

Housing Project

Submitted by:

Megha Shah

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same

purpose, the company has collected a data set from the sale of houses in Australia.

There is a need to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

Once the company knows all these things it can profitably invest in housing business.

* Conceptual Background of the Domain Problem

Housing prices can be very illusive as the price depends on a vast majority of different factors, like locality, type of housing, year built, size, shape, locality type, street/road etc. There are some other factors like availability of parking place e.g. garage etc. Some other factors are like agriculture land or residential or commercial etc.

Some factors may be the key factors while some others may also affect housing pricing but up to some less extent. Here we have data available that we need to pre-process for conforming according to the needs of machine learning models.

* Review of Literature

I read the data description and tried to search the terms from internet and that helped me a lot in preparing the data for machine learning. I went through

* 1. Housing terms from Internet.
  2. Visualization techniques taught by Institute.
  3. Data pre-processing taught by Institute.
  4. Machine learning taught by the institute.
* Motivation for the Problem Undertaken

As a problem put forward by my institute, I find it interesting as well as necessary compulsion to complete this project. It is very interesting to resolve such problems while learning the subject of Machine Learning. As per knowledge that I retained from what I was I taught at the Institute I was motivated to complete this project.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The project utilizes statistical methods like mean, median, mode at the step of data-preparation, cleaning, filling and during machine learning score calculations. I used Regression Model for this continuous data prediction.

* Data Sources and their formats

Data was provided by my Institute. In this data many columns were numerical like integer, float. But some features of type categorical that needs some special attention and required conversion from categorical to numerical. Null values were present in many features.

The data was provided in two CSV files. One for the Training Data and another one for Test Data that we have to use for prediction.

* Data Preprocessing Done

Some features having null values more than 80% were simply discarded as filling them using any method could jeopardize the model into giving very wrong predictions.

Some categorical feature had NaNs that could be easily replaced with string ‘NA’ as per description of the feature available in the description.txt file.

For numerical features, most of the time Mode was used because it could easily safeguard us from filling wrong values as Mean and Median get affected easily due to the presence of outliers.

For some features like GarageYrBlt I used YearBuilt field values to replace NaNs as Mean, Median, Mode could give very wrong results like a Garage being built many years before the house was built.

There was another problem too that I encountered while converting the feature from categorical to numerical ones using One hot encoder was that the number of columns differed when used on training data and testing data. To overcome this problem I had to club both the datasets together to form a single dataset and then process it. Later after creation and training of the Model I recreated the Test Dataset by separating the same from the clubbed dataset.

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

Assumptions made were like for GarageYtBlt year I took it as the YearBuilt (House) for replacing the NaN values in the feature GarageYtBlt.

Similarly the NaNs present in many features were taken as NA string.

One assumption was that the feature having NaNs more than 80% were taken as useless and discared.

Later when all the features were converted to Numerical features, the columns having correlation more than 60% were dropped to reduce the no of features to be used while training the Model.

* Hardware and Software Requirements and Tools Used

I coded this project using Jupyter notebook on my Laptop with I-3, with 4 GB RAM and 500 GB HDD. The project occupied space in HDD in kbs only (150 kb).

Libraries used for accomplishing the above task were

1. Numpy used for some function like isnan() etc.
2. Pandas used for handling DataFrames
3. Matplotlib for data visualization.
4. Seaborn also for data visualization.
5. Sklearn for machine learning Regression Model, and score etc.
6. Pickle for saving the Trained Model in a file.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I clubbed both the datasets : training data as well as test dataset. Because of some missing unique values in some categorical features that created problems while converting Categorical Data into Numerical Data. The major problematic area was the preprocessing/cleaning and conversion of dataset into conforming to the Machine learning model requirement. Statistic mean, median and mode were used to guess missing values in some features that had NaNs in some records.

* Testing of Identified Approaches (Algorithms)

The Regression Model was created for the predictions.

* Run and Evaluate selected models

I used Score value for evaluating my model. I trained my model on training dataset and used it to predict values on set of training dataset extracted as a test set just to check Score.

* Key Metrics for success in solving problem under consideration

Most of the time I used mode to replace NaNs as mean and median could give wrong values due to presence of outliers in some features.

* Visualizations

I used distplot and boxplot to visualize data skewness and decide about what strategy to follow to fill NaNs.

* Interpretation of the Results

Visualization techniques show us various problems and correlation between different features and some problematic records having outliers too. Based on that we can apply different techniques to fill NaNs present in a feature.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Key finding is that the categorical data may present some peculiar problems while handling the training and test datasets separately.

* Learning Outcomes of the Study in respect of Data Science

Yes I learned to visualize various problems in the data and learnt how to deal with missing values and created regression model.

I encountered problem at every step, first was the filling of NaNs and had to make some assumptions about the missing data, but I took the help of statistical methods like mean, median and mode to guess various missing variables.

The other type of problem was encountered when converting categorical features to numerical features due to different no of mismatching columns created in the datasets so that both the dataset looked different. Hence to overcome this I had to club both the datasets together to make one big DataSet having 1460 rows.

* Limitations of this work and Scope for Future Work

This is my first model for this problem and it needs further observation and modification like using various other Models and comparing their output. Hyper parameter tunning can also be used to improve the outcome further. Currently this is in a very primitive stage and there is scope for improvement on various fronts.