“Being a good Party official”: Forecasting government officials’ promotion in China

Name of Instructor: Dr. Chow Yan Chi Vinci

ZHANG, Shuoxun (1155211160)

LIU, Chingnga (1155212828)

Abstract:

Machine learning has been used for election result prediction. This research focused on Chinese official in *Zhengting* and *Fubu* level. The result finds that the number of positions, year of birth, department and places working serve as important indicators for promotion.

1. Introduction

Political selection and promotions have always been widely discussed and followed, since it is closely related to the development and survival of regimes (Lee, 2023). In democracies, political elections have been systematically studied, owing in part to a higher degree of political transparency, which enables political scientists to utilize abundant public surveys and public historical election turnouts (Hummel & Rothschild, 2014). Comparatively, authoritarian regimes exhibit lower level of political transparency, leading to scarce research on relevant topics. In recent years, political scientists are inclined to employ machine learning methods to study political patterns empirically. It has been discovered that machine learning methods could overcome problems caused by limited data, through incorporating multiple features (Lee, 2023; Vabalas et al., 2019). For example, Lee (2023) employed machine learning methods to predict the probability of promotion for members of Central Political Bureau. He investigated that ensembled model’s predictive accuracy was 20% higher than that of traditional econometrics models. However, similar predicting officials’ promotion with the background of China remains few, while most using qualitative or econometrics

methods. This study aims to utilize machine learning methods to predict the promotion of Chinese officials, thereby partially filling the research gap.

2. Background

The system for selecting Chinese officials is extensive and complex. In terms of selection methods, there are multiple ways including appointment, evaluation, election, and examination. For high-ranking officials, appointment and evaluation are more common. The hierarchy of Chinese officials is closely linked to the administrative division. Apart from central government and central party, each tier of administrative division has its corresponding officials' rank. According to Civil Servant Law of the People's Republic of China (2018), for instance, provincial-level leaders usually hold the rank of *Shengbu* level. It is worth noting that prefecture-level cities occupy an intermediate position within the administrative hierarchy. That is to say, the corresponding leader (*Zhengting* officials) serves as a vital conduit for transmitting information and resources to both higher and lower levels. Therefore, this research selected *Zhengting* officials as our primary research subjects, and aims to examine them from the following two perspectives: Predicting the pace to which officials are promoted to *Zhengting* level and discussing how one can be promoted further after attaining the rank of *Zhengting* level.

3. Data

The data for this study is mainly from Chinese Political Elite Database (CPED) which is constructed by Jiang (2018). It is a biographical database of all officials of prefecture-level city and above levels, which means most of them basically reached *Zhengting* level. The officials it records are from 2000-2015, centering on the Hu Jintao era. It includes extensive demographic and career information of officials.

4. Promotion speed

To minimise bias in this paper's promotion rate projections, we first applied age restrictions to ensure no officials received further promotions after 2015 (beyond the data's time scope). According to Kou and Tsai (2014), we only select officials whose age was larger than 57 in 2015. Additionally, to improve the quality of data, we select officials who have more than three records of career information in the database.

Regarding the independent variable, *promotion speed*, it is defined as the years an officials used from the year he or she got the first job to the year he or she first reached the *Zhengting* level. The distribution is as Figure 1 shows. The mean of promotion speed is 26.33 years.

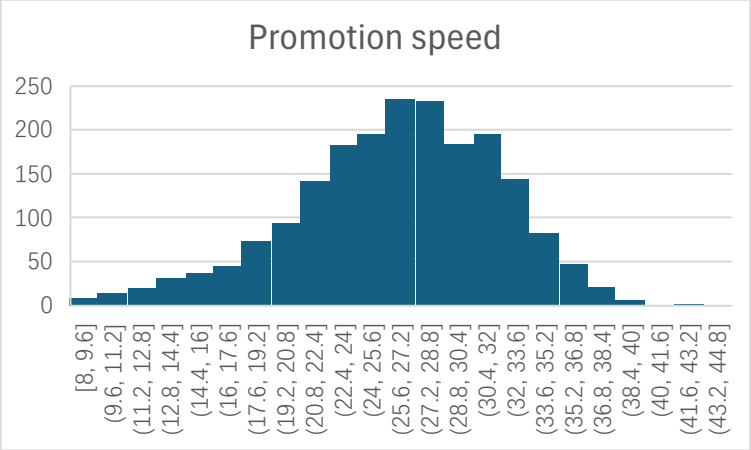


Figure 1. Distribution of Promotion speed

Regarding the selection of independent variables, as demonstrated in the appendix 1, we generated 17 variables from the CPED, covering six categories including personal characteristic, educational background, career development, work experience, network background and geographic mobility.

