Nashville housing Dataset

2023 October 20th, ALY6020

Presented to Professor Behzad Ahmadi
Prepared by Heejae Roh

Data Set
& Data Cleaning

01

Target Variables

O	P	Q	R	S	T	U	V	W	X	Y	Z
Neighborhood	Land Value	Building Value	Finished Area	Foundation Type	Year Built	Exterior Wall	Grade	Bedrooms	Full Bath	Half Bath	Sale Price Compared To Value
3127	32000	134400	1149	PT BSMT	1941	BRICK	C	2	1	0	Over
9126	34000	157800	2090.82495	SLAB	2000	BRICK/FRAME	C	3	2	1	Over
3130	25000	243700	2145.60001	FULL BSMT	1948	BRICK/FRAME	B	4	2	0	Under
3130	25000	138100	1969	CRAWL	1910	FRAME	C	2	1	0	Under
3130	25000	86100	1037	CRAWL	1945	FRAME	C	2	1	0	Under
3179	16000	68100	1216	CRAWL	1932	FRAME	D	2	1	0	Under
3179	16000	68100	1216	CRAWL	1932	FRAME	D	2	1	0	Under
3131	25000	57100	1152	CRAWL	1945	FRAME	C	2	1	0	Under
3131	25000	80100	1300	CRAWL	1955	BRICK	C	2	1	0	Under
3926	21500	87900	1175	CRAWL	1968	BRICK	C	3	1	1	Under
3926	21500	107800	1825	SLAB	1978	BRICK	C	7	2	0	Under
3926	21500	81700	1474	FULL BSMT	1960	BRICK	C	2	1	1	Over
3926	21500	92000	1600	SLAB	1962	BRICK	C	4	1	1	Over
4026	47000	213600	1436	CRAWL	1955	BRICK	C	2	2	0	Under
3927	28000	71100	1436	CRAWL	1970	BRICK	C	4	2	0	Over
3927	21000	83100	1455	CRAWL	1958	BRICK	C	3	2	0	Over

For linear Regression

For dt, rf, XGBoost

Target Variables

#	column name	Description
17	**Building Value**	worth of a building
18	**Finished Area**	finished area of the property in square feet
19	**Foundation Type**	type of foundation the property has (e.g., PT BSMT, SLAB)
20	**Year Built**	year the building was built
21	**Exterior Wall**	material of the property's exterior wall (e.g., BRICK, BRICK/FRAME)
22	**Grade**	grade or rating for the property (e.g., C, B)
23	**Bedrooms**	The number of bedrooms in the property
24	**Full Bath**	a bathroom that includes a shower, a bathtub, a sink, and a toilet.
25	**Half Bath**	a half bathroom only contains a sink and a toilet
26	**Sale Price Compared To Value**	evaluation of the sale price in comparison to its value (e.g., Over, Under)

Whole Variables

```
Data columns (total 26 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Unnamed: 0                       22651 non-null  int64
 1   Parcel ID                        22651 non-null  object
 2   Land Use                         22651 non-null  object
 3   Property Address                 22649 non-null  object
 4   Suite/ Condo   #                 0 non-null      float64
 5   Property City                    22649 non-null  object
 6   Sale Date                        22651 non-null  object
 7   Legal Reference                  22651 non-null  object
 8   Sold As Vacant                   22651 non-null  object
 9   Multiple Parcels Involved in Sale  22651 non-null  object
 10  City                             22651 non-null  object
 11  State                            22651 non-null  object
 12  Acreage                          22651 non-null  float64
 13  Tax District                     22651 non-null  object
 14  Neighborhood                     22651 non-null  int64
 15  Land Value                       22651 non-null  int64
 16  Building Value                   22651 non-null  int64
 17  Finished Area                    22650 non-null  float64
 18  Foundation Type                  22650 non-null  object
 19  Year Built                       22651 non-null  int64
 20  Exterior Wall                    22651 non-null  object
 21  Grade                            22651 non-null  object
 22  Bedrooms                         22648 non-null  float64
 23  Full Bath                        22650 non-null  float64
 24  Half Bath                        22543 non-null  float64
 25  Sale Price Compared To Value     22651 non-null  object
dtypes: float64(6), int64(5), object(15)
```

```
Number of Uniques:
Unnamed: 0                         22651
Parcel ID                          19720
Land Use                               4
Property Address                   20448
Suite/ Condo   #                       0
Property City                         10
Sale Date                           1044
Legal Reference                    22452
Sold As Vacant                         2
Multiple Parcels Involved in Sale      2
City                                  10
State                                  1
Acreage                              407
Tax District                           7
Neighborhood                         189
Land Value                           886
Building Value                      4298
Finished Area                       5955
Foundation Type                        6
Year Built                           125
Exterior Wall                          9
Grade                                 10
Bedrooms                              12
Full Bath                             11
Half Bath                              4
Sale Price Compared To Value           2
```

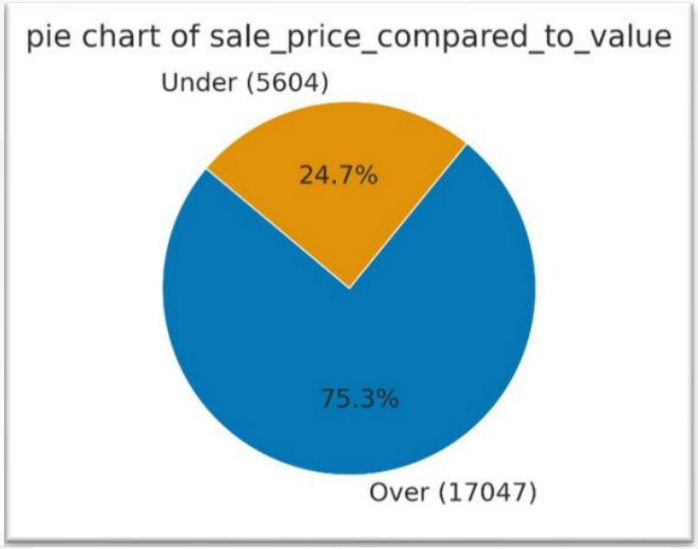
Exploratory Data Analysis

Descriptive Analysis of Dataset 1

	acreage	land_value	building_value	finished_area	year_built	bedrooms	full_bath	half_bath
count	22651	2.27E+04	2.27E+04	22650	22651	22648	22650	22543
mean	0.454705	7.01E+04	1.72E+05	1915.37715	1961.94768	3.10491	1.887285	0.270239
std	0.611818	1.03E+05	1.90E+05	1079.09452	25.843908	0.829287	0.95122	0.480186
min	0.04	9.00E+02	1.40E+03	450	1832	0	0	0
25%	0.2	2.20E+04	8.55E+04	1250	1947	3	1	0
50%	0.28	3.00E+04	1.19E+05	1645.825	1959	3	2	0
75%	0.46	6.03E+04	1.88E+05	2213.375	1977	4	2	1
max	17.5	1.87E+06	5.82E+06	19728.2499	2017	11	10	3

