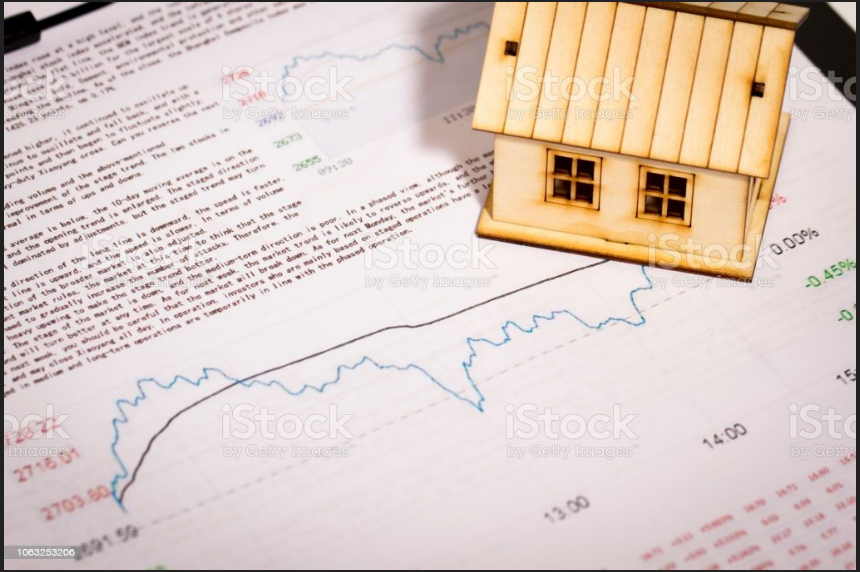
Predicting Housing Prices in the Bay Area

By Patrick Anderson, Amanda Delfino, and Mario Morales

Table of Contents

Current Issues Concerning Machine Learning in Real Estate

- Too many systems, collect data in different ways
- Timely and complete data is the biggest barrier to entry



Problem Statement

San Francisco MLS Data

Can we accurately predict housing prices in the Bay Area using Machine Learning?

Is there a day of the week that is best to list your property?

Business Case

BUYER

- Allows buyers to gain confidence around the list price of their property

SELLER

- Enable sellers to feel comfortable with pricing estimates regarding their investment

INVESTORS

- what goes in to pricing a real investment
- Helps appraisers appraise true value of properties for taxes

Data Source and Suggested Approach

- Collected data from SF MLS
- Predict prices of residential properties in SF, sold in the year 2020.
- Use this finding to predict prices on currently active homes in San Francisco.

Model - Data Cleansing

```
[ ] df.describe(include = 'all') #Listing all columns with various datatypes
```

	# of Rooms	# of Units	San Francisco	San Jose	Sacramento	Oakland	Walnut Creek
count	14174.000000	14174.000000	14174.000000	14174.000000	14174.000000	14174.000000	14174.000000

MAPE	*	lower	12.87551	0
GINI		higher	0.9728788	0
MAE		lower	77558.6	0.0078125
MER		lower	8.298738	0
MSE		lower	2.443724e+10	4096
R2		higher	0.8961563	2.220446e-16

Normalizing non-binary columns (Continuous 0 to 1)

