STA303 Final Project Report

Guanlin (Tony) Chen   ID. 1007822260

April 7, 2024

1. Introduction

The goal of this study is to use the Generalized Linear Model (GLM) for binomial response variable to predict whether or not a nearby subway route is available to adjacent medium-size residences in Beijing (the proximity of subway route to residences) in terms of residence information including location by relative latitude difference, area of m², house age, community average price, floor of residence, and ownership of five-year-property. With this goal, the study provides insights into the relationship between the subway network and various patterns of housing information for better urban planning of the subway network to satisfy the demand of city residents.

Previous studies inform the selection of interesting variables in my study and provide acknowledged relationships between housing information and the expansion of subway. Additionally, the price and building structure can be potential confounders that both influence the subway route proximity and other predictors. However, they used various models including spatial lag and quantile regression different than GLM, and the response variables are usually the housing price rather than binary categorical variables which can be applied with logistic link function in constructing GLM models (Ahn et al., 2020; Wen et al., 2018; Xu et al., 2018).

2. Methods

2.1 Data cleaning and EDA analysis

The original data was sourced from Kaggle and fetched by the author based on the Chinese housing platform Lianjia.com from 2011 to 2017 (Kaggle, 2017), and the data was collected from different residence houses that are independent of each other. To make the data align with my research question, I chose the subset of traded houses in 2017 that are between 90 and 130 square meters and removed all housing items that contain missing values in their associated variables and obtain 12296 observations. The response variable is a binary outcome (subway proximity) such that Y = 1 if the residence is near a subway route or Y = 0 if it is not near a subway route, satisfying the binomial distribution, and a logit link is appropriate to construct the GLM. The EDA includes numerical summary tables that illustrates the descriptive statistics of both continuous and categorical predictors, the histograms of some continuous predictors, and the boxplots of important numerical predictors related the subway proximity.

2.2 Variable Selection

Firstly, building the preliminary model with the logit where P is the probability that the residence is nearby a subway route, and the 11 candidate predictors from the dataset: latitude difference in degree(1), price per m²(2), area in m²(3), house age in years(4), whether or not elevator is available (5), whether or not the residence is in concrete-steel structure(6), floor(7), community average price which is a representative price level of the affiliated community for each residence(8), the number of followers focusing on the residence(9), days on market, DOM(10), and whether or not the residence is a legitimate five-year property, FYP (11).

The preliminary logistic model:

$$log(\frac{P}{1-P}) = log(\frac{E(Y|x_i)}{1-E(Y|x_i)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + +\beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

The AIC and BIC are used to make automatic stepwise selection of variables for the second and third model. Starting from the AIC with the preliminary full model, AIC at each step (either forward or backward selection) is computed and model with the smallest AIC is chosen to become the starting model for the next step until no smaller AIC value is obtained, then the last model is the preferred AIC model, and replicating the same process with BIC to obtain the preferred BIC model. Besides, apply the LASSO shrinkage method on the cleaned data set to shrink the regression coefficients towards zero by imposing a penalty which increases as the absolute value of the coefficients increases. At the end, remove the predictors whose regression coefficients are reduced to 0, and form the LASSO model. Finally, compare the four constructed models concerning the significance of predictors by testing $H_0\colon \beta_i = 0\ vs.\ H_a\colon \beta_i \neq 0$ with the significance level at 0.05, and compare the VIFs that measure the multicollinearity of numerical predictors.

2.3 Diagnostics and Validation

For each of the four models after variable selection, conduct diagnoses of problematic observations. The influential points on all fitted values of models are identified by the Cook's distance larger than the cutoff, and the influential points on the own fitted value of an observation are identified by larger than the absolute value of DFFITS at the cutoff. Moreover, outliers are the observations with standardized residuals outside the range (-4,4).

Furthermore, cross-validation is applied by splitting the data into 50 parts to fit the models with 49 parts (training set) and predict the outcomes for the remaining part (testing set), and the calibration plot indicates good prediction accuracy if the predicted probability fits the actual probability close to a 45 degree line or else. In addition, the ROC-AUC graph shows the percentage of discrimination between living near subway route and not living near subway route by AUC value.

The model with the most significant predictors, less multicollinearity, better prediction accuracy and discrimination ability is chosen as the final model.

