

**HOUSING: PRICE PREDICTION**

Submitted by:

NARENDRAN SURESH

**ACKNOWLEDGMENT**

I would like to take the opportunity to thank the organisation ‘FlipRobo Technologies”, “Datatrained “and my mentor Ms. Astha Mishra for their immense support and guidance for making this feat achievable. I would also like to thank my dad Mr. Suresh Kumar S who is a statistics subject expert for guiding me along the right path and techniques to be used.

FlipRobo Technologies who are specialised in making ML / AI models provided me the Data set as a part of my internship.

References were made to several articles among Medium, KdNuggets, towardsdatascience, realpython, machinelearningmastery, python and sklearn documentation for the successful completion of the project.

**INTRODUCTION**

* HOUSING – PRICE PREDICTION (Business Problem Framing )

Surprise House, a US based real estate investor is our client who is looking to enter the Australian real estate market and invest in houses which they could resell later at a margin.

* Conceptual Background of the Domain Problem

Real estate is one of the very profitable industries provided the investor has done his research. There are numerous factors like the age, quality, finishes, type of materials used, amenities and several other factors that affect the pricing of a house. The client is motivated to build a machine learning / deep learning model that predicts the prices of houses and factors affecting the pricing.

* Review of Literature

Housing and real estate research was done for the purpose of the dataset. The research showed that general factors include the climatic conditions of the area, the level of the ground, the locality and other factors. These factors vary based on region and it was noted that luxury facilities like swimming pool and garage parking were factors that affected prices of houses in the area under the scope of our research. Overall quality is a factor that affects the prices irrespective of region.

* Motivation for the Problem Undertaken

The project being a typical linear regression problem with the price of the housing units to be predicted provided the factors caught interest. The factor that deep learning was to be implemented and this being the first ever project where I have got a chance to implement deep learning and Neural Networks also is a motivation factor.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The dataset is a csv file with 1460 rows and 81 columns, 80 features that could either affect the price positively / negatively and one target feature which is the price.

* Data Sources and their formats

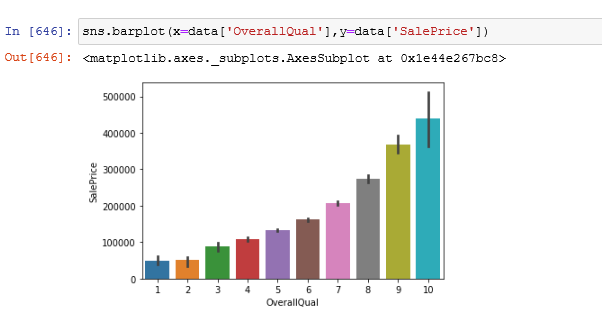
The dataset contains 81 columns with each row explaining features and characteristic of homes. FlipRobo, as a part of the internship provided the data for analysis and modelling.

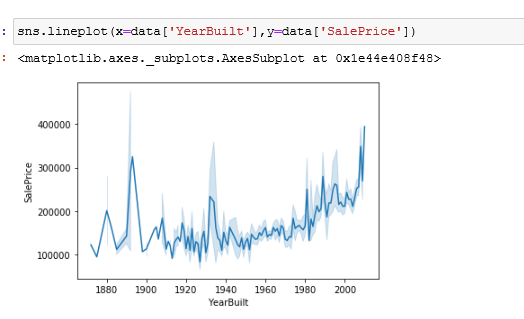
The datatypes are spread out, there are equivalent amount of object type attributes and integer type and a few date attributes.

Processing needs to be done to convert all the data into system understandable language.

* Data Preprocessing Done

None of the features were removed for the scope of the project. All features have been adjusted, null values treated with either their modes, replaced NA’s with their typical values (No, Not included).Univariate Analysis was done for the variables and replaced the errors in the data. Multivariate analysis was then done and a few graphical representations were done to get insights into the data.Below are a few charts for illustration purpose.



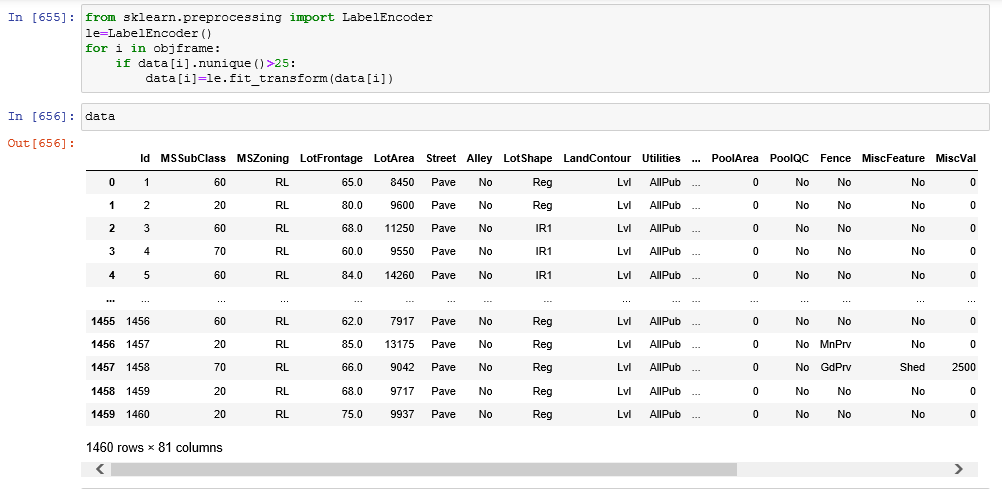


* Data Inputs- Logic- Output Relationships

Correlation plots were done to see how the data interacted with each other and affected the prices.