To predict promotion speed, we employ both single models and ensemble techniques. The model's performance is shown in Figure 2. The ensemble model demonstrated the most effective performance in the task of predicting the promotion speed of officials. Among these, the voting ensemble which gives same weight to the single models and weighted model which gives the weight according to scores of single models, proved particularly effective with Cross-validation R-squared reaching about 0.614 and 0.615 respectively. Following closely behind is Random Forest and Gradient boosting.

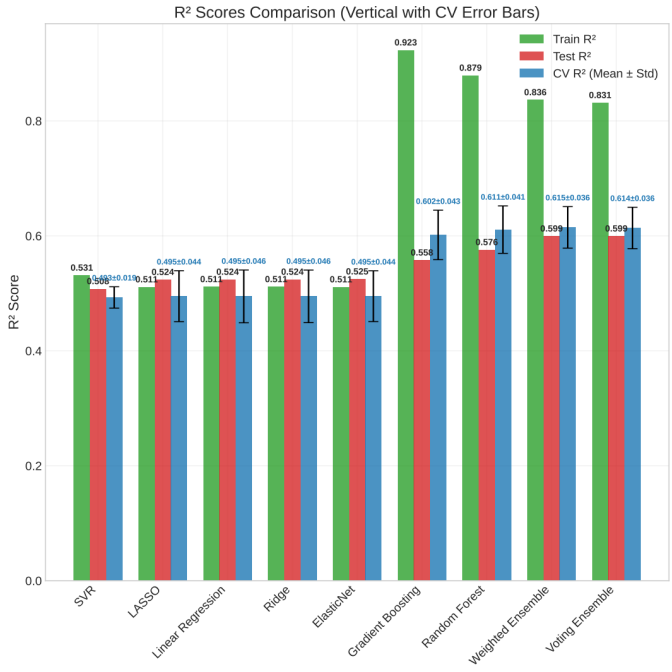


Figure 2. Model performance

This paper also examined the importance of independent variables. Features like total_positions_count, youth_league_experiece, party_standing_years, and education_level, contribute most to explaining the pace of promotion.

5. Further promotion

In most cases, the authority to appoint and remove *Zhengting* level officials is controlled by provincial party committee. However, the authority to appoint and remove *Futing* officials, which ranks one level higher than *Zhengting,* is largely administered by the Central organization department, which is commonly called “Zhongguanganbu”. Thus, this leap in rank represents the transition from a local official to a central official. It is important to investigate what may determine whether *Zhengting* officials can be promoted to *Fubu* level

5.1 Data

Officials born before 1950 are selected for predicting Fubu promotion since people are unlikely to be promoted to Fubu after age 65 (Kou and Tsai.,2014). After data cleaning, the data set includes 1150 observations ,695 promoted, and 455 not promoted. 81 variables included in the dataset, missing value imputed using median.

5.2 Variable selection

To find out possible variables for promotion prediction. Correlation analysis for every variable with Y/N (promoted or not) is conducted. 16 variables are shown to be highly correlated. Within the variables, some of them are also highly correlated. For example, the correlation between Pleaxp and Pleaxp 10 is higher than 0.9. This may be a problem for further analysis as multicollinearity may not affect accuracy of classification model but may confound judgment for allocation of variable importance within the model (Lieberman & Morris, 2014) and classification models may have different ability dealing with multicollinearity (Araveeporn &Wanitjirattikal, 2024). Since we are also interested in feature importance, a VIF and correlation analysis is conducted for the 16 variables with reference to previous literature (Sundus et al., 2022). Final variables are selected according to domain knowledge, correlation between variables and Y/N, VIF, correlation between variables.11 variables are selected for final model. The result is shown in Appendix 3 with the first 11 as selected variables.

5.3 Model building

Models selected are K-Nearest Neighbors, Logistic Regression, Decision Tree, Voting Classifier. Before training the model, we standardized the data and selected the best hyperparameter by 10 K-fold cross-validation. The best hyperparameter is then used to train the model, 80% of data is used to train the models while 20% is used to test the performance. A logistic model includes all 16 variables specially built to compare performance to the logistic model that includes only 11 variables. For K-Nearest Neighbors, Decision Tree and Voting Classifier, only model with 11 variables is built. Voting Classifier gives the same weight for all models with 11 variables.

