Predicting the house price in Astana

Myrzakhmetov Bagdat

November 20, 2017

Abstract

Prediction of the house price is an attractive tasks in Machine learning community. House price affected by many factors of the house, as well as dependent on the time, and other economic factors. Multivariate linear regression model can be used to predict the house price. In this project we attempted applying the regression model to our collected datasets. For this project, all of the selling houses in Astana were collected from "Krysha" web-page. Overall, the fitted models to these datasets showed high accuracy in both intrinsic and extrinsic evaluations. Extrinsic evaluation performed on hold-on validation datasets, and showed lower errors.

1 Introduction

In this project, the analysis of the house price in Astana has been implemented. Prediction of house prices is an important information for buyers as well as for the sellers. If someone wants to buy a house, one can estimate the price of flat with the given parameters, attributes. Also, prediction of the house price is important for sellers, as sellers can put their price based on the attributes of their apartments and estimate their house's price by using the predictive models. In this project, we attempted to predict the house price in Astana. To predict, firstly, the datasets were collected and preprocessed for the models. There are many factors which affects to the house price. Based on the availability of the parameters, we extracted several attributes from our collected datasets.

1. X_1- Number of rooms in the house
2. X_2- Apartment area in square meters
3. X_3- Residential area in square meters
4. X_4- Year of construction
5. X_5- Floor of the house
6. X_6 - X_7- Is the apartment fully furnished?
7. X_8- Is the apartment close to school? (0 no, 1 yes)
8. X_9- Is it a kitchen-studio? (0 no, 1 yes)
9. $X_{10}-$ Is it an urgent sale? (0 no, 1 yes)
10. $X_{11}-$ Number of comments
11. $X_{12}-$ Number of views
12. X_{13} - $X_{14}-$ District of the house

Here, all of the predictors were chosen artificially with some sense. Factors X_1, X_2 X_4, X_5, X_{13}-X_{14} could be considered as one of the most important information. Most people will look at the area of an apartment, number of rooms, in which year it was built and in which floor it is located. Residential area, or in other words useful area of the house can be considered as important. Because, some apartments might have a large area, but their balcony, kitchen, hallways can be large. So considering this fact separately might help us to predict the price more accurately. To give more information about the house, some additional features were added. For example, it's possible to extract information about furnishings of an apartment. It could be one the tree cases: fully furnished, partially furnished and no furnishings. Respectively, the house price might differ depending on these values.

Another factor, that people may look at during the buying of the house is it's location and proximity to the building such as school, kindergarten, bus stops and etc. The additional feature "Is the house near to School?" also added as a predictor to the model.

If the advertisements contains the information about urgency (urgent, bargaining), this kind of an apartment could be cheaper than the others. Sellers wants to sell in a short time of period, they put lower cost to their apartments. Considering these facts also might help to model well the house price.

Factor X_9 is about the houses which is a kitchen-studio. This parameter is true mainly for one-roomed or two-roomed flats, where living room and kitchen are one room. This feature is also considered as important, because, in many cases kitchen-studio apartments are cheaper than the other apartment with one or two-rooms, without a kitchen-studio. For instance, two-roomed kitchen studio might cost between one-roomed and two-roomed apartments. So, it might help to predict more accurately.

Last two attributes are: number of comments on the advertisement and number of views. Here, we assume, that if there are many comments, then there are many interests for that apartment. So, this shows that people are interested to buy for the given price, price for the apartments are just normal, or even could be cheaper than others. More views also indicates the interests of other people, who are more likely to buy the apartment.

There are some categorical attributes. In regression models and in many machine learning tasks, modeling of the categorical datasets are not so easy. In many cases, "one hot encodings" or just giving the "dumb" values will be used to model the datasets. In our case, we have categorical attributes, such as "Is the apartment close to school?", "Is the apartment an urgent sale?", "District", "furnishings", "Kitchen-studio". For many categorical attributes, there are only 2 variables, so they are easy to model. For "District" and "Furnishings" attributes, there are 3 variables. Here, for these attributes, we can choose just one attribute with "dumb" variables, or additional 2 attributes with "one-hot encodings".

Our regression analysis showed that these attributes are indeed good for the prediction.

Fitted best model was

$$ln(Y) = 1.633099 + 0.022851X_2 + (-0.000443)X_4 + 0.097862X_6 + 0.158212X_7 + 0.056313X_8 +$$

$$(-0.080342)X_{10} + 0.008484X_{11} + (-0.000071383)X_{12} + 0.078418X_{13} + 0.100452X_{14} + (-0.000044282)X_2X_2 +$$

$$(-0.001147)X_2X_7 + (-0.000917)X_2X_8 + (-0.000147)X_2X_{11} + 0.000000513X_2X_{12} + 0.000994X_2X_{13} +$$

$$0.000059344X_4X_4 + (-0.003372)X_4X_6 + (-0.004782)X_4X_7 + (-0.000000857)X_4X_{12} + (-0.002630)X_4X_{14} +$$

$$0.000023616X_6X_{12} + 0.073992X_6X_{13} + 0.061545X_7X_{13} + (-0.068428)X_8X_{13}$$

This means that the price of the house were depend on the factors such as depends on the factors such as Apartment area, year of construction, furnishings, closeness to the school, district, number of comments, number of views and urgency of the selling. Our study shows these factors as an important attributes.

2 Preparation of the datasets

2.1 Preprocessing datasets

Datasets contain information about 17582 apartments in whole Astana city. Datasets were collected from the www.krysha.kz web-page by using own Python Script. Collected datasets has been processed further by using BeautifulSoup, Requests libraries in Python. Firstly, all of the advertisements in "Krysha" web-page stored in a separate files. This file contains: Title, url, image links, owner, address, district, and all available parameters, such as, number of rooms, apartment area, residential area, balcony and other parameters. One of this file example shown in Figure 1.

Then from these whole datasets were extracted attributes, which are common for all advertisements. For example, one dataset might contain information about balcony, but others don't. So the attributes were chosen in that way, so we might have less data sparsity. Also, many of the advertisements came from the agencies, one flat might be advertised several times with different agencies. To exclude this, only chosen advertisements, whose owner is houses' real owner. (I exclude all advertisements, which came from the agencies). Also some of the datasets, which might not have enough information, were excluded. At the end, obtained one single file with all advertisements with only appropriate attributes for each districts. This obtained document can be seen in figure 2.

The last step is prepare these datasets for SAS. To do so, we have to take integer values. The final output for SAS is given figure 3. Here, we also have to consider the categorical datasets. I will describe later how do I dealt with the categorical datasets.

