Surprise Housing

Submitted by:

Sabuj Ghosh

ACKNOWLEDGMENT

- Stackoverflow.
- Github.

INTRODUCTION

- Business Problem Framing

  Prediction of residential real estate properties is needed to assist business decisions regarding investments.

- Conceptual Background of the Domain Problem

  My experience in real estate business in non existent. All I have is from the data.

- Review of Literature

  Performance of various similar github projects suggest that regularization is very much needed.

- Motivation for the Problem Undertaken

  To assist business decision making.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Clearly this is a linear regression problem; but the data had categorical variables as well. To begin with the data had 81 features so regularization is the need of the hour.

- Data Sources and their formats

  Data was provided by fliprobo in csv format.

- Data Preprocessing Done

  The categorical variables were not ordinal in nature so label encoding was avoided and dummy variables were created. Missing values in continuous variables were replaced by median and the same in categorical variables were replaced by most frequent elements. Features with more than 85% missing values were removed. The data was scaled using standard scaler. Target variable had high skewness that was removed using logarithm.

- Data Inputs- Logic- Output Relationships

  Most dominant features exhibit linear relationship with target variable.

- State the set of assumptions (if any) related to the problem under consideration

  None.

- Hardware and Software Requirements and Tools Used

  Any modern laptop will work as number of instances is not very high. Anaconda with jupyter notebook is sufficient.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Mainly linear regression.

- Testing of Identified Approaches (Algorithms)

Sci kit learn linear regression algorithm, ridge regularization, lasso regularization .

- Run and Evaluate selected models

  Cross validation with ridge and lasso with various learning rates.

- Key Metrics for success in solving problem under consideration

  r2 score.

- Visualizations

  All categorical variables were visualized for effect on mean of sales price as well as count of the categorical variable in different categories. If any category contained more than 90-95% of the whole data the feature was dropped and not used in final model building.

  All continuous variable were visualised by histogram and scatter plots. Their correlations were also bar plotted.

- Interpretation of the Results

  Lasso regression gives around 85-90% accuracy. So the investment done based on this model should consider at least 10% risk and normalize the same.

CONCLUSION

- Key Findings and Conclusions of the Study

  Overall quality of the property, a good garage, high living space, quality basements are the ones a property dealer should look for in an investment.

- Learning Outcomes of the Study in respect of Data Science

  Tried various algorithms, often the simplest one is the best one.

- Limitations of this work and Scope for Future Work

  None as of now, pumping up the accuracy will come with the problem of overfitting.