



Predicting Russian Housing Market Prices

A CSCI 271 Data Mining Project by:

AQUINO, Alec
ARROBIO, Anne
VILLAREAL, Rosiel



Outline

Introduction

Methodology

Pre-processing

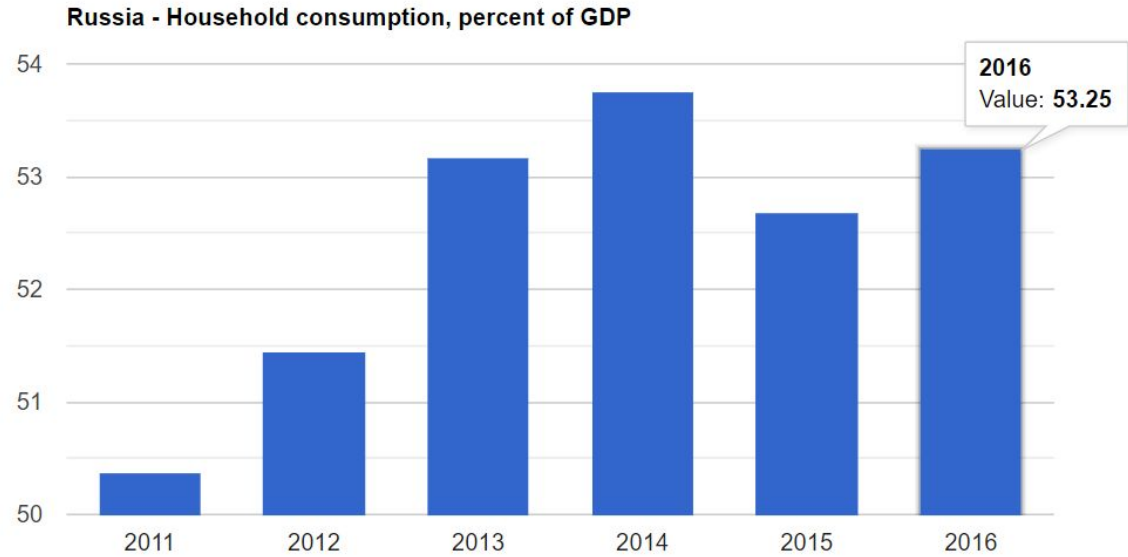
Exploratory Data Analysis

Regression & Evaluation

Insights

Conclusion

Household spending accounts for more than half of GDP, so changes in household spending will also affect the economy.




Introduction

Methodology

Insights

Conclusion



One factor that affects household consumption is their amount of debt, which is largely due to mortgages. When households have large amounts of debt, they might suddenly hold back on spending because they worry about repaying their debts.




Mortgage lending is the main activity for banks, and it carries a lot of risk. If house prices go down and the borrower defaults, the bank will lose money if the house is worth less than the loan.

Introduction

Methodology

Insights

Conclusion



How might we predict house prices accurately to reduce financial risks for banks and give more certainty for borrowers?

Introduction

Methodology

Insights

Conclusion



We want to know...

What housing and neighborhood characteristics affect house prices?

What macroeconomic and financial characteristics affect house prices?

...so that we can

Adjust mortgage interest rates

**Limit proportion of borrowers based on ratio of
loan value to house value**

Introduction

Methodology

Insights

Conclusion



Dataset

30,471 real estate transactions from August 2011 to June 2015 with 390 features representing:

1. Characteristics of the property itself
2. Characteristics of the local area surrounding the property
3. Macroeconomic conditions in Russia



Total area
Living area
Number of floors
Wall material
Year built
Number of rooms
Kitchen area
District



Number of cafes & restaurants
Number of markets
Number of schools
Share of industrial zones
Time to nearest public transport
Inflation PPI
Annual GDP
GDP Growth

Introduction

Methodology

Insights

Conclusion



Pre-processing



Exploratory Data
Analysis



Regression &
Evaluation



Recommendations



Introduction

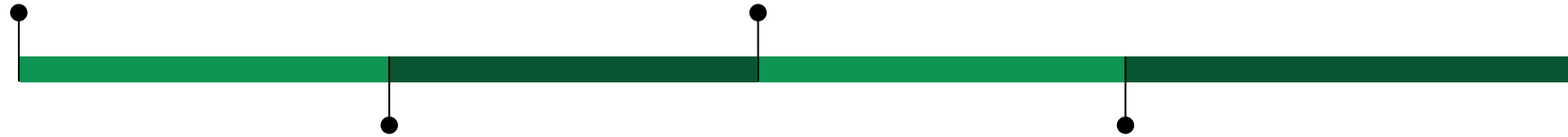
Methodology

Insights

Conclusion



Pre-processing

- 
- Dealing with **categorical** features
 - Dealing with features with high % **null** values
 - Dealing with features with **low variance**
 - Dealing with **multicollinear** features
 - Dealing with **outliers** per feature

