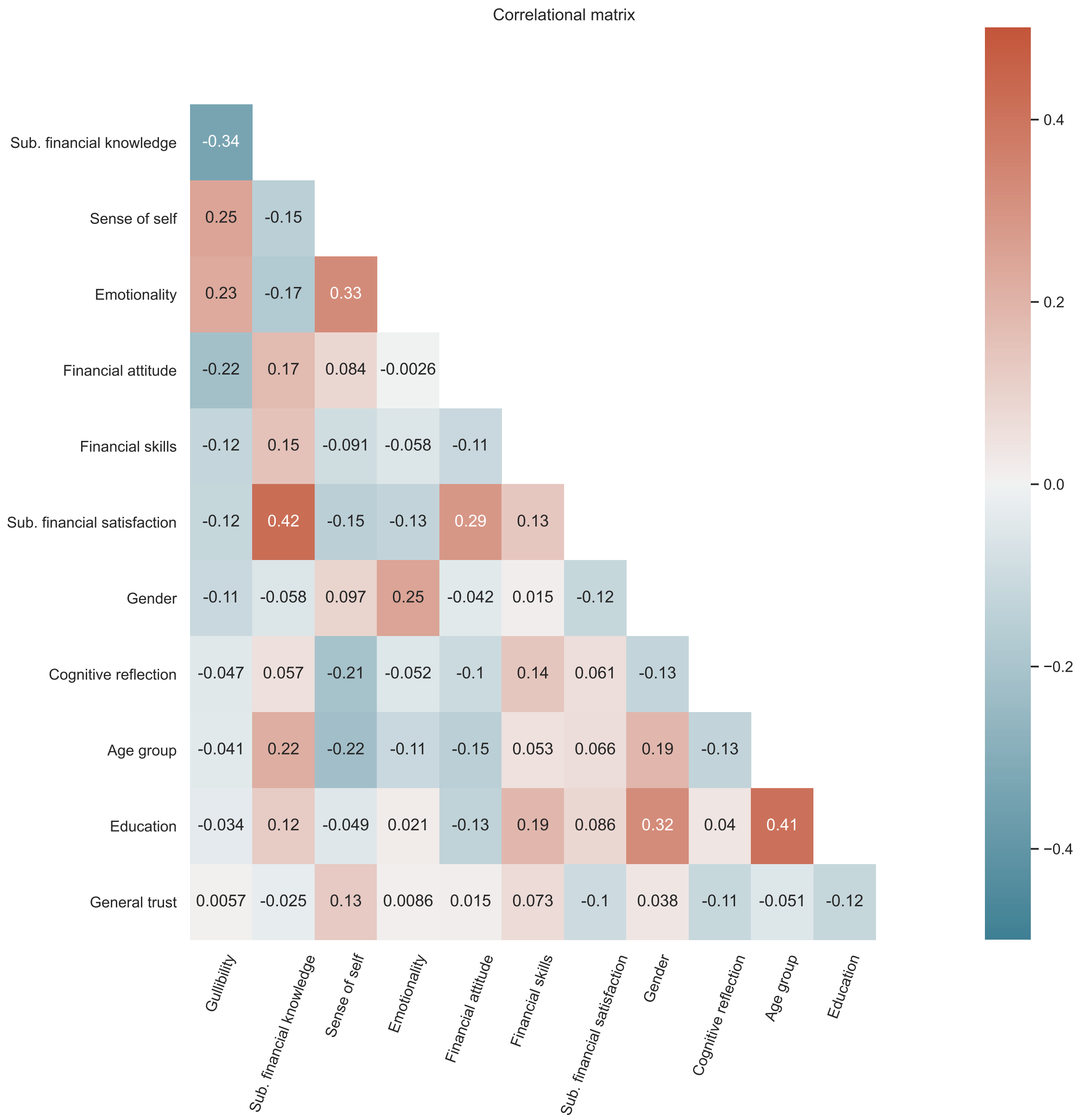


Figure 12: Gullibility correlation matrix created from the survey answers

sets. From the figure, we can conclude that most models perform best with third set of features.

4.4 REGRESSION

In Tabs. 6, 7, 8, and 9 we summarized the results of the regression tasks where we predicted the value of the gullibility score. The baseline algorithm predicted the mean value of the gullibility score in the training set. Each table presents the model's performances with a different set of features. We reported the average RMSE and MAE scores from the 5 splits.

In Fig. 19 we can see the performance of the regression models used with the first

Table 2: Results of classification models using first set of features, bolded models achieved higher accuracy than the baseline

Model name	Set of features	Accuracy (mean)	Accuracy (std)
Baseline	1	0.516	0.014
Logistic Regression	1	**0.572**	0.076
Gaussian Naive Bayes classifier	1	**0.566**	0.026
KNN classifier	1	0.460	0.079
SVC	1	**0.579**	0.053
Gradient Boosting classifier	1	**0.522**	0.044
Random forest classifier	1	**0.566**	0.045
Decision Tree classifier	1	0.471	0.101
SGD classifier	1	0.491	0.046
Ridge classifier	1	**0.522**	0.036
Extra trees classifier	1	**0.542**	0.082
MLP classifier	1	0.491	0.044
Majority voting classifier	1	**0.591**	0.020
Bagging + Logistic Regression	1	**0.572**	0.054
Ada Boost + Gradient Boosting classifier	1	**0.566**	0.026
Ada Boost + Logistic Regression	1	**0.578**	0.059
Ada Boost + Random forest classifier	1	**0.572**	0.062
Ada Boost + SGD classifier	1	**0.573**	0.043
Ada Boost + Decision Tree classifier	1	0.516	0.044
Ada Boost + Ridge classifier	1	**0.585**	0.034

set of features. Out of 18 models, 7 of them were better than the baseline. The top 3 models were PLS regressor (9.825), extra trees regressor (10.04), and ada boost + random forest regressor (10.119). The bottom 3 models were SVR (11.026), decision tree regressor (13.774), and Huber regressor (18.258).

In Fig. 20 we can see the performance of the regression models used with the second set of features. Out of 18 models, 0 of them were better than the baseline. The bottom 3 models were linear regression (13.147), decision tree regressor (15.023), and Huber regressor (18.737).

In Fig. 21 we can see the performance of the regression models used with the third set of features. Out of 18 models, 5 of them were better than the baseline. The top 3 models were extra trees regressor (10.349), ada boost + random forest regressor (10.452), and bagging + random forest regressor (10.604). The bottom 3 models were linear regression (11.676), decision tree regressor (14.532), and Huber regressor (17. 595).

In Fig. 22 we can see the performance of the regression models used with the fourth set of features. Out of 18 models, 0 of them were better than the baseline. The bottom 3 models were decision tree regressor (15.411), Huber regressor (18.022), and linear

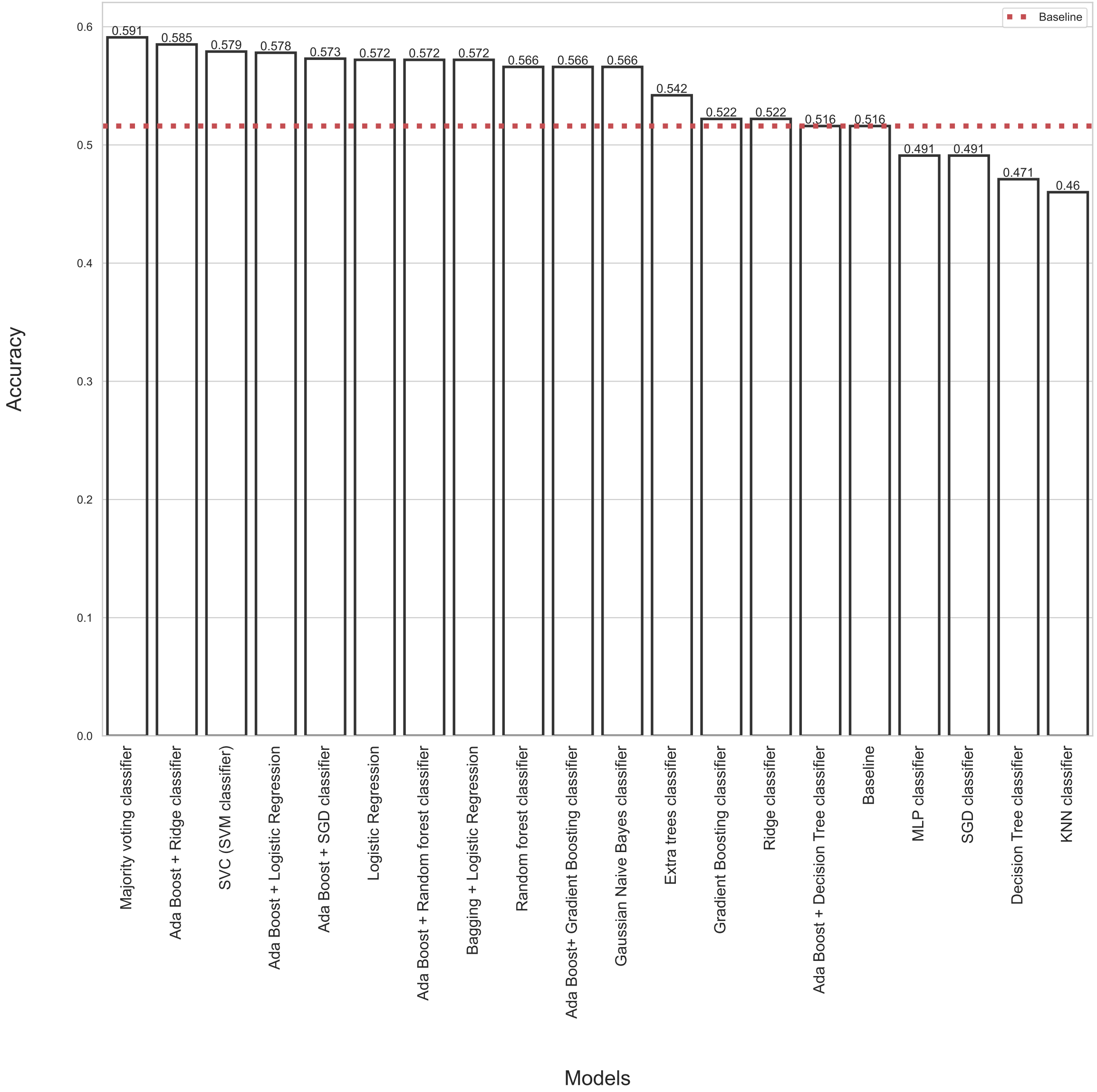


Figure 13: Bar chart showing the performance of classification models that were used with the first set of features, sorted in descending order

regression (25.963).

In Fig. 23 we can see the performance of top 10 regression models across all feature sets. All 10 models performed better than the baseline model. The top 3 models were PLS regressor with dataset 1 (9.825), extra trees regressor with dataset 1 (10.04), and ada boosting + random forest regressor with dataset 1 (10.119). The bottom 3 models were extra trees regressor with dataset 3 (10.349), ada boost + random forest regressor with dataset 3 (10.452), and bagging + random forest regressor with dataset 3 (10.604).

In Fig. 24 we show how each regression model performs across a different set of features. From the figure, we can conclude that most regression models perform similarly or worse than the baseline regardless of the features used.

Table 3: Results of classification models using the second set of features, bolded models achieved higher accuracy than the baseline

Model name	Set of features	Accuracy (mean)	Accuracy (std)
Baseline	2	0.516	0.014
Logistic Regression	2	**0.578**	0.052
Gaussian Naive Bayes classifier	2	**0.579**	0.014
KNN classifier	2	0.503	0.029
SVC	2	**0.553**	0.042
Gradient Boosting classifier	2	**0.603**	0.069
Random forest classifier	2	**0.604**	0.059
Decision Tree classifier	2	**0.541**	0.094
SGD classifier	2	**0.535**	0.029
Ridge classifier	2	**0.585**	0.077
Extra trees classifier	2	**0.636**	0.090
MLP classifier	2	0.515	0.073
Majority voting classifier	2	**0.578**	0.068
Bagging + Logistic Regression	2	**0.585**	0.059
Ada Boost + Gradient Boosting classifier	2	**0.541**	0.079
Ada Boost + Logistic Regression	2	**0.579**	0.024
Ada Boost + Random forest classifier	2	**0.585**	0.055
Ada Boost + SGD classifier	2	**0.553**	0.077
Ada Boost + Decision Tree classifier	2	**0.597**	0.037
Ada Boost + Ridge classifier	2	**0.566**	0.049