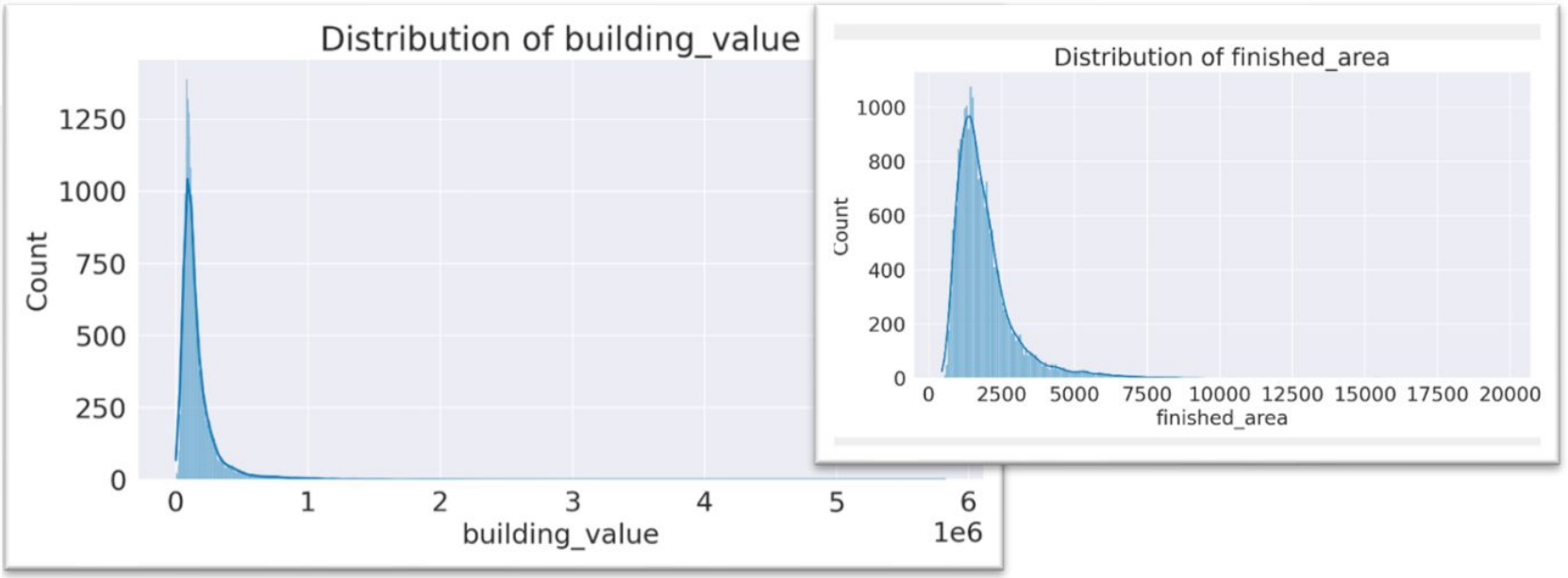
- The target variable building_value shows a mean of 172,000 and the max value is high at 582,000.
- The average value of year_built is 1961.94. The oldest house was built in 1832, and the most recent house was built in 2017, so you can see that the range is quite wide
- Bedrooms are distributed from 0 to 11, but you can see that most of them are distributed between 3 and 4

Pie Chart of Target variable



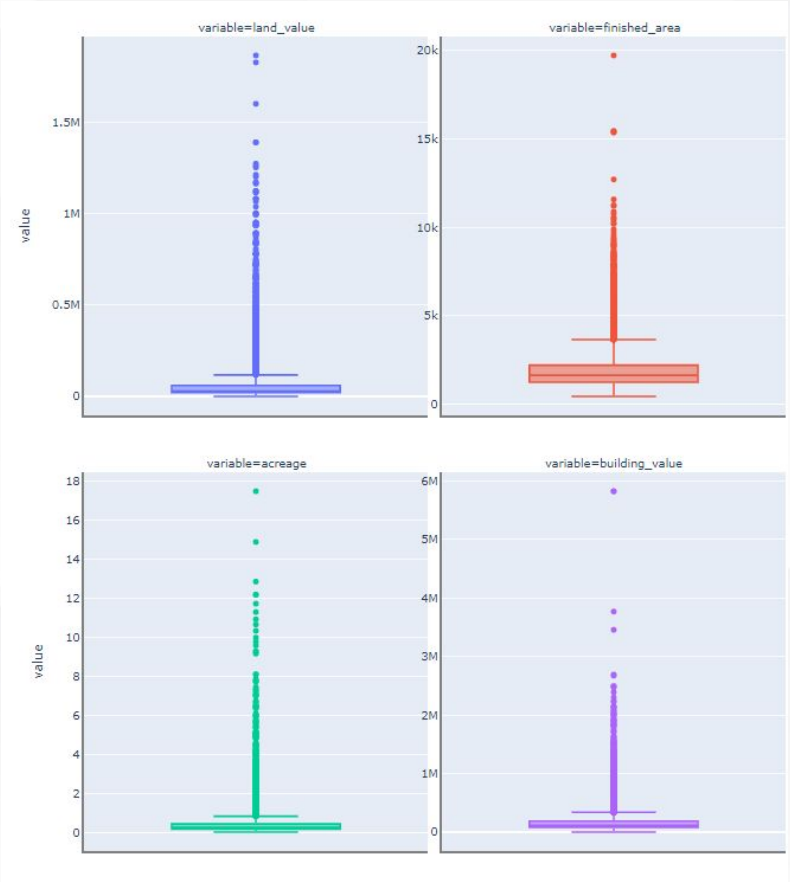
- The chart underscores an imbalance, with 'Over' representing a significant 75% of the data

