UCL DEPARTMENT OF GEOGRAPHY

YEAR 2023-24

EXAM CANDIDATE ID:	**DWKB3**
MODULE CODE:	**GEOG0178**
MODULE NAME:	**Machine Learning for Social Sciences with Python**
COURSE PAPER TITLE:	**GEOG0178 Data Science Project**
WORD COUNT:	**1994**

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this ☐

Introduction

For the first time in Switzerland's history, in March 2024, a referendum approved an increase in the public pension scheme, costing 4.1bn for the first year, with rising costs. Despite unclear funding plans and governmental opposition, 58.25% of the population voted in favour (VoteInfo, 2024; SRT, 2024a, 2024b). This study uses logistic regression (logit) and random forest (RF) models to explore municipal-level factors behind the referendum's success. It will enhance our understanding of pension reform buy-in and offer guidance for campaign strategies.

Literature Review

Logit models are widely employed to analyse voter preferences and participation. Alabrese et al. (2019) utilised them to investigate Brexit referendum determinants, while Carreras and Castañeda-Angarita (2014) explored predictors of voter turnout in Latin America. RF models are another popular approach, applied by Montfort (2023) to study predictors of support for Swiss climate policies, and Marozzo and Bessi (2017) to examine how news websites impact voter polarisation. Similarly, Breiman (2001) classified American congressmen by party based on their voting records; Kovacs and Uwe (2022) analysed Swiss demographic tendencies toward electronic voting; and Brinksma (2019) identified tones predictive of European elections in news reports.

Studies on voting behaviour extend to pension reform support. Bonoli and Häusermann (2009) found older individuals typically oppose measures benefiting younger demographics but support increases in their pensions. Additionally, they identified minor relationships between income and voting decisions. Van Dalen et al. (2019) and Fernández and Jaime-Castillo (2012) further suggest that left-wing voters and nations with higher elderly poverty tend to resist pension cuts. They also reveal that preferences for delaying retirement are positively correlated with higher retirement ages while individuals in countries with higher social security contributions are less supportive of further increases. Riekhoff (2020) discusses a cyclical relationship, where changes in pension policies affect public opinion and the absence of such changes in Switzerland helps explain consistent attitudes during the 2010s.

Data and Method

Municipal-level Variables

Data for the 2126 Swiss municipalities was collected from the Federal Statistical Office. Municipal-level variables were selected as the smallest unit of public vote data. Individual-level variables were restricted in access or not yet published (e.g., MoSaiCH). Additionally, social desirability and hypothetical bias do not affect referendum results, unlike survey responses (Brunner and Kuhn, 2018).

Predictor Variables

- **Percentage of Votes for the Four Largest National Parties:** Reflects political leanings, associated with views on redistribution and pension policies.
- **Average Income**: Indicates economic status, where higher income levels may lessen sensitivity to pension increases due to diminished relative benefit.
- **Percentage of Population Aged 20-39 and 65+**: Younger voters bear the tax burden longer before benefiting, whereas older voters benefit immediately, influencing support or opposition (Bonoli and Häusermann, 2009).
- **Mean Tax Burden at 80k and 150k Annual Salary**: Earners already facing substantial taxes might oppose further increases (Jaime and Castillo, 2012).
- **Linguistic Region**: Captures cultural disparities. Historically, municipalities vote similarly within-region (VoteInfo, 2024).
- **Municipality Type**: Urban, suburban, and rural municipalities. Reflects socioeconomic background and local industry characteristics, influencing economic interests and perspectives on pensions.
- **Total Population**: Larger urban areas generally have more diverse economies, potentially impacting opinions on pensions differently compared to smaller urban areas.

Data Cleaning and Wrangling

Small municipalities' average income, unavailable due to privacy concerns, was imputed using the median of same-type municipalities within the same canton (FSO, 2024). Vote data from one hamlet was included in another municipality's total, and the average income of Vaux-sur-Morges was skewed by the billionaire owner of Roche, resulting in their exclusion from the dataset (VoteInfo, 2024; Péclet, 2010). Referendum outcomes were recoded from percentages to “support” and “against” for results above and below 50%. The linguistic region and municipality type variables were one-hot encoded. Municipality subtypes (three per municipality type) were merged to prevent dimensionality issues.

The Swiss Municipality Merger Tool provided mapping tables for data aggregation, accounting for municipality mergers. Numeric data was reformatted in Excel, considering municipality size for accurate average and percentage aggregations. Tax burdens for the six 2023-2024 merging municipalities were directly averaged due to unavailable taxpayer counts. Linguistic regions of the 15 merged-municipalities over 2022-2024 were assigned based on the first municipality's region in the dataset.