"Krysha" provides some attribute information, parameters that can extracted easily from the web-page only considering the HTML structure of the texts. However, these attributes does not tell much about the whole flat. Many parts of the advertisement information are unstructured, we cannot get more useful predictors just by these structure features. In order to add some other attributes, I processed the texts and extracted these attributes from the input texts. For example, attribute "Is the apartment close to school?" can be extracted from. In the text, one can see the word "near school", "School 50 is not far away from the flat". If the texts contains such as information, then these datasets are considered to close to the school. At the beginning, initial idea was using some lemmatization or word2Vec approach to find the exact meanings of the words. But for this case, only using a bag-of-words approach solved the problem well. For instance, if there are words like "school" (in Russian) in the text, then an attribute will be 1, and 0 otherwise.

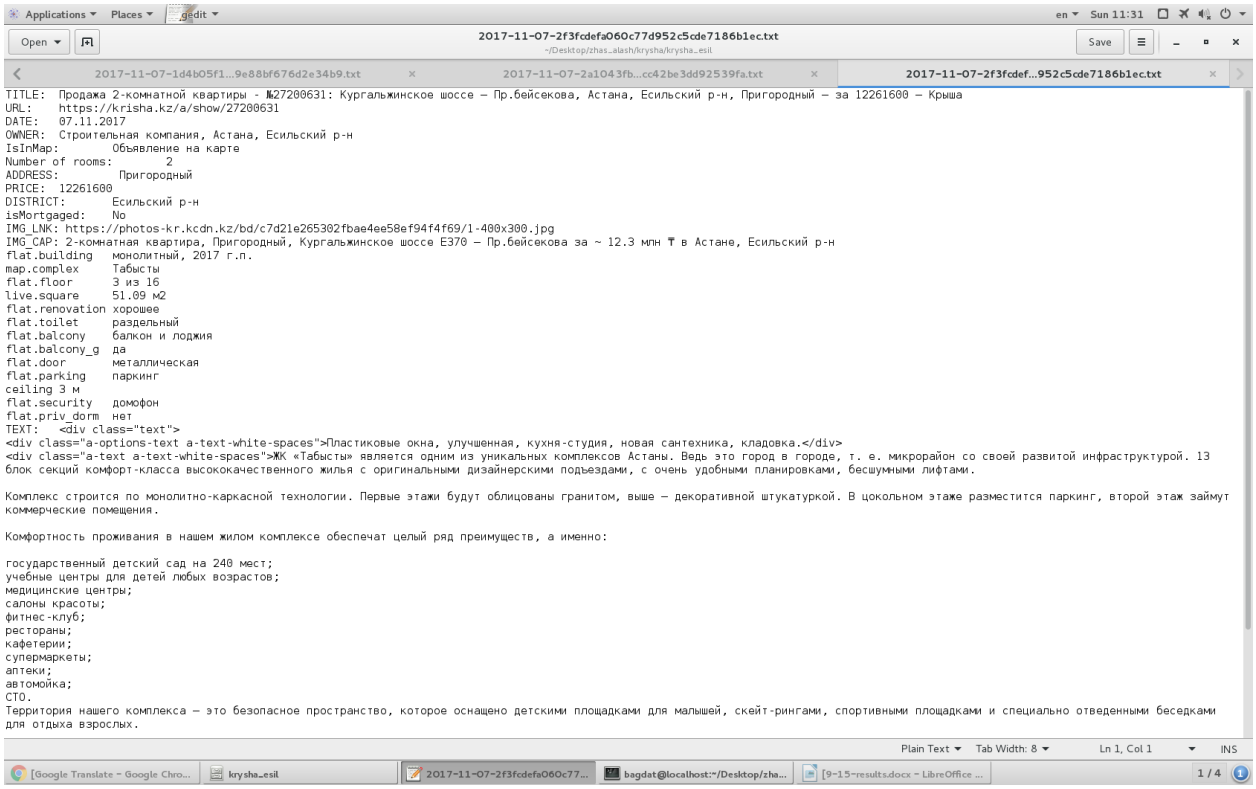


Figure 1: Initial form of the crawled datasets.

Another attribute "Is a kitchen studio?". This is true mainly for one-roomed or two-roomed flats, where the living room is together with the kitchen. This attribute extracted from the text considering the words itself and also considering the surrounded words (could be written "not a kitchen studio", so don't have to take this case into account.)

The district information modeled with categorical variables. To model district properly, at the beginning, I chose, 1 for Esil district, 2 - Almaty and 3 for Saryarka district. This is just using the "dump" values to indicate the categorical datasets. However, after first order models, if we want to check interactions between the attributes, this "dump" encoding might not work properly, as we indicate them as a continuous variables and values might give some significance. In order to solve this problem. one can use "one-hot encodings" to model the categorical datasets. For district, I also considered "one-hot encodings". For three values, we might need 2 attributes, like "Is the apartment in the district Esil?", (1 if Esil district and 0 otherwise) "Is the apartment in the district Almaty?" (1 if Almaty district and 0 otherwise). We consider that if the both values are zero, then the apartment is in the Saryarka district.

For the "Furniture" attribute, one can also use continuous variables. such as 0-2. If 0, then this means that the apartment is not furnished, 1 if the apartment is partially furnished and 2 for fully furnished. But for this task I used "one-hot encoding" approach. So to encode categorical data with 3 variables, we need two additional attributes.

Last attribute, which I played around with was an attribute "Year of construction." At the beginning I chose the year itself, for example, 2017, 1985. Just for interest I fitted the model with this values. After that, I subtracted these values from 2017, just to indicate how many years was for the buildings, also fitted the model, and checked the R^2_{adj} values for both cases. For the second case, R^2_{adj} were higher than the first case. So I decided to use the age of the buildings.

For the predicted value, I divided it to 1 million, in order to get small numbers.

Collected datasets are up-to-date. In future, we are planning to put datasets into open-source tool, that any people can access and build their own models.

