Unraveling China's "high level of corruption, high economic growth" puzzle: linking corrupt officials and their monetary gains

Group Koi

Group Members	Work Division in the Assignment
YUAN Minjun Erin (PG)	○ Introduction ○ Literature Review ○ Overall Revision
CHING Cho Yu Gary (UG)	○ Methodology ○ Prediction Models
WONG Kan Hei Angel (UG)	○ Methodology ○ Discussion
KWAN Ho Wun Cassie (UG)	○ Background ○ Conclusion

1. Introduction

As an "economic and political evil," corruption could hamper economic development and undermine political trust in one society (Seligson, 2002). According to the Corruption Perception Index (CPI) rankings published by Transparency International, the vast majority of nations are still plagued by corruption today. China's performance regarding CPI rankings is not yet satisfactory, regardless of its improving anti-corruption efforts in recent years. However, unlike countries that suffered from low economic growth because of widespread corruption, the coexistence of severe corruption and booming economic development in China has been puzzling. On the one hand, corruption has led to enormous economic loss and eroded the legitimacy of the Chinese Communist Party (CCP) (Zhu & Wu, 2014). On the other hand, it seems that corruption does little

harm to China's rapid GDP growth momentum. Corruptive officers play a leading role in the corruption chain. The concept of corruption is complex, multifaceted, ambiguous, and specific to different national contexts (Ko & Weng, 2011). Although there lacks a uniform definition of corruption in academia (Song & Chen, 2012), corruption is generally characterized as public power abuse for private gain (Philp, 2006). These private gains could accumulate to an enormous amount, causing significant economic loss to the national economy. Hu and Guo (2001) estimated that the direct economic loss caused by corruption accounts for 13.2 % to 16.8 % of annual GDP since the second half of the 1990s. Even worse, the phenomenon of corrupt officials fleeing abroad began in the 1980s and became more common in the 1990s. Since fleeing officials often transfer vast amounts of money abroad, this aggravates economic losses caused by corruption and makes anti-corruption work more challenging.

What are the characteristics of corrupt officials who are involved in the abuse of public office for monetary gains? The complexity of corruptive behaviors and the long-standing guiding philosophy of anti-corruption governance have led the CCP to choose a relatively broad and vague definition of corruption in practice. That is to say, the types of crimes for corruption differ in different situations and are not necessarily in monetary forms. This article mainly focuses on corruptive behaviors that exchange monetary bribes for political favors. Based on the big data method, the article intends to shed light on the puzzle of "high level of corruption, high economic growth" in China by analyzing the characteristics of corruptive officers punished since Xi's anti-corruption campaign in 2012.

The rest of the paper is structured as follows. The next section briefly introduces the background of China's anti-corruption campaign since 2012. This is followed by a literature review regarding the impacts of anti-corruption on China's economic growth and traditional methods

employed in this research field. The following section utilizes a Logistic Regression model with Cross Validation to explore the characteristics of corrupt officers who were punished because of monetary corruption from 2012 to 2018. The paper then elaborates on the key findings generated from the classification models and compares differences among them. It concludes by highlighting the theoretical implications of this research and suggesting pathways for future research.

2. Background of China's Anti-corruption Campaign since 2012

After the 18th National Congress of the Chinese Communist Party in 2012, Xi Jinping pledged to crack down on both "tigers" and "flies," namely senior officials and local civil servants in corruptive cases. In Xi's view, the existence of corruption will undoubtedly cause the collapse of the Party and the state (Yuen, 2014). The CPC Central Committee is committed to the strict governance of the party. The majority of the investigated officials were suspended from office and faced allegations of bribery and authority abuse. In order to inhibit corruption, Xi vowed to avoid extravagance and to minimize bureaucratic visits and meetings. The "Eight-point Regulation"(Zhong yang ba xiang gui ding 中央八項規定) lays out the detailed criteria for how government officials can improve their style of work in eight aspects. The ruling party has also made great efforts to rectify unacceptable working styles of formalism, bureaucratism, hedonism, and extravagance (Si feng 四風) since it was introduced in December 2012. It intends to reverse the injustice and establish institutional protection by tightening internal regulations.

