Larry-On-Ballots - Progress Report

This project looks at referendum results in Switzerland at municipality level. First, we will focus on predicting the outcome of the face covering referendum with a series of machine learning models. Second, we will analyze the efficiency of these models in predicting outcomes for a politically similar referendum.

1) Exploratory Data Analysis

We have a total of 2,179 observations, one per municipality. Regarding the voting outcome for the face covering referendum, the results follow the distribution shown in Figure 1, with a mean of 57.33% 'Yes' and a standard deviation of 7.67 percentage points. This reflects 83.1% of municipalities with a 'Yes' majority of votes. Find the summary statistics for the individual features considered in the Appendix I.

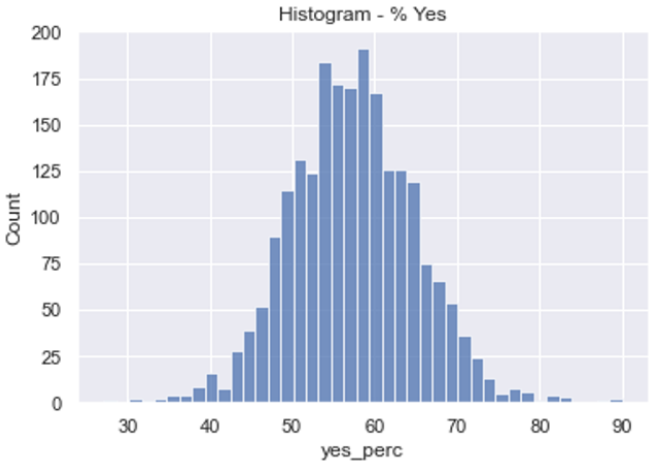


Figure 1 - Histogram of Referendum Outcome at Municipality Level

For further exploration, we have created a correlation matrix between all the features. Due to the large dimensionality, we created an excel file with the results, which can be found in the file *data/correlations.xlsx.* The figure in Appendix II depicts a sample of the correlation matrix highlighting possible collinearity concerns. The figure in Appendix III depicts the results of a Variance Inflation Factor (VIF) test performed as an additional method to identify collinearity. Many of the collinearity issues are due to the values of features being correlated to the population level, (e.g., registered voters). Therefore, we will be looking at adjusting these variables per capita. The existence of collinearity can incorrectly increase independent variables' standard error reducing the reliability of the model and increasing uncertainty on the statistical significance of each variable.

To better understand the relationship of the features with the outcome variable, we have also plotted a scatterplot for each (Appendix IV). We found no significant outliers that need to be treated. We also analyzed missing data. While the referendum data does not contain any missing data, we find missing data for ten features in the demographics data. We are most concerned about missing data in 'social aid' for almost 25% of the observations.

2) Baseline ML models

After the initial data cleaning and exploration, we implemented three baseline machine learning models to predict the outcomes and test the accuracy of the models. We used the Scikit-learn Python Machine learning library for our implementation:

A. K-nearest neighbors (KNN) classification
B. Decision tree classification
C. Logistic regression

Of the attributes in the processed data, the following attributes were initially shortlisted for the modeling: population_density, foreigner_percentage, age_percentage_between_20_64, agriculture_surface_perc, participation_rate (Refer definitions in Appendix V). We then defined the output variable as a binary indicator if the referendum was passed in a municipality. Thus, we have output label y = 1 if 'yes_perc' > 50 and y = 0 otherwise.

The detailed results including Accuracy, Confusion matrix and Classification report along with the associated error plots were included in the file *models/baseline_analysis.ipynb*. Here are the main results:

1. KNN classifier
    a. Development data:
        i. Accuracy score: 0.825 and Error rate was minimum at K = 13 with increasing trend after K > 20
    b. Test data:
        i. Accuracy score: 0.871 and Error rate was minimum at K = 25 and stable with minor variations for K > 25
2. Decision tree with Gini Index
    a. Common trends:
        i. ‘population_density’, ‘participation_rate’ and ‘agriculture_surface_perc’ emerged as key attributes for classification
        ii. ‘Population_density’ emerged a a key factor for classification as municipalities with low population density (rural) tend to pass the referendum
    b. Accuracy score for Development data is 0.852 and for Test data is 0.875
3. Logistic regression: Accuracy score for Development data is 0.843 and for Test data is 0.885

3) Future Plans

Since the base rate is 83.1%, we see that it is difficult to train a model with better accuracy. Thus, going forward, there are a few things we would like to explore before trying to generalize our model.

