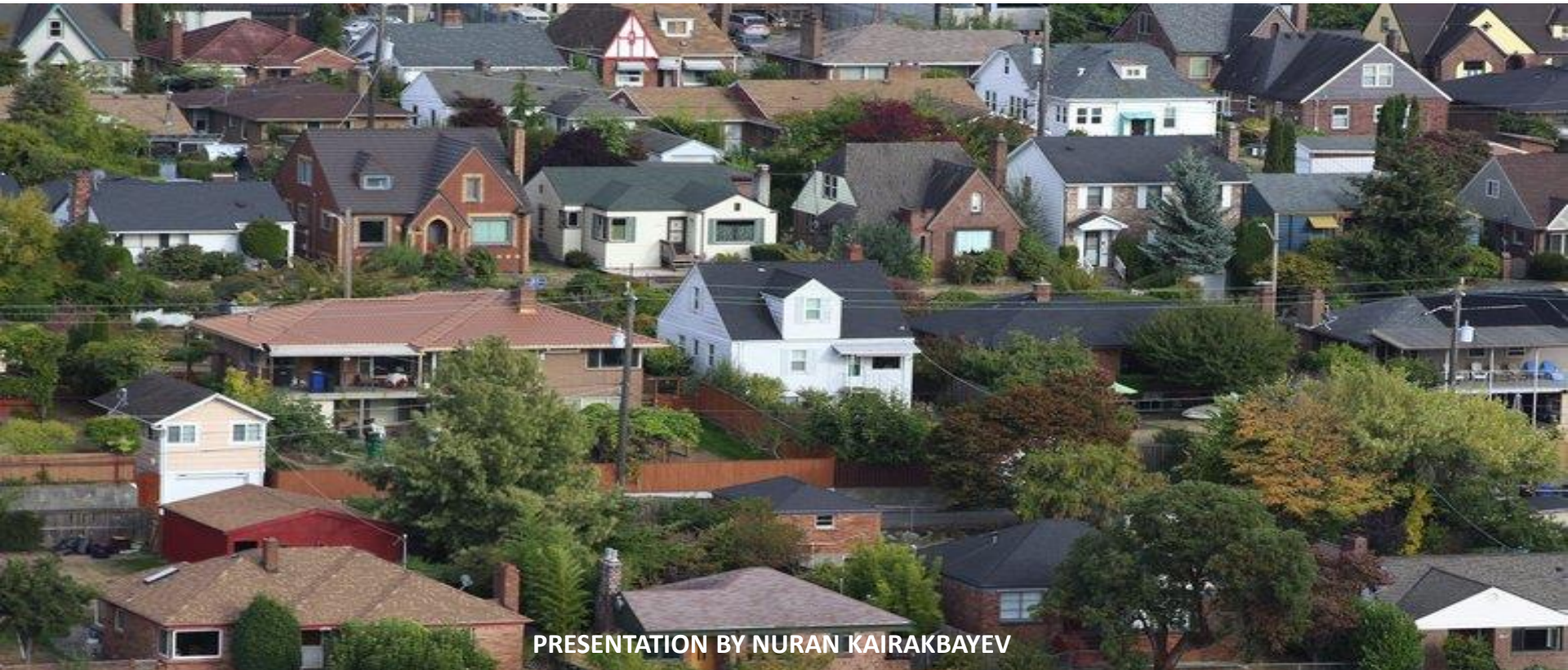
Module 1 Project

Flatiron School. Data Science program

King County House Sale

Research on price prediction.



PRESENTATION BY NURAN KAIRAKBAYEV

General concept of analysis

General target of analysis – create predictive model for Prices of houses by using multivariate linear regression models and machine learning methods

DATASET:

KC_HOUSE_DATA

consists of:

- ~21 600 samples;
- 21 variables with different data types;
- stored in CSV

PREPARING DATASET TO MODEL:

- Missing data;
- Outliers;
- Non-continuous data;
- Correlated variables
- Adjusting skewness and scaling

MODELLING:

- Feature selection;
- Testing p-values;
- Model fit;
- Residuals testing;
- Interpretation of coefficients

APPLIED METHODS CONCEPTS

DATA SCRUBBING

- Main goal for this step – is to clean dataset from incorrect data types and formats and identify wrong, missing or confused data sampes.