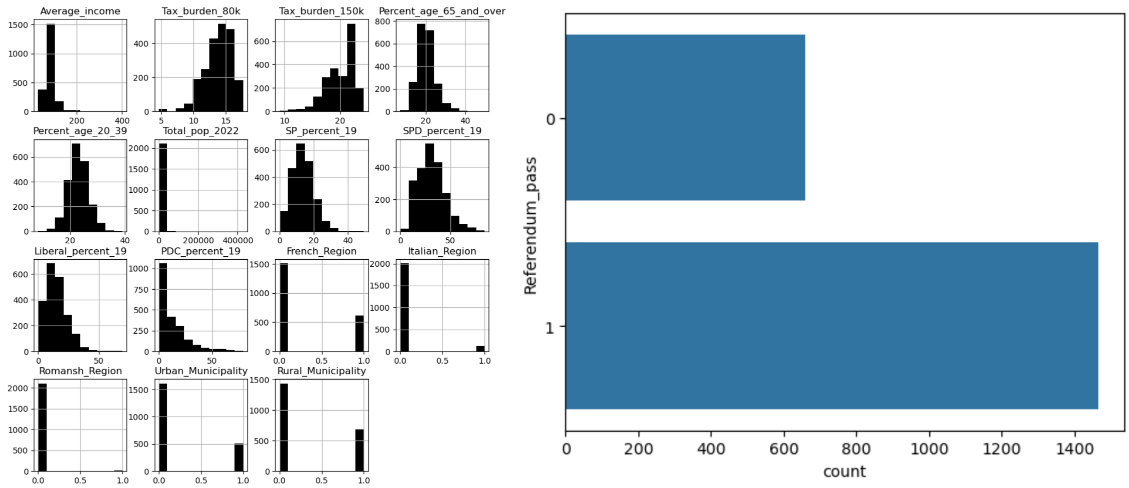
Model Selection and Premises

Logit and RF models were selected for their suitability for classification, handling large datasets and dimensionality, and conveying feature importance (Prinzie and Van den Poel, 2008; Yadav and Sharan, 2020). Logit models require no outliers or multicollinearity and assume linear relationships between predictors and vote outcomes' log-odds. RF models offer a complementary, potentially more robust approach, not requiring these assumptions (Biau and Scornet, 2016; Zhao et al., 2020). Utilising both models enhances reliability and insights, leveraging their distinct methodologies.

Modelling

Default models were run to set a benchmark. Figure 1 shows predictors' distribution and target class imbalance, municipalities supporting the referendum forming the majority class. The data was split into 80/20 for training and testing, stratified to preserve the original class distribution (Sadaiyandi et al., 2023).

Figure 1: Distribution of the Features and Target Classes



Logit Model Tuning

The German region and Suburban municipality type were selected as reference categories, representing majorities in their respective groups (Pillai et al., 2024).

To satisfy logit assumptions, influential outliers were removed based on Cook's distance, leverage, and standardised residuals. Addressing imbalance was crucial for correct classifications of municipalities *for* and *against* the referendum. Inverse-proportional reweighting was initially considered to maintain predictors' distribution. However, an unidentified issue prevented weights from being applied. SMOTE was then preferred over undersampling to preserve observations (Fernandez et al., 2018; Qazi and Raza, 2022). Numeric features were standardised to ensure comparability.

VIF tests revealed strong multicollinearity between the tax burden variables. Initially, Lasso regularisation was used to preserve all features, reducing coefficient instability by penalising larger coefficients (Xi et al., 2023; Perlato, n.d.). However, the penalty entailed by the severe multicollinearity removed four variables (IBM, 2024). To retain more features, two alternatives were evaluated: using only the 80k tax burden, approximately the average Swiss salary, and applying PCA to the tax burdens (FSO, 2024; Kyriazos and Poga, 2023). PCA marginally improved performance (AUC +0.3%) but reduced interpretability, prompting the selection of the 80k model.

Figure 2 reveals the linearity assumption was generally met, except for average income and age-share variables, exhibiting low predictive power. Log and squared transformations were attempted; these did not improve the income variable and only slightly improved the age-share variables and performance (AUC +0.2%). Considering the interpretability loss associated with squared standardised variables, the original 80k model was retained.