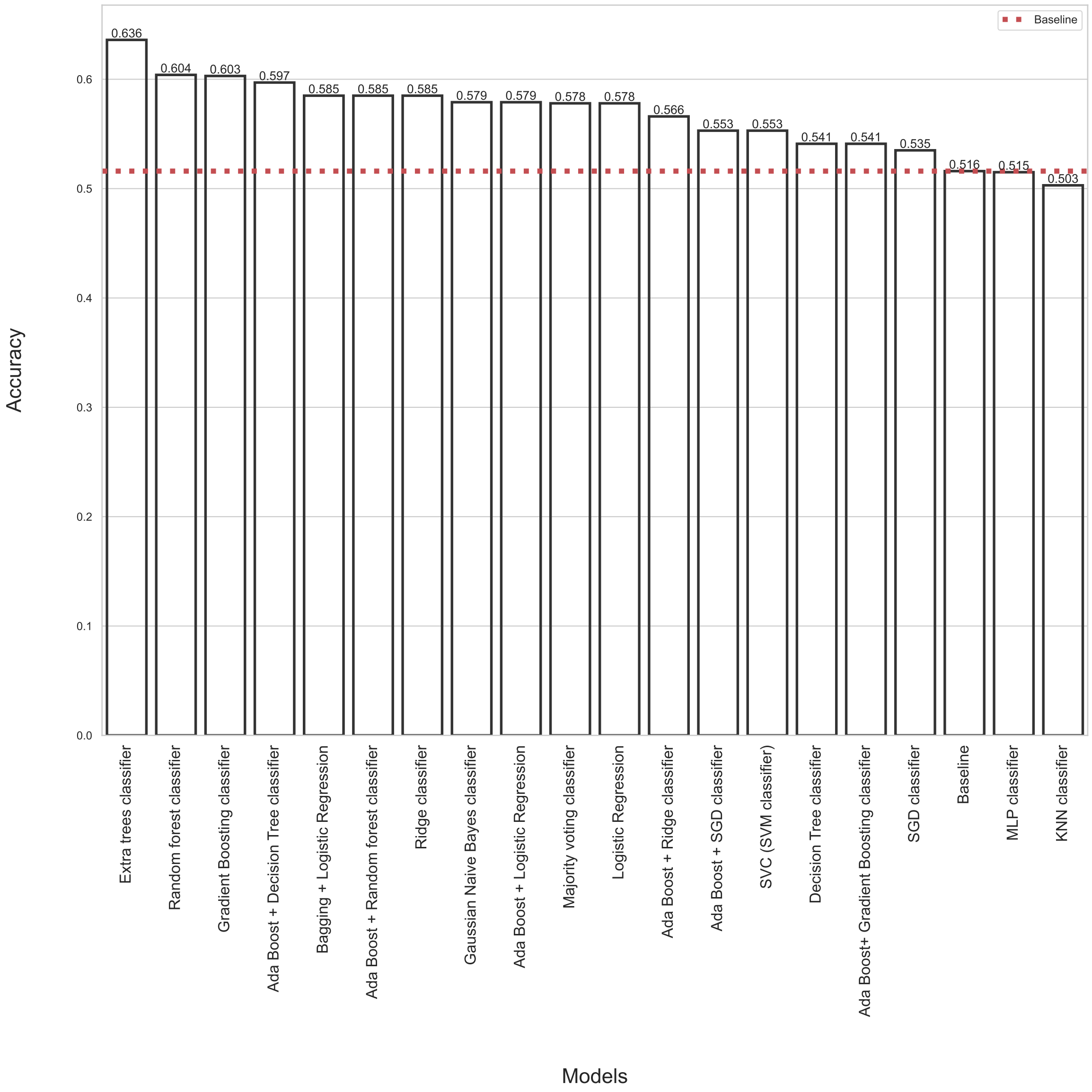


Figure 14: Bar chart showing the performance of classification models that were used with the second set of features, sorted in descending order

Table 4: Results of classification models using the third set of features, bolded models achieved higher accuracy than the baseline

Model name	Set of features	Accuracy (mean)	Accuracy (std)
Baseline	3	0.516	0.014
Logistic Regression	3	**0.642**	0.046
Gaussian Naive Bayes classifier	3	**0.604**	0.050
KNN classifier	3	0.510	0.039
SVC	3	**0.553**	0.042
Gradient Boosting classifier	3	**0.592**	0.069
Random forest classifier	3	**0.616**	0.064
Decision Tree classifier	3	**0.534**	0.082
SGD classifier	3	**0.522**	0.023
Ridge classifier	3	**0.598**	0.081
Extra trees classifier	3	**0.629**	0.097
MLP classifier	3	0.497	0.035
Majority voting classifier	3	**0.610**	0.116
Bagging + Logistic Regression	3	**0.667**	0.080
Ada Boost + Gradient Boosting classifier	3	**0.553**	0.064
Ada Boost + Logistic Regression	3	**0.635**	0.053
Ada Boost + Random forest classifier	3	**0.591**	0.090
Ada Boost + SGD classifier	3	**0.540**	0.069
Ada Boost + Decision Tree classifier	3	**0.598**	0.073
Ada Boost + Ridge classifier	3	**0.566**	0.049

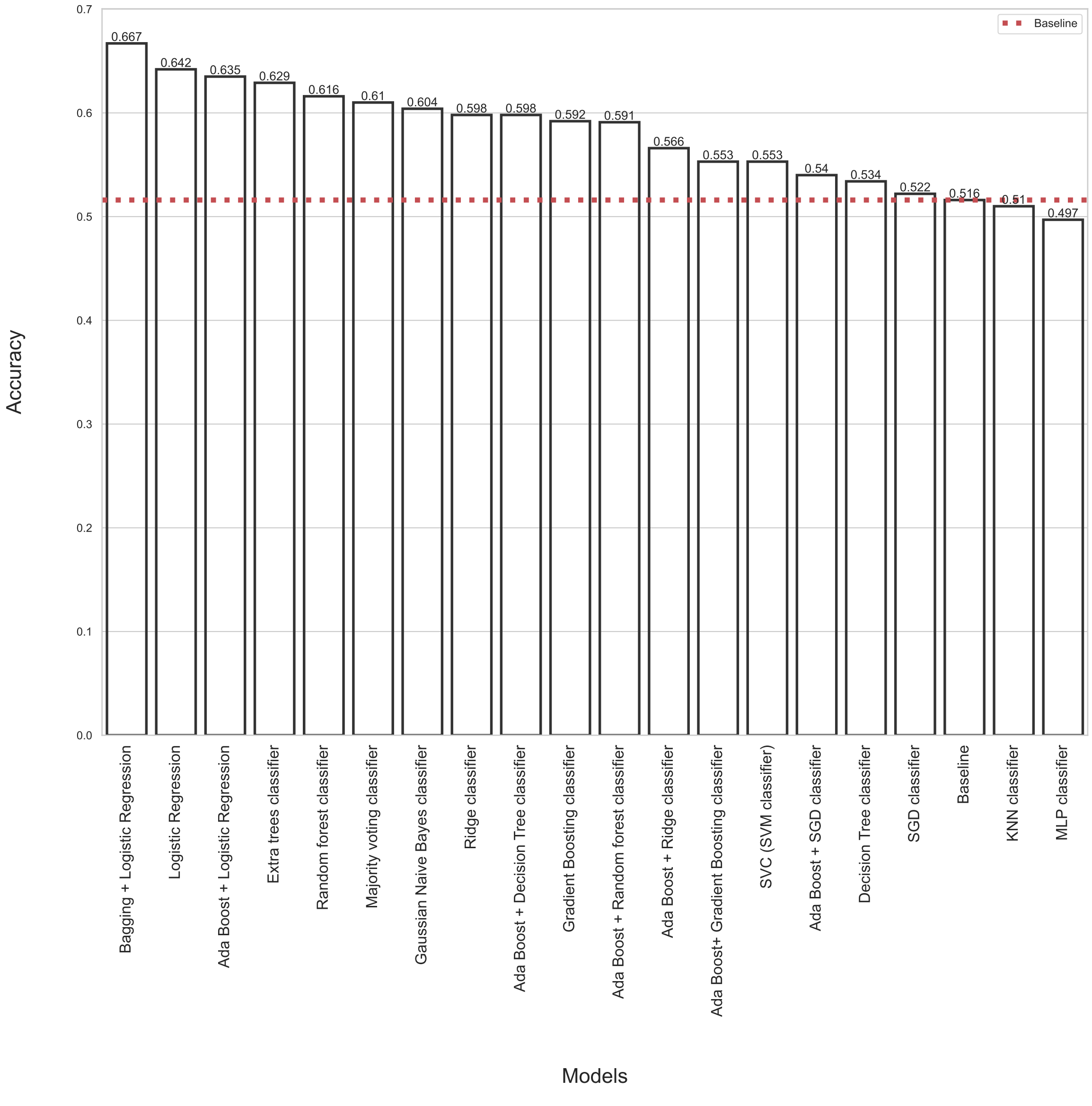


Figure 15: Bar chart showing the performance of classification models that were used with the third set of features, sorted in descending order

Table 5: Results of classification models using the forth set of features, bolded models achieved higher accuracy than the baseline

Model name	Set of features	Accuracy (mean)	Accuracy (std)
Baseline	4	0.516	0.014
Logistic Regression	4	**0.579**	0.054
Gaussian Naive Bayes classifier	4	**0.585**	0.020
KNN classifier	4	0.503	0.029
SVC	4	**0.553**	0.042
Gradient Boosting classifier	4	0.510	0.053
Random forest classifier	4	0.491	0.094
Decision Tree classifier	4	**0.541**	0.037
SGD classifier	4	**0.535**	0.029
Ridge classifier	4	**0.573**	0.069
Extra trees classifier	4	**0.523**	0.118
MLP classifier	4	0.516	0.014
Majority voting classifier	4	**0.591**	0.045
Bagging + Logistic Regression	4	**0.598**	0.063
Ada Boost + Gradient Boosting classifier	4	**0.541**	0.079
Ada Boost + Logistic Regression	4	**0.566**	0.035
Ada Boost + Random forest classifier	4	0.504	0.101
Ada Boost + SGD classifier	4	**0.553**	0.077
Ada Boost + Decision Tree classifier	4	0.510	0.039
Ada Boost + Ridge classifier	4	**0.566**	0.049

Table 6: Results of regression models using the first set of features, bolded models achieved smaller (better) RMSE than the baseline

Model name	Set of features	RMSE (mean)	RMSE (std)	MAE (mean)	MAE (std)
Baseline	1	10.633	1.113	8.885	1.194
KNN regressor	1	10.861	1.023	8.885	0.926
PLS regressor	1	**9.825**	1.609	8.049	1.613
Decision Tree regressor	1	13.774	1.235	10.341	1.012
Bayesian Ridge regressor	1	10.763	1.060	8.963	1.149
Huber regressor	1	18.258	2.085	14.898	1.878
Ridge regressor	1	10.943	1.169	9.068	1.242
Linear Regression	1	**10.290**	1.664	8.507	1.651
SVR	1	11.026	1.187	9.050	1.200
Extra Trees regressor	1	**10.040**	1.183	8.286	1.178
Bagging + Random Forest regressor	1	**10.161**	1.127	8.329	1.226
Ada Boosting + Random Forest regressor	1	**10.119**	1.162	8.240	1.292
Ada Boost + Decision Tree regressor	1	10.970	1.596	8.684	1.682
Ada Boost + Ridge regressor	1	10.834	1.044	9.154	1.216
Ada Boost+ SVR	1	10.661	1.145	8.945	1.254
Random Forest regressor	1	**10.165**	1.138	8.282	1.162
Gradient Boosting regressor	1	**10.267**	1.164	8.447	1.207
Bagging + SVR	1	10.646	1.093	8.880	1.171
Bagging + Ridge regressor	1	10.751	1.018	9.050	1.100

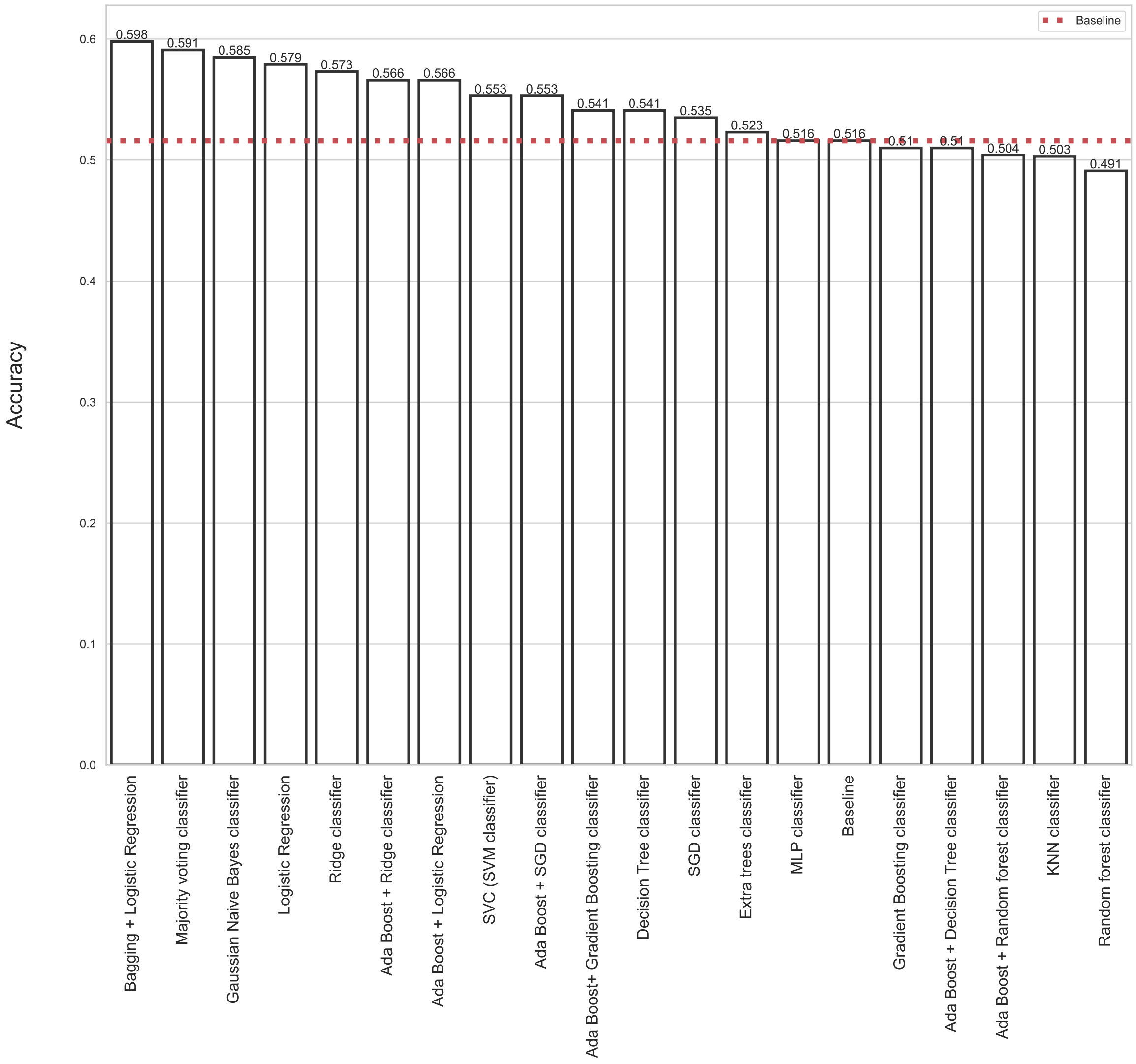


Figure 16: Bar chart showing the performance of classification models that were used with the forth set of features, sorted in descending order

