FLIP ROBO

HOUSING PROJECT

Submitted by:

VAISHNAVI BALAJI

ACKNOWLEDGMENT

[1] Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study

[2] https://www.greenscapegroup.co.in/blog/indian-real-estate-sector-and-the-current-challenges/

[3] https://www.thebalancemoney.com/real-estate-what-it-is-and-how-it-works-3305882

[4] https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20will%20give%20you,data%20and%20predictions%20are%20biased.

[5] Lifelong Property Price Prediction: A Case Study for the Toronto Real Estate Market

INTRODUCTION

- Business Problem Framing

Housing has been basic security for humans for generations to come. But every requirement being fulfilled comes with its own challenges. Real Estate market forms a major section of the world’s economy. Like how every market utilises the analysis and results of Data Science, real estate market too uses Data Science tool to solve problems in the domain and help increase the revenue of the companies. It is also used to model marketing strategies and focus on changing trends in house sales and purchases. There are several algorithms in Machine Learning such as predictive modelling, market mix modelling, recommendation systems which can be implemented in for achieving business goals for housing.

Some of the challenges in Indian Real-Estate market are Unavailability of Land, long pending infrastructure projects, overpopulation, outdated building techniques. For

- Conceptual Background of the Domain Problem

Real property is nothing but a piece of land, any improvements on the land and the legal rights to use and operate the land granted to the property owner. Ownership of real estate includes the ability to sell, lease or enjoy the property.

Real Estate is the transactional sale of real property, including everything attached to the land example-natural resources, houses. Real Estate refers to the physical property rather than the rights belonging to the owner. It is an immovable property attached to a piece of land. There are five main categories of real estate which include residential, commercial, industrial, raw land, and special use. Investing in real estate includes purchasing a home, rental property, or land. Indirect investment in real estate can be made via REITs or through pooled real estate investment. A real estate agent is a licensed professional who arranges real estate transactions, matching buyers and sellers and acting as their representatives in negotiations.

There are many types of real estate. Residential real estate: Any property used for residential purposes. Examples include single-family homes, condos, cooperatives, duplexes, townhouses, and multifamily residences. Commercial real estate: Any property used exclusively for business purposes, such as apartment complexes, gas stations, grocery stores, hospitals, hotels, offices, parking facilities, restaurants, shopping centres, stores, and theatres. Industrial real estate: Any property used for manufacturing, production, distribution, storage, and research and development. Land: Includes undeveloped property, vacant land, and agricultural lands such as farms, orchards, ranches, and timberland. Special purpose: Property used by the public, such as cemeteries, government buildings, libraries, parks, places of worship, and schools.

• Review of Literature

The real estate business is heavily dependent on finance because it is a capital-intensive sector of the economy. In the context of real estate development, building, operation, circulation, and consumption, the term "real estate finance linked to real estate" is used to refer generally to investments, financing, and related financial services provided through channels of currency circulation and credit. Real estate finance realises and strengthens financial operations while assisting the real estate industry's growth through the unique carrier of real estate, which serves as the primary source of financial risks.

China currently has a real estate finance system that was first built on the foundation of commercial banks and serves real estate development and housing consumption. The housing vestment, domestic demand, and national economic growth have all benefited from the real estate finance industry's assistance to urban inhabitants in their home purchases. But as real estate finance has grown quickly, a number of issues have also come to light, including a flawed market, a lone financing source, and a limited capacity for product innovation. In comparison to bank deposit rates and personal house loan rates, rental income is superior on one hand. On the other hand, the index and the housing price are both still rising quickly. When other investment avenues are barred, a significant portion of social funds shift to property purchases to sustain and increase value, which exacerbates the financial risks associated with real estate.

Due to the real estate industry's influence on the total national economy, the real estate market also carries significant financial risks that could jeopardise the safety of the entire financial system and potentially compromise the macroeconomic stability. Therefore, it is crucial from both a theoretical and practical standpoint to advance research on real estate financial risk.

Starting with the land rent theory, domestic scholars researched the financial risk associated with real estate before examining the connections between macroeconomics, financial systems, and real estate at both the macro and micro levels. This essay initially explains the idea of real estate financial risk and its various classifications before summarising the study on real estate bubbles, the creation and transmission of real estate financial risk, as well as the management of real estate financial risk.