According to the statistics of the Central Commission for Discipline Inspection (CCDI), during the five years from the 18th to the 19th National Congress of the Communist Party of China, a total of 440 Party members and officials at or above the province- or military-level (省軍級) and other centrally-administered officials were investigated. More than 8,900 cadres at the

department- and bureau- level (廳局級), and 63,000 officials at the county and department level (縣處級) were punished (Deng, 2018). Since the 19th National Congress of the Communist Party of China, more than 70 centrally administered officials have been under investigation. From January to September 2018, the national discipline inspection and supervision organs filed 464,000 cases and sanctioned 406,000 people. Take the centrally administered officials as an example. During 2012-2017, the average number of people investigated and punished each year was 88. However, the number of people investigated and punished has dropped by about 20% in 2018 (Deng, 2018).

To track down corrupt officials who fled overseas, China's Central Anti-Corruption Coordination Group deployed the "SkyNet" operation (天網) in April 2015. Different ministries are responsible for different dedicated projects, and the special project of the Ministry of Public Security to hunt down the people involved is called the "Fox Hunt" operation (獵狐行動). According to the website of the State Supervision Commission of the CCDI, the "SkyNet 2018" operation arrested a total of 1,335 people who had fled across the country. These include 307 party members and state officials, five "the country's list of the 100 most-wanted fugitives" (百名紅通人員) and the amount of ill-gotten money recovered was 3.541 billion renminbi. As of 2019, "Fox Hunt" has captured more than 4,900 fugitive officials from more than 120 countries and regions and recovered more than 17 billion renminbi in ill-gotten money, which strongly supports the central anti-corruption campaign. It has made positive contributions to the party's overall strict administration and the maintenance of social stability.

3. Literature Review

The economic consequences of corruption have received extensive attention from academia. Both political scientists and economists have comprehensive discussions about how corruption

affects private businesses in China. They analyze the economic factors in corruption activities and reveal their behavior patterns in order to enlighten the effective anti-corruption methods and measures. Traditionally, data sources include in-depth interviews, government publications, news from media reports, financial market databases, statistical data from government websites, and surveys. For example, Zhu and Wu (2014) utilized the results of two survey questions from the All-China Federation of Industry and Commerce (ACFIC) between 1997 and 2006 to identify how corruption has spread across industries. Although the survey samples are relatively small, this study reveals the most corrupt sectors, such as the real estate (RE) sector. It is noteworthy that the data sources employed in the field could be mixed instead of simplex. In the study of exploring the relationship between corruption and private business, Zhu and Zhang (2017) used three data sources, including field interviews to explore possible channels, 2012 World Bank Enterprise Survey data of 2700 firms in 25 prefectures in China, and self-compiled data set which contains 25 municipal party committees from 2002 to 2011. Based on the three-dimensional data sources, Zhu and Zhang (2017) argue that corruption predictability is a crucial factor determining corruption's effects on the private sector and is remarkably affected by government leadership stability.

Studying corruption in the Chinese RE sector is a critical branch in the anti-corruption branch. The RE sector is one of the most vulnerable sectors in terms of corrupt behaviors. Scholars such as Zhu (2012) mainly use a qualitative method to study the Chinese RE industry's corruptive chain. Other scholars such as Lu (2017) and Chen and Kung (2019) mainly use the quantitative method by gathering a massive amount of data and using an explicit causal mechanism to explore the effects of China's anti-corruption campaign on the land market. Lu (2017) used the information collected from China's land website to analyze residential land sales during the anti-corruption

campaign. Chen and Kung (2019) have done an anti-corruption investigation using the financial market and stock market data. Most of the data used are collected from the WIND database and China Stock Market & Accounting Research (CSMAR). They could apply the big data method to identify princeling firms as land transaction data could be traced back easily through online platforms and databases such as "A Million Transactions." Studies in other economic sectors, such as estimating loss in the financial market because of corruption (Kim et al., 2018), corrupt collusion between companies and government officials (Hao et al., 2020), and the effect of public governance on firms' incentives to commit fraud (Zhang, 2016) grasp data from official websites and financial databases.