Histogram of variables

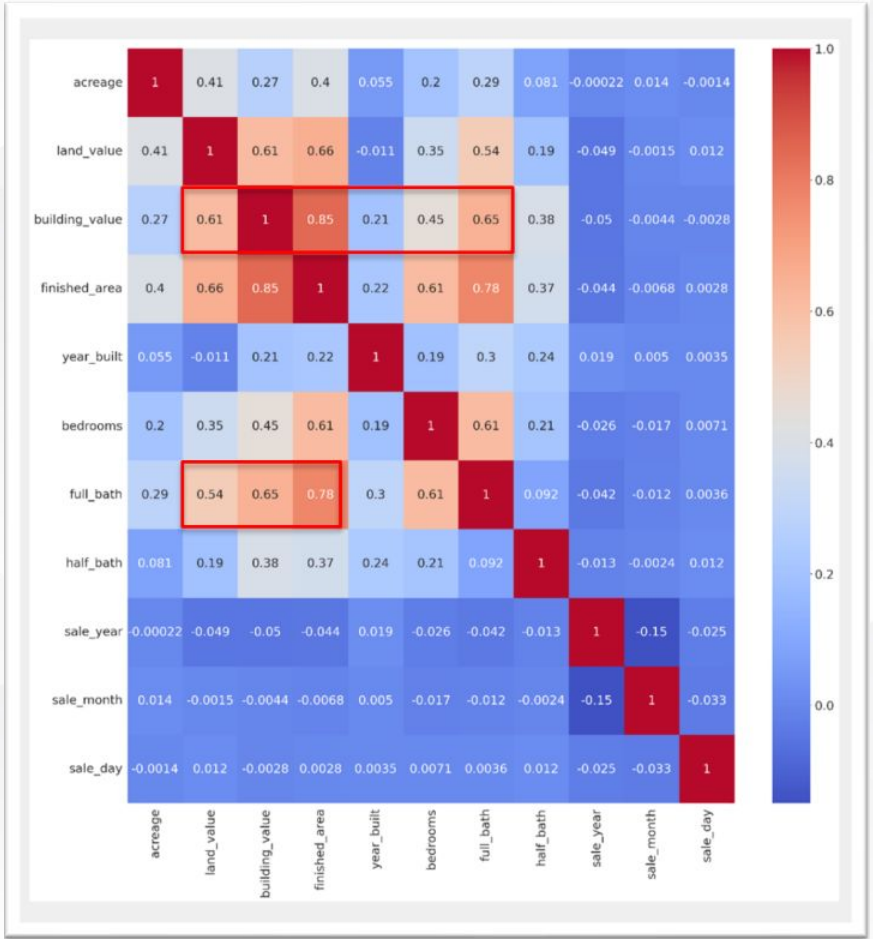


- A right-skewed distribution

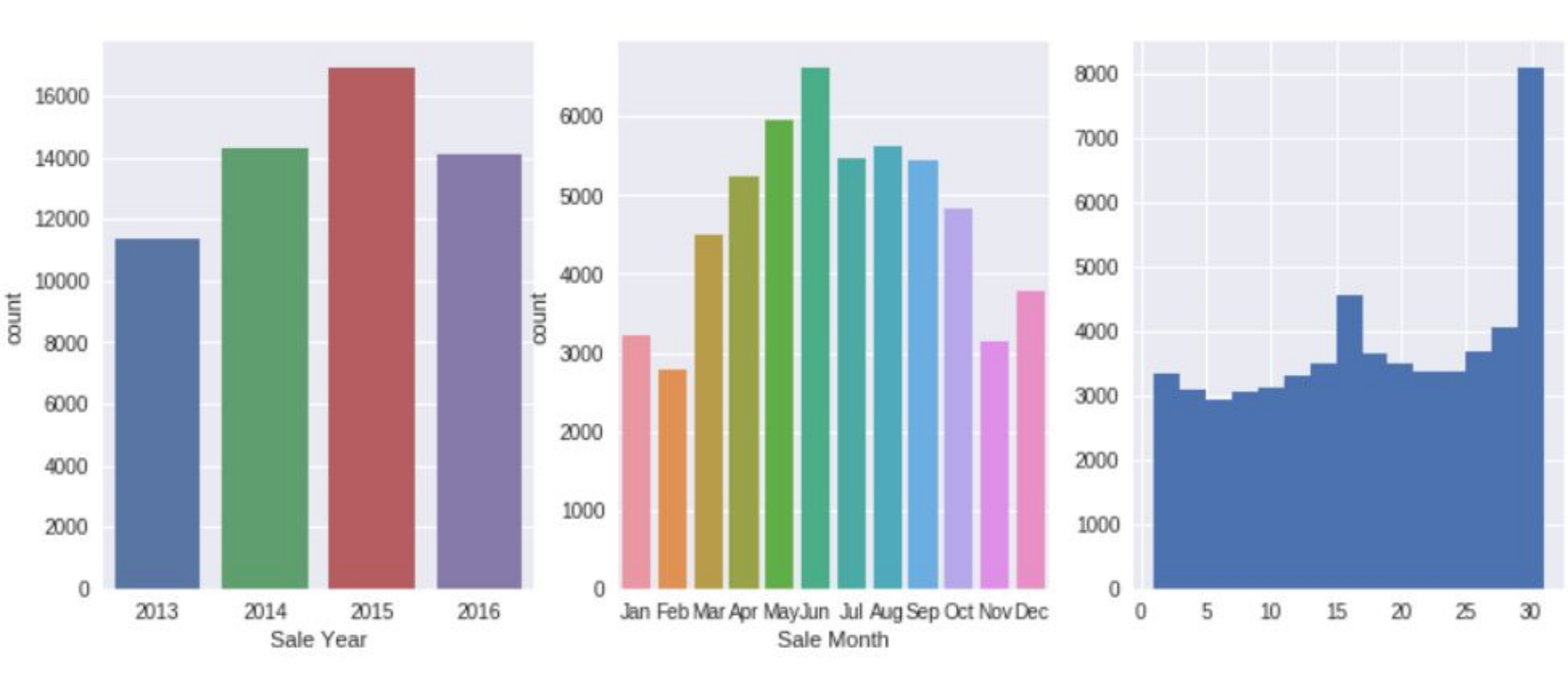
Boxplots of each



Correlation matrix of numerical variables



Data Cleansing 1. Date -> Year/Month/Day



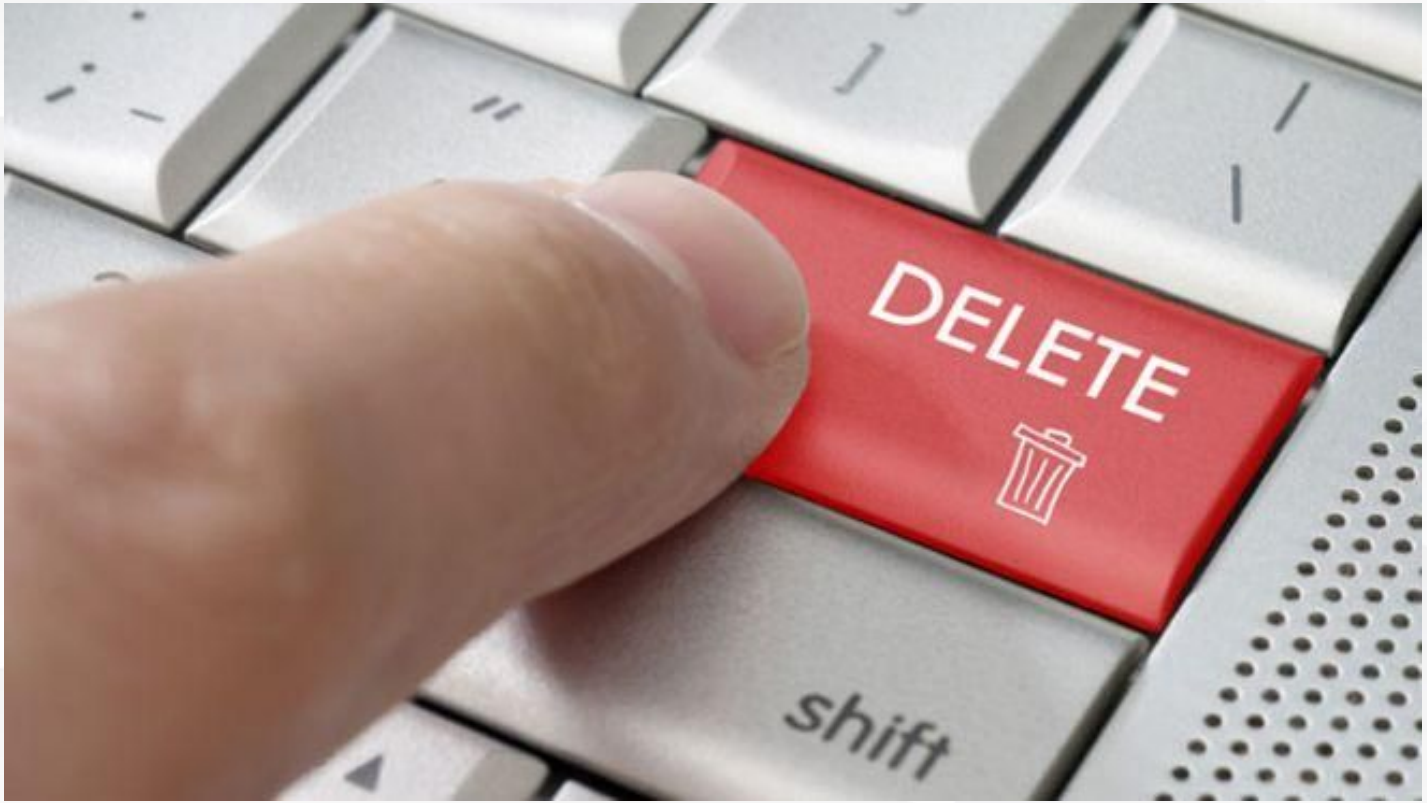
Cleansing 2. Check missing value & Drop

Missing Values:

