



PREDICTING UK POLITICAL DONATIONS IN A BROADER CONTEXT USING MACHINE LEARNING

MARTIJN FAES

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2012824

COMMITTEE

prof. dr. Afra Alishahi
dr. Giacomo Spigler

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 24th, 2025

WORD COUNT

8799

ACKNOWLEDGMENTS

First I would like to thank the Office for National Statistics in the UK for making their Census data available for scientific research under the Open Government Licence v3.0.

I want to thank my supervisor for her guidance. I am also very grateful for my study advisor Shinta The for supporting me through difficulties I have been experiencing during this process. I would like to thank all Tilburg University staff who have played a role in me successfully completing this thesis project.

I would also like to thank my partner for all her love and support during this process. Thank you so much to my family and my partner's family for all their love and continued support.

Finally I would like to thank Alexander Marcus, Alligatoah, Rammstein, Gorb and Pedalem for creating the music that has helped keep me energized and focused throughout the final stretch of this process.

PREDICTING UK POLITICAL DONATIONS IN A BROADER CONTEXT USING MACHINE LEARNING

MARTIJN FAES

Abstract

Despite their impact on societal and political developments, research on political donations is underrepresented in scientific literature, especially so regarding studies using machine learning (ML) for their analyses. On top of this, when ML is used the ways it is implemented vary heavily across studies. This study investigates a broad range of ML methods in an attempt to create a larger consensus on appropriate ML implementations in the field. To achieve this, the predictive performances of Decision Trees, Random Forests, K-nearest neighbors, Support Vector Machine, and XGBoost algorithms are compared across both regression and classification tasks. In addition to this, some donation patterns regarding gender, racial distribution and donor types found in other studies are investigated on the largely unused UK political donations dataset combined with UK Census data. The analyses find optimal performance of the Random Forest and XGBoost algorithms when used in classification tasks. Further, differences between male and female donors were identified, suggesting higher donation spending and frequency patterns for male donors. Finally, individual people were found to be the largest donor group in amount and frequency, overshadowing contributions from other donor groups such as Trade Unions. The findings of this study promote citizen and official understanding of political donations, fortify the literature base of ML studies on political donations and contribute to a more substantially supported range of ML methods by comparing a larger amount of approaches at once.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

This study makes use of datasets from two different sources and owners. Firstly, there is the dataset containing information of and around donations

made to political entities in the UK valued £0 or above starting in 2001. This data is available publicly by law and is held and provided by the UK Electoral Commission on their website. The data contains data from natural persons, as well as organisations of many natures that have made donations. The data has not been anonymised, since it is legally required for these donations to be reported publicly and transparently. For this reason, consent to have your data recorded and published publicly is always implied and necessary when donating any amount of £500 or higher. Secondly, data from the UK 2021 Census is used. This data is gathered, owned and made available publicly by the Office for National Statistics. This data is anonymised and contains population information from natural persons in the UK, although often reported in the form of pooled data instead of entries for individuals. The Census data can be found through a link on the front page of the website of the Office for National Statistics.

All figures found in this study are made by the author using software listed or mentioned in this thesis. All libraries and frameworks used in this study can be found [through the author's GitHub repository](#) as well as in Appendix A (page 31). This includes version numbers and, the license of the code. Parts of the code from other people that were used and adapted for this study are noted in the notebook at the places where that person's code is used. The code used for this study can be found on the author's GitHub: <https://github.com/Martijns-projects/U.K.-political-donations-analysis>. The author used the online Thesaurus (<https://www.thesaurus.com/>) in rare cases to help paraphrasing. For spelling and grammar checking, only Microsoft Word's built-in software for this was used. The author has not used any tools or services to typeset the text in this work. All reference data was obtained from the source websites of the articles if possible and otherwise obtained through Google Scholar or manually formatted in accordance with the BibTeX format. Then, these references were added to the LaTeX template provided by Tilburg University. No other reference management software was used.

2 PROBLEM STATEMENT & RESEARCH GOAL

Transactions to political actors have been part of political landscapes for a long time. These transfers of resources take place both privately and in the public eye. For the purpose of fairness and regulation, governing bodies can state laws and regulations to document and publish these transactions. There is an upward trend in countries that require the registration and publication of transfers of most commonly financial resources, such as the United States with the Federal Election Campaign Act or the UK with the

Political Parties, Elections and Referendums Act 2000. There are financial resources of various natures available to political entities, which span from government funding to donations from private persons. In this study, the focus will be on the greatest financial contributions political entities receive, which are political donations of any kind.

Apart from government support and private individuals, there are many other sources of political donations, e.g. trusts, businesses and advocacy groups. Frequent donations from all sources account for a large amount of data that can be studied and learned from. In the current literature, very few quantitative studies with a focus on donation amount and factors related to it have been performed. The articles studying or related to this topic often do not make use of machine learning (from now: ML), opting instead for more traditional statistical approaches or other non-ML methods. The studies that do utilize ML methodology use a wide variety of ML models, making for a thinly established literature base for ML-based research of political donations. For this reason, this study is expected to strengthen and further establish the use of state-of-the-art ML methods in this sub-field of political finances. As will be highlighted in the Literature Review of this thesis, the UK provides a unique opportunity to do so, and also to contribute to a broader knowledge base on donation behaviour worldwide.

2.1 *Relevance*

Most studies on the patterns and prediction of political donations focus on a single or limited set of factors influencing the donation amount or characteristics of donors (some of these studies are included in the section *Literature Review*). This makes for a scattered view of many different factors affecting donations. Different studies focusing on closely related topics such as crowdfunding for business or charity also experience similar situations. This study attempts to create a framework in which many different features are used in predicting donation amount, using frequently successful models found in other studies. In addition to this, patterns found in studies related to the used features will be targeted in this study to see if similar results are found despite using a different method. Through these efforts, factors from different studies can possibly be synthesized, a larger overview of features influencing donation amount can be built, and a broader consensus can be created on appropriate ML models for use with donation predictions. Patterns found in previous studies can consequently be reinforced further or challenged. Not only will this advance literature

on political donations, adjacent fields such as the ones described previously can also benefit from the results of this study.

Further, this study offers benefits to both political entities and the public. Firstly, political entities might find this study's results interesting for guiding their approach to donation acquisition. This study could provide them with new information that will help them with better allocation of their resources, to maximize the effects of their campaigns and other actions. Secondly, the information presented in this study will provide a further level of transparency for the public on the origins of donations made to the entities that will end up affecting their lives on sometimes rather fundamental issues. This allows citizens to view political finances more critically and more consciously, and possibly even take action should they feel unhappy with the current insights. One thing that should be noted is that while this study tries to incorporate a larger range of features than previous studies, there are many factors unaccounted for in this research, that may affect the donations and the patterns found therein. Therefore, this study should not be seen as a complete view, rather it is an attempt to get better grip on political finance using ML methods.

2.2 *Strategy*

In order to study the large aggregation of features that are similar to features from related studies or approach them, data from UK political donations is combined with UK Census data. Though often not identically representing the features from other studies, an approach can be made by using features that hold data of a similar nature. The chosen data are then processed by and used for the training of five different generally high-performing ML models used in other studies and, based on whether they are regression or classification tasks, are reflected against a baseline Linear Regression or Logistic Regression model, respectively. For regression tasks, error analysis will be performed through RMSE and R^2 , so an accurate view can be provided on both error size and goodness-of-fit of the models. Classification task errors are analysed using accuracy. Only one metric is chosen here, since there are already a large number of models and configurations used in this study. Therefore, a more simplistic approach for the error analysis was chosen for the sake of the nature of the study and to keep the study within reasonable bounds. These steps and analyses will lead to the answer of the main research question (RQ) of this study:

Out of five pre-selected models (CART, random forest, XGBoost, SVM, KNN), which ML model performs best in predicting expected donation amounts for specific donor groups?

