Enhancing Real Estate Investment Decisions with Predictive Modeling

SUBMITTED BY:

BHUVANA SHREE N

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to Intrnforte for giving this internship project on the topic (Micro-Credit Defaulter), which helped me in doing a lot of Research and which help me learn way too many things. The references I used for completion of this project are mentioned here Kaggle, [www.stackoverflow.com](http://www.stackoverflow.com) **,** [www.geeksforgeeks.org](http://www.geeksforgeeks.org) are the websites which helped me in completing the project.

“Enhancing Real Estate Investment Decisions with Predictive Modeling”

Project Question: How can data science empower real estate companies to strategically enter new markets, maximize revenue, and optimize investment decisions?

Project Objective: Create a robust machine learning model capable of forecasting house prices in the Australian real estate market. This predictive tool will serve as a vital asset for Surprise Housing, aiding them in identifying prospective properties for acquisition. Furthermore, it aims to unveil the significance of variables that influence house prices and how these variables collectively describe the price of a house.

**Data Understanding and Cleaning:**

The dataset consists of 1460 entries with 81 variables.

The dataset contains missing values that need to be addressed using domain knowledge and appropriate techniques.

Both numerical and categorical variables are present, and they need to be handled accordingly.

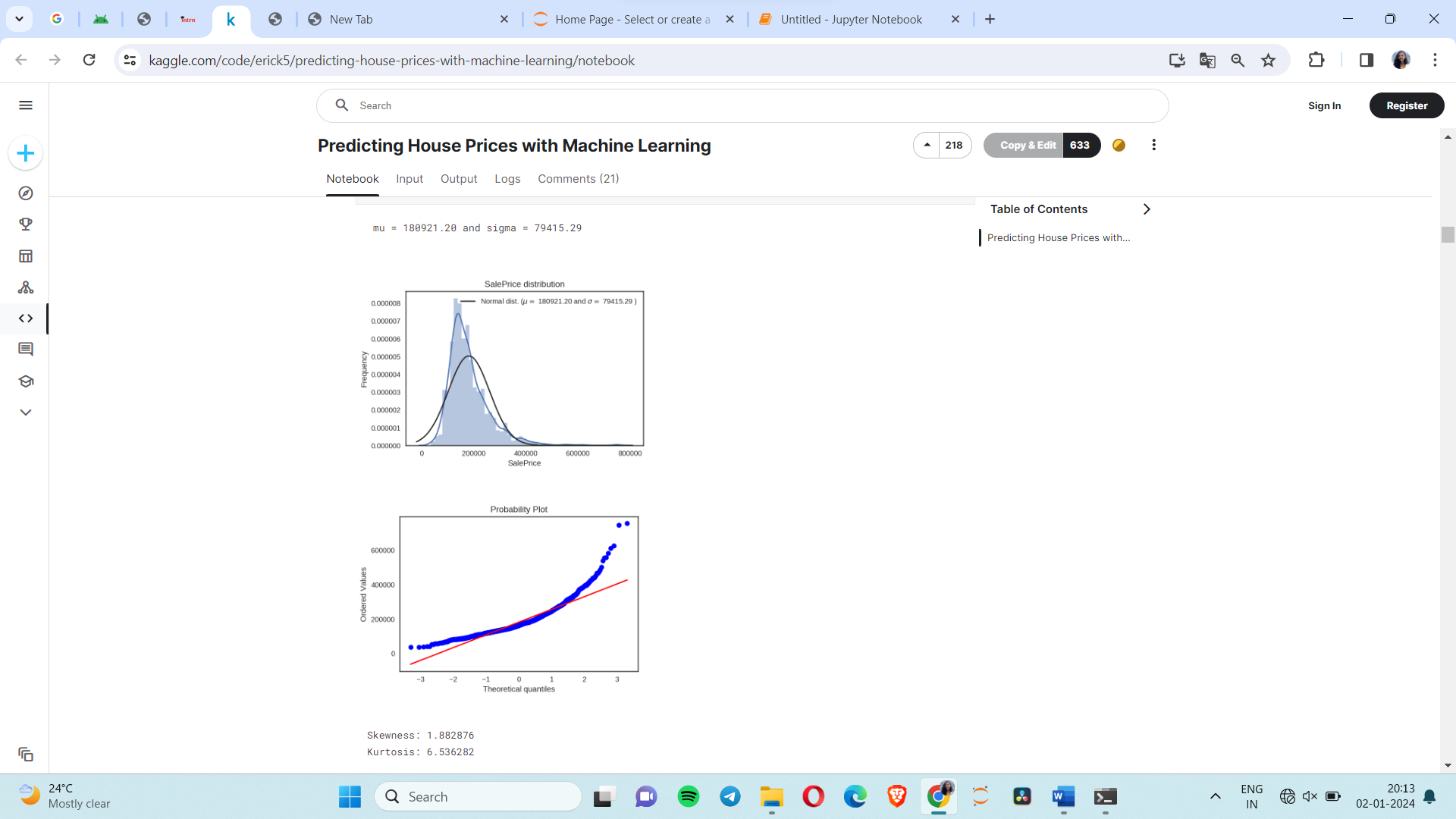
in this study let’s assume we can have two clients at the same time

Client House buyer: This client wants to find their next dream home with a reasonable price tag. They have their locations of interest ready. Now, they want to know if the house price matches the house value.