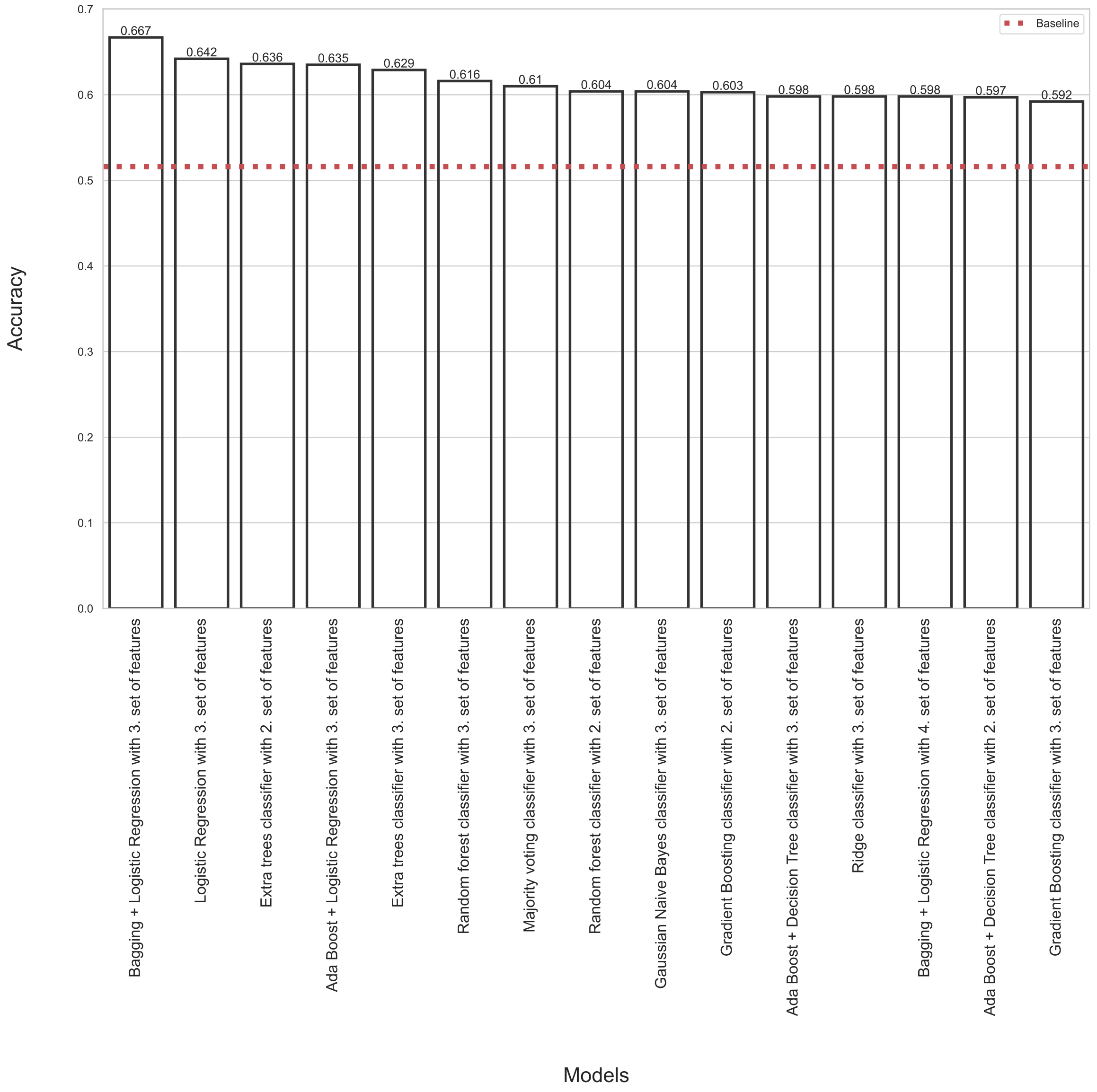


Figure 17: Bar chart showing the performance of top 15 classification models regardless of used set of features, sorted in descending order

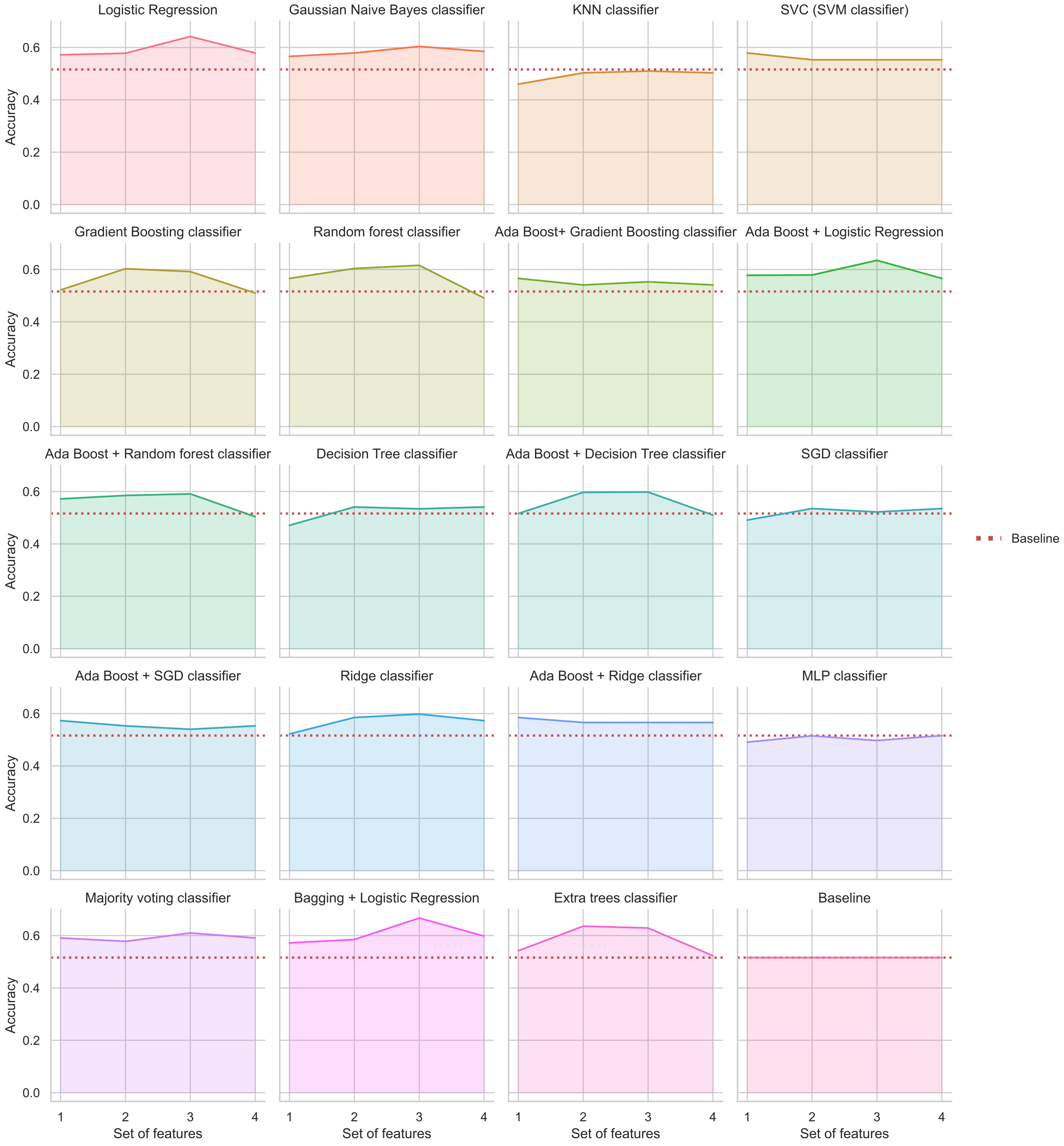


Figure 18: Performance of each classification model across different set of features

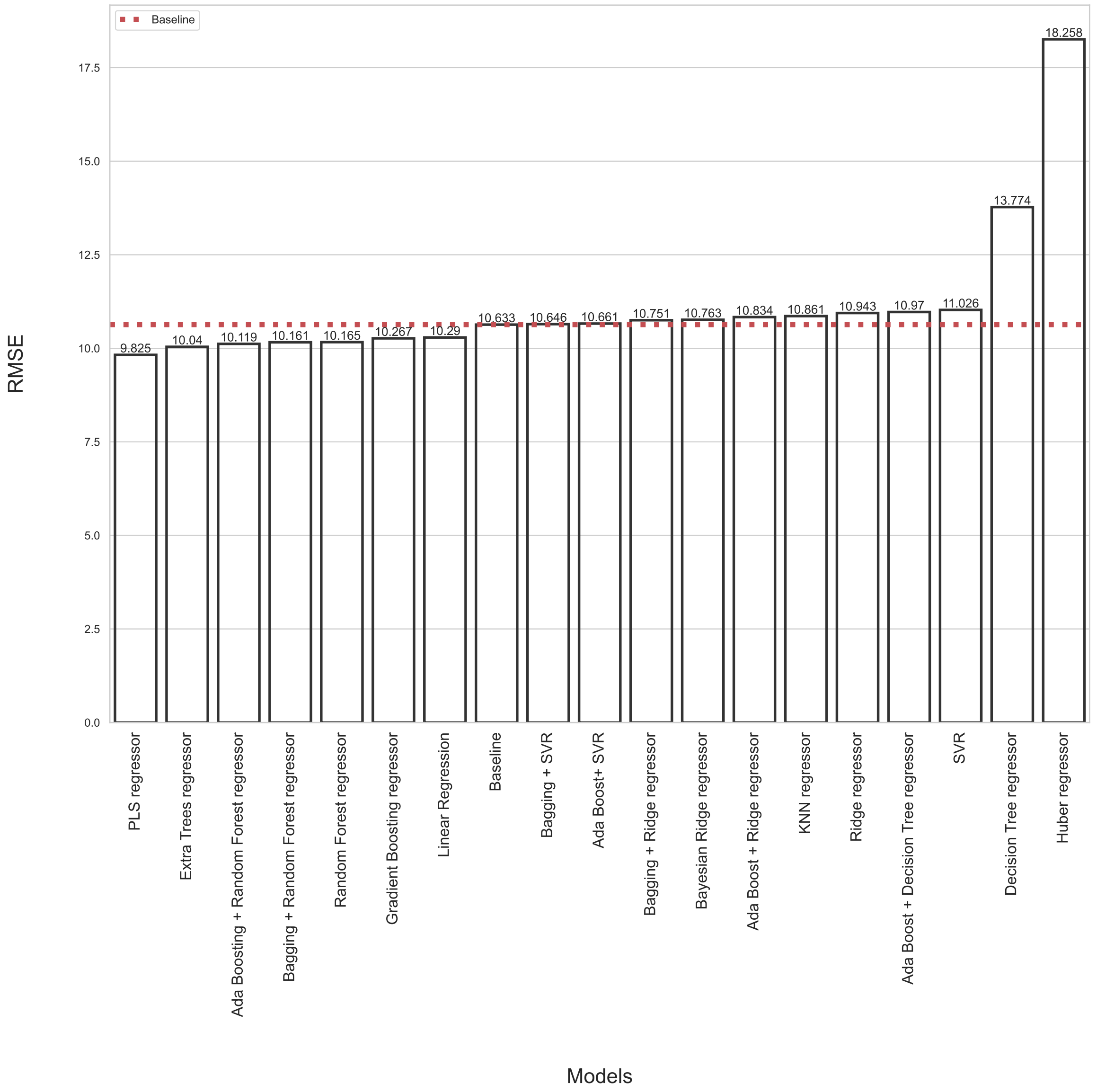


Figure 19: Bar chart showing the performance of regression models that were used with the first set of features, sorted in ascending order (lower RMSE is better)

Table 7: Results of regression models using the second set of features

Model name	Set of features	RMSE (mean)	RMSE (std)	MAE (mean)	MAE (std)
Baseline	2	10.633	1.113	8.885	1.194
KNN regressor	2	10.879	1.118	8.826	1.124
PLS regressor	2	11.732	0.868	9.573	0.849
Decision Tree regressor	2	15.023	1.801	12.085	1.625
Bayesian Ridge regressor	2	10.792	1.078	9.024	1.132
Huber regressor	2	18.737	2.211	15.067	2.030
Ridge regressor	2	10.785	1.031	9.064	1.127
Linear Regression	2	13.147	0.574	10.592	0.333
SVR	2	11.030	1.185	9.054	1.198
Extra Trees regressor	2	10.912	1.449	8.994	1.271
Bagging + Random Forest regressor	2	10.800	1.100	8.908	1.028
Ada Boosting + Random Forest regressor	2	10.935	1.122	9.122	1.055
Ada Boost + Decision Tree regressor	2	11.542	1.130	9.135	0.908
Ada Boost + Ridge regressor	2	10.791	1.050	9.116	1.208
Ada Boost+ SVR	2	10.661	1.145	8.945	1.254
Random Forest regressor	2	11.021	1.025	9.116	1.059
Gradient Boosting regressor	2	11.257	1.583	9.246	1.171
Bagging + SVR	2	10.646	1.093	8.880	1.171
Bagging + Ridge regressor	2	10.752	1.018	9.052	1.100

