Sberbank Presentation

Fouad, John, Jason, Chris

Correcting bad data

Examples:

kitch_sq > full_sq, kitch_sq== 0 | kitch_sq== 1, kitch_sq should’ve been built_year

Life_sq == 0 | life_sq == 1 as well as full_sq == 0 or 1

Full_sq values too high (divided by 10)

Build_year (20052009, 4965, values below 1800)

State (categorical from 1 to 4, 33 found and converted to 3)

Cleaning test set

Number of rooms between 1 and 10

Imputation

- MAR compared to MNAR
  - Missingness related to time, related to other variables
  - Patterns of missingness: build_count_(by year, by material type), num_room & kitch_sq (9572)
- MICE or KNN
- Median, mean by Sub Area
  - All features with missingness under 10%
  - Variables with low correlation

#(Let’s talk about missingness based on time, sub_area)

Modeling - Feature Selection

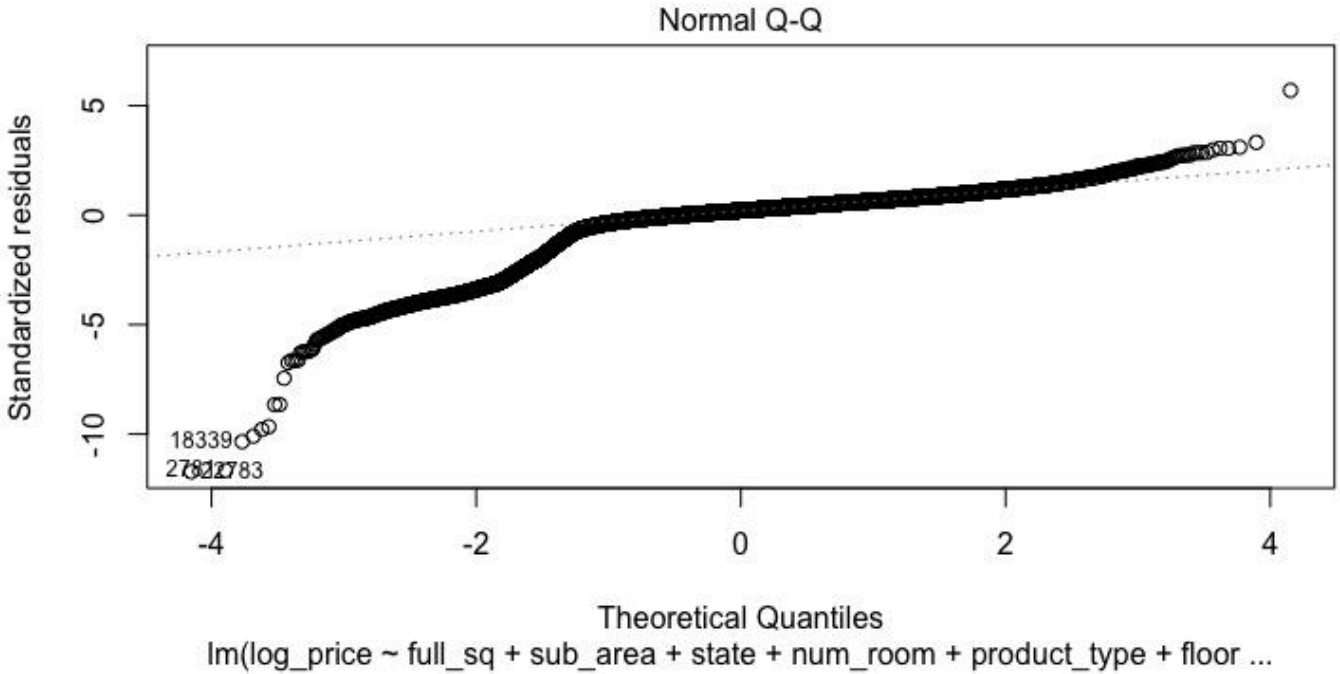
Feature Selection

- Bottom-up approach
  - Start with a subset of variables
  - Dependent variable: log(price)
- Feature Engineering
  - PCA (distance to transportation)
  - Clustering (presence of risk factors)
- Feature Selection
  - Least Absolute Shrinkage and Selection Operator Regression (Lasso)
  - L1 shrinkage penalty (Bs toward 0)
  - Cross validation to identify shrinkage penalty