EXPLORATION DATA ANALYSIS

- This stage use statistical and visual tools and methods for understanding variables probabilistic distributions (shape), multicollinearity (dependency between features), categorical types, outliers of data and etc.
- The reason of making all this steps is to prepare correct dataset for regression modelling. All factors above could bias results of modelling.

MODEL FITTING

- Instead of running one-time multivariate linear regression, ML-approaches has methods to do it more sophisticated ways;
- Main principle of those, is to run model on train subset and check efficiency on test subset. Comparing to classical econometric approach, this principle gives more adaptive opportunities, especially for non-fixed and updatable datasets.
- 2 important steps within ML-approaches: feature selections (which is actual, when number of predictors is big), and defining optimal train/test ratio.

Obtained results of the modelling is not fixed set of parameters and needs to be updated (“learned”) time-to-time

DESCRIPTION of INITIAL VARIABLES

PREDICTION TARGET – "PRICE"

NON-NUMERICAL VARIABLES

- "ID" – unique identifier for House (but not unique in Data Set);
- "Date" – date of House sold;
- "zipcode" – House's zip adress;

ORDINAL CATEGORICAL VARIABLES and DUMMIES

- "waterfront" – is a House has view to waterfront (1-yes, 0-no);
- "conditions" – how good is condition (Ordinal variable);
- "grade" – overall grade based on King County grading system;

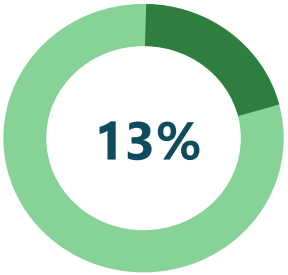
NUMERICAL VARIABLES WITH DISCRETE VALUES

- "bedroom" – number of bedrooms per house;
- "bathroom" – number of bathroom per bedroom;
- "floors" – number of floors in House;
- "view" – how many times a House has been viewed;

NUMERICAL VARIABLES WITH CONTINUOUS VALUES

- "sqft_living" – footage of the home;
- "sqft_lot" – footage of the lot;
- "sqft_above" – footage of the home apart from basement;
- "sqft_basement" – footage of the home's basement;
- "sqft_living15" – footage of the nearest 15 homes (average);
- "sqft_lot15" – footage of the nearest 15 lots (average);
- "lat" and "long" – latitude and longitude (location parameters);
- "yr_built" and "yr_renovated" – year, when house was built and renovated

INITIAL DATA vs DATA for MODEL

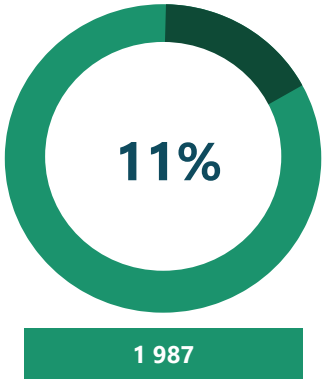


1. MISSING DATA

- Missing, undefined and incorrect data samples was founded in some variables;
- Some of missed variables replaced by median values – for age_after_ren (ex-yr_renovated);
- For other variables missing samples dropped.
- In total, dataset size reduced from 21 597 to 18 749 samples

2. OUTLIERS

- Outliers identified mainly by visual analysis and quantile calculations;
- Most of all outliers are “right-based” – lies of right side of range;
- Most of categorical outliers tested on p-values using single-OLS.
- In total, dataset size reduced from 18 749 to 16 762