Facing considerable amounts of online data generated in the process, relying on the traditional methodologies may be troublesome and time-consuming. The big data method has its edge in collecting a considerable amount of data in this kind of research; thus, it could be utilized as a complementary method for traditional methodologies. Admittedly, the big data methodology alone is inadequate to conclude sound causal inferences, and there are several limitations for it (Jiang et al. 2019). However, having as much data as possible can help social scientists infer causal relations using massive datasets (Grimmer 2015). Existing research in this field rarely utilizes the big data approach. Therefore, this article intends to have a preliminary attempt to fill this gap and shed new lights on characteristics of corrupt officers related with monetary gains in China's anti-corruption campaign.

4. Methodology

4.1 Preliminary Design

This research project aims to provide references for future quantitative studies on China's Anti-corruption campaign. We intend to explore critical and common features shared by targeted

officials in the previous anti-corruption campaigns based on the official data from the CCDI. We also aim to reveal the pattern of investigation of China's anti-corruption agency at the same time. In other words, we would like to classify the specific characteristics and backgrounds of targeted officials and how these traits affect their certain corruptive behaviors. We narrowed down our research scope to corruptive behaviors related with monetary gains through abuse of power. That is to say, these corruptive officials were targeted by the CCDI mainly because of illegal monetary gains when they were in position. Therefore, our project focused less on other corruptive behaviors. We mark other types of corruption as "No" to attain a binary categorization. What is the probability of Chinese officials getting attention from the CCDI? How does the probability change along with every additional similarity that targeted corruptive officers in the future shared with their predecessors? With all the proposed predictors, this section intends to explore the above-mentioned questions by using predictors to predict the possibility of a current official being targeted by the CCDI.

4.2 Data sources

We compile our primary dataset from a database of the Visualizing China's Anti-Corruption Campaign of the China File, an open database available on the internet. This database is a digital magazine issued by the Center on U.S.-China Relations at Asia Society, aiming to provide a detailed, integrated, and inspiring platform about China for the West. The first attempt of data collection was in January 2016, while the second attempt was conducted by July 31, 2018. In the second attempt, anti-corruption cases shed light on the continuous anti-corruption campaign led by Secretary of the CCDI Wang Qishan and President Xi Jinping. The China File has provided raw data online for readers to explore the relevant issues further. We utilize the raw data as the primary dataset while combining several other datasheets to enrich our dataset's

comprehensiveness. These complementary datasets include Prefectural Party Secretary (Chen, 2015), Mayor of P. R. China, 2000-2010 from Fudan WTF SOSC Lab, and Wang(2020) China's Corruption Investigations Dataset. More predictors have been added, and some of the missing entries in the original dataset have been filled by comparison among different datasets. After combining and cleaning data, our current dataset contains 2500 observations with 14 predictors.

4.2.1 List of Predictors

Predictors (Number of factors)	Description	
Rank (9 Categorical factors)	Highest rank of the official obtained	
Age (6 interval categories)	NA	
Native born province (31 Categorical factors)	To find out possible factions of hometown	
Corruption location (34Categorical factors)	To see which provinces suffered from corruption	
Corrupted amount (Categorical factors)	Venus Sentence detail	
Gender (Binary)	NA	
Sector (12 Categorical factors)	Divided into: 1. real estate 2. agriculture 3. mining 4. infrastructure 5. energy 6. manufacturing	7. transport 8. technology 9. CEHS 10. catering 11. social service 12. finance

Overseas (Binary)	To see whether the officials are targets of global search
Head or Vice (Binary)	Leading position in workplace
Sentence detail (Categorical factors)	Imprisonment, fine, death sentence
Local official (Binary)	To see whether local people will be involved more than counterpart
Status type (Binary)	Label as Tiger (deputy provincial level, ranked 4 or above) and Flies
Connections with tigers (Binary)	To find out possible factions
Monetary gain (Binary)	Predicted value