Figure 2: Linearity in the Numeric Features of the Final Logit Model

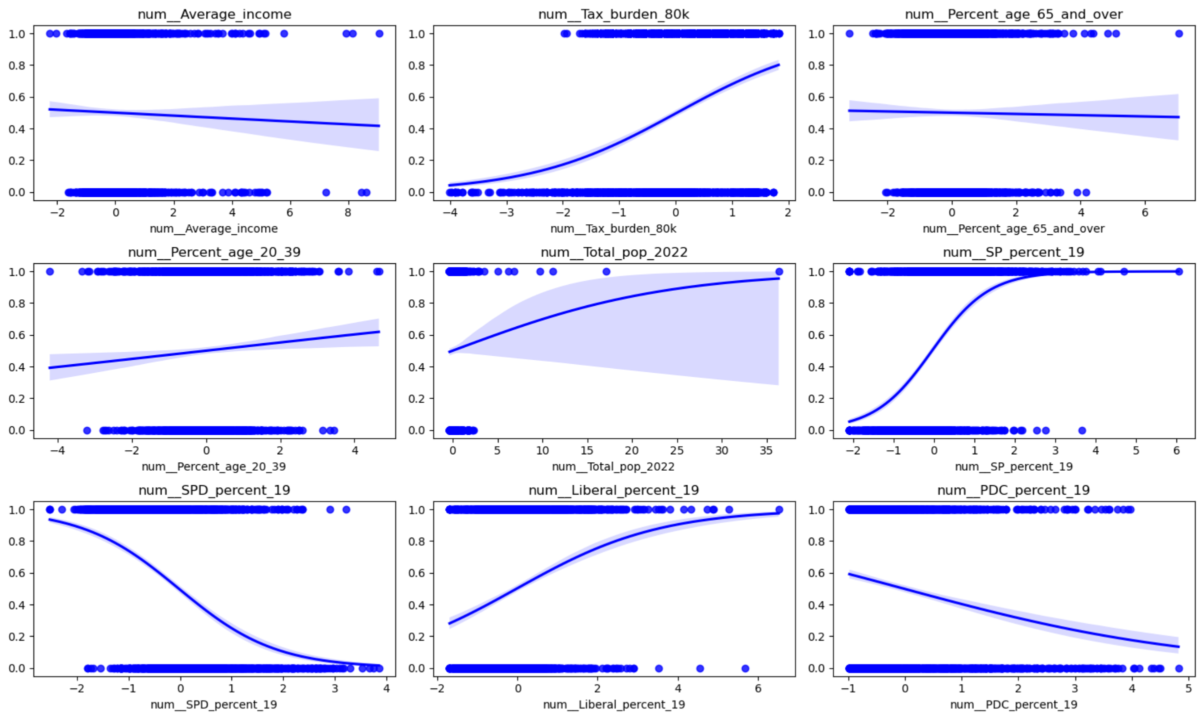
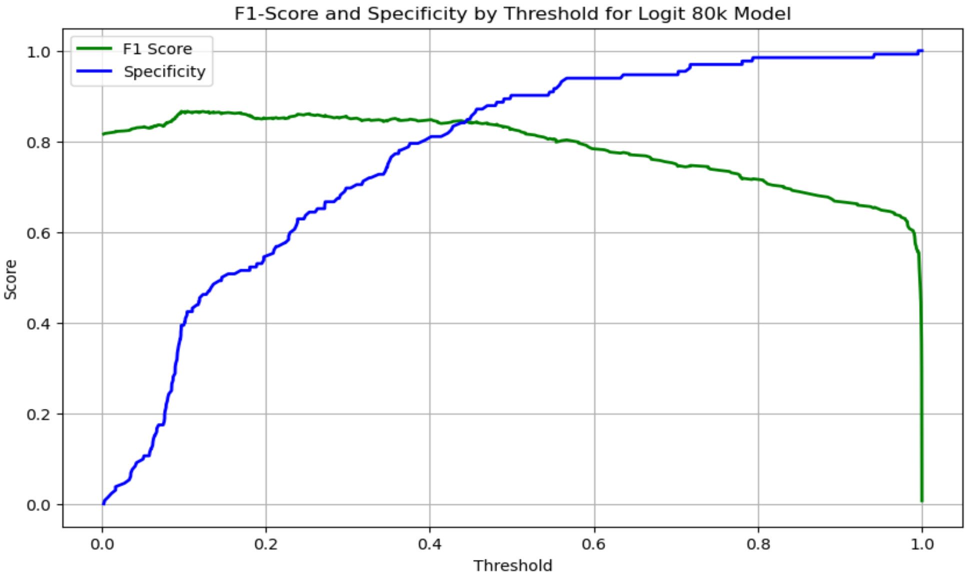


Figure 3 illustrates that lowering the threshold from 0.5 to 0.1 would increase the F1-score by approximately 0.03 but decrease specificity by 0.45. Accordingly, the default threshold was preferred.

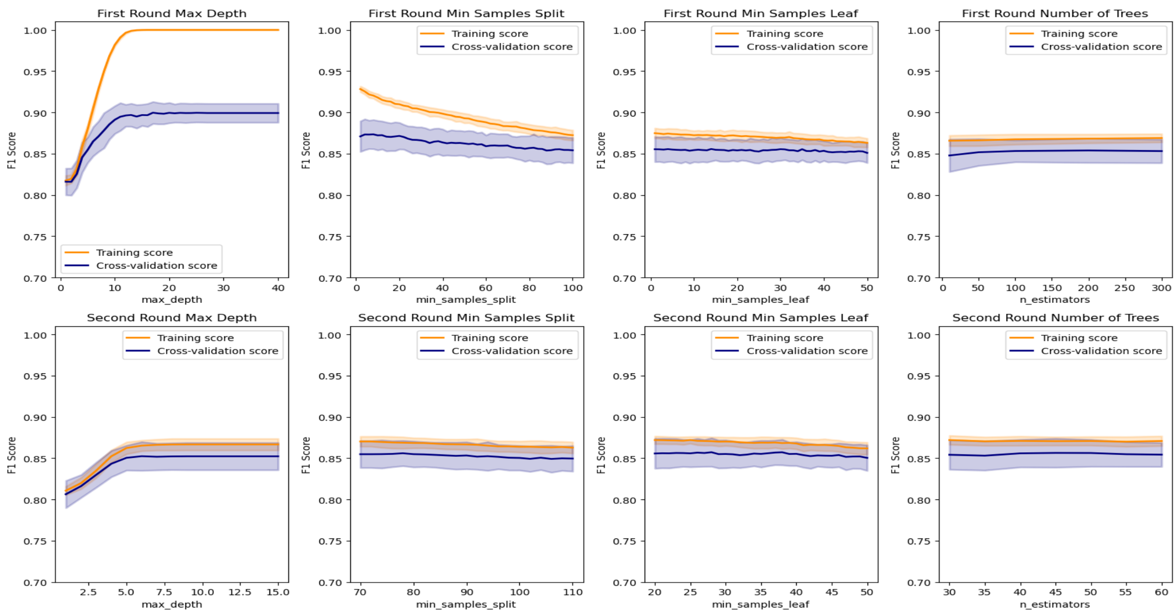
Figure 3: F1-Score and Specificity by Logit Model Threshold



RF Model Tuning

Bootstrap-aggregation makes RFs outlier-robust. Consequently, outliers were retained to enhance generalisability (Brence and Brown, 2006). Two rounds of inverse-proportionally reweighted validation curves were employed to refine the tree depth, minimum samples per split and leaf, and number of trees (see Figure 4). Objectives included minimising overfitting and standard deviation and maximising F1-scores. The F1-score, as the harmonic mean of precision and recall, was selected to balance the importance of false positives and negatives, ensuring the significance of both classes (Cahyana et al., 2019; Czakon, 2019).