Table 8: Results of regression models using the third set of features, bolded models achieved smaller (better) RMSE than the baseline

Model name	Set of features	RMSE (mean)	RMSE (std)	MAE (mean)	MAE (std)
Baseline	3	10.633	1.113	8.885	1.194
KNN regressor	3	10.864	1.100	8.811	1.102
PLS regressor	3	10.688	1.380	8.834	1.238
Decision Tree regressor	3	14.532	1.717	11.469	1.697
Bayesian Ridge regressor	3	10.792	1.078	9.025	1.131
Huber regressor	3	17.595	3.247	13.570	2.454
Ridge regressor	3	11.109	1.307	9.147	1.226
Linear Regression	3	11.676	1.287	9.570	1.063
SVR	3	11.027	1.188	9.052	1.200
Extra Trees regressor	3	**10.349**	1.152	8.570	0.942
Bagging + Random Forest regressor	3	**10.604**	1.129	8.808	1.048
Ada Boosting + Random Forest regressor	3	**10.632**	1.207	8.839	1.069
Ada Boost + Decision Tree regressor	3	11.355	1.200	9.184	0.975
Ada Boost + Ridge regressor	3	10.791	1.050	9.116	1.208
Ada Boost+ SVR	3	10.661	1.145	8.945	1.254
Random Forest regressor	3	**10.452**	1.227	8.749	0.980
Gradient Boosting regressor	3	**10.622**	0.990	8.838	0.890
Bagging + SVR	3	10.646	1.093	8.880	1.171
Bagging + Ridge regressor	3	10.752	1.018	9.052	1.100

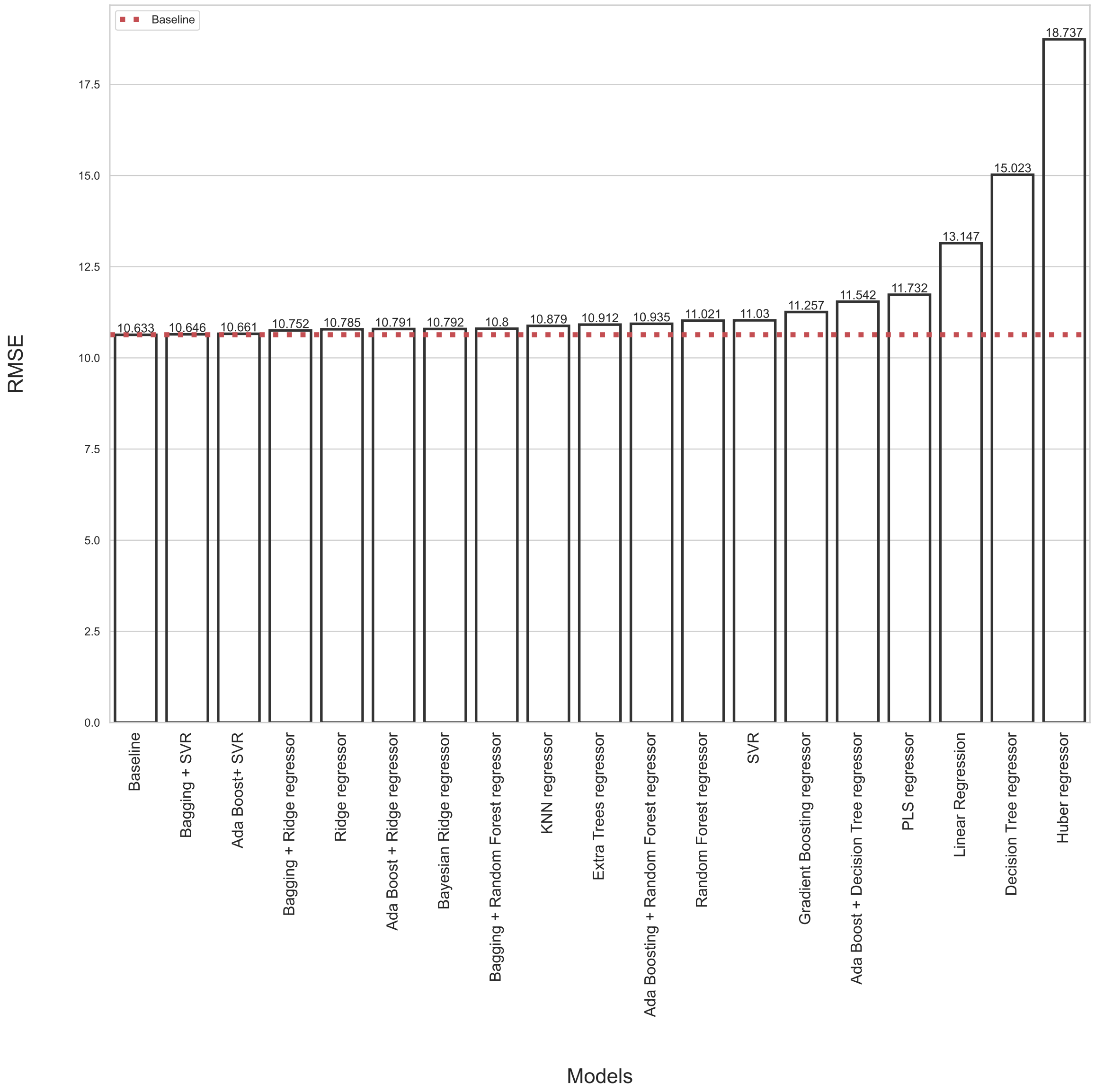


Figure 20: Bar chart showing the performance of regression models that were used with the second set of features, sorted in ascending order (lower RMSE is better)

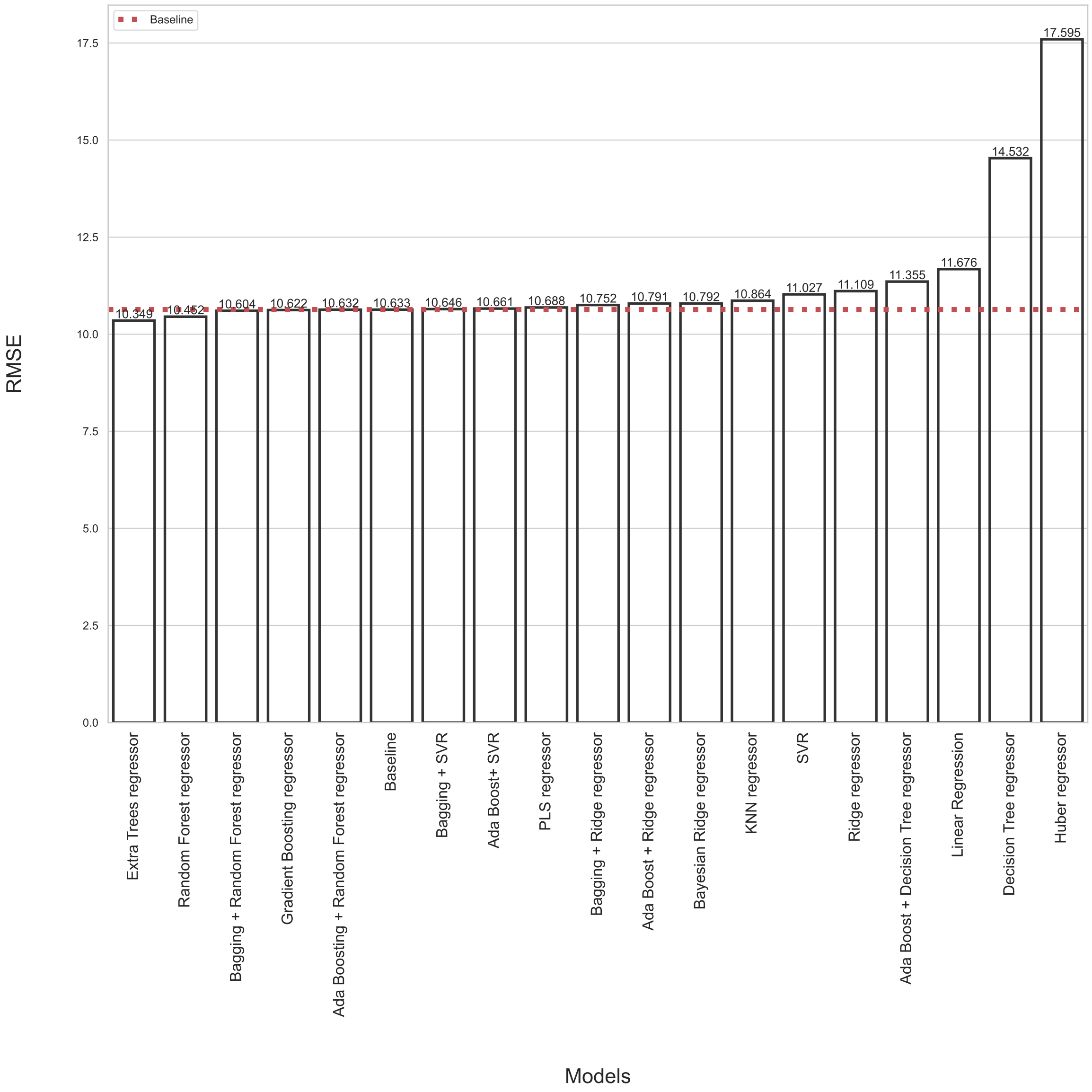


Figure 21: Bar chart showing the performance of regression models that were used with the third set of features, sorted in ascending order (lower RMSE is better)

Table 9: Results of regression models using the forth set of features

Model name	Set of features	RMSE (mean)	RMSE (std)	MAE (mean)	MAE (std)
Baseline	4	10.633	1.113	8.885	1.194
KNN regressor	4	10.879	1.118	8.826	1.124
PLS regressor	4	11.576	1.474	9.490	1.571
Decision Tree regressor	4	15.411	1.158	12.504	1.167
Bayesian Ridge regressor	4	10.792	1.079	9.024	1.132
Huber regressor	4	18.022	1.391	14.285	1.376
Ridge regressor	4	10.785	1.031	9.064	1.127
Linear Regression	4	25.963	4.569	18.805	2.633
SVR	4	11.045	1.188	9.071	1.192
Extra Trees regressor	4	10.824	1.153	8.947	1.177
Bagging + Random Forest regressor	4	10.699	1.045	8.929	1.102
Ada Boosting + Random Forest regressor	4	10.998	1.089	9.120	1.166
Ada Boost + Decision Tree regressor	4	11.694	0.959	9.693	1.171
Ada Boost + Ridge regressor	4	10.791	1.050	9.116	1.208
Ada Boost+ SVR	4	10.661	1.145	8.945	1.254
Random Forest regressor	4	11.061	0.985	9.137	1.052
Gradient Boosting regressor	4	10.966	1.091	9.228	1.152
Bagging + SVR	4	10.646	1.093	8.880	1.171
Bagging + Ridge regressor	4	10.752	1.018	9.052	1.100

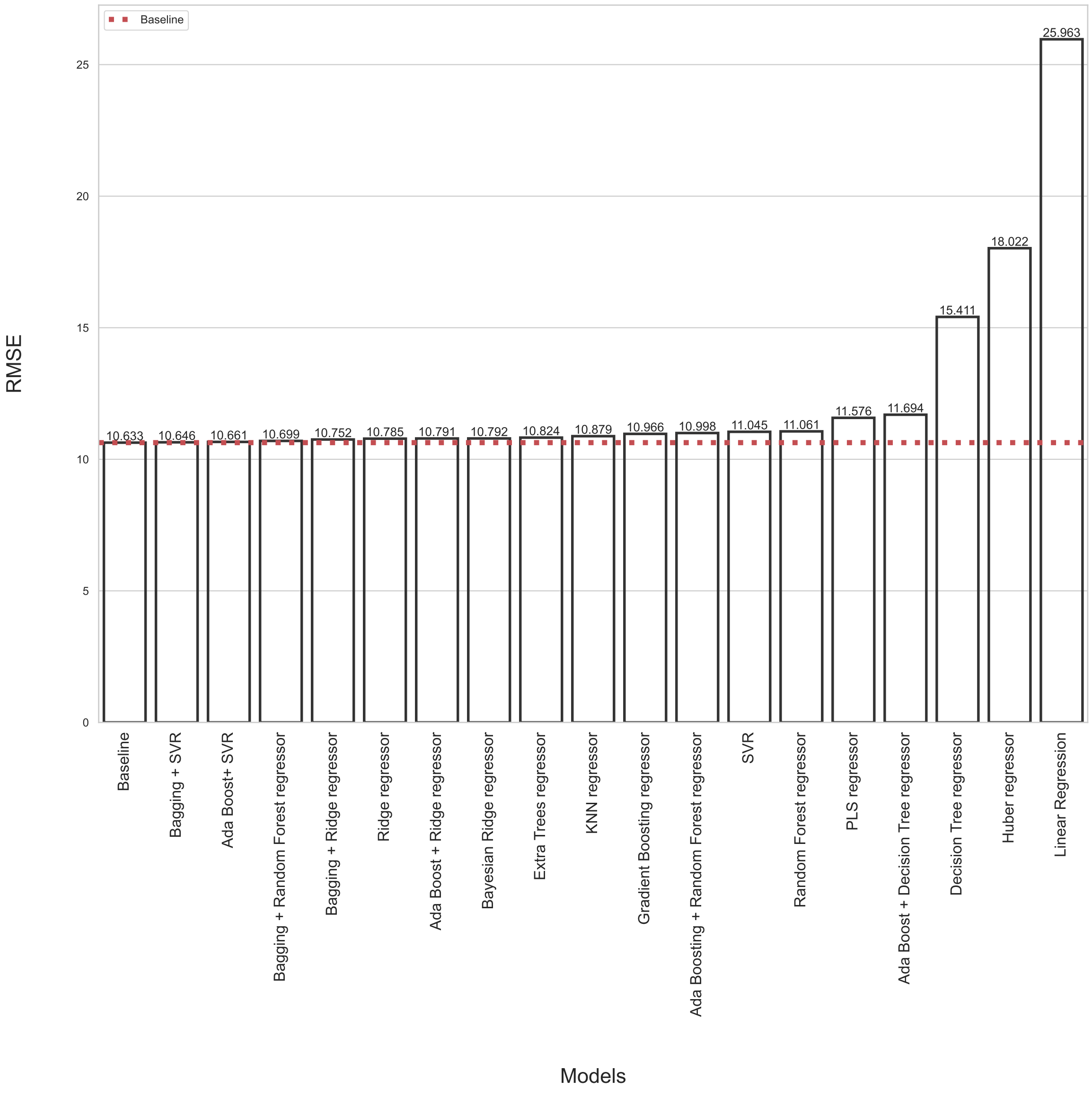


Figure 22: Bar chart showing the performance of regression models that were used with the forth set of features, sorted in ascending order (lower RMSE is better)