5.4 Result

Figure 3 - results

To evaluate the performance of models, Test accuracy, Precision(Y), Recall (Y), F1-Score (Y), Precision(X), Recall (X), F1-Score (X) are used, formula shown in Appendix 4, result shown in Figure 3. Comparing performance of model build by 11 variables. The models that perform best are logistic regression, which get the highest score in Test Accuracy, Precision (Y), F1-Score (Y), Recall (N), F1-Score (N), and Voting regression, which get the highest score in Recall (Y) and Precision (N). When the logistic model with 16 variables is included, it outperforms Test Accuracy, precision(N), recall(Y), and f1-score(Y). Regarding feature importance, the results are shown in Figure 4 to 7, noflt2, Born, cenexp are features identified as important. Moreover, the logistic model with 16 variables shows that the coefficient is negative for noflt2 but positive for noflt . Suggesting a possible non-linear relationship.

6.Conclusion and discussion

While there are a lot of previous research using machine learning to predict election result (de Slegte et al., 2025), little has been done on China. Our research may provide insight on what feature is important for being a Chinese official. Variables such as Noflt2, Total_position, cenexp indicating the potential importance of the number of position and the department worked before. Interestingly, Variables such as Prov_Dicty also suggested the importance of the working place. Future research may

focus on Possible promotion patten based on time period differences since variables based on time period is not available now, possible reason for higher promotion rate with working experience in certain areas since people are likely to work in the same district before and after promotion. It may also be interesting to further investigate whether the significance of the born variable is due to time period differences or age.

References

Araveeporn, A., & Wanitjirattikal, P. (2024). Comparison of Machine Learning Methods for Binary Classification of Multicollinearity Data. Proceedings of the 2024 7th International Conference on Mathematics and Statistics, 44-49. https://doi.org/10.1145/3686592.3686600

Chien-wen Kou, & Wen-Hsuan Tsai. (2014). "Sprinting with Small Steps" Towards Promotion: Solutions for the Age Dilemma in the CCP Cadre Appointment System. The China Journal, 71, 153–171. https://doi.org/10.1086/674558

de Slegte, J., Van Droogenbroeck, F., Spruyt, B., Verboven, S., & Ginis, V. (2025). The Use of Machine Learning Methods in Political Science: An In-Depth Literature Review. Political Studies Review, 23(3), 764–784. https://doi.org/10.1177/14789299241265084

Hummel, P., & Rothschild, D. (2014). Fundamental models for forecasting elections at the state level. Electoral Studies, 35, 123–139. https://doi.org/10.1016/j.electstud.2014.05.002

Jiang, J. (2018). Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China. American Journal of Political Science, 62(4), 982–999. https://doi.org/10.1111/ajps.12394

Kou, C., & Tsai, W.-H. (2014). "Sprinting with Small Steps" Towards Promotion: Solutions for the Age Dilemma in the CCP Cadre Appointment System. *The China Journal (Canberra, A.C.T.)*, *71*(1), 153–171. https://doi.org/10.1086/674558

Lee, J. (2023). Machine-learning applications to authoritarian selections: The case of China. Research & Politics, 10(4), 20531680231211640. https://doi.org/10.1177/20531680231211640

Lieberman, M. G., & Morris, J. D. (2014). The precise effect of multicollinearity on classification prediction. General Linear Model Journal, 40(1), 5-10. https://ojs.lib.ua.edu/glmj/article/view/275

Sundus, K. I., Hammo, B. H., Al-Zoubi, M. B., & Al-Omari, A. (2022). Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. Informatics in Medicine Unlocked, 33, Article 101088. https://doi.org/10.1016/j.imu.2022.101088

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PloS One, 14(11), e0224365.

Statement of Use of AI

Throughout the project's advancement, we employed AI tools for auxiliary support, including Qwen Code and Grok4. We utilized these to generate required variables from raw data, facilitate programming tasks, and assist in verifying accuracy.