First, with this approach of predicting the referendum outcomes of certain municipalities with the result of others (for the same referendum), we want to investigate the following, hoping that we find a way to increase our accuracy.

- Regularization: It seems like our model could suffer from overfitting. Thus, we will try some regularization techniques and check if we get better results.
- Missing Values: There are certain missing values in our demographic datasets. For some, filling them with the appropriate data. When a party’s percentage of seats in a municipality is missing for instance, that means that it is not present and thus, we should

fill it with 0. However, in many other instances, inferences are more difficult. For example, the percentage of the population receiving social aid is much more difficult to infer. Should we use the mean, the median, KNN, or a regression to infer the data? We will explore those options and use the one giving us the highest testing accuracy.

- Feature engineering: As mentioned above, there are certain variables, which could become a problem going forward. A small municipality, for instance, will have fewer employment positions than a large one in absolute terms. Therefore, absolute employment positions will not only inform about the job market, but also about the population size. We think that transforming certain features from absolute into per-capita variables will allow the models to perform better by focusing on one mechanism at a time.
- Different elections: Our baseline models were not too performant with the face covering referendum. If time permits, we will expand our analysis to other votes as well to see if there are certain referendums we can predict better. For example, we could explore different types of voting objects; environment, national security, culture. This would allow us to further understand how generalizable our method is.
- Cross-validation: We will implement cross-validation to see if it allows us to get better results.
- More complex models: Neural nets and random forests might give us better testing accuracy.

While predicting the voting outcome of certain municipalities with the votes of others for the same object is interesting, it does not allow us to predict voting patterns of future referendums. To do so, we might want to train a better model.

The issue is that referendums are by their nature different from each other. One cannot simply train a model on a few referendums and hope for accurate predictions. One must find a way to classify referendums to then train a model on only certain past ones to predict future voting patterns. To do so, we might exploit the fact that the major political parties always publish recommendations (yes, no, free) about what to vote on a given object in advance. We hypothesize that a municipality with certain characteristics will vote similarly as municipalities in previous referendums with similar characteristics when the party recommendations were the same.

Hence, to predict a future referendum, we should be able to train machine learning models on previous referendums where the party recommendations were the same. If time permits, we will try to implement such a model.