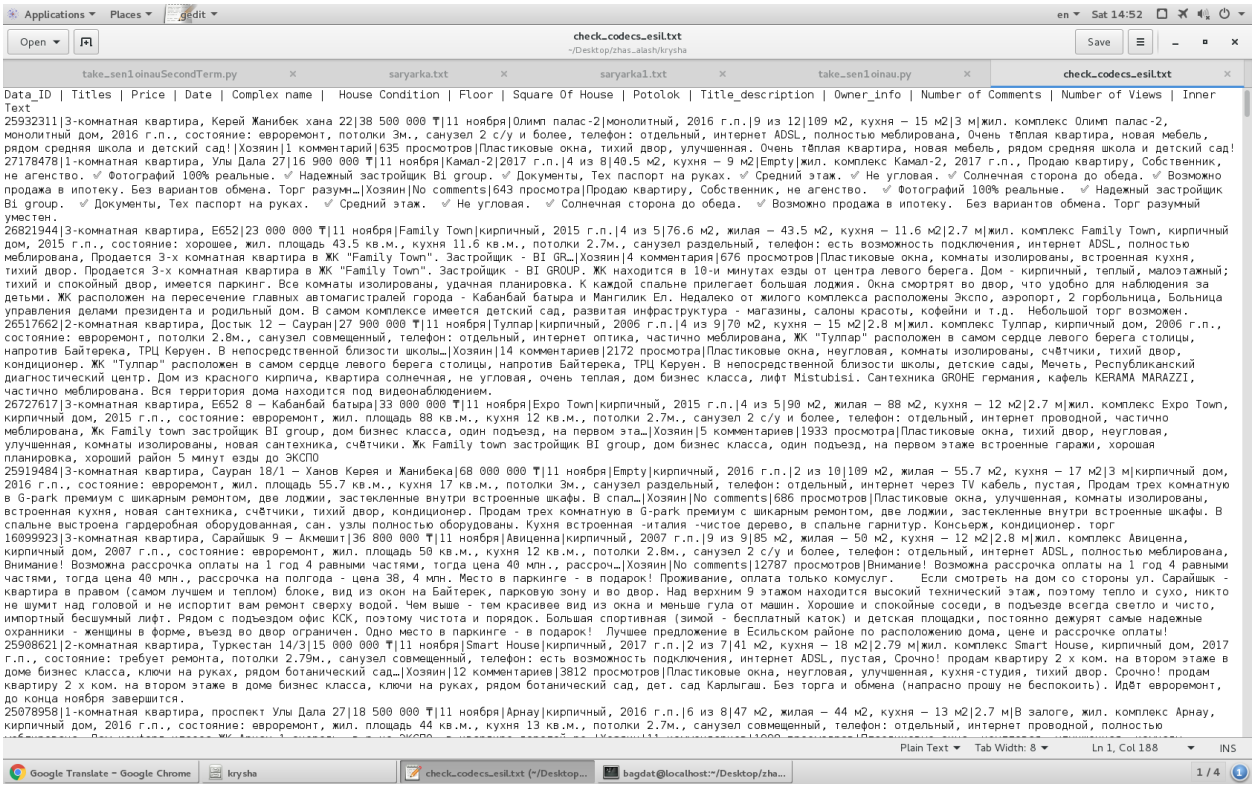


Figure 2: Transformed into one file.

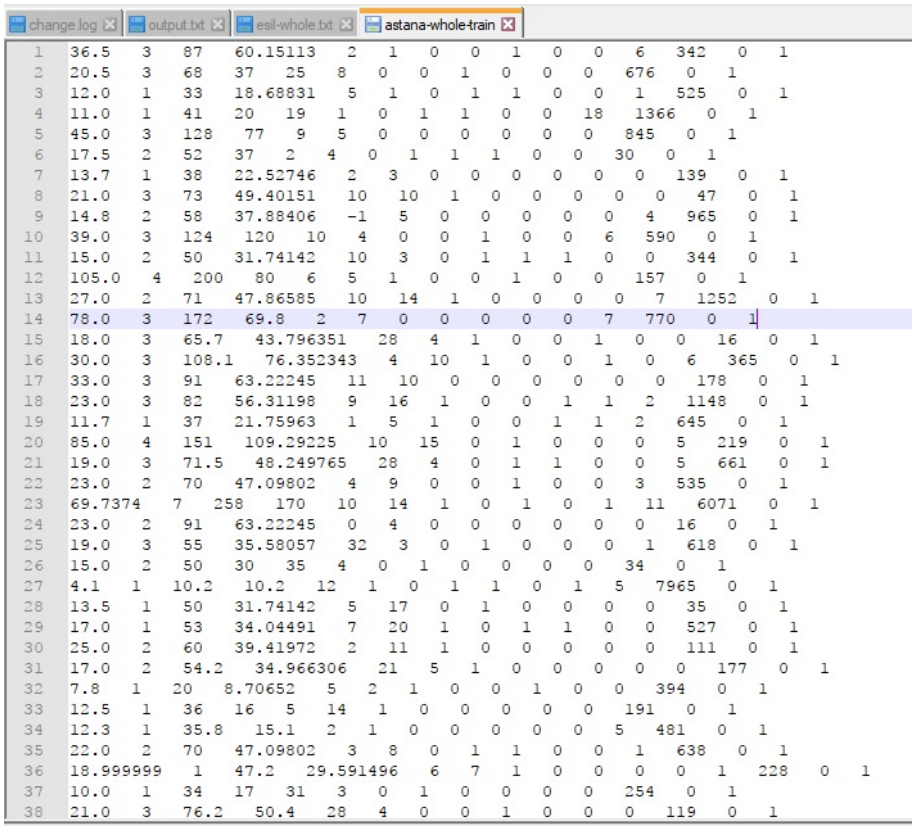


Figure 3: Feeding into SAS.

2.2 Threating the Missing values

Adding an extra predictor for a residential area might be useful. Some flats might have a large area, but their living area could be small, as full area consists a kitchen, balcony, hallway. In the datasets, some values for this attribute is missing. But all flats have total square meters attribute.

One idea is this feature might be correlated with the total square meters, so we can use regression to fill these fields. For example, X - is the total area of the flat, and Y is the square meters of the living room. After putting the regression, I explored that living room and total area is related with:

$$Y = -6.65 + 0.76783X$$

In some examples, there were not enough information about the year of buildings, number of rooms, total area. These attributes are important, if they are missing, then there are something wrong. I have looked at some of these examples, and noticed, that in this advertisements, there is no photos, no years, short texts or very short descriptions. These kind of advertisements could be considered as "spam" or authors did very quick postings, by not paying more attention on the advertisement content, many information are missing. Or its actually fake information. These kind of advertisements are useless, and could be considered as a "spam". I guess, Krysha should have some kind of filtering function, which sorts out these kind of advertisements. So I removed these cases. All of elements with missing values are removed.

3 Building of the model

3.1 Initial sanity checks

After removing and threating the missing values, initial edit checks should be performed before building the regression models. We assume, that our data is an error free, and there is no missing values. But before the building of the models with all these datasets, we have to do some initial sanity checks for the datasets. As collected and preprocessed datasets are large (3293 entries overall), so it's likely that the datasets might have some error.There are might be some extreme outliers, which might not be fitted with any kind of models. They will heavily mislead whole models. I looked at the initial regression and also looked at the residuals. From here we can see many outliers. Firstly, I looked at the values of Y (price). For one example, Cook's distance was 12.4237. This is very high value, this should be removed. This can be seen here: https://krisha.kz/a/show/27148501 Here, the apartment with a total area of 768 square meters, but that is impossible for 4-roomed apartment, as there are some typo in the advertisement. For this example, Studentized residuals and hat matrix leverage values are high.