• Motivation for the Problem Undertaken

Surprise Housing is a US based housing company and they have decided to make their mark in the Australian Market. The company makes use of data analytics to buy houses at lower rates and sell them at an upscaled price.

The company requires a model wherein prices of properties are predicted which can thereby help them in purchasing properties and implementing their business strategies accordingly.

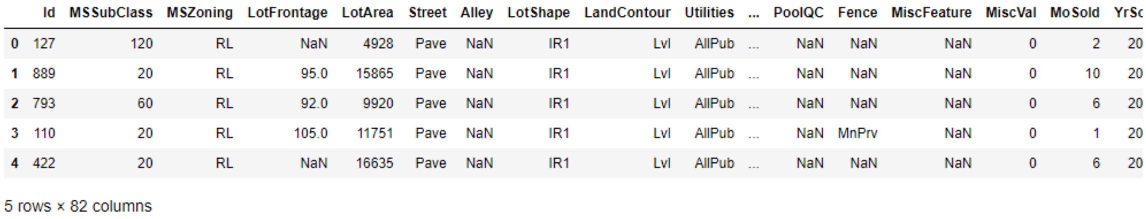
Analytical Problem Framing

- ### Mathematical/ Analytical Modelling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- ### Data Sources and their formats

US based housing company Surprise Housing has retrieved a dataset from the sale of houses in Australia. Data is in Comma Separated Format (CSV). Data contains 1460 entries each having 81 variables. There are two datasets which are provided- train.csv and test.csv. Model training is performed on training data and model testing is done on testing data.



	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSo
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	2	20
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	10	20
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	6	20
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	NaN	MnPrv	NaN	0	1	20
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	6	20

5 rows × 82 columns

- Data Features
  - The training data has 1168 rows and 82 columns, whereas the test data has 292 rows and 80 columns.
  - There are 5558 total null values in training data and 1407 null values in test data.
  - Number of categorical columns in train data are 44 and number of numerical columns are 38.

- Data Inputs- Logic- Output Relationships
    - Floating Village Residential has more sale price followed by Residential Low Density, Residential High Density, Residential Medium Density and Commercial.
    - Paved Access to property has higher sale price compared to gravel.
    - No alley access gives higher sale price followed by Paved and gravel.
    - Moderately irregular properties give higher returns followed by irregular, slightly irregular and regular
    - Hill Side land contour has higher sale price followed by depressed contour, levelled and banked.
    - With respect to LotConfiguration CullDSac yields a higher sale price followed by Frontage on 3 sides of property, Corner, Inside and Frontage on 2 sides.
    - When a property has severe slope, it gives higher returns compared with Moderate Sloping and Gentle Sloping
    - North Ridge neighbourhood has highest sale price followed by Northridge Heights, Stone Brook and Veenker. Meadow Village neighbourhood has the lowest sale price of houses.
    - With respect to condition 1, houses which are adjacent to positive off-site feature have higher sale price followed by houses which are within 200' of North-South Railroad and houses which are near positive off-site feature--park, greenbelt, etc.
    - Townhouse end unit buildings have higher sale price followed by houses which are single-family detached, townhouse inside unit, townhouse inside unit and two-family Conversion; originally built as one-family dwelling
    - Two and one-half story: 2nd level finished buildings have higher sale price followed by 2 story and 1 story. One and one-

half story: 2nd level unfinished buildings have the least sale price.

- Houses with shed type of roof have high sale price followed by hip, flat, mansard, gabel and gambrel.
- Wood Shingle roof material houses have high sale price followed by Wood Shakes, membrane,tar and gravel, metal and composite shingle.
- Houses which are made of imitation stucoo,stone and cement board have high sale price.
- When the material of the exterior is excellent, the sale price is high followed by good, average and fair.
- Poured concrete foundation has high sale price followed by stone, wood, cinder block, brick and tile and slab.
- For heating type in houses, gas forced warm air furnace ranks high sale price followed by steam heat, hot water, wall furnace.
- Houses with heating quality Excellent have high selling price.
- Houses which have air conditioning have high selling price compared with houses which don't.
- Houses with electrical systems-Standard Circuit Breakers & Romex have high selling price followed by Fuse Box over 60 AMP and all Romex wiring (Average), 60 AMP Fuse Box and mostly Romex wiring (Fair).
- Excellent kitchen quality has proportionally high sale price followed by good, average and fair.
- With respect to fire place quality, houses which have poor quality fire place have lesser selling price compared to houses which do not have fire place.
- Built-in garage houses have high selling price followed by attached category, basement, 2types, detached and car port.
- Houses with paved driveway have higher selling price.
- Houses which do not have fence have the highest selling price.