Client House seller: Think of the average house-flipper. This client wants to take advantage of the features that influence a house price the most. They typically want to buy a house at a low price and invest on the features that will give the highest return.

**Data Collection and Preprocessing:**

With an average house price of $181477.005993



Let's check out all the variables! There are two types of features in housing data, categorical and numerical.

Categorical data is just like it sounds. It is in categories. It isn't necessarily linear, but it follows some kind of pattern.

Numerical data is data in number form.These features are in a linear relationship with each other.

Total Features: 43 categorical + 37 numerical = 80 features

linkcode

With 81 features, how could we possibly tell which feature is most related to house prices? Good thing we have a correlation matrix. Let's do it!

**Exploratory Data Analysis (EDA):**

| Most Correlated Features |
| --- |
| 0 | SalePrice |
| 1 | OverallQual |
| 2 | GrLivArea |
| 3 | GarageCars |
| 4 | GarageArea |
| 5 | TotalBsmtSF |
| 6 | 1stFlrSF |
| 7 | FullBath |
| 8 | TotRmsAbvGrd |
| 9 | YearBuilt |

Well, the most correlated feature to Sale Price is... Sale Price?!? Of course. For the other 9, they are as listed. Here is a short description of each. (Thank you, data\_description.txt!)

OverallQual: Rates the overall material and finish of the house (1 = Very Poor, 10 = Very Excellent)

GrLivArea: Above grade (ground) living area square feet

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

TotalBsmtSF: Total square feet of basement area

1stFlrSF: First Floor square feet

FullBath: Full bathrooms above grade

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

YearBuilt: Original construction date

Let's take a look at how each relates to Sale Price and do some pre-cleaning on each feature if necessary.

**Impute Missing Data and Clean Data**

Important questions when thinking about missing data:

How prevalent is the missing data?

Is missing data random or does it have a pattern?

The answer to these questions is important for practical reasons because missing data can imply a reduction of the sample size. This can prevent us from proceeding with the analysis. Moreover, from a substantive perspective, we need to ensure that the missing data process is not biased and hiding an inconvenient truth.

Let's combine both training and test data into one dataset to impute missing values and do some cleaning.

Missing Ratio

PoolQC 99.585921

MiscFeature 96.342305

Alley 93.788820

Fence 80.676329

MasVnrType 59.765355

FireplaceQu 47.412008

LotFrontage 17.805383

GarageType 5.590062

GarageYrBlt 5.590062

GarageFinish 5.590062

GarageQual 5.590062

GarageCond 5.590062

BsmtFinType 22.622498

BsmtExposure 2.622498

BsmtFinType 12.553485

BsmtCond 2.553485

BsmtQual 2.553485

MasVnrArea 0.552105

Electrical 0.069013

Train data size is : (1157, 80)

Test data size is : (292, 79)

Combined dataset size is : (1449, 79)

**Imputing Missing Values**

PoolQC : data description says NA means "No Pool"

MiscFeature : data description says NA means "no misc feature"

Alley : data description says NA means "no alley access"

Fence : data description says NA means "no fence"

FireplaceQu : data description says NA means "no fireplace"

LotFrontage : Since the area of each street connected to the house property most likely have a similar area to other houses in its neighborhood , we can fill in missing values by the median LotFrontage of the neighborhood.

GarageType, GarageFinish, GarageQual and GarageCond : Replacing missing data with "None".

GarageYrBlt, GarageArea and GarageCars : Replacing missing data with 0.

BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath and BsmtHalfBath: Replacing missing data with 0.

BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1 and BsmtFinType2 : For all these categorical basement-related features, NaN means that there isn't a basement.

MasVnrArea and MasVnrType : NA most likely means no masonry veneer for these houses. We can fill 0 for the area and None for the type.

MSZoning (The general zoning classification) : 'RL' is by far the most common value. So we can fill in missing values with 'RL'.

Utilities : For this categorical feature all records are "AllPub", except for one "NoSeWa" and 2 NA . Since the house with 'NoSewa' is in the training set, this feature won't help in predictive modelling. We can then safely remove it.

Functional : data description says NA means typical.

Electrical : It has one NA value. Since this feature has mostly 'SBrkr', we can set that for the missing value.

KitchenQual: Only one NA value, and same as Electrical, we set 'TA' (which is the most frequent) for the missing value in KitchenQual.