Another observation was with Cook's distance 0.81186. 5 roomed apartment in Paris Quarter for 500 million tenge. This is a real advertisement, not a "spam" or a typo, but it's price is comparatively very high compared to other houses.

This highly priced apartments could be modeled separately. One idea was adding the separate attribute to indicate whether the apartment is "luxury", expensive one or equipped with expensive furniture, or in the expensive apartments (HighVill, Paris, Italian quarter) compared to normal apartments. But observations were not many. So I decided to remove all of these datasets as outliers (Overall 8 highly expensive apartments) from my training and testing sets.

At the beginning, the R^2_{adj} was 0.6387 and without outliers - 0.7605. This shows that removing the outliers helps us to fit the regression model well.

Now our observations are free from the outliers, we can safely split our datasets into training and validation sets. One can estimate parameters only based on the whole training set, but in this case, predicted model might work well on the training data and there are might happen over-fitting of the datasets. To check that model is indeed good for other datasets and to see that there is no over-fitting, we can check our model estimations for different unseen datasets. I divided datasets into training and testing sets randomly. 80% for training and 20% for testing the model, (2612 and 651 entries respectively).

3.2 Transformations of the variables

After data preparation, removing the outliers and splitting the datasets into training and validation set, we can build our models: multivariate linear regression model with all attributes. The

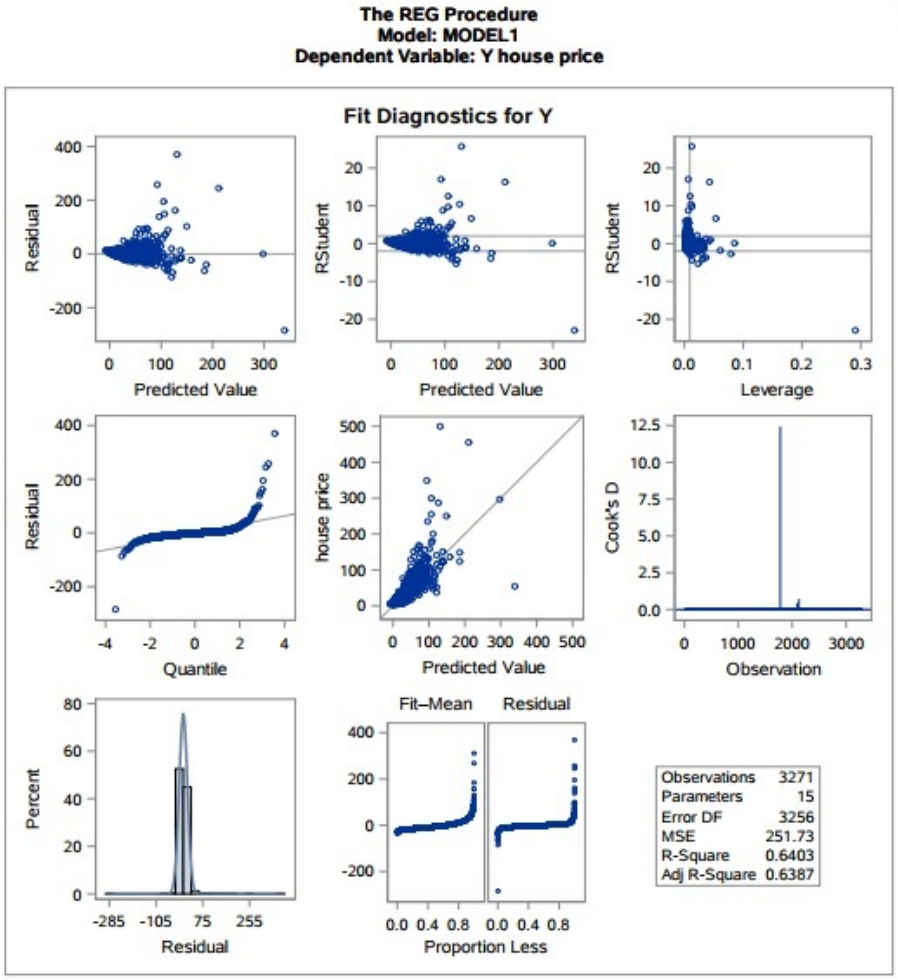


Figure 4: Initial plot of the regression.

