Prediction of real estate prices in Istanbul by web scraping, machine learning and regression techniques

Final Report

Cem Kıraç (0055081)

Ali Duymaz (0080797)

Hande Kastan (0065740)

Introduction

Although all three of us are working as a member of different data analytics and data science departments in QNB Finansbank of Turkey, we wanted to test our capabilities in a different field other than banking/finance. There is a real estate assessment process within the bank but it's done by experts only. When people are using mortgage loans the real value of the real estate is an important concept in banking industry because it is considered as a collateral. However as mentioned before we did not start this project for our company as a work goal. It was a special field of interest for us.

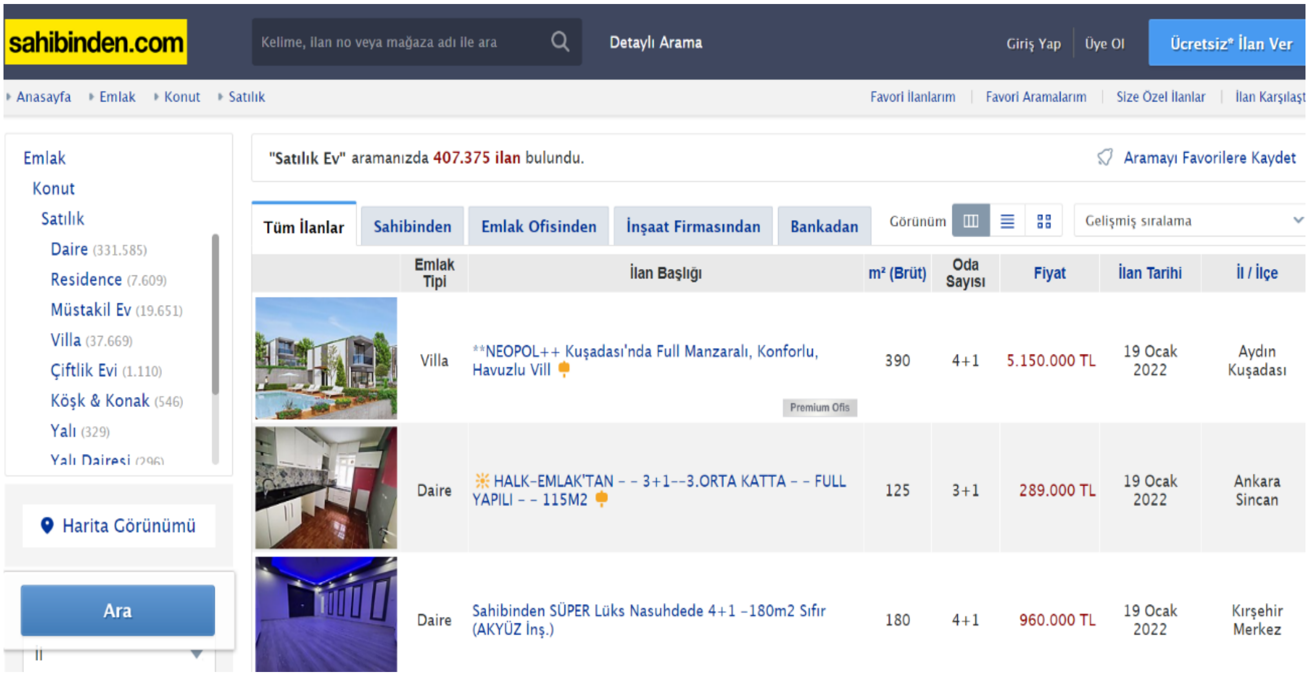
Problem Definition

The real estate industry is huge in Turkey and the grand share of that industry is Istanbul. Turkish people really think buying houses is the best way of investing your money. Especially the mid-aged and older generation closely monitors house prices but when buying or selling a house its really difficult to know the exact value of the house. The real estate agencies determine the buyer-seller prices most of the time. Usually the district/neighborhood, size, closeness to transportation are the main factors which effect the prices.

