Housing Project.

Submitted by:

Harshita Panchamia

Internship 25

ACKNOWLEDGMENT

For this particular task, I referred the following websites and articles when stuck:

- https://www.codingem.com/python-maximum-recursion-depth/
- https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/
- https://medium.com/analytics-vidhya/hyperparameter-tuning-in-linear-regression-e0e0f1f968a1

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 objective was to model the prices of houses with the available independent variables.

Conceptual Background of the Domain Problem

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.

The company is looking at prospective properties to buy houses to enter the market. You 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. 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?

Technical Requirements

- Data contains 1460 entries each having 81 variables.
- Data contains Null values. You need to treat them using the domain knowledge and your own understanding.
- Extensive EDA has to be performed to gain relationships of important variable and price.
- Data contains numerical as well as categorical variable. You need to handle them accordingly.
- Need to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem
  - Linear Regression with Lasso, Ridge
  - Random Forest Regression
  - XGBoost

This is a Regression problem, where our end goal is to predict the Prices of House based on given data. I will be dividing my data into Training and Testing parts. A Regression Model will be built and trained using the Training data and the Test data will be used to predict the outcomes. This will be compared with available test results to find how

well the model has performed. The ‘r2’ score will be used to determine the best model among all the models.

- # Data Sources and their formats

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 data is provided in the CSV file.
The first 5 rows of data looks like:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	Mc
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	

]:

otArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2007	WD	Normal	128000
15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	10	2007	WD	Normal	268000
9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2007	WD	Normal	269790
11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	1	2010	COD	Normal	190000
16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2009	WD	Normal	215000

- # Data Description

MSSubClass: Identifies the type of dwelling involved in the sale.
20 1-STORY 1946 & NEWER ALL STYLES 75 2-1/2 STORY ALL AGES
30 1-STORY 1945 & OLDER 80 SPLIT OR MULTI-LEVEL
40 1-STORY W/FINISHED ATTIC ALL AGES 85 SPLIT FOYER
45 1-1/2 STORY - UNFINISHED ALL AGES 90 DUPLEX - ALL STYLES AND AGES

50 1-1/2 STORY FINISHED ALL AGES 150 1-1/2 STORY PUD - ALL AGES
60 2-STORY 1946 & NEWER 160 2-STORY PUD - 1946 & NEWER
70 2-STORY 1945 & OLDER 180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
190 2 FAMILY CONVERSION - ALL STYLES AND AGES
MSZoning : Identifies the general zoning classification of the sale.
A: Agriculture C : Commercial FV: Floating Village Residential I: Industrial RH: Residential High
Density
RL: Residential Low Density RP: Residential Low Density Park RM: Residential Medium Density
LotFrontage : Linear feet of street connected to property
LotArea : Lot size in square feet
Street: Type of road access to property
Grvl Gravel Pave Paved
Alley : Type of alley access to property
Grvl : Gravel Pave: Paved NA: No alley access
LotShape : General shape of property
Reg : Regular IR1: Slightly irregular IR2:Moderately Irregular IR3: Irregular
LandContour : Flatness of the property
Lvl: Near Flat/Level Bnk: Banked - Quick and significant rise from street grade to building
HLS Hillside - Significant slope from side to side Low: Depression
Utilities : Type of utilities available
AllPub : All public Utilities (E,G,W,& S) NoSewr : Electricity, Gas, and Water (Septic Tank)
NoSeWa : Electricity and Gas Only ELO: Electricity only
LotConfig : Lot configuration
Inside: Inside lot Corner: Corner lot CulDSac : Cul-de-sac FR2: Frontage on 2 sides of property
FR3 Frontage on 3 sides of property
LandSlope : Slope of property

