Table of Contents

Introduction

The rapid increase in population in Turkey and the effects of the pandemic have a significant influence on the real estate market. The supply and demand gap in the real estate market is increasing fast and this situation causes an increase in property prices that is difficult to predict. Especially, in big cities, this problem is nonnegligible. İstanbul and İzmir are the two of the biggest cities in Turkey and the fact that the population in these cities is increasing day by day increases the demand for properties.

Zingat, one of the biggest real estates in Turkey, shared a set of data in Kaggle (https://www.kaggle.com/zingatbi/zingat-real-estate) about 3 months ago which includes property prices and some variables given by sellers. The purpose of this project is to predict the future sales price of properties in these with the help of these variables.

Data Description

Data represents information of properties from January 2019 to April 2020. The dataset contains 19 inputs, 1 target variable, and 64573 rows. Some columns consist of Turkish words and missing values are labeled as "-". The description of the columns is given below.

Columns of the Dataset

1. date: the date when listing is started to active on the website
2. price: the price of the property
3. path: address column shows the city/district/neighbourhood of the listing.
4. Property_type: the type of property.
5. Building_age: age of the building
6. Totalfloorcount: the total number of floors
7. Room_count: the number of rooms (i.e. 3+1 represents room count + living room count)
8. grossm^2: the gross size.
9. netm^2: the net size.
10. floor_no: column shows the floor info.
    - Yüksek giriş: High Entrance
    - Giriş kat: Entrance Floor
    - Kot: Basement
    - Zemin kat: Ground floor
    - Çatı Katı: Penthouse apartment
    - Bahçe katı: Garden floor flat
11. heating_type:
    - Kalorifer (Doğalgaz): central heating/natural gas,
    - Kalorifer (Kömür): central heating/coal,

- Kombi (Elektrikli): combi boiler
- Klima: air-conditioning
- Kombi (Doğalgaz): gas-combi boiler
- Merkezi Sistem: central heating
- Merkezi Sistem (Isı Payı Ölçer): central heating
- Yerden Isıtma: floor heating
- Soba (Kömür): heating stove/coal,
- Soba (Doğalgaz): heating stove/natural gas
- Güneş Enerjisi: solar energy,
- Jeotermal: geothermal energy
- Fancoil: a type of air conditioner
- Kat Kaloriferi: central heating
- Kalorifer (Akaryakıt): central heating/fuel oil
- Yok: None

12. bath_count: the number of baths.
13. landscape: the landscape type (i.e. sea, street, nature or lake)
14. car_park: the park type. Open or underground with paid/free info.
15. Intercom: whether there is an intercom. (Var: yes, Yok: no)
16. Earthquake_reg: suitable to earthquake regulations.
17. Elevator: whether there is an elevator. (Var: yes, Yok: no)
18. children_playground: whether there is a playground for children. (Var: yes, Yok: no)
19. dressing_room: whether there is a room for dressing. (Var: yes, Yok: no)
20. parents_bathroom: whether there is a bath for parent. (Var: yes, Yok: no)

Data Preprocessing

Preprocessing of Input Columns

1. Path

Path column consists of three information such as city, distinct and neighbourhood. I separated them into three different columns to encode them easily.

2. Room Count

There are some missing or irrational values (i.e. 1149+0), a total of 51. I decided to fill these values with the average room number of property types. The room count column consists of two information such as the number of rooms and number of living rooms. I first separated them into two columns. Then, I calculated the median of these two new columns in terms of property types (I preferred to use the median to decrease the outlier effect). Finally, I filled these values with these medians. Additionally, I added one more column named total room as the sum of these two columns.

3. Gross and Net m²

There are some missing or irrational values in these columns. While some properties' gross and net m^2 are too small, some values are extremely high (Figure 1 and 2). Furthermore, for some properties, net m^2 is bigger than gross m^2 which is impossible in real life.

count	64522.000000
mean	185.825160
std	2672.887722
min	1.000000
25%	100.000000
50%	125.000000
75%	155.000000
max	370000.000000

Figure 1 Statistics for the column grossm2

count	64522.000000
mean	127.800719
std	765.843472
min	1.000000
25%	90.000000
50%	110.000000
75%	135.000000
max	145000.000000

Figure 2 Statistics for the column netm2

First of all, I eliminated the columns which have lower than 5 in both gross and net m^2 (4 rows). After that, I equalized the gross and net m2 in the cases where the net is bigger than the gross. After these corrections, as there is no extreme low value in gross m^2, there are still some extreme low values in net m^2 as shown in Figure 3. I equalized the net m^2 of these properties to their gross m^2.