In order to further explore the many facets of the main RQ, the following sub-RQs are studied. After model performances are obtained and analysed, the best performing model from this study will then be used to answer sub questions 2 and 3:

RQ1 *How do the results of regression and classification analysis compare to one another?*

RQ2 *Are there noticeable differences by gender and ethnicity in donation amount and frequency, and how are these reflected in predicted versus real data?*

RQ3 *What are the predicted donation amount patterns in different donor types (public fund, individual, etc.) and how do these differ from each other?*

Through these sub-RQs, the effectiveness of use of either classification and regression models in political funding studies can be justified and verified [RQ1], providing further insights in appropriate ML methods. Different features that can be analysed from donor data (name, ethnic distribution on the level of local areas) can be used [RQ2] to provide further insights into relevant features for donation research. Lastly, various patterns might be found in terms of sources of funding [RQ3], further mapping out the landscape of financial resourcing for political ends.

2.3 Findings

Out of all models present in this study, it was found that the Random Forest and XGBoost models performed best. This is consistent with Liu and Hu (2024) and supports the idea that these models are very suitable for classification tasks, more so than regression tasks in this field. The models show around 43% accuracy when predicting on various categorical features not including regional data and show significant gender differences. Looking further into classification task performance, all other models exceed baseline performance but show little variation among each other. Individuals are the dominant group of donors, often donating amounts on the higher end of the measured value ranges. More publicly oriented organizations follow in donation frequency, showing different behaviour in spending amount than individuals.

3 LITERATURE REVIEW

3.1 *Exploring relevant features*

Globally, an increasing number of frameworks for transparency of political financing are adopted into legislation. There are many reasons for this, e.g. increased trust in political parties (Crepaz & Arian, 2024), less abuse of administrative resources, promotion of democracy, reduction of corruption (de la Rama et al., 2022; Sule, 2021) and reduction of inequality (Lupu & Warner, 2022). Lupu and Warner (2022) found that increased regulation is positively correlated with reductions in inequality on a global scale, but also to a larger effect on the scale of Western Europe. These inequalities arise when wealthier groups of society find ways to use their resources to influence politics to work in their favour. This leaves other societal groups with an unfair disadvantage, since the politics will be skewed in a way that does not serve all people equally and favours the wealthy. In a 2021 study by Maclean et al. (2021) on power disparities through philanthropy in the US and the UK, similar effects were found. Richer people who engage in philanthropy are able to leverage their financial influence into political power. The implications of these findings could materialize in regions with significant financial polarization. Agrawal and Phillips (2020) find that in the UK there are large income differences on the local authority level. This could mean that even in close geographical proximity, large differences in political representation may occur. Following the results of these studies on inequality and the findings from Lupu and Warner (2022), further emphasizes the importance of regulations and restrictions for political donations and encourages research into potential localized asymmetries of political power. Despite the larger body of research and media attention for financial inequalities in political funding, there are other inequalities that should also be paid attention to and investigated further.

There are several societal patterns that show in research on political donations, which are less broadly represented in current literature, though they are still societally important influences to study. One such pattern is the differences in gender representation in donation behaviour. Tolley et al. (2022) researched donation patterns between male/female donor groups in Canadian donation data spanning a 25-year period. Their study concluded that female donors tend to donate smaller amounts of money than male donors and donate less frequently. It was also found that female donors tend to vote for national parties more often than local candidates. Though there is limited literature on gender disparity in political donations, it is possible to look into related fields. A good example of this can be found in

the study of Hsu et al. (2021). Hsu and colleagues researched patterns in online donations made through a platform called NPO Channel, which has a wide variety of options for charitable donations. The donation data was investigated for multiple features, including demographic variables like gender, age and education level. Some of the results agreed with Tolley et al. in that female donors were found to contribute smaller amounts at a time than male contributors. What differed between the studies is that Hsu et al. found that the female group contributed more money in total than the male group, with a dominant in-group representation of females between 30 and 40 years old, who have completed high levels of education. Even though the natures of the donations vary, with a limited literature base, it is important to still take findings from related fields seriously. To better determine the behaviours of female donors in political settings, this study will explore gender disparities in donation amount and frequency. An unassuming stance is adopted where no effect of gender is taken as the baseline to measure gender differences in this new context (RQ2).

Apart from gender as discussed above, there are other variables that could be taken into account when researching donation patterns, e.g. age and education. Garnett (2024) performed a study on Canadian political donation data which has a goal akin to this study, which is to gain more insights into financial support for political entities using ML models. Using a collection of variables including age, income, racial/ethnic distribution and education, likelihood of donations in small localized areas as well as amount of donations received per 100 people were predicted. This was done using logistic, poisson and multi-level OLS regression models. Notable findings are positive relations between donation amount, age and income. Further, level of education was mostly related to likelihood to donate rather than amount donated, groups of South and Southeast Asian people were found to have significantly above average campaign contributions, while Black and Latin Americans were less likely to donate with lower total amounts donated as well. This confirms patterns previously by Besco and Tolley (2022), also performed on Canadian donation datasets. Using people with European surnames as a control group, they find that South Asians donate at a very high rate compared to other groups. Following these findings it becomes clear that it is important to consider that many different factors play into donation behaviours on any scale. This study will use features such as racial distribution (RQ2), education level, and age in its estimations to see if they have any significant effect on amount donated in the UK political donations dataset. Because this study focuses on a topic that has been underrepresented in the body of research on political donations, no effect is assumed while working with

these variables, as it is not yet clear whether any effect will reflect the findings from the studies above.

Lastly, it is important to consider donating groups that are not individuals. These can really be organizations or any kind, whether these are businesses, advocacy groups, charities, etc. Many of these organizations have larger financial resources than the vast majority of individuals who donate and often have people in executive positions who are financially well-resourced themselves. There are common suspicions based on the effect that these organizations may have on the democratic process at the national level. In the UK, historically, the Conservative Party has been the primary recipient of corporate donations, while Labour would largely receive their funds from trade unions. This, however, is slowly changing. There is a rise of contributions from individuals (Draca et al., 2022; Fisher, 1997) as well as a blurring of the lines between campaign funding sources, with corporate donations mainly swaying with expected returns from donations and these expectations leaning away from the Conservative Party more (McMenamin, 2020). Still, contributions from organizations are a large part of a political entity's financial resources and even harmful industries try to influence politics with large sums of money (Lacy-Nichols & Cullerton, 2023). For this reason, it is not only important to consider contributions from individual donors, it is also imperative to take the whole picture and pay attention to donations from organizations as well (RQ3). Donation data from the UK has data available from both sources, though only data from non-individuals have their postal codes published and are thus viable for location-based analysis. The data that falls in this latter group, combined with features from Census data, will thus be used as a proxy for analysis on heavily localized data in this study.

In summary, several features that might be of effect on the amount of money political entities receive are identified. Some main variables of interest are levels of income across small-scale geographical areas, gender, income, racial distribution, age, education, and type of donor. Because of the sparse literature on these possible contributing factors to donation amount, one aim of this study is to synthesize a dataset that allows for analysis of these features and find whether these effects are reflected in the UK political donation dataset as well.