- Houses which are just constructed and sold, have high selling price followed by house sales which have a contract 15% Down payment regular terms.
- For sale conditions, when new buildings were not completed when last assessed have high sale price followed by normal sales.

Hardware and Software Requirements and Tools Used

List of libraries used along with their functions-

1) Pandas- In pandas library the functions used are-
    - Pd.read function is used to load the data.
    - The DataFrame's information is printed via the info() method. The data includes the total number of columns, their labels, data kinds, memory use, range index, and the number of cells in each column (non-null values)
    - The DataFrame's data is described via the describe() method. If the DataFrame includes numerical data, each column's description will provide the following details, count is the total amount of non-empty values, mean – a measure of average value, the standard deviation, or std, min is the lowest value, 25 percent – the 25 percentile*, the fifty percentile is 50%, the 75% percentile* is at 75%, max is the highest value.
    - The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True for NULL values, and otherwise False.
    - Pandas DataFrame.dtypes attribute return the dtypes in the DataFrame. It returns a Series with the data type of each column.
    - DataFrame from Pandas. The DataFrame's dtypes are returned via the dtypes attribute. A Series with the data types of each column is what is returned.

2) Python scripts can be used to create 2D graphs and plots using the Matplotlib module. With features to control line styles, font attributes, formatting axes, and other features, it offers a module called pyplot that makes things simple for plotting.
3) A package called Seaborn uses Matplotlib as its foundation to plot graphs. In order to see random distributions, it will be used.
4) A free machine learning library for Python is called Scikit-learn. It supports Python's NumPy and SciPy libraries as well as a number of methods, including support vector machines, random forests, and k-neighbors.

Model/s Development and Evaluation

- Problem-solving approaches

After the analysis of the data, before the models are developed, the data is cleaned.

- The columns- "MiscFeature","Fence","PoolQC","FireplaceQu" had high number of missing values hence were eliminated from the column.
- In the given dataset, the columns with high number of missing values are MiscFeature, Fence, PoolQC and FireplaceQu, hence they are eliminated from the dataset.
- The rest 14 columns which have null values are imputed with mean of the column for numerical type and mode of the column for categorical type.
- When box plot was plotted for the data, as per the observation there were many datapoints beyond the range. Using the z-score the values above value 3 were eliminated.
- 40 categorical columns were encoded using the get_dummies function in pandas.

- Since there are several columns after encoding, the feature selection is done using Chi2 test and Principal Component Analysis.
- From the analysis it is observed that information can be retained with 12 features.
- Using the 12 features, models were trained using different algorithms and tested and their efficiency on basis of r2 score was analysed.
- After selection of the efficient model, hyperparameter tuning was performed and the model is saved.

• Testing of Identified Algorithms

The features and target are split using the train_test_split function in sklearn, and a loop is defined to obtain the random state value for the function which yields best results.

The algorithms which were used in training and testing the model are:

1. Linear Regression
2. Random Forest Regression
3. Decision Tree Regression
4. K Neighbours Regression
5. Support Vector Regressor
6. Gradient Boosting Regressor
7. XGB Regressor
8. AdaBoost Regressor