Most important feature was found to be overallQuality which had a strong positive correlation to the prices.

The object data types were treated using either one hot encoding / label encoder.



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

N/A

* Hardware and Software Requirements and Tools Used
* Hardware:
* Inter Core (i7) – 5500U, clock speed at 2.40GHz
* RAM – 12.0 GB
* Software:
* Jupyter Notebook (Anaconda 3) – Python 3.7.6
* Microsoft Excel
* Libraries & Packages used – Pandas, numpy, sklearn, matplotlib, seaborn, sklearn, scipy, imblearn, tensorflow, keras

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The approach was pretty straight forward which started off with the cleaning of the data, checking for correlation and then directly applying the object encoders followed by implementation to the model.

* Testing of Identified Approaches (Algorithms)

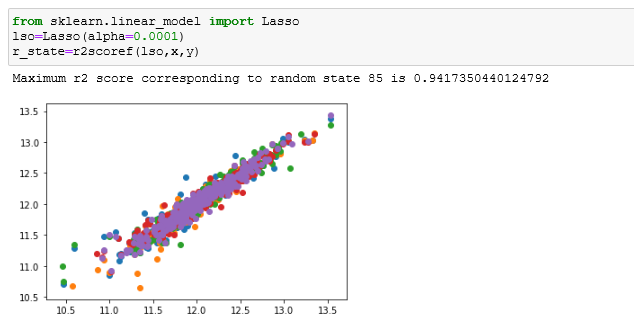
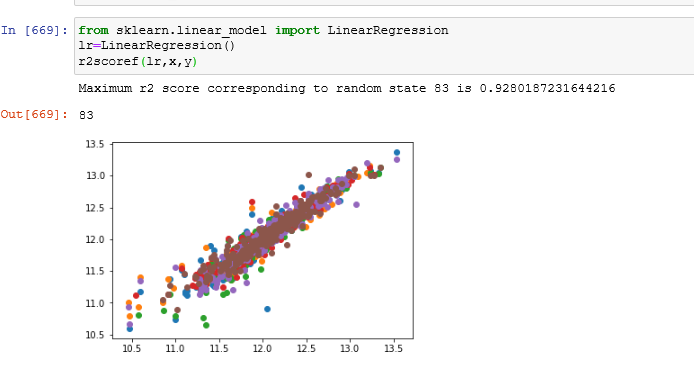
Linear Regression

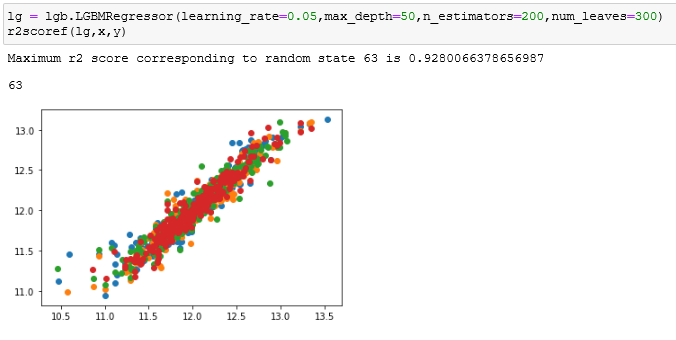
Lasso

Lightgbm

* Run and Evaluate selected models

The above models were evaluated to check for their accuracy and it was noted that the models performed well. Below are the screenshots of the model accuracy scores.





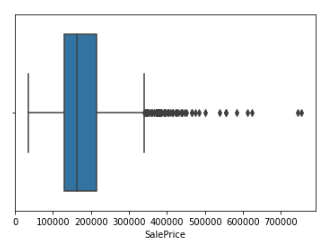
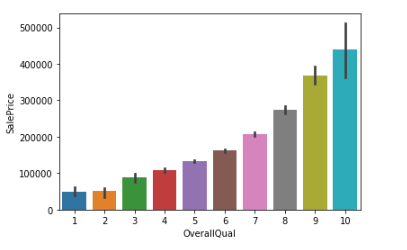
* Key Metrics for success in solving problem under consideration

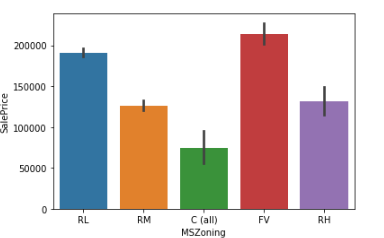
The key metrics used were accuracy. This was further validated with cross validation and the scores were satisfactory.

The cross validation scores for the Lasso Model was given as 0.89 with a standard deviation of 0.042

* Visualizations

Box Plots, heat maps, scatter plots,histograms and line plots were used for the scope of the project.

Few snippets are as below :



* Interpretation of the Results
* From the results, it is clear that the model performance is above 90% which is a desirable score for any model. To further enhance the results, we can try hyperparameter tuning using GridsearchCV.

The score was found to be 92% for hyperparameter tuning of the lasso model.

Deep Learning model was also implemented which also gave a fairly good performing model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

House prices were highly dependent on the overall quality, but not on the overall condition.

Except for the antique values and the exceptional traditional value, almost all house pricing was observed to increase the later it was built.

* Learning Outcomes of the Study in respect of Data Science

The major learning opportunity in this project was the chance to work on deep learning on a live project. Also got to learn the data management technique without having to delete a single row or column in the dataset.

Better use of encoders were done so that the dataset gave a better outcome.

* Limitations of this work and Scope for Future Work

The dataset was limited and there were only 1600 entries. A wider dataset would have meant more scope for learning and implementation of new techniques.