Prediction of House Sales Price in Turkey

Cumhuriye Tüluğ Küçüköğüt

Özge Akat

***Abstract*— This article reports a case study evaluating the impact of different metrics on home prices using different types of machine learning algorithms.**

Today, when those who want to buy a new house are more conservative about their budgets and market strategies, our aim is to find out what factors the house prices depend on by estimating the house prices with different features. For this reason we use Zingat house selling data that has 199308 rows. We collected 18 attributes. Linear Regression and Random Forest Regression are used for prediction, analysis provided by Mean Errors, R2 Score, Mean Errors that are mean squared error,mean absolute error, root mean squared error.

The functioning of this project involves some good amount of dataset on which prediction can be done. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

***Keywords*—housepricing,prediction,machine learning,statistics**

I. INTRODUCTION

It is difficult to calculate manually and estimate the factors affecting house prices. Customers who depend on real estate agents are defrauded because the real estate agent can give a much higher price than the real price. People who have the budget to buy a house cannot buy it due to the difference in the prices given to them by the agency. Manual information is confusing as the data can vary from person to person. Linear Regression and Random Forest Regression are applied to eliminate such problems, to estimate the dependent value from the independent values and to provide efficient outputs.

The most important purpose of the study is to prevent buyers from wasting time and to prevent exaggerated prices. For this reason we used Zingat house selling information which includes price,building age,floor count etc attributes.Zingat is a real estate site.

In the second part of the report, models that were used in our project are explained. Section 3 covers experimental data,observation, separation,modelling and selection of features. In the next section the results are shown with algorithms and feature sets. Finally conclusion and the future work are explained in Section 6.

II. BACKGROUND

For forecasting the price of houses regression algorithms were used because regression is a machine learning technique in which the model predicts its output as a numerical value.[1]

A.Linear regression

This algorithm is a basic and commonly used type of predictive analysis. Regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. Three major uses for regression analysis are: determining the strength of predictors, forecasting an effect., trend forecasting.[2]

B. Random Forest

Random Forest is a "Tree" based algorithm that uses attribute properties of multiple decision trees to make decisions. Random Forest Regression is a technique that combines predictions from multiple machine learning algorithms to generate more predictions from a model.[3]

We choose supervised machine learning algorithms because these algorithms receive a set of labeled data as training and make predictions for all unknown points. This method is generally used with classification, regression, and ranking problems. [4]

Linear Regression and Random Forest Regression are used in this project because both of them can be used for predicting the dependent value according to independent variables in data.In our project, start date,end date, heating type etc are independent variables and price is dependent variable.

According to articles, Random Forest Regression give better results in large datasets and missing data by creating estimates for them.[5]

III. EXPERIMENTAL DETAILS AND DESIGN

3.1. Environment Setup

Our project consists of text data and 18 different attributes. Attributes are divided into categorical and numerical. Numerical attributes are start date,end date,building age,floor no, room count,size and price. Categorical attributes are type, sub type, listing type,student available,mortgage available, address,furnished and currency.

Our data set consists of independent and dependent variables, and our aim is to estimate the price that is dependent variable according to independent variables for this reason regression methods are suitable for this. As explained before Random Forest Generator is mostly used for unknown and very large datasets, we think it be more effective in our training data.