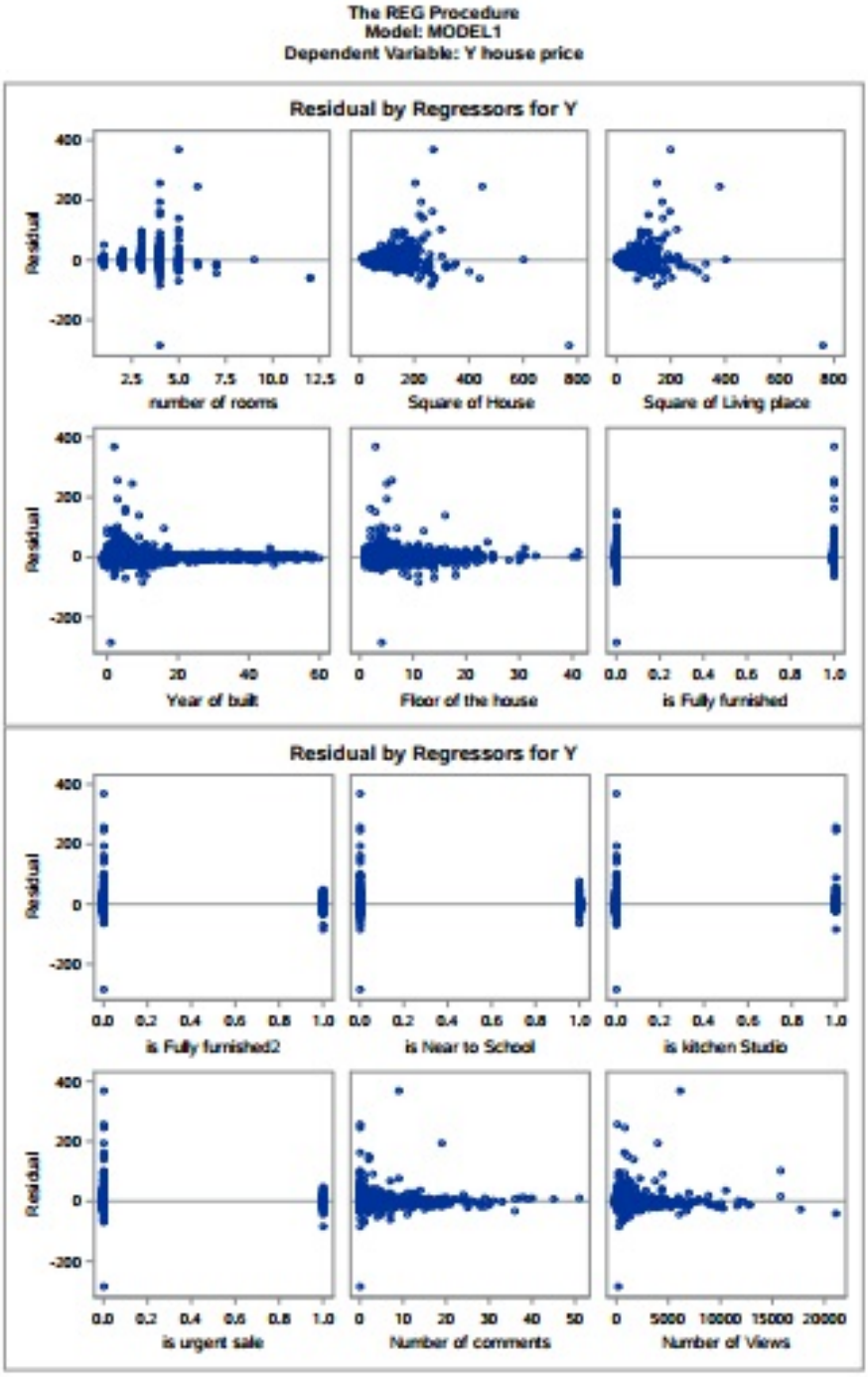
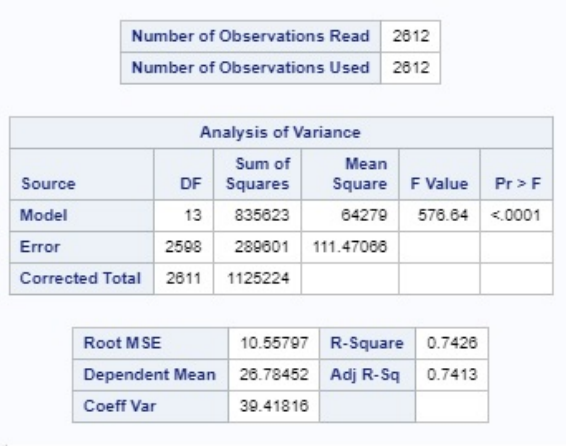


Figure 5: Initail residual plots.



Number of Observations Read	2612
Number of Observations Used	2612

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	835623	64279	576.64	<.0001
Error	2598	289601	111.47066		
Corrected Total	2611	1125224			

Root MSE	10.55797	R-Square	0.7426
Dependent Mean	26.78452	Adj R-Sq	0.7413
Coeff Var	39.41816		

Figure 6: ANOVA table for initial model.

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6.80269	0.86501	-7.86	<.0001
x1	number of rooms	1	-2.46319	0.37383	-6.59	<.0001
x2	Square of House	1	0.49591	0.02040	24.31	<.0001
x3	Square of Living place	1	0.01414	0.02403	0.59	0.5564
x4	Year of built	1	0.02266	0.02076	1.09	0.2752
x5	Floor of the house	1	-0.17326	0.04854	-3.57	0.0004
x6	is Fully furnished	1	3.61117	0.50773	7.11	<.0001
x7	is Fully furnished2	1	1.00348	0.54702	1.83	0.0667
x8	is Near to School	1	-1.73091	0.44025	-3.93	<.0001
x9	is kitchen Studio	1	1.95120	0.54987	3.55	0.0004
x10	is urgent sale	1	-2.32107	0.62075	-3.74	0.0002
x11	Number of comments	1	-0.20751	0.04826	-4.30	<.0001
x13	Is the district Esil	1	5.27636	0.63762	8.28	<.0001
x14	Is the district Almaty	1	1.22401	0.56575	2.16	0.0306

Table 1: Parameter estimates.

parameter estimations are given in table 1. ANOVA table is given in 6 figure. From, here, we can see that, R^2_{adj} value for the fitted model is 0.7412 and MSE for the model is 111.47. This model can be considered as a moderate fitting model. We might need some transformations or higher order terms to get the better fittings.

By looking just at p-values for each estimators, we don't just drop the first order terms. In some cases, variables with the second order values could have a significant impact to the overall regression model. So, in this stage, I decided to keep all attributes.

In order to check the models fittings and predictors' dependence from the attribute parameters, we can do testing for attribute parameters, or do some transformations. By looking at the fitted model, we can see from the figure 7 that the relations are non-linear, we can also look at Q-Q plot and the normality of the residuals are not so well fitted. For these reasons, we might need some transformations. In order to transform the datasets into correct form, we can test our model for Box-Cox testings with SAS easily. Box-cox transformation will give us suitable λ parameter that can be used to transform the datasets. λ parameter learned from the datasets. Depending on the value of λ parameter, we can to some transformation. After applying the Box-Cox test in SAS, we see that the appropriate λ value will be near to zero. Maximum likelihood estimation and R^2 estimation for $\lambda = 0$ shows good results. The results of λ calculation and Box-Cox plot itself are given in figures 8 and 9.

Now, to implement the transformation for our predictor Y, we have to take a logarithm for the predictor. After applying the log transformation, adjusted R^2_{adj} value increased up to 0.7876 and mean squared error of the model decreased to 0.07493. Also, we can see the fitted model against the real house price (from figure 10). Also, if we look at the Q-Q plot for the residual, we can see

