NYC Housing Price Prediction

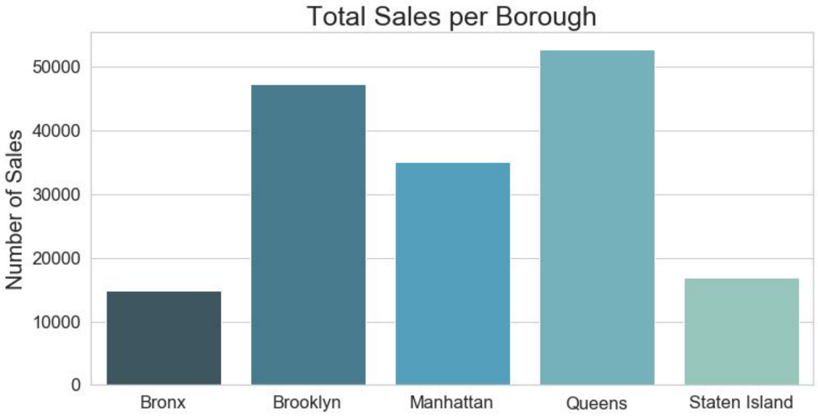
Alex Rudra, Kathleen Zhang, Pratyush Kundu, Nicole Ng

AGENDA

1. DATA EXPLORATION
2. MODEL APPROACH
3. MODEL EVALUATION
4. CONCLUSION

Our Dataset

166,968 properties sold in September 2016 to December 2018 from NYC Department of Finance



Bulk of sales in 3 boroughs

Highly skewed distribution a lot of outliers

Dealing with Messy Data

Instances with missing target value must be dropped

Drop instances for features with fewer missing values

40% of square footage values are missing, but we want to find a way to impute these values

Feature	Missing Values
LAND SQUARE FEET	63913
GROSS SQUARE FEET	58900
SALE PRICE	49741
TOTAL UNITS	38183
AGE	13267
RESIDENTIAL UNITS	49
COMMERCIAL UNITS	49
ZIP CODE	1
BLOCK	0
BOROUGH	0
BUILDING CLASS CATEGORY	0
Month	0
LOT	0
NEIGHBORHOOD	0
Year	0
TAX CLASS AT TIME OF SALE	0
30 Year Rate	0
15 Year Rate	0

Dealing with Categorical Data

Models can only interpret numerical data

<u>CATEGORICAL COLUMNS</u>

BLOCK
BOROUGH
LOT
NEIGHBORHOOD
BUILDING CLASS CATEGORY
TAX CLASS AT TIME OF SALE
YEAR

Borough
Manhattan
Queens
Queens
Brooklyn
Bronx

One-hot Encoding

Manhattan	Queens	Brooklyn	Bronx
1	0	0	0
0	1	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Linear Regression: Assumptions

- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence**: Observations are independent of each other.
- **Normality**: For any fixed value of X, Y is normally distributed.

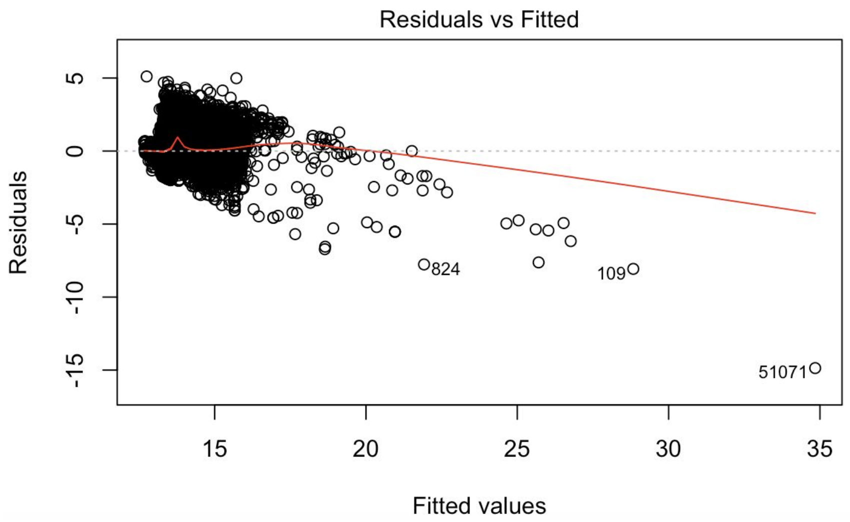
Linear Regression: Multicollinearity

Variable	VIF
Commercial Units	30.425479
Residential Units	83.192938
15 Year Rate	38.222793
30 Year Rate	**38.146073**
Total Units	**112.900570**

