In a world where downtime is more wishful thinking than reality, it's essential that your home offers true tranquility, space and comfort.

Introducing the Fit Analysis for living:

House Hunters Evaluation

Designed to help you make the most efficient use of your time while searching for your dream home. Whether it's natural light, soaring ceilings, 8ft doors and enough space for you family; we're here to make your dreams come true.

House Hunters Presents: The Ultimate in Modern Home Search

House Hunters

- Thomas Clemons: Chairman and CEO
- Gi'Anna Cheairs: CFO and Comptroller
- Vineet Duggi: CTO
- Timothy Carter: COO

Executive Summary

- **Goal:** The aim of our project is to identify which attributes have the most impact on price and then predict the sale price of homes in Ames, IA.
- We'll examine relationships between various home attributes and the sale price.
- We will use this modeling to predict the sale price of homes that we have in our Real Estate portfolio.
- We identified an appropriate data file.
- We analyzed data file attributes, cleaned, and transformed them for modeling.
- We ran cleansed, transformed data file through a variety of regression models to determine best model for predicting sale prices.

Project Approach

1. **Data File**
   1. Read in CSV
   2. Reviewed Length
   3. Reviewed Type
2. **Data Preprocessing**
   1. Remove/Rename missing Values
   2. Train, test, split
   3. Categorical Variables - LabelEncoder
   4. Numeric variables - StandardScaler
   5. Variance Inflation Factor (VIF)
   6. Probability Value (p-value)
   7. Linear Regression, OLS
3. **Regression Analysis**
   1. Lasso, Random Forest, Gradient Boost and CatBoost
   2. Coefficient (R-Squared), Evaluate Mean Absolute Error (MAE) and Mean Squared Error (MSE)
4. **Best Model and Validation**
   1. CatBoost

Data Collection, Cleanup, and Exploration

Reviewed Value_Counts for all variables

Dropped variables with high percentage of same values or null values

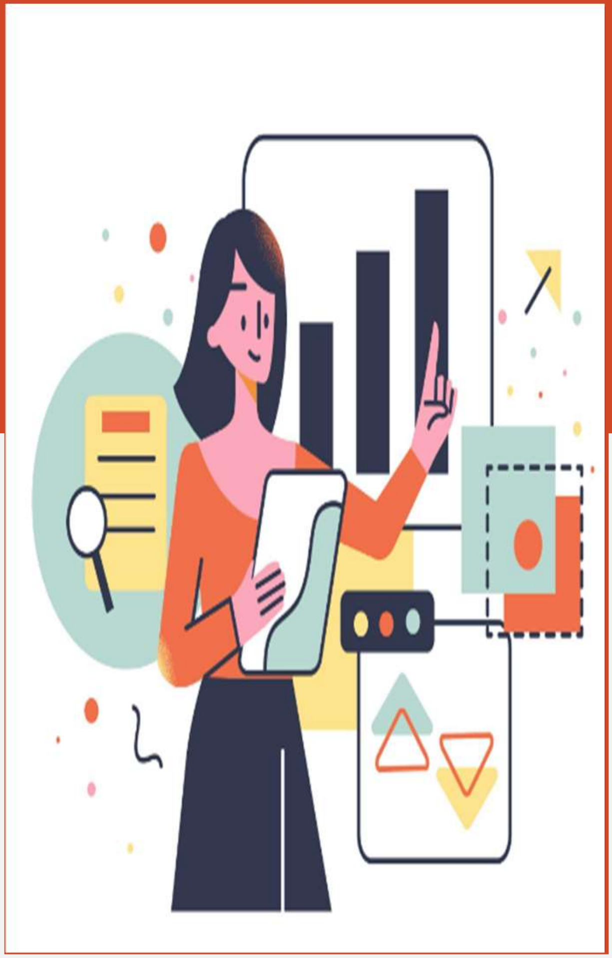
Transformed remaining null values based on analysis of home variables