3.2 *Suitability of UK datasets with regards to literature*

Ever since the Political Parties, Elections and Referendums Act 2000, the U.K. has been in the highest margins of regulation on political finances (Ewing & Ghaleigh, 2007) worldwide. Unlike many other countries that only require donation amounts to be recorded from a set threshold and only

deal with that data, The Electoral Commission keeps track of donations starting at £0, with a mandatory minimum for registration of £500. This means that there is a rather complete view of this manner of political financing for this area. The data reflects a period of over 20 years, providing a substantial dataset that offers an abundance of opportunities for ML studies. This also means that this study has the opportunity to verify ML methods used in related studies and with that, create a larger consensus on appropriate ML methods to be used in the field of political financing. One important discrepancy exists between the datasets used in other studies and the present study, which is that not all features used are found in the U.K. donations dataset or if they are, they are not of an entirely identical nature. Because of this, UK 2021 Census data from the Office of National Statistics is used to approach desired features to the largest possible extent. Combining datasets in this way, this study takes a first step to relate ML findings across multiple countries with regard to political donations specifically.

The geographical area and nature of the data used for this study were deemed fit for its largely unexplored nature in the field of Machine Learning. Being able to analyse data from the UK not only facilitates studying of appropriate ML methods for political finance data, it also provides opportunity to explore effects in a relatively new area, furthering research efforts in that way too. In fledgling findings across other geographical areas, different patterns have been identified that have not been explored in a broader geographical context. They have only pertained to specific countries or areas. In order to strengthen the literature base of ML studies on political donations, this study is one of the first to take the opportunity to look at the United Kingdom.

Since the UK has adopted stricter political donation policy starting 2001 and donation data has not been extensively studied yet, it is an interesting focus area for this study. The UK is also part of Western Europe where restrictions and regulations on these donations should provide larger measures of equality. Furthermore, the UK has detailed (1) Census data on its local authorities with information on many socio-economic features and (2) donation data that can be either localized or gendered through name interpretation algorithms. This makes it suitable for the purposes of this study. For example, the Census data allows for an opportunity to approach features used in other studies, to see if similar patterns can be identified. Since a large variety of features is used in this study, different sets of features are used with different analyses, as explained in the methods section of this thesis. The best performing model will be used to report on any patterns found regarding topics mentioned in the sub-RQs.

3.3 *ML models*

Even though ML is underrepresented in the literature on political donations, some researchers make a case for an increased use of it in the social sciences. They argue for example that the right parameters can be identified from very large datasets and that ML is well-suited as a tool for social science research (Grimmer et al., 2021). The fact that there is reason to use ML in this field, but there is only limited research published using it, makes it hard to find an established manner of approaching ML tasks. On top of this, as is the case with ML, there are many possible approaches that can be taken to study donation amounts in political funding or similar research goals. Consequently, what follows now is an overview of some closely related literature and the way that existing studies approach their research goals with ML. This way, it is determined what ML methods would be suitable for this study.

3.3.1 *Closely related work*

In studies involving political donations as either an independent or dependent variable, many different models are chosen to conduct analyses. Titl et al. (2024) use multiple ML methods including tree-based algorithms to identify politically connected firms, using donation data alongside other features to perform their predictions. This study uses a baseline logistic regression model and contrasts with ridge regression, Least Absolute Shrinkage and Selection Operator and random forests without and with boosting. The data is split using k-Fold Cross Validation to make the study more reproducible and efficient, as well as to maintain large training data size. Random forest with boosting was found to be the best at the prediction task, achieving the highest degree of accuracy. Other types of models are also used. In their study partially aiming to predict donation amount in political settings, Garnett (2024) uses multi-level Ordinary Least Squares regression models to predict expected donation amount in small geographical areas and logistic regression to predict donor presence in these areas. Lastly, there is a trend of using clustering methods in the research on political donations. Horncastle (2022) introduces the use of a combination of multiple correspondence analysis and model-based clustering, the latter involving multiple ML methods, which are ranked using the Bayesian Information Criterion. Waggoner (2021) makes a case for exploratory analysis of data in political and social research and explores a wide variety of methods including k-means clustering, fuzzy C-means and DBSCAN. Finally, Wahl et al. (2019) applies Fuzzy c-Means clustering to spectral k-Means clustering and then uses these clusters in random forests alongside voting behaviour to analyse donor behaviours. Therefore, what

becomes apparent from literature on similar topics to this study is that logistic regression is a popular baseline model, tree-based algorithms and clustering algorithms are commonly used focus models, and most studies approach their analyses as classification tasks.

3.3.2 *Other fields*

Since the literature that includes political donations is rather limited, especially when ML are included, closely related research fields like crowdfunding and charity financing can be used to get a broader idea of how to approach donation research with ML. A trend that is also visible in research on politically oriented studies is that there is a focus on classification tasks e.g. whether a person will donate or not (Nabar, 2020; Schetgen et al., 2021), whether a fundraising campaign is likelier to have high or low performance (Liu & Hu, 2024) or which people to target for highest return of acquisition costs in fundraising campaigns (Cagala et al., 2021). Commonly used ML models in these studies are decision trees, random forest, KNN, logistic regression, XGBoost and Support Vector Machine (SVM). Apart from these classification tasks, literature adjacent to the topic of political donations which focuses on regression tasks is also available, providing further background for ML options to be used in this study, despite being scarcely available.

3.3.3 *Regression*

Regression models in ML are used to predict continuous numerical values. Since political finance handles income streams for political entities, it would seem likely for regression models to be appropriate for use on such data. Even though regression models are only very rarely used in the field of political donations or any related fields, in the scope of this study it would seem valuable to explore the usefulness of each type of approach. For this reason, regression models and classification models will be contrasted as well (RQ₁).

Many of the models used in aforementioned classification tasks also have applications in the regression domain. In a study by Peng et al. (2021), ML methods commonly applied to classification problems are applied to a regression context and tested for their efficacy for financial contribution prediction tasks in the field of medical crowdfunding. This study contrasts five different models: KNN, Linear regression, ANN, Classification and Regression Tree (CART), and XGBoost. For the training of the models, a 10-fold cross-validation was used with MAE, MSE and R-squared as metrics. Ultimately, the predictive powers of XGBoost were found to be the strongest overall, performing close to CART and ANN. Alazazi et al. (2020)

researches six ML models and their predictive performance of amount of money donated per day in crowdfunding campaigns. Linear regression, CART, SVM, KNN, NN and random forest models were contrasted using root mean square error (RMSE), R-squared and adjusted R-squared as evaluation metrics. SVM was found to perform the best across all models. The fact that many different models are contrasted by researchers who are trying to work towards a greater consensus of what ML models are appropriate to use in a large spectrum of highly specific prediction tasks in the social and political financing domain proves a lack of knowledge in this regard. This study constitutes an attempt to contribute to addressing this issue.

To summarize, there is a need for a greater consensus on what ML methods are effective and appropriate in the analysis of political donations. Many studies in the field of political finance and also in adjacent fields focus on classification tasks in an attempt to predict many different facets of donation behaviours and amounts. Models used in these classification tasks often also have their adaptations for regression tasks, even if the model is used in a slightly different form. Though less commonly represented, these regression tasks cannot be skipped over, so they will also be included in this study. Commonly well-performing models are CART, SVM, KNN, random forests and XGBoost. For this reason, this study will use these models in an attempt to explore their effectiveness and performance on the UK political donations dataset, supplemented by UK Census data.