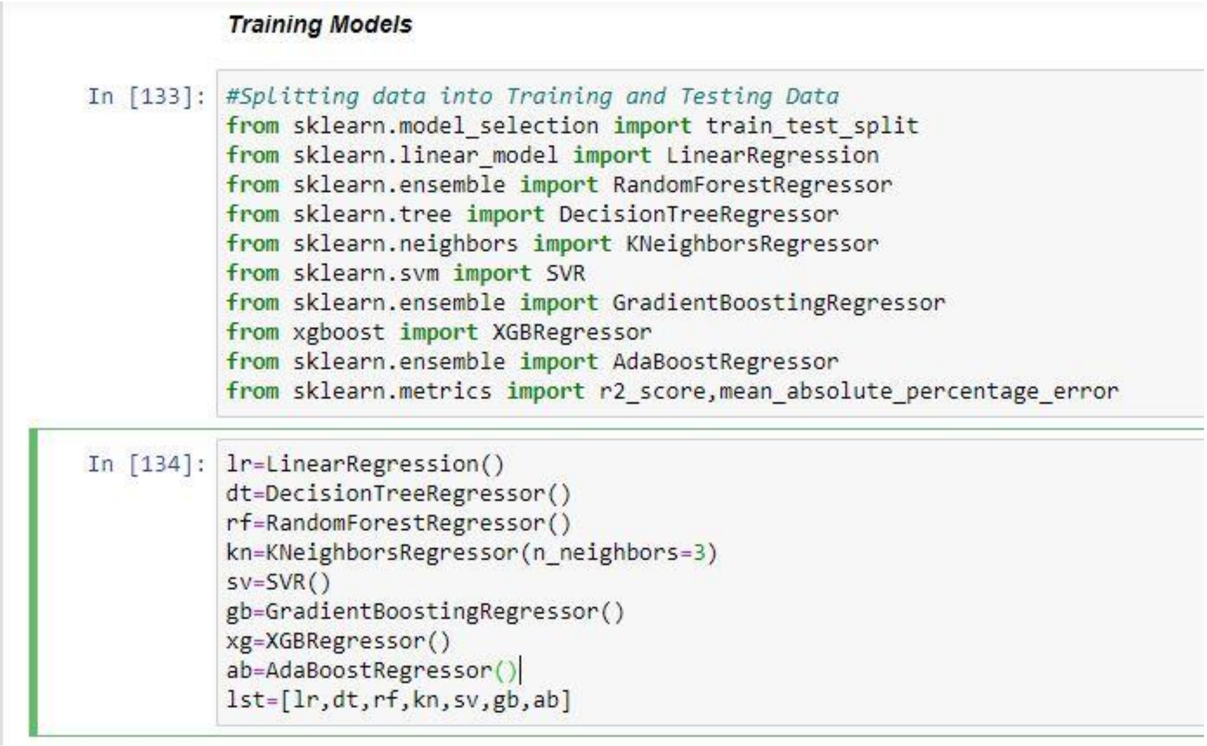
• Run and evaluate selected models

When the features and target were split using the train_test_split function and results were analysed for different random state values, the following observations were made:

- Linear Regression models had minor difference in r2 score of training and testing data. The r2 score for training data obtained is 0.8 and for testing data 0.81.

- Decision Tree Regressor had high difference of r2 score between training and test data.
- Random Forest Regressor has better results compared to Decision Tree Regressor.
- K Neighbours Regressor has approximately 0.20 difference of r2 score between training and testing data.
- Support Vector Regressor has negative r2 score for both training and test data.
- Gradient Boosting Regressor has an average difference of 0.16 in r2 score value between training and test data.
- Adaboost Regressor has an r2 score of 0.82 for training data and 0.73 for testing data.