Encoded categorical variables using LabelEncoder

Scaled numeric variables using StandardScaler

Performed VIF and P-Value analysis to identify any variables that should be removed (VIF > 10; p-value >= 0.05)

The home sales file was reduced from 81 to 24 variables



Data Collection, Cleanup, and Exploration - Examples

```
-------------------- Street START --------------------
Street
Pave     1454
Grvl        6
Name: count, dtype: int64
-------------------- Street END ----------------------


-------------------- Alley START --------------------
Alley
NaN      1369
Grvl       50
Pave       41
Name: count, dtype: int64
-------------------- Alley END ----------------------
```

value_counts

In [25]:
```
# Use loc to filter to columns with p-values below 0.05
select_cols = p_values.loc[p_values < 0.05]

# Show the index of the results
select_cols.index
```

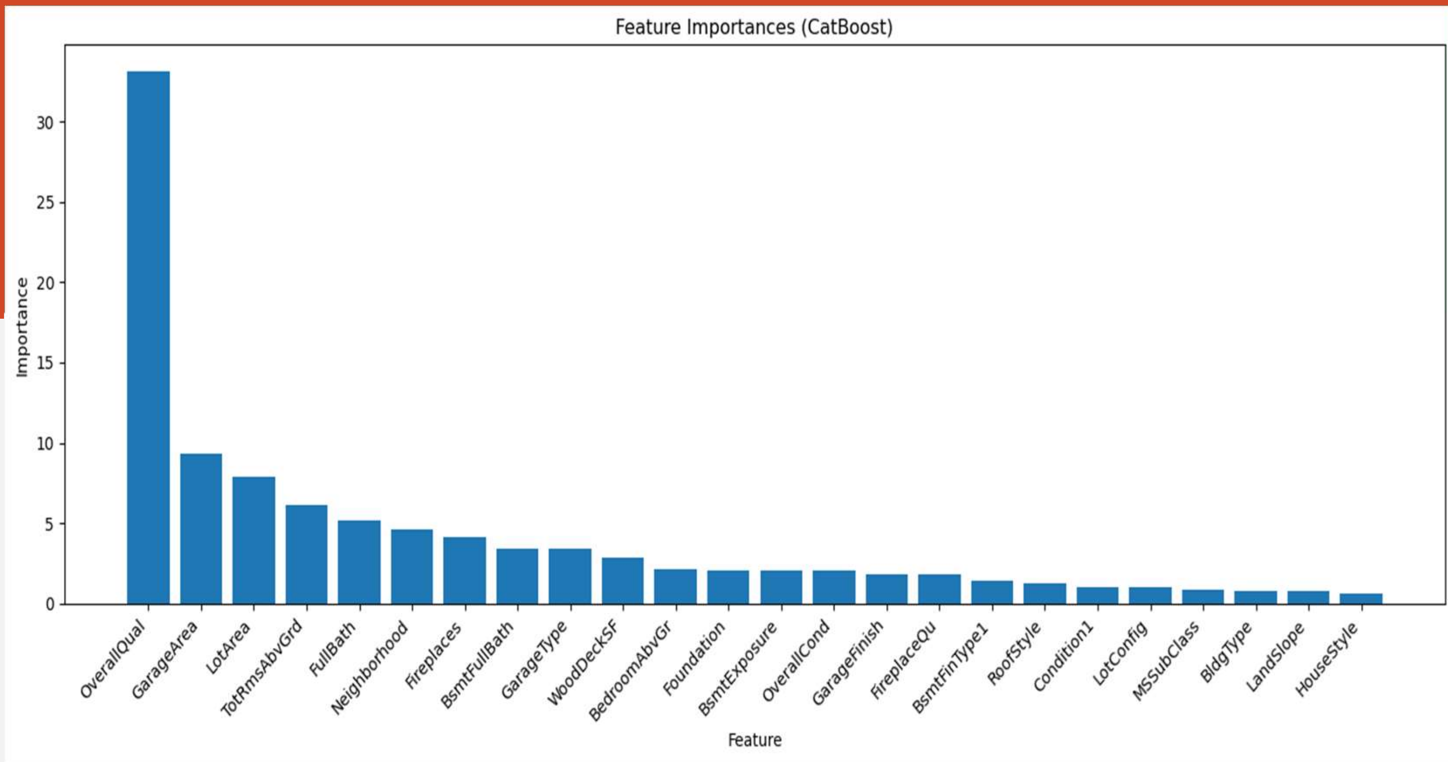
p_values

Out[25]:
```
Index(['OverallQual', 'Foundation', 'Neighborhood', 'BsmtFullBath',
       'MSSubClass', 'Condition1', 'RoofStyle', 'GarageType', 'TotRmsAbvGrd',