4.3 Data analytic model

As we have narrowed the scope down into a binary classification problem, we proposed to use logistic regression to build our classification predictive model. Logistic regression is a classic statistical model often used to predict categorical outcomes. It is suitable for conducting a regression analysis when the predicted variable is dichotomous, i.e., Yes or No outcomes are favored. Logistic regression is employed in this project to depict our data and reveal the relationship between the predicted binary variables (i.e., whether or not the officials were involved in the abuse of public office for monetary gains) and other 13 independent variables (i.e., the rank of the officials, their corruption location, sectors they were in charge in, etc.). We use these binary, interval, categorical predictors, and corresponding data types to make predictions. Some techniques from random forest models are used to cope with data-cleaning problems. Even though

logistic regression modeling is regarded as a basic and simple model, it is the best tool to be used to carry out data regression analysis in our project. Furthermore, to develop an optimal model and achieve better results, we decided to test and train our logistic regression model using 10- Fold cross-validation.

4.4 Problems encountered in the data-cleaning stage

One of the significant factors producing variation and inconsistency in our project are the methods we employed to clean our dataset. Therefore, it is better to state what methods we have used in the report, though they may not be the best or suitable measure to solve the problem and demonstrate the problem-solving approach. Our lesson learned from this computational social science project is also noteworthy for future studies.

Though we have integrated several datasets to build up our dataset, we noticed that since the overlapping among different datasets cannot fully cover all the predictors, not all the entries in our dataset have been filled. Our input data should be more comprehensive and cleaner to train the classification predictive model by feeding historical data, as the logistic regression lacks capability to handle missing entries. Thus, the data cleaning process is crucial for our project. We have tried several approaches to fill in missing entries in our datasets, including data cleaning in the pre-model training stage and the R package and command in the training stage.

In the pre-model training stage, we did manual labeling of categorical factors. As one of our predictors needs to find out which sectors corrupt officials were involved in, we had to figure out their categories based on their names. Input methods such as the random and equal distribution of categorical factors in some predictors, inputting means, and medians to minimize the effect of filling in suspicious entries regardless of the exact value. We have also excluded the datum, which contains missing entries using R commands, but concerning the size of our dataset (2500

observations in total), which is already a small size, removing missing entries had led to the vanishing of 1000+ observations. Training and testing results of the logistic regression in this trial are less effective and less desired than inputting missing entries by other measures.

The method we have applied to deal with the problem of missing value is using RThe R command *data.imputed <- rfImpute(NAME:columns of predictors ~ ., data = data, iter=6)* to fill in value. Based on the concept of a decision tree and random forest algorithm, the filled-in values are tested by creating five trees for each round. Therefore, we can preserve some of the information, and the uniqueness of the individual data is less affected. Ideally, using the predicted value to impute possible values for missing values for predictors is the best approach. In our case, the problem we faced is missing values in our predicted value due to the anti-corruption agency and media's insufficient official information. In logistic regression, the predicted value must not contain any missing entry, so we have to judge our predicted value to fix this problem. Thanks to our preliminary work in the data-cleaning stage, we have several predictors without missing values that can be tested. However, the predictors' choice to be spun by other columns of predictors will lead to the result's variation. The critical insight we obtained from the missing value problem is that we have to know the model's limitations. It is a reminder to our study that our model and conclusion may be biased.

4.5 Comparison of trials and proposed results

As mentioned above, we proposed two reasons regarding the measures employed to deal with missing entries producing possible variation of the results and models. The two approaches are the data input approaches used in the data-cleaning stage and choosing different predictors to fill missing values with command rfimpute.

Data used for comparison are extracted from the inbuilt tools in R package, Caret, using command like *summary(mod_fitcv) and caret::confusionMatrix(table((mod_fitcv$pred)$pred, (mod_fitcv$pred)$obs)),* as the models are built using same packages and went through same standardized test using the tools. Therefore, it is fair to conduct comparisons by using the extracted data. P-value is used to identify the statistical significance of a feature. The predictors which contain the largest number of parameters which are statistically significant are regarded as good predictors of the logistic models.