Introduction

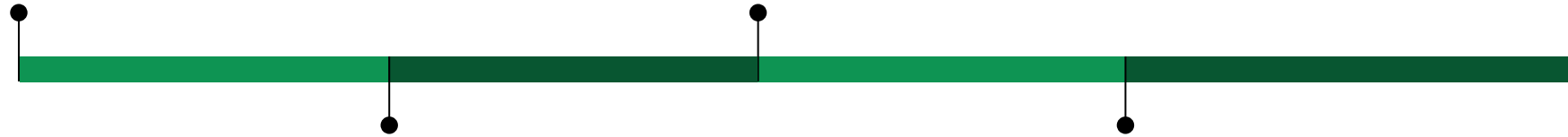
Methodology

Insights

Conclusion

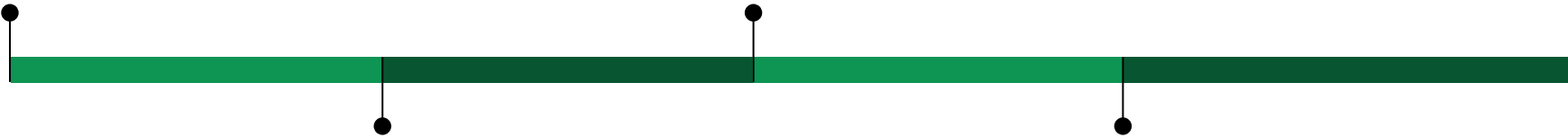


Pre-processing - Categorical Features

- 
- Dealing with **categorical** features
 - Each feature was checked for its distinct values
 - The sub-area, with 133 distinct values, was dropped
 - One-hot encoding
 - `pd.get_dummies()`

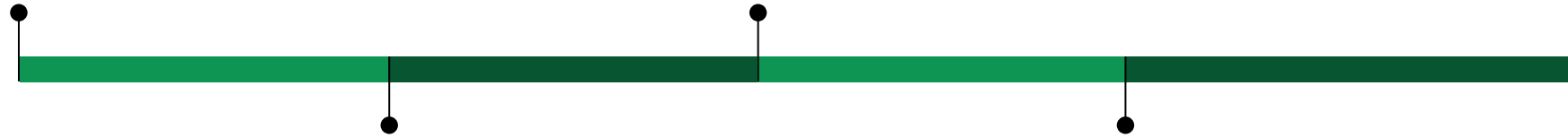


Pre-processing - Numeric Features

- 
- Dealing with features with high % null values
 - Dropped features with > 10% missing values
Remaining features out of all original features: **80.46%**
 - Dropped observations with missing values
Remaining obs. out of all original obs.: **93.65%**



Pre-processing


$$\text{Var}[X] = p(1 - p)$$

p is the probability that a feature contains a certain value

- Dealing with features with low variance
 - VarianceThreshold
 - Dropped numeric features whose variance was lower than 0.19
 - Threshold = $.75 * (1 - .75)$