In this project, we are aiming to learn which factors has the most impactful on the house prices. So, we use the house prices as our continuous target and build supervised learning models. Besides of the predicting the target, we also aim to improve the model output, in order to do that we make some preprocessing on the row data, split the data of %60 as train and test the prediction score on test data (%40) and evaluate different algorithms to challenge the best fit for this prediction.

About Data

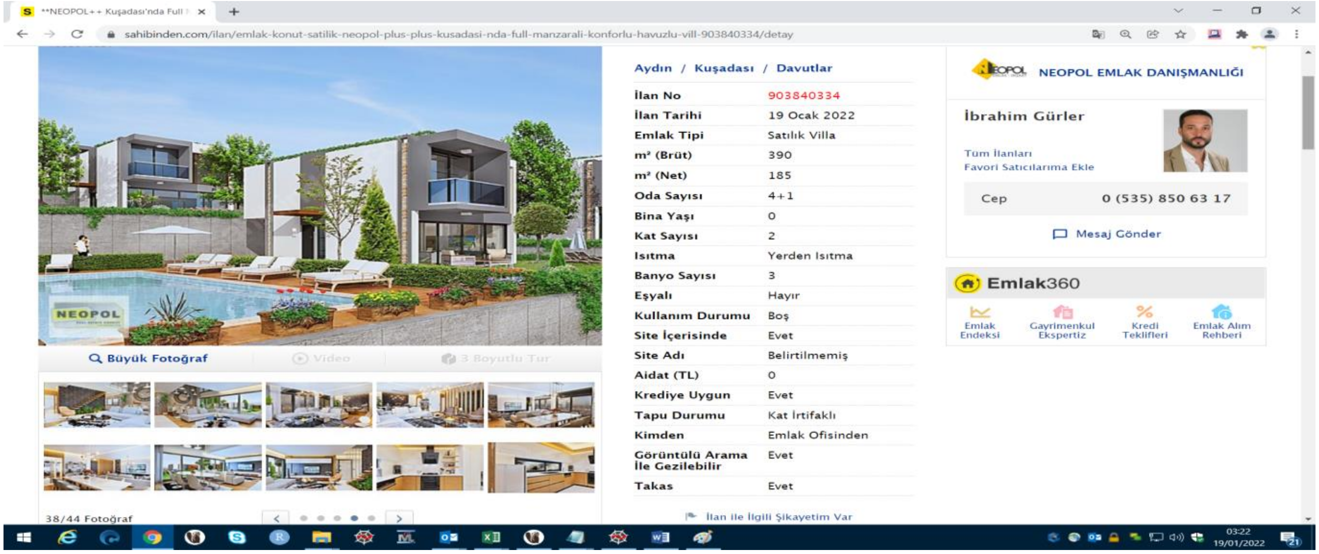
Collection of the data was absolutely the most challenging part. There is no bulk data available for this kind of project which is publicly available. We thought we can do some web-scraping from the largest real-estate ads website of Turkey which is sahibinden.com. There are hundreds of thousands of ads from all cities of Turkey. When you click on flats for sale, the page below appears. This is actually a summary page.



"Satılık Ev" aramanızda 407.375 ilan bulundu.

	Emlak Tipi	İlan Başlığı	m² (Brüt)	Oda Sayısı	Fiyat	İlan Tarihi	İl / İlçe
	Villa	**NEOPOL++ Kuşadası'nda Full Manzaralı, Konforlu, Havuzlu Vill	390	4+1	5.150.000 TL	19 Ocak 2022	Aydın Kuşadası
	Daire	HALK-EMLAK'TAN - - 3+1--3.ORTA KATTA - - FULL YAPILI - - 115M2	125	3+1	289.000 TL	19 Ocak 2022	Ankara Sincan
	Daire	Sahibinden SÜPER Lüks Nasuhdede 4+1 -180m2 Sıfır (AKYÜZ İnş.)	180	4+1	960.000 TL	19 Ocak 2022	Kırşehir Merkez

When you enter one sale ad a detailed page appears and you can find the detailed features of just 1 single flat/house. You can see sample page below.



sahibinden.com/ilan/emlak-konut-satilik-neopol-plus-plus-kusadasi-nda-full-manzarali-konforlu-havuzlu-vill-903840334/detay

