Zillow's Home Value Prediction Challenge

Harish Pullagurla: (hpullag@ncsu.edu)
Pooja Mehta: (pmehta@ncsu.edu)
Ratika Kapoor: (rkapoor3@ncsu.edu)

Importance & Conventional Methods

- Helps in prediction of value of a house for future sale
- Key to prospective homeowners, real estate participants
- Important in real estate transactions such as loans

- Conventional methods: involved professional appraisers
  - Disadvantage:
    - Vested interest from the seller, buyer, mortgage broker or lender
    - Biased estimate of the house price
- Led to creation of automated house prediction systems
  - Independent of bias
  - More accurate prediction

Related Work

- *Park et.al* [1]: Analysis of townhomes in FairFax County, VA
  - RIPPER outperformed other house prediction models such as C4.5, Bayesian and AdaBoost.
  - Housing markets influenced by macroeconomic variables
- *Pow et.al* [2]: Predicted selling prices of properties using Support Vector Regressor (SVR), KNN and Random Forest Regression
  - KNN and Random Forest Regression performed better than linear regression and linear SVR
  - Ensembling KNN with Regression Forest Regression improved prediction

[1] Park.B & Jae.K, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,(2016) *Expert Systems with Applications*
[2] Pow. N, Janulewicz. E & Liu. D, Prediction of Real Estate Property Prices in Montreal

Zillow® Home Value Prediction (Zestimate)

- Zestimate: house prediction model by Zillow
  - Computed using a proprietary formula
  - Based on public and user-submitted data
  - Takes into account special features, location and market conditions
- Zillow Challenge : an open competition on Kaggle
  - develop a model to improve the Zestimate residual error

$$logerror = log(Zestimate) - log(Sale\ Price)$$

Data Set Description

3 Million Data Points

58 Dim Vector Each

Train Set

3 - Counties
Los Angles
Ventura
Orange

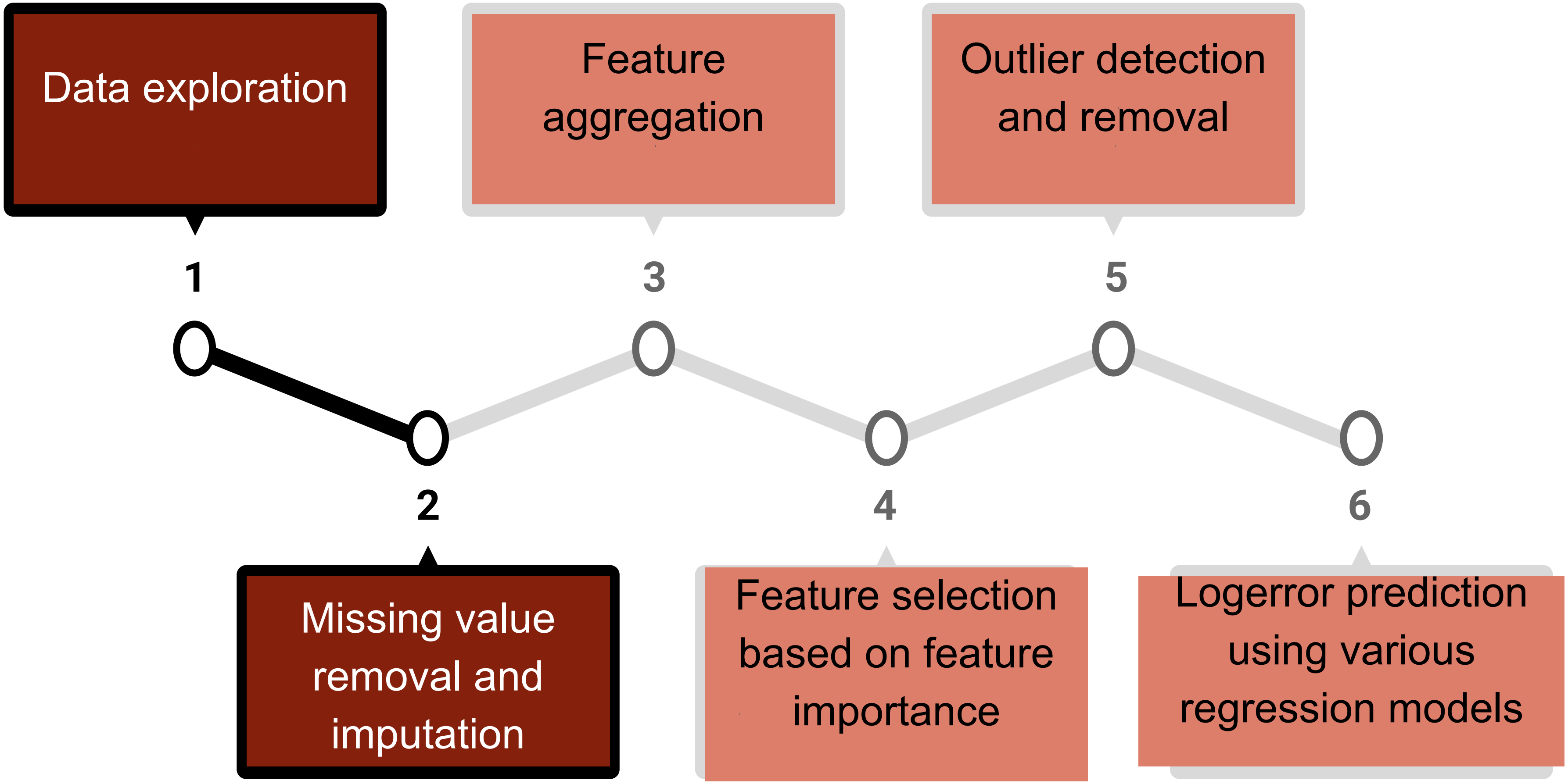
170,000 training samples with log errors

