

NAME OF THE PROJECT

Housing price prediction

Submitted by:

Sharanya Susheel

**ACKNOWLEDGMENT**

Firstly, would like to thank Flip Robo for giving me this opportunity and this learning will stay with me for a life time , special thanks to my mentors and career coach at Data Trained who have been my constant source of support and last but not the least would like to thank my family without whose support I would ‘nt be able to finish this project.

Below are the references which helped me in competing my project:

https://m2pi.ca/project/2020/bc-financial-services-authority/BCFSA-final.pdf

https://www.researchgate.net/publication/349477129\_House\_Price\_Prediction

**INTRODUCTION**

* Business Problem Framing
* The goal is to build a regression model to forecast the price ofhouses.The dataset is preprocessed by analysing the missing values.The various regression algorithms are used to predict the price.
* Conceptual Background of the Domain Problem
* The company wants to know:
* 1)Which variables are significant in predicting the price of the house
* 2)How well those variables describe the price of the house.
* Also,determine the lambda for ridge and lasso regression.
* Review of Literature
* We were required to model the price of houses with the available independent variables. The 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 management to understanding the pricing dynamics of a new market.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This problem is about house price prediction. Houses are one of the necessary need 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 comes as a very important tool to 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. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

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 project has been divided into two parts,

Part 1

Exploratory Data Analysis and Data Cleaning :

Part 2:

Training a machine learning model

* This data set contains 1168 rows and 81 columns.
* Data Sources and their formats
* The data is obtained from the US-based housing company named **Surprise Housing.**
* Data Preprocessing Done

The steps followed for data cleaning are described as follows:

1)Firstly the information about the dataset was checked and it gave the data shape i.e the total number of rows and columns.

2)The datatypes of all the columns were checked to find out if they are object, integer or float.

3)Duplicate rows were dropped in the dataset

4)Null values were checked using df.isnull().sum() function.

* 5) After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not

**Model/s Development and Evaluation**

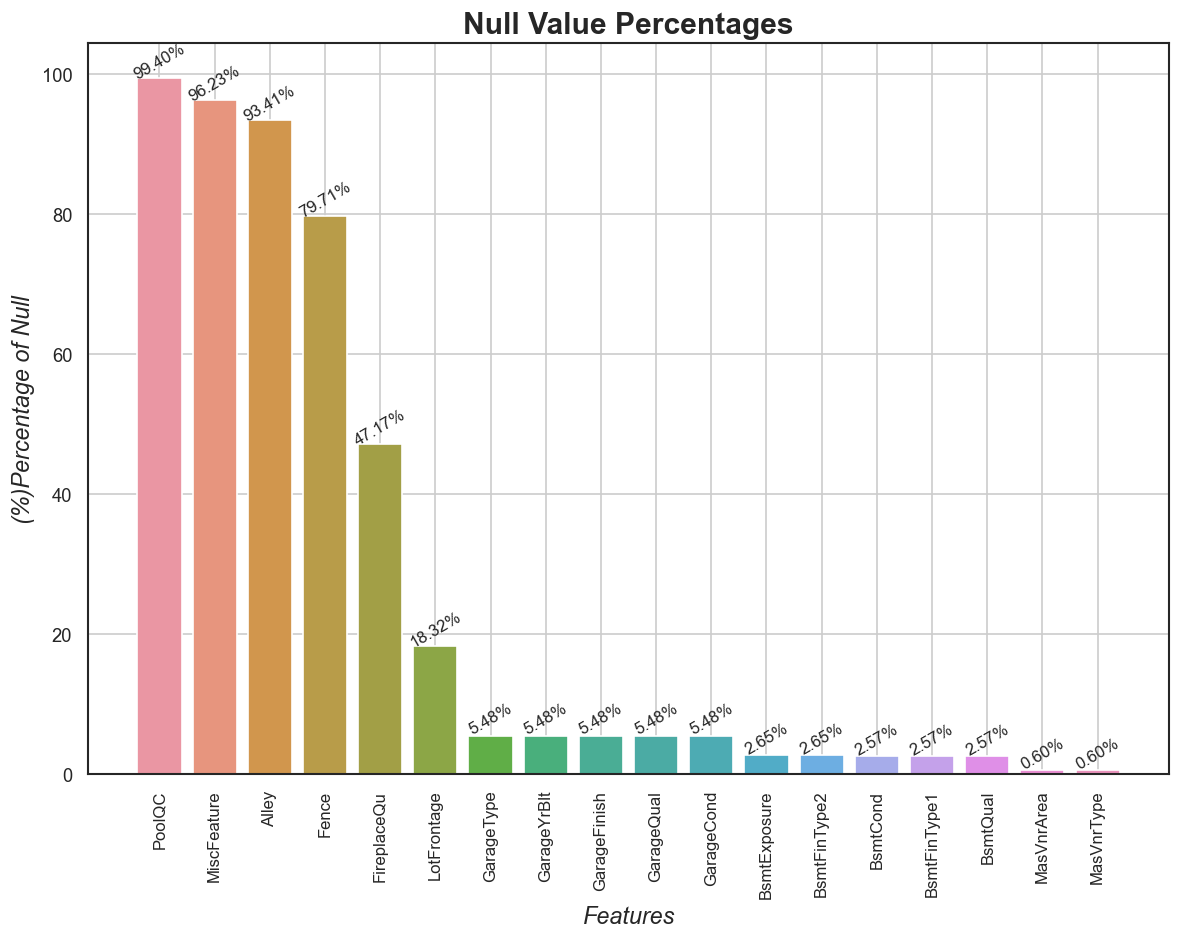
* Identification of possible problem-solving approaches (methods)

. We know that this is regression problem. We will use linear regression models,lasso,ridge for our final outcome of the model.