Aydın / Kuşadası / Davutlar

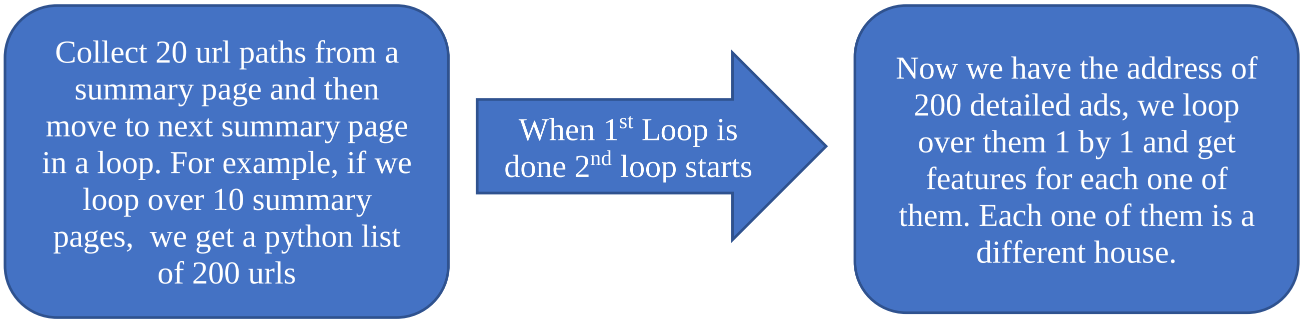
İlan No	903840334
İlan Tarihi	19 Ocak 2022
Emlak Tipi	Satılık Villa
m² (Brüt)	390
m² (Net)	185
Oda Sayısı	4+1
Bina Yaşı	0
Kat Sayısı	2
Isıtma	Yerden Isıtma
Banyo Sayısı	3
Eşyalı	Hayır
Kullanım Durumu	Boş
Site İçerisinde	Evet
Site Adı	Belirtilmemiş
Aidat (TL)	0
Krediye Uygun	Evet
Tapu Durumu	Kat İrtifaklı
Kimden	Emlak Ofisinden
Görüntülü Arama İle Gezilebilir	Evet
Takas	Evet

İbrahim Gürler

Cep 0 (535) 850 63 17

Scraping Algorithm Design

After doing some digging in sahibinden.com html codes, we designed the web scraping like below in our minds at first, there was a pattern in the url of summary page. Each page shows 20 ads and when you move from 1-20 to 21-40 (from 1st page to 2nd page) the url is almost same but only the number at the end changes from 20 to 40. We planned to gather the urls of the detail pages, in each summary page we would collect 20 urls. After collecting let's say 20.000 ad urls, then we would iterate over them 1 by 1.



Collect 20 url paths from a summary page and then move to next summary page in a loop. For example, if we loop over 10 summary pages, we get a python list of 200 urls

When 1st Loop is done 2nd loop starts

Now we have the address of 200 detailed ads, we loop over them 1 by 1 and get features for each one of them. Each one of them is a different house.

We wrote the code and used Beautifulsoup, regexp(re) libraries of python. The algorithm started working well and there were no errors at first. However, after 10 minutes the website blocked our ip address because of too many requests in a given time of period. There was a warning which said unusual automated access was detected. For one day, our wifi ip was blocked. When we changed connection to our smartphones, we were able to open the site but same thing happened after 10 minutes.

We came up with a very simple solution. The solution was to slow down our web mining drastically. We put some waiting periods after some scraping is done. For example, after loading the first 30 ads, the code waits for 15 minutes then continues. This worked nicely but its time cost is huge. We are still working on this part for a better solution. Up to now we were able to get only 2300 rows of data. Moreover, the connection time outs somehow, we couldn't solve that problem either.

As a result, we decided to model on the data available but we are still web-mining. In the sample code for replication we reduced the number of rows to 25, but in our real code it continues until 300-400 rows are reached each time executed. The data was written into a csv file. (real_estate.csv).