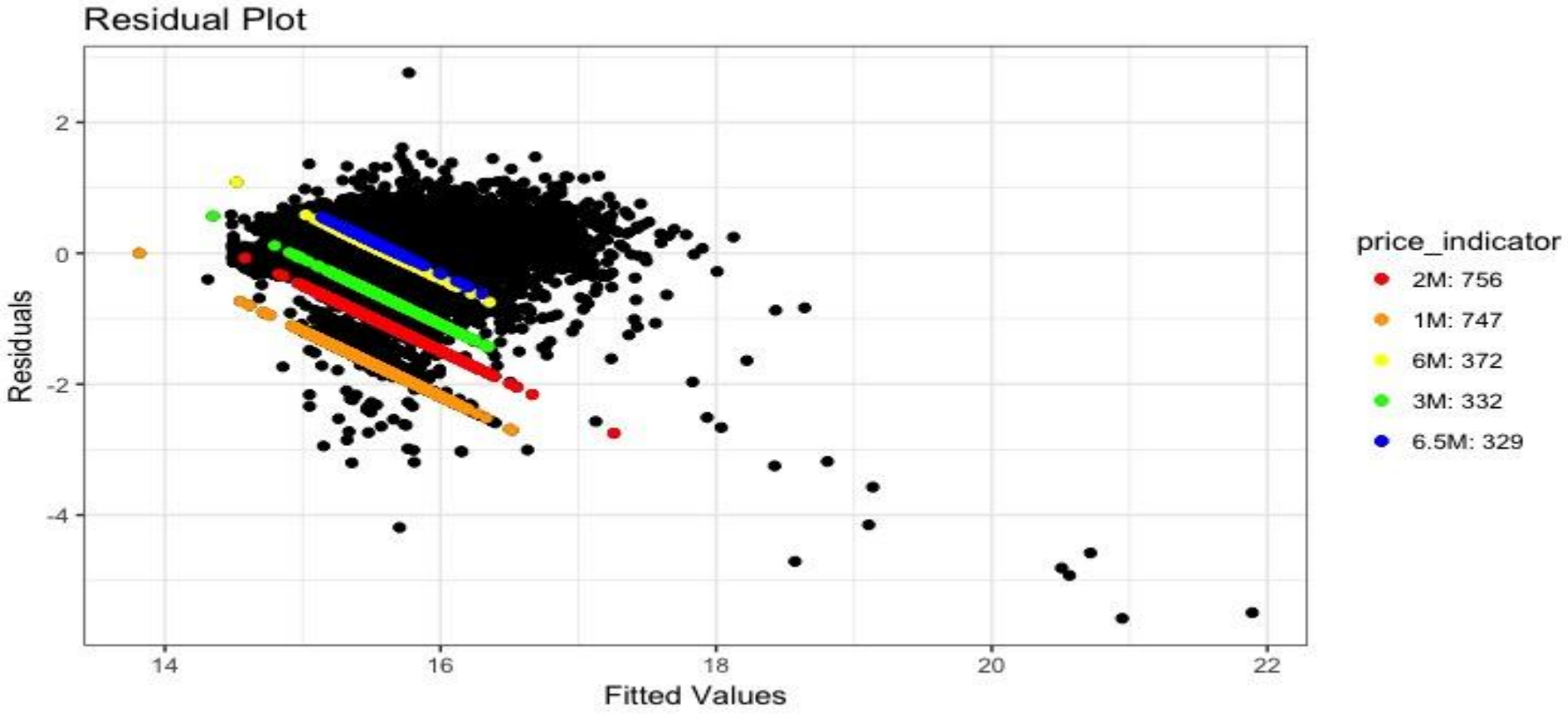
Multiple Linear Regression

- Best Subset Selection
  - Forward Selection
- Evaluation
  - R2
  - AIC
  - Training set RMLSE calculation
- Submission
  - Test set RMLSE calculation
  - Trap of Kaggle
  - Assumptions?

