Peer consulting report

Project reviewed:

Predicting votes in the Swiss National Council

December 2022

By: **Marc Bratschi**
bratschi.marc@gmail.com

For: **Kim Lan Vu & Emilie Zucchinetti**
kmln.vu@gmail.com, zucchinetti@outlook.com

Table of Contents

1 Project Description and objectives

This project aims to predict the votes in the Swiss National Council, the first chamber of the Swiss Parliament. The Parliament holds four regular sessions a year, in March, June, September and December. The Swiss National Council is composed of 200 members and the seats are allotted to the 26 cantons on the basis of their resident population. Each Council member belongs to one of the six parliamentary groups, representing the major parties in Switzerland. During each session, the Council members vote on hundreds of business items.

This project uses the data related to the sessions in 2022. The data was collected through the public API and three tables were used: the **Voting table**, which contains data about the votes (how each Council member voted on each items), the **Council member table**, which contains the personal data (such as gender, age, canton and parliamentary group affiliation) of the Council members and the **Business table**, which contains data about the vote topics (theme and responsible department). The business table is grouped into seven categories:

- Politics
- Law
- Social
- Energy, environment & country
- Finance & economy
- Culture
- Education & Science

Seven departments are responsible for these topics.

- EDA (Eidgenössisches Departement für auswärtige Angelegenheiten)
- EDI (Eidgenössisches Departement des Innern)
- EFD (Eidgenössisches Finanzdepartement)
- EJPD (Eidgenössisches Justiz- und Polizeidepartement)
- WBF (Eidgenössisches Departement für Wirtschaft, Bildung und Forschung)
- UVEK (Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation)
- VBS (Eidgenössisches Departement für Verteidigung, Bevölkerungsschutz und Sport)

2 Solutions

First, the data was cleaned and filtered. The data from the last 4 sessions were taken into account. The data was requested in batches to overcome limits. All categorical variables have been transformed to numerical variables (using boolean variables). The final dataset consists of 49’871 lines with 24 variables. Based on these attributes of the council members and depending on the vote topic the model tries to predict if the council votes yes or no for various political questions.

Several descriptive statistics were implemented to get an overview of some dimensions. The analysis of the votes related to topic category and responsible department shows that the count of votes is not evenly distributed. Obviously, the biggest part of votes concerns the four categories Finance & Economy, Politics, Social and Energy, Environment & Country. In addition, the biggest part of the votes concerns the department EFD. From totally 267 votes, more than 80 votes concerned the EFD.

Another interesting statistic shows that there seems to be a difference in left and right parties in relation to gender. Whereas, in right parties’ men seem to outnumber women, in left parties seem to be more women.

Several machine learning models have been applied. First, a **linear regression** was implemented with very little success. The R^2-Value of 0.06 does not indicate a linear relationship. For this reason, a more complex model had to be applied. The implementation of a **logistic regression, a random forest and K-means** all resulted in about 65 % accuracy. However, knowing the voting dataset contains 65% of “yes” answers, the algorithm could always predict a yes and still get correct results in 65% of cases. A good model should thus achieve a better accuracy than 65%. Finally, a neural network algorithm, a **multilayer perceptron**, has been applied, giving better predictions with a test score of 73%.

3 Suggestions for improvement

Data and approach

An obvious weakness of the model is that the votes and the political views of the council members are known, however, when voting on a political issue, it also depends on how the question is phrased. For instance, the descriptive statistic of the project shows that in political questions regarding education and science, council members from left parties vote yes with 100%. But if the question is phrased in favor of right-wing parties, left council members will probably vote no. For instance, when the question is "Should the education budget be increased?", left council members will probably vote yes. But if the question is "Should the education budget be cut?", left council members will probably vote no. Thus, when applying the model, great attention must be paid to how the political question has been formulated.

Another limitation could be that the political orientation of the council members. The policitical orientation is not necessarily a decisive factor, as council members can vote against the tendency of their political orientation. By consideration of additional data like personal interests, lobbying activities or alliances this uncertainty could be reduced.

Model and methodology

Another weak point concerns the linear regression. The linear regression is probably difficult to implement on that dataset. The model tries to predict the votes (yes or no) and abstentions. However, by assigning values to yes, no and abstention, a ranking between the observation is implicitly defined which should not exist.

To get a better feeling for the correlations between the variables, it makes sense to implement a confusion matrix. The confusion matrix helps to detect issues such as an algorithm that gets 65% of correct predictions by only predicting yes-votes. Without the confusion matrix it is unclear how the algorithm reached 65% correct predictions.

An addition, it could be helpful to cluster the data and then to take a closer look at the various clusters. By using clusters, it is possible to detect potential patterns and samples on which the model resulted in higher accuracy. The model should then be trained by implementing with different clusters or under changing parameters.