Pre-processing & Feature Generation

We used seaborn, matplotlib, NumPy and pandas for this stage. First, we dropped some unnecessary columns like date and id and then we checked the unique elements of each column. There were some non-informative columns that have the same value for each house, because of these type of columns' zero variance we dropped those as well.

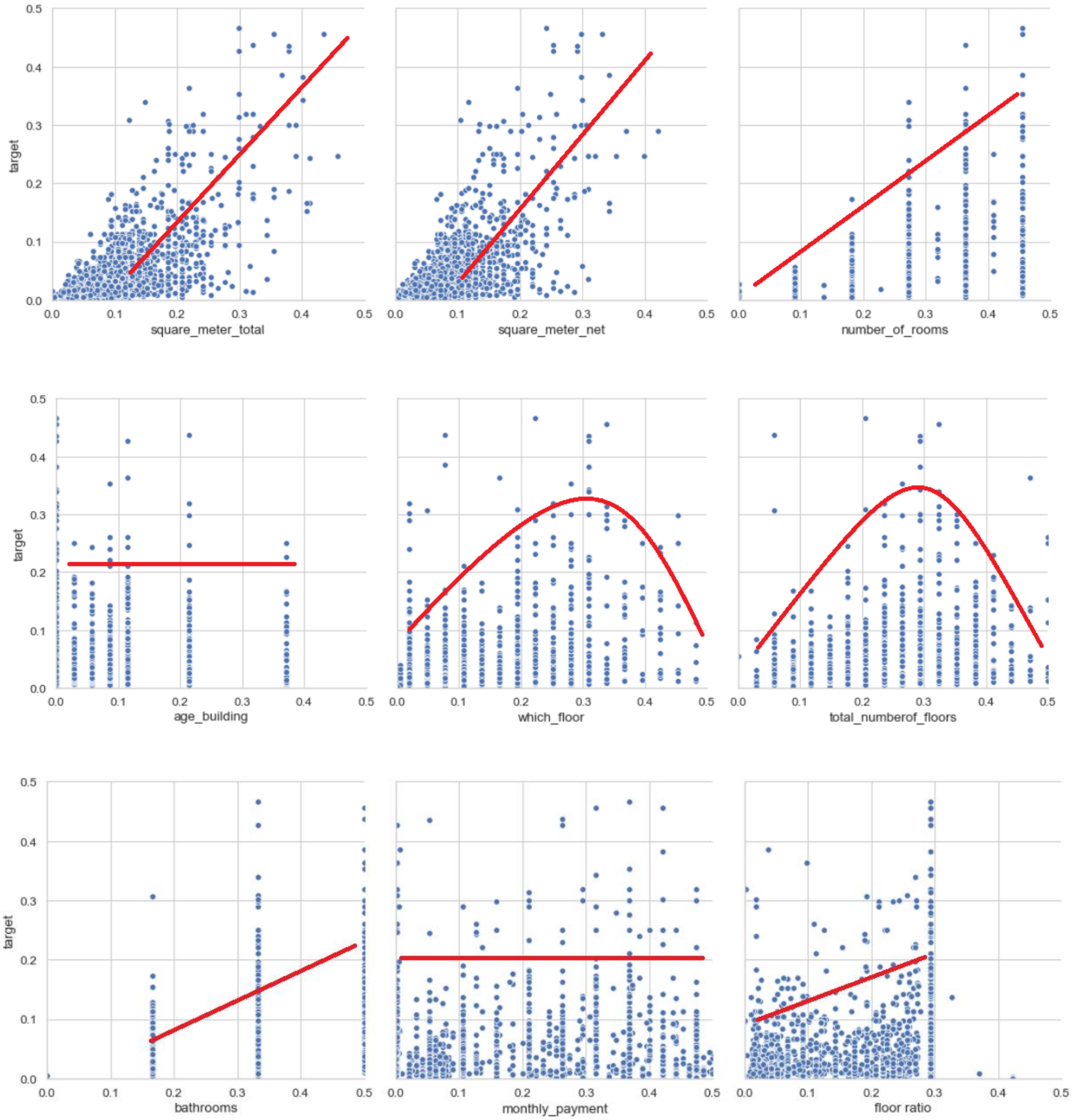
Then we made some preprocessing on the variables; some Turkish categorical data was transformed to None, and some ordinal categorical data was converted to numerical (opposite of binning). Actually, some numerical features are binned in the website, but not binned completely either. For example, the values of the number of floors are 1,2,3,4 and then 5-10, 10+. So, we replaced '5-10 floors' with 7.5 and '10+ floors' with 12 by intuition. Similar to those, instead of 5+1 rooms we used 6, 7 for 5+2 etc… After that we made some mapping like yes=1 and no=2, and for the remaining categorical. At this stage we had 21 inputs and 1 output listed below with descriptions on the right column.