House Attributes - 38 dim

Location Attributes - 13 dim

Tax Info Attributes - 7 dim

Pipeline



1

Data exploration

2

Missing value removal and imputation

3

Feature aggregation

4

Feature selection based on feature importance

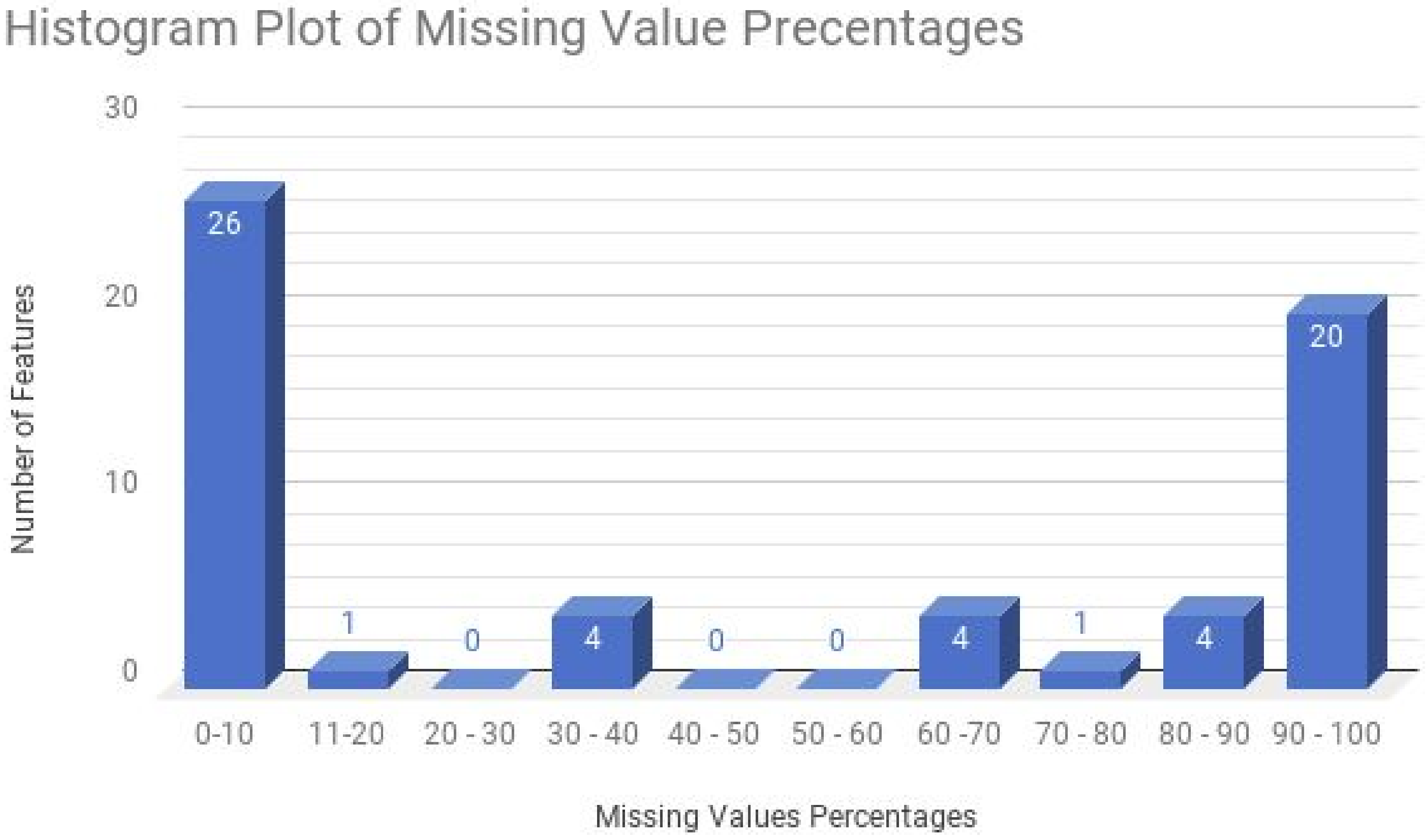
5

Outlier detection and removal

6

Logerror prediction using various regression models

Missing Values



Data Imputation

- Imputation of missing values based on
  - Most frequently occurring data, correlation between properties
  - Nearest neighbour derived from latitude and longitude
- Eg: imputation for ‘unitcnt’ - used to define number of stories

Step 1

‘numberofstories’ used to impute ‘unitcnt’ where respective ‘unitcnt’ values were missing‘

Step 2

‘PropertyLandUseTypeID’ used to impute further missing values of ‘unitcnt’.information could be derived from the type of property (eg. duplex, triplex)

Step 3

Imputation of remaining missing values in ‘unitcnt’ was done considering the mode of the respective property type.

- Eg: imputation for ‘regionidcity’ and ‘regionidzip’