	Missing_Number	Missing_Percent
Suite/ Condo #	22651	1.000000
Half Bath	108	0.004768
Bedrooms	3	0.000132
Property Address	2	0.000088
Property City	2	0.000088
Full Bath	1	0.000044
Foundation Type	1	0.000044
Finished Area	1	0.000044

	name	age	marks
0	Joe	NaN	85.10
1	Sam	NaN	NaN
2	NaN	NaN	NaN
3	Harry	NaN	91.54

Cleansing 3-5. Delete meaningless Columns



Linear Regression Process and Results

02

Before feature selection and normalization

OLS Regression Results

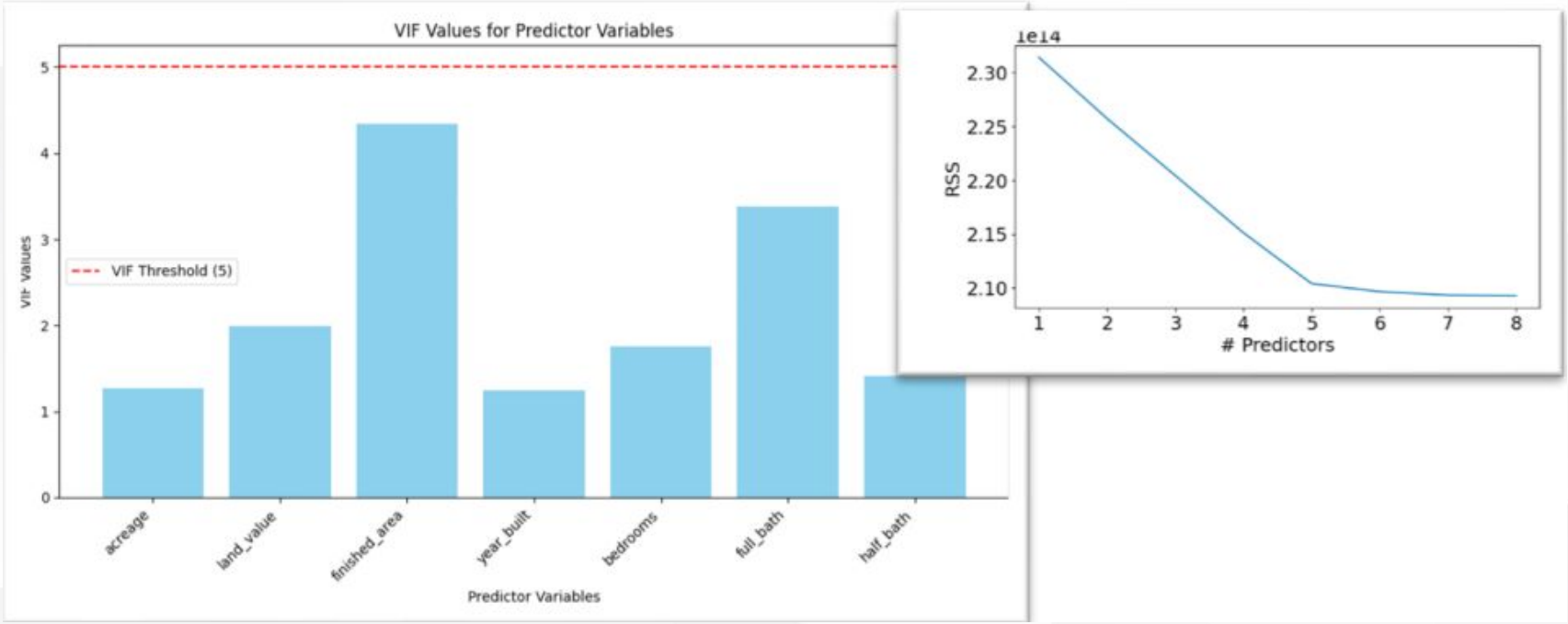
Dep. Variable:	building_value	R-squared:	0.735
Model:	OLS	Adj. R-squared:	0.735
Method:	Least Squares	F-statistic:	4999.
Date:	Thu, 19 Oct 2023	Prob (F-statistic):	0.00
Time:	21:32:37	Log-Likelihood:	-2.3289e+05
No. Observations:	18028	AIC:	4.658e+05
Df Residuals:	18017	BIC:	4.659e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.181e+06	1.39e+06	2.295	0.022	4.64e+05	5.9e+06
acreage	-3.103e+04	1407.094	-22.055	0.000	-3.38e+04	-2.83e+04
land_value	0.2246	0.010	22.398	0.000	0.205	0.244
finished_area	143.1306	1.438	99.569	0.000	140.313	145.948
year_built	182.7186	31.859	5.735	0.000	120.273	245.164
bedrooms	-2.553e+04	1168.102	-21.857	0.000	-2.78e+04	-2.32e+04
full_bath	6821.6508	1425.523	4.785	0.000	4027.489	9615.812
half_bath	3.39e+04	1822.694	18.596	0.000	3.03e+04	3.75e+04
sale_year	-1779.0124	688.062	-2.586	0.010	-3127.679	-430.345
sale_month	25.6320	242.105	0.106	0.916	-448.918	500.182
sale_day	-135.9338	81.563	-1.667	0.096	-295.806	23.938

Omnibus:	25570.369	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46611780.720
Skew:	7.705	Prob(JB):	0.00
Kurtosis:	251.626	Cond. No.	2.37e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.37e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Checking VIF and Best Subset method