3. Results

3.1 Data Description

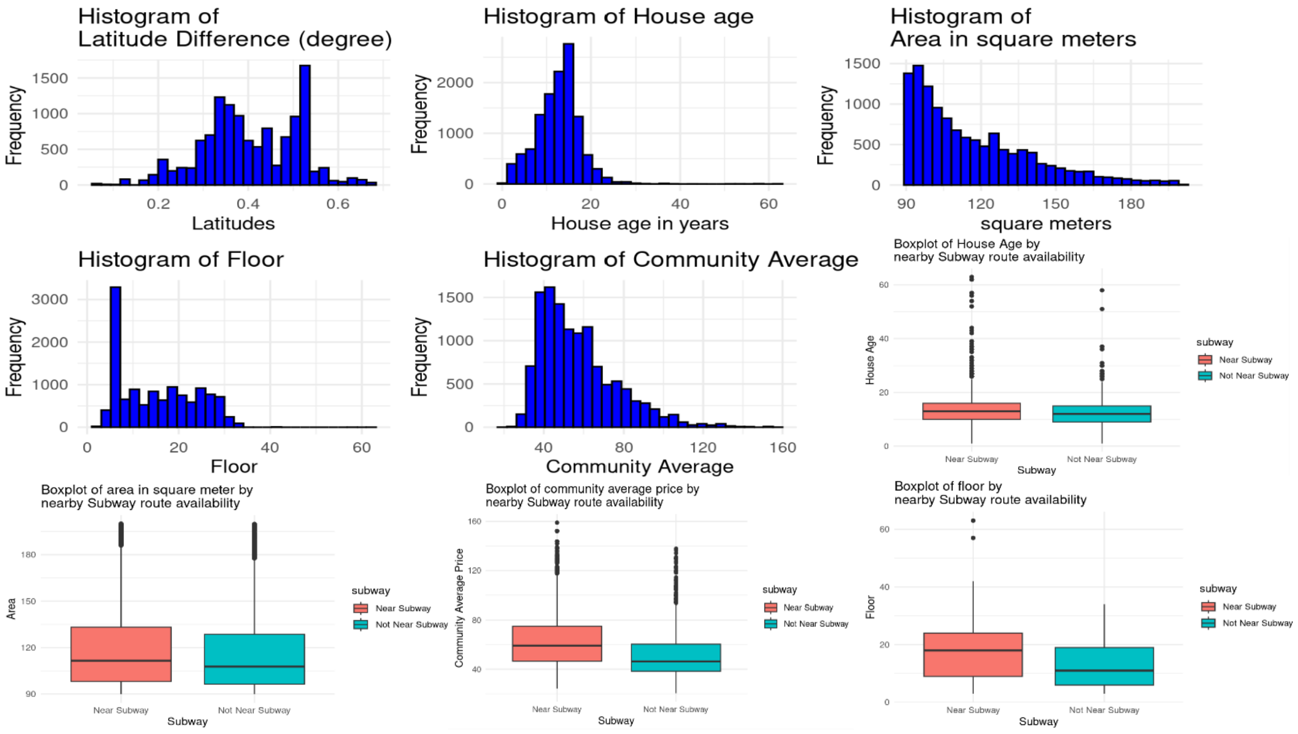
Table 1 provides the descriptive statistics, and Figure 1 shows the EDA graphs demonstrating that: The latitude difference has two peaks in the distribution, suggesting distinct clusters of residential regions at locations that are around 0.3 and 0.5 degrees from the city center. Area and community average show strong right skewness, and both house age and floor follow approximately normal distribution though most residences seem to be located at a low-level floor. Moreover, the boxplots show that house age and area indicate little group difference between the proximity to subway by the small box width differences, and outliers are all detected in the boxplots.

Table1. Summary Statistics of Continuous and Categorical Predictors

Names	Lat Difference	Price	Area	HouseAge	Floor	CommunityAve Price	Followers	DOM	Near Subway	Elevator Available	Concrete-Steel	FYP
Mean	0.4	59.71	116.92	12.67	15.18	57.2	43.42	58.42	/	/	/	/
Median	0.39	55.24	109.57	13	14	52.98	29	38	/	/	/	/
Standard Deviation	0.11	21.22	24.1	5.08	8.26	19.37	51.62	60.2	/	/	/	/
IQR	0.17	26.43	33.48	6.00	16	23.3	46	68	/	/	/	/
Range	0.61	147.55	110	62.00	60	138.47	908	650	/	/	/	/
Proportion of Yes	/	/	/	/	/	/	/	/	0.51	0.71	0.73	0.61
Proportion of No	/	/	/	/	/	/	/	/	0.49	0.29	0.27	0.39