4 METHODOLOGY & EXPERIMENTAL SETUP

4.1 General setup

This section is a brief description of the research flow visualised in *figure 1*. First, donation data from the Electoral College are pre-processed and made suitable for processing. The same is done with UK Census data, which also involves the extraction of relevant features. After this, both datasets are merged. The complete data can then be used for model training. To facilitate this and to promote model robustness, out-of-sample generalization and reduce the effects of imbalanced groups in the data, this study employs k-fold cross-validation for the regression tasks to split its data with $k=10$. A 10-fold approach is accepted for this study because of the general acceptance of this distribution, as well as the support from related literature on similar topics. The splits that will be applied for the training of the models will all have the same random state (`random_state=26`) to ensure a more reliable comparison between model performances. For the classification tasks, a stratified and shuffled train_test_split are performed

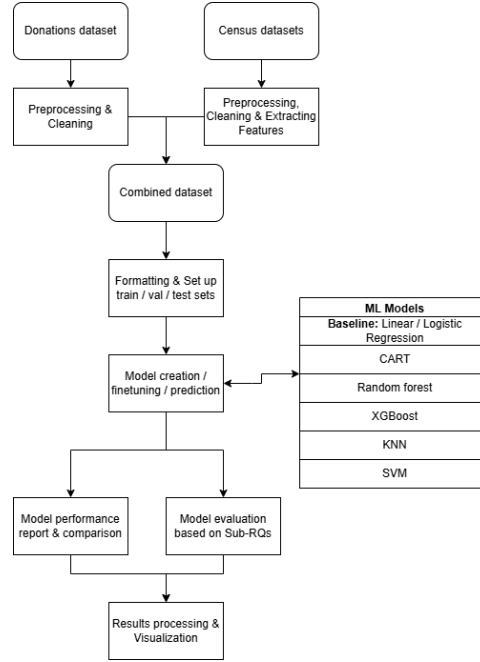


Figure 1: General setup flowchart

for ease of use, while still maintaining generalizability through the proportional representation of each category, based on occurrence percentage in real-life data. Train, validation and test sets of the data are obtained this way (random_state=21). Fine-tuning of parameters is performed on the validation set of this data. Six ML models selected from related literature are trained on the combined dataset that is created. The models are subsequently used to predict new instances of the data. The results of these predictions are used to answer all RQs. The best performing model is used to give insights on sub-RQ 2 and 3 of this study. Learning curves and a table with performances are created for the regression tasks, while for the classification tasks, confusion matrices and a table and graph for accuracy results are created.

4.2 Data

4.2.1 General

The main dataset used in this study was obtained on 8/9/2024 from the website of the UK Electoral Commission (The Electoral Commission, n.d.) where data can be obtained regarding political donations, loans and more. The website can provide data from 2001 onward and all information is available freely to anyone. The dataset that was downloaded contains

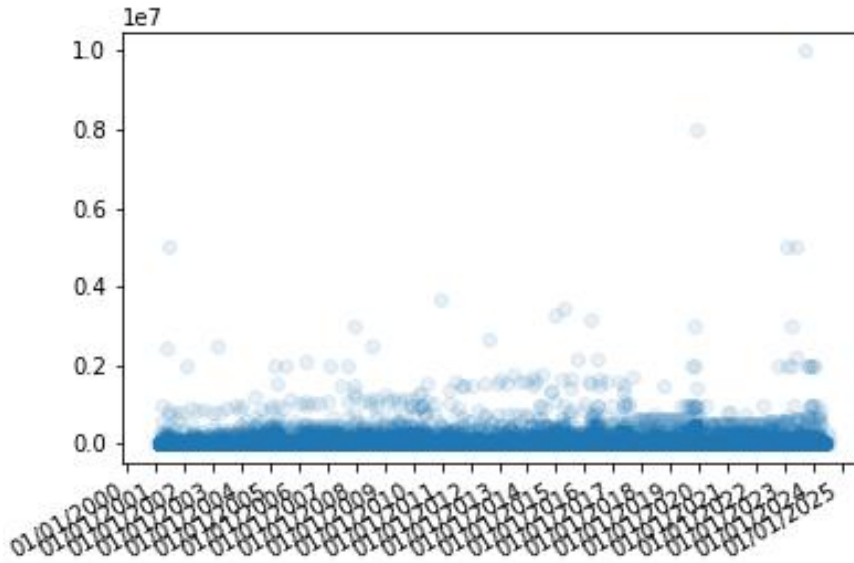


Figure 2: All donations across all years

84470 rows and 29 columns excluding index, of which 4 are bool type, 2 are floats, 1 is integers and 22 are object type. The dataset is 16.4MB. This dataset was chosen as it is data that is easily available, yet largely unused in the context of machine learning studies. The dataset also has a large number of rows and is thus appropriate for prediction tasks. Because of these factors, it is expected that with this dataset, a valuable addition to the literature on political funding can be made. The donation spread over the years can be seen in *figure 2* and *figure 3*.

4.2.2 Features

Some examples of features are 'DonorStatus' which identifies the type of donor to be an individual, company, etc., 'Value', that indicates how much was donated and 'RegulatedEntityType', identifying the type of receiving entity. One extra feature is added to the dataset regarding gender. This could only be done through gender estimation of individual donors and not with the other donor groups. The value for gender for the non-individual donors was set to 'not a person'. Gender estimations were made using the Gender Guesser by David Arcos ([found through this link](#)). As for non-individual donors, through the other donors having postal code data available, other data is connected to these groups that could not be connected to the individuals. In the following paragraph, this additional data is addressed.

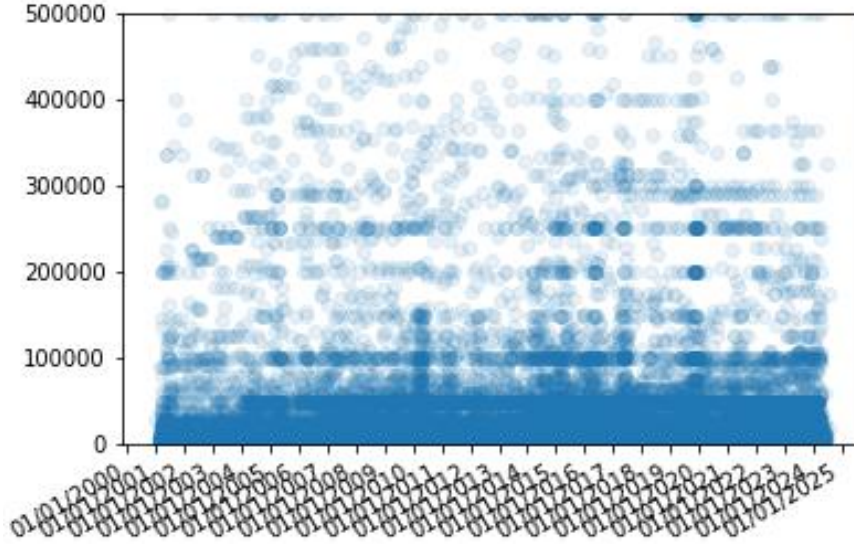


Figure 3: All donations across all years, no values > £ 500,000

4.2.3 Obtaining supplementary data

Supplementary data was obtained and all updated on December 2, 2024 from the Office for National Statistics (ONS). Since this study partially focuses on geographical areas on the local authority level, a postcode lookup from ONS (attribution statements: (1) Contains OS data © Crown copyright and database right [2024], (2) Contains Royal Mail data © Royal Mail copyright and database right [2024], (3) Source: Office for National Statistics licensed under the Open Government Licence v.3.0) updated November 28, 2024 (retrieved on December 2, 2024, from [here](#)) was used to match postal codes from the donation dataset to one of 359 different local authorities. Just over 2.7 million postal codes are recorded in the dataset. Secondly, 2021 racial distribution data on local authority level was obtained via [this website](#) on December 2, 2024. Third, data on education level distribution per local authority was obtained from [this source](#) on December 2, 2024. Further, 2020 income data tied to local authority districts were obtained through [here](#) on December 2, 2024. This includes income after housing costs. From the literature, it seems most appropriate to use this, since there are imbalances in housing costs across different regions. These imbalances are mitigated this way. Finally, a 2023 dataset containing data on population estimates regarding (median) age, gender, obtained on December 2, 2024 through [this link](#). All datasets from the Office for National Statistics (all supplementary datasets) fall under the Open

Government License v3.0. This means that it is appropriate to use this data for the purposes of this study. More information on the license can be found [by following this link](#).