Figure 4: Iterative Validation Curves for RF Hyperparameters



The validation curves reduced the hyperparameter combination search-range for subsequent Grid Search, enhancing the likelihood of optimal settings. Repeated stratified k-fold validation addressed k-fold validation's sensitivity to imbalance while minimising variance (Thölke et al, 2023). Similarly to the logit model, the default threshold was retained.

Performance Comparison

Figure 5 shows the RF models outperform the Logit models in all metrics throughout. Precision is consistently higher for class 1. In the final models, recall for class 0 increases, especially in the RF model, attributed to F1-score optimisation. The final Logit model's AUC decline suggests diminished sensitivity, possibly due to reduced precision for class 0, and recall and f-1score for class 1 after variable elimination. The baseline RF model exhibits extreme overfitting with perfect scores, mitigated in the final model, with a maximum difference of 0.3% between test and training. Conversely, Logit models show minimal under and overfitting, except for the final model's class 0 precision and f1-score.

Figure 5: Classification Reports, Specificities, and AUCs

Models & Classes	Precision	Recall	F1-Score	Specificity	ROC-AUC
Baseline Logit Class 0	0.690 (0.760)	0.730 (0.760)	0.710 (0.760)	0.735 (0.756)	0.902 (0.921)
Baseline Logit Class 1	0.880 (0.890)	0.850 (0.890)	0.870 (0.890)	0.735 (0.756)	0.902 (0.921)
Final Logit Class 0	0.600 (0.810)	0.900 (0.910)	0.720 (0.850)	0.902 (0.906)	0.885 (0.913)
Final Logit Class 1	0.940 (0.890)	0.730 (0.780)	0.830 (0.830)	0.902 (0.906)	0.885 (0.913)
Baseline RF Class 0	0.720 (1.000)	0.750 (1.000)	0.740 (1.000)	0.750 (1.000)	0.924 (0.999)
Baseline RF Class 1	0.890 (1.000)	0.870 (1.000)	0.880 (1.000)	0.750 (1.000)	0.924 (0.999)
Final RF Class 0	0.660 (0.690)	0.920 (0.910)	0.770 (0.790)	0.924 (0.907)	0.939 (0.949)
Final Logit Class 1	0.960 (0.950)	0.790 (0.820)	0.870 (0.880)	0.924 (0.907)	0.939 (0.949)

Note: The Test scores are listed first, with the Train scores in brackets.

Feature Importance

Figure 6 illustrates the final logit model's coefficients, indicating features' relative impact on predictions. Key predictors include the French and Italian regions, with odds of municipalities voting in favour approximately 900 and 330 times higher than in the German region, respectively. Municipalities with higher percentages of votes for the socialist party, and those in rural or urban areas compared to semi-urban areas, are more likely to vote in favour. The opposite holds for municipalities with higher average incomes, percentages of younger adults, and shares of retirees (65+).

Figure 6: Feature Importance of the Final Logit Model (Train Set)

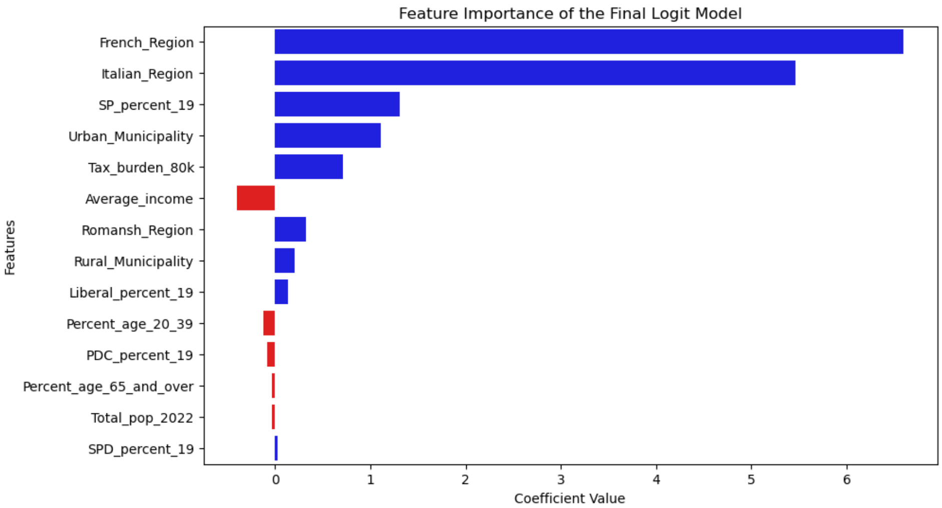
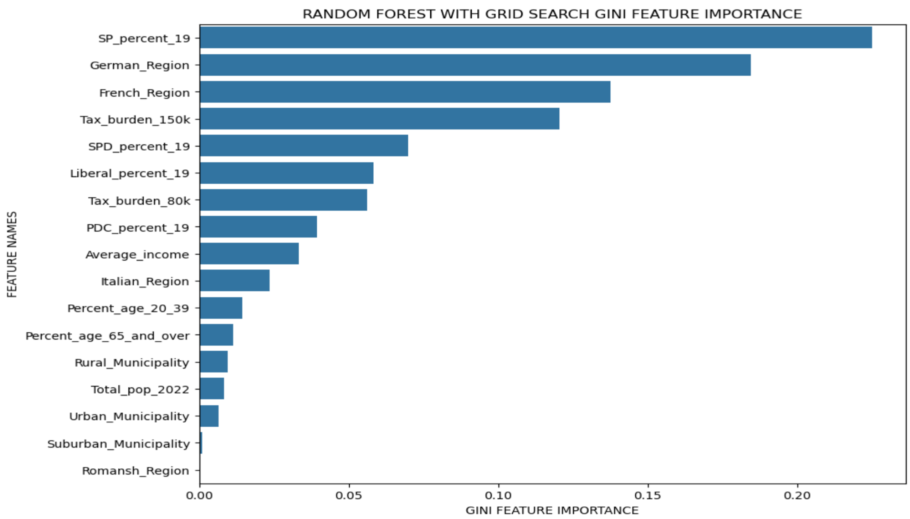


Figure 7 highlights features' contribution to reducing RF trees' impurity (Biau and Scornet, 2016). Socialist party support emerges as the most effective in separating classes, followed by the German and French regions, tax burdens, and other-party support.