Below are our comparison criteria:

1. Akaike information criterion (model with lowest score are selected)
2. Accuracy
3. Kappa coefficient
4. Sensitivity (Recall or True positive rate)
5. Specificity (True negative rate)

4.5.1 Table of comparison

Description of the model	Logistic regression using all predictor, Status type using Gender	Logistic regression using all predictors, rfimputed using Status type	Logistic regression using all predictors, rfimputed using rank	Logistic regression using all predictors, rfimputed using sector	Logistic regression using all predictors, rfimputed using corruption location	Logistic regression using all predictors, rfimputed using Head or vice

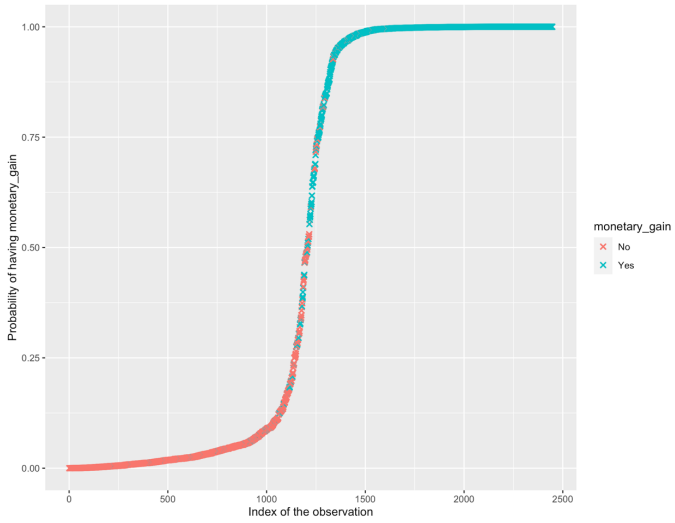
Akaike information criterion (AIC)	1633	842.11	1629.8	1615.7	1418.3	2005.3
Accuracy	0.8664	0.9395	0.8578	0.8548	0.8884	0.8026
Kappa coefficient	0.6284	0.879	0.645	0.6293	0.6566	0.5229
Sensitivity	0.6613	0.9524	0.7065	0.6913	0.6741	0.6243
Specificity	0.9360	0.9276	0.9199	0.9186	0.9487	0.8823

The model in which the predictor status_types was predicted by other columns of predictors results in lowest Akaike information criterion, highest accuracy, the best sensitivity and the best specificity. Therefore, we chose it to be our base model.

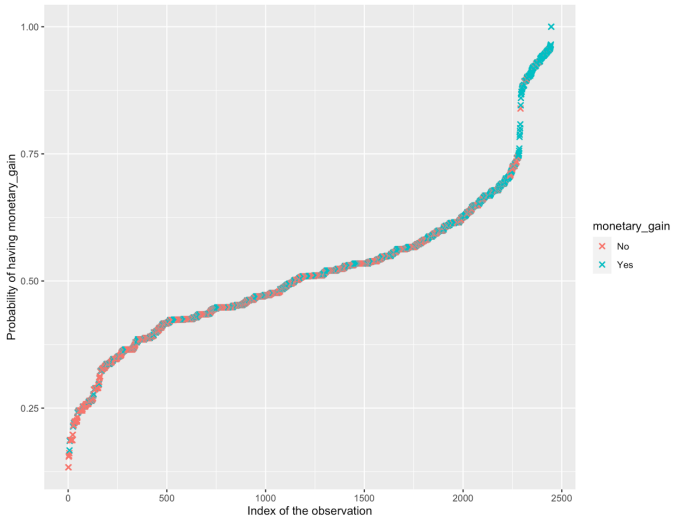
Concerning the important predictors, those contains more parameters which are statistically significant, we found out that these predictors have been overweight, they are

1. native born province of officials
2. corruption location a.k.a provinces that they were in position in the end
3. local officials or not
4. connection with tigers,
5. targets of overseas searching
6. Head or Vice