With categorical variables with normalization

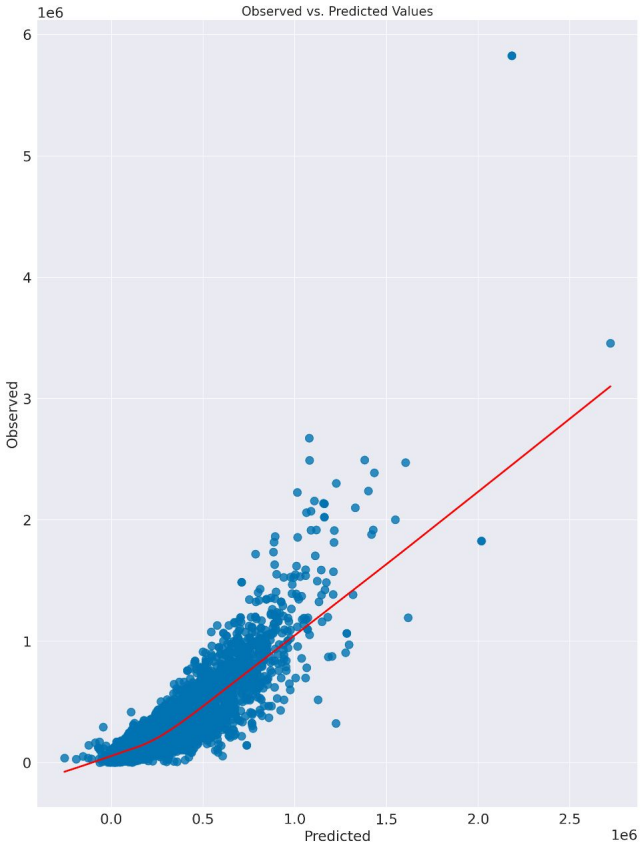
OLS Regression Results

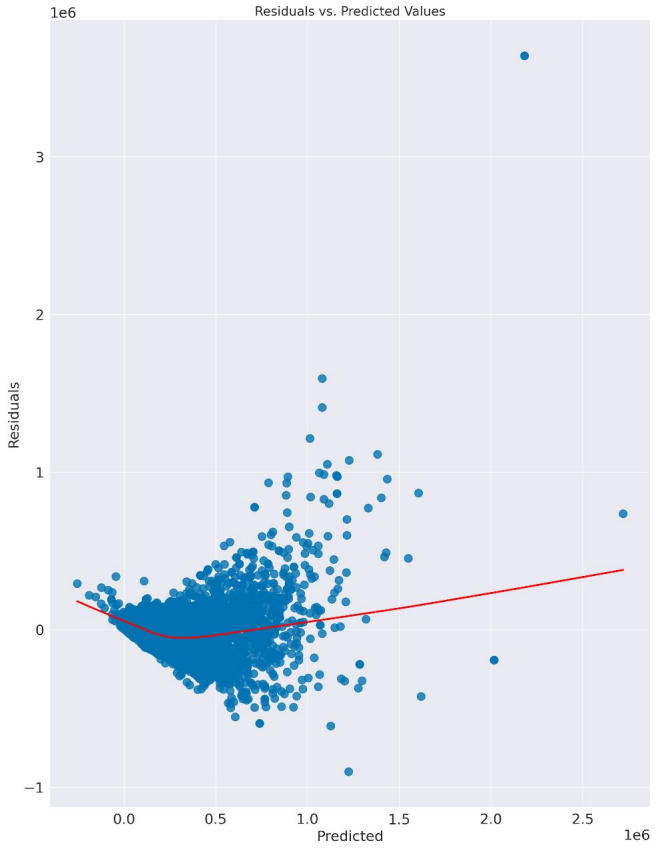
Dep. Variable:	building_value	R-squared:	0.735
Model:	OLS	Adj. R-squared:	0.735
Method:	Least Squares	F-statistic:	8313.
Date:	Thu, 19 Oct 2023	Prob (F-statistic):	0.00
Time:	21:33:00	Log-Likelihood:	-2.3285e+05
No. Observations:	18028	AIC:	4.657e+05
Df Residuals:	18021	BIC:	4.658e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.625e+04	2975.826	5.461	0.000	1.04e+04	2.21e+04
acreage	-5.298e+05	2.32e+04	-22.814	0.000	-5.75e+05	-4.84e+05
land_value	4.27e+05	1.82e+04	23.448	0.000	3.91e+05	4.63e+05
finished_area	2.864e+06	2.24e+04	127.883	0.000	2.82e+06	2.91e+06
bedrooms	-2.594e+05	1.23e+04	-21.039	0.000	-2.84e+05	-2.35e+05
half_bath	9.45e+04	4974.609	18.997	0.000	8.48e+04	1.04e+05
sold_as_vacant	6.133e+04	9807.383	6.253	0.000	4.21e+04	8.05e+04

Omnibus:	25315.703	Durbin-Watson:	2.018
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43579468.002
Skew:	7.564	Prob(JB):	0.00
Kurtosis:	243.389	Cond. No.	37.7