Figure 7: GINI Feature Importance



SHAP values quantify each feature's marginal impact on the model by comparing predictions with and without the feature (Awan, 2023). The sum of SHAP values for all features plus the expected model output equals the prediction (Lundeberg, 2018). In Figure 8, positive SHAP values increase the likelihood of support while negatives reduce it, revealing how feature values affect the final RF model outcome. Higher socialist party backing and shares of younger and older age demographics increase the likelihood of predicted support. Against initial expectations, this also holds for tax burdens. Conversely, higher support for the SPD and PDC parties, and belonging to the German region, lower it. Notably, the effects of age groups are reversed compared to logit coefficients.

Figure 8: SHAP Feature Importance for Class 1 (Train Set)

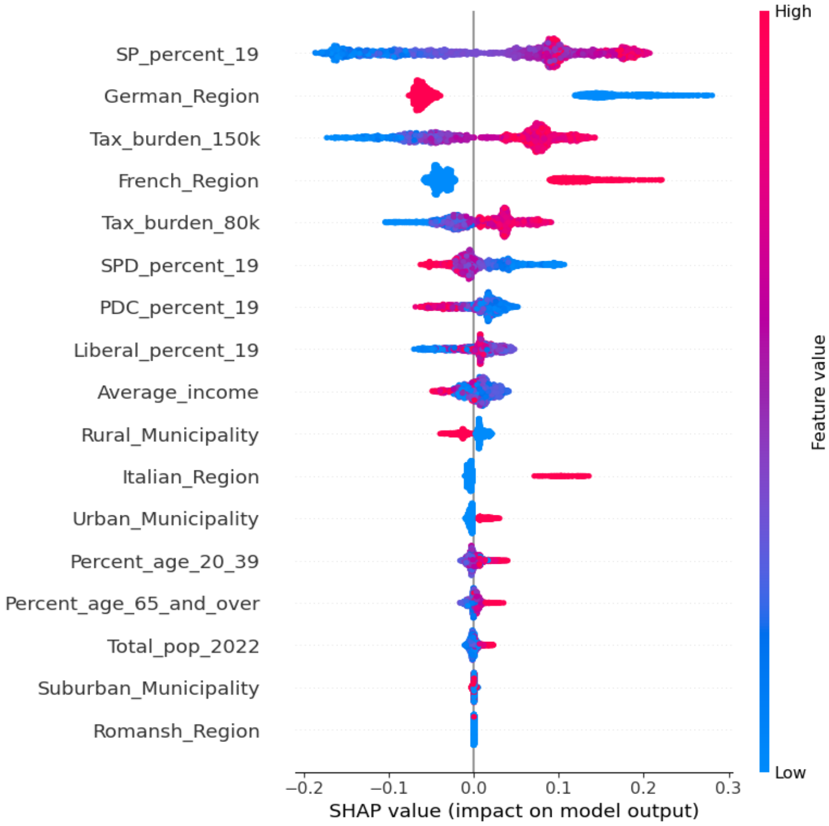


Figure 9 reveals the most influential support factors for Zug. Its sub-national average tax burdens emerge as the top predictors of unfavourable votes, followed by its location in the German region and average income, more than double the national average. Conversely, the 12.7% and 20.6% vote shares for the socialist and SPD parties increase the likelihood of support. However, these factors are insufficient to counterbalance the negative effects, resulting in an 18% support probability and (correct) prediction of no support.