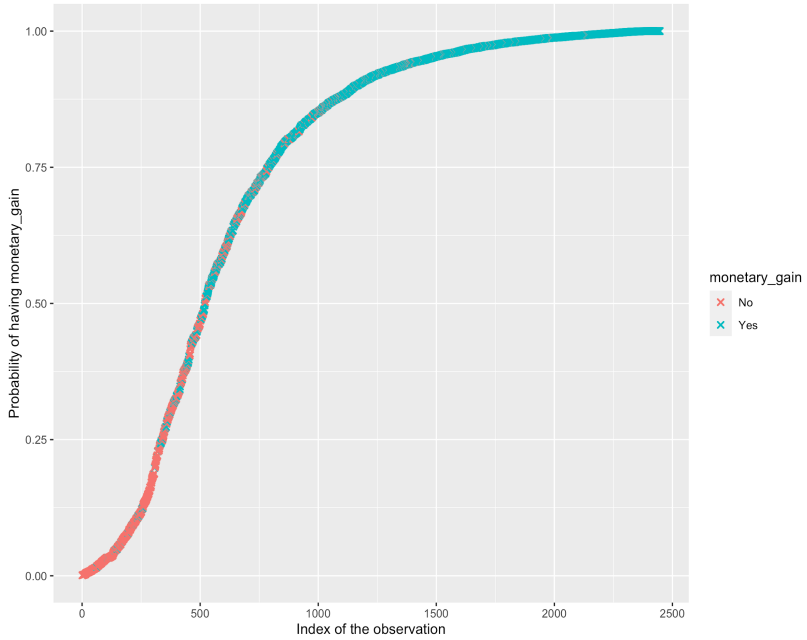
These six predictors are then implemented to train and test a logistic model, forming our proposed simple model in the end. The proposed model, beyond our expectation, did not perform as we thought. A proposed explanation to this problem is the artificial input of missing value has led to missing or even distortion of information about the officials.



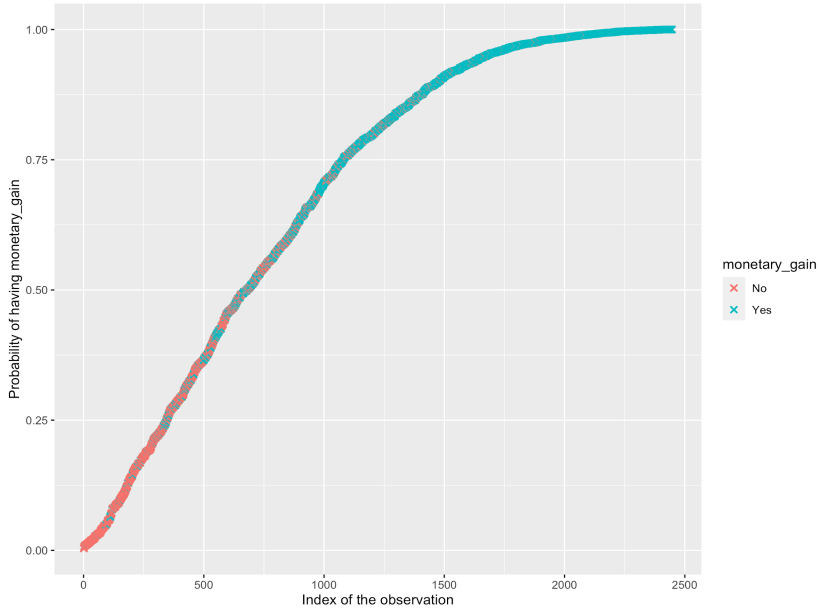
"Index of observation versus Probability of having monetary_gain" graph of logistic model using rfimputed: Status



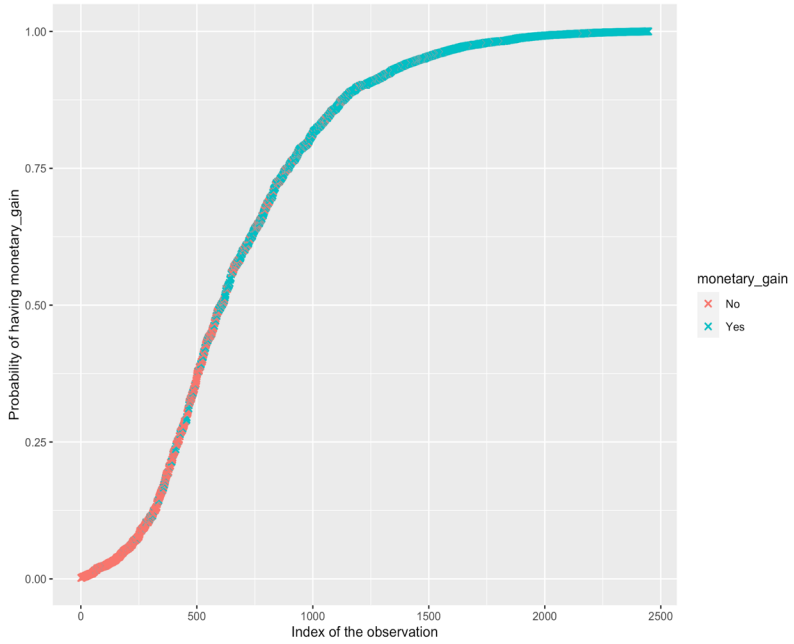
"Index of observation versus Probability of having monetary_gain" graph of our proposed logistic model



