AIST4010
Project Report

Name: Kwok Chun Kiu (1155141911)

Prediction of Hong Kong Legislative Council election:
An attempt on machine political forecast

I. INTRODUCTION

With the popularization of artificial intelligence, society has advanced in automating: autonomous driving and computer vision, just to name a few. However, the industry is focused on making these technologies more accessible, instead of focusing on social usage, which aims to shape public opinions on government policy or public discourse. As the use of artificial intelligence and machine learning becomes more versatile, this project will attempt to apply these technologies to the social science subjects. By utilizing the socio-economic indices such as education level, income, employment rate and the vote counts of each candidate, we attempt to predict the political party a constituency will elect. Figure 1 provides a brief view of the structure of this project.

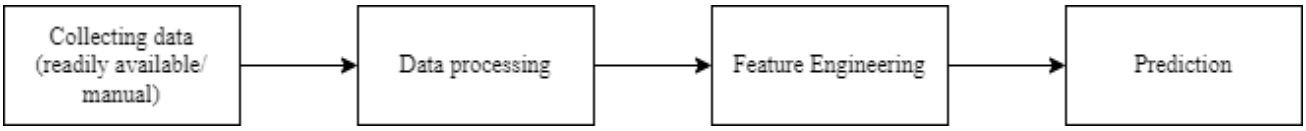

Fig. 1. Structure of this project.

II. RELATED WORKS

Several related works like [6, 7] can be seen, which provides inspiration for this project. However, they tackle this problem from different aspects. In [6], Gilbert first passed the constituency data to affinity propagation model for clustering, classifies the data by a decision tree, then merge the results to get the final prediction. While in [7], Jayawickrama used feature engineering and normalization to improve the prediction accuracy of the *k*-nearest neighbours model. Still, what they aim to achieve is to predict the winner in the constituency. Artificial intelligence is also used for political analysis during the election period. In the latest South Korean presidential election, SBS collaborated with Underscore to investigate polling results during the official “no polling results period”. [8] SBS also collaborated with data scientists to provide predicted winner of the election while the vote counting is underway. [9] Other advanced works include [5], which uses a hierarchical graph convolutional network to predict Australian election results, suggesting further investigations in this aspect. All in all, these works aim to find a convincing prediction to predict the voters’ preference, and thus

helping to shape public opinion.

III. Data Set

The data sets are taken from the 2011 census, 2016 by-census, the candidates' description, and vote counts from the Electoral Affairs Commission.

A. *2011 census statistics and 2016 by-census statistics*

The most part of the data sets are taken from the census statistics conducted in 2011 and 2016. [1, 2] This data contains 194 socioeconomic indices such as age, educational level, work force and median income. Census data compiles anonymized features where populations are grouped into geographical areas. The census data in these two data sets are listed according to the District Council Constituency Areas (DCCA), which are drawn for the District Council Election.

B. *Candidates' description*

These descriptions are published on the official websites for the Legislative Council Elections. [3, 4] Since this project aims to predict the political party instead of the candidate elected, only their political affiliation is concerned. For those who claims to be independent, unaffiliated or is affiliated to parties that did not obtain 5% of the vote, they will be identified as 'IND'.

C. *Vote Counts*

Votes counts are obtained from the official records of the Electoral Affairs Commission. These vote counts are grouped by voting stations. Although in the elections, the votes are accumulated and legislators are elected using proportional representation system; in this project, only the most voted party is concerned.

IV. Methodology

The project can mainly be divided into four parts: data pre-processing, coarse prediction, feature engineering, and refined prediction. In data pre-processing, we remove redundant columns that does not have any prediction power, such as the name and code of the constituencies. These data will then be passed for a rough prediction to evaluate the performance.

Since the data set is full of correlated or similar features, feature engineering is required to reduce the complexity of our composed model. In this step, we first calculate feature importance using the chi-square distribution, and only retain the top 30 features with highest importance. These features are then saved for refined prediction.

In each prediction stage, we normalize the data to ensure the distribution of the feature values does not sacrifice the accuracy of the model. The values are scaled using the standard score with respect to its corresponding column of feature values.

Then, with the use of various models such as decision trees and discriminant analysis, we wish to find comprehensible relationships among the features to explain the factors electors will consider when casting their votes. The baseline model is the logistic regression model, to which we will compare the results of other models.