*Notes: The money unit is in thousand CNY where 1,000 CNY = 1 thousand CNY. The relative **latitude difference** between the residence house and the center reference coordinate of Beijing city (39.56 °N), the **house age** and the **floor** of residence show similar mean and median so normal distribution is implied. The mean of **area**, residence **price** per square meter, **community average price,** number of **followers** and **DOM** is larger than the median indicating that the distribution with right-skewness. In addition, with the large standard deviation that is close or larger to the interquartile range (IQR), the price, area, followers, and DOM demonstrate a relatively large spread. The response variable **subway** proximity shows an even split between residences near a subway route and not near a subway route, and the categorical variables show that a majority of residences are installed with **elevators**, built in **concrete-steel** structures, and under the **FYP** ownership.*

Figure1. EDA plots

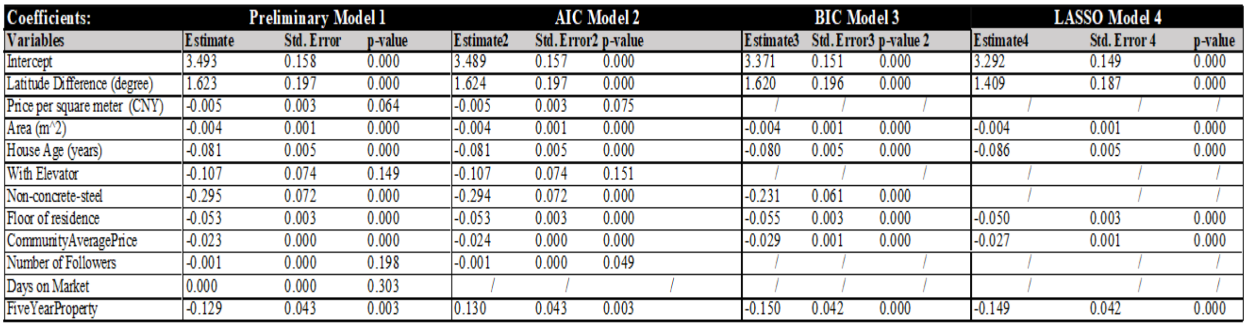


Notes: The histograms show the distribution of continuous predictors, and the boxplots relate the housing predictors with the subway route availability.

3.2 Statistical Analysis Process (Selection, Diagnostics, and Validation)

AIC stepwise selection removes DOM, and BIC stepwise selection removes price, elevator availability, followers and DOM with a greater penalty, and obtain the preferred AIC model 2 and BIC model 3. The LASSO process shrinks the non-concrete-steel structure indicator into 0 and also the same predictors that are removed as in BIC selection. According to Table 2 of the four constructed model summary, The BIC model and LASSO model has the most significant predictors as the p-value of estimated coefficients are all smaller than 0.05, and Table A1 in the Appendix shows that the multicollinearity among numerical predictors is small at VIF = 1 in both of two models, although the LASSO model has slightly smaller VIFs.