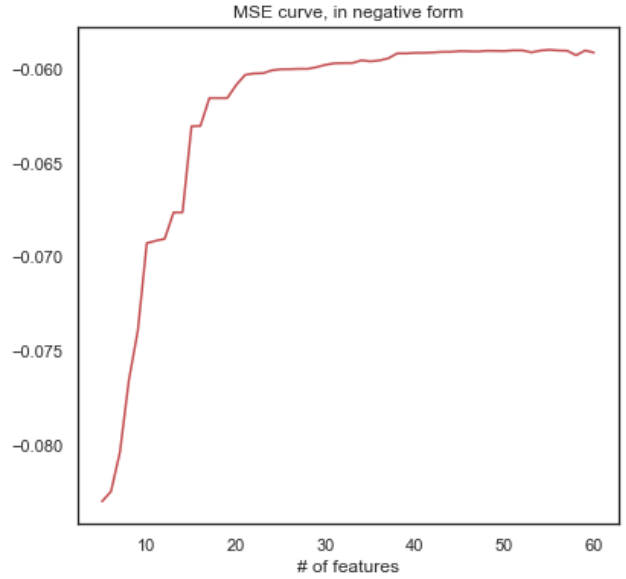
3. PREDICTORS

From initial 20 variables – candidates for predictors list:

- 3 predictor were dropped as useless (zipcode, id, date);
- 8 variables were transformed to dummy-sets of variables;
- After single-OLS and outlier analysis part of dummy-vars dropped;
- Multicollinearity test identified 2 dependent variables: sqft_above and sqft_lot15
- In total, 64 variables were prepared for model fit stage

MODEL FIT STAGE

Main results of iterative modelling multiple linear regressions:

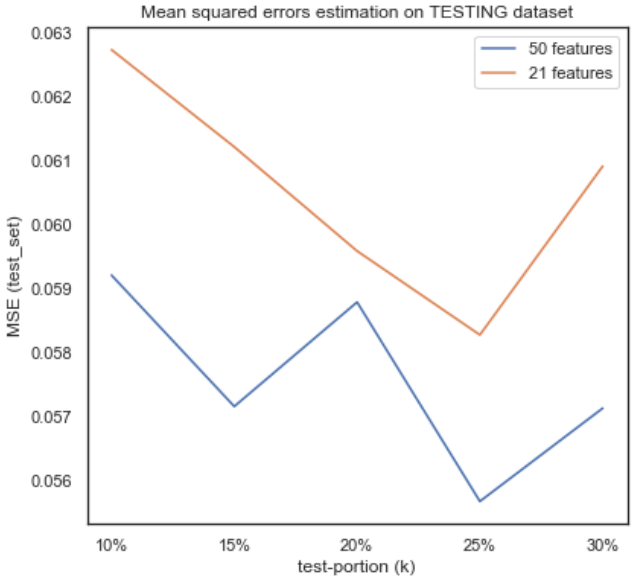


< = RFE outcomes

optimal range number of features vary from 23 to 55
(after p-values test – adjusted to range from 21 to 50)

TTS outcomes = >

Optimal test size in dataset is 25% for both selected cases (50 and 21 features)



Obtained from train R-squared-adjusted on train and test subsets of data varies in range 70%-75% for the case with 21 features.

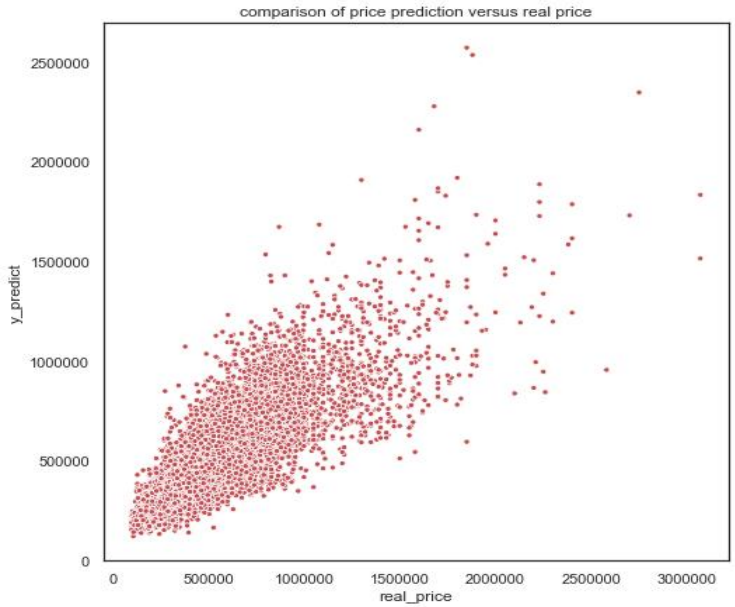
This result demonstrate, that Model is not overfitted and relatively viable.