4.2.4 *Implementing supplementary data*

All these supplementary datasets can be merged together through the local authority district area codes. From the postal code lookup dataset, only postal codes, area code (in this study, local authorities) and name of the area are preserved. The dataset containing age data did not provide a recent age average. For this reason, all ages times their occurrences per area are averaged to obtain localized mean age data. This is preserved alongside the area codes. From the education data, area code and qualification index score are taken, since the latter is a good measure of score distance between all areas respective to each other. For the ethnicity distribution dataset, the percentages per area are kept alongside area codes. Finally, the income data needs to be processed to reflect local authorities. This is done by grouping on area codes and then averaging grouped incomes after housing costs to have a better idea of the area this study focuses on. Both the area codes and the net annual income excluding housing costs were used. After all this pre-processing, these data are linked to the donation dataset by matching the postal codes with the right local authority district in the postcode lookup dataset. All numerical data is scaled with Sklearn's StandardScaler algorithm before use in regression models.

4.2.5 *Encoding for classification*

Due to the fact that individual donors do not have postal codes related to them, it is not possible to study these across the same features as donations of non-individuals. One advantage is that gender interpretations can be made of the names of the donors. This facilitates a way for this study to analyse gender based patterns as well, although in a way that is disjointed from the analysis that also handles localized data. However, gendered data can be used for the classification models. This is done by using the Gender Guesser algorithm as described before, but replacing all 'NaN' values of non-individual rows in the gender column with 'not a person'. For the classification task, the columns ["RegulatedEntityType", "DonorStatus", "DonationType", "IsAggregation", "gender"] are one-hot encoded in order to be processed by the ML models. This results in an encoded dataframe of shape: (75918, 30). The columns can be interpreted in the following ways: "RegulatedEntityType" is the type of political entity receiving the donation, "DonorStatus" is the type of donor, such as individuals or companies, "DonationType" describes the way the money was transferred,

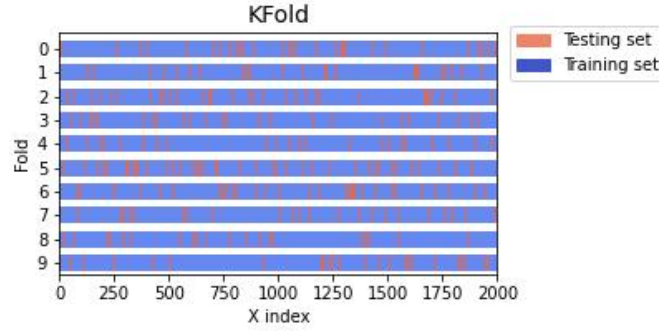


Figure 4: KFold split distribution

“IsAggregation” describes whether the payment or group of payments passed a legal threshold and needs to be reported.

4.2.6 Data selection & general flow for regression and classification

After all preprocessing and merging, the data for the regression analysis is chosen. The columns all hold continuous numerical values and make up data on donation amount as target, and as features on average age per local authority, qualification index score per local authority, net annual income after housing costs (£) per local authority, and racial distribution per local authority. The features are scaled and a KFold split is initiated with 10 splits. This split structure was chosen to improve out-of-sample generalizability and robustness. Split distribution can be seen in [figure 4](#). Using Sklearn’s crossvalidate function, the six models (1 baseline, 5 to evaluate) are trained and tested on the data and RMSE and R2 scores are obtained. RMSE is favoured over MSE, which will be explained in the evaluation section.

For classification prediction, all one-hot encoded columns are used alongside a categorized donation amount column. Donations were categorized into buckets of intervals between the following numbers and anything past the final: [0, 1000, 2500, 5000, 10000, 100000, 1000000]. These numbers were chosen in order to keep the amount of labels to predict limited and were tailored to evenly spread out donation occurrences over the intervals as much as possible. For this data, a stratified and shuffled train_test_split was performed in order to keep target class percentages and counteract order of donations by date. These measures were taken to improve out-of-sample generalization and improve robustness. Since the classification problem is simpler than the regression problem, a more rudimentary manner of splitting the data was chosen.

Regression	Classification
LinearRegression	LogisticRegression
DecisionTreeRegressor	DecisionTreeClassifier
RandomForestRegressor	RandomForestClassifier
KNeighborsRegressor	KNeighborsClassifier
SVR	SVC
XGBRegressor	XGBClassifier

Figure 5: All models used

4.3 Software, programs & ML models

All ML models that were used can be found in *figure 5*. Other software and programs that were used, including licence agreement and version, can be found [through the author's GitHub repository](#) or in Appendix A. A note about the use of the XGBoost algorithm: XGBoost was found to be a high-performing model in various studies among those discussed in the literature review section. It has an affinity for handling large datasets, flexibility with mixed data and NaN values, as well as being recognised for its performance in both regression and classification tasks. More information about XGBoost can be found [through this link](#).

4.4 Evaluation

Apart from using k-fold cross-validation and stratified and shuffled train_test_split to promote generalization strength of the models, other measures are also taken in terms of evaluation. First, a simple linear and logistic regression model is used as a baseline to compare other models with. Due to their linear and rigid nature, these model are used as a dummy model to provide a standard that all other models can be contrasted with. This allows for insights into the performance of the other models in comparison to one another in terms of their performance compared to the baseline. As the error analysis metric in model evaluation, Root Mean Squared Error (RMSE) is chosen. Some advantages of RMSE over MSE are that it takes outliers into account, providing better coverage for extreme values since these also hold valuable information in donation behaviour research. Another advantage is that the unit of the outcome will represent pounds, rather than squared pounds, making the outcome more easily interpretable and explainable in a social and political context. For more precise training due to its larger punishment of errors, though slightly less easily interpretable, RMSE is also chosen over the linear MAE. Another metric which is used is

R-squared. This metric gives a measure of how well the features explain the effect found in donation amount, providing a goodness-of-fit measure that can justify the use and impact of the chosen features for this study. For classification, accuracy was used as the metric for error analysis. Other studies mentioned in the literature review used this metric as well, and to reduce complexity of the results, only accuracy was chosen. When models have been compared, the best performing model is used to reflect on the sub-RQs of this study. This is done by selecting sub-groups from the data used for prediction and contrasting values according to the relevant RQs.

5 RESULTS

5.1 Regression results

In this section, performance and results for the regression models are described. As seen in *figure 6*, none of the regression models achieved a passable performance. Even though not all models were ultimately trained, it can be seen that only very low R2-scores were able to be obtained for the models that were used. Hyperparameter tuning and feature reduction did not help model performance and RMSE stayed very high. As can be observed in *figure 7*, learning behaviour was slightly different between baseline model and the CART model, but ultimately the learning curve would hover around very similar poor scores. Because of the low performance, variables and patterns that were planned to be investigated in the sub-RQs that occurred in these models will not be able to be investigated well. Reasons for the model not performing up to a good standard could be bugs in the code, problems with the form of or uninformative features used for training. Another issue could be that the data was all area-focused and derived from non-individual data. Ways to possibly increase model performance are preparing the data in different ways for the models to process, introducing different data that might help reduce prediction error and goodness-of-fit. Lastly, working with data from individuals, rather than anything but them, might improve performance as well. Companies and other organizations might not offer donation data that is generalizable with regards to the area they are from.