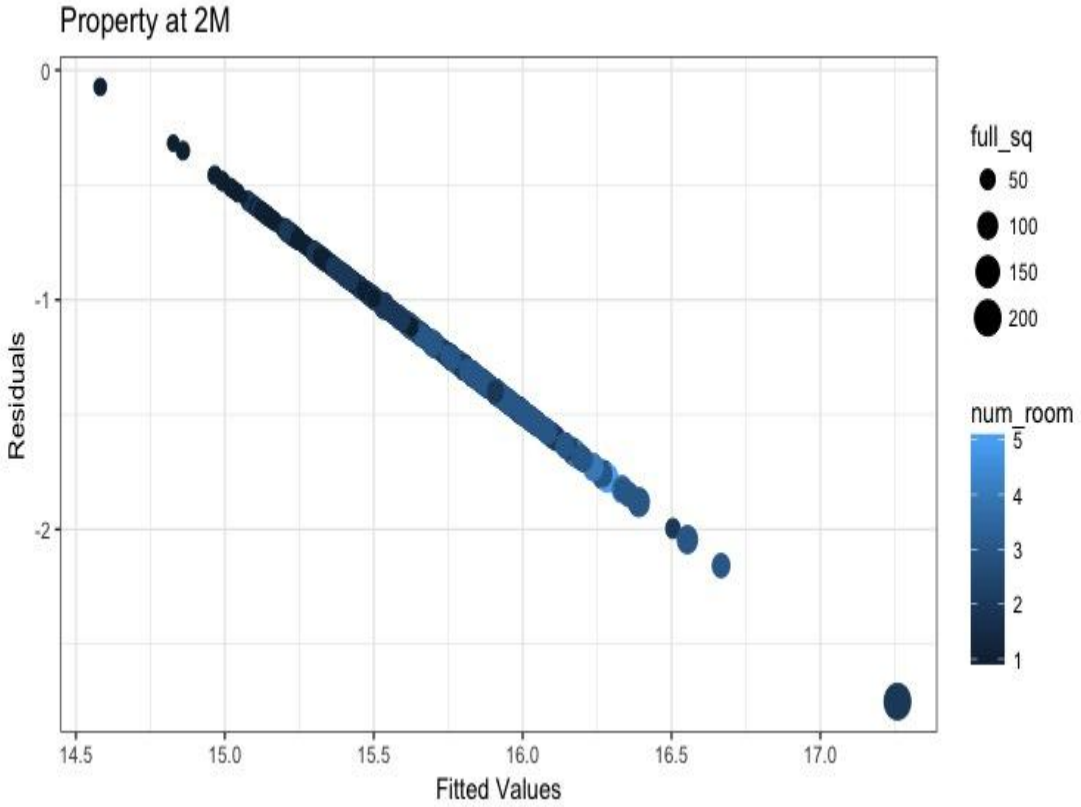
Assumptions?