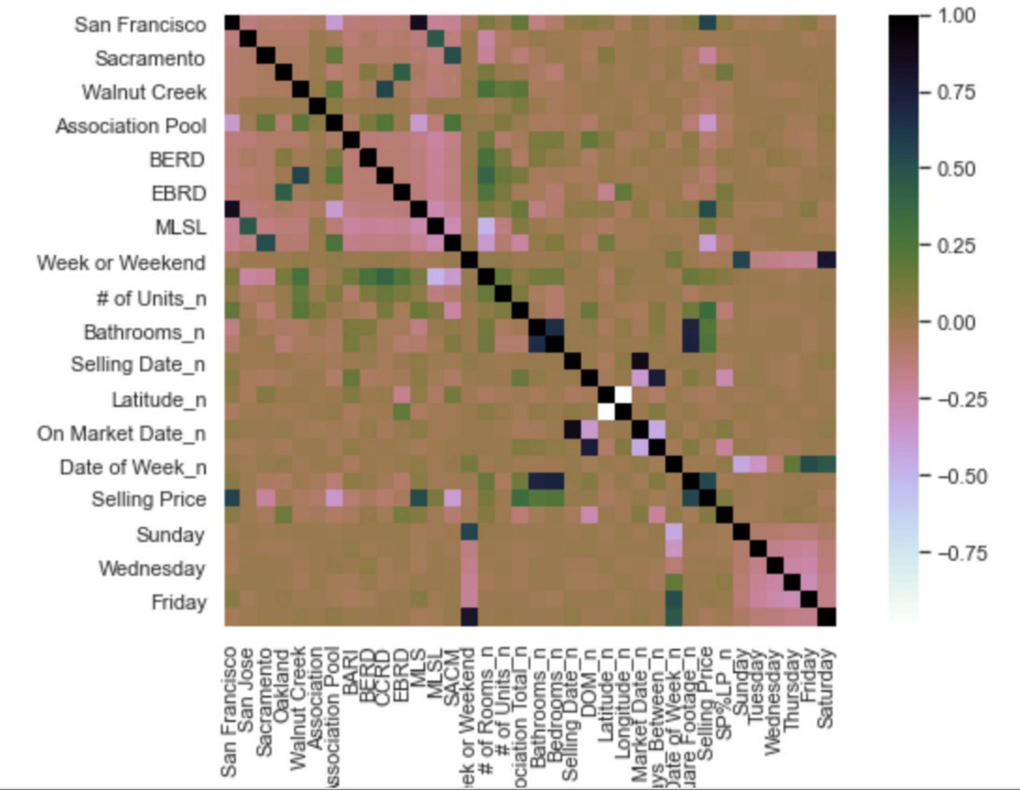
- Normalized columns will have **'_n'** appended to name

```
[ ]
X = np.array(df['# of Rooms']).reshape(-1,1)
scaler = MinMaxScaler()
scaler.fit(X)
X_scaled = scaler.transform(X)
df['# of Rooms_n'] = X_scaled.reshape(1,-1)[0]
```

```
[ ]
X = np.array(df['# of Units']).reshape(-1,1)
scaler = MinMaxScaler()
scaler.fit(X)
X_scaled = scaler.transform(X)
df['# of Units_n'] = X_scaled.reshape(1,-1)[0]
```

Correlation Heat Map

	Selling Price
San Francisco	0.56
San Jose	-0.01
Sacramento	-0.22
Oakland	-0.01
Walnut Creek	-0.05
Association	-0.07
Association Pool	-0.35
BARI	-0.10
BERD	-0.04
CCRD	-0.10
EBRD	-0.06
MLS	0.52
MLSL	0.10
SACM	-0.38
Week or Weekend	-0.04
# of Rooms_n	0.11
# of Units_n	-0.01
Association Total_n	0.34
Bathrooms_n	0.24
Bedrooms_n	0.25
Selling Date_n	0.02
DOM_n	0.02
Latitude_n	-0.04
Longitude_n	-0.01
On Market Date_n	0.03
Days_Between_n	-0.02
Date of Week_n	-0.01
Square Footage_n	0.54
Selling Price	1.00
SP%LP_n	0.08
Sunday	-0.01
Tuesday	-0.02
Wednesday	-0.01
Thursday	0.01
Friday	0.02
Saturday	-0.04



Model - Using Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Selling Price_n   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.715
Method:                 Least Squares   F-statistic:                     1272.
Date:                Thu, 02 Dec 2021   Prob (F-statistic):               0.00
Time:                        15:05:27   Log-Likelihood:                 34940.
No. Observations:               14174   AIC:                        -6.982e+04
Df Residuals:                   14145   BIC:                        -6.960e+04
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
```

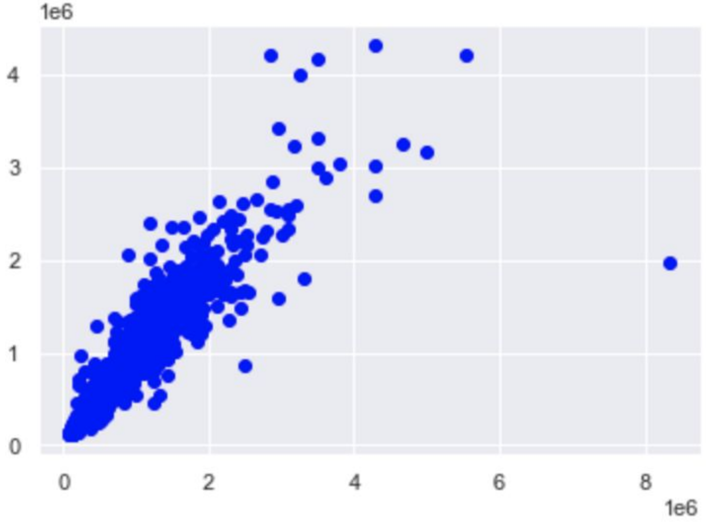
Key Takeaways:

- $R2 = 0.716$
- Adj. $R2 = 0.715$
- LR fits 72% of data

Model - Random Forest on Days of the Week