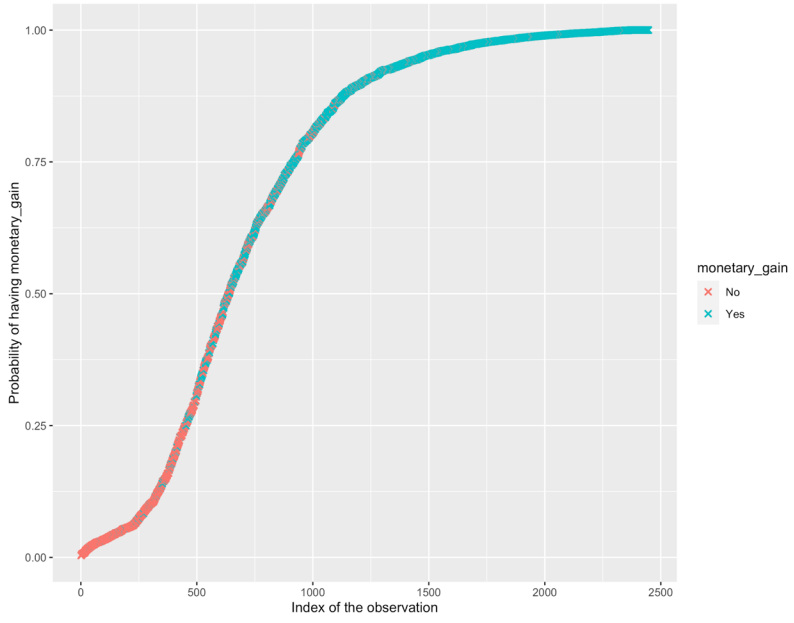
"Index of observation versus Probability of having monetary_gain" Graph of logistic model using rfimputed: corruption location



"Index of observation versus Probability of having monetary_gain" Graph of logistic model using rfimputed: Head or vice



"Index of observation versus Probability of having monetary_gain" Graph of logistic model using rfimputed: Sector



"Index of observation versus Probability of having monetary_gain" Graph of logistic model using rfimputed: Rank

5. Discussion

Although one of our goals is to provide references for China's Anti-corruption campaign's quantitative studies, there are a few important and common features of the targeted officials in previous anti-corruption campaigns that have been shown in our results. For officials that their native provinces are from Sichuan, Ningxia, and Heilongjiang, they are more likely to be targeted in these campaigns with slightly outweighed data. As for differences among different parts of the country, five southern provinces and eleven northern provinces are classified as statistically significant. At the province level, it was found that the top four provinces that suffered from corruption involving monetary gain are also in the northern part, which is Beijing, Hebei, Jilin, and Shandong. For characteristics of the targets, results showed that local officials who grew up and worked in their hometown, namely a native province, are more likely to become targets using the predictor Native-born province. Those who have connections with higher-level officials or tigers, at least sub-provincial levels, also increase the chance. Comparing the officials' leading position by head or vice, targets are usually the head in the workplace. The last common feature is the sentence detail of the targets. 5 to 10 years and 10 to 15 years imprisonment are the most frequently used sentences and punishment by the court. The above are the patterns of investigation of the Anti-corruption agency that have been discovered in our results. We have discussed how the officials' characteristics and specific backgrounds, such as a native province or working position in the workplace, will affect corruption-related behaviors.