A. *Logistic Regression and Multi-layer Perceptron*

Logistic regression and multi-layer perceptron (MLP) are the basic models in deep learning. They make use of all the features to classify samples. Using scikit-learn, we can implement logistic regression and multi-layer perceptron easily. A version of MLP is also implemented with PyTorch.

B. *Decision Tree and Random Forest Classifier*

Decision tree classifies data by choosing a best cut at each level, considering every feature. Random forest provides a more robust mechanism than decision trees by implementing bootstrap aggregating, or “bagging” as it is often called, so that the classification results is less sensitive to noise. Since these models are dividing data according to some given conditions, they are discriminative models. These models are implemented by scikit-learn.

C. *Linear Discriminant Analysis and Quadratic Discriminant Analysis*

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are two generative models that focus on the probability distribution of the data. Given the samples, they create a probability distribution function of each class. These models are implemented by scikit-learn.

V. RESULTS

A. *Prediction Results*

Since the elections in 2012 and 2016 are conducted the same way, we can use the election results in 2016 to evaluate the prediction of the model. We expect each constituency to be classified with the correct political party, and the aggregated result for all Legislative Council Geographical Constituency to be the same as the overall result. If there is any discrepancy, they will be measured by cross entropy. Figure 2 shows the prediction accuracy of the mentioned models.

B. *Analysis of the Results*

The results are, objectively, not promising: the accuracies are only around 50% for guessing the blocs, and only around one-third of the predictions are correct for the parties. This can be analysed in terms of external factors and implementation choices.

1) *External factors*: These are the factors we cannot control. For example, some data must be removed because the counting results of that constituency area is not explicitly listed, due to the arrangement of the counting stations. Also, the constituencies are amended every four years to address population changes, which mostly affects the predictions in urban areas. Also, political events may impart more influence on the voters than the underlying social factors.

2) *Implementation choices*: These are the choices made to enhance the comprehensibility of this project. For simplicity, we did not consider the characteristics of political parties, which is an information loss, and the results above shows that the model cannot distinguish the preferred parties or blocs clearly. Also, the geographical relation between each pair of constituencies is not considered, which may affect the result of neighbouring constituencies.