5.2 Classification results

In this section, the performance and results achieved by the classification model will be discussed. As can be seen in *figure 8*, the classification models yielded much better results than the regression tasks. All models perform above baseline level, with Random Forest and XGBoost achieving

Model	RMSE	R2
Linear Regression	-94193.2	0.01388
Decision Tree	-94016.6	0.017592
Random Forest	-94013.3	0.017672
KNN	-102328	-0.17434

Figure 6: Regression scores

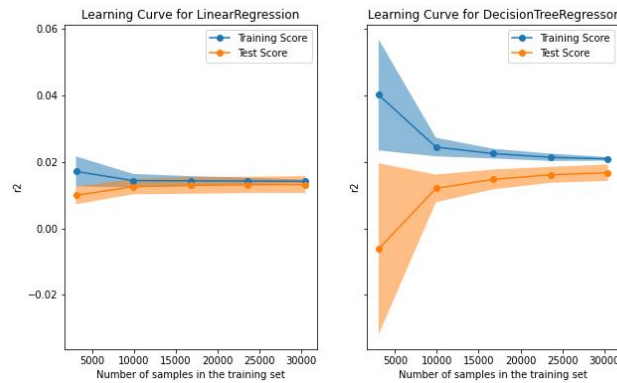


Figure 7: Learning curves for regression tasks (Linear Regression & Decision Tree)

the highest accuracy scores. It can also be observed in *figure 9* that the Decision Tree, Random Forest, SVM and XGBoost perform rather similarly. Contrastingly, the KNN model performs worse than the other models though still performs better than the baseline model. With scores well above chance level, it becomes apparent that the chosen features were appropriate for use in a classification case on this data. One way performance could be improved further, is by adding features that add predictive power to the model. The features used for the training were still rather limited in scope, with a limited amount of research backing them up. This means that relevant features may be yet to be identified. Looking at confusion matrices in *figure 10* for the predictions, it becomes apparent that the higher value classes are poorly predicted by the models compared to the more common lower value classes. Random Forest predicts the most high value classes correctly, whereas logistic regression performs best on the lower value and common values. A tendency to predict lower value classes might occur because of the larger amount of data in these categories that the model has the chance to train on. If this is true, a larger amount of data for higher value classes, or by using data augmentation to increase the amount of training data in these classes might help teach the model to predict higher values with higher accuracy. One last manner in which

ML model	highest accuracy
LogisticRegression	0.421891465
DecisionTreeClassifier	0.430057956
RandomForestClassifier	0.430321391
KNeighborsClassifier	0.425579557
SVC	0.430189673
XGBClassifier	0.430321391

Figure 8: Highest achieved accuracy scores per model

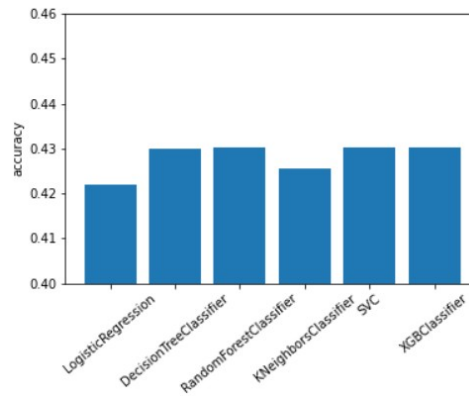


Figure 9: Accuracy scores per model

performance on higher value targets may be improved, is by using other sampling methods for model training data. Perhaps a KFold split would improve performance.

5.3 Gender differences

In this section, the results of the predicted values by gender are discussed and reflected on with the true labels. Important to note is that even though the Gender Guesser also measures a 'neutral' category, previous literature has only focused on male and female differences, so in this study the focus will be on 'male' and 'mostly male' categories, contrasted with 'female' and 'mostly female' categories. The results can be seen in [figure 11](#). Predictions for this analysis and the donor type patterns analysis were done with XGBoost, because of its high performance in this study and the support it receives in contemporary literature. Findings show low total donation frequencies for female donors and an underestimation for donations with high value (£10,000 – £100,000) in predicted values. Male donors show high total donation frequencies, but the inverse to what happens in predicted values for female donors occurs here, namely a significant overestimation of donations in the (£10,000 – £100,000) range, while many donations were

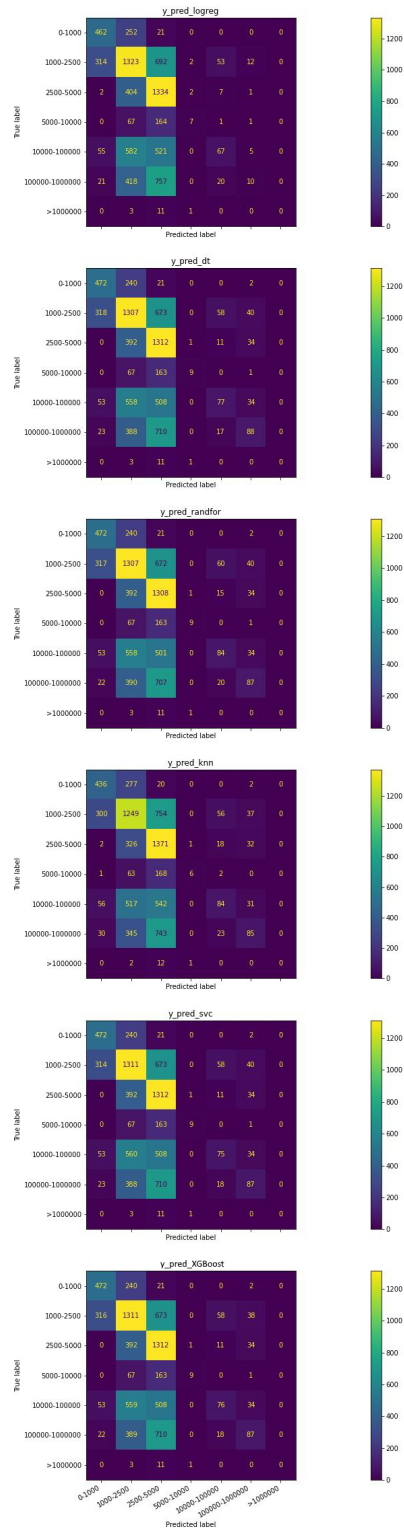


Figure 10: Confusion Matrices per model

gender_female		gender_mostly_female		gender_male		gender_mostly_male	
y_pred		y_pred		y_pred		y_pred	
1000-2500	333	0-1000	26	10000-100000	1822	1000-2500	66
10000-100000	18	5000-10000	22	1000-2500	525	10000-100000	12
0-1000	11	2500-5000	2	0-1000	93	0-1000	3
Name: count, dtype: int64		1000-2500	1	5000-10000	31	5000-10000	2
y_true		Name: count, dtype: int64		2500-5000	5	Name: count, dtype: int64	
1000-2500	135	y_true		Name: count, dtype: int64		y_true	
10000-100000	70	1000-2500	21	y_true		1000-2500	26
5000-10000	62	0-1000	18	1000-2500	715	10000-100000	19
2500-5000	60	5000-10000	6	10000-100000	707	2500-5000	15
0-1000	29	2500-5000	4	5000-10000	424	0-1000	12
100000-1000000	6	100000-1000000	1	2500-5000	369	5000-10000	11
Name: count, dtype: int64		10000-100000	1	0-1000	163	Name: count, dtype: int64	
		Name: count, dtype: int64		100000-1000000	94		
				>1000000	4		
				Name: count, dtype: int64			

Figure 11: Gender differences in donating behaviour (female, mostly female, male, mostly male)

lower. One possible explanation for this could have to do with gender-specific donating behaviour. For example, categories of features other than gender could coincide with gender-specific behaviours, even if these donations do not come from individuals.