Table 2. Model Summary

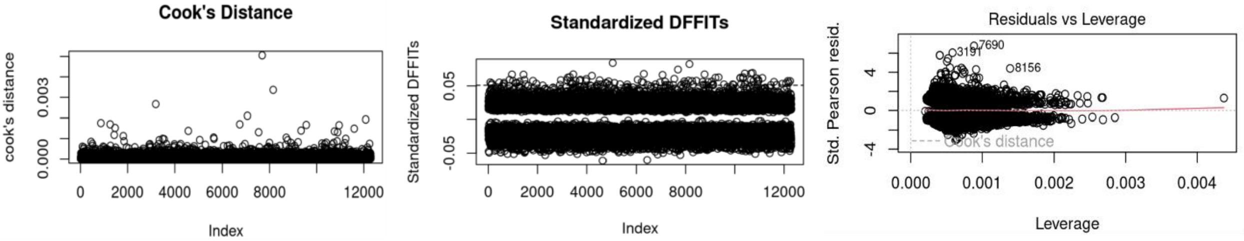


Coefficients:	Preliminary Model 1			AIC Model 2			BIC Model 3			LASSO Model 4		
Variables	Estimate	Std. Error	p-value	Estimate2	Std. Error2	p-value	Estimate3	Std. Error3	p-value 2	Estimate4	Std. Error 4	p-value
Intercept	3.493	0.158	0.000	3.489	0.157	0.000	3.371	0.151	0.000	3.292	0.149	0.000
Latitude Difference (degree)	1.623	0.197	0.000	1.624	0.197	0.000	1.620	0.196	0.000	1.409	0.187	0.000
Price per square meter (CNY)	-0.005	0.003	0.064	-0.005	0.003	0.075	/	/	/	/	/	/
Area (m^2)	-0.004	0.001	0.000	-0.004	0.001	0.000	-0.004	0.001	0.000	-0.004	0.001	0.000
House Age (years)	-0.081	0.005	0.000	-0.081	0.005	0.000	-0.080	0.005	0.000	-0.086	0.005	0.000
With Elevator	-0.107	0.074	0.149	-0.107	0.074	0.151	/	/	/	/	/	/
Non-concrete-steel	-0.295	0.072	0.000	-0.294	0.072	0.000	-0.231	0.061	0.000	/	/	/
Floor of residence	-0.053	0.003	0.000	-0.053	0.003	0.000	-0.055	0.003	0.000	-0.050	0.003	0.000
CommunityAveragePrice	-0.023	0.000	0.000	-0.024	0.000	0.000	-0.029	0.001	0.000	-0.027	0.001	0.000
Number of Followers	-0.001	0.000	0.198	-0.001	0.000	0.049	/	/	/	/	/	/
Days on Market	0.000	0.000	0.303	/	/	/	/	/	/	/	/	/
FiveYearProperty	-0.129	0.043	0.003	0.130	0.043	0.003	-0.150	0.042	0.000	-0.149	0.042	0.000

Notes: The table summarizes the estimated coefficients, standard errors of coefficients, and p-values for testing the estimated coefficients with $H_0: \beta_i = 0\ vs.\ H_a: \beta_i \neq 0$ *into 3 decimals, where "/" means the predictor is removed during selection.*

The BIC model attains no influential points on all fitted values based on cook's distance, most observations are within the acceptable cutoff of standardized DFFITS of 0.05, and a number of outliers are detected to be outside the range of (-4,4) standardized residuals based on the residuals vs leverage plot, and a prominent leverage point is observed.

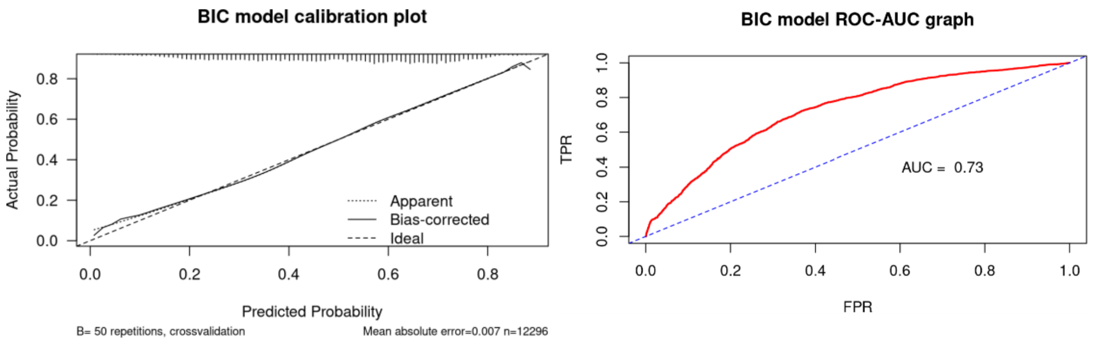
Figure 2. BIC Model Diagnostics Plots



Notes: Cook's distance plot: all cook's distance for observations are below the cutoff at quantile $F_{0.5,n=12296,n-p-1=12288}$*=0.9. Standardized DIFFTs plot: most of observations are within the standardized DFFITs cutoff at* $\left|2\sqrt{(8)/12296}\right| = \pm 0.05$. *The residuals vs leverage plot shows though most points are within the residuals = (-4,4) range, the observations with low leverage tend to be the outliers.*

The BIC model has the medium mean absolute error (MSE = 0.008) in cross-validation compared to the AIC model (MSE = 0.007) and the LASSO model (MSE = 0.005), though the error difference is very small. The BIC model has the highest AUC region of 0.73 so the model discriminates between the nearby subway route availability and not nearby subway route availability of residences 73% of the times. Therefore, the BIC model is selected to be the final model.