APPENDIX

I. Summary statistics - Demographics and Referendum Results

	count	mean	std	min	25%	50%	75%	max
population	2172.0	3962.261971	12875.284703	32.0	720.2500	1555.50	3834.250	420217.0
population_variation	2172.0	9.209899	11.342446	-30.3	2.4000	7.95	14.400	92.8
population_density	2172.0	437.707643	792.787000	1.0	80.7500	185.00	467.000	12811.0
foreigner_percentage	2172.0	16.947744	9.702113	0.0	9.6000	15.20	23.125	57.8
age_percentage_less_20	2172.0	20.471455	3.367026	2.1	18.7000	20.60	22.500	37.2
age_percentage_between_20_64	2172.0	60.154880	3.197797	39.5	58.3000	60.25	62.200	81.1
age_percentage_more_64	2172.0	19.374540	4.413891	6.5	16.4000	19.00	21.600	40.3
marriage_rate	2172.0	4.168692	2.590607	0.0	2.8000	4.00	5.200	36.7
divorce_rate	2172.0	1.918692	1.762745	0.0	1.0000	1.80	2.500	38.5
birth_rate	2172.0	9.224401	3.879388	0.0	7.0000	9.30	11.200	47.0
death_rate	2172.0	7.624448	3.952898	0.0	5.5000	7.30	9.400	57.7
private_households	2172.0	1754.744936	6161.564561	14.0	306.7500	673.50	1636.750	204411.0
avg_household_size	2172.0	2.323849	0.193711	1.5	2.2000	2.30	2.400	3.3
total_surface	2172.0	18.415746	33.450123	0.3	4.4000	8.30	16.800	438.6
housing_and_infrastructure_surface	2172.0	14.925046	14.776796	0.1	5.8000	10.00	18.900	97.3
housing_and_infrastructure_surface_variation	2172.0	26.891805	30.481848	-37.0	8.7500	18.00	35.000	395.0
agriculture_surface_perc	2172.0	45.981584	19.259597	0.0	33.0000	47.65	60.800	91.5
agriculture_variation_surface_perc	2172.0	-39.160221	61.395848	-616.0	-44.0000	-20.00	-9.000	95.0
forest_surface_perc	2172.0	32.484484	16.096235	0.0	20.4000	31.00	42.800	88.2
unproductive_surface_perc	2172.0	6.609300	14.006514	0.0	0.3000	1.10	4.700	95.0
employment_total	2076.0	2498.994701	14026.548188	11.0	215.0000	598.00	1620.000	491193.0
employment_primary	2136.0	75.572566	82.941617	0.0	24.0000	47.00	96.000	751.0
employment_secondary	2115.0	516.084161	1417.563517	0.0	40.0000	149.00	476.000	34946.0
employment_tertiary	2166.0	1845.255771	12618.036130	4.0	97.0000	313.50	986.250	462410.0
establishments_total	1875.0	346.909867	1411.799868	9.0	69.0000	133.00	287.500	45057.0
establishments_primary	2022.0	26.276954	28.328050	0.0	9.0000	16.00	32.000	272.0
establishments_secondary	2009.0	47.456944	102.259616	0.0	11.0000	24.00	52.000	2637.0
establishments_tertiary	2151.0	250.039981	1227.898035	4.0	33.0000	74.00	186.000	42368.0
empty_housing_units	2172.0	1.963315	1.694620	0.0	0.7875	1.55	2.680	13.1
new_housing_units_per_capita	2172.0	6.218370	8.603579	0.0	0.7000	3.40	8.300	96.0
social_aid_perc	1692.0	2.223286	1.617768	0.2	1.1000	1.75	2.800	11.2
PLR	2172.0	14.923343	9.179349	0.0	8.7000	13.80	20.000	69.7
PDC	2172.0	12.609162	13.882447	0.0	2.3000	7.90	18.300	79.5
PS	2172.0	13.621731	6.374857	0.0	9.2000	13.30	17.700	49.3
UDC	2172.0	31.226750	13.469845	0.0	21.4000	30.40	39.500	84.1
PEV_PCS	2172.0	2.336096	2.638545	0.0	0.5000	1.60	3.400	30.4
PVL	2172.0	6.255479	4.025103	0.0	3.4000	6.10	8.900	23.8
PBD	2172.0	2.853085	4.862200	0.0	0.0000	1.00	3.200	64.6
PST_Sol	2172.0	0.767495	1.986292	0.0	0.0000	0.00	0.500	32.0
PES	2172.0	10.988168	6.113617	0.0	6.9000	10.00	14.600	38.4
small_right_parties	2172.0	2.209438	4.222997	0.0	0.0000	0.60	2.225	27.2

	count	mean	std	min	25%	50%	75%	max
canton_id	2179.0	13.704452	8.458174	1.0	3.0	17.0	22.0	26.0
registered_voters	2179.0	2522.651675	7173.123155	30.0	530.0	1099.0	2563.5	234028.0
cast_ballots	2179.0	1297.025241	3874.621587	17.0	277.5	576.0	1310.0	128959.0
participation_rate	2179.0	53.257458	9.361085	19.5	46.8	52.1	58.9	94.2
blank_votes	2179.0	13.356586	48.254538	0.0	2.0	5.0	12.0	1203.0
invalid_votes	2179.0	4.139514	13.116165	0.0	0.0	0.0	2.0	202.0
valid_ballots	2179.0	1279.529142	3828.651792	16.0	275.5	568.0	1298.5	127806.0
yes	2179.0	655.045434	1393.930347	9.0	163.0	331.0	720.5	39255.0
no	2179.0	624.483708	2488.276033	4.0	110.0	241.0	560.0	88551.0
yes_perc	2179.0	57.332171	7.672309	27.1	52.2	57.2	62.3	90.0