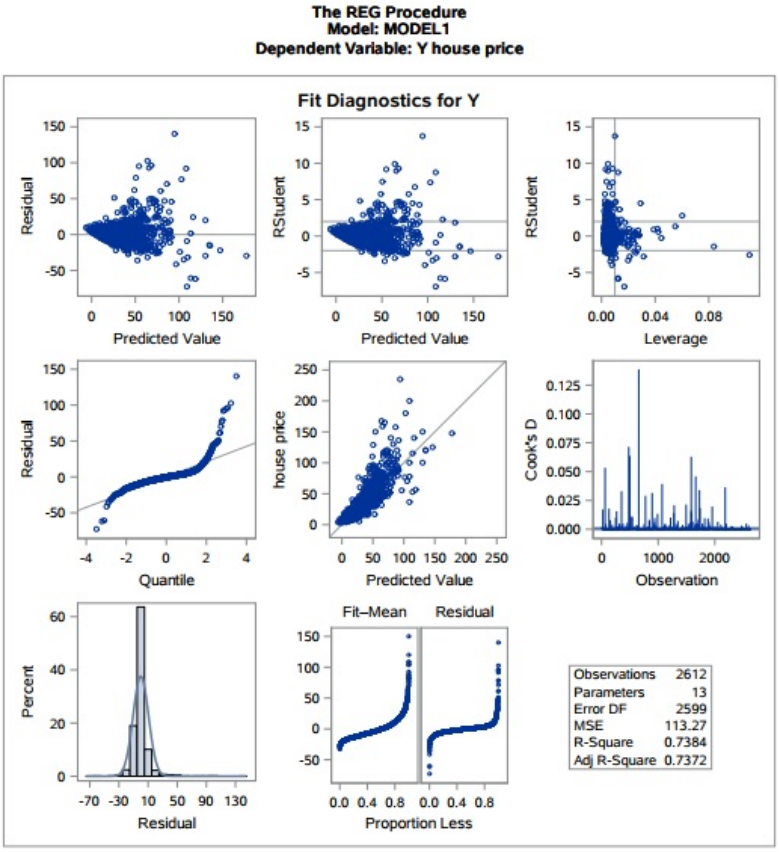


Figure 7: .Residual plots of the fitted models

The TRANSREG Procedure

Box-Cox Transformation Information for house price

Lambda		R-Square	Log Like	
-3.0		0.13	-12183.7	
-2.5		0.19	-10327.4	
-2.0		0.29	-8651.4	
-1.5		0.43	-7191.7	
-1.0		0.58	-5989.8	
-0.5		0.71	-5108.5	
0.0	+	0.79	-4691.7	<
0.5		0.80	-5008.6	
1.0		0.74	-6156.3	
1.5		0.63	-7863.1	
2.0		0.49	-9878.0	
2.5		0.36	-12092.4	
3.0		0.26	-14457.9	

< - Best Lambda
* - 95% Confidence Interval
+ - Convenient Lambda

Figure 8: Box-cox parameter estimation table

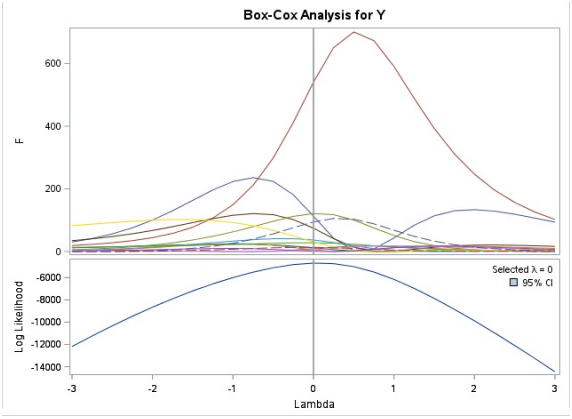


Figure 9: Box-cox figure.

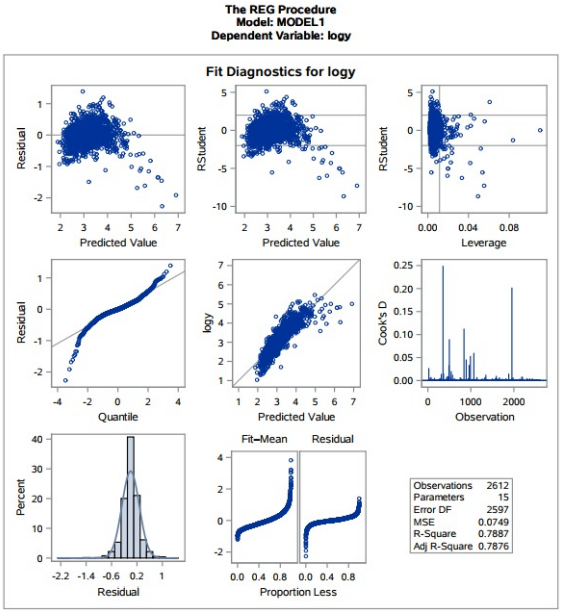


Figure 10: Fitted model with after applying the logarithm function to Y.

the residuals are well normally distributed. Overall, transforming the Y into log(Y) will increase the models performance.

we can also test for the normality of the attributes with Kolgomorov-Smirnov, Anderson tests. The results of these tests are given in Figure 11.

By looking at Kolgomorov-Smitnov, Anderson tests, we can say that the normality assumption is hold.

3.3 Removing correlated attributes

Checking for the collinearity is one of important factors. We might have unnecessary attributes that can be modeled by other attributes. If we look at the table 1, we can see that the p-value for X_3 is relatively high and Standard error (SSR) for X_3 is small. This means that extra sum of squares for adding X_3 into model is small, because, the information that X_3 carries is already in the model. To check whether the elements are indeed correlated with each other.

We can check the diagonal kernel matrix to see the relationships between the elements. This graph indicated very strong correlation between the parameters x2 and x3 and also between x2 and x1. In figure 12 we can see the correlations between the elements.

Also Variance inflation factor (VIF) again indicates the fact that attributes X_2 and X_3 are highly correlated. For VIF criteria, we may want to remove the attributes which has VIF value above than 1. VIF graph can be seen from the figure 13.

The UNIVARIATE Procedure
Variable: residual (Residual)

Moments			
N	2612	Sum Weights	2612
Mean	0	Sum Observations	0
Std Deviation	0.27462122	Variance	0.07541681
Skewness	-0.6802643	Kurtosis	6.04414382
Uncorrected SS	196.913299	Corrected SS	196.913299
Coeff Variation	.	Std Error Mean	0.00537338

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	0.27462
Median	-0.00201	Variance	0.07542
Mode	-0.03508	Range	3.66793
		Interquartile Range	0.26746

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-12	Pr >= |M|	0.6527
Signed Rank	S	32731	Pr >= |S|	0.3959

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.086494	Pr > D	<0.0100
Cramer-von Mises	W-Sq	5.712267	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	32.88535	Pr > A-Sq	<0.0050

Figure 11: Tests for normality.

