HOUSING PRICE PREDICTION

Submitted by:

LALIT JOSHI

ACKNOWLEDGMENT

I would like to thank our SME(Khushboo Garg) for her expert advice and encouragement throughout this project, and I also took some help from googled documents and few YouTube channels.

INTRODUCTION

- Business Problem Framing

- A US-based housing company named 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.
- The company is looking at prospective properties to buy houses to enter the market. We are required 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.
- We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

- Conceptual Background of the Domain Problem

- Firstly, we will check whether the problem is supervised or not, mostly we will have a supervised one where there will be a target variable.
- After that we will check whether the project is a regression type or a classification type.
- If we have a train and a test dataset, then we will do all the procedures with the train dataset and the same with the test dataset.
- We will also check whether our dataset is balanced or imbalanced. If it is an imbalanced one, we will apply sampling techniques to balance the dataset.
- Then we will do model building and check its accuracy.
- Our main motto is to build a model with good accuracy and for that we will also go for hyperparameter tuning.

- Review of Literature
    - I have imported important libraries for my project.
    - I have created the data frame for the train dataset. I have analysed my data by checking its shape, number of columns, presence of null values if any and checking the datatypes.
    - Then I have done some data cleaning steps, e.g. checking the value counts of the target variable, dropping some irrelevant columns from the dataset, checking correlation between the dependant and independent variables, visualizing data using bar plots, splitting the data into independent and dependant variables and finally scaling the data. Now I have used DecisionTreeRegressor, KNeighborsRegressor, AdaBoostRegressor, LinearRegression, GradientBoostingRegressor, RandomForestRegressor for model building and found GradientBoostingRegressor with the highest accuracy. I have seen 91% accuracy after hyperparameter tuning (RandomizedSearchCV).
    - Then I have imported the test dataset and follow all the cleaning steps as I have done in the training dataset.
    - I have trained my model on my train dataset and predict on my test dataset.
- Motivation for the Problem Undertaken

  My main objective is to predict the sale price of the houses so that a builder or housing company could easily identify the factors that could you used to decide the prices of the house. Also, objective behind to make this project is to contribute to the world's economy. Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science provides motivation as it can solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  While Working on the project I used many mathematical, statistical and analytical modelling. I had used plotting libraries like matplotlib and seaborn that helped to me analyse the data in the form of graphs. Several, methods in pandas' libraries that helped to identify the mathematical aspects of the data like mean median mode etc.

- Data Sources and their formats
  - **Data Source**: The read_csv function of the panda's library is used to read the content of a CSV file into the python environment as a pandas Data Frame. The function can read the files from the OS by using proper path to the file.
  - **Data description**: Pandas describe () is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values. When this method is applied to a series of string, it returns a different output

- Data Preprocessing Done
  - I checked the shape and datatypes of the train dataset.
  - I have dropped some columns which are not efficient.
  - I have checked the value counts of the necessary columns and also checked if any null values are present.
  - I used Label Encoder to encode the object type columns and used mean to replace the null values.
  - I did some visualizations using heatmap, barplots, cat-plot and scatter-plots.
  - I checked the correlation between dependant and independent variables using heatmap.
  - I split the dependant and independent variables into x and y.
  - I scaled the data using StandardScaler method and made my data ready for model building.

- Data Inputs- Logic- Output Relationships

  The logic between the data input and data output is that how the input variables are influencing the output variable. Here the output variable is the **'Sale Price'** and there are 80 input variables which are responsible for predicting the output variable in an efficient way.

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

  After converting my dataset to numerical (**Encode**) form I build a heatmap which shows the correlation between the inputs and output variable. After checking that I found there are few columns having negative correlation with the output hence I dropped them.

- Hardware and Software Requirements and Tools Used

  ## **Hardware Requirements**:

  Processor: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz (4 CPUs), ~2.7GHz

  RAM: 8192MB

  ## **Software Requirements:**

  Python: Programming language

  ## **Tools used**:

  - Jupyter notebook: NumPy
  - Pandas
  - Matplotlib
  - Seaborn
  - Scikit-learn
  - Scipy.stats

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  1. I have used heatmap to visualize the correlation among the data.

  2. I have used bar plots and scatter plots to get a clear view of the input variable columns and the target column.

  3. For scaling the data, I have used StandardScaler method.

  4. For training and testing the data, I have imported train_test_split library from scikit-learn.

  5. For model building, I have used GradientBoostingRegressor on my train dataset.

  6. For better accuracy of the model, I have used hyperparameter tuning (**RandomizedSearchCV**).

- Testing of Identified Approaches (Algorithms)

  Among the six chosen algorithms for testing, Gradient Boosting Regressor gives a higher accuracy score. So, I have selected Gradient Boosting Regressor algorithm for this project.

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

- Key Metrics for success in solving problem under consideration

  What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

- Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

- Interpretation of the Results

  Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

CONCLUSION

- Key Findings and Conclusions of the Study

  The key findings are we have to study the data very clearly so that we are able to decide which data are relevant for our findings. The techniques that I have used are heatmap, Label Encoder, etc. The conclusion of our study is we have to achieve a model with good accuracy.

- Learning Outcomes of the Study in respect of Data Science

  We will execute statistical analyses with professional statistical software. We will develop the ability to build and assess data-based models.

- Limitations of this work and Scope for Future Work

  The results were promising for the public data due to it being rich with features and having strong correlation, whereas the local data gave a worse outcome when the same pre-processing strategy was implemented due to it being in a different shape compared with the public data in terms of the number of features and the correlation strength. Hence, the local data needs more features to be added preferably with a strong correlation with the house price. Future scope of this work is we can try different algorithms like Lasso Regression, Ridge Regression etc for model building and try to achieve a good accuracy and f1-score.