SUMMARY

- Obtained regression formula evaluate house price dependency on 21 factors;
- Price is predicted in form of log(price) as positive variable;
- General formula is:

*Log(Price) = Const + {a*Log(predictor)} + {b*(predictors)}*

- Achieved Rsq ~ 0,7 says, that prediction level is moderate and works good when input data not abnormal, it can be viewed on the plot



21

PREDICTORS

Optimal set according to RFE method

75%

TRAIN SET

Optimal train portion

70%-75%

R-squared

Achieved range

Model results could be optimized and re-trained

THANK YOU

APPENDIX A. OUTCOME FORMULA

PREDICTION TARGET:

LN(**PRICE**) =

PREDICTORS (coefficient given as rounded numbers):

= – 49.63 + 2.21 LN(**SQFT_LIVING**) – 0.46 LN(**SQFT_LOT**) + 1.14 LN(**SQFT_LIVING15**) –

– 5.85 **YR_BUILT** + 64.41 **LAT** + 0.52 **WATERFRONT_1.0** – 0.48 **BATHROOM_0.5** +

+ 0.21 **BATHROOM_3.75** + 0.12 **BATHROOM_4.0** + 0.19 **BATHROOM_4.25** +

+ 0.09 **CONDITION_3** + 0.14 **CONDITION_4** + 0.21 **CONDITION_5** – 0.34 **GRADE_4** –

– 0.41 **GRADE_5** – 0.33 **GRADE_6** – 0.18 **GRADE_7** + 0.21 **GRADE_9** + 0.36 **GRADE_10** +

+ 0.48 **GRADE_11** – 0.15 **VIEW_0**

RESULTS INTERPRETATIONS:

- There are 13 predictors with positive influence 8 – with negative effect on price
- We could state, that in general space, latitude, year of building, number of bathroom, condition, grade rank, is house viewed or not and view on waterfront – are the main predictors for prices.
- Negative correlation with SQFT_LOT and YR_BUILD was surprised here, but we need to look deeper for house market specifics for understanding this point.
- In the group of dummy-variables the TOP3 strongest features are: WATERFRONT_1.0, BATHROOM_0.5 and GRADE_11.
- Some subgroups of dummies has clear picture of the point, dividing negative and positive dummies: for example, we can state, that all houses with grade less that 8 traded with discounts, and graded higher than 8 uses some premium on prices.

APPENDIX B. APPLIED METHODS DETAILS

DATA SCRUBBING

- Missing or non-defined Data supposed to dropped or replaced by median values;
- Uncategorized and non-sufficient Data dropped from analysis;

EXPLORATION DATA ANALYSIS

- Visual analysis at first step used for identifying category-type data and outliers;
- Unique values and visual multi-modal KDE-plots used for assuming bin sizes;
- Converting to natural Log was used to neutralizing skewness of continuous data;
- Statistical dependency was tested on correlation matrix with excluding variables with more than $|R| > 0{,}75$;
- Min-max-scaling method was used for scaling features, since almost all variables are positive.

MODEL FITTING

- Recursive Features Selection approach (RFE) was used for identifying optimal set of predictors (from 5 to 61, step 1);
- Rsq and Mean Squared Errors (average in K-fold runs, k=5) were used as main quality measures of RFE process;
- Based of learning curve – 2 cases was chosen for model fit: best (by MSE and Rsq) and balanced (based on visual analysis);
- Based on t-stats (p-values) additional feature filtering was applied (by using statsmodels OLS) before final fit;
- Train-test-split approach applied to define optimal train portion for test Rsq and MSE for final model;
- Visual analysis (QQ-plot) and key statistics (skew, kurtosis and JB-test) were applied on final models result.

EXPECTED RESULTS:

Linear regression formula, derived from train set and usable on test set