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

Visualizations

Heat map for the representation of null values



A screenshot of a computer

Description automatically generated with medium confidence

. Plot describing Null value percentageChart, bar chart

Description automatically generated

Bar graph for categorical data

Chart, bar chart

Description automatically generated

Numeric categorical variables

Chart, bar chart

Description automatically generated

Outliers

Chart, box and whisker chart

Description automatically generated

**Visualising target variable-‘SalePrice’**

**Chart, line chart, histogram

Description automatically generated**

**SalePrice after log transformation**

**Chart, histogram

Description automatically generated**

**scatterplots to understand relationship between 'dependent' and 'numeric' variables.**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Heatmap for correlation among numeric variables**

**Chart, bar chart

Description automatically generated**

**Interpretaion of visualizations**

From the heat map for null values we can see the presence of nullvalues in our dataset.

Missing values encountered in the above mentioned columns and their respective percentages. But there are some columns that do have meaningful missing.

From the barplots of object typecategorical variables the following can be observed,

Most preferred choices: (1) (MSZoning) zone of the sale:'RL'(Residential Low Density) followed by 'RM'(Residential Medium Density). (2) LotShape: Reg(Regular) type (3) LotConfig: Inside(Inside lot) (4) No significant preference w.r.t following features: --(i) FireplaceQu (Quality), MasVnrType: 'None' --(ii) KitchenQual (Quality), ExterQual (Quality of exterior material), BsmtQual (Height of basement):'TA'(Typical/Average) (5) RoofStyle: 'Gable' (6) GarageType: Attached (7) GarageFinish: Unfinished (8) Foundation: Poured Concrete (9) (BsmtFinType1) Rating of Basement Fin. Area: Good Living Quarters and Unfinished (10) HeatingQC: Excellent (11) Neighborhood: North Ames, followed by College Creek (12) HouseStyle: One story, followed by Two story. (13) Exterior1st (Exterior covering on the house) and Exterior2nd (if more than one material on exterior): Vinyl Siding, followed by Metal Siding (14) Basement Exposure: No Exposure

From the plots of numeric categorical variables,

Observations: Most preferred choices: (1) Month (House) Sold: June, followed by July and May (2) MSSubClass: Most preferred type of dwelling involved is 20 i.e. 1-Story 1946 & Newer All Styles, followed by 60 i.e. 2-Story 1946 & Newer. (3) Overall Quality (material and finish): Average and Above Average (4) Overall Condition: Mostly Average, followed by Above Average (5) (i) Basement Full Bathrooms: Maximum 0(None), followed by 1. ---(ii) Full Bathrooms: Preferred 2, followed by 1 and (iii) Half Bathrooms: Preferred 0(None), followed by 1. (6) Bedrooms Above Ground:3, followed by 2. (7) Total Rooms Above Ground: 6, followed by 7. (8) No significant preferences for Fireplaces (0 0r 1) and WhetherRemodelled (House) or not. (9) GarageCars space: 2 car spaces, followed by 1 car space.

The outlier plot shows the presence of outliers which are removed to improve the proper efficiency our model.

From the saleprice plot,

The points on the probplot don't lie along the red line. It shows the skewness in the data of the target variable. Hence, transforming the variable using log-transform.

After the removal of skewness from saleprice ,

1) The normal distribution plot now shows that the new distribution of SalePrice (log-transformed) is almost normally distributed. (2) The points on the probplot almost lie along the red line. It shows that the skewness issue in the data of the target variable, 'SalePrice' is almost resolved with some noise.

From the scatter plots,

1) '1stFlrSF','GrLivArea' show a positive correlation with the SalePrice. (2) Other continuous variables are too scattered. Let's understand them from further analysis. (3) SalePrice vs ('LotFrontage', 'MasVnrArea', '1stFlrSF', and 'GarageArea' scatterplots do have some datapoints with some eccentric values/outliers.

From the heatmap for the correlation among numeric variables,

(1)The predictors 'GrLivArea' and 'TotRmsAbvGrd' shows (0.83) strong positive correlation. Hence, dropping 'TotRmsAbvGrd'. (2)The predictors 'GarageCars' and 'GarageArea' shows (0.89) strong positive correlation. Hence, dropping 'GarageCars'. (1)The predictors 'GrLivArea' and '2ndFlrSF' shows (0.71) strong positive correlation. Hence, dropping '2ndFlrSF'. (2)The predictors '1stFlrSF' and 'TotalBsmtSF' shows (0.77) strong positive correlation. Hence, dropping '1stFlrSF'.