	Feature	Description
1	**Price (TARGET OUTPUT)**	selling price of the house – **TARGET OUTPUT**
2	**neighborhood**	neighborhood or district
3	**square_meter_total**	size of the house , balcony, yard inclued
4	**square_meter_net**	net size of the house , balcony etc. others excluded.
5	**number_of_rooms**	number of rooms in house
6	**age_building**	Ag of the building in years
7	**which_floor**	At what floor is the house ?
8	**total_numberof_floors**	What is the total number of floors in the building
9	**heating_system**	central ground heating / natural gas
10	**bathrooms**	number of bathrooms
11	**bathroom_yes_no**	is there a bathroom
12	**house_appliances**	Tv, dishwasher, fridge etc. Also included in sale?
13	**usage_st**	Is the facility occupied right now ?
14	**in_a_compound**	Is it part of a bigger compund like istinyepark or akasya mall?
15	**compound_name**	compound name if available
16	**monthly_payment**	monthly maintanence amount paid
17	**loan_status**	eligible for loan application from the bank ?
18	**realestate_document_type**	land %, property %, flat ownership %100, needs approval
19	**seller_type**	owner/ contractor firm/ real estate agent
20	**online_tour_available**	Online visual tour (yes/no)
21	**trade**	instead of cash payment is there an exchange possibility?
22	**floor ratio**	which_floor divided by total_number_of_floors

Data visualizations and feature hypothesis

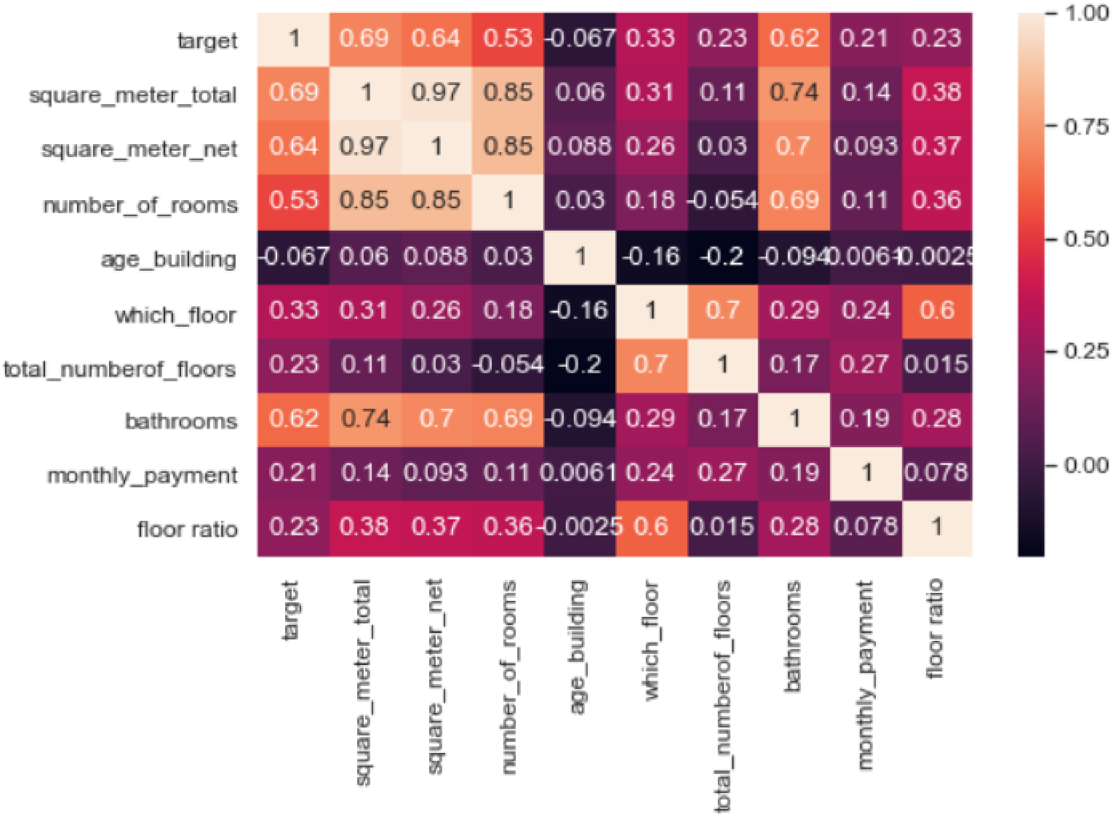
For the numerical features as you can see below age of the building and monthly maintenance payments has no correlation with target. The floor number is an interesting finding which is parallel with our expectations. People does not like very high floors nor entrance level floors. The middle levels are more valuable. Some of those variables are correlated with each other as well.