Assumptions Violation

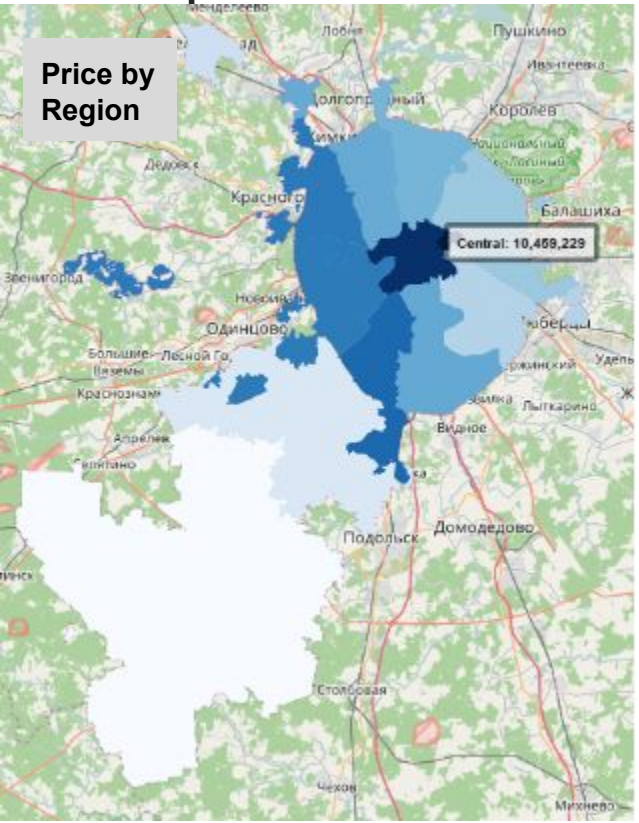


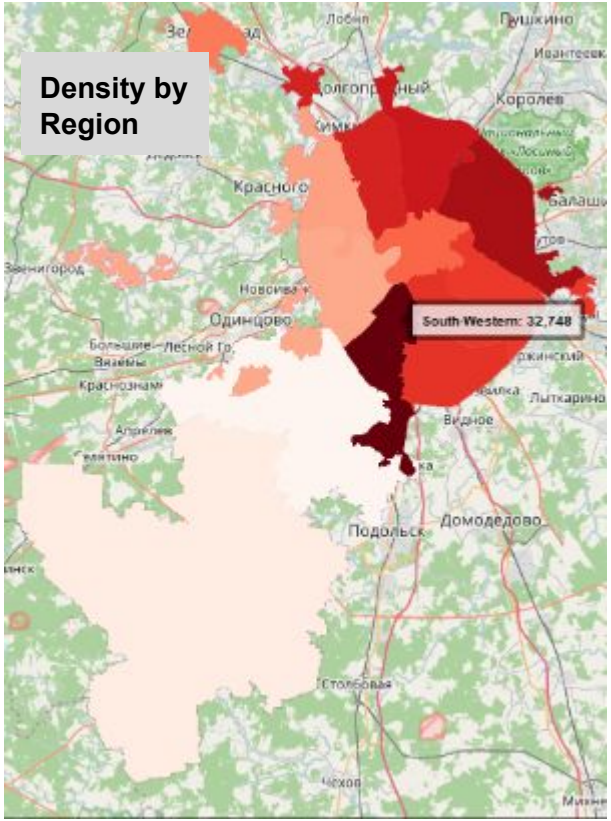
Property at 2M

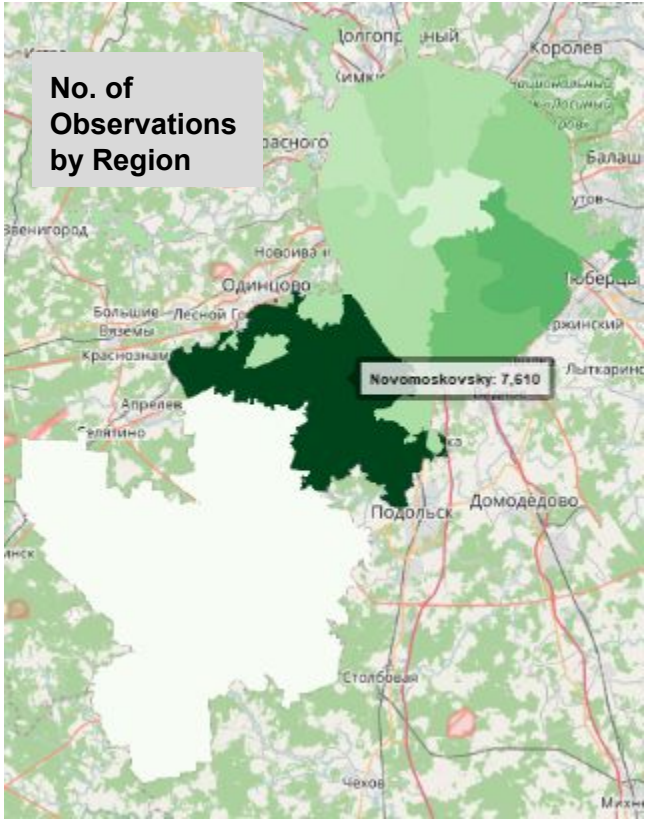


- Violating the basic assumptions of linear regression:
- Especially for properties with similar price and different characteristics
    - Linear relationship might not apply

Maps


Price by Region
Central: 10,459,229
Пушкино
Королев
Балашиха
Одинцово
Подольск
Домодедово
Видное
Лыткарино
Чехов


Density by Region
South Western: 32,748
Пушкино
Королев
Одинцово
Подольск
Домодедово
Видное
Лыткарино
Чехов


No. of Observations by Region
Novomoskovsky: 7,610
Королев
Одинцово
Подольск
Домодедово
Чехов

Tree Model - Random Forest

- Goal :

  Trying to get rid of the effect of strong predictor.

  Investigate nonlinear pattern from the data set.

  Feature Selection

- Bottom-up approach

  Start with a subset of variables

Ramdom Forest- R V.S. Python

Variables: full_sq, life_sq, floor ,num_room,

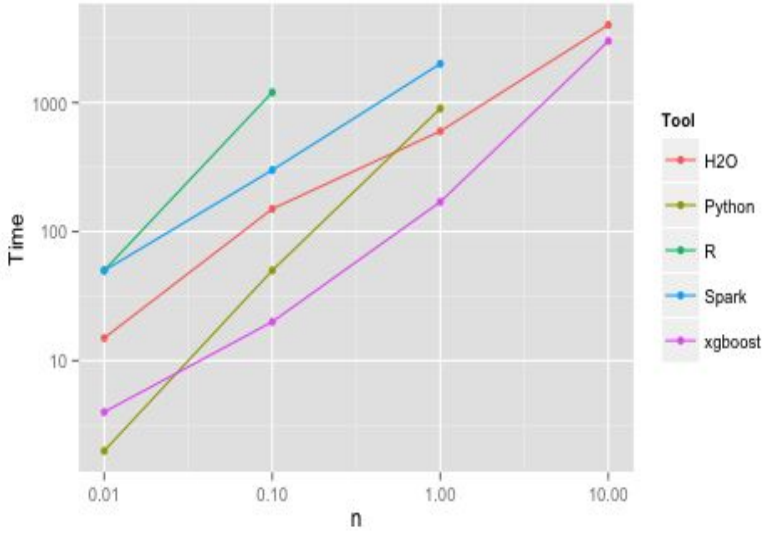
product_type, sub area

10 fold Cross-Validated

Rsquared 0.6172676

Processing time > 6 hr

Score :0.34



Random Forest : Region V.S. Sub_area