       'HouseStyle', 'GarageFinish', 'LotConfig', 'WoodDeckSF', 'LandSlope',
       'BldgType', 'FullBath', 'BsmtFinType1', 'LotArea', 'BsmtExposure',
       'OverallCond', 'FireplaceQu', 'GarageArea', 'BedroomAbvGr',
       'Fireplaces'],
      dtype='object')
```

VIF

13	YearBuilt	8.647230
20	Foundation	9.264507
26	BsmtUnfSF	10.417220
25	BsmtFinSF1	10.547944
43	GarageFinish	13.699053
27	TotalBsmtSF	14.348297
54	SaleCondition	15.793707
36	KitchenQual	18.648691
21	BsmtQual	25.134984
53	SaleType	27.531083
19	ExterCond	29.135861
22	BsmtCond	30.047846
1	MSZoning	31.254992
38	Functional	33.668657
17	Exterior2nd	36.454375
18	ExterQual	36.796721
16	Exterior1st	38.127610
29	1stFlrSF	77.061843
46	GarageQual	91.903634
30	2ndFlrSF	95.096910
47	GarageCond	102.141085
31	GrLivArea	133.668210

Price Impact - Features



Overall Quality, incorporating the latest technology of a home, is by far the most important feature for homebuyers.

Garage Space has also played an integral role for many.

One must not forget about Lot Area and Total Rooms.

Expansive Full Baths most certainly has its place on the list.