Checking residuals Assumption



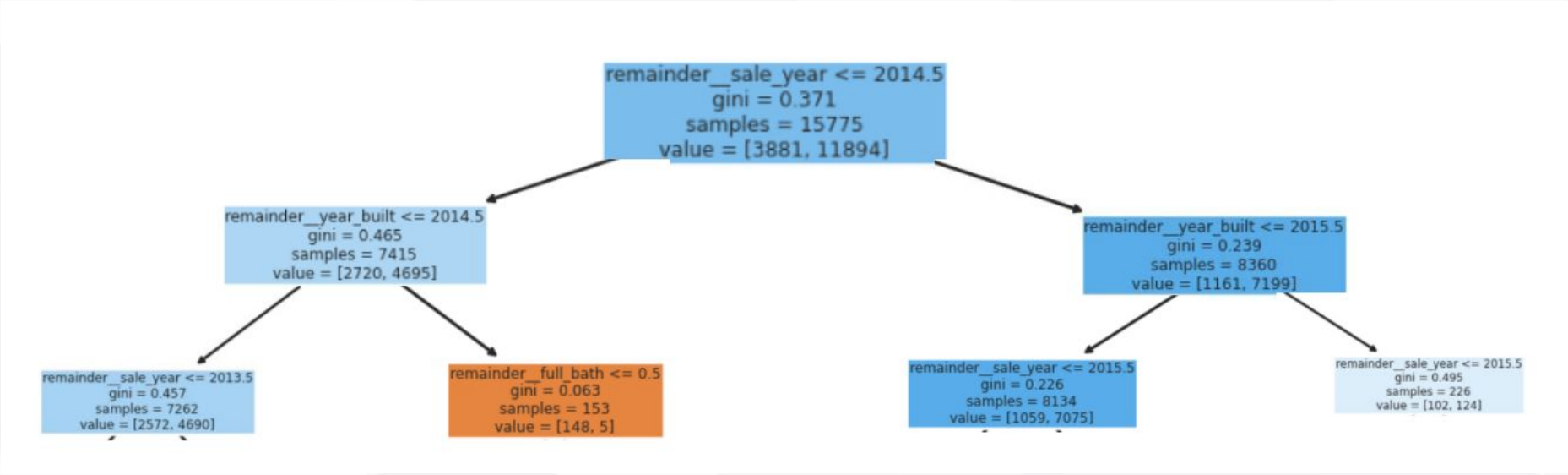


Decision Tree and Random Forest

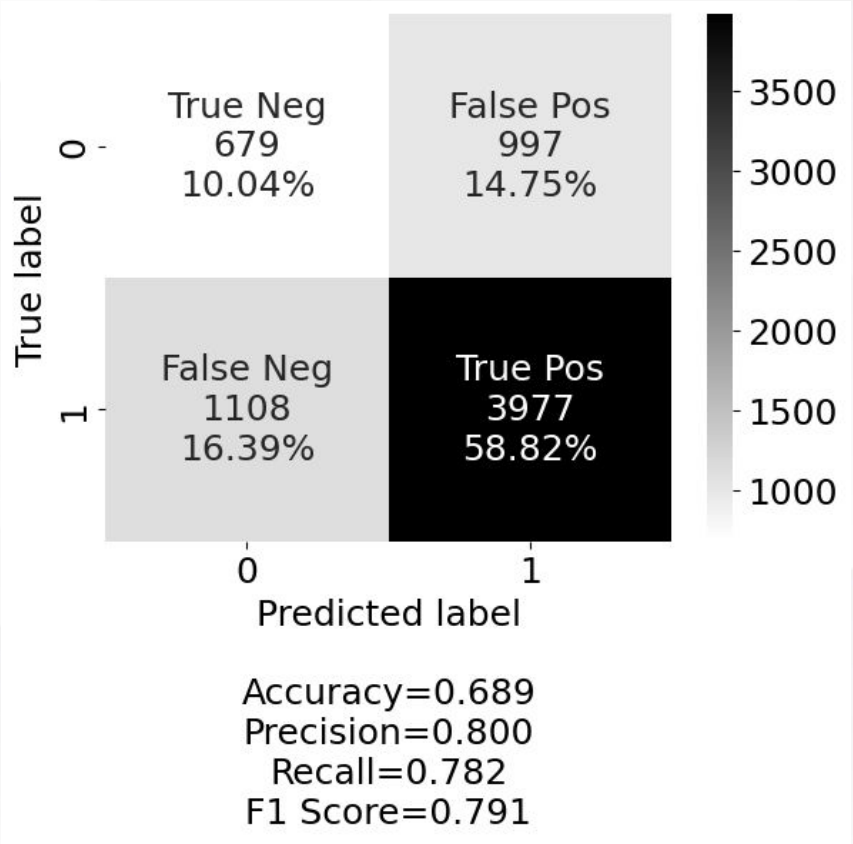
03

DT: Features importance and selection

DT: Layer 1 & Layer 2



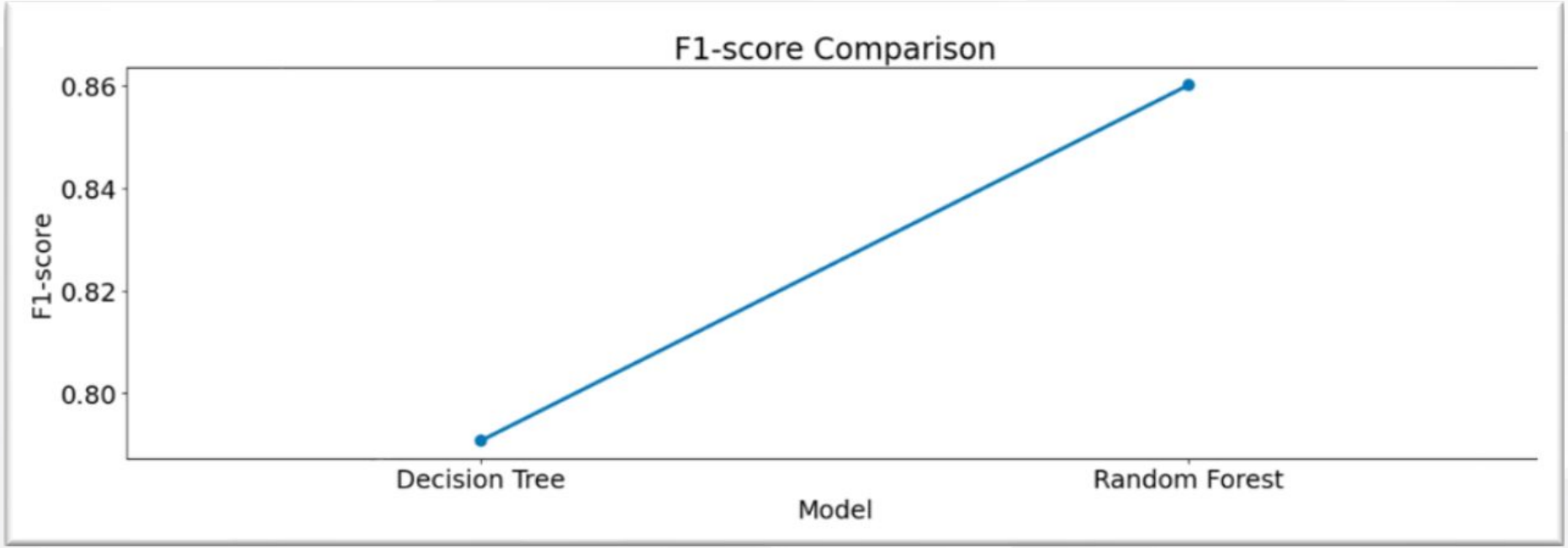
Decision Tree: confusion matrix



Random Forest Tree Number and Depth

Tree Number	Tree Depth
1	32
2	37
3	32
4	33
5	32
...	...
96	33
97	36
98	33
99	32
100	34