	date	path	price	room_count	grossm²	netm²
6931	2/23/2019	İzmir/Seferihisar/Payamlı	330000 TRY	3+1	130	1
19798	6/1/2019	İzmir/Menderes/Çile	280000 TRY	2+1	1800	1
326	1/3/2019	İstanbul/Beylikdüzü/Kavaklı	445000 TRY	4+1	10	1
38130	10/25/2019	İstanbul/Bağcılar/15 Temmuz	900000 TRY	10+0	4400	1
11840	4/2/2019	İstanbul/Beşiktaş/Vişnezade	132500000 TRY	9+2	1300	1

Figure 3 First 5 columns of Minimum Net m2

To handle extremely high values in the data, I decided to apply a capping to these values. I could get rank all the data together and cap extreme values with the same value but instead of doing this, I preferred to observe the net and gross m2 of different property types. The reason is that the property type is one of the important features for the size of the house because it can be assumed that on average the size of an apartment is smaller than the size of a villa. When we look at the data to see that difference (Figure 4), we can observe that the mean size of the villa is almost twice the mean size of the apartment (Daire in Turkish). Thus, we need to cap extreme values by considering the types of property.

property_type	count	mean	std	min	25%	50%	75%	max
Daire	56262.0	117.359141	809.947414	7.0	85.00	103.0	130.00	145000.0
Köşk / Konak / Yalı	48.0	488.625000	474.777199	90.0	230.75	400.0	562.25	3000.0
Müstakil Ev	1433.0	204.704815	310.405898	7.0	100.00	150.0	220.00	5999.0
Prefabrik Ev	181.0	41.265193	51.217362	8.0	20.00	20.0	45.00	500.0
Rezidans	1106.0	164.154611	147.537208	23.0	70.00	120.0	200.00	1300.0
Villa	4321.0	222.740801	246.156048	39.0	134.00	170.0	250.00	7195.0
Yalı Dairesi	72.0	156.069444	70.793953	80.0	105.00	137.5	180.00	500.0
Yazlık	1053.0	125.313390	63.491329	14.0	100.00	120.0	140.00	1200.0
Çiftlik Evi	93.0	758.150538	2291.311097	22.0	100.00	175.0	300.00	18000.0

Figure 4 Statistics of Net Area of Property Type

After capping, I added two columns to the data: area_diff as the difference between gross and net area and ave_area as the average of gross and net area (because as seen in Figure 3, some gross and net areas are not supporting each other. To add this column may provide a more accurate area size. I will observe this in the feature importance part).

4. *Building age*

In this column, there are many categories which are not organized properly. I grouped after 30 years because of low frequency and first years (1 to 5).

0	28088
6-10 arası	5748
16-20 arası	3808
1	3523
21-25 arası	3242
11-15 arası	3211
26-30 arası	3133
2	2795
4	2598
-	2450
3	2406
31-35 arası	1731
5	1309
36-40 arası	298
40 ve üzeri	229

0	28088
1-5 arası	12631
6-10 arası	5748
16-20 arası	3808
21-25 arası	3242
11-15 arası	3211
26-30 arası	3133
Unknown_building_age	2450
30+	2258

Figure 5 Previous and Current Categories of Building Age

5. ***Floor No***

The column of Floor number also contains many unorganized categories and some columns express similar things. For example, both “Kot” and “Bodrum Katı” indicate the basement. After grouping, the category number decreased significantly.

2	10587
3	8588
1	6064
4	5871
-	4502
5	3823
Müstakil	3765
Bahçe katı	3517
Yüksek Giriş	3324
6	2043
7	1749
Giriş Katı	1735
8	1379
9	1131
10	885
Komple	698
11	652
Kot 1	600
Zemin Kat	561
Kot 2	484
12	429
Çatı Katı	348
13	337
Kot 3	280
14	267
20 ve üzeri	213
Kot 4	175
15	134
En Üst Kat	95
16	87
17	83
Teras Kat	52
18	47
Bodrum Kat	41
19	23

Ara Kat	44179
Giriş Katı	9137
Unknown_floor_no	4502
Müstakil	4463
Bodrum Kat	1580
En Üst Kat	495
20 ve üzeri	213

Figure 6 Previous and Current Categories of Floor No

6. ***Detached House***

Some categories in property types and floor numbers specify that the property is a detached house (i.e. villa in property type). I created a new column by using these categories.

7. ***Total Floor Count***

I applied similar prosses to this column as the building age column.

8. ***Bath Count***

I decided to define this column as ordered. So, I converted the category of “6 ve üzeri” (above 6) to 6. Since the frequency of the missing value is low, I filled them with the mode of the column.