3.2. Data Description and Collection

Our dataset comes from a zingat.com competition. Our dataset contains house sale prices and house features. There are 18 total attributes in the dataset and 199308 instances. Attributes are type,sub_type,start_date,end_date,listing_type,building_age,total_floor_count,room_count,student_avaliable,mortgage_available,address,furnished,currency,floor_no,price,heating_type. Type is type of property which is Konut. Sub_type is a group of property types which is Daire. Start_date is the date when listing starts to be active on the market. End_date is the date when the listing is not active on the market anymore. Listing_type is the type of listing which are Satılık, Kiralık. Building_age is building age which are 0,1,2,3,11-15 arası,40 üzeri etc. total_floor_count is the count of total floors in the building. Room_count is the count of rooms in the flat for example 2+1. Size is house size, unit m2. Student_ available is availability for students which are Evet,Hayır. Mortgage_ available is availability for mortgages which are Evet,Hayır. Address is the address of the housing city/county/district. Furnished is whether the house is furnished Eşyalı, Eşyasız, Sadece Beyaz Eşya, Sadece Mutfak. Currency is house price currency which are TRY,USD,EUR,GBP. Floor_no is the floor number information of the given listing which are yüksek giriş, giriş kat, zemin kat,çatı katı, bahçe katı. Price is house price. Heating_type is different types of heating systems which are Kalorifer (Doğalgaz), Kalorifer (Kömür), Kombi (Elektrikli), Klima, Kombi (Doğalgaz), Merkezi Sistem, Merkezi Sistem (Isı Payı Ölçer), Yerden Isıtma, Soba (Kömür), Soba (Doğalgaz), Güneş Enerjisi = solar energy, Jeotermal, Kat Kaloriferi, Kalorifer (Akaryakıt), Yok.Table 1 shows features detailly.

Attribute	Type
listing_id	discrete
type	categorical
sub_type	categorical
start_date	continous
end_date	continous
listing_type	categorical
building_age	discrete
total_floor_count	discrete
room_count	discrete
student_avaliable	categorical
mortgage_avaliable	categorical
address	-
furnished	categorical
currency	categorical
floor_no	discrete
price	continous
size	continous
heating_type	categorical

Table 1:Table of attributes with their types.

3.3.Data prepation and Cleaning

This part explaines the data preperation.We started our project with reading csv file and observing this.Figure 1 shows the example part of dataset.

	listing_id	type	sub_type	start_date	end_date	listing_type	building_age	total_floor_count	floor_no	room_count	size	student_avaliable	mc
0	1.0	Konut	Daire	1/3/19	1/3/19	satılık	3	4	Kot 3	3+1	130.0	NaN	
1	2.0	Konut	Daire	1/2/19	1/2/19	satılık	0	5	1	4+1	175.0	NaN	
2	3.0	Konut	Daire	1/2/19	1/2/19	satılık	0	3	1	3+1	125.0	NaN	
3	4.0	Konut	Daire	1/2/19	1/2/19	satılık	0	10	7	2+1	72.0	NaN	
4	5.0	Konut	Daire	1/2/19	1/2/19	satılık	0	4	3	1+1	75.0	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
199303	199304.0	Konut	Daire	2/9/18	11/18/18	satılık	0	5	3	3+1	130.0	NaN	
199304	199305.0	Konut	Daire	2/9/18	12/3/18	satılık	6-10 arası	3	Kot 3	3+1	240.0	NaN	

Figure 1

Figure 2, 3, 4 and 5 shows some histogram graphs of data.