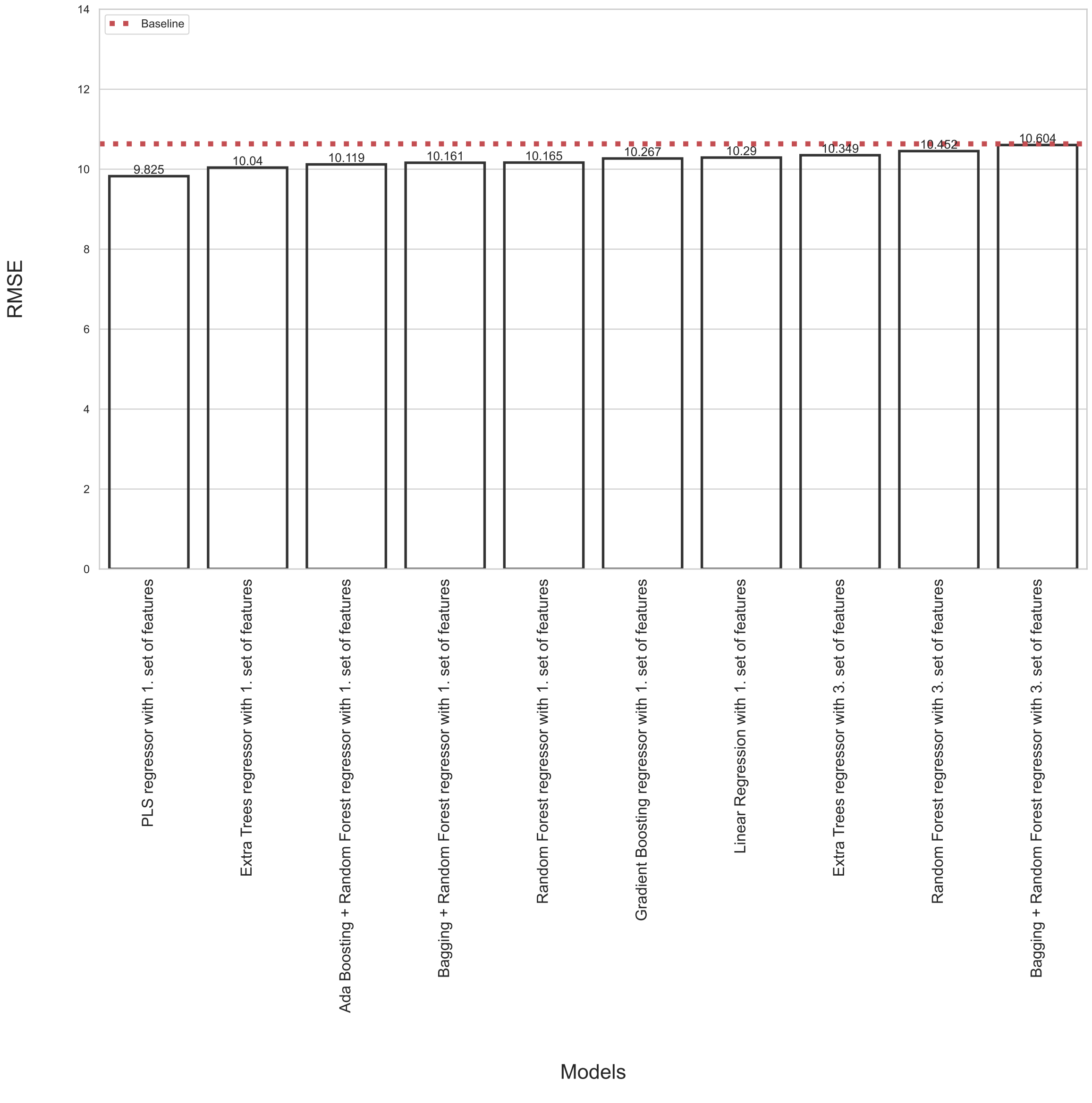


Figure 23: Bar chart showing the performance of top 10 regression models regardless of used set of features, sorted in ascending order (lower RMSE is better)

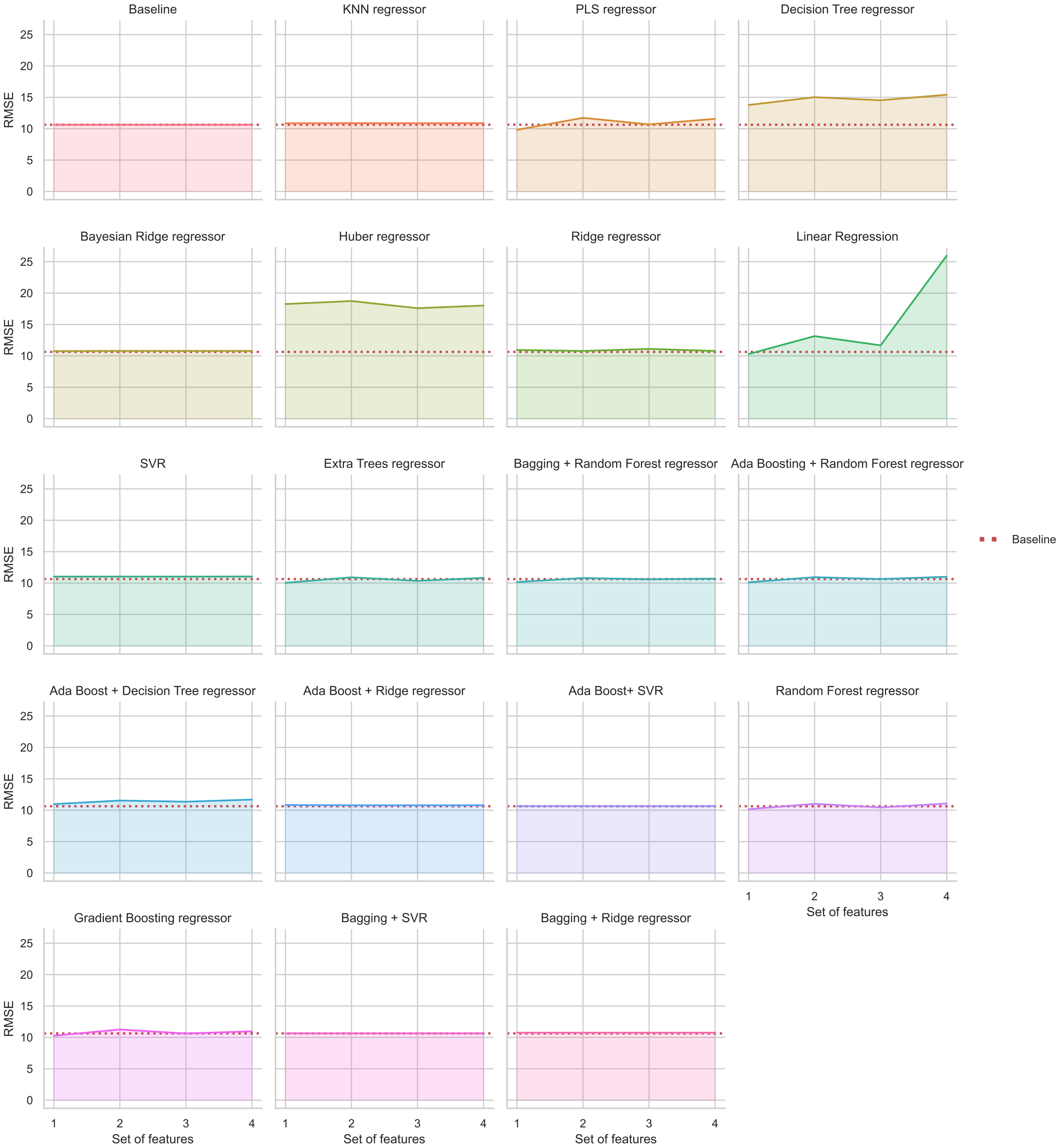


Figure 24: Performance of each regression model across different set of features, every facet is a different model

5 DISCUSSION

The goal of this study was to investigate if gullibility can be predicted from the social media traces of Twitter users. Besides this, we tested our hypothesis that high gullibility is correlated to low financial literacy, with the analogy that financial scam victims might be more gullible due to their lack of financial expertise. Alongside financial literacy and gullibility questions, we included emotionality, sense of self, general trust, cognitive reflection, and basic demographic questions in the survey. Next, we have generated three sets of features from the users' Twitter profiles: LIWC, fastText, and basic profile metrics and punctuation. Finally, in combination with survey answers, we created four different sets of features that we used to evaluate classification and regression models. By analyzing the results covered in chapter 4. we report the following findings.

In the correlational matrix displayed in Fig. 12 we can notice that features from the financial literacy questionnaire, subjective financial knowledge, and financial attitude were both negatively correlated to the gullibility score. However, this does not mean that all the gullible people are financially illiterate, yet people with lower financial literacy scores are more likely to have higher gullibility scores. Besides this, we recorded a positive correlation between emotionality and gullibility and a weak sense of self and high gullibility, which was also done in other studies [11]. From the survey, the least correlated feature to the gullibility score was general trust, which was also expected, bearing in mind that other studies on gullibility failed to report a correlation between those two features.

With regard to the model's performance, we can report that we were more successful using the classification approach. It is hard to compare classification and regression approaches because there are many factors that we need to take into account. But if we take the total number of models that were better than their respective baseline, the classification approach has clearly achieved better performance. Furthermore, other classification metrics followed the accuracy results without a significant number of anomalies suggesting that models are also stable. Tables with results that include all classification metrics are displayed in appendix D. Nevertheless, our findings indicate that gullibility can be predicted from the social media traces, but also that it requires further research.

5.1 LIMITATION OF THE CURRENT STUDY

From the age-gender distribution displayed in figure 11 we can see that genders were not balanced across all age groups. Additionally, we can see that there is only one data point in the age group of 65+ years. This requires our attention since other research on gullibility emphasized that elderly people belong to the group of extremely gullible people. Thus we might have missed recruiting participants with extremely high gullibility scores.

In Figs. 18 and 24, we can see the performance of the classification and regression models using four different sets of features. We can also see the model's results grouped by the input set of features in figures 13, 14, 15, 16, 19, 20, 21, 22 and tables 2, 3, 4, 5, 6, 7, 8, 9. These results suggest that in the classification task, the best feature set was the third one. Whereas in the regression approach, there was no clear winner. Furthermore, the elements that raised the accuracy of classification models the most were the LIWC features. On the other hand, the fastText vectors seem to add more noise to the data, but that could also be the case because of a small sample size of 159 participants. FastText vectors had a large number of features (300) which naturally requires a larger number of data points to support it. We noticed a sllightly better performance for regression models when we used the first and third feature set, suggesting that survey answers were the most influential segment for this approach. However, both RMSE and MAE of the best regression models were very close to the baseline, so we cannot confidently say that we managed to predict the gullibility score using the regression approach. The classification approach shows undoubtingly a greater potential for further research.

To summarize, the limitations of our experiment could be the small sample size and the possibility that we did not manage to recruit the participants with high gullibility scores. There is a chance that the highly gullible people are not using Twitter or they have protected (private) Twitter accounts.

5.2 FUTURE WORK

For future research, we suggest testing this or a similar experiment pipeline on different social media platforms, having a larger sample size, and using a more up-to-date financial literacy questionnaire containing questions regarding cryptocurrencies. Additionally, we encourage modifying the classification task from a binary to a multi-class problem. Finally, we believe that we do not need to predict the exact value of the gullibility score for our model to be effective. But we also think that different levels of gullibility need to be properly addressed so that users can get adequate treatment.

6 CONCLUSION

Until recently human gullibility was hard to measure, it required a lot of resources and it was not efficient. Fake news, the spread of misinformation, and financial scams alerted the researchers about the importance of preventing gullible outcomes. Combining the usage of psychology models, machine learning algorithms, and social media data, we managed to address this problem. With this experiment, we made a first step toward predicting gullibility from user's online behavior. We successfully built an experiment pipeline capable of classifying users by their gullibility levels. Over 15 classification models were better than the baseline with 10 of them achieving an accuracy of over 60%. Regression models were not as successful but we did manage to achieve results better than the baseline model as well. We compared different feature engineering techniques LIWC and fastText, and found LIWC to be more useful in classification tasks. In the end, we addressed our limitations, mainly the lack of data, and suggested improvements for future research on this topic.