1	35473	1	35600
2	23099	2	23099
3	3914	3	3914
4	1099	4	1099
5	333	5	333
0	318	0	318
6 ve üzeri	206	6	206
-	127		

Figure 7 Previous and Current Categories of Bath Count

9. ***Dropped Columns***
   I dropped some columns such as heating type, landspace, and carpark because of hig frequency of missing data and/or dirtiness of the column.

Target Column

As aforementioned, the data is given for a period of one and a half years. In Turkey, inflation increase is unstable, and it naturally affects the price of properties. Thus, we should handle this effect, otherwise, we may estimate prices incorrectly.

To do this, I decided to determine a base date to calculate all prices according to that base and chose the date where the data ends. I got values of CPI per month from the TCMB (The Central Bank of the Republic of Turkey) website and calculated the coefficients of each date to convert that date to the end date. After I multiplied the coefficients with prices, I obtained prices in terms of the end date.

Separation of the Dataset

As mentioned before, the data consists of two cities as İstanbul and İzmir. Since these cities have different characteristics, I decided to separate the data into two parts as data of İstanbul and İzmir. It also makes it easy to apply models in terms of computational power. While there are 33177 rows in İstanbul data, this number is 31392 in İzmir.

After separating the dataset, I examined the statistics of prices in both İstanbul and İzmir (Figure 8) and realized that there are some extreme values in both upper and lower bonds in prices. Thus, I decided to eliminate the 0.1% percentile from both the above and below datasets.

count	33177
mean	1357880
std	6707580
min	252
25%	234702
50%	417297
75%	934002
max	582141304

count	31392
mean	671490
std	1025116
min	173
25%	273948
50%	398870
75%	675461
max	35297500

Figure 8 Statistics of price in İstanbul (left) and İzmir (right)

Modelling

For both datasets, I applied three different machine learning models such as Decision Tree, Random Forest and Xgboost and evaluated models with the three metrics as r-squared, mean absolute percentage error (MAPE) and mean absolute error (MAE). After choosing the model which has better metrics, I applied parameter tuning.

Case of İstanbul

Model Selection and Parameter Tuning

The results show that Random Forest is a better model for each metric. So, I decided to continue with Random Forest.

	R^2	MAPE	MAE
Decision Tree	0,75	0,33	427239,04
Random Forest	0,85	0,27	338291,12
Xgboost	0,84	0,33	362700,27

I tuned six parameters and best parameters are {'n_estimators': 50, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 100, 'bootstrap': False}.

	R^2	MAPE	MAE
Random Forest	0,85	0,27	338291,12
Random Forest (tuned)	0,86	0,27	321406,89

Feature Importance

The first ten most important features in tuned Random Forest are given below. It shows that size, distinct, room numbers of properties are plying significant roles to estimate price properly in İstanbul.

```
netm²                          0.153101
ave_area                       0.144876
distinct_Beşiktaş              0.128624
grossm²                        0.120568
bath_count                     0.054302
num_room                       0.053577
area_diff                      0.048006
total_room                     0.031331
property_type_Daire            0.021420
distinct_Beylikdüzü            0.019148
```

Case of İzmir

Model Selection and Parameter Tuning

The results show that while Random Forest gives better estimations in terms of R^2 and MAE, Xgboost is better for MAPE. Thus, I decided to choose Random Forest.

	R^2	MAPE	MAE
Decision Tree	0,30	0,44	241505,02
Random Forest	0,68	0,37	184578,52
Xgboost	0,63	0,33	193858,60

I again tuned the same parameters and best parameters are {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 80, 'bootstrap': True}.

	R^2	MAPE	MAE
Random Forest	0,68	0,37	184578,52
Random Forest (tuned)	0,71	0,37	183146,36

Feature Importance

The first ten most important features in tuned Random Forest are given below. It seems that important features are similar to the case of İstanbul. However, being a detached house is a significant feature in İzmir different from İstanbul.

```
ave_area              0.107094
netm²                 0.097800
bath_count            0.094046
grossm²               0.090203
total_room            0.077104
area_diff             0.063916
distinct_Çeşme        0.059656
num_room              0.048784
detached_house        0.028820
property_type_Daire   0.028360
```

Conclusion

In this project, I studied applying a proper model to predict the price of properties given by Zingat. The metrics are promising (R2s of the models are above 0,7 and MAPE scores are about 0,30) and important features to predict the price are very similar. Furthermore, popular districts in the cities (Beşiktaş in İstanbul and Çeşme in İzmir) are playing significant roles in the prediction which is meaningful. Unlike the case of İstanbul, whether the property is detached is also an important feature for the case of İzmir.