Gtl : Gentle slope Mod: Moderate Slope Sev : Severe Slope
Neighborhood : Physical locations within Ames city limits
Blmngtn: Bloomington Heights Blueste : Bluestem BrDale : Briardale
BrkSide: Brookside
ClearCr: Clear Creek CollgCr : College Creek Crawfor : Crawford
Edwards:Edwards
Gilbert: Gilbert IDOTRR: Iowa DOT and Rail Road MeadowV:
Meadow Village
Mitchel: Mitchell
Names: North Ames NoRidge: Northridge NPkVill: Northpark Villa
NridgHt: Northridge
Heights
NWAmes: Northwest Ames OldTown : Old Town SWISU: South &
West of Iowa State
University
Sawyer: Sawyer SawyerW: Sawyer West Somerst: Somerset
StoneBr: Stone Brook
Timber: Timberland Veenker: Veenker
Condition1: Proximity to various conditions
Artery: Adjacent to arterial street Feedr: Adjacent to feeder street
Norm: Normal
RRNn: Within 200' of North-South Railroad RRAn: Adjacent to
North-South Railroad
PosN: Near positive off-site feature--park, greenbelt, etc.
PosA: Adjacent to postive off-site feature RRNe: Within 200' of East-
West Railroad
RRAe: Adjacent to East-West Railroad
Condition2: Proximity to various conditions (if more than one is present)
Artery: Adjacent to arterial street Feedr: Adjacent to feeder street
Norm: Normal
RRNn: Within 200' of North-South Railroad RRAn: Adjacent to
North-South Railroad
PosN: Near positive off-site feature--park, greenbelt, etc. PosA:
Adjacent to postive off-site

feature
RRNe: Within 200' of East-West Railroad RRAe: Adjacent to East-West Railroad
BldgType : **Type of dwelling**
1Fam: Single-family Detached 2FmCon: Two-family Conversion; originally built as one-family
dwelling
Duplx: Duplex TwnhsE: Townhouse End Unit TwnhsI: Townhouse Inside Unit
HouseStyle : Style of dwelling
1Story:One story 1.5Fin:One and one-half story: 2nd level finished
1.5Unf:One and one-half story: 2nd level unfinished 2Story:Two story
2.5Fin: Two and one-half story: 2nd level finished 2.5Unf: Two and one-half story: 2nd level
unfinished
SFoyer Split Foyer SLvl Split Level
OverallQual : Rates the overall material and finish of the house
10 Very Excellent 9 Excellent 8 Very Good 7 Good 6 Above Average
5 Average 4 Below Average 3 Fair 2 Poor 1 Very Poor
OverallCond : Rates the overall condition of the house
10 Very Excellent 9 Excellent 8 Very Good 7 Good 6 Above Average
5 Average 4 Below Average 3 Fair 2 Poor 1 Very Poor
YearBuilt : Original construction date
YearRemodAdd : Remodel date (same as construction date if no remodeling or additions)
RoofStyle : Type of roof
Flat: Flat Gable: Gable Gambrel: Gabrel (Barn) Hip: Hip Mansard: Mansard Shed : Shed
RoofMatl: Roof material
ClyTile: Clay or Tile CompShg: Standard (Composite) Shingle Membran: Membrane
Metal: Metal Roll: Roll Tar&Grv: Gravel & Tar WdShake: Wood Shakes WdShngl: Wood
Shingles

Exterior1st: Exterior covering on house
AsbShng Asbestos Shingles AsphShn Asphalt Shingles BrkComm Brick Common
BrkFace: Brick Face CBlock: Cinder Block CemntBd Cement Board HdBoard: Hard Board
ImStucc: Imitation Stucco MetalSd Metal Siding Other Other Plywood Plywood
PreCast PreCast Stone: Stone Stucco: Stucco VinylSd: Vinyl Siding Wd Sdng Wood Siding WdShing Wood Shingles
Exterior2nd : Exterior covering on house (if more than one material)
AsbShng Asbestos Shingles AsphShn Asphalt Shingles BrkComm Brick Common
BrkFace: Brick Face CBlock: Cinder Block CemntBd Cement Board HdBoard Hard Board
ImStucc Imitation Stucco MetalSd Metal Siding Other: ther Plywood Plywood PreCast PreCast Stone Stone Stucco Stucco VinylSd: Vinyl Siding Wd Sdng Wood Siding
WdShing Wood Shingles
MasVnrType : Masonry veneer type
BrkCmn Brick Common BrkFace Brick Face CBlock Cinder Block None None
Stone Stone
MasVnrArea: Masonry veneer area in square feet
ExterQual : Evaluates the quality of the material on the exterior
Ex Excellent Gd Good TA Average/Typical Fa Fair Po Poor
ExterCond: Evaluates the present condition of the material on the exterior
Ex Excellent Gd Good TA Average/Typical Fa Fair Po Poor
Foundation: Type of foundation
BrkTil Brick & Tile CBlock Cinder Block PConc Poured Contrete Slab Slab Stone Stone Wood Wood
BsmtQual : Evaluates the height of the basement
Ex Excellent (100+ inches) Gd Good (90-99 inches) TA Typical (80-89 inches)