Appendices

Appendix 1. Independent variable for predicting promotion speed

No.	Variable Name	Description
1	gender	Male = 0; Female = 1
2	Ethnicity_han	Han = 1; others = 2
3	Party_standing_years	Years from becoming CCP member to the year first time reaching *Zhengting* level
4	Education_level	High school and lower = 1 Bachelor =2 Master & Phd =3
5	Total_positions_count	The number of positions
6	Avg_position_duration	The average duration of all positions
7	Slow_promotion_count	The number of slow promotions (>5 years)
8	Organization_diversity	The number of different types of organizations
9	Legislative_experience_years	Years served in legislative
10	Enterprise_experience_years	Years served in enterpriese
11	Education_experience_years	Years served in education
12	Military_experience_years	Years served in military
13	Youth_league_experience	Having experience in Youth league = 1
14	Secreteary_experience	Having experience being secretary = 1
15	Grassroots_experience	Having experience being grassroots officials = 1
16	Total_cities_worked	The number of cities one have worked in
17	Corss_province_movers	The number of times one worked cross provinces

Appendix 2. Feature importance in predicting promotion speed

Model	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Linear	youth_league_experience	education_level	grassroots_experience	organization_diversity	slow_promotion_count
LASSO	total_positions_count	party_standing_years	youth_league_experience	education_level	grassroots_experience
Ridge	total_positions_count	party_standing_years	youth_league_experience	education_level	grassroots_experience
ElasticNet	total_positions_count	party_standing_years	youth_league_experience	education_level	grassroots_experience
RandomForest	party_standing_years	avg_position_duration	total_positions_count	education_experience_years	enterprise_experience_years
GradientBoosting	party_standing_years	total_positions_count	avg_position_duration	education_experience_years	youth_league_experience

Appendix 3. Variables

Name	Meaning
1. Edu	1=High School or below; 2=College or tech school; 3=post-graduate
2. prov_auto	Whether had ever worked in Autonomous regions
3. Born	Year of birth
4. Gender	0=Male,1=Female
5. prov_dcity	Whether had ever worked in Directly Administered Municipalities
6. prov_coast	Whether had ever worked in the coastal regions
7. plaexp	Whether had ever worked in the People’ s Liberation Army
8. Same_location	Same location as birth province in 正厅
9. noflt2	Square of the number of local positions held before
10. cenexp	Whether had ever worked in the central departments

11.	cparexp	Whether had ever worked in central party organizations
12.	noflt	The number of local positions held before
13.	plaexp10	Whether had worked in the People’ s Liberation Army in the first 10 years
14.	cylexp	Whether had ever worked in the Communist Youth League
15.	prov_west	Whether had ever worked in the western regions
16.	cenexp10	Whether had worked in the central departments in the first 10 years

Appendix 4. Formula

- Precision(Y) = True Positives(Y) / [True Positives(Y) + False Positives(Y)]
- Recall (Y) = True Positives(Y) / [True Positives(Y) + False Negatives(Y)]
- F1-Score (Y) = 2 × [Precision(Y) × Recall(Y)] / [Precision(Y) + Recall(Y)]
- Precision(N) = True Positives(N) / [True Positives(N) + False Positives(N)]
- Recall (N) = True Positives(N) / [True Positives(N) + False Negatives(N)]
- F1-Score (N)= 2 × [Precision(N) × Recall(N)] / [Precision(N) + Recall(N)]
- Test Accuracy = [True Positives(Y) + True Positives(N)] / [Total Test Samples]