7 DALJŠI POVZETEK V SLOVENSKEM JEZIKU

V kolikor je študentu v skladu s pravili fakultete odobrena priprava magistrskega dela v angleškem jeziku, mora študent pripraviti povzetek dela v obsegu od 4.000 do 10.000 znakov (s presledki) v slovenskem jeziku. Povzetek v slovenskem jeziku je zadnje poglavje magistrskega dela ter je ustrezno oštevilčeno (pred poglavjem Literatura in viri).

Cilj tega magistrskega dela je bil raziskati, ali je mogoče na podlagi spletnega vedenja uporabnikov Twitterja napovedati njihovo lahkovernost. Opravili smo pregled literature o lahkovernosti in ugotovili, da so razredi ljudi, ki so zaradi lahkovernosti še posebej izpostavljeni izkoriščanju, otroci, starejši in osebe z motnjami v razvoju. Ugotovitve drugih raziskovalcev kažejo, da obstaja povezava med čustvenostjo in lahkovernostjo ter šibkim občutkom lastne vrednosti in lahkovernostjo. Poleg tega naj bi bila lahkovernost zelo kontekstualna in odvisna od mikrosituacije. Trenutno so najpogostejše manifestacije lahkovernosti v spletnem okolju širjenje dezinformacij (lažne novice), politično izkoriščanje ter finančne in romantične prevare. Ker je lovljenje prevarantov in odstranjevanje vira napačnih informacij pogosto zelo zahtevno, menimo, da bi nam lahko prepoznavanje zelo naivnih posameznikov in uporaba ustrezne obravnave pri njih pomagala v boju proti vstajniškim težavam. Da bi se tega lotili, smo razvili orodje, ki predvideva naivnost na podlagi vedenja v družabnih medijih na Twitterju.

Najprej smo izdelali predhodno študijo, v kateri smo preizkusili in potrdili našo raziskavo. Vprašalnik smo posodobili, da je bil časovno učinkovitejši, in nadaljevali z glavno študijo. V glavni študiji je sodelovalo 159 uporabnikov Twitterja (81Ž, 72M, 6O), pri čemer je večina udeležencev (103) spadala v starostno skupino od 21 do 40 let. Raziskava je vsebovala vprašanja o lahkovernosti, občutku lastne vrednosti, čustvenosti, finančni pismenosti, kognitivni refleksiji, zaupanju in demografskih podatkih. Za merjenje lahkovernosti smo uporabili vnaprej pripravljeno in preizkušeno 12-stopenjsko samoocenjevalno lestvico lahkovernosti, iz katere smo izračunali oceno lahkovernosti. Ocena lahkovernosti je v razponu od 12 (nizka lahkovernost) do 84 (visoka lahkovernost), vendar nam je v našem poskusu uspelo zabeležiti le vrednosti od 12 do 60. Druge lestvice in posamezna vprašanja iz ankete smo uporabili za ponovitev ugotovitev iz prejšnjih študij ter za preverjanje naše hipoteze. Hipoteza je, da bi lahko bila finančna pismenost negativno povezana z lahkovernostjo, glede na analogijo, da bodo posamezniki, ki imajo več praktičnega in na terenu usmerjenega znanja s področja financ, manj nagnjeni k temu, da bodo nasedli finančnim prevaram, ki so največji primer lahkovernega vedenja v resničnem svetu.

Poleg socialne ankete smo zbrali javno dostopne podatke iz profilov udeležencev na Twitterju z uporabo vmesnika Twitter API. Zbrali smo vse osnovne značilnosti profila in tvite uporabnikov. Besedilne podatke iz tvitov smo dodatno očistili in obdelali s standardnimi tehnikami NLP (tokenizacija, odstranjevanje stop besed in lemmatizacija). Za postopek oblikovanja značilnosti smo uporabili programa LIWC in fastText.

Ko so bili podatki očiščeni in predobdelani, smo nato ustvarili štiri različne nabore značilnosti, ki smo jih uporabili kot vhodne podatke za naše napovedne modele. To smo storili, da bi primerjali in se odločili, kateri nabor značilnosti je najboljši za napovedovanje naivnosti uporabnikov Twitterja. Ker smo na to temo predhodno opravljali raziskovalno raziskavo, smo uporabili tako klasifikacijski kot regresijski pristop. Skupino modelov smo preizkusili s štirimi različnimi nabori značilnosti. Ker je bil naš nabor podatkov majhen, smo se odločili, da podatke razdelimo s tehniko gnezdenega navzkrižnega preverjanja. Za klasifikacijsko nalogo smo predhodno izvedli mediano delitve ocene zanesljivosti in uporabili osnovni model, ki napoveduje najpogostejši razred iz učnega nabora. Za nalogo regresije smo rezultate primerjali z osnovnim modelom, ki vedno napoveduje srednjo vrednost iz učnega nabora. Metrike, ki smo jih uporabili za klasifikacijske modele, so bile natančnost, f1, odpoklic, natančnost in površina pod ROC, za regresijske modele pa smo uporabili RMSE in MAE. Glede na vrsto naloge smo optimizirali hiperparametre za natančnost in RMSE.

Obe skupini modelov sta dosegli boljše rezultate od svojih osnovnih modelov, vendar lahko poročamo, da smo bili uspešnejši pri uporabi klasifikacijskega pristopa. Rezultati kažejo, da je bila pri klasifikacijski nalogi najboljša tretja kombinacija funkcij (odgovori na anketo, osnovne funkcije Twitterja in funkcije LIWC), medtem ko pri regresijskem pristopu praktično ni jasnega zmagovalca. Nazadnje naše ugotovitve kažejo, da je naivnost mogoče napovedati na podlagi spletnega vedenja in da ima prihodnje delo na tem področju velik potencial. V prihodnjih raziskavah bi predlagali, da se ta ali podoben potek poskusa preizkusi na različnih platformah družabnih medijev, da se zagotovi večja velikost vzorca in da se naloga razvrščanja spremeni iz binarnega v večrazredni problem.

8 REFERENCES

[1] Michael C. Ashton and Kibeom Lee. The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91(4):340–345, 2009. ISSN 00223891. doi: 10.1080/00223890902935878. *(Cited on page 13.)*

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. *(Cited on pages 11 in 17.)*

[3] GC Cawley and NLC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, July 2010. ISSN 1532-4435. *(Cited on page 21.)*

[4] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2):317–332, 03 2014. ISSN 0021-9916. doi: 10.1111/jcom.12084. URL `https://doi.org/10.1111/jcom.12084`. *(Cited on page 10.)*

[5] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014. *(Cited on pages 10 in 11.)*

[6] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137, Aug. 2021. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14432`. *(Cited on page 10.)*

[7] Tao Ding, Warren K Bickel, and Shimei Pan. Social media-based substance use prediction. *arXiv preprint arXiv:1705.05633*, 2017. *(Cited on page 10.)*

[8] Judith M. Flury and William Ickes. Having a weak versus strong sense of self: The sense of self scale (SOSS). *Self and Identity*, 6(4):281–303, 2007. ISSN 1529-8868. doi: 10.1080/15298860601033208. *(Cited on page 13.)*

[9] Joseph P. Forgas and Roy F. Baumeister. *The social psychology of gullibility: Conspiracy theories, fake news and irrational beliefs.* Routledge, 2019. ISBN 9780367190149. *(Cited on page 1.)*

[10] Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, 2005. ISSN 08953309. doi: 10.1257/089533005775196732. *(Cited on page 13.)*

[11] Madeline S. George, Alessandra K. Teunisse, and Trevor I. Case. Gotcha! Behavioural validation of the Gullibility Scale. *Personality and Individual Differences*, 162(February):110034, 2020. ISSN 01918869. doi: 10.1016/j.paid.2020.110034. URL `https://doi.org/10.1016/j.paid.2020.110034`. *(Cited on pages 1, 3, 9, 10, 13 in 45.)*

[12] David Glodstein, Susan Glodstein, and Jim Fornaro. Fraud trauma syndrome: The victims of the bernard madoff scandal. *Journal of Forensic Studies in Accounting & Business*, 2, 01 2010. *(Cited on page 1.)*

[13] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156, 2011. doi: 10.1109/PASSAT/SocialCom.2011.33. *(Cited on pages 10 in 11.)*

[14] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 253–262, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302685. doi: 10.1145/1979742.1979614. URL `https://doi.org/10.1145/1979742.1979614`. *(Cited on page 10.)*

[15] Stephen Greenspan. *Chapter 5 Foolish Action in Adults with Intellectual Disabilities. The Forgotten Problem of Risk-Unawareness*, volume 36. Elesvier Inc., 1 edition, 2008. ISBN 9780123744760. doi: 10.1016/S0074-7750(08)00005-0. URL `http://dx.doi.org/10.1016/S0074-7750(08)00005-0`. *(Cited on pages VII, 1, 3, 4, 6, 7 in 8.)*

[16] Stephen Greenspan. *Annals of gullibility: Why we get duped and how to avoid it.* Praeger, 2008. ISBN 9780313362163. *(Cited on page 1.)*

[17] Stephen Greenspan, Gail Loughlin, and Rhonda S Black. Credulity and gullibility in people with developmental disorders: A framework for future research. In *International review of research in mental retardation*, volume 24, pages 101–135. Elsevier, 2001. *(Cited on pages VII, 3, 4, 5 in 9.)*

[18] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013. doi: 10.1073/pnas.1218772110. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1218772110`. *(Cited on page 10.)*

[19] Hugo Mercier. How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology*, 21(2):103–122, 2017. ISSN 10892680. doi: 10.1037/gpr0000111. *(Cited on page 3.)*

[20] Gian Paolo Stella, Umberto Filotto, and Enrico Maria Cervellati. A Proposal for a New Financial Literacy Questionnaire. *International Journal of Business and Management*, 15(2):34, 2020. ISSN 1833-3850. doi: 10.5539/ijbm.v15n2p34. *(Cited on page 13.)*

[21] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001, 2001. *(Cited on page 11.)*

[22] Julian B. Rotter. Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35(1):1–7, 1980. doi: 10.1037/0003-066x.35.1.1. URL `https://doi.org/10.1037/0003-066x.35.1.1`. *(Cited on pages 3 in 7.)*

[23] Matthew J. Salganik. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press, 2017. ISBN 0691158649,9780691158648. *(Cited on page 10.)*

[24] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16, 09 2013. doi: 10.1371/journal.pone.0073791. URL `https://doi.org/10.1371/journal.pone.0073791`. *(Cited on page 10.)*

[25] Alessandra K. Teunisse, Trevor I. Case, Julie Fitness, and Naomi Sweller. I Should Have Known Better: Development of a Self-Report Measure of Gullibility. *Personality and Social Psychology Bulletin*, 46(3):408–423, 2020. ISSN 15527433. doi: 10.1177/0146167219858641. *(Cited on pages 1, 3, 9, 11, 13 in 14.)*

[26] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):178–185, May 2010. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14009`. *(Cited on pages 10 in 11.)*

[27] Toshio Yamagishi, Masako Kikuchi, and Motoko Kosugi. Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, 2(1):145–161, 1999. doi: https://doi.org/10.1111/1467-839X.00030. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-839X.00030`. *(Cited on pages 1, 3, 7, 8 in 9.)*

Appendices

APPENDIX A Pre-study survey

Question 1: I guess I am more gullible than the average person.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 2: If anyone is likely to fall for a scam, it's me.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 3: My friends think I'm easily fooled.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 4: My family thinks I am an easy target for scammers.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 5: People think I'm a little naïve.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 6: Overall, I'm pretty easily manipulated.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 7: I'm pretty good at working out when someone is trying to fool me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 8: I'm not that good at reading the signs that someone is trying to manipulate me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 9: I'm pretty poor at working out if someone is tricking me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 10: It usually takes me a while to 'catch on' when someone is deceiving me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 11: I'm usually quick to notice when someone is trying to cheat me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 12: I quickly realize when someone is pulling my leg

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 13: A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?

Question 14: If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

Question 15: In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Question 16: If three elves can wrap three toys in hour, how many elves are needed to wrap six toys in 2 hours?

Question 17: Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class?

Question 18: In an athletics team, tall members are three times more likely to win a medal than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?

Question 19: Please rate your opinion on the following scale:

In general, most people can be trusted	1 2 3 4 5 6 7	You can't be too careful in dealing with people.

Question 20: I wish I were more consistent in my feelings

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 21: It's hard for me to figure out my own personality, interests, and opinions

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 22: I often think how fragile my existence is

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 23: I have a pretty good sense of what my long-term goals are in life

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 24: I sometimes wonder if people can actually see me

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 25: Other people's thoughts and feelings seem to carry greater weight than my own

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 26: I have a clear and definite sense of who I am and what I'm all about

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 27: It bothers me that my personality doesn't seem to be well-defined

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 28: I'm not sure that I can understand or put much trust in my thoughts and feelings

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 29: Who am I? is a question that I ask myself a lot

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 30: I need other people to help me understand what I think or how I feel

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 31: I tend to be very sure of myself and stick to my own preferences even when the group I am with expresses different preferences.

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 32: I would feel afraid if I had to travel in bad weather conditions.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 33: I sometimes can't help worrying about little things.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 34: When I suffer from a painful experience, I need someone to make me feel comfortable.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 35: I feel like crying when I see other people crying.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 36: When it comes to physical danger, I am very fearful.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 37: I worry a lot less than most people do.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 38: I can handle difficult situations without needing emotional support from anyone else.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 39: I feel strong emotions when someone close to me is going away for a long time.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 40: Even in an emergency I wouldn't feel like panicking.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 41: I remain unemotional even in situations where most people get very sentimental.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 42: Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

- More than $102
- Exactly $102
- Less than $102
- Do not know
- Refuse to answer

Question 43: Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?

- More than today
- Exactly the same
- Less than today
- Do not know
- Refuse to answer

Question 44: Please tell me whether this statement is true or false. “Buying a single company’s stock usually provides a safer return than a stock mutual fund”.