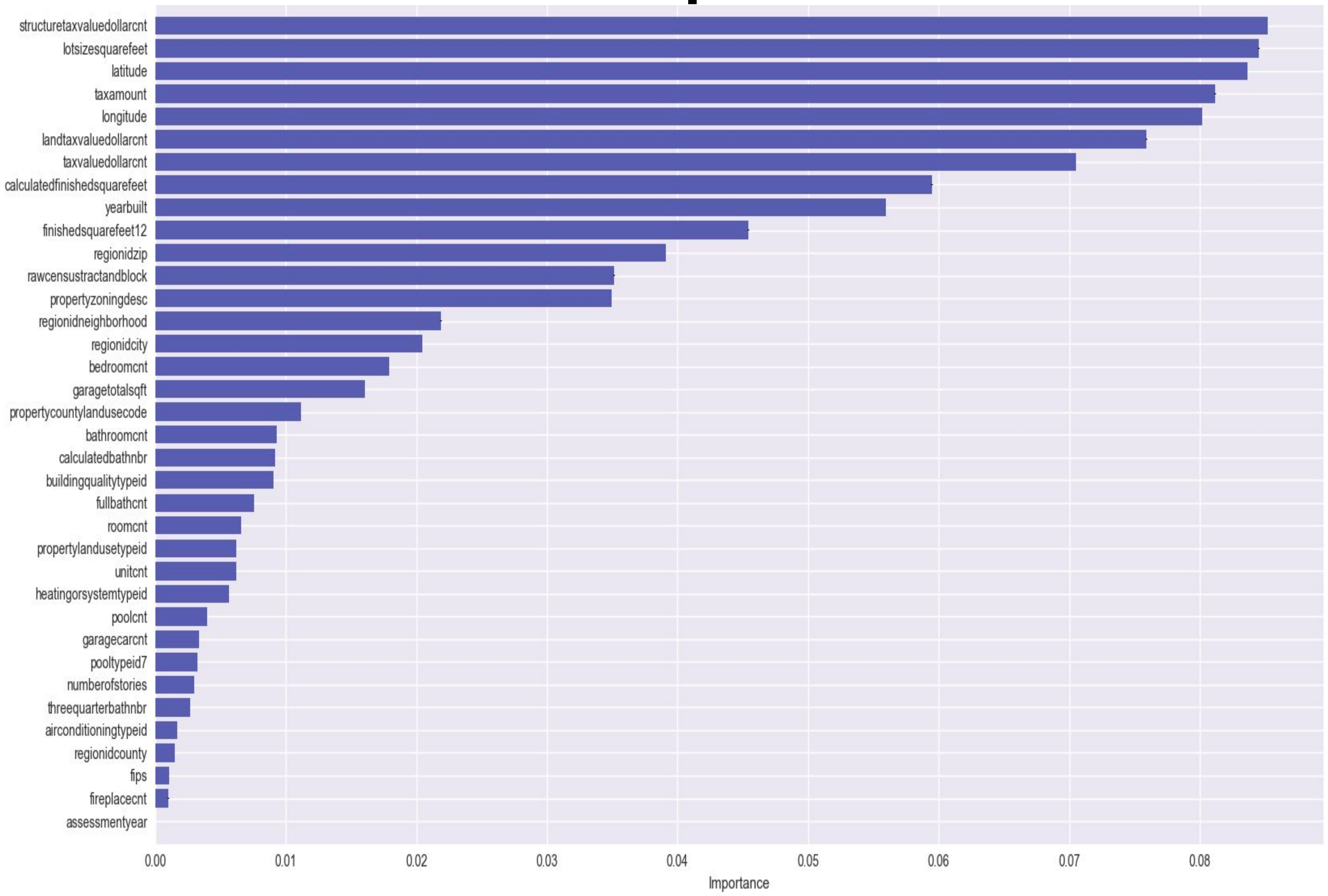
Step 1

KNN was used to find the nearest neighbour by using latitude and longitude for properties with missing ‘regionidcity’ and ‘regionidzip’