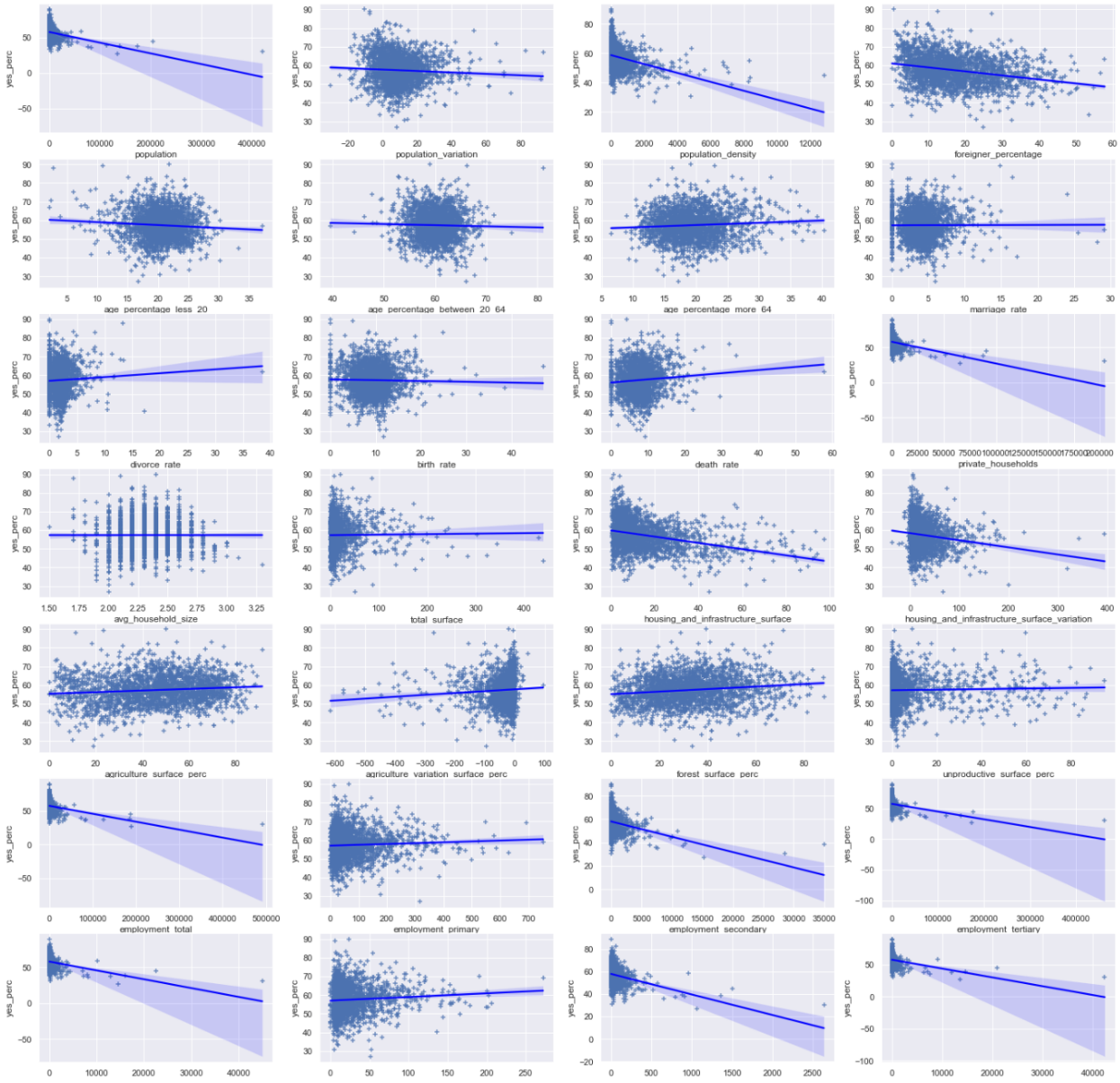
II. Abbreviated Correlation Matrix to identify possible collinearity.