- True
- False
- Do not know
- Refuse to answer

Question 45: Please tell me whether this statement is true or false. “A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest over the life of the loan will be less.”

- True
- False
- Do not know
- Refuse to answer

Question 46: If interest rates rise, what typically happens to bond prices?

- They false
- They rise
- They stay the same
- There is no relationship between bond prices and interest rates
- Do not know
- Refuse to answer

Question 47: You moved to a city where the cost of living is one-third higher than where you used to live. For the same salary, how will you be able to keep your savings ratio constant?

- Increasing purchases by 1/3
- Decreasing purchases by 1/3
- Decreasing purchases by 2/3
- Do not know

- Refuse to answer

Question 48: You have recently become a parent. You would like to find a solution that would allow your family to have more economic peace of mind in case something happens to you; what do you do?

- Buy a house by taking out a mortgage
- Buy shares in a company
- Subscribe an insurance policy
- Do not know
- Refuse to answer

Question 49: You have decided to invest 10,000€ in financial assets. You are offered three different funds; which fund would you choose? [Level 1 indicate low risk, level 5 medium risk and level 9 high risk]

- Asset A: 2% return, risk level 3
- Asset B: 4% return, risk level 3
- Asset C: 5% return, risk level 9
- Do not know
- Refuse to answer

Question 50: You have the opportunity to invest 20,000€. You are a risk-averse person and have a long-term investment horizon. Which investment do you think is the closest to your needs?

- Investment in Bitcoin
- Investment in government bonds
- Investment in derivatives
- Do not know
- Refuse to answer

Question 51: You have just turned 42, and your company is in a bad economic condition. Fortunately, you won a lottery prize of 200,000€.

- Using 90% of the amount to fulfill my long-desired wishes and save the remaining 10%Using 90% of the amount to fulfill my long-desired wishes and save the remaining 10%
- Using 30% for my wishes, Using 40% for a supplementary pension plan and 30% for savings
- Using 70% of the amount for my wishes, and Using 30% for savings
- Do not know
- Refuse to answer

Question 52: Before buying something I ask myself if I have paid my necessary expenses

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 53: Before buying something, I compare prices.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 54: Before signing a financial contract, I carefully read its contents.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 55: I am careful to distinguish between necessary and unnecessary expenses.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 56: Before making a major purchase, I make sure that my savings are sufficient to cover any sudden expense.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 57: The first thought I have when I borrow money is that I want to return the money on time.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 58: If I know the costs I will have to incur tomorrow, I'll think about it today.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 59: Before making online payments, I concern about the security of my data.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 60: Overall, thinking of your assets, debts, and savings, how satisfied are you with your current personal financial condition?

Not at all satisfied 1 2 3 4 5 6 7 Extremely satisfied

Question 61: How would you use your overall financial knowledge?

Very low 1 2 3 4 5 6 7 Very high

Question 62: In which age group do you belong?

- up to 20 years of age
- 21 - 40 years of age
- 41 - 60 years of age
- 61 years of age or more

Question 63: What is your gender?

- Male
- Female
- I do not wish to say

Question 64: What country do you currently reside in?

Question 65: What is the highest level of education you have completed?

- Primary/Elementary school
- High school
- Trade qualification or Certificate (e.g., carpentry, hairdressing)
- Diploma
- Some university
- Bachelor degree
- Postgraduate degree

APPENDIX B Main study survey

Question 1: I guess I am more gullible than the average person.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 2: If anyone is likely to fall for a scam, it's me.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 3: My friends think I'm easily fooled.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 4: My family thinks I am an easy target for scammers.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 5: People think I'm a little naïve.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 6: Overall, I'm pretty easily manipulated.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 7: I'm pretty good at working out when someone is trying to fool me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 8: I'm not that good at reading the signs that someone is trying to manipulate me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 9: I'm pretty poor at working out if someone is tricking me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 10: It usually takes me a while to 'catch on' when someone is deceiving me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 11: I'm usually quick to notice when someone is trying to cheat me

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 12: I quickly realize when someone is pulling my leg

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Question 13: A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?

Question 14: In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Question 15: In an athletics team, tall members are three times more likely to win a medal than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?

Question 16: Please rate your opinion on the following scale:

In general, most people can be trusted 1 2 3 4 5 6 7 You can't be too careful in dealing with people.

Question 17: I often think how fragile my existence is.

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 18: I sometimes wonder if people can actually see me.

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 19: I have a clear and definite sense of who I am and what I'm all about.

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 20: It bothers me that my personality doesn't seem to be well-defined

Very uncharacteristic of me 1 2 3 4 Very characteristic of me

Question 21: When I suffer from a painful experience, I need someone to make me feel comfortable.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 22: I feel like crying when I see other people crying.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 23: I can handle difficult situations without needing emotional support from anyone else.

Strongly disagree 1 2 3 4 5 Strongly agree

Question 24: You have recently become a parent. You would like to find a solution that would allow your family to have more economic peace of mind in case something happens to you; what do you do?

- Buy a house by taking out a mortgage
- Buy shares in a company
- Subscribe an insurance policy
- Do not know
- Refuse to answer

Question 25: You have the opportunity to invest 20,000€. You are a risk-averse person and have a long-term investment horizon. Which investment do you think is the closest to your needs?

- Investment in Bitcoin
- Investment in government bonds
- Investment in derivatives
- Do not know
- Refuse to answer

Question 26: Before buying something I ask myself if I have paid my necessary expenses

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 27: Before buying something, I compare prices.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 28: Before signing a financial contract, I carefully read its contents.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 29: I am careful to distinguish between necessary and unnecessary expenses.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 30: Before making a major purchase, I make sure that my savings are sufficient to cover any sudden expense.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 31: The first thought I have when I borrow money is that I want to return the money on time.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 32: If I know the costs I will have to incur tomorrow, I'll think about it today.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 33: Before making online payments, I concern about the security of my data.

Completely disagree 1 2 3 4 5 6 7 Completely agree

Question 34: Overall, thinking of your assets, debts, and savings, how satisfied are you with your current personal financial condition?

Not at all satisfied 1 2 3 4 5 6 7 Extremely satisfied

Question 35: How would you use your overall financial knowledge?

Very low 1 2 3 4 5 6 7 Very high

Question 36: In which age group do you belong?

- up to 20 years of age
- 21 - 40 years of age
- 41 - 60 years of age
- 61 years of age or more

Question 37: What is your gender?

- Male
- Female
- I do not wish to say

Question 38: What country do you currently reside in?

Question 39: What is the highest level of education you have completed?

- Primary/Elementary school
- High school
- Trade qualification or Certificate (e.g., carpentry, hairdressing)
- Diploma
- Some university
- Bachelor degree
- Postgraduate degree

APPENDIX C Hyper-parameters

1. Logistic regression

   - penalty=(”l1”, ”l2”, ”elasticnet”, ”none”)
   - solver= (”newton-cg”, ”lbfgs”, ”liblinear”, ”sag”, ”saga”)
   - tol= (1e-5,1e-4, 1e-3, 1e-2,1e-1)
   - C=(0.1, 1, 2, 3, 5, 10)

2. Gaussian naive Bayes classifier:

   - var_smoothing= (1e-8, 1e-9, 1e-10, 1e-11)

3. K nearest neighbors classifier:

   - n_neighbors=(3, 4, 5, 10)
   - weights= (”uniform”, ”distance”)
   - algorithm= (”auto”, ”ball_tree”, ”kd_tree”, ”brute”)
   - leaf_size=(100, 50, 40, 30, 20, 10, 5)
   - p=(1, 2, 3)

4. SVC (SVM classifier):

   - C=(1, 2, 3, 5)
   - kernel= (”rbf”)
   - degree= (2, 3)
   - gamma=(”scale”, ”auto”)
   - tol=(1e-3, 1e-4)

5. Gradient boosting classifier:

   - n_estimators=(10, 100, 250)
   - learning_rate= (0.01, 0.1, 1)
   - criterion= (”friedman_mse”, ”squared_error”)
   - min_samples_split=(2)
   - min_samples_leaf=(1)
   - max_features=(”auto”)
   - tol=(1e-5, 1e-4)

- max_depth=(1, 2, 3, 4,5)

6. Random forest classifier:

   - bootstrap=(True)
   - max_depth= (1, 2, 3, 4)
   - max_features= (”auto”, ”sqrt”)
   - min_samples_leaf=(1, 2)
   - min_samples_split=(2, 5)
   - n_estimators=(10, 50, 100, 200])

7. Extra trees classifier:

   - criterion=(”gini”, ”entropy”)
   - n_estimators= (100)
   - max_features= (”auto”, ”sqrt”, ”log2”)
   - bootstrap=(”True”, ”False”)

8. Decision tree classifier:

   - criterion = (”gini”, ”entropy”)
   - splitter = (”best”, ”random”)
   - max_features = (auto”, ”sqrt”, ”log2)

9. Stochastic gradient descent classifier:

   - penalty=(”l2”, ”l1”, ”elasticnet”)
   - alpha= (0.01, 0.1, 0.0001, 0.00001, 0.001)
   - max_iter= (500, 1000, 1200, 1500, 2000)
   - tol=(1e-3, 1e-2, 1e-4, 1e-5)

10. Ridge classifier:

    - solver=(”auto”, ”svd”, ”cholesky”, ”lsqr”, ”sparse_cg” ,”sag”, ”saga”, ”lbfgs”)
    - alpha= (1,2, 0.1, 0.01)
    - max_iter= (100, 250, 500, 1000, 1200, 1500, 2000, None)
    - tol=(1e-3, 1e-2, 1e-4, 1e-5)

11. Multi-layer perceptron classifier:

    - hidden_layer_sizes=((5,2), (7,5))
    - solver= (”lbfgs”, ”sgd”, ”adam”)
    - alpha= (1e-3, 1e-5)

12. Majority voting:

    - estimators=(
      MLPClassifier(alpha= 0.001, hidden_layer_sizes= (7, 5), solver= 'adam', random_state=1),

      RidgeClassifier(alpha= 1, max_iter= 100, solver= 'sag', tol= 0.01, random_state= 1),

      SGDClassifier(alpha= 0.001, max_iter= 100, penalty= 'l2', tol= 0.001, random_state=1),

      DecisionTreeClassifier(criterion= 'entropy', max_features= 'auto', splitter= "random", random_state= 1),

      LogisticRegression(C= 1, penalty= 'l2', solver= 'newton-cg', tol= 0.1, random_state= 1),

      KNeighborsClassifier(algorithm= 'auto',leaf_size= 1000, n_neighbors= 6, p= 1, weights= 'distance')
)

13. Ada boost classifier:

    - base_estimator=(
      RidgeClassifier(alpha= 1, max_iter= 100, solver= 'sag', tol= 0.01, random_state= 1),

      SGDClassifier(alpha= 0.001, max_iter= 100, penalty= 'l2', tol= 0.001,random_state= 1),

      DecisionTreeClassifier(criterion= 'entropy', max_features= 'auto', splitter= "random", random_state= 1),

      RandomForestClassifier(bootstrap= True, max_depth= 1, max_features= 'auto', min_samples_leaf= 1, min_samples_split= 2, n_estimators= 10, random_state= 1),

      LogisticRegression(C= 1, penalty= 'l2', solver= 'newton-cg', tol= 0.1, random_state= 1),

GaussianNB(var_smoothing= 1e-08)
)

- n_estimators=(10, 100, 250, 500)
- learning_rate=(0.01, 0.1, 1, 3, 5, 10, 100)
- algorithm= ("SAMME")

14. Bagging classifier:

    - base_estimator= (LogisticRegression(C= 1, penalty= 'l2', solver= 'newton-cg', tol= 0.1, random_state= 1))

15. KNN regressor:

    - weights= ("uniform", "distance")
    - algorithm= ("ball_tree", "auto", "kd_tree","brute")
    - n_neighbors= (3, 5, 10),
    - p= (1,2)

16. PLS regressor:

    - n_components= (2, 3, 4, 5)
    - scale= ("True", "False"),
    - max_iter= (250, 500, 1000),
    - tol= (1e-06, 1e-05, 1e-07),
    - copy= ("True", "False")

17. Decision tree regressor:

    - criterion= ("squared_error", "frieman_mse", "absolute_error", "poisson")
    - splitter = ("best", "random")
    - max_features = ("auto", "sqrt", "log2")

18. Bayesian Ridge regressor:

    - No hyper-parameters

19. Huber regressor:

    - max_iter= (100, 200, 300)
    - tol= (1e-3, 1e-2, 1e-4, 1e-5),
    - alpha= (1e-01, 1e-02, 1, 10)

20. Ridge regressor:

    - max_iter= (25, 50, 100, 200)
    - tol= (1e-3, 1e-4, 1e-5)
    - alpha= (1e-01, 1e-02, 1, 10)
    - solver= (”auto”, ”svd”, ”cholesky”, ”lsqr”, ”sparse_cg”, ”sag”, ”saga”, ”lbfgs”)

21. Linear regression:

    - No hyper-parameters

22. SVR:

    - kernel= (”rbf”)
    - degree= (1, 2, 3)
    - gamma= (”scale”, ”auto”)
    - tol= (1e-3, 1e-2, 0.1, 1, 10)
    - C= (1e-03, 1e-01, 1e-02, 1, 10, 25, 100)

23. Random Forest regressor:

    - n_estimators= (50, 100, 200, 300)
    - max_features= (”auto”,”sqrt”,”log2”)
    - bootstrap= (”True”, ”False”)

24. Gradient boosting regressor:

    - n_estimators= (100, 200)
    - learning_rate= (1, 0.1, 0.01)
    - criterion= (”friedman_mse”, ”squared_error”, ”mse”,”mae”)
    - tol= (1e-4, 1e-5)
    - alpha= (1e-03, 1e-01, 1e-02, 1)
    - max_features= (”auto”, ”sqrt”, ”log2”)