5.4 Donor type patterns

This section presents the differences between the predicted values of the different types of donor groups as they are found in the ‘DonorStatus’ features. *Figure 12* shows the donation behaviours per donor group. While individuals have the highest frequency of donations for the highest amount, the trade unions and unincorporated associations are the next most frequent donors, however, donating on the lower end value categories. Other donor types may donate less frequently, but are often in the habit of donating larger amounts at once. Although there could be a skewedness in the predicted values, just as with the gender data, these findings are consistent with ongoing trends in UK donating behaviour.

6 DISCUSSION

The goal of this study was to analyse the appropriateness of the use of a large number of sparsely used ML methods in the field of political funding fields adjacent to it. This was done while also investigating some more common features and their effects as found in other literature on donation amount. As its main question, this thesis studied five different ML models and compared them to a baseline to find the best performing model for the prediction of political donations. This was done as both a regression task and a classification task. Due to the regression models not learning from

DonorStatus_Building Society		DonorStatus_Friendly Society		DonorStatus_Individual	
y_pred		y_pred		y_pred	
10000-100000	1	10000-100000	42	10000-100000	2078
Name: count, dtype: int64		1000-2500	8	1000-2500	1424
		5000-10000	8	0-1000	146
		2500-5000	7	5000-10000	59
		Name: count, dtype: int64		2500-5000	7
				Name: count, dtype: int64	
DonorStatus_Limited Liability Partnership				DonorStatus_Other	
y_pred				y_pred	
10000-100000	30			2500-5000	96
5000-10000	9			1000-2500	17
0-1000	8			100000-1000000	8
1000-2500	1			5000-10000	5
2500-5000	1			Name: count, dtype: int64	
Name: count, dtype: int64					
DonorStatus_Public Fund		DonorStatus_Registered Political Party		DonorStatus_Trade Union	
y_pred		y_pred		y_pred	
10000-100000	200	1000-2500	61	1000-2500	620
100000-1000000	1	0-1000	55	0-1000	470
Name: count, dtype: int64		2500-5000	24	10000-100000	29
		5000-10000	4	2500-5000	2
		Name: count, dtype: int64		Name: count, dtype: int64	
DonorStatus_Trust		DonorStatus_Unincorporated Association			
y_pred		y_pred			
1000-2500	14	1000-2500	597		
0-1000	3	0-1000	186		
Name: count, dtype: int64		10000-100000	13		
		2500-5000	4		
		Name: count, dtype: int64			

Figure 12: Donor type differences in donating behaviour

the (localized) input data, the model was unable to explain the variance of the donation amount, while the high RMSE scores indicate poor prediction performance and high errors. These findings could mean that classification is more suitable for the prediction of the movement of political resources, but it could also mean that if data is handled and prepared differently, or when other data is used which would allow the model to learn more effectively, regression might hold more value in these kinds of tasks than this study was able to highlight. In this study, classification offered the best model performance for the prediction of political donation amounts, with the two best-performing models being Random Forest and XGBoost.

This study also identified significant gender differences in donating behaviours. Additionally, it found that the ML models underestimate female donor contribution and overestimate male contributions. This might have happened through the influence of other training data. Finally, from all donor types, individuals were predicted to donate the most often and very commonly, the greatest sums of money to political entities. The other higher frequency donors do not normally donate as much at once, while other donor types may contribute less often, yet with larger amounts. To summarize, classification tasks using Random Forest and XGBoost were found to perform the best on the data provided. Using XGBoost, predictions were made to analyse gender differences and donor type differences for donation amount. Significant differences between male and female donors were found, and patterns were identified by contrasting different donor types.

6.1 *Results in the light of the literature*

The fact that most ML research in the field of political donations has been performed as classification tasks seems like an optimal decision that would ultimately lead to better results, according to the findings of this study. The model performance findings of this study support the findings of Peng et al. (2021) and Titl et al. (2024), though the latter was performed in this study without the boosting. Both XGBoost and Random Forest are commonly used algorithms across studies, although are not always found to be the most powerful in terms of prediction. This study does make a case solidifying the position of these two algorithms as benchmark models in classification approaches to political finance. Regression models performed poorly on localized data, but as Garnett (2024) establishes, these data might be more suitable for prediction of likelihood to donate, rather than continuous amount donated. In terms of donor groups, some interesting effects were found. A dominant presence of donations from individuals could be explained by patterns identified by Draca et al. (2022) and Fisher

(1997), identifying a notable rise in donations originating from individuals. It could also mean that while trade unions and similar organisations are still donating regularly, the amount of resources obtained and with it possibly the importance from these sources are starting to be overshadowed by contributions made by individuals.

As far as the findings on gender disparity in donations go, this study tends to lean more towards support of behaviours found in Tolley et al. (2022). This means that female donor behaviour did indeed show lower amounts donated per donation, as well as a lower donation frequency compared to male donor groups. The first part of these findings, although exaggerated in model prediction, agrees with Hsu et al. (2021) in that females donate in smaller amounts, but the second part contrasts the study's findings that females would donate in larger frequencies than males. This could be explained by the possibility that female donors treat political donations differently to charitable donations, which Hsu et al. research.

6.2 *Limitations*

A notable limitation of this study is that through its efforts to pull together a broad range of features and ML methods, it became a study itself that was not able to go very in depth on very specific configurations of models, parameters and features. Other findings might have been obtained, for example with the regression models, if the study focused on a smaller scale approach that only aimed to produce a high-performing regression model. Another limitation is the amount of data available for donation values at the extreme ends of the labels, most notably data on very high value donations. This might have come at the cost of predictive performance. One more limitation is the features used for the training of the models. Although this study attempted to take a broad approach in terms of features, it still could have analysed further possible features and interactions. Other selections may yield stronger results for the chosen models or might find different patterns entirely. For better comparison between regression and classification tasks, the models could have also been run on the same data, though with different target structures. While this study does offer insights into future potential research about a broad range of approaches that can be taken when considering this type of data, it is important to take these limitations into account.

6.3 *Impact*

Overall, this study achieves what it set out to do, providing information on the appropriateness of various ML implementations and finding patterns in donor behaviour that extend the current literature base. Firstly, gender-based donating patterns previously identified in other countries have now been confirmed in the UK. Furthermore, this study presented an overview of donating patterns between all discernible donor types. These findings provide discernible insights some basic disparities in financial resourcing of political entities and their activities. This empowers citizens, since it gives them the power to form critical opinions about financial flows in the political system. Secondly, the model performances for classification tasks offer guidance and justification on what models future research in similar areas might choose to use. Finally, donor behaviour patterns that were confirmed provide grounds for further investigation, since for example the gender disparity is now found in a geographical area where it was not explicitly identified before, extending insights to UK political finance.

6.4 *Future research*

Beyond the scope of this study, there is more to analyse with the UK donation dataset, especially given that it is so large and has not been used to its potential. Patterns found in this study may be further investigated since they hold significant social importance. For example, an explanation may be found for the exaggeration of the gender-specific donation amount predictions. Future studies may also challenge the idea that classification is more suitable for political donation analysis, since some studies did manage to build functioning regression structures.