We visualize the target vs variables as shown below. To clarification reasons, we scale the numbers between 0-1 in the both axes.



Correlation Matrix of numerical features and dependent variable

Square meter total and & net are highly correlated, number of rooms and number of bathrooms too. It makes sense as expected, if a house is big it will have more rooms and more bathrooms. We decided to proceed with square meter total only and dropped the remaining 3.
Age of building has no correlation and the others have weak correlations with the dependent variable. We dropped age of building as well. After this stage, we checked the correlation of all features including the Booleans and the on-hot encoded columns as well. All features with a correlation less than 0.15 with the dependent variable were dropped.

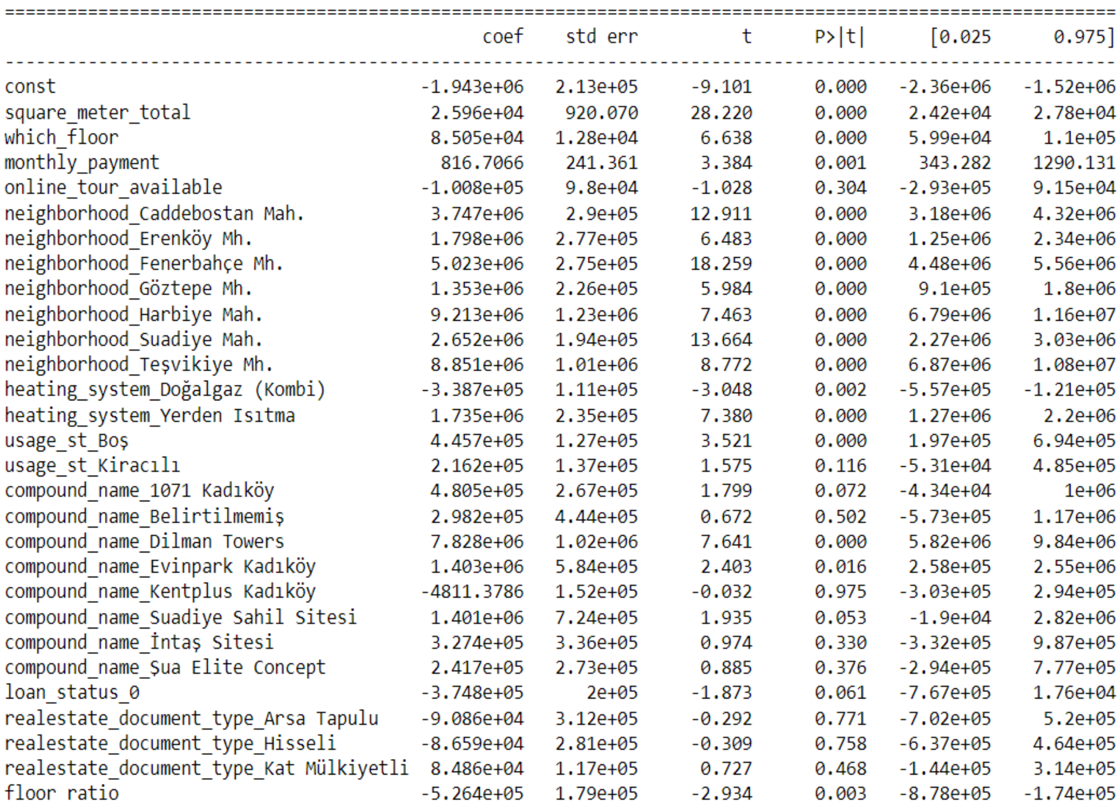


	target	square_meter_total	square_meter_net	number_of_rooms	age_building	which_floor	total_numberof_floors	bathrooms	monthly_payment	floor ratio
target	1	0.69	0.64	0.53	-0.067	0.33	0.23	0.62	0.21	0.23