25. Extra Trees regressor:

    - n_estimators= (100, 200, 500)
    - bootstrap= (”True”, ”False”)
    - max_features= (”auto”, ”sqrt”, ”log2”)

26. Bagging regressor:

    - base_estimator= (

      RandomForestRegressor(bootstrap = ’True’, max_features= ’sqrt’, n_estimators= 200, random_state= 1,

SVR(C=10, degree= 1, gamma= 'auto', kernel= 'rbf',
tol= 10),

Ridge(alpha= 10, max_iter= 25, solver= 'saga',
tol= 0.001, random_state=1)

)

- n_estimators= (10, 30),
- max_samples= (60, 100),
- bootstrap= ("True", "False"),

27. Ada boost + Random Forest regressor:

    - base_estimator= RandomForestRegressor(bootstrap = 'True', max_features= 'sqrt', n_estimators= 200, random_state= 1)
    - n_estimators= (10, 30)
    - learning_rate= (0.1, 0.01)
    - loss= ("linear")

28. Ada boost + Decision tree regressor:

    - base_estimator = DecisionTreeRegressor(criterion= 'friedman_mse', max_features= 'auto', splitter= 'random', random_state= 1)
    - n_estimators= (10, 30)
    - learning_rate= (0.1, 0.01)
    - loss= ("linear", "square", "exponential")

29. Ada boost + Ridge regressor:

    - base_estimator=Ridge(alpha= 10, max_iter= 25, solver= 'saga', tol= 0.001,random_state= 1)
    - n_estimators= (10, 30)
    - learning_rate= (0.1, 0.01)
    - loss= ("linear", "square", "exponential")

30. Ada boost + SVR regressor:

    - base_estimator=SVR(C=10, degree= 1, gamma= 'auto', kernel= 'rbf', tol= 10)
    - n_estimators= (10, 30, 50)
    - learning_rate= (0.1, 0.01, 0.001)
    - loss= ("linear", "square", "exponential")

APPENDIX D Detailed classification results

Table 10: Classification results with the first set of features

Model name	Set of features	Accuracy (mean)	Accuracy (std)	Precission (mean)	Precision (std)	F1 (mean)	F1 (std)	Recall (mean)	Recall (std)	AUC (mean)	AUC (std)
Baseline	1	0.516	0.014	0.516	0.014	0.680	0.012	1.000	0.000	0.500	0.000
Logistic Regression	1	**0.572**	0.076	0.593	0.069	0.581	0.062	0.575	0.075	0.625	0.055
Gaussian Naive Bayes classifier	1	**0.566**	0.026	0.547	0.019	0.693	0.011	0.951	0.047	0.633	0.052
KNN classifier	1	0.460	0.079	0.451	0.109	0.399	0.137	0.365	0.153	0.465	0.062
SVC	1	**0.579**	0.053	0.560	0.034	0.675	0.046	0.854	0.084	0.546	0.120
Gradient Boosting classifier	1	**0.522**	0.044	0.535	0.060	0.537	0.074	0.549	0.114	0.576	0.053
Random forest classifier	1	**0.566**	0.045	0.575	0.040	0.591	0.047	0.610	0.064	0.632	0.075
Decision Tree classifier	1	0.471	0.101	0.492	0.105	0.487	0.094	0.486	0.093	0.470	0.101
SGD classifier	1	0.491	0.046	0.457	0.333	0.300	0.283	0.364	0.397	0.480	0.082
Ridge classifier	1	**0.522**	0.036	0.527	0.028	0.558	0.079	0.608	0.157	0.595	0.042
Extra trees classifier	1	**0.542**	0.082	0.543	0.086	0.561	0.114	0.585	0.146	0.593	0.115
MLP classifier	1	0.491	0.044	0.506	0.049	0.577	0.120	0.753	0.303	0.494	0.036
Majority voting classifier	1	**0.591**	0.020	0.589	0.007	0.625	0.085	0.693	0.170	0.500	0.000
Bagging + Logistic Regression	1	**0.572**	0.054	0.591	0.043	0.567	0.067	0.551	0.100	0.624	0.066
Ada Boost + Gradient Boosting classifier	1	**0.566**	0.026	0.547	0.019	0.693	0.011	0.951	0.047	0.553	0.033
Ada Boost + Logistic Regression	1	**0.578**	0.059	0.590	0.052	0.597	0.059	0.610	0.089	0.612	0.053
Ada Boost + Random forest classifier	1	**0.572**	0.062	0.584	0.055	0.596	0.053	0.611	0.065	0.626	0.044
Ada Boost + SGD classifier	1	**0.573**	0.043	0.579	0.053	0.623	0.024	0.684	0.067	0.589	0.044
Ada Boost + Decision Tree classifier	1	0.516	0.044	0.534	0.050	0.538	0.031	0.548	0.055	0.513	0.045
Ada Boost + Ridge classifier	1	**0.585**	0.034	0.565	0.021	0.679	0.028	0.854	0.073	0.556	0.058

Table 11: Classification results with the second set of features

Model name	Set of features	Accuracy (mean)	Accuracy (std)	Precission (mean)	Precision (std)	F1 (mean)	F1 (std)	Recall (mean)	Recall (std)	AUC (mean)	AUC (std)
Baseline	2	0.516	0.014	0.516	0.014	0.680	0.012	1.000	0.000	0.500	0.000
Logistic Regression	2	**0.578**	0.052	0.606	0.065	0.566	0.059	0.537	0.075	0.632	0.058
Gaussian Naive Bayes classifier	2	**0.579**	0.014	0.624	0.042	0.529	0.071	0.475	0.116	0.654	0.042
KNN classifier	2	0.503	0.029	0.526	0.037	0.480	0.091	0.471	0.160	0.548	0.042
SVC	2	**0.553**	0.042	0.540	0.022	0.667	0.046	0.877	0.105	0.619	0.156
Gradient Boosting classifier	2	**0.603**	0.069	0.616	0.065	0.614	0.081	0.623	0.117	0.652	0.048
Random forest classifier	2	**0.604**	0.059	0.622	0.039	0.599	0.085	0.588	0.124	0.665	0.087
Decision Tree classifier	2	**0.541**	0.094	0.576	0.105	0.555	0.059	0.550	0.075	0.542	0.100
SGD classifier	2	**0.535**	0.029	0.471	0.246	0.450	0.241	0.500	0.342	0.530	0.096
Ridge classifier	2	**0.585**	0.077	0.603	0.065	0.598	0.061	0.599	0.083	0.637	0.106
Extra trees classifier	2	**0.636**	0.090	0.665	0.094	0.620	0.107	0.587	0.127	0.690	0.073
MLP classifier	2	0.515	0.073	0.527	0.055	0.607	0.071	0.745	0.181	0.503	0.087
Majority voting classifier	2	**0.578**	0.068	0.583	0.055	0.612	0.077	0.660	0.155	0.538	0.109
Bagging + Logistic Regression	2	**0.585**	0.059	0.598	0.040	0.588	0.079	0.588	0.132	0.647	0.080
Ada Boost + Gradient Boosting classifier	2	**0.541**	0.079	0.553	0.063	0.541	0.130	0.577	0.249	0.587	0.063
Ada Boost + Logistic Regression	2	**0.579**	0.024	0.601	0.034	0.570	0.052	0.549	0.086	0.640	0.037
Ada Boost + Random forest classifier	2	**0.585**	0.055	0.614	0.079	0.598	0.028	0.599	0.083	0.654	0.062
Ada Boost + SGD classifier	2	**0.553**	0.077	0.590	0.090	0.547	0.055	0.526	0.091	0.562	0.075
Ada Boost + Decision Tree classifier	2	**0.597**	0.037	0.609	0.049	0.608	0.051	0.609	0.068	0.596	0.038
Ada Boost + Ridge classifier	2	**0.566**	0.049	0.552	0.026	0.662	0.045	0.829	0.100	0.544	0.088

Table 12: Classification results with the third set of features

Model name	Set of features	Accuracy (mean)	Accuracy (std)	Precission (mean)	Precision (std)	F1 (mean)	F1 (std)	Recall (mean)	Recall (std)	AUC (mean)	AUC (std)
Baseline	3	0.516	0.014	0.516	0.014	0.680	0.012	1.000	0.000	0.500	0.000
Logistic Regression	3	**0.642**	0.046	0.658	0.020	0.642	0.064	0.637	0.113	0.692	0.072
Gaussian Naive Bayes classifier	3	**0.604**	0.050	0.637	0.039	0.570	0.087	0.524	0.121	0.709	0.029
KNN classifier	3	0.510	0.039	0.542	0.055	0.457	0.088	0.424	0.157	0.550	0.046
SVC	3	**0.553**	0.042	0.540	0.022	0.667	0.046	0.877	0.105	0.619	0.156
Gradient Boosting classifier	3	**0.592**	0.069	0.596	0.059	0.591	0.117	0.599	0.161	0.672	0.044
Random forest classifier	3	**0.616**	0.064	0.628	0.061	0.629	0.067	0.635	0.095	0.671	0.089
Decision Tree classifier	3	**0.534**	0.082	0.557	0.073	0.537	0.066	0.524	0.078	0.533	0.082
SGD classifier	3	**0.522**	0.023	0.424	0.213	0.535	0.268	0.741	0.388	0.579	0.119
Ridge classifier	3	**0.598**	0.081	0.612	0.070	0.616	0.064	0.623	0.076	0.659	0.121
Extra trees classifier	3	**0.629**	0.097	0.648	0.094	0.643	0.080	0.648	0.100	0.707	0.081
MLP classifier	3	0.497	0.035	0.430	0.225	0.447	0.237	0.540	0.362	0.512	0.091
Majority voting classifier	3	**0.610**	0.116	0.620	0.115	0.598	0.158	0.601	0.211	0.470	0.042
Bagging + Logistic Regression	3	**0.667**	0.080	0.684	0.061	0.669	0.080	0.660	0.112	0.707	0.069
Ada Boost + Gradient Boosting classifier	3	**0.553**	0.064	0.607	0.068	0.464	0.157	0.418	0.192	0.612	0.072
Ada Boost + Logistic Regression	3	**0.635**	0.053	0.651	0.038	0.633	0.074	0.623	0.114	0.671	0.049
Ada Boost + Random forest classifier	3	**0.591**	0.090	0.603	0.086	0.616	0.077	0.635	0.087	0.675	0.100
Ada Boost + SGD classifier	3	**0.540**	0.069	0.575	0.084	0.541	0.065	0.539	0.146	0.529	0.092
Ada Boost + Decision Tree classifier	3	**0.598**	0.073	0.601	0.086	0.599	0.119	0.607	0.156	0.595	0.074
Ada Boost + Ridge classifier	3	**0.566**	0.049	0.552	0.026	0.662	0.045	0.829	0.100	0.544	0.088

Table 13: Classification results with the forth set of features

Model name	Set of features	Accuracy (mean)	Accuracy (std)	Precission (mean)	Precision (std)	F1 (mean)	F1 (std)	Recall (mean)	Recall (std)	AUC (mean)	AUC (std)
Baseline	4	0.516	0.014	0.516	0.014	0.680	0.012	1.000	0.000	0.500	0.000
Logistic Regression	4	**0.579**	0.054	0.607	0.046	0.551	0.081	0.515	0.124	0.625	0.087
Gaussian Naive Bayes classifier	4	**0.585**	0.020	0.629	0.039	0.538	0.081	0.487	0.133	0.653	0.042
KNN classifier	4	0.503	0.029	0.526	0.037	0.480	0.091	0.471	0.160	0.548	0.042
SVC	4	**0.553**	0.042	0.540	0.022	0.667	0.046	0.877	0.105	0.618	0.157
Gradient Boosting classifier	4	0.510	0.053	0.522	0.043	0.561	0.073	0.625	0.142	0.500	0.072
Random forest classifier	4	0.491	0.094	0.503	0.087	0.521	0.110	0.553	0.152	0.512	0.106
Decision Tree classifier	4	**0.541**	0.037	0.554	0.034	0.571	0.039	0.599	0.091	0.541	0.036
SGD classifier	4	**0.535**	0.029	0.471	0.246	0.450	0.241	0.500	0.342	0.530	0.096
Ridge classifier	4	**0.573**	0.069	0.578	0.056	0.611	0.070	0.662	0.134	0.653	0.111
Extra trees classifier	4	**0.523**	0.118	0.524	0.095	0.555	0.139	0.603	0.193	0.528	0.143
MLP classifier	4	0.516	0.014	0.516	0.014	0.680	0.012	1.000	0.000	0.507	0.013
Majority voting classifier	4	**0.591**	0.045	0.586	0.031	0.647	0.056	0.745	0.161	0.507	0.013
Bagging + Logistic Regression	4	**0.598**	0.063	0.609	0.038	0.598	0.088	0.601	0.154	0.647	0.081
Ada Boost + Gradient Boosting classifier	4	**0.541**	0.079	0.553	0.063	0.541	0.130	0.577	0.249	0.587	0.063
Ada Boost + Logistic Regression	4	**0.566**	0.035	0.586	0.038	0.555	0.071	0.537	0.106	0.598	0.030
Ada Boost + Random forest classifier	4	0.504	0.101	0.512	0.094	0.530	0.125	0.566	0.184	0.517	0.121
Ada Boost + SGD classifier	4	**0.553**	0.077	0.590	0.090	0.547	0.055	0.526	0.091	0.562	0.075
Ada Boost + Decision Tree classifier	4	0.510	0.039	0.532	0.061	0.490	0.059	0.465	0.089	0.513	0.041
Ada Boost + Ridge classifier	4	**0.566**	0.049	0.552	0.026	0.662	0.045	0.829	0.100	0.544	0.088