Figure 9: SHAP Force Plot for Zug

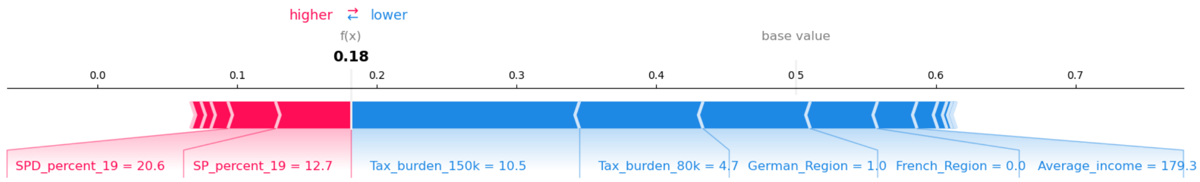
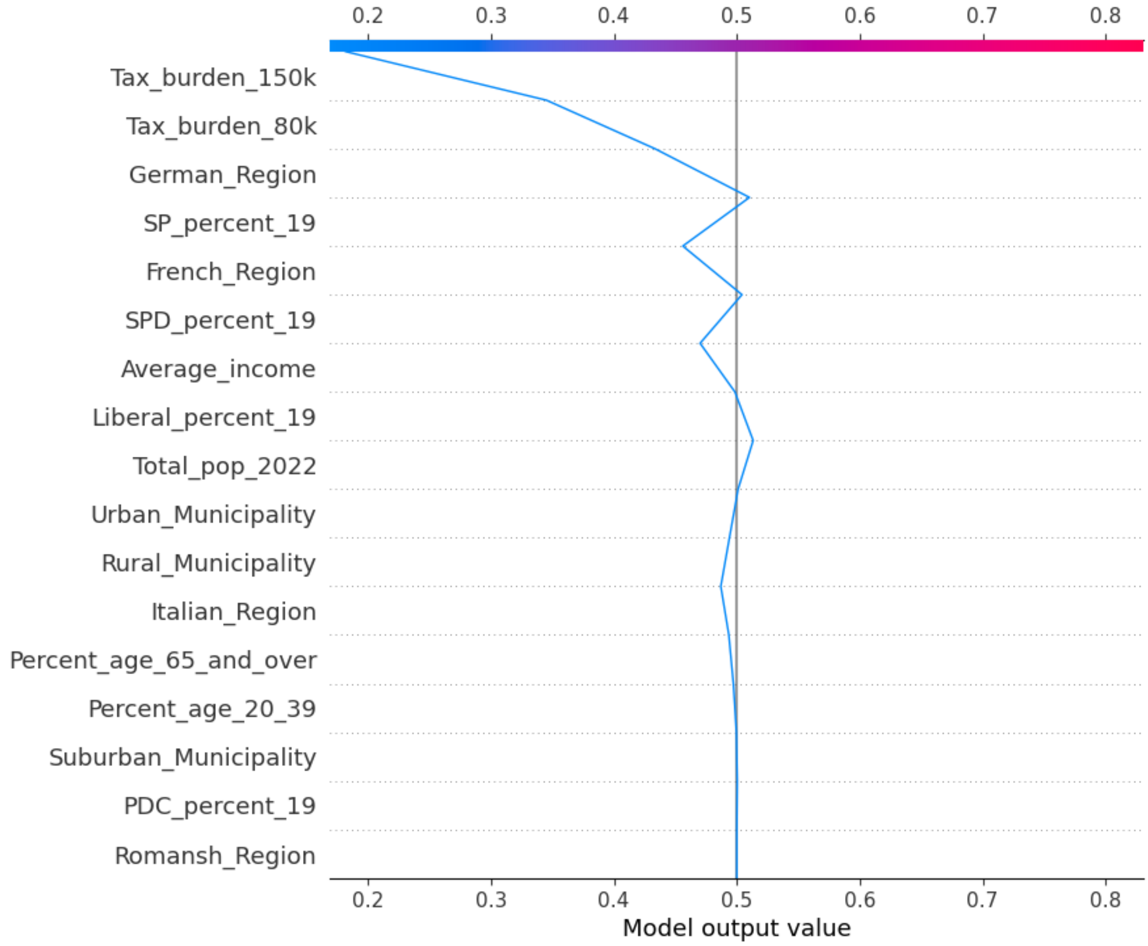


Figure 10 supplements Figure 9's insights with the cumulative impact of *all* features on the model's prediction, highlighting that 6 out of 17 features predominantly contributed to the final prediction.

Figure 10: Decision Plot for Zug



Limitations

Logit models' linearity assumption might not capture complex patterns in voting data (Ansolabehere and Hersh, 2011). Multicollinearity led to dropping the 150k tax burden variable, highlighted as a key predictor in RF models. Continuous variables' Gini importance could be upwardly biased, as they can be split in many ways to reduce impurity, even if weakly linked to outcomes (Loecher, 2020; Li, 2022). Feature importance relies on the model's performance; errors can undermine its insights. Conclusions from the final logit model should be approached with caution, 8 out of 14 coefficients being insignificant at conventional levels[1].

Referendum results provide insights into political dynamics but have limited applicability for understanding attitudes' in the broader population, excluding the non-Swiss. Nonetheless, studies suggest referenda generally reflect majority opinions. (Leininger, 2017).

Conclusion

RFs outperform logit models across all metrics, with political leanings, regional location, and tax burdens as key predictors. Socialist party support and the municipality's region—French, German or Italian—are particularly influential. However, municipality types' and age groups' influence is uncertain, varying between models and feature importance assessments.

Cross-validation could be employed to ensure logit results are not sample-dependent. Switching the RF splitting criterion from Gini Index to Entropy could improve performance and impact feature importance. RF prediction accuracy could be enhanced by utilising tree-level weights or training set resampling instead of tree prediction averages (Biau and Scornet, 2016). Time-variant predictors like recent changes in poverty or party support could capture trends, enhancing findings' future relevance. Utilising past pension-referendum results as predictors could highlight local vote tendencies. Classifying municipalities into high, medium, and low likelihoods of support or reject would enable swing areas' identification.

Further research could use pre- and post-referendum surveys to gain deeper insights into individual-level behaviours and characteristics. While the current method suits within-municipality campaigns, survey-based approaches could enhance targeted digital campaigns.

[1] Cf. Appendix 1.

Appendix

Appendix 1: Summary Table of the Final Logit Model (Train Set)

Variables	Coefficient	Std. Error	P-Value	[0.025	0.975]
Constant	-0.840	0.093	0.000	-1.022	-0.658
Average Income	-0.401	0.137	0.003	-0.671	-0.132
Tax Burden 80k	0.714	0.087	0.000	0.543	0.886
Percent Age 65+	-0.033	0.095	0.728	-0.219	0.153
Percent Age 20-39	-0.123	0.095	0.195	-0.310	0.063
Total Pop. 2022	-0.031	0.056	0.580	-0.140	0.078
SP Percent 19	1.308	0.112	0.000	1.089	1.526
SPD Percent 19	0.030	0.131	0.819	-0.0227	0.287
Liberal Percent 19	0.139	0.103	0.178	-0.063	0.341
PDC Percent 19	-0.083	0.096	0.385	-0.271	0.105
French Region	6.510	1.081	0.000	4.482	8.718
Italian Region	5.467	1.040	0.000	3.428	7.506
Romansh Region	0.330	0.819	0.687	-1.275	1.936
Urban Municipality	1.110	0.186	0.000	0.745	1.476
Rural Municipality	0.209	0.157	0.184	-0.100	0.517

Reference List

Alabrese, E., Becker, S.O., Fetzer, T. and Novy, D. (2019). Who voted for Brexit? Individual and regional data combined. *European Journal of Political Economy*, [online] 56(56), pp.132–150. doi:https://doi.org/10.1016/j.ejpoleco.2018.08.002.