square_meter_total	0.69	1	0.97	0.85	0.06	0.31	0.11	0.74	0.14	0.38
square_meter_net	0.64	0.97	1	0.85	0.088	0.26	0.03	0.7	0.093	0.37
number_of_rooms	0.53	0.85	0.85	1	0.03	0.18	-0.054	0.69	0.11	0.36
age_building	-0.067	0.06	0.088	0.03	1	-0.16	-0.2	-0.094	0.0061	-0.0025
which_floor	0.33	0.31	0.26	0.18	-0.16	1	0.7	0.29	0.24	0.6
total_numberof_floors	0.23	0.11	0.03	-0.054	-0.2	0.7	1	0.17	0.27	0.015
bathrooms	0.62	0.74	0.7	0.69	-0.094	0.29	0.17	1	0.19	0.28
monthly_payment	0.21	0.14	0.093	0.11	0.0061	0.24	0.27	0.19	1	0.078
floor ratio	0.23	0.38	0.37	0.36	-0.0025	0.6	0.015	0.28	0.078	1

Model Algorithms & Comparison

Model Type	Hyperparameters	Train R-squared	Test R-squared	Overfit ?
Linear Regression		68.6%	72.5%	No
Neural Network	(32.16.10.5) max_iter=5000 alpha=0.001	88.0%	76.6%	Yes
Neural Network	(32.16.8.6.4) max_iter=5000 alpha=0.001	89.2%	75.8%	Yes
Neural Network	(32.16.8) max_iter=5000 alpha=0.001	79.5%	78.6%	No
Random Forest	max depth =2	52.1%	50.0%	No
Random Forest	max depth =3	68.0%	59.0%	Yes
Random Forest	max depth =4	77.0%	64.0%	Yes
Random Forest	max depth =5	83.0%	68.0%	Yes