```
plt.scatter(Y_test, rand_pred_pca, c='blue')
```

```
<matplotlib.collections.PathCollection at 0x2c5bf4abd60>
```

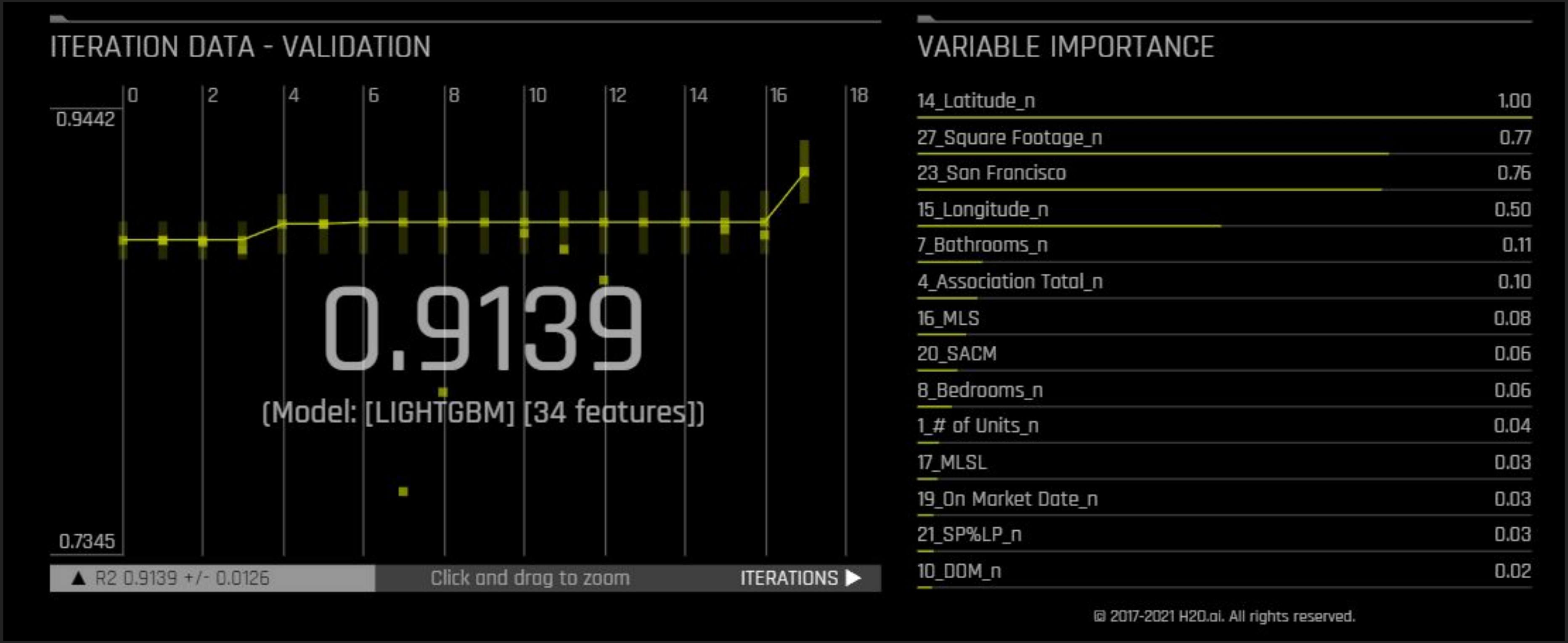


Key Takeaways:

- Train Score: 0.986
- Test Score: 0.869
- Prediction Accuracy 86%
- Higher Variance with prediction of higher Selling Prices

```
Best parameters: {}
Score on test set: 0.8687344674001293
MSE 30344061244.803455
Train score: 0.9862530715568794
Test score: 0.8687344674001293
```

H20.AI Results



Your Takeaway

- The price of a listing is heavily dependent on the location and square footage of the property.
- The day of the week a property is listed has a small, but noticeable, effect on the sale price, with Sunday being the best day of the week to list a property.

	coef	std err	t	P>|t|	[0.025	0.975]
Sunday	6.715e+04	1.25e+04	5.393	0.000	4.27e+04	9.16e+04
Tuesday	-2.254e+04	9097.542	-2.477	0.013	-4.04e+04	-4705.252
Wednesday	-3.717e+04	1.1e+04	-3.379	0.001	-5.87e+04	-1.56e+04
Thursday	-8.222e+04	1.4e+04	-5.858	0.000	-1.1e+05	-5.47e+04
Friday	-1.099e+05	1.79e+04	-6.133	0.000	-1.45e+05	-7.48e+04
Saturday	-1.028e+05	1.63e+04	-6.306	0.000	-1.35e+05	-7.09e+04

Our GitHub Links

Project's:

https://github.com/panderson-scu/FNCE-2431-Final-Project

Patrick's:

https://github.com/panderson-scu

Amanda's:

https://github.com/adelfino22/RealEstateProjections

Mario's:

https://github.com/mcmorales415/FNCE-2431-Final-Project

Thank You!

If you have further ideas, reach out at: panderson@scu.edu, adelfino@scu.edu, and mcmorales@scu.edu.