Ansolabehere, S. and Hersh, E. (2011). *Gender, Race, Age, and Voting: A Research Note*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1901654.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), pp.197–227. doi:https://doi.org/10.1007/s11749-016-0481-7.

Bonoli, G. and Häusermann, S. (2009). WHO WANTS WHAT FROM THE WELFARE STATE? *European Societies*, 11(2), pp.211–232. doi:https://doi.org/10.1080/14616690801942116.

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

Brence, J.R. and Brown, D.E. (2006). *Improving the Robust Random Forest Regression Algorithm*. [online] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0fff413767a2a63b1b8df71dccffb8ba40dc6e84 [Accessed 19 Apr. 2024].

Brunner, B. and Kuhn, A. (2018). Immigration, Cultural Distance and Natives' Attitudes Towards Immigrants: Evidence from Swiss Voting Results. *Kyklos*, 71(1), pp.28–58. doi:https://doi.org/10.1111/kykl.12161.

Cahyana, N., Khomsah, S. and Aribowo, A.S. (2019). *Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICSITech46713.2019.8987499.

Carreras, M. and Castañeda-Angarita, N. (2013). Who Votes in Latin America? A Test of Three Theoretical Perspectives. *Comparative Political Studies*, 47(8), pp.1079–1104. doi:https://doi.org/10.1177/0010414013488558.

Czakon, J. (2019). *F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?*[online] neptune.ai. Available at: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc.

Engist, D. (2021). *Swiss municipality mergers since 2000: A dataset to easily aggregate municipality-level data to 2021 municipalities*. [online] www.research-collection.ethz.ch. doi:https://doi.org/10.3929/ethz-b-000492382 The SMMT is updated on a monthly basis and can currently aggregate municipality data up to march 2024.

Federal Council - VoteInfo (2024). *VoteInfo*. [online] www.admin.ch. Available at: https://www.admin.ch/gov/fr/accueil/documentation/votations/voteinfo.html [Accessed 19 Apr. 2024]. VoteInfo is an application developed by the Federal Council and Federal Statistical Office to track vote results and get reliable background information.

Federal Statistical Office (FSO) (2019a). *Karte: Force du 'PLR.Les Libéraux-Radicaux' (PLR), en 2019 [Communes (politique)]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/12/fr/14581_14573_14572_242/23153.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2019b). *Karte: Parti le plus fort en 2019 [Communes (politique)]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/12/fr/14580_14573_14572_242/23151.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2019c). *Karte: Participation, en 2019 [Communes (politique)]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/12/fr/14591_14573_14572_242/23175.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2023a). *Karte: Les 4 régions linguistiques de la Suisse par commune, en 2020 [Communes]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/13/fr/17138_17137_235_227/26599.html.

Federal Statistical Office (FSO) (2023b). *Karte: Population résidante permanente de 20 à 39 ans, en 2022 [Communes]*. [online] www.atlas.bfs.admin.ch. Available at:

https://www.atlas.bfs.admin.ch/maps/13/fr/17500_3008_3005_70/27110.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2023c). *Karte: Population résidante permanente de 65 ans ou plus, en 2022 [Communes]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/13/fr/17503_3012_3005_70/27119.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2024a). Charge fiscale* des célibataires**, revenu brut de 80'000 francs, en 2022 [Communes]. *Federal Statistical Office (FSO)*. [online] doi:https://doi.org/maps/13/fr/17827_8287_8282_8281/27601.html.

Federal Statistical Office (FSO) (2024b). *Karte: Charge fiscale* des célibataires**, revenu brut de 150'000 francs, en 2022 [Communes]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/13/fr/17828_8287_8282_8281/27602.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2024c). *Karte: Initiative populaire 'Mieux vivre à la retraite (initiative pour une 13e rente AVS)', votation du 03.03.2024 [Communes (politique)]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/12/fr/17887_17886_15863_259/27647.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2024d). *Karte: Revenu net moyen* par contribuable, en 2020 [Communes]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/13/fr/17823_9161_8282_8281/27593.html [Accessed 5 Apr. 2024].

Federal Statistical Office (FSO) (2024e). *Karte: Typologie des communes 2020 en 9 catégories [Communes]*. [online] www.atlas.bfs.admin.ch. Available at: https://www.atlas.bfs.admin.ch/maps/13/fr/17845_17843_3191_227/27616.html [Accessed 5 Apr. 2024].

Fernandez, A., Garcia, S., Herrera, F. and Chawla, N.V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, pp.863–905. doi:https://doi.org/10.1613/jair.1.11192.

Fernandez, J.J. and Jaime-Castillo, A.M. (2012). Positive or Negative Policy Feedbacks? Explaining Popular Attitudes Towards Pragmatic Pension Policy Reforms. *European Sociological Review*, 29(4), pp.803–815. doi:https://doi.org/10.1093/esr/jcs059.

IBM (2024). *What is lasso regression? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/lasso-regression#:~:text=Lasso%20regression%20can%20handle%20some [Accessed 19 Apr. 2024].