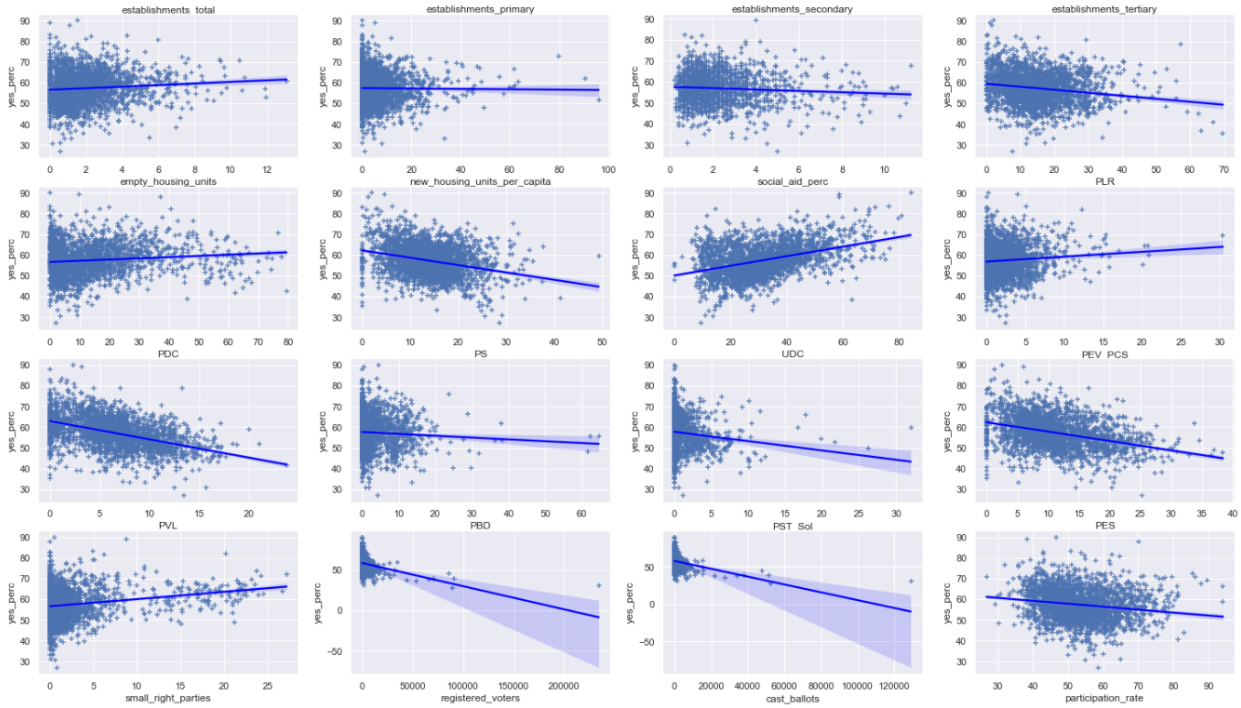
	population	population_density	private_households	total_surface	housing_and_infrastructure_surface	agriculture_variation_surface_perc	employment_total	employment_primary	employment_secondary	employment_tertiary	establishments_total	establishments_primary	establishments_secondary	establishments_tertiary
population	1.000	0.491	0.999	0.086	0.337	-0.194	0.974	0.146	0.837	0.966	0.984	0.113	0.957	0.979
population_density	0.491	1.000	0.475	-0.155	0.861	0.028	0.428	-0.125	0.441	0.398	0.515	-0.145	0.489	0.436
private_households	0.999	0.475	1.000	0.087	0.324	-0.189	0.979	0.136	0.838	0.972	0.985	0.105	0.952	0.982
total_surface	0.086	-0.155	0.087	1.000	-0.269	-0.807	0.075	0.386	0.090	0.069	0.094	0.429	0.138	0.082
housing_and_infrastructure_surface	0.337	0.861	0.324	-0.269	1.000	0.082	0.279	-0.185	0.350	0.255	0.324	-0.221	0.344	0.277
agriculture_variation_surface_perc	-0.194	0.028	-0.189	-0.807	0.082	1.000	-0.153	-0.349	-0.218	-0.138	-0.184	-0.344	-0.280	-0.167
employment_total	0.974	0.428	0.979	0.075	0.279	-0.153	1.000	0.098	0.815	0.998	0.986	0.066	0.915	0.987
employment_primary	0.146	-0.125	0.136	0.386	-0.185	-0.349	0.098	1.000	0.177	0.083	0.127	0.937	0.228	0.105
employment_secondary	0.837	0.441	0.838	0.090	0.350	-0.218	0.815	0.177	1.000	0.775	0.803	0.147	0.891	0.788
employment_tertiary	0.966	0.398	0.972	0.069	0.255	-0.138	0.998	0.083	0.775	1.000	0.984	0.051	0.896	0.986
establishments_total	0.984	0.515	0.985	0.094	0.324	-0.184	0.986	0.127	0.803	0.984	1.000	0.097	0.943	0.999
establishments_primary	0.113	-0.145	0.105	0.429	-0.221	-0.344	0.066	0.937	0.147	0.051	0.097	1.000	0.192	0.074
establishments_secondary	0.957	0.489	0.952	0.138	0.344	-0.280	0.915	0.228	0.891	0.896	0.943	0.192	1.000	0.930
establishments_tertiary	0.979	0.436	0.982	0.082	0.277	-0.167	0.987	0.105	0.788	0.986	0.999	0.074	0.930	1.000