3 of the models were robust however random forest seems to be performing poorly. Some of the algorithms were performing very high on train set but couldn't perform well on test set. This is a signal of overfitting which is failing to generalize. Overfitted models actually memorize the train set so we don't want to proceed with them. This leaves us with two choice. 3 hidden layers MLP with 32 16 8 hidden nodes and the classic linear regression. NN clearly has better results and wins with %78.6 against %72.5 of linear regression. However, we would like to explain the significance variables with their p values as you can see below.



```
==========================================================================================================
                                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                    -1.943e+06   2.13e+05     -9.101      0.000   -2.36e+06   -1.52e+06
square_meter_total                        2.596e+04    920.070     28.220      0.000    2.42e+04    2.78e+04
which_floor                               8.505e+04   1.28e+04      6.638      0.000    5.99e+04     1.1e+05
monthly_payment                            816.7066    241.361      3.384      0.001     343.282    1290.131
online_tour_available                    -1.008e+05    9.8e+04     -1.028      0.304   -2.93e+05    9.15e+04
neighborhood_Caddebostan Mah.             3.747e+06    2.9e+05     12.911      0.000    3.18e+06    4.32e+06
neighborhood_Erenköy Mh.                  1.798e+06   2.77e+05      6.483      0.000    1.25e+06    2.34e+06
neighborhood_Fenerbahçe Mh.               5.023e+06   2.75e+05     18.259      0.000    4.48e+06    5.56e+06
neighborhood_Göztepe Mh.                  1.353e+06   2.26e+05      5.984      0.000     9.1e+05     1.8e+06
neighborhood_Harbiye Mah.                 9.213e+06   1.23e+06      7.463      0.000    6.79e+06    1.16e+07
neighborhood_Suadiye Mah.                 2.652e+06   1.94e+05     13.664      0.000    2.27e+06    3.03e+06
neighborhood_Teşvikiye Mh.                8.851e+06   1.01e+06      8.772      0.000    6.87e+06    1.08e+07
heating_system_Doğalgaz (Kombi)          -3.387e+05   1.11e+05     -3.048      0.002   -5.57e+05   -1.21e+05
heating_system_Yerden Isıtma              1.735e+06   2.35e+05      7.380      0.000    1.27e+06     2.2e+06
usage_st_Boş                              4.457e+05   1.27e+05      3.521      0.000    1.97e+05    6.94e+05
usage_st_Kiracılı                         2.162e+05   1.37e+05      1.575      0.116   -5.31e+04    4.85e+05
compound_name_1071 Kadıköy                4.805e+05   2.67e+05      1.799      0.072   -4.34e+04       1e+06
compound_name_Belirtilmemiş               2.982e+05   4.44e+05      0.672      0.502   -5.73e+05    1.17e+06
compound_name_Dilman Towers               7.828e+06   1.02e+06      7.641      0.000    5.82e+06    9.84e+06
compound_name_Evinpark Kadıköy            1.403e+06   5.84e+05      2.403      0.016    2.58e+05    2.55e+06
compound_name_Kentplus Kadıköy           -4811.3786   1.52e+05     -0.032      0.975   -3.03e+05    2.94e+05
compound_name_Suadiye Sahil Sitesi        1.401e+06   7.24e+05      1.935      0.053    -1.9e+04    2.82e+06
compound_name_İntaş Sitesi                3.274e+05   3.36e+05      0.974      0.330   -3.32e+05    9.87e+05
compound_name_Şua Elite Concept           2.417e+05   2.73e+05      0.885      0.376   -2.94e+05    7.77e+05
loan_status_0                            -3.748e+05      2e+05     -1.873      0.061   -7.67e+05    1.76e+04
realestate_document_type_Arsa Tapulu     -9.086e+04   3.12e+05     -0.292      0.771   -7.02e+05     5.2e+05
realestate_document_type_Hisseli         -8.659e+04   2.81e+05     -0.309      0.758   -6.37e+05    4.64e+05
realestate_document_type_Kat Mülkiyetli   8.486e+04   1.17e+05      0.727      0.468   -1.44e+05    3.14e+05
floor ratio                              -5.264e+05   1.79e+05     -2.934      0.003   -8.78e+05   -1.74e+05
```

Findings and Conclusion

Most of the features were included in the model just as we expected, however some of them were strangely insignificant. For example, real estate document type is a major indicator in Turkey. It’s so important that even bank don’t grans loans to shared type ownership documents. Neighborhood is also very significant, especially if it’s an elite neighborhood like Etiler, Bebek or Suadiye, any neighborhood around Bagdat caddesi etc.
There is a lot of potential for development in our models. First bottleneck was lack of data, we are searching ways to bypass website security issues. Moreover, there are comments in every page which is in freetext form. We didn’t have time to use text analytics or NLP for that field, it’s another opportunity for improvement as well.

	Feature	comments
2	**neighborhood**	for some neigborhoods we can conclude it is very decisive
4	**square_meter_net**	most powerful feature without question
7	**which_floor**	it is significant , people avoid top floors and floors at the very bottom
9	**heating_system**	ground heating system increases house prices
13	**usage_st**	if its unoccupied it has an effect
15	**compound_name**	similar to neighborhood some of the componds effects the price greatly
16	**monthly_payment**	if its a hous with good maintenance the monthly payment and price are correlated
17	**loan_status**	insignificant
18	**realestate_document_type**	insignificant
20	**online_tour_available**	insignificant
22	**floor ratio**	it is significant , people avoid top floors and floors at the very bottom