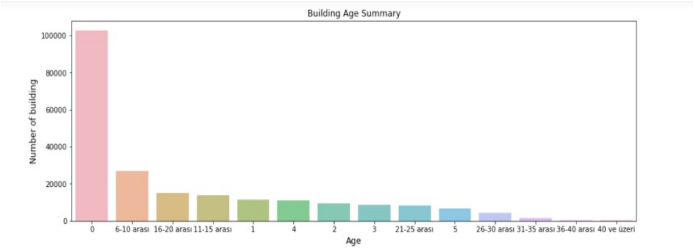

Figure 2- Building Age Summary

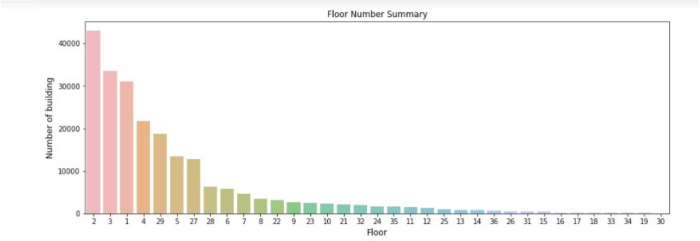

Figure 3-Floor Count Summary

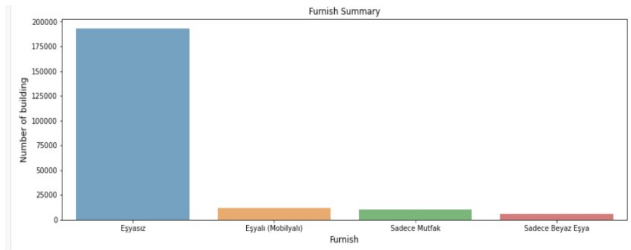

Figure 4-Furnished Summary

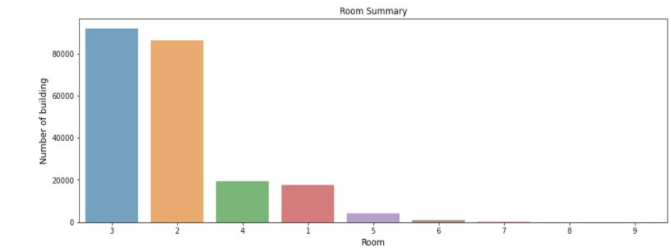

Figure 5 – Room Count Summary

House selling data consists null values firstly these were eliminated. Figure 6 shows null value counts according to columns.

```
listing_id              32
type                    32
sub_type                32
start_date              32
end_date                32
listing_type            32
building_age            32
total_floor_count       32
floor_no                32
room_count              32
size                   190
student_avaliable    199308
mortgage_avaliable   199308
address                162
furnished               32
heating_type            32
price                   53
currency                53
dtype: int64
```

Figure 6

After observation of null values,firstly student_available and mortgage_available columns were dropped because most of rows are null. Secondly rows were cleaned which have null values so the cleanup of null values is complete.

Secondly, listing type was removed because listing type is increasing order and all of them unique. Meanwhile type,subtype were removed because type's all rows are Satılık and subtype's all rows are Konut.

After cleanup and columns disappear, we converted some columns. For example currency has 4 different categorical data that are TRY,EUR,Dolar and Sterlin.The current Turkish Lira values of these currencies were taken and multiplied by the price. In this way, all values were created in Turkish Lira and currency column was dropped. The house price dataset includes columns named start and end days. First of all, we found the differences between these days in terms of days, we created a new column named differenceOfDates and wrote them in this column. After this process, we dropped them because the start day and end day columns were no longer needed.

Address column was split into city,province and neighbourhood. Province and neighbourhood was dropped because of unique items.City names were converted to licence plate codes.Room counts include 3+1 or 2+1, this columns was splited into room count and living room count.For 3+1, 3

is room count and 1 is living room count.Some building age total floor count have string values.For example 15-20 arası, last number was assigned for them.Heating type and furnished columns all categorical data they were encoded and new columns created like furnished_eşyalı, furnished_eşyasız,furnished_beyaz_eşya,heating_type_merkezi_sistem,heating_type_kombi etc. All duplicated values were eliminated and graphs were created. Some of them are shown in Figure 7,8,9 and 10.