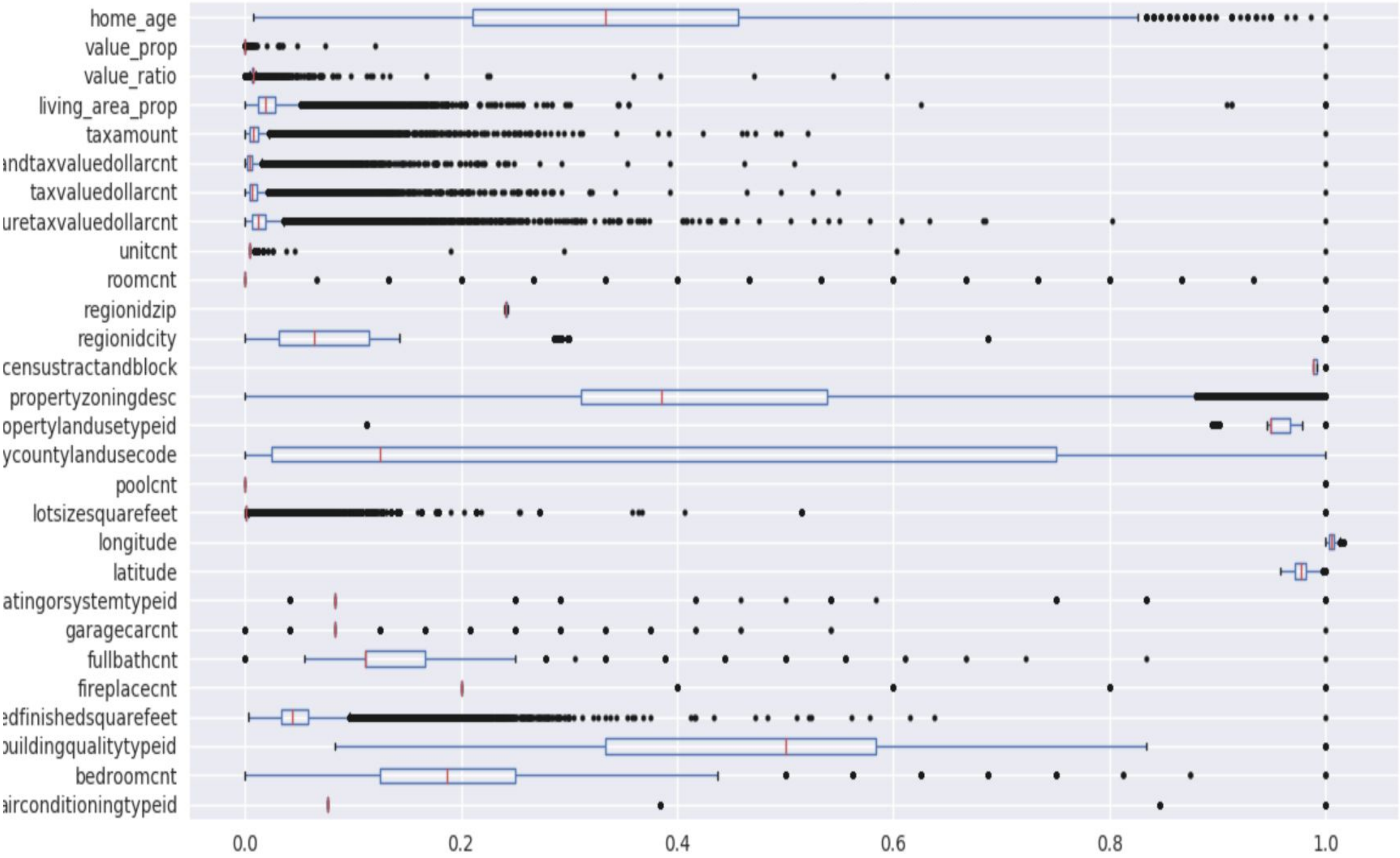
Step 2

‘Regionidcity’ and ‘regionidzip’ of missing properties could be imputed based on the city and zip of their nearest neighbour

Feature Importance



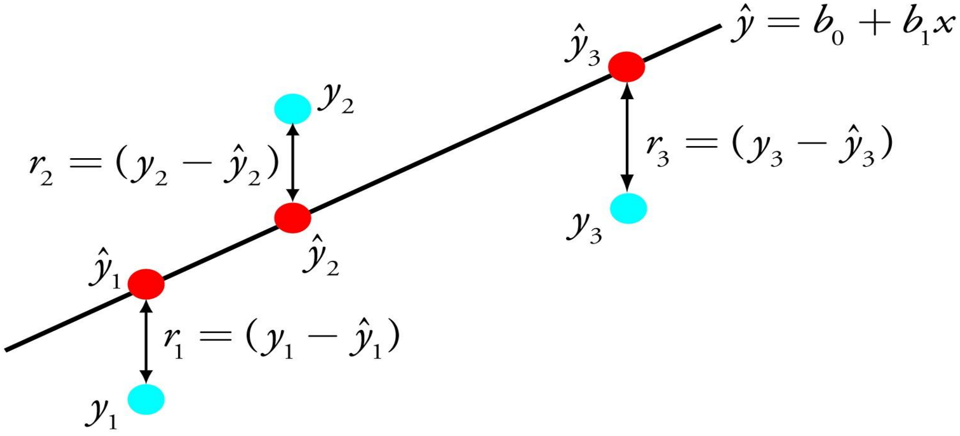
Outlier Detection and Removal



Model 1 - Lasso Regression

- Supervised machine learning algorithm for predictive analytics
- Uses regularized linear regression model
- Method to predict a target variable
  - Shrinks the parameters; used to prevent multicollinearity