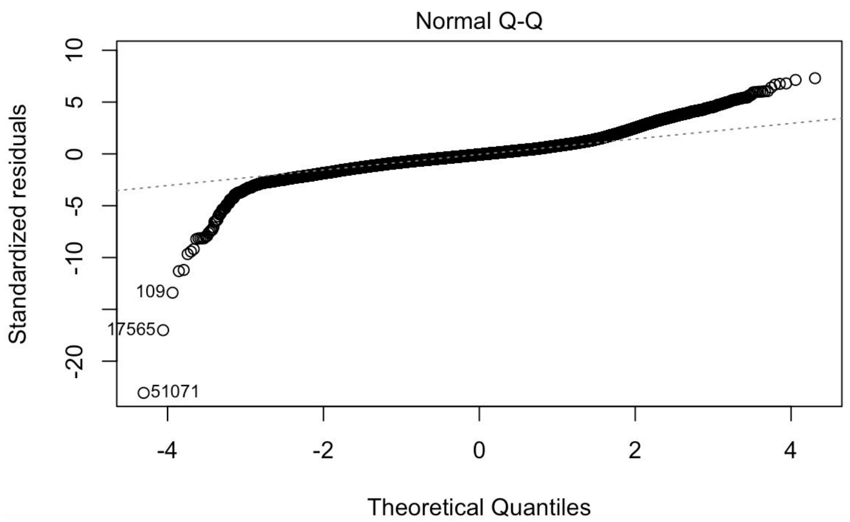
Variable	VIF
Gross Square Feet	2.218817
Land Square Feet	1.301346
Age	0.163727
Zip Code	~1

Removed 30 Year Rate and Total Units, reducing VIF in the remaining variables to be < 10