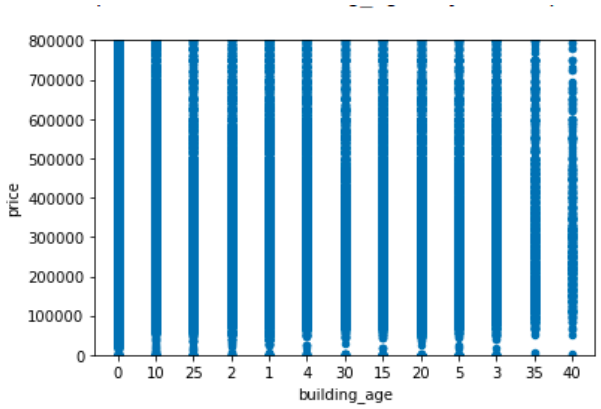

Figure 7-Price according to building age

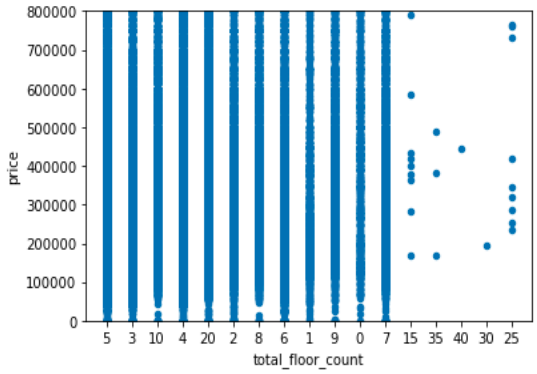

Figure 8-Price according to total floor count

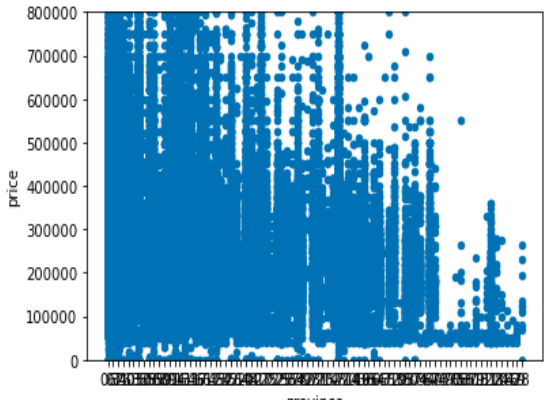

Figure 9-Price according to city

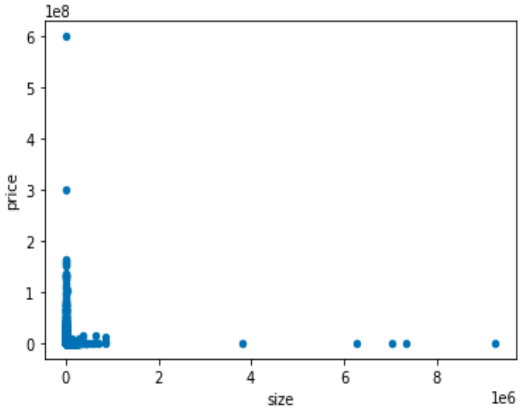

Figure 10-Price according to size

Graphs observed and some outliers were cleaned up. For example, size values greater than 550 were cleared. Mean is also considered for this transaction, price values equal to 0 were also cleared.Size graph became like in Figure 11.

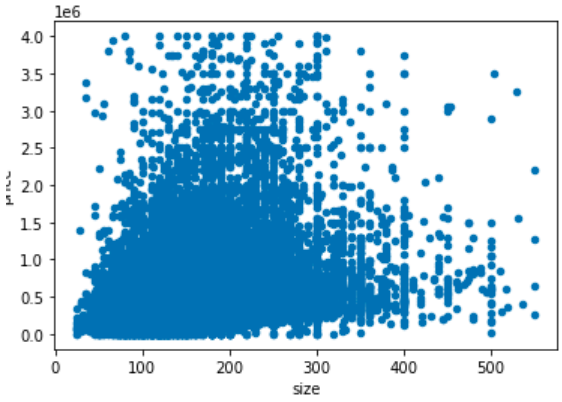

Figure 11-Price according to size after cleaned up

When all these processes were finished,our dataset shape is 193518 rows and 29 columns now.

3.4.Feature Selection

Features were selecting according to graphs, graphs were considered and our first feature was created with *"size","province","furnished_Eşyalı (Mobilyalı)"*

Secondly for choosing feture set, SelectKBest method was used this method selects the features according to the k highest score.

By changing the score_func' parameter we can apply the method for both regression data. We defined the model by using SelectKBest class. For regression, we set the 'f_regression' method as a scoring function. The target number of features to select is 3

"size","roomCount","total_floor_count"

#Finally feature set was selected via Pearson Correlation. Correlation is a measure of the linear relationship of 2 or more variables. If two variables are correlated, we can predict one from the other. Variables should be correlated with the target. Heating map was shown in Figure 12