Fa Fair (70-79 inches) Po Poor (<70 inches) NA No Basement
BsmtCond : Evaluates the general condition of the basement
Ex Excellent (100+ inches) Gd Good (90-99 inches) TA Typical (80-89 inches)
Fa Fair (70-79 inches) Po Poor (<70 inches) NA No Basement
BsmtExposure : Refers to walkout or garden level walls
Gd Good Exposure Av Average Exposure (split levels or foyers typically score average or
above)
Mn Mimimum Exposure No No Exposure NA No Basement
BsmtFinType1: Rating of basement finished area
GLQ Good Living Quarters ALQ Average Living Quarters BLQ Below Average Living Quarters
Rec Average Rec Room LwQ Low Quality Unf Unfinshed NA No Basement
BsmtFinSF1: **Type 1 finished square feet**
BsmtFinType2: Rating of basement finished area (if multiple types)
GLQ Good Living Quarters ALQ Average Living Quarters BLQ Below Average Living Quarters
Rec Average Rec Room LwQ Low Quality Unf Unfinshed NA No Basement
BsmtFinSF2: Type 2 finished square feet
BsmtUnfSF: Unfinished square feet of basement area
TotalBsmtSF : Total square feet of basement area
Heating: Type of heating
Floor : Floor Furnace GasA: Gas forced warm air furnace GasW: Gas hot water or steam heat
Grav Gravity furnace OthW : Hot water or steam heat other than gas Wall Wall furnace
HeatingQC : Heating quality and condition
Ex Excellent Gd Good TA Average/Typical Fa Fair Po Poor
CentralAir : Central air conditioning
N No Y Yes
Electrical: Electrical system
SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix Mixed
1stFlrSF : First Floor square feet
2ndFlrSF : Second floor square feet
LowQualFinSF : Low quality finished square feet (all floors)
GrLivArea : Above grade (ground) living area square feet
BsmtFullBath : Basement full bathrooms
BsmtHalfBath: Basement half bathrooms
FullBath: Full bathrooms above grade
HalfBath: Half baths above grade
Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
Kitchen: Kitchens above grade
KitchenQual : Kitchen quality
Ex Excellent Gd Good TA Typical/Average Fa Fair Po Poor
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
Functional : Home functionality (Assume typical unless deductions are warranted)
Typ Typical Functionality
Min1: Minor Deductions 1 Min2: Minor Deductions 2 Mod: Moderate Deductions
Maj1: Major Deductions 1 Maj2: Major Deductions 2 Sev: Severely Damaged Sal: Salvage
only
Fireplaces : Number of fireplaces
FireplaceQu : Fireplace quality
Ex Excellent - Exceptional Masonry Fireplace
Gd Good - Masonry Fireplace in main level
TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove NA No Fireplace