There are a few remarks that should be noted in the results. Although some common features have been discovered for targets in anti-corruption campaigns, there are no specific sectors outweighed in the current algorithm. Moreover, the absconded officials are usually being targeted in order to get their corrupted amount retrieved. With the dataset's limitations as the data is not

good enough, including problems of missing data and limitations of variables, we only have six predictors in our current algorithms. Further investigation is needed with a more comprehensive dataset and casual relations. Our proposed hypothesis is that corruption is here and there in China. No matter what sectors, northern provinces are often targeted in the anti-corruption campaign, but not those provinces with more robust economic performance.

6. Conclusion

After conducting social network and statistical analysis of officials in corruption cases, we found that there are several apparent categories, such as the geographic location of work (hometown or other provinces), interpersonal relationship in the workplace, position level, etc., are the collaborative characteristics of corrupt officials who participate in the abuse of public office for monetary gain. In short, big data on anti-corruption is an essential means and an imminent development to boost the efficacy of corruption governance and strengthen governance capabilities. We believe that big data analysis of the characteristics of the main body of corruption and bribery crimes will help improve the accuracy of corruption punishment, the scientific nature of corruption prevention, and the objectivity of corruption research.

Reference:

Chen, T., & Kung, J. K. S. (2019). Busting the "Princelings": The campaign against corruption in China's primary land market. The Quarterly Journal of Economics, 134(1), 185-226.

China Files.(2018). Visualizing China's Anti-Corruption Campaign. Retrieved November 30,2020, from https://www.chinafile.com/infographics/visualizing-chinas-anti-corruption-campaign

Deng, Y. (2018, December 27). Opinion: What signal does Xi send when he claims an "overwhelming victory in the fight against corruption"? BBC News. Retrieved October 30, 2020, from https://www.bbc.com/zhongwen/simp/chinese-news-46691256

Grimmer, J. (2015). We're All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. PS: Political Science & Politics 48(1): 80–83.

Hao, Z., Liu, Y., Zhang, J., & Zhao, X. (2020). Political connection, corporate philanthropy and efficiency: Evidence from China's anti-corruption campaign. Journal of Comparative Economics, 48(3), 688–708.

Hu, A., & Guo, Y. (2001). Comprehensive strategies of corruption control and institutional design during transition [Zhuanxingqi fangzhi fubai de zonghe zhanlue yu zhidu sheji]. Management World, 6, 44–55.

Jiang, W., Yang, T., Sun, G., Li, Y., Tang, Y., Lv, H., & Xiang, W. (2019). The Analysis of China's Integrity Situation Based on Big Data. Journal on Big Data, 1(3), 117.

Kim, D. S., Li, Y., & Tarzia, D. (2018). Value of corruption in China: Evidence from anti-corruption investigation. Economics Letters, 164, 112-116.

Ko, K., & Weng, C. (2011). Critical Review of Conceptual Definitions of Chinese Corruption: A formal-legal perspective. Journal of Contemporary China, 20(70), 359-378.

Lu, X. (2017). Essays in China's Anti-corruption Campaign (Doctoral dissertation, UC Berkeley).

Seligson, M. (2002). The impact of corruption on regime legitimacy: a comparative study of four Latin American countries. The Journal of Politics, 64(2), 408–433.

Shuo Chen. (2015). From Governance to Institutionalization: Political Selection from the Perspective of Central-local Relations in China--Past and Present (1368-2010), Department of Economics, Fudan University Working Paper.

Wang, Yuhua. (2020). China's Corruption Investigations Dataset. Retrieved November 30, 2020, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9QZRAD

Zhang, J. (2016). Public Governance and Corporate Fraud: Evidence from the Recent Anti-corruption Campaign in China. Journal of Business Ethics, 148(2), 375-396.

Zhu, J. (2012). The Shadow of the Skyscrapers: real estate corruption in China. Journal of Contemporary China, 21(74), 243-260.

Zhu, J., & Wu, Y. (2014). Who pays more "tributes" to the government? Sectoral corruption of China's private enterprises. Crime, Law and Social Change, 61(3), 309-333.

Zhu, J., & Zhang, D. (2017). Does corruption hinder private businesses? Leadership stability and predictable corruption in China. Governance, 30(3), 343