Remaining features out of all original features: **68.89%**

Introduction

Methodology

Insights

Conclusion



Pre-processing

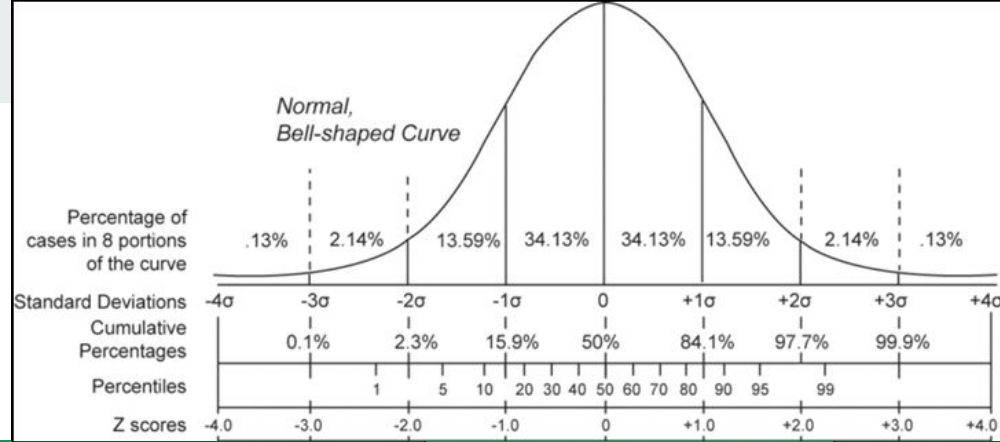
$$VIF_i = \frac{1}{1 - R_i^2}$$

- Dealing with multicollinear features
 - Dropped numeric features whose variance inflation factor (VIF) scores exceeded 10

R_i^2 is the regression coefficient of one feature X_i onto all of the other predictors

Remaining features out of all original features: **23.65%**

Pre-processing



- Dealing with outliers per feature
 - RobustScaler was used to scale the numeric features before dealing with outliers
 - For each numeric feature, dropped observations whose z scores were beyond -4 to + 4

Remaining obs. out of all original obs.: **78.82%**



Pre-processing Summary



	Before	After	% Remaining
Columns or Features	389	92	23.65%
Rows or Obsevation	30,471	24,018	78.82%

Introduction

Methodology

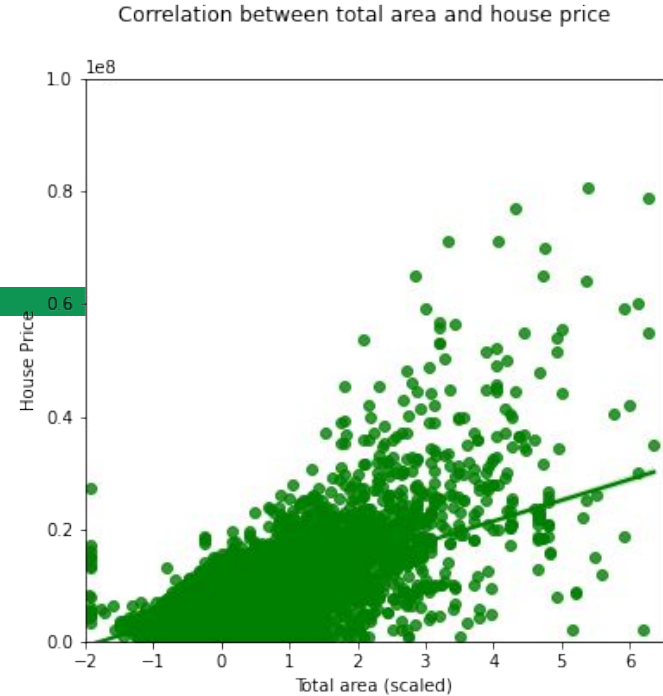
Insights

Conclusion



Exploratory Data Analysis

- Total area and house price have the strongest positive correlation among all numeric features

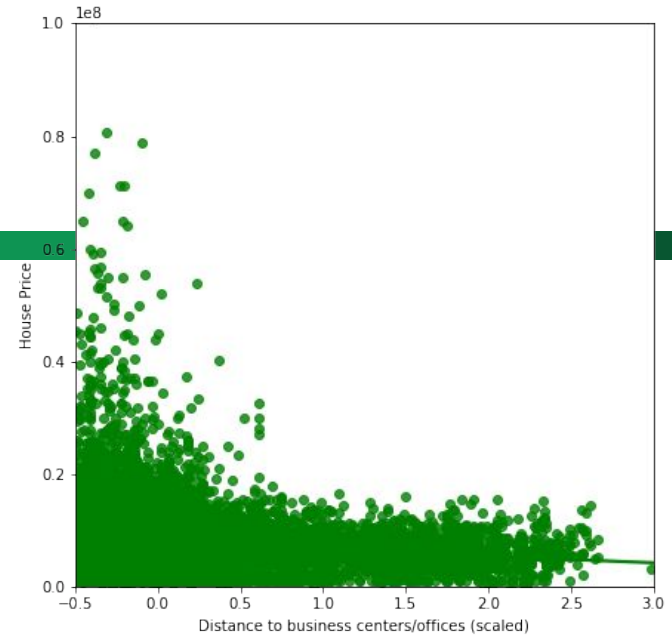




Exploratory Data Analysis

- Distance to business centers/offices and house price have a negative correlation