GarageType: Garage location
2Types More than one type of garage
Attchd Attached to home Basment Basement Garage
BuiltIn Built-In (Garage part of house - typically has room above garage)
CarPort Car Port Detchd Detached from home NA No Garage
GarageYrBlt: Year garage was built
GarageFinish: Interior finish of the garage
Fin Finished RFn Rough Finished Unf Unfinished NA No Garage
GarageCars: Size of garage in car capacity
GarageArea: Size of garage in square feet
GarageQual: Garage quality
Ex: Excellent Gd: Good TA Typical/Average Fa: Fair Po: Poor NA: No Garage
GarageCond: Garage condition
Ex : Excellent Gd : Good TA: Typical/Average Fa: Fair Po: Poor NA: No Garage
PavedDrive: Paved driveway
Y Paved P Partial Pavement N Dirt/Gravel
WoodDeckSF: Wood deck area in square feet
OpenPorchSF: Open porch area in square feet
EnclosedPorch: Enclosed porch area in square feet
3SsnPorch: Three season porch area in square feet
ScreenPorch: Screen porch area in square feet
PoolArea: Pool area in square feet
PoolQC: Pool quality
Ex Excellent Gd Good TA Average/Typical Fa Fair NA No Pool
Fence: Fence quality
GdPrv Good Privacy MnPrv Minimum Privacy
GdWo Good Wood MnWw Minimum Wood/Wire NA No Fence
MiscFeature: Miscellaneous feature not covered in other categories
Elev Elevator Gar2 2nd Garage (if not described in garage section)
Othr Other Shed Shed (over 100 SF) TenC Tennis Court NA None
MiscVal: $Value of miscellaneous feature
MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)
SaleType : Type of sale
WD Warranty Deed - Conventional CWD Warranty Deed - Cash
VWD Warranty Deed - VA Loan New Home just constructed and sold
COD Court Officer Deed/Estate Con Contract 15% Down payment regular
terms
ConLw Contract Low Down payment and low interest ConLI Contract Low
Interest
ConLD Contract Low Down Oth Other
SaleCondition: Condition of sale
Normal Normal Sale Abnorml Abnormal Sale - trade, foreclosure, short sale
AdjLand Adjoining Land Purchase
Alloca Allocation - two linked properties with separate deeds, typically condo
with a garage unit
Family Sale between family members
Partial Home was not completed when last assessed (associated with New
Homes)

Data Preprocessing Done

This data has many null values present in it, which cannot be processed. Also, there ae lots of variables labelled as NaN, but they are actually not null values and they have a certain meaning,
For Example,

- NA in feature 'Alley' means No_Alley
- In case of PoolQC, NA means 'No Pool'

I've replaced them with actual variables.

After treating the null values, I found that there's some kind of order present in the data, hence it is ordinal in nature. And replaced the ordinal data with numeric one by using OrdinalEncoder and labelEncoder.
Then I checked if there are any outliers present in the data. MiscVal, GrLiveArea, LowQualityFinSF, GarageArea, TotalBsmtSF, LotArea, MasVsrArea, 1stFlrSF are the some of the columns which has outliers present in them.
To treat this outliers, we will use zscore.

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.

Hardware and Software Requirements and Tools Used

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, Ridge, Lasso, ElasticNet
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier,DecisionTreeRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.metrics import mean_squared_error,mean_absolute_error
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix,classification_report
from sklearn.metrics import r2_score
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.ensemble import GradientBoostingRegressor
```

These are all the pacakges used for building the model.

- # Testing of Identified Approaches (Algorithms)

1. Random Forest Regressor
   Score : 0.9784613384922712
   Mean absolute error : 17458.718717948715
   Mean squared error : 643990684.7064768
   Root mean squared error: 25376.971543241263
   r2 score : 0.8810763178491386

2. XGBoost Regressor
   Score : 0.9719628188682605
   Mean absolute error : 15466.27027578889
   Mean squared error : 500001357.1843611
   Root mean squared error: 22360.710122542197
   r2 score : 0.9076663624352046

3. Linear Regression
   Accuracy score : 0.8213234246452539
   mean squared error : 691757925.9923393
   mean absolute error : 19707.506079972773
   r2 Score : 0.8722552955039279

4. Lasso Regression
   Accuracy score : 0.8213234246394986
   mean squared error : 691757056.7205957
   mean absolute error : 19707.48676891729
   r2 score: 0.8722554560295364

5. Ridge Regression
   Accuracy score : 0.8213234246452409
   mean squared error : 691757824.045911
   mean absolute error : 19707.50478941513
   r2 score : 0.8722553143300459

6. ElasticNet
   Accuracy score : 0.8213234218055451
   mean squared error : 691710405.7308168
   mean absolute error : 19706.904115686328
   r2 score : 0.8722554560295364

CONCLUSION

Key Findings and Conclusions of the Study

- MS Sub Class seems to have the biggest impact on House Prices, followed by Basement Full Bath and Basement Half Bath
- Other than the Basement related features, Condition 2, Exterior Quality and Lot Area are some of the other important features.

Learning Outcomes of the Study in respect of Data Science

- All transformation must be done after splitting the data to test and train, otherwise the parameters are affected.

Limitations of this work and Scope for Future Work

- The biggest limitation I observed was that not all categories of a particular feature were available in the training data. So, if

there is a new category in the test data/new data, the model would not be able to identify the new categories.