Correlation															
Variable	Label	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
x1	number of rooms	1.0000	0.8103	0.7889	0.0651	0.0551	-0.0062	0.0740	-0.0189	-0.0662	-0.0002	0.0835	0.0847	-0.0007	0.0389
x2	Square of House	0.8103	1.0000	0.9588	-0.1583	0.1492	0.0140	0.0039	-0.0712	-0.1076	0.0000	0.0757	0.0913	0.0770	0.0169
x3	Square of Living place	0.7889	0.9588	1.0000	-0.1376	0.1363	0.0199	0.0037	-0.0669	-0.0888	0.0070	0.0691	0.0892	0.0821	0.0103
x4	Year of built	0.0651	-0.1583	-0.1376	1.0000	-0.2601	0.0315	0.1490	0.1061	-0.0250	-0.0390	-0.0014	-0.0460	-0.3976	0.0992
x5	Floor of the house	0.0551	0.1492	0.1363	-0.2601	1.0000	0.0488	-0.0119	-0.0072	0.0652	0.0411	-0.0153	-0.0068	-0.0318	0.0708
x6	is Fully furnished	-0.0062	0.0140	0.0199	0.0315	0.0488	1.0000	-0.5181	0.0069	0.1157	0.0007	0.0075	-0.0102	-0.0285	0.0536
x7	is Fully furnished2	0.0740	0.0039	0.0037	0.1490	-0.0119	-0.5181	1.0000	0.0580	-0.0173	-0.0377	0.0061	-0.0400	-0.1342	0.0671
x8	is Near to School	-0.0189	-0.0712	-0.0669	0.1061	-0.0072	0.0069	0.0580	1.0000	0.0142	0.0209	0.0017	0.0094	-0.1352	0.0704
x9	is kitchen Studio	-0.0662	-0.1076	-0.0888	-0.0250	0.0652	0.1157	-0.0173	0.0142	1.0000	0.0409	0.0098	0.0497	-0.0236	0.0396
x10	is urgent sale	-0.0002	0.0000	0.0070	-0.0390	0.0411	0.0007	-0.0377	0.0209	0.0409	1.0000	0.0317	0.0603	0.0275	-0.0350
x11	Number of comments	0.0835	0.0757	0.0691	-0.0014	-0.0153	0.0075	0.0061	0.0017	0.0098	0.0317	1.0000	0.3510	0.0034	0.0043
x12	Number of Views	0.0847	0.0913	0.0892	-0.0460	-0.0068	-0.0102	-0.0400	0.0094	0.0497	0.0603	0.3510	1.0000	0.0491	-0.0209
x13	Is the district Esil	-0.0007	0.0770	0.0821	-0.3976	-0.0318	-0.0285	-0.1342	-0.1352	-0.0236	0.0275	0.0034	0.0491	1.0000	-0.6473
x14	Is the district Almaty	0.0389	0.0169	0.0103	0.0992	0.0708	0.0536	0.0671	0.0704	0.0396	-0.0350	0.0043	-0.0209	-0.6473	1.0000

Figure 12: Correlation matrix.

Root MSE	10.55864	R-Square	0.7427
Dependent Mean	26.78452	Adj R-Sq	0.7413
Coeff Var	39.42066		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-6.85779	0.86767	-7.90	<.0001	0
x1	number of rooms	1	-2.46797	0.37390	-6.60	<.0001	3.33962
x2	Square of House	1	0.49577	0.02040	24.30	<.0001	14.53673
x3	Square of Living place	1	0.01401	0.02403	0.58	0.5599	12.56100
x4	Year of built	1	0.02297	0.02076	1.11	0.2688	1.55451
x5	Floor of the house	1	-0.17251	0.04855	-3.55	0.0004	1.12823
x6	is Fully furnished	1	3.62990	0.50827	7.14	<.0001	1.44011
x7	is Fully furnished2	1	1.02826	0.54789	1.88	0.0607	1.46432
x8	is Near to School	1	-1.73895	0.44039	-3.95	<.0001	1.02996
x9	is kitchen Studio	1	1.92413	0.55089	3.49	0.0005	1.04921
x10	is urgent sale	1	-2.34474	0.62146	-3.77	0.0002	1.01169
x11	Number of comments	1	-0.22207	0.05143	-4.32	<.0001	1.14656
x12	Number of Views	1	0.00014399	0.00017578	0.82	0.4128	1.16078
x13	Is the district Esil	1	5.26478	0.63782	8.25	<.0001	2.20277
x14	Is the district Almaty	1	1.22266	0.56579	2.16	0.0308	1.83191

Figure 13: VIF values.

So considering the above criteria, I removed the attribute X_3 - Residential area of the flat, then I noticed that R^2_{adj} increased up to 0.78. But now, VIF for X_1 and X_2 were high. X_1 - number of rooms, also highly correlated with the square meters of the apartment. If we think, of course, this is true, more rooms corresponds to more area of the apartment, or if the area of the flat is large, then one might have a large residential (living) area. After removing the X_3 attribute, VIF values for X_1 and X_2 were 3.32990 and 3.45225 respectively. These values are not larger than 10, so I decided to leave both attributes.

3.4 Second order terms.

In this part, we explored the interaction between the predictions. In some cases, one predictor might not be a good predictor, instead, their interaction with other predictors might be more useful.

There are two methods that can automatically pick the best parameters for the models: **best subset** method and **stepwise** procedures.*proc glmselect* command in SAS will do backward elimination and forward selection stepwise procedures.

Also, we can get the best attributes in a hierarchical order. For stepwise procedure, I chose cutoffs suggested in the textbook select=sl, also sle=0.1 is entry cutoff and sls=0.15 is cutoff for staying in the model. We can also choose (AIC, BIC, Cp, R^2 instead of p-value cutoffs. So, I did backward and forward stepwise methods to choose the best model among the all possible combinations.

Final estimation of the parameter values by forward and backward selections are given in Figure 14 and in Figure 15 given test statistics.

Also MSE decreased to 0.05243 and R_adj^2 values increased up to 0.8528. Also, if we look at general fittings from the Figure 16, normal qq plots and Q-Q plots are well modeled. Also if we look at the prediction versus real price, they are all lies one the line, at fits the data well.

The second method for selecting the best subsets is best subsets according to c^p value. But for this task, best subset selection based on c^p value wasn't implemented in SAS, as I have too many instances and attributes, computationally it wasn't possible.

I also extracted third order terms with Python script, and tried to model with SAS. But, implementing with third order terms also was impossible, because of the computation limitations.

3.5 Testing on the validation set

As we said in 3.1 Section, checking the quality of the model should not only be done with intrinsic evaluation of the regression models. To check the robustness of the proposed model, we also have to estimate our models on the validation - unseen data sets. I wrote Python script to test the model's performance on the test set. Estimated R_2 value for the validation set with my best model, and this was 0.835. This is a good result, this means that there is no over-fittings. Model can be used in a real world tasks.

3.6 Interpretation of the best model

Obtained best model for predicting the house price in Astana were somehow complicated:

$$ln(Y) = 1.633099 + 0.022851X_2 + (-0.000443)X_4 + 0.097862X_6 + 0.158212X_7 + 0.056313X_8 +$$

$$(-0.080342)X_{10} + 0.008484X_{11} + (-0.000071383)X_{12} + 0.078418X_{13} + 0.100452X_{14} + (-0.000044282)X_2X_2 +$$

$$(-0.001147)X_2X_7 + (-0.000917)X_2X_8 + (-0.000147)X_2X_{11} + 0.000000513X_2X_{12} + 0.000994X_2X_{13} +$$

$$0.000059344X_4X_4 + (-0.003372)X_4X_6 + (-0.004782)X_4X_7 + (-0.000000857)X_4X_{12} + (-0.002630)X_4X_{14} +$$

$$0.000023616X_6X_{12} + 0.073992X_6X_{13} + 0.061545X_7X_{13} + (-0.068428)X_8X_{13}$$

where X_2 is an apartment area, X_4 is year of construction, X_7 is about "Furnishings", X_8 is about closeness to the school, X_{10} is about "urgency", X_{11} is a number of comments, X_{12} is a number of views, X_{13} - X_{14} is location. By looking at this model, we can notice that some of the attributes are not needed to get the better predictions, instead we might need to add some extra high order interactions.