Linear Regression: Diagnostic Plots

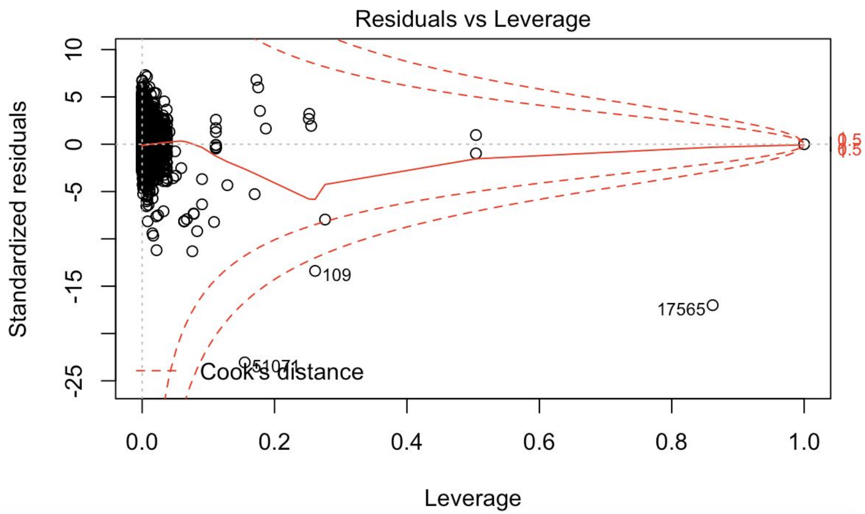


Non-Constant Variance
(Heteroscedasticity)
Possible Non-Linear Relationship



Highest and Lowest Quantiles
not normally distributed

Linear Regression: Diagnostic Plots



Rightmost points have high leverage
Points far from 0 may be outliers

Linear Regression: Model

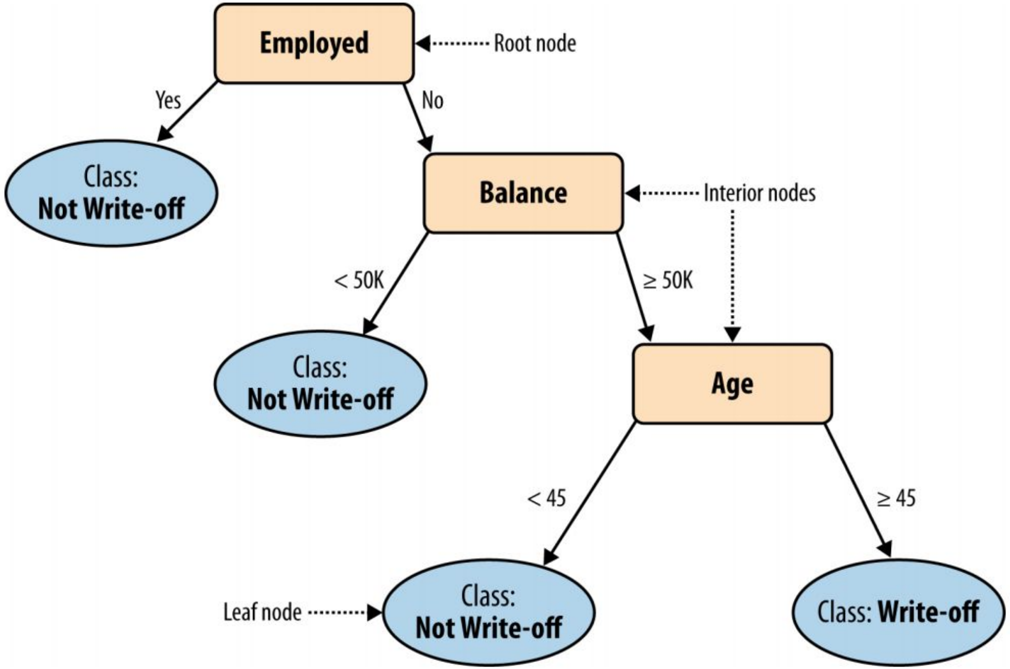
Formula

log(Sale Price) ~ Gross Square Feet + Land Square Feet + Residential Units + Commercial Units + 15 Year Rate + Age + Zip Code (One-Hot Encoded)

R-Squared: 0.3565

Variable	P-Values
Gross Square Feet	<0.05
Land Square Feet	<0.05
Residential Units	<0.05
Commercial Units	<0.05
15 Year Rate	<0.05
Age	<0.05
Zip Code	<0.05

Decision Tree



Decision Tree: Methodology

- Instances with missing values were dropped.
- Categorical Features with too many variables were dropped.
- Hyperparameter Tuning using Grid Search

LAND SQUARE FEET	63913
GROSS SQUARE FEET	58900
SALE PRICE	49741
TOTAL UNITS	38183
AGE	13267
RESIDENTIAL UNITS	49
COMMERCIAL UNITS	49
ZIP CODE	1
BLOCK	0
BOROUGH	0
BUILDING CLASS CATEGORY	0
Month	0
LOT	0
NEIGHBORHOOD	0
Year	0
TAX CLASS AT TIME OF SALE	0
30 Year Rate	0
15 Year Rate	0

Decision Tree : Hyperparameter Tuning

Hyperparameter tuning was performed using Grid Search with 4-fold cross validation.

Hyperparameters used (with values):

- Max Depth : 5,10,15,20
- Max Leaf Nodes : 3,5,7,10,100,1000,100000
- Min Impurity Decrease : 0.1, 0.2, 0.3, 0.4, 0.5, 0.6

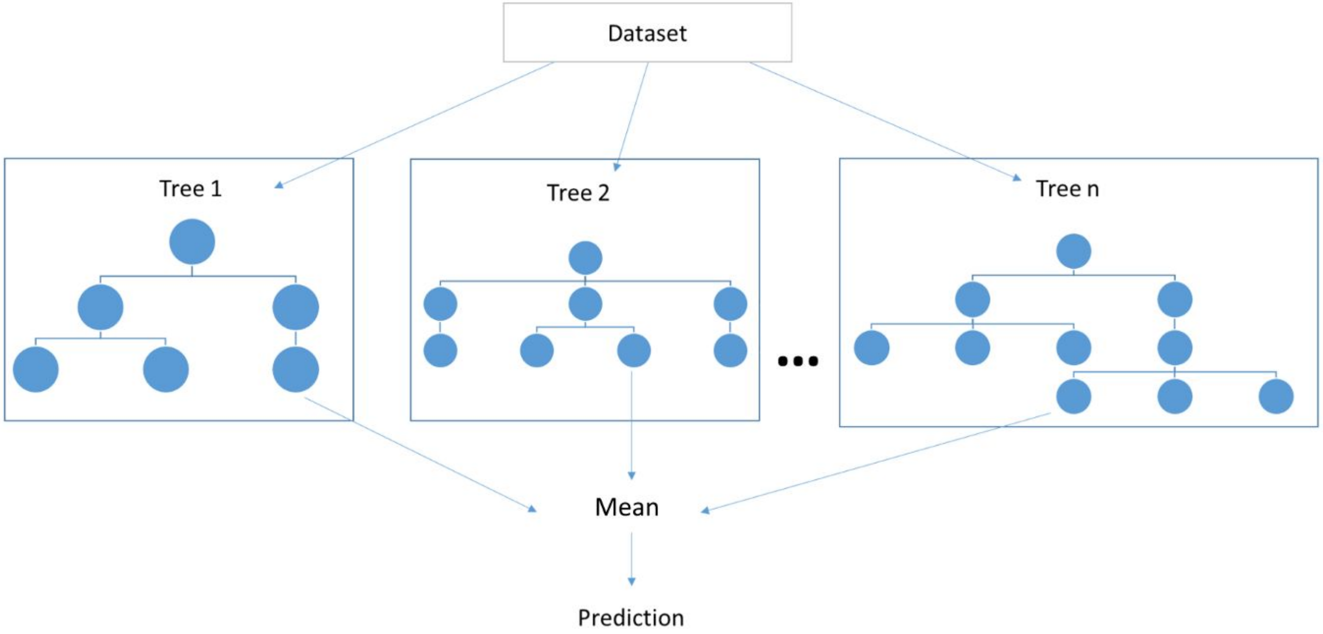
Decision Tree: Accuracy and Evaluation

Holdout Accuracy at each stage of the model:

1. At the beginning, without any changes to the dataset : 22%
2. After dropping null values and instances with missing values : 28%
3. After performing Grid Search : 40%

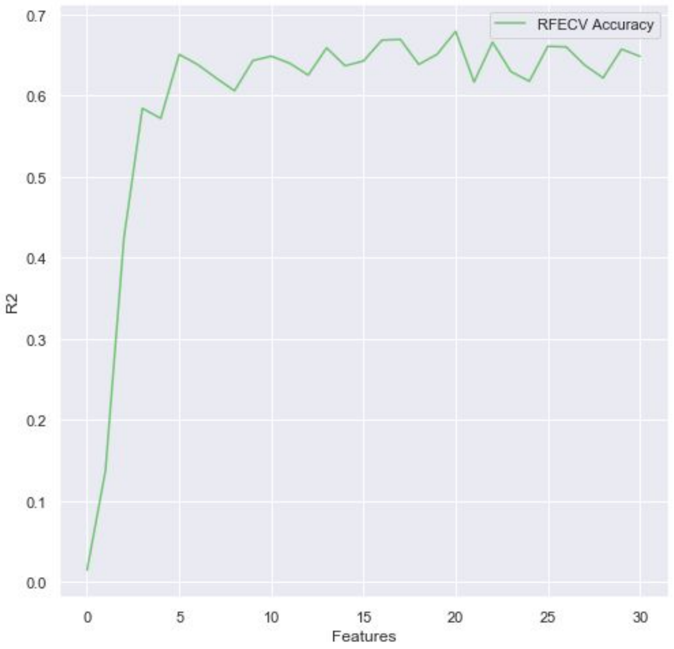
Verdict: The model does not perform satisfactorily on the dataset.

Random Forest



Random Forest : Methodology

- Rows with missing values were dropped
- Hyperparameter Tuning using Grid Search
- Complexity control using the mean test scores from Grid Search
- Features were selected through RFECV and Feature Importances



Random Forest : Hyperparameter Tuning

Grid Search with 3-fold Cross Validation was used.

Hyperparameters selected:

- Max_features : auto
- N_estimators : 10
- Min_samples_leaf : 5
- Min_samples_split : 8
- Bootstrap : True
- Max_depth : 10

Random Forest : Feature Selection

- RFECV : Recursive Feature Elimination with Cross Validation
- Number of Features selected : 20
- Feature Importances

Model achieves holdout accuracy of 77%

Verdict: Best performing model

Feature	Importance
GROSS SQUARE FEET	0.903005
BOROUGH_Manhattan	0.026631
COMMERCIAL UNITS	0.019930
AGE	0.018885
30 Year Rate	0.010933
RESIDENTIAL UNITS	0.007118
15 Year Rate	0.006853
ZIP CODE_11201.0	0.001962
TAX CLASS AT TIME OF SALE_Class 3	0.001782
BOROUGH_Brooklyn	0.000500
BOROUGH_Bronx	0.000285
TAX CLASS AT TIME OF SALE_Class 2	0.000198
ZIP CODE_11101.0	0.000187
ZIP CODE_10022.0	0.000175
ZIP CODE_10012.0	0.000171

IMPUTATION OF DATA

How to deal with lots of missing square footage values?

1. Created model to predict square footage
2. Used Linear Regression, Decision Trees, and Random Forest models
3. Imputation model accuracy of 83.86% with Random Forest
4. With imputed values, accuracy on sales price model decreased to 48%

Verdict: Square footage model accuracy handicaps the overall model

Takeaway + Prediction

Overall best model was the Random Forest

Used to predict on out-of-sample data:

```
# Model prediction on a property w/ Effective Market Value $804,533
model.predict([[4, 1323, 79, 3.0, 0.0, 3.0, 1125.0, 3240.0, 68.0, 1, 3.5
```

Predicted Value: **$1,029,660**

```
# Model prediction on a property w/ Effective Market Value $904,066
model.predict([[4, 1336, 72, 3.0, 0.0, 3.0, 2280.0, 3430.0, 57.0, 1, 3.5
```

Predicted Value: **$1,053,319**

Conclusion + Further Analysis

Tool for real estate investors and homeowners to check property valuations

Ensure accurate property tax and asset management

Identify good deals for primary residence or investment

Model predictions were very close to neighborhood median -- perhaps should do neighborhood analysis rather than specific properties

Thank you,
Questions?