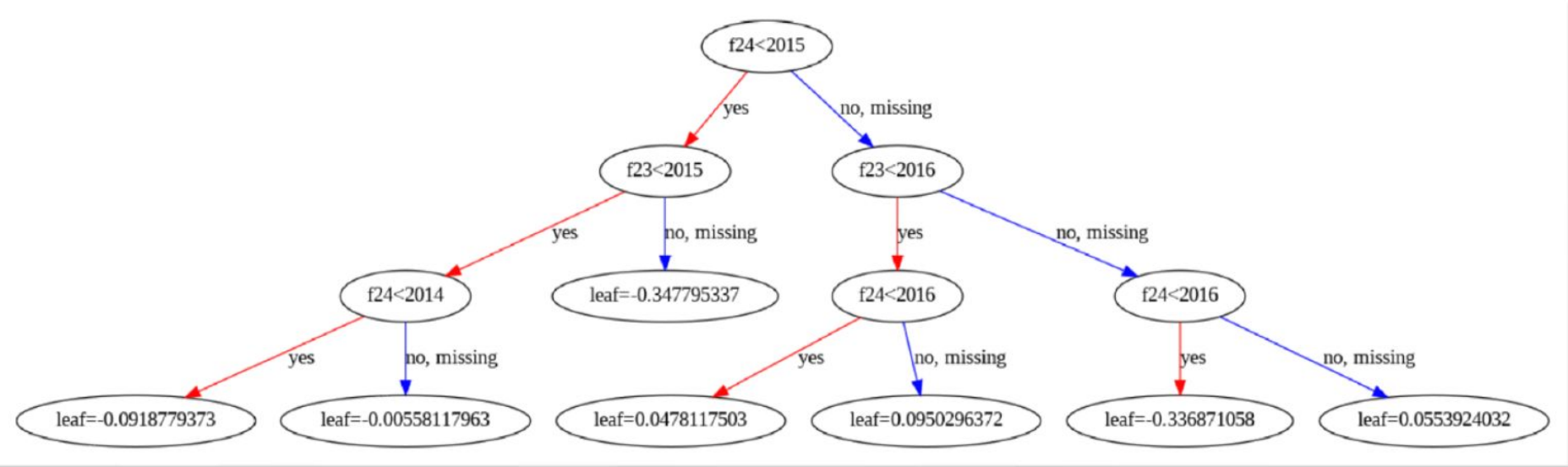
Decision Tree & Random Forest F1 score



XGBoost Result and Visualization

04

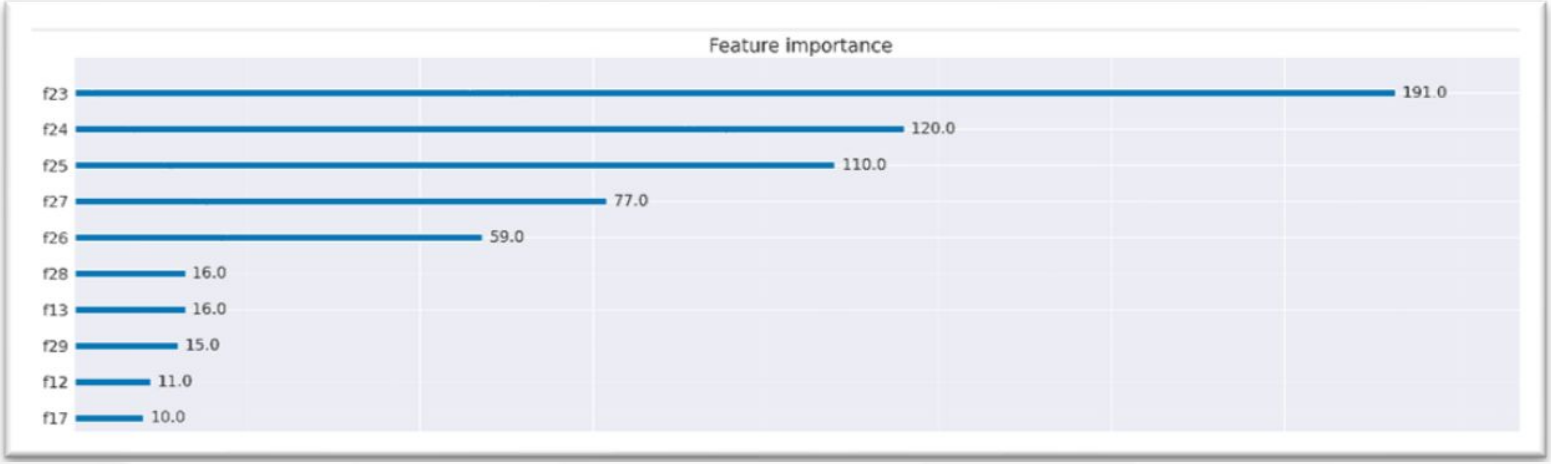
Number 0 Tree



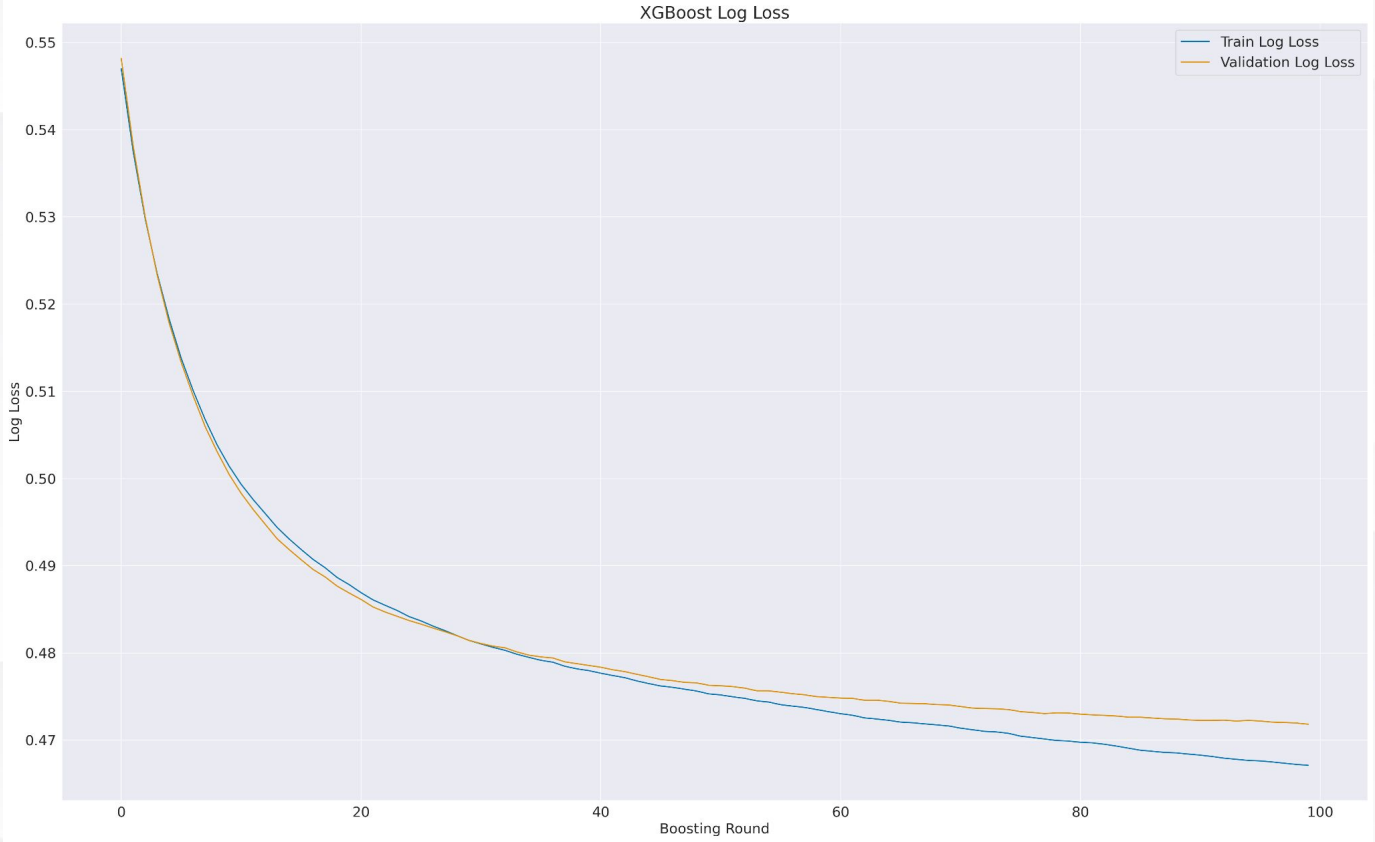
- 'max_depth': 3
- 'learning_rate': 0.1: Choose a relatively high learning rate. A value of 0.1 means that each tree's contribution to the final prediction will be reduced or "shrunk" by 10%
- 'n_estimators': 100
- 'eval_metric': 'logloss: measures the performance of a classification model where the prediction input is a probability value between 0 and 1

Features and importance

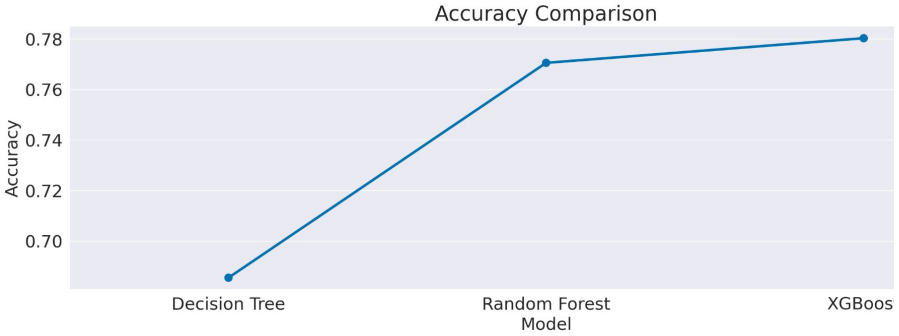
	Default Feature Name	Original Feature Name
0	f23	year_built
1	f24	sale_year
2	f25	finished_area
3	f27	sale_month
4	f26	sale_day