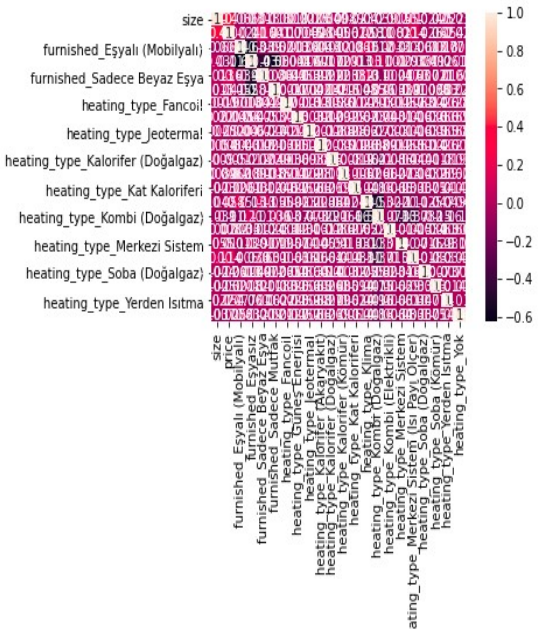

Figure 12-Heating map

Correlation results according to price are shown in Figure 12.

```
size                                          0.402853
price                                         1.000000
furnished_Eşyalı (Mobilyalı)                  0.002369
furnished_Eşyasız                             0.106970
furnished_Sadece Beyaz Eşya                   0.164018
furnished_Sadece Mutfak                       0.041434
heating_type_Fancoil                          0.016648
heating_type_Güneş Enerjisi                   0.000451
heating_type_Jeotermal                        0.031646
heating_type_Kalorifer (Akaryakıt)            0.008422
heating_type_Kalorifer (Doğalgaz)             0.004987
heating_type_Kalorifer (Kömür)                0.006192
heating_type_Kat Kaloriferi                   0.012057
heating_type_Klima                            0.057697
heating_type_Kombi (Doğalgaz)                 0.113474
heating_type_Kombi (Elektrikli)               0.008182
heating_type_Merkezi Sistem                   0.011163
heating_type_Merkezi Sistem (Isı Payı Ölçer)  0.142877
heating_type_Soba (Doğalgaz)                  0.026401
heating_type_Soba (Kömür)                     0.061719
heating_type_Yerden Isıtma                    0.054386
heating_type_Yok                              0.024514
Name: price, dtype: float64
```

Figure 12. Correlation between features and price

Correlation values were observed for feature selection and different feature combinations were put into the linear regression algorithm and the feature combination with the highest score was selected.That is *'size', 'furnished_Eşyalı (Mobilyalı)', 'furnished_Eşyasız', 'furnished_Sadece Beyaz Eşya', 'furnished_Sadece Mutfak', 'heating_type_Fancoil', 'heating_type_Jeotermal', 'heating_type_Kalorifer (Akaryakıt)', 'heating_type_Kalorifer (Doğalgaz)', 'heating_type_Kalorifer (Kömür)', 'heating_type_Kat Kaloriferi', 'heating_type_Klima', 'heating_type_Kombi (Doğalgaz)', 'heating_type_Kombi (Elektrikli)', 'heating_type_Merkezi Sistem', 'heating_type_Merkezi Sistem (Isı Payı Ölçer)', 'heating_type_Soba (Doğalgaz)', 'heating_type_Soba (Kömür)', 'heating_type_Yerden Isıtma', 'heating_type_Yok'*

3.5.Train- Test Split

Our target (Y) value is price and the other feature combinations which are explained in 3.4 Feature Selection part are our data(X).For seperation sklearn.model_selection's train_test_split method was used. 30% of X and Y reverted to test set and 70% to train set.Figure13 shows X_train data for Feature set of size,province and furnished_Eşyalı (Mobilyalı),Figure 14 is X_test and Figure 15 is Y_train.

```
        size province  furnished_Eşyalı (Mobilyalı)
143765  90.0       10                             0
3575    90.0       34                             0
118487  130.0      45                             0
51553   145.0      07                             0
126532  110.0      06                             0
...       ...     ...                           ...
157145  140.0      06                             0
182425  140.0      35                             0
121884   55.0      34                             0
179108   60.0      09                             0
45177   180.0      55                             0

[135462 rows x 3 columns]
```