Kovacs, M. and Uwe, S. (2022). A deep learning approach to predict individual internet voting use based on electoral register data. *2022 International Conference on Digital Government Technology and Innovation*. [online] doi:https://doi.org/10.1109/dgti-con53875.2022.9849183.

Kyriazos, T. and Poga, M. (2023). Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, [online] 13(3), pp.404–424. doi:https://doi.org/10.4236/ojs.2023.133020.

Leininger, A. and Heyne, L. (2017). How representative are referendums? Evidence from 20 years of Swiss referendums. *Electoral Studies*, 48, pp.84–97. doi:https://doi.org/10.1016/j.electstud.2017.05.006.

Li, S. (2022). *Best Practice to Calculate and Interpret Model Feature Importance*. [online] Medium. Available at: https://towardsdatascience.com/best-practice-to-calculate-and-interpret-model-feature-importance-14f0e11ee660#:~:text=The%20default%20feature%20importance%20is.

Loecher, M. (2020). From unbiased MDI Feature Importance to Explainable AI for Trees. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2003.12043.

Lundberg, S. (2018). *An introduction to explainable AI with Shapley values — SHAP latest documentation*. [online] Readthedocs.io. Available at: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html.

Marozzo, F. and Bessi, A. (2017). Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, 8(1). doi:https://doi.org/10.1007/s13278-017-0479-5.

Montfort, S. (2023). Key predictors for climate policy support and political mobilization: The role of beliefs and preferences. *PLOS climate*, 2(8), pp.e0000145–e0000145. doi:https://doi.org/10.1371/journal.pclm.0000145.

OpenAI (2024). *ChatGPT*. [online] chat.openai.com. Available at: https://chat.openai.com/ ChatGPT was used to debug my code.

Péclet, J.-C. (2010). Vaux-sur-Morges, c'est si dur d'être riche! - Le Temps. *www.letemps.ch*. [online] 12 Nov. Available at: https://www.letemps.ch/suisse/vauxsurmorges-cest-dur-detre-riche#:~:text=L [Accessed 19 Apr. 2024].

Perlato, A. (n.d.). *Deal Multicollinearity with LASSO Regression - Andrea Perlato*. [online] www.andreaperlato.com. Available at: https://www.andreaperlato.com/mlpost/deal-multicollinearity-with-lasso-regression/.

Pillai N. , V. and A. , R.M. (2024). *Perfect Multicollinearity and Dummy Variable Trap: Explaining the Unexplained*. [online] mpra.ub.uni-muenchen.de. Available at: https://mpra.ub.uni-muenchen.de/id/eprint/120376 [Accessed 19 Apr. 2024].

Prinzie, A. and Van den Poel, D. (2008). Random Forests for multiclass classification: Random MultiNomial Logit. *Expert Systems with Applications*, 34(3), pp.1721–1732. doi:https://doi.org/10.1016/j.eswa.2007.01.029.

Qazi, N. and Raza, K. (2022). Effect of Feature Selection, SMOTE and under Sampling on Class Imbalance Classification. *International Conference on Business Analytics for Technology and Security (ICBATS)*. [online] doi:https://doi.org/10.1109/uksim.2012.116.

Riekhoff, A. (2020). Pension reforms, the generational welfare contract and preferences for pro-old welfare policies in Europe. *Social Policy & Administration*. doi:https://doi.org/10.1111/spol.12678.

Sadaiyandi, J., Arumugam, P., Sangaiah, A.K. and Zhang, C. (2023). Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced

Dataset. *Electronics*, [online] 12(21), p.4423. doi:https://doi.org/10.3390/electronics12214423.

Swiss Center of Expertise in the Social Sciences (FORS) (2018). *MOSAiCH | FORS*. [online] FORS. Available at: https://forscenter.ch/projects/mosaich/.

Swiss Ratio and Television (SRT) (2024a). *La 13ème rente AVS, ça change quoi pour vous?* [online] rts.ch. Available at: https://www.rts.ch/info/suisse/2024/article/la-13eme-rente-avs-ca-change-quoi-pour-vous-28435791.html [Accessed 15 Apr. 2024].

Swiss Ratio and Television (SRT) (2024b). *La Suisse accepte une 13e rente AVS, la Suisse romande championne du 'oui'*. [online] rts.ch. Available at: https://www.rts.ch/info/suisse/2024/article/la-suisse-accepte-une-13e-rente-avs-la-suisse-romande-championne-du-oui-28424839.html [Accessed 15 Apr. 2024].

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J. and Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, [online] 277, p.120253. doi:https://doi.org/10.1016/j.neuroimage.2023.120253.

van Dalen, H.P., Henkens, K. and Oude Mulders, J. (2019). Increasing the Public Pension Age: Employers' Concerns and Policy Preferences. *Work, Aging and Retirement*, 5(3), pp.255–263. doi:https://doi.org/10.1093/workar/waz004.

Xi, L., Guo, Z., Yang, X. and Ping, Z. (2023). Application of LASSO and its extended method in variable selection of regression analysis. *PubMed*, 57(1), pp.107–111. doi:https://doi.org/10.3760/cma.j.cn112150-20220117-00063.

Yadav, C.S. and Sharan, A. (2020). Feature Learning Using Random Forest and Binary Logistic Regression for ATDS. *Applications of Machine Learning*, pp.341–352. doi:https://doi.org/10.1007/978-981-15-3357-0_22.

Zhao, X., Yan, X., Yu, A. and Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, 20, pp.22–35. doi:https://doi.org/10.1016/j.tbs.2020.02.003.