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	1.633099	0.027755	58.84
x2	1	0.022851	0.000374	61.06
x4	1	-0.000443	0.001603	-0.28
x6	1	0.097862	0.019737	4.96
x7	1	0.158212	0.030335	5.22
x8	1	0.056313	0.021430	2.63
x10	1	-0.080342	0.013490	-5.96
x11	1	0.008484	0.002447	3.47
x12	1	-0.000071383	0.000009433	-7.57
x13	1	0.078418	0.028587	2.74
x14	1	0.100452	0.017121	5.87
x2*x2	1	-0.000044282	0.000001336	-33.16
x2*x7	1	-0.001147	0.000291	-3.94
x2*x8	1	-0.000917	0.000262	-3.50
x2*x11	1	-0.000147	0.000027473	-5.36
x2*x12	1	0.000000513	8.8300636E-8	5.80
x2*x13	1	0.000994	0.000245	4.05
x4*x4	1	0.000059344	0.000028274	2.10
x4*x6	1	-0.003372	0.001035	-3.26
x4*x7	1	-0.004782	0.001043	-4.58
x4*x12	1	-0.000000857	0.000000381	-2.25
x4*x14	1	-0.002630	0.000869	-3.02
x6*x12	1	0.000023616	0.000007578	3.12
x6*x13	1	0.073992	0.024090	3.07
x7*x13	1	0.061545	0.026990	2.28
x8*x13	1	-0.068428	0.020616	-3.32

Figure 14: Final estimation of the values.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 25).

Effects:	Intercept x2 x4 x6 x7 x8 x10 x11 x12 x13 x14 x2*x2 x2*x7 x2*x8 x2*x11 x2*x12 x2*x13 x4*x4 x4*x6 x4*x7 x4*x12 x4*x14 x6*x12 x6*x13 x7*x13 x8*x13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	25	785.57092	31.42284	599.36
Error	2586	135.57775	0.05243	
Corrected Total	2611	921.14867		

Root MSE	0.22897
Dependent Mean	3.09178
R-Square	0.8528
Adj R-Sq	0.8514
AIC	-5061.14801
AICC	-5060.56287
SBC	-7522.58335

Figure 15: ANOVA table of the final model.

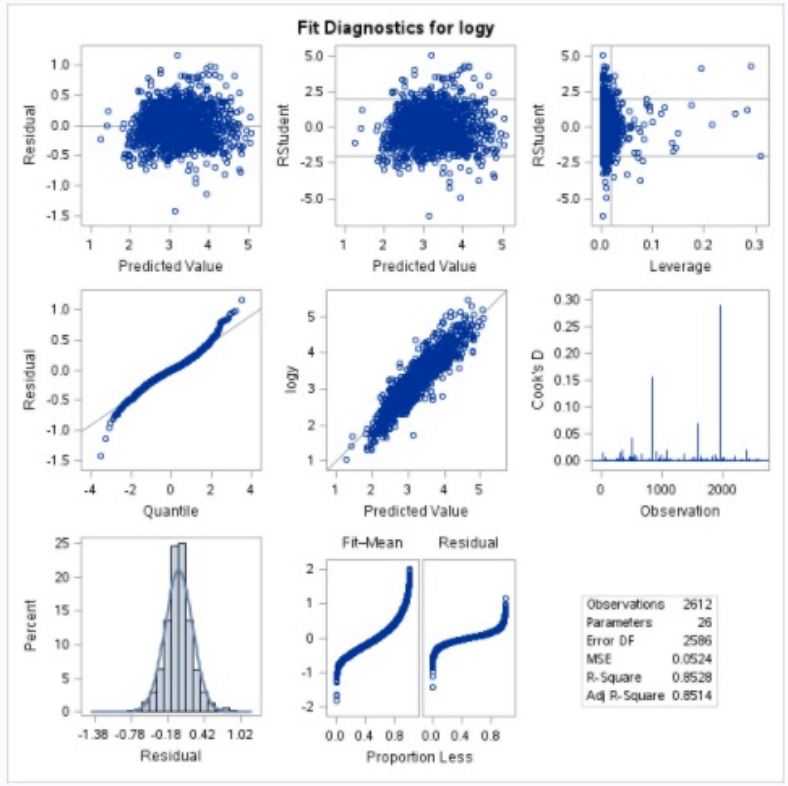


Figure 16: Overall fits of the final model.

4 How this model can be used

This model is very useful in the real world. My initial goal was to create some useful model, that can be used and applied in the real world applications. This study also important for Kazakhstan, as far as I know, there are not so many such as projects that predicts the price of the apartments, and there are no publicly available datasets to play around with. Suppose, someone wants to buy some apartment in Astana. Then, he/she might thought about his/her desired flat: flat situated in the Esil district, with 70 square meters area, with some furniture. This model predicts the price of this apartment. Some attributes, such as number of comments and number of views has lower coefficients and they might not be taken into consideration.

Also, if someone wants to sell an apartment in Astana, then by putting the attributes into this model, one can estimate the price of the flat.

Of course, purposed model is not 100% accurate. But, this model might give some kind of measurement to reduce the amount of manual work for people who consider to buy or to sell an apartment. Also to compare prices on the market.

5 Conclusion

At the beginning, some attributes were considered as a significant predictors. But, regression analysis shows that this attributes are less important compared to others. Also interaction terms are useful for some parameters. For instance, I thought that the residential area can be a good predictor for the house price. Because, when I look at some example datasets, I saw that the residential areas differ from the total area, as many apartments might have kitchen, hallways, balcony. Even, treated for the missing values (additional work). But this attribute was highly correlated with the total square of the flat and was useless attribute in overall. Also, the attributes, such as the number of rooms, the floor number, "a kitchen-studio" information were excluded from the final best model. Overall, predicting the house price is not easy task. We might consider a lot features in order to estimate the house price correctly. I learned that, choosing the right parameters are better than choosing the large amount of useless features. Also, was surprised that

simple linear regression might work well in these kind of complicated tasks. Finally, using the text processing, many useful features (attributes) can be extracted.