Figure 3. BIC model validation plots



Notes: The calibration plot for the selected BIC model shows that the predicted probability fits the actual probability well since the bias-corrected line is roughly close to the ideal 45-degree straight line though slight variation at the tails is observed, so the prediction accuracy of the model from the training set is good. The ROC-AUC graph shows that the area under the red ROC curve is 0.73 which is a good though not highly reliable indication of discrimination ability.

4. Discussion

Final fitted Model:

$$log(\frac{P}{1-P}) = log(\frac{E(Y|x_i)}{1-E(Y|x_i)}) = 3.371+1.62x_{Lat\ Diff}-0.004x_{Area}-0.08x_{HouseAge}-0.231x_{non-Concrete-Steel}$$

$$-0.055x_{floor} - 0.029x_{communityAveprice} - 0.15x_{withFYP}$$

While holding other predictors fixed, the fitted model tells that: an increase in odds ratio of subway proximity by a factor of exp(1.62) = 5.053 for one-degree increase in latitude difference, an increase in odds ratio by a factor of exp(-0.08) = 0.923 for one year increase in house age, and the odds of subway proximity in residences with the FYP ownership is exp(-0.15) = 0.86 times of the odds in residences without the ownership. In general, if the residence is newer, with an FYP and lower community average price further from the city center, the higher the odds of living nearby a subway route. However, the effects of house size, floor level, and community average price are relatively limited on the odds of response as predicted by the small coefficients. Therefore, the results suggest city dwellers purchase the new FYP houses located in the subordinate regions of city to have more chances to access a subway route, and urban designers to strengthen the subway network between the downtown and the suburbs.

Some limitations should be discussed. Firstly, the concrete-steel structure indicator is a confounder, which previous study found that the newer houses with a more modern structure are likely located near a subway route, is kept in the final model and it still significantly influences the odds of response with a negative effect because removing it may alter the estimates for other predictors and reduce the accuracy of model inference (Xu et al., 2018). Moreover, the outliers and influential points of data are not removed to ensure the intactness of residence observations for predicting the chance of subway proximity for residences in extreme cases such as the relatively large residences that are far from downtown or with a competitive community average price.

References

Ahn, K., Jang, H., & Song, Y. (2020). Economic impacts of being close to subway networks: A case study of Korean metropolitan areas. *Research in Transportation Economics*, *83*, 100900. https://doi.org/10.1016/j.retrec.2020.100900

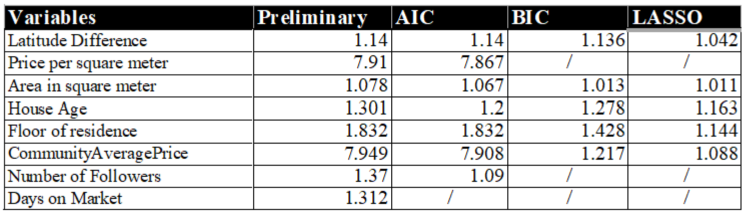
Housing Price in Beijing. (2017). Kaggle, Retrieved April 6, 2024 from https://www.kaggle.com/datasets/ruiqurm/lianjia

Wen, H., Gui, Z., Tian, C., Xiao, Y., & Fang, L. (2018). Subway opening, traffic accessibility, and housing prices: A quantile hedonic analysis in hangzhou, china. Sustainability, 10(7), 2254. https://doi.org/10.3390/su10072254

Xu, Y., Zhang, Q., Zheng, S., & Zhu, G. (2018). House age, price and rent: Implications from land-structure decomposition. *The Journal of Real Estate Finance and Economics*, *56*(2), pp. 303–324. https://doi.org/10.1007/s11146-016-9596-6

Appendix

Table A1. Variance Inflation Factor (VIF) of Models



Variables	Preliminary	AIC	BIC	LASSO
Latitude Difference	1.14	1.14	1.136	1.042
Price per square meter	7.91	7.867	/	/
Area in square meter	1.078	1.067	1.013	1.011
House Age	1.301	1.2	1.278	1.163
Floor of residence	1.832	1.832	1.428	1.144
CommunityAveragePrice	7.949	7.908	1.217	1.088
Number of Followers	1.37	1.09	/	/
Days on Market	1.312	/	/	/

Notes: This table shows the VIF which measures the multicollinearity of numerical predictors in each constructed model during variable selection, if VIF is around 1 then multicollinearity is small.