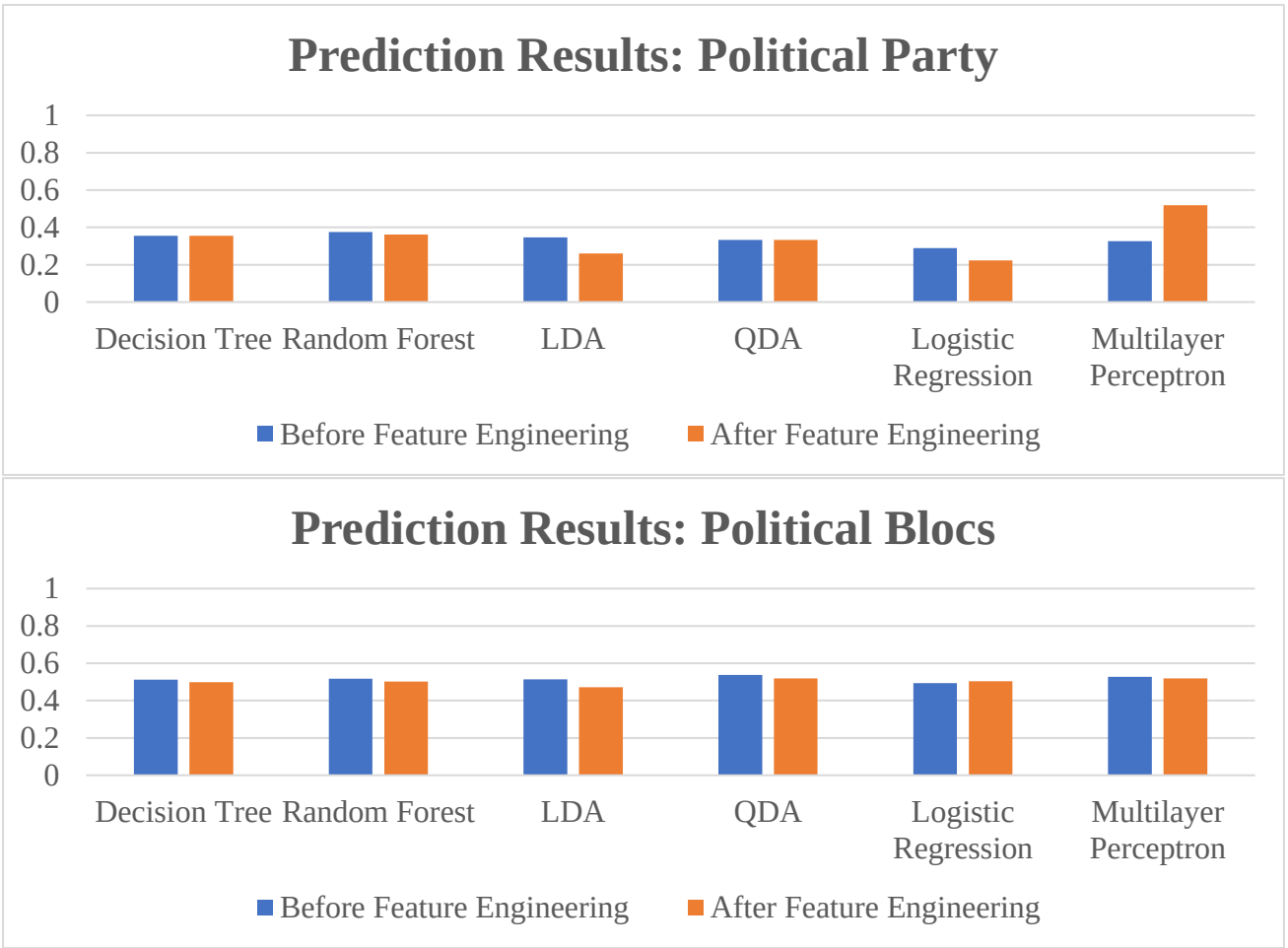

Figure 2. Upper: Accuracy of the models predicting for political parties.
Lower: Accuracy of the models predicting for political blocs.

VI. Conclusion

This project is only a beginning of an attempt to apply artificial intelligence in societal context. Although this project did not provide a satisfactory prediction to the Legislative Council result, it allows further investigations and improvements for this project, which will continue even after the submission of this project.

References

[1] Census and Statistics Department. "2011 Population Census Statistics (By District Council Constituency Area)." *Hong Kong GeoData Store*. https://geodata.gov.hk/gs/view-dataset?uuid=7ae5c7ff-5418-4f39-afbf-341268f826dc&sidx=0 (retrieved 13 March 2022).

[2] Census and Statistics Department. "2016 Population Census Statistics (By District Council Constituency Area)." https://geodata.gov.hk/gs/view-dataset?uuid=e7c0c11c-207d-4acc-aa29-1faa6559b416&sidx=0 (retrieved 13 March 2022).

[3] Electoral Affairs Commission. "2012 LegCo Election – Introduction to Candidates." *2012 Legislative Council Election*. https://www.elections.gov.hk/legco2012/eng/introd.html (retrieved 13 March 2022).

[4] Electoral Affairs Commission. "2016 Legislative Council Election – Introduction to Candidates." *2016 Legislative Council Election*. https://www.elections.gov.hk/legco2016/eng/intro_to_can.html (retrieved 13 March 2022).

[5] M. Li, E. Perrier, and C. Xu. "Deep Hierarchical Graph Convolution for Election Prediction for Geospatial Census Data," presented at the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, Jan. 27 – Feb. 1, 2019.

[6] P. Gilbert. "Analysing UK general election results using machine learning." *Towards Data Science*. https://towardsdatascience.com/analyzing-uk-general-election-results-using-machine-learning-137d130634b1 (retrieved 13 March 2022).

[7] T. D. Jayawickrama. "Feature engineering for election result prediction (in Python)." *Towards Data Science*. https://towardsdatascience.com/feature-engineering-for-election-result-prediction-python-943589d89414 (retrieved 13 March 2022)

[8] SBS and Underscore. "Poliscore." http:// poliscore.kr (retrieved 9 March 2022).

[9] SBS News, Seoul, Republic of Korea. AI 유확당…꼼꼼이 · 화끈이의 분석은? / SBS 선거방송 / 2022 국민의 선택 (Mar. 9, 2022) (In Korean) Accessed: Mar. 9, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=NeOU5Kt-dWo