Correlation between distance to business centers/offices and house price



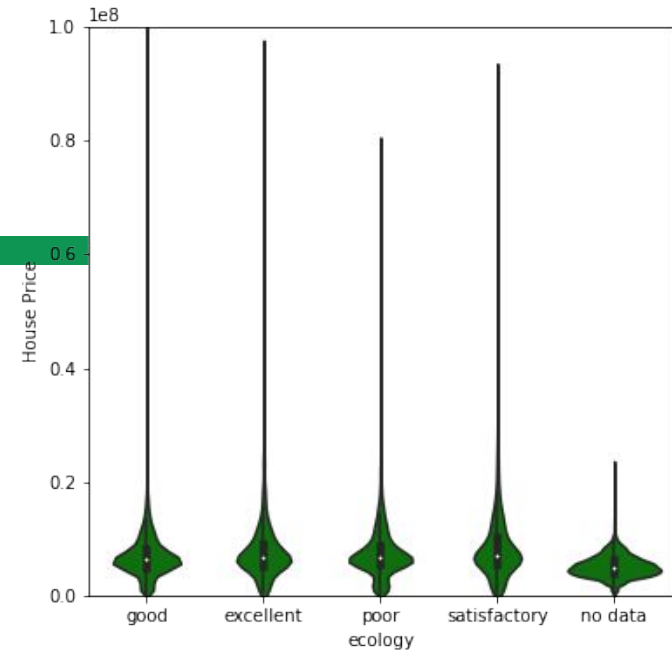
Introduction

Methodology

Insights

Conclusion

Ecological conditions and house price



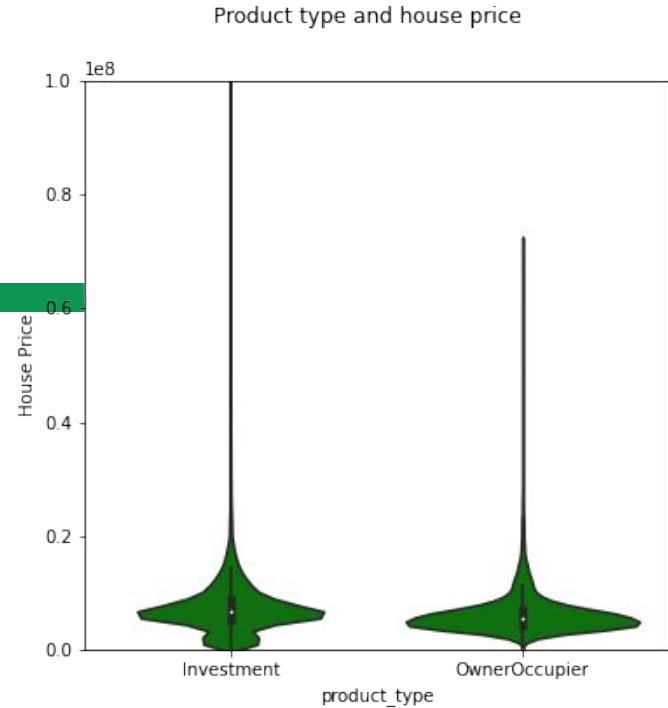
Exploratory Data Analysis

- Better ecological conditions have more houses in the higher price range



Exploratory Data Analysis

- Investment properties (ex. apartments for rent) have more houses in the higher price range





Regression & Evaluation

● Splitting into train and test sets

- Train: 80%
- Test: 20%

Introduction

Methodology

Insights

Conclusion



Regression & Evaluation



Experiment with different regressors

- Linear regression (OLS, Ridge, Lasso)
- Random forest
- Neural network

Introduction

Methodology

Insights

Conclusion



Linear Regression

Concepts

Assumptions:

1. Relationship between features and response is linear
2. No highly correlated features

Introduction

Methodology

Insights

Conclusion



Random Forest Model

Implementation

n_estimators: 150

max_depth: 50

max_features: 0.5

min_samples_leaf: 3

Introduction

Methodology

Insights

Conclusion



Neural Network

Implementation

Loss Function: Mean Absolute Error

Hidden Activation Function: ReLu

Final Activation Function: Linear

Optimizer: Adam

Learning Rate: 0.001

Batch Size: 32

Validation Split: 0.2

Early Stop Patience: 100

Introduction

Methodology

Insights

Conclusion



Regression

Input Layer : 91

Layer1 : 512

Layer2 : 256

Layer3 : 128

Layer4 : 64

Output Layer : 1

Drop Out : 0.2

Layer (type)	Output Shape	Param #
=====	=====	=====
dense (Dense)	(None, 512)	47104
dense_1 (Dense)	(None, 256)	131328
dropout (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 128)	32896
dropout_1 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 64)	8256
dropout_2 (Dropout)	(None, 64)	0
dense_4 (Dense)	(None, 1)	65
=====	=====	=====