Sales price prediction model
CatBoost

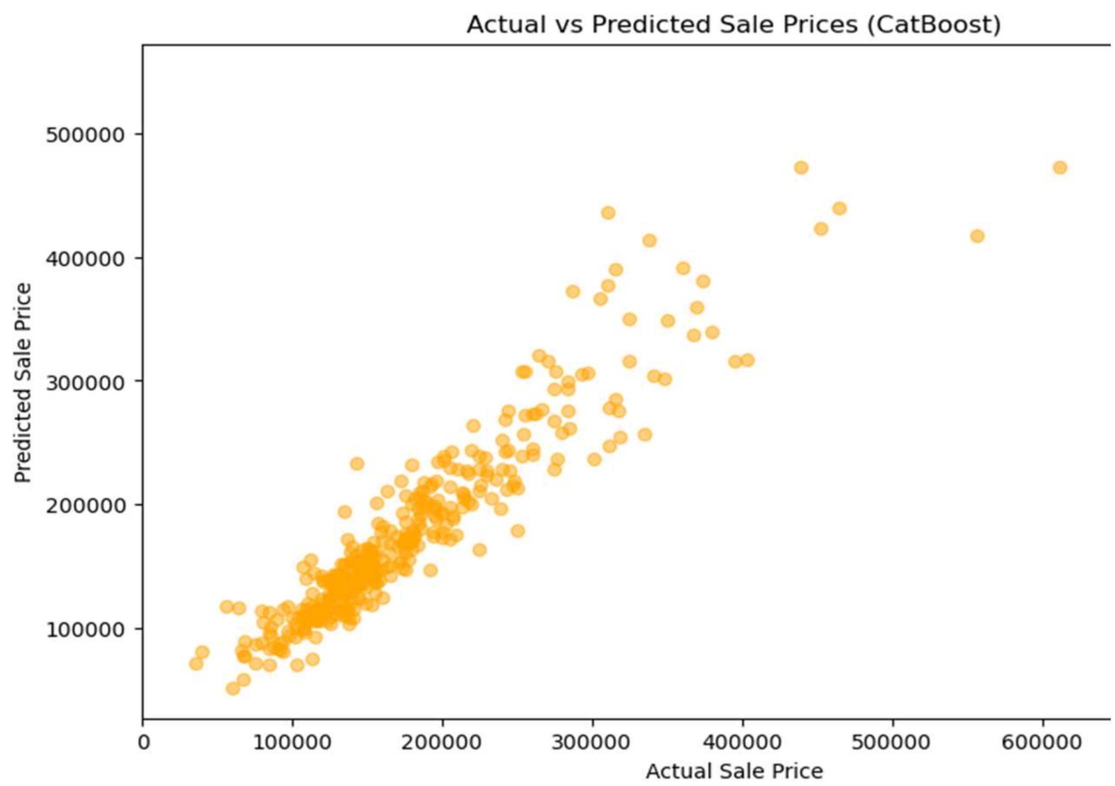
- Values Narrative

- R-squared: 0.8774344325281175

- Mean Absolute Error:
- 19291.19610825505

- Mean Squared Error:
- 858609679.3158845

- Root Mean Squared Error:
- 29302.04223797182



Conclusion

Lasso R-Squared - 0.7830943994223807

Random R-Squared - 0.8536607463238572

XGBoost R-Squared - 0.8553621589835805

CatBoost R-Squared - 0.8774344325281175

CatBoost in our data file provided the most accurate model.

Lasso as you can see did not meet or fulfill the project requirements.

We wanted to view several models to determine the best fit.

