HOUSE SALES ANALYSIS

Volha Puzikava
April, 2022

Overview

- The project analyzes the King County House Sales dataset
- The project proposes the pricing algorithm that can help real estate agencies and homeowners to sell houses
- The project answers the question how the number of floors influences on the estimated value of the house

Outline

- Business Problem
- Data
- Data Preparation
- Modeling
- Results
- Conclusion

Business Problem

The XYZ Realty asked for help.

Task:

- to analyze the house sales in a northwestern county
- to provide an advice how the number of floors in homes may increase or decrease the estimated value of those homes, and by what amount
- to create an algorithm that would predict the best price for selling homes based on the known features of those homes

Data

- the King County House Sales dataset was analyzed
- the data provided various information about houses: square footage of the living space, lot, basement; number of bedrooms and bathrooms, number of floors, condition and grade of the houses, etc.
- the dataset contained 21,596 houses sold from 2014 till 2015 in the King County, Washington

Data Preparation

- Missing values were replaced to “Unknown” in categorical columns and filled in with appropriate measurements in numerical columns
- All categorical variables were transformed to numbers that reflected some kind of intensity
- The columns were checked for having a direct connection with the column “price” (linear relation)
- The strength of the resulted relations was then visualized (correlation)

Modeling

- Dependent variable/ target was assigned to “price”, all other columns were represented as predictors
- The data was split into training set (to build up a model) and testing set (to implement a model) to estimate how well the learned model will generalize to new data and to improve accuracy
- The strength of the relations between the target and predictors was checked; the strongest relation was used in a simple model with minimum expected performance (baseline)

Modeling Cont’d

- To improve the model, some predictors were excluded due to the lack of connection with the target (no linear relation), high relation strength between one another (multicollinearity) or they were not strong enough to produce an effect in the final model (high p-value)
- Extreme data values were excluded from the model
- Graphs were plotted

Results

According to the analysis:

1. The price of the house is predicted with square footage of the living space, square footage of the lot, and the quality of view from the house:

a. the base price for a house is about $148,475.59

b. for every square footage of living space, the price goes up by $163.06

c. for every square footage of the lot, the price goes down by $0.06

d. the quality of view from the house increases its price by $35,332.79 per numerical value of the feature

2. The number of floors in the house is not a statistically significant feature

Conclusions

- The coefficients of the resulted model should be used only for predictive purposes
- Not the strongest model
- Only 39.4% of the variance of the target variable “price” is explained by the variance of the predictors

Next Steps:

- Further analysis of the dataset
- The use of only numerical data for the development of algorithm

Thank you!

Email: helga.mikel@gmail.com

GitHub: @VolhaP87

LinkedIn: linkedin.com/in/volha-puzikava-2319294a