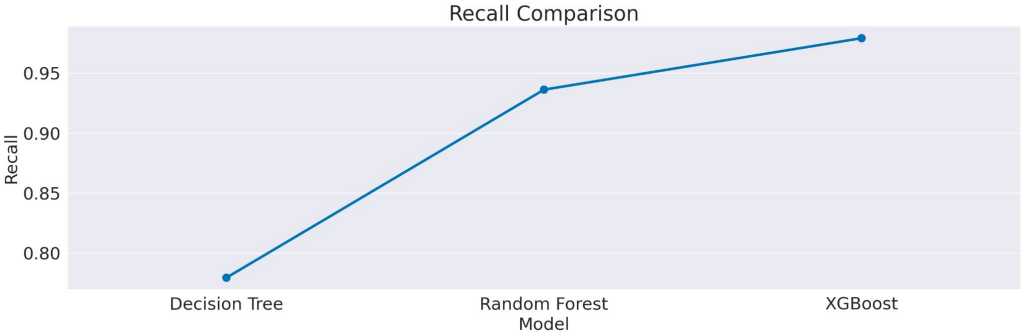


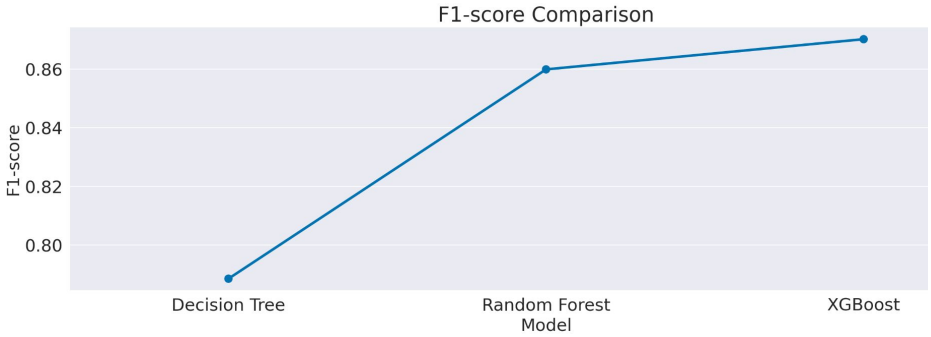
Log Loss Visualization

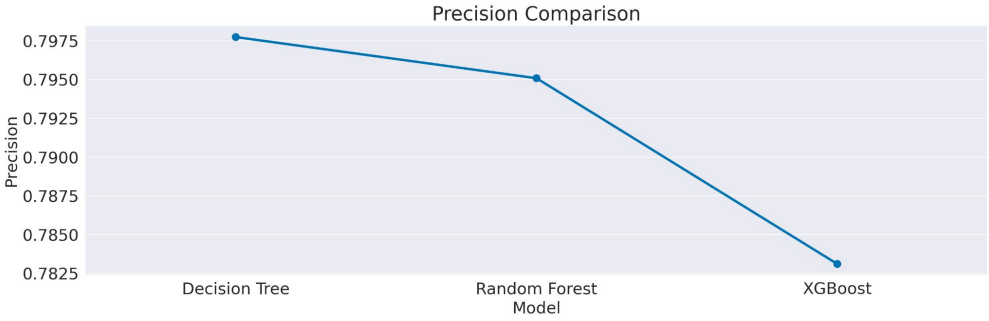


Comparing Three method

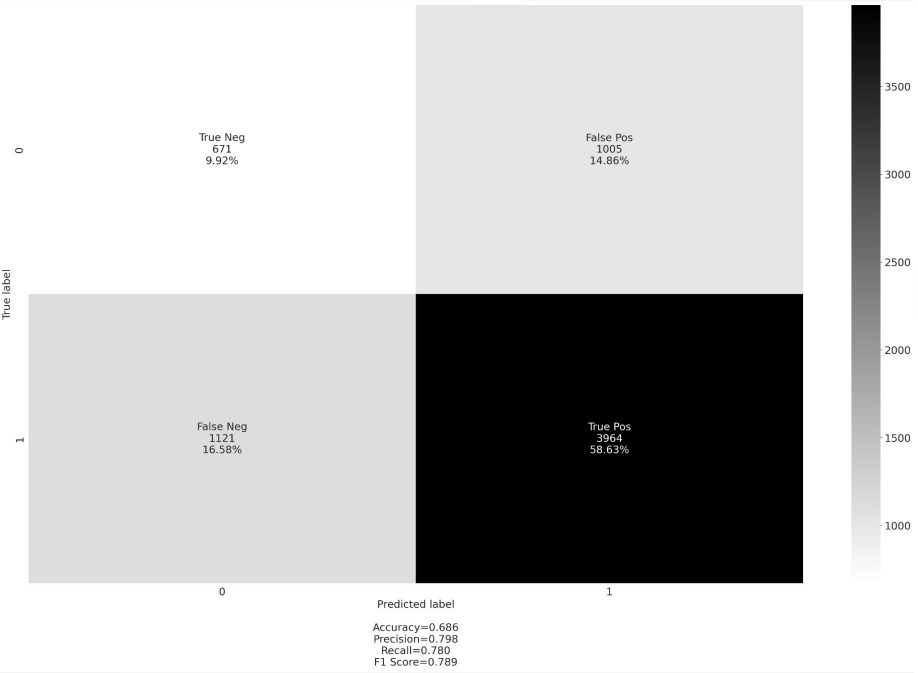


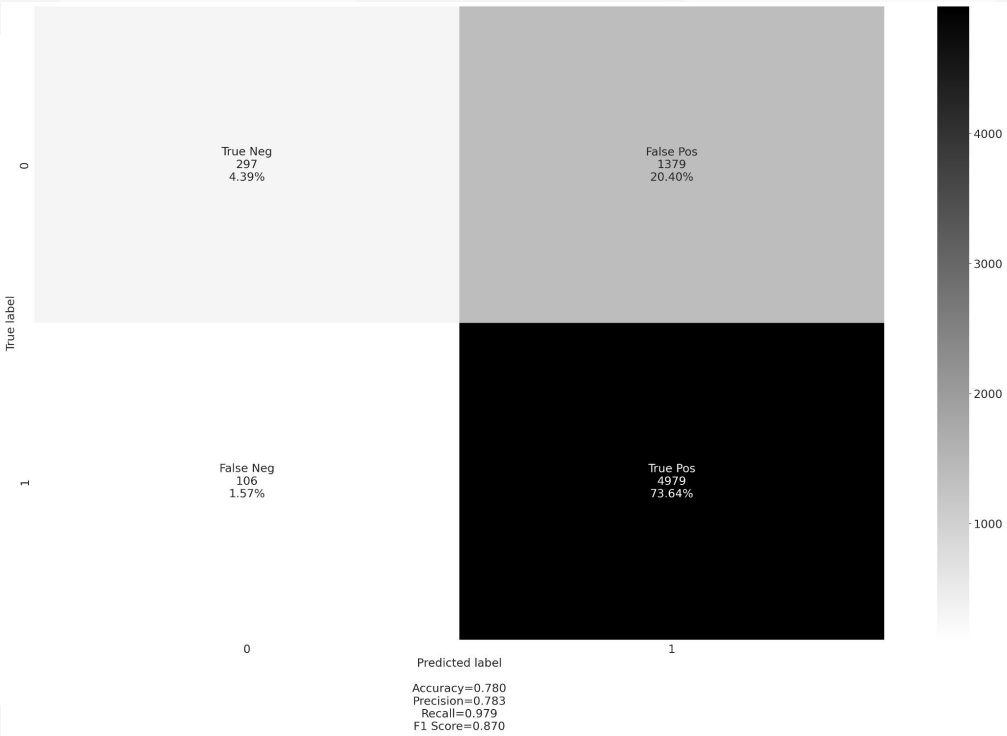




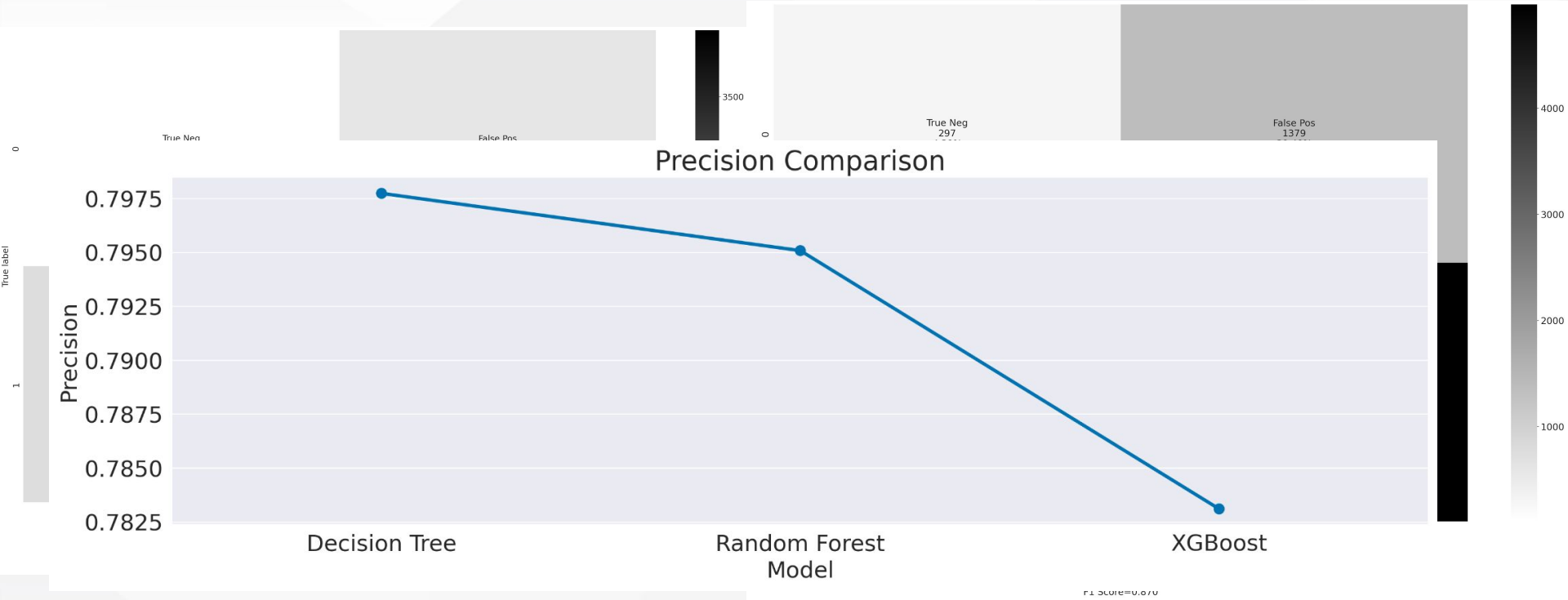


Comparing Three method





Comparing Three method



Thanks! Do you have any questions?