Figure 4. Feature Importance-Decision Tree

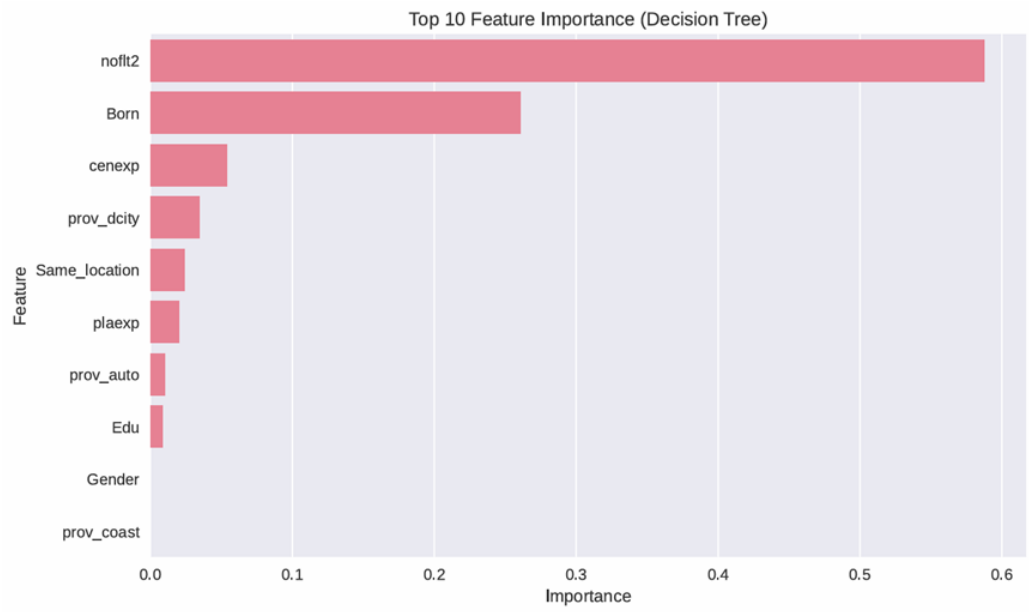


Figure 5. Feature Importance-KNN

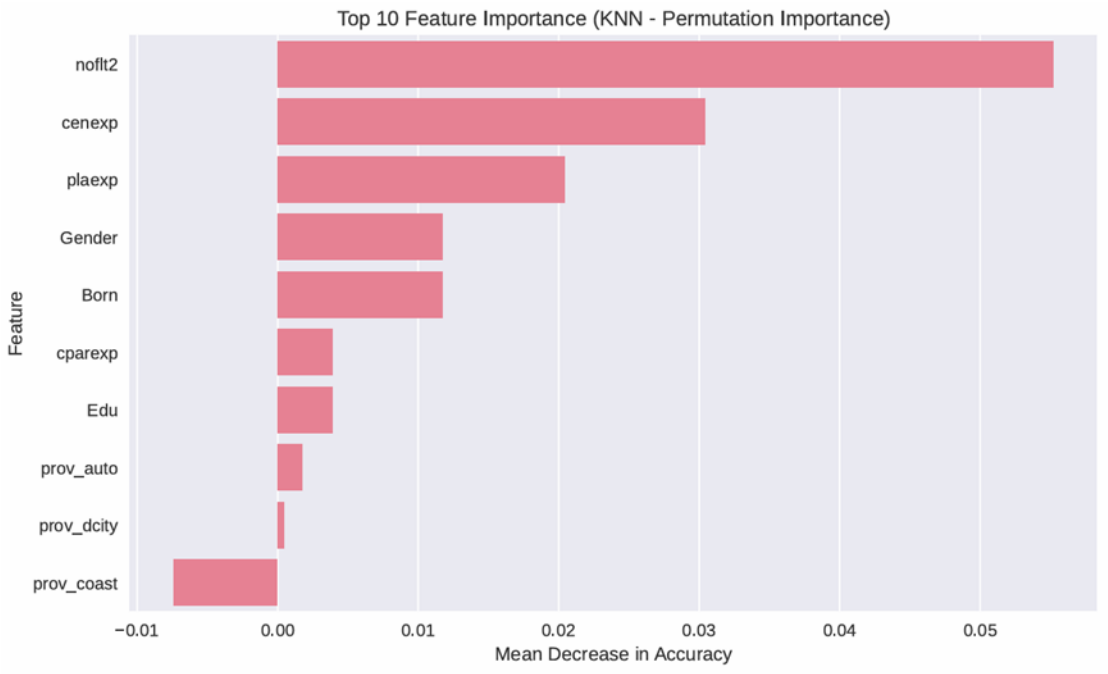


Figure 6. Feature Importance-Logistic-11 variables

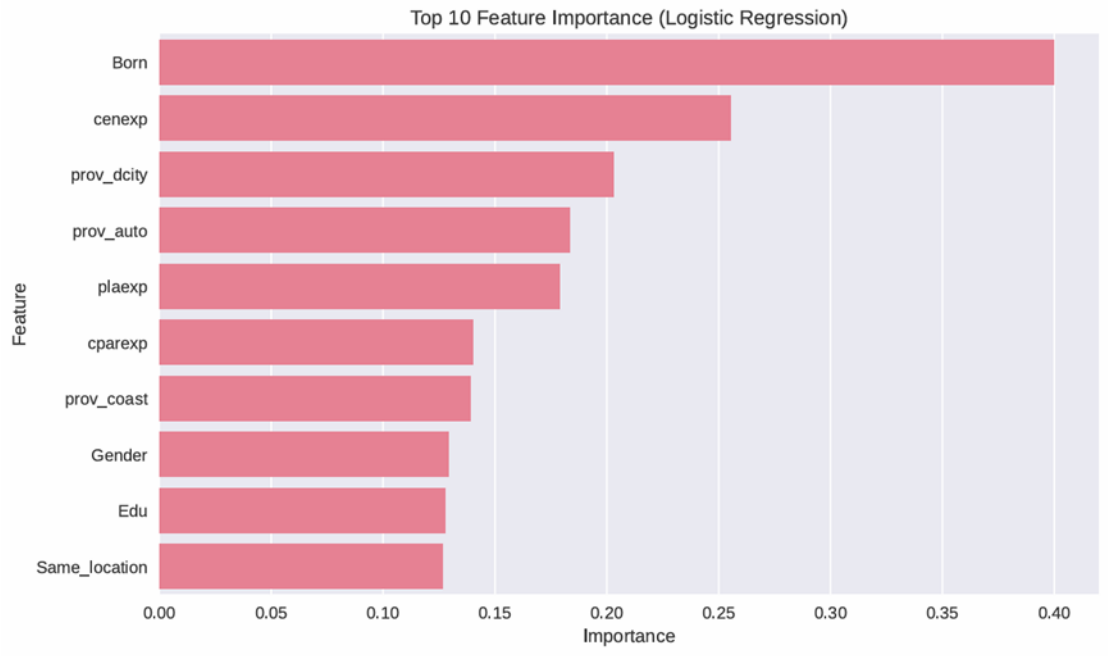