III. Variance Inflation Factor values for each of the demographic variables.

	VIF	features
0	844.381903	population
1	2.903329	population_variation
2	9.833690	population_density
3	10.136636	foreigner_percentage
4	79916.603926	age_percentage_less_20
5	672634.575217	age_percentage_between_20_64
6	73142.485990	age_percentage_more_64
7	4.094369	marriage_rate
8	2.348625	divorce_rate
9	8.411329	birth_rate
10	6.296188	death_rate
11	863.769642	private_households
12	531.921483	avg_household_size
13	7.123426	total_surface
14	81854.349151	housing_and_infrastructure_surface
15	8.026231	housing_and_infrastructure_surface_variation
16	461021.346296	agriculture_surface_perc
17	7.590454	agriculture_variation_surface_perc
18	243763.683947	forest_surface_perc
19	44486.314936	unproductive_surface_perc
20	1200.236090	employment_total
21	13.874228	employment_primary
22	21.635748	employment_secondary
23	1055.144811	employment_tertiary
24	226.719107	establishments_total
25	15.040481	establishments_primary
26	43.236438	establishments_secondary
27	252.781598	establishments_tertiary
28	2.778755	empty_housing_units
29	1.798396	new_housing_units_per_capita
30	5.670716	social_aid_perc
31	15.522057	PLR
32	14.676447	PDC
33	16.336256	PS
34	45.937831	UDC
35	2.815561	PEV_PCS
36	7.243754	PVL
37	3.235819	PBD
38	2.064433	PST_Sol
39	13.155562	PES
40	2.688230	small_right_parties

IV. Scatterplots - Features and output





V. Definitions for shortlisted baseline model attributes

S. No	Attribute	Definition
1.	'population_density'	Population density per km² 2019
2.	'foreigner_percentage'	Foreigners in % 2019
3.	'age_percentage_between_20_64'	20-64 years old in 2019
4.	'agriculture_surface_perc'	Agricultural area in % 2004/09
5.	'participation_rate'	Turnout rate during elections
6.	‘yes_perc’	Vote percentage of ‘Yes’