7 CONCLUSION

Political donations are of great importance to the political entities that receive them, as well as the citizens that are governed by these entities. Despite this, these resources remain largely underrepresented in scientific research. In existing literature on these donations, newer analysis tools are rarely discussed. Powerful ML methods including state-of-the-art algorithms remain largely unused, and when they are used, there is no broader consensus on what ML methods would be most appropriate or effective to use. This study aimed to reinforce the literature base on ML methods in the field of political financing. It did so by contrasting and evaluating the performance of a large selection of ML models and features across both regression and classification tasks. Additionally, this thesis

analysed donor patterns found by other studies on the UK dataset through the predictions of the highest performing model of this study.

This study found support for the use of classification models over the use of regression models, with the Random Forest classifier and the XG-Boost classifiers achieving the highest accuracies across all models. These findings could support the identification of these models as a benchmark for future research on political donation data, though additional research is needed to further substantiate these results. While donations from individuals behaved differently to public organization donors and are of greater amount both in donation value and in donation frequency compared to non-individual donor types, gender differences within the individual donor group were found. Male donors were identified to donate more frequently and in larger amounts than female donors. Since these findings were not identified in the UK before, they could provide a rudimentary step in establishing well-understood donor patterns in this area. Future research may adopt a deeper focus on the inner workings of the patterns found in this study. In conclusion, the findings in this study provide easily understandable insights in donation patterns that can be interpreted and used by citizens and political entities alike. Furthermore, this study helps converge the many possible ML methods for political donation research and provides insights that can be used and more closely investigated in future research.

REFERENCES

- Agrawal, S., & Phillips, D. (2020). Catching up or falling behind? geographical inequalities in the uk and how they have changed in recent years. *The Institute for Fiscal Studies*, 2.
- Alazazi, M., Wang, B., & Allan, T. (2020). Success factors of donation-based crowdfunding campaigns: A machine learning approach. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.
- Besco, R., & Tolley, E. (2022). Ethnic group differences in donations to electoral candidates. *Journal of Ethnic and Migration Studies*, 48(5), 1072–1094. <https://doi.org/10.1080/1369183X.2020.1804339>
- Cagala, T., Glogowsky, U., Rincke, J., & Strittmatter, A. (2021). Optimal targeting in fundraising: A causal machine-learning approach. <https://arxiv.org/abs/2103.10251>
- Crepaz, M., & Arikan, G. (2024). The effects of transparency regulation on political trust and perceived corruption: Evidence from a survey experiment. *Regulation & Governance*, 18(3), 896–913. <https://doi.org/https://doi.org/10.1111/regu.12555>

- dela Rama, M. J., Lester, M. E., & Staples, W. (2022). The challenges of political corruption in australia, the proposed commonwealth integrity commission bill (2020) and the application of the apuncac. *Laws*, 11(1). <https://doi.org/10.3390/laws11010007>
- Draca, M., Green, C., & Homroy, S. (2022). *Financing uk democracy : A stocktake of 20 years of political donations* (Working Paper) [Unpublished], University of Warwick. <https://warwick.ac.uk/fac/soc/economics/research/workingpapers/2022/twerp%5C%5f1431%5C%5f-%5C%5fdraca.pdf>
- Ewing, K., & Ghaleigh, N. (2007). The cost of giving and receiving: Donations to political parties in the united kingdom. *Election Law Journal: Rules, Politics, and Policy*, 6(1), 56–71. <https://doi.org/10.1089/elj.2006.6005>
- Fisher, J. (1997). Donations to political parties. *Parliamentary Affairs*, 50(2), 235–246.
- Garnett, H. A. (2024). Where do donors come from? using census data to predict donations to canadian federal election candidates. *Political Geography*, 110, 103077. <https://doi.org/10.1016/j.polgeo.2024.103077>
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(Volume 24, 2021), 395–419. <https://doi.org/10.1146/annurev-polisci-053119-015921>
- Horncastle, W. C. (2022). Model based clustering of political finance regimes: Developing the regulation of political finance indicator. *Electoral Studies*, 79, 102524. <https://doi.org/10.1016/j.electstud.2022.102524>
- Hsu, C.-W., Chang, Y.-L., Chen, T.-S., Chang, T.-Y., & Lin, Y.-D. (2021). Who donates on line? segmentation analysis and marketing strategies based on machine learning for online charitable donations in taiwan. *IEEE Access*, 9, 52728–52740. <https://doi.org/10.1109/ACCESS.2021.3066713>
- Lacy-Nichols, J., & Cullerton, K. (2023). A proposal for systematic monitoring of the commercial determinants of health: A pilot study assessing the feasibility of monitoring lobbying and political donations in australia. *Globalization and health*, 19(1), 2. <https://doi.org/10.1186/s12992-022-00900-x>
- Liu, Z., & Hu, S. (2024). Predicting the fundraising performance of environmental crowdfunding projects: An interpretable machine learning approach. *Information Processing & Management*, 61(2), 103587.

- Lupu, N., & Warner, Z. (2022). Why are the affluent better represented around the world? *European Journal of Political Research*, 61(1), 67–85. <https://doi.org/https://doi.org/10.1111/1475-6765.12440>
- Maclean, M., Harvey, C., Yang, R., & Mueller, F. (2021). Elite philanthropy in the united states and united kingdom in the new age of inequalities. *International Journal of Management Reviews*, 23(3), 330–352. <https://doi.org/https://doi.org/10.1111/ijmr.12247>
- McMenamin, I. (2020). Party identification, the policy space and business donations to political parties. *Political Studies*, 68(2), 293–310. <https://doi.org/10.1177/0032321719841243>
- Nabar, R. (2020). *A comparative study of machine learning models for fundraising success* [Doctoral dissertation, Dublin Business School].
- Peng, N., Zhou, X., Niu, B., & Feng, Y. (2021). Predicting fundraising performance in medical crowdfunding campaigns using machine learning. *Electronics*, 10(2). <https://doi.org/10.3390/electronics10020143>
- Schetgen, L., Bogaert, M., & Van den Poel, D. (2021). Predicting donation behavior: Acquisition modeling in the nonprofit sector using facebook data. *Decision Support Systems*, 141, 113446. <https://doi.org/https://doi.org/10.1016/j.dss.2020.113446>
- Sule, B. (2021). How accountable and transparent is the african democracy? reviewing political party financing and regulations. *African Journal of Social Sciences and Humanities Journal (ASSHJ)*, 2(3), 168–184.
- Titl, V., Mazrekaj, D., & Schiltz, F. (2024). Identifying politically connected firms: A machine learning approach. *Oxford Bulletin of Economics and Statistics*, 86(1), 137–155. <https://doi.org/https://doi.org/10.1111/obes.12586>
- Tolley, E., Besco, R., & Sevi, S. (2022). Who controls the purse strings? a longitudinal study of gender and donations in canadian politics. *Politics & Gender*, 18(1), 244–272. <https://doi.org/10.1017/S1743923X20000276>
- Waggoner, P. D. (2021). *Unsupervised machine learning for clustering in political and social research*. Cambridge University Press.
- Wahl, S., Sheppard, J., & Shanahan, E. (2019). Legislative vote prediction using campaign donations and fuzzy hierarchical communities. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 718–725. <https://doi.org/10.1109/ICMLA.2019.00129>

APPENDIX A

Library Name	Version	License
Numpy	1.24.4	Open Source, modified BSD
Pandas	2.2.3	Open Source, BSD
Matplotlib	3.9.2	Open Source, PSF-like
Scikit-Learn	1.5.2	Open Source, BSD
gender-guesser	0.4.0	GPLv3
XGBoost	2.1.3	Apache License 2.0