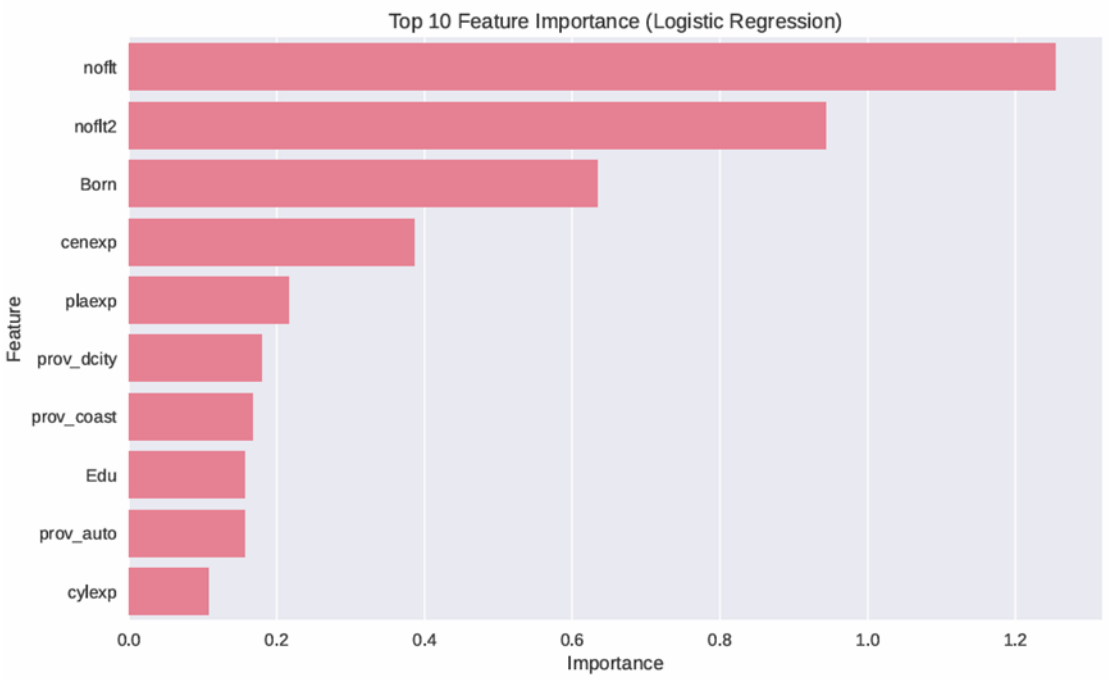


Figure 7. Feature Importance-Logistic-16 variables


Top 10 Feature Importance (Logistic Regression)
nofit
nofit2
Born
cenexp
plaexp
prov_dcity
prov_coast
Edu
prov_auto
cylexp
Feature
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Importance