Exterior1st and Exterior2nd : Both Exterior 1 & 2 have only one missing value. We will just substitute in the most common string

SaleType : Fill in again with most frequent which is "WD"

MSSubClass : Na most likely means No building class. We can replace missing values with None

count 1449.000000

mean 56.832298

std 42.277695

min 20.000000

25% 20.000000

50% 50.000000

75% 70.000000

max 190.000000

Name: MSSubClass, dtype: float64

So, the average is a 56.832298 typeThis feature was interpreted as numerical when it is actually categorical. The types listed here are codes, not values. Thus, we need to feature transformation with this and many other features.

*Kitchen Quality:*

array(['Gd', 'TA', 'Ex', 'Fa'], dtype=object)

Ex: Excellent

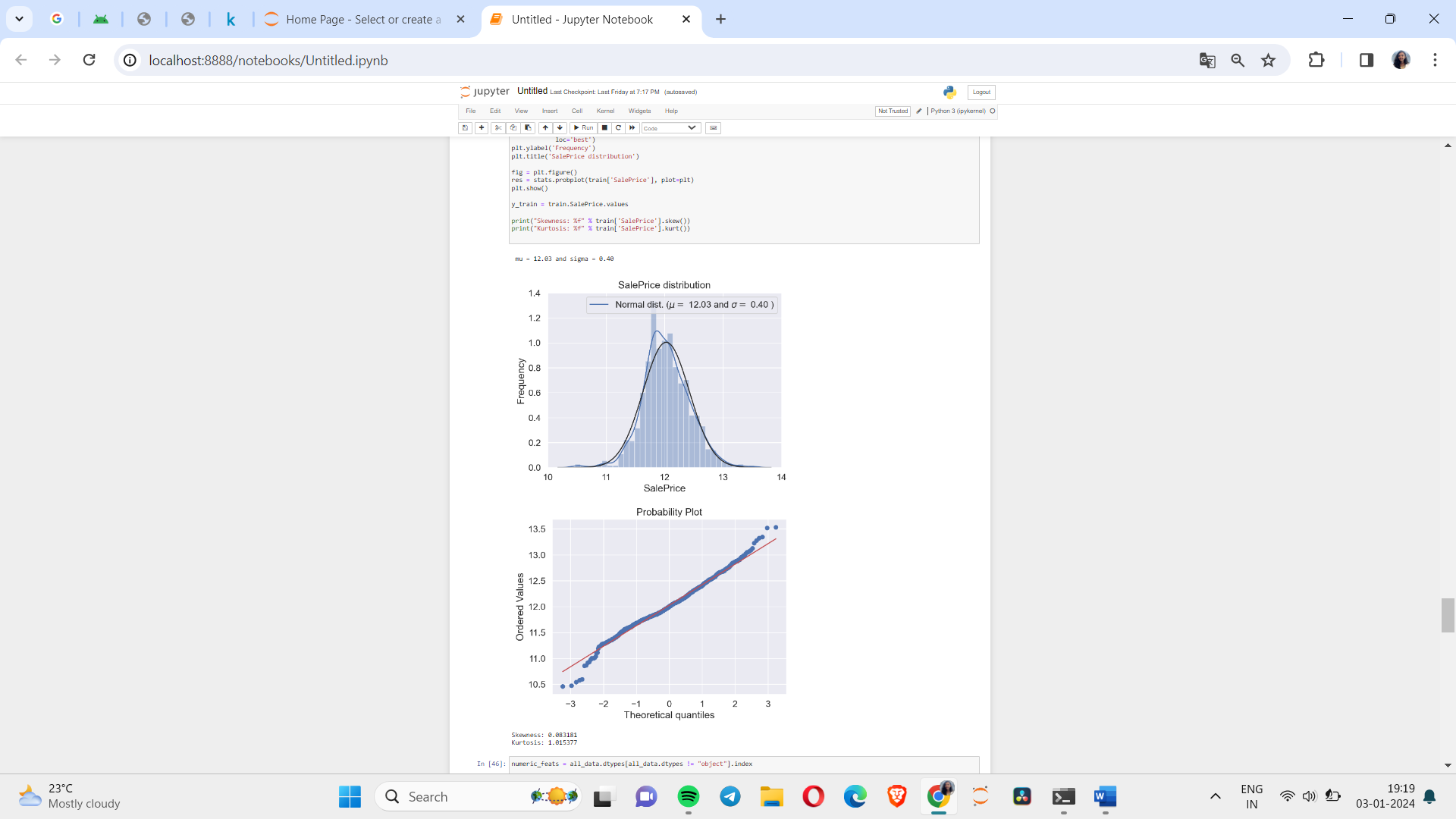
Gd: Good

TA: Typical/Average

Fa: Fair

Po: Poor

Is a score of "Gd" better than "TA" but worse than "Ex"? I think so, let's encode these labels to give meaning to their specific orders.



Skewed Features

MiscVal 24.388024

PoolArea 15.882700

LotArea 12.5952713

SsnPorch 10.253854

LowQualFinSF 8.966866

There are 59 skewed numerical features to Box Cox transform

Averaged base models score: 0.1135 (0.0124)

Stacking Averaged models score: 0.1135 (0.0117)

RMSLE score on train data:

0.0735590585928247