Figure 13-X_train

```
        size province  furnished_Eşyalı (Mobilyalı)
166460  185.0      01                             0
51517   105.0      50                             0
121983   55.0      35                             0
106190  140.0      59                             0
21864    85.0      34                             0
...       ...     ...                           ...
52273   150.0      20                             0
106223   60.0      07                             0
116819  105.0      06                             0
83650   160.0      07                             0
44493   145.0      59                             0

[58056 rows x 3 columns]
```

Figure 14-X_test

```
143765    120000.0
3575      127000.0
118487    200000.0
51553     275000.0
126532    205000.0
             ...
157145    169000.0
182425    390000.0
121884    165000.0
179108    159000.0
45177     295000.0
Name: price, Length: 135462, dtype: float64
```

Figure 15-Y_train

After dataset seperation, StandardScalar was used for Linear-Regression Algorithm.

The standardization is about feature scaling. It is useful to speed up the learning algorithm. Indeed, when we have several variables with different scales like size has a bigger range and it is likely that it will have a bigger impact on the result because of distance measures used in many learning algorithms. We used StandardScaler for standardization.

```
from sklearn.preprocessing import StandardScaler

myScaler = StandardScaler()

scaled_x_train = myScaler.fit_transform(X_train)

scaled_x_test= myScaler.transform(X_test)
```

Figure 16. StandardScaler implementation

When all these processes were finished, the models were created with Linear Regression and Forrest RandomForestRegressor. Linear regression helps understand which variables are significant and which not. Linear regression is a good approach to use as a starting step.[6] Random forest has been used to rank the importance of variables in a regression problem in a natural way. In a regression tree, for each independent variable, the data is split at several split points.[7]

IV. RESULT

1) According to the graphs features are selected firstly size,province and furnished

	LinearRegression	RandomForestRegressor
Mean-squared error	56389749408.525444	43606229311.77778
Mean-absolute error	127311.56235993443	102174.03752662153
Root-mean-squared error	237465.25937181935	208821.0461418527
r2 score	0.16262149662876868	0.35245466735061526

2) Features are selected according to the k highest scores

	LinearRegression	RandomForestRegressor
Mean-squared error	56044105191.034096	158346030866.16476
Mean-absolute error	126774.69682954082	344925.5801221703
Root-mean-squared error	236736.3622070638	397927.1677909976
r2 score	0.1677542564047545	1.3514125125981944

3) Dataset is created according to correlation

	LinearRegression	RandomForestRegressor
Mean-squared error	53395264149.32889	53143704526.03356
Mean-absolute error	124226.78242094706	120522.80843057459
Root-mean-squared error	231074.15292353425	230529.18367537236
r2 score	0.20708911017582032	0.21082472920362583

V. Conclusion and Future Work

We decided on the evaluation approach using different combinations of features according to algorithms and correlation. We compared these feature combinations with mean errors and r2 scores. According to MAE, MSE and RMSE information, Random Forest Regressor gives better results and the results are primarily that the selected size,province and furnished feature set is better. When we compare the algorithms, we mostly see that the results of the Random Forest Regressor algorithm are smaller, we can interpret that this algorithm is better in line with these results.

This work uses regression techniques to develop a price prediction model in house pricing problems. Looking at the results, we can say that our values are not linearly related to the dependent value.

Also this study can be enlarged with separating dataset according to cities and we can observe prices according to neighbourhoods.

VI. References

[1] 5 Regression Algorithms you should know – Introductory Guide! ,Author Gaurav Sharma, Published date: May 26, 2021, Access Link: https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/

[2] What is Linear Regression? Access Link: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

[3] Random Forest Regression, Author Chaya Bakshi, Published date:Jun 9, 2020, Access Link: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

[4] Foundations of Machine Learning By Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar

[5] Random Forest Regression: When Does It Fail and Why?, Author Derrick Mwiti, Published date: Updated October 26th, 2021, Access Link: https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why

[6] Understanding the Fundamentals of Linear Regression,Author Peter Grant, Published date:August 5,2019,Access Link:https://towardsdatascience.com/understanding-the-fundamentals-of-linear-regression-7e64afd614e1

[7] 4 Simple Ways to Split a Decision Tree in Machine Learning, Author Abhishek Sharma, Published date:June 30, 2020, Access Link: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/