Regression & Evaluation



	Linear Regression	Random Forest	Neural Networks
r^2 score (train)	59.6%	89.1%	
r^2 score (test)	55.5%	71.5%	

Introduction

Methodology

Insights

Conclusion



Regression & Evaluation



	Linear Regression	Random Forest	Neural Networks
MAE (train)	1,791,005.925	748,951.17	1,576,457.75
MAE (test)	1,759,614.235	1,410,964.57	1,484,579.63

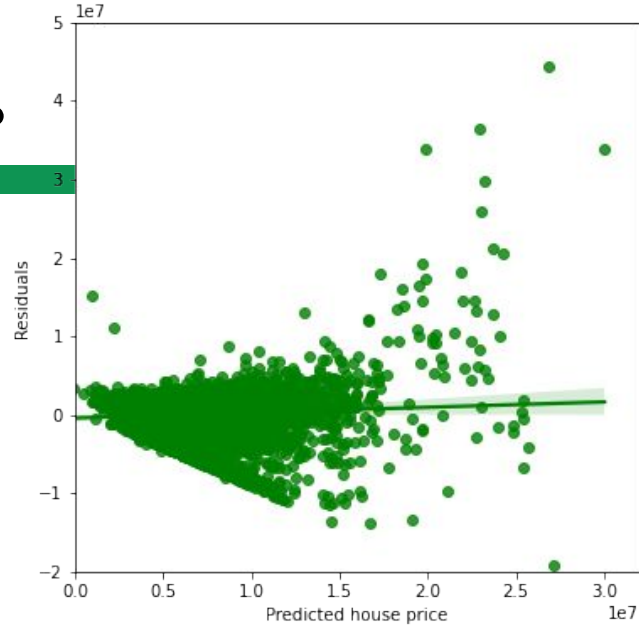
Introduction

Methodology

Insights

Conclusion

Plot of linear regression fitted values vs. residuals

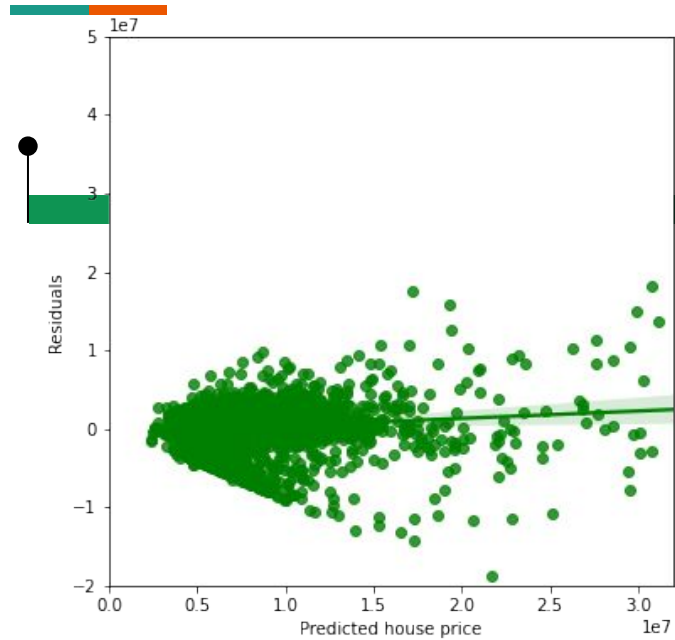


Regression & Evaluation: Linear regression

The residual plot shows a funnel shape where the higher the predicted price, the greater the error. This is often solved by transforming the target house price before modeling.

However, transforming it led to worse r^2 performance, so the original response was kept.