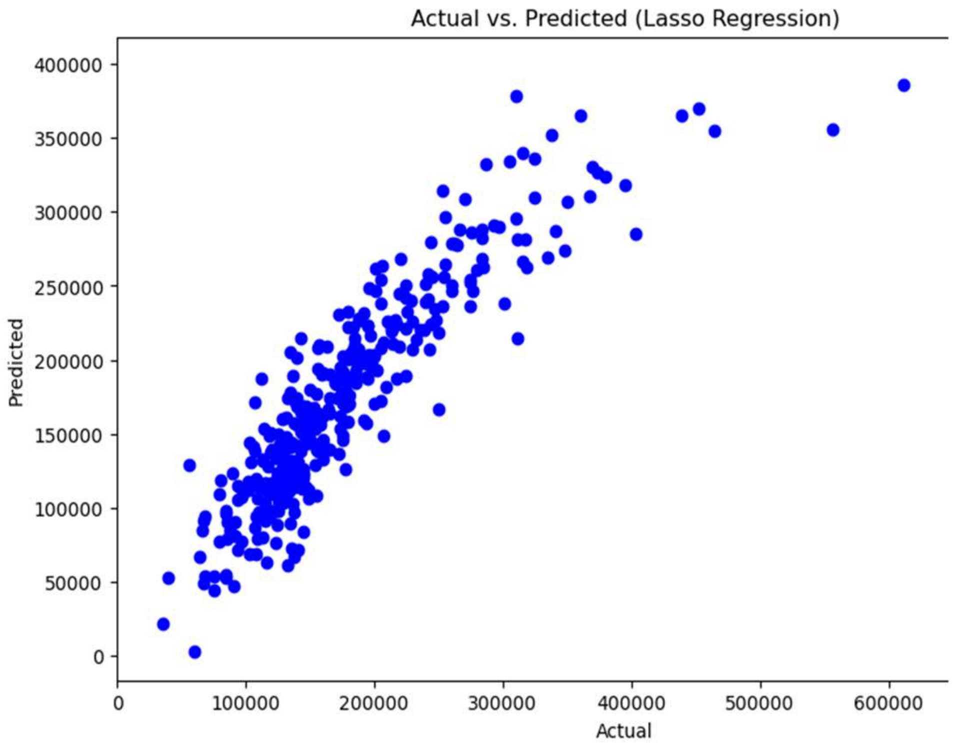
Additional Questions…

- Tune model to take better advantage of optimal home selling season.
- Refactor notebooks for pipeline processing.
- Bring in Real Estate SME to verify model predictions.
- Further dive into the statistical aspects of this model, (e.g. t-statistic).

APPENDIX

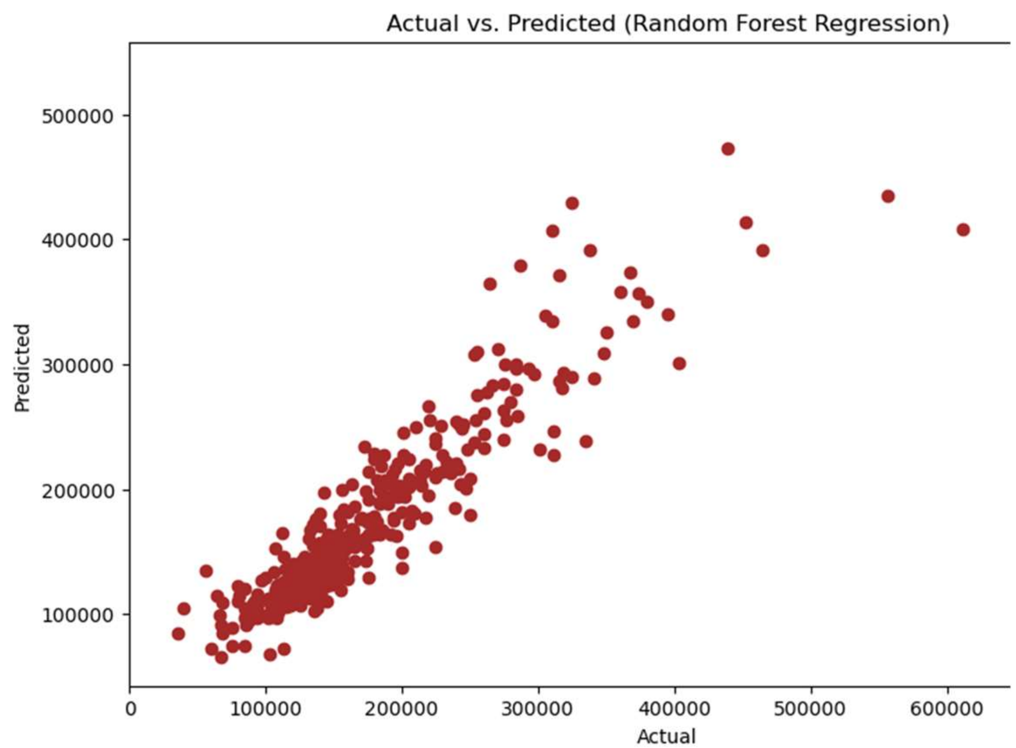
Lasso

- The Values Narrative
- R-squared: 0.7830943994223807
- Mean Absolute Error:
- 25173.89159828941
- Mean Squared Error:
- 1519490767.2294931
- Root Mean Squared Error: 38980.64605967291



Random Forest

- The Values Narrative
- R-squared: 0.8536607463238572
- Mean Absolute Error:
- 20753.749406392697
- Mean Squared Error:
- 1025151698.4900634
- Root Mean Squared Error: 32017.990231900305



XGBoost

- Values Narrative

- R-squared: 0.8553621589835805

- Mean Absolute Error:
- 21126.00980308219

- Mean Squared Error:
- 1013232776.9831389

- Root Mean Squared Error:
- 31831.317550223066