    $Y = a_o + a_1X_1 + a_2X_2 + a_3X_3$

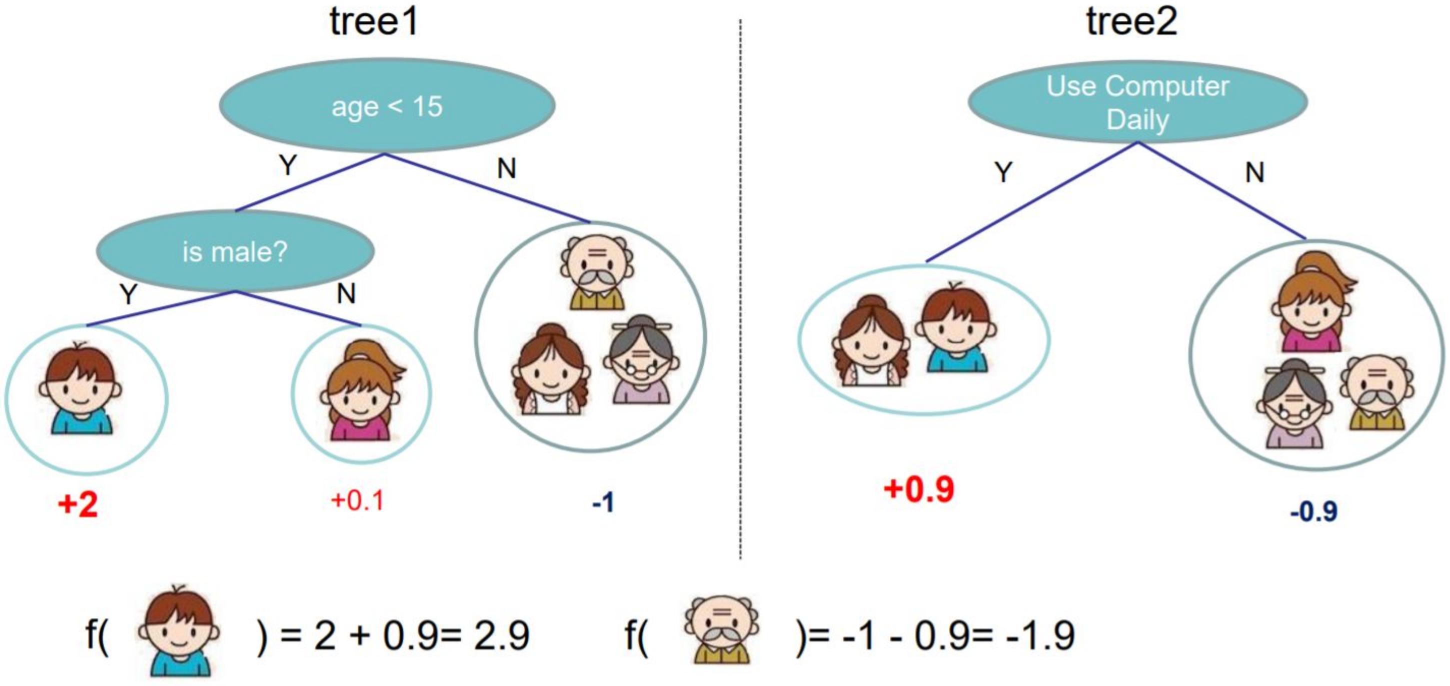

$\hat{y} = b_0 + b_1 x$
$\hat{y}_3$
y_2
$r_3 = (y_3 - \hat{y}_3)$
$r_2 = (y_2 - \hat{y}_2)$
y_3
$\hat{y}_2$
$\hat{y}_1$
$r_1 = (y_1 - \hat{y}_1)$
y_1

Model 2 - Random Forests Regression

- Is an ensemble algorithm that combines multiple Regression Trees (RTs)
- Additive model that makes predictions by combining decisions from a sequence of base models
- RT is trained using a random subset of the features, and the output is the average of the individual RTs

Model 3 - Xgboost

- Based on gradient boosting trees: similar to random forests but training the model is different
- Uses regression tree: CART
  - Decision rules same as in decision tree
  - Contains one score in each leaf value



Prediction of is sum of scores predicted by each of the tree

Results and Conclusion

Used R-Square and MAE to evaluate model performance

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \qquad \text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

ML Model	Hyper Parameters used	R squared	MAE
Lasso Regression	Alpla	1.000021	0.070792
Random Forests	Number of trees in forest	1.0000163	0.0707384
Xgboost	Eta, max_depth	**1.00**	**0.0706256**

- Xgboost performs best with minimum MAE: reduces bias and variance by combining outputs
- Feature selection for understanding their relative importance for predicting a house price are key to any model's forecast.

Future Work

- Include temporal information to predict price at different time points
- Come up with innovative imputation techniques
- Apply and combine different ML techniques (e.g., SVR, Clustering, NN)

Project Learnings

- Understood real life(big) data and how it is used
- Problem of missing values & methods of dealing with it
- ML model is a very small part of complete pipeline, feature engineering plays an important role
- Got introduced to data science packages like pandas and scipy