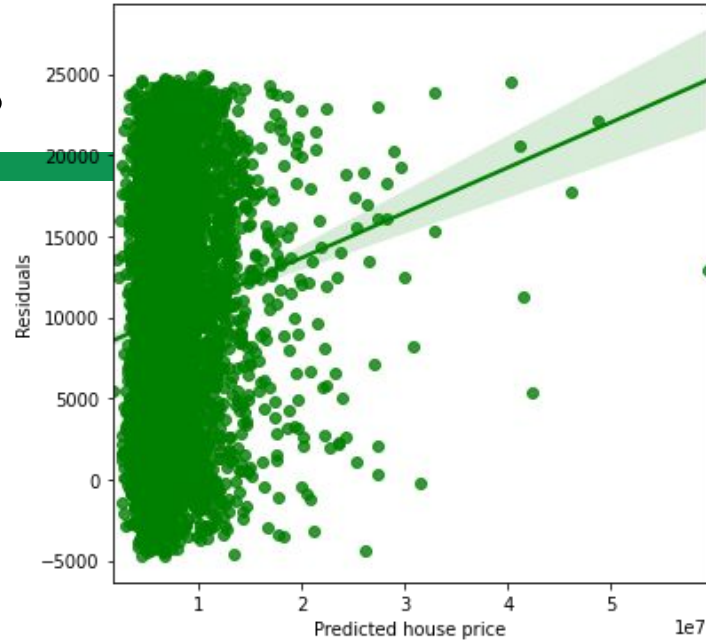
Plot of random forest regression fitted values vs. residuals



Regression & Evaluation: Random Forest Regression

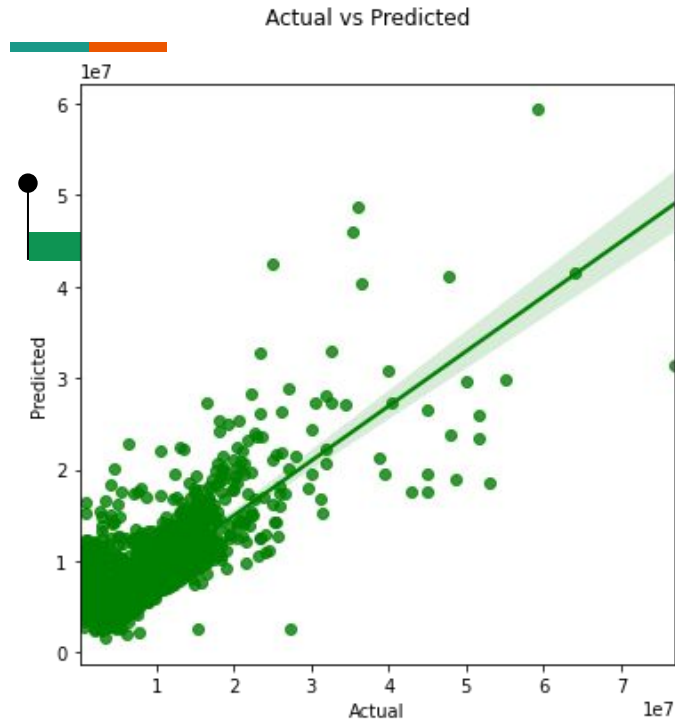
The residual plot shows a funnel shape where the higher the predicted price, the greater the error.

Plot of neural network regression fitted values vs. residuals



Regression & Evaluation: Neural Network

The residual plot does not show any discernible pattern across different predicted prices.



Regression & Evaluation: Neural Network

The residual plot does not show any discernible pattern across different predicted prices.

Most Significant Features

Linear Regression	Random Forest	Neural Network
full_sq	full_sq	full_sq
office_sqm_5000	trc_sqm_3000	ecology_no data
ecology_satisfactory	office_sqm_2000	ecology_poor
railroad_1line_no	industrial_km	gdp_quart_growth
culture_objects_top_25_yes	cemetery_km	office_km

Insights and Observations

The following factors greatly impact housing prices:

- The higher the total area in square meters is
- The closer the property is to different commercial places such as business centers & industrial zone
- The better the ecological condition
- The presence of significant cultural objects
- The higher the GDP growth in the country's economy

Recommendations to Sberbank based on most significant features



Banks can

Adjust mortgage interest rates

**Limit proportion of borrowers based on ratio of
loan value to house value**



How Can Banks Calculate These?



Banks can look at the most important features to help them conduct risk assessment

- Monitor property development in business districts
- Monitor GDP growth
- Monitor possible adverse outcomes

Recommendations to improve modeling



- Time Series Modeling - include timestamp of transaction
- Feature Selection - use other methods
- Feature Engineering - e.g. include sub-area (name of district where house is located) and categorize it to reduce the number of classes
- Try polynomial regression instead of linear regression



Thank you for listening!

Questions?