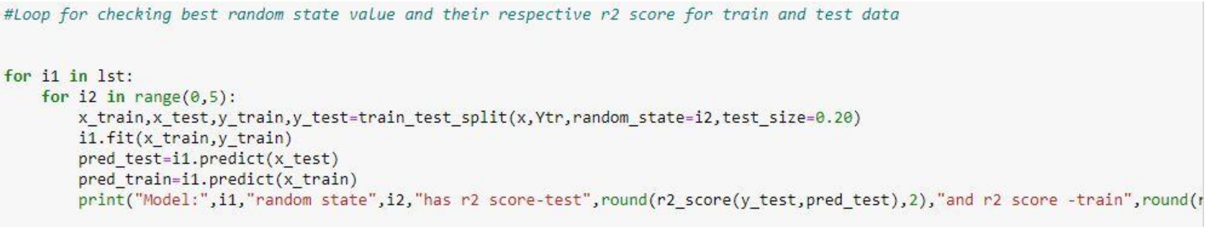
Code-



Training Models

```
In [133]: #Splitting data into Training and Testing Data
          from sklearn.model_selection import train_test_split
          from sklearn.linear_model import LinearRegression
          from sklearn.ensemble import RandomForestRegressor
          from sklearn.tree import DecisionTreeRegressor
          from sklearn.neighbors import KNeighborsRegressor
          from sklearn.svm import SVR
          from sklearn.ensemble import GradientBoostingRegressor
          from xgboost import XGBRegressor
          from sklearn.ensemble import AdaBoostRegressor
          from sklearn.metrics import r2_score,mean_absolute_percentage_error
```

```
In [134]: lr=LinearRegression()
          dt=DecisionTreeRegressor()
          rf=RandomForestRegressor()
          kn=KNeighborsRegressor(n_neighbors=3)
          sv=SVR()
          gb=GradientBoostingRegressor()
          xg=XGBRegressor()
          ab=AdaBoostRegressor()
          lst=[lr,dt,rf,kn,sv,gb,ab]
```



```
#Loop for checking best random state value and their respective r2 score for train and test data


for i1 in lst:
    for i2 in range(0,5):
        x_train,x_test,y_train,y_test=train_test_split(x,Ytr,random_state=i2,test_size=0.20)
        i1.fit(x_train,y_train)
        pred_test=i1.predict(x_test)
        pred_train=i1.predict(x_train)
        print("Model:",i1,"random state",i2,"has r2 score-test",round(r2_score(y_test,pred_test),2),"and r2 score -train",round(
```

Result:

```
Model: LinearRegression() random state 0 has r2 score-test 0.79 a
nd r2 score -train 0.81 and Mean Absolute Percentage Error is 0.1
4
Model: LinearRegression() random state 1 has r2 score-test 0.83 a
nd r2 score -train 0.8 and Mean Absolute Percentage Error is 0.15
Model: LinearRegression() random state 2 has r2 score-test 0.81 a
nd r2 score -train 0.8 and Mean Absolute Percentage Error is 0.15
Model: LinearRegression() random state 3 has r2 score-test 0.77 a
nd r2 score -train 0.81 and Mean Absolute Percentage Error is 0.1
5
Model: LinearRegression() random state 4 has r2 score-test 0.78 a
nd r2 score -train 0.81 and Mean Absolute Percentage Error is 0.1
4
Model: DecisionTreeRegressor() random state 0 has r2 score-test 0
.58 and r2 score -train 1.0 and Mean Absolute Percentage Error is
0.19
Model: DecisionTreeRegressor() random state 1 has r2 score-test 0
.53 and r2 score -train 1.0 and Mean Absolute Percentage Error is
0.21
Model: DecisionTreeRegressor() random state 2 has r2 score-test 0
.52 and r2 score -train 1.0 and Mean Absolute Percentage Error is
0.21
Model: DecisionTreeRegressor() random state 3 has r2 score-test 0
.57 and r2 score -train 1.0 and Mean Absolute Percentage Error is
0.2
Model: DecisionTreeRegressor() random state 4 has r2 score-test 0
.5 and r2 score -train 1.0 and Mean Absolute Percentage Error is
0.19
Model: RandomForestRegressor() random state 0 has r2 score-test 0
.81 and r2 score -train 0.97 and Mean Absolute Percentage Error i
s 0.13
Model: RandomForestRegressor() random state 1 has r2 score-test 0
.83 and r2 score -train 0.97 and Mean Absolute Percentage Error i
s 0.14
Model: RandomForestRegressor() random state 2 has r2 score-test 0
.78 and r2 score -train 0.97 and Mean Absolute Percentage Error i
s 0.16
Model: RandomForestRegressor() random state 3 has r2 score-test 0
.76 and r2 score -train 0.97 and Mean Absolute Percentage Error i
s 0.16
Model: RandomForestRegressor() random state 4 has r2 score-test 0
.78 and r2 score -train 0.97 and Mean Absolute Percentage Error i
s 0.12
Model: KNeighborsRegressor(n_neighbors=3) random state 0 has r2 s
core-test 0.65 and r2 score -train 0.84 and Mean Absolute Percent
age Error is 0.17
Model: KNeighborsRegressor(n_neighbors=3) random state 1 has r2 s
core-test 0.63 and r2 score -train 0.85 and Mean Absolute Percent
age Error is 0.18
Model: KNeighborsRegressor(n_neighbors=3) random state 2 has r2 s
core-test 0.66 and r2 score -train 0.83 and Mean Absolute Percent
age Error is 0.18
Model: KNeighborsRegressor(n_neighbors=3) random state 3 has r2 s
core-test 0.66 and r2 score -train 0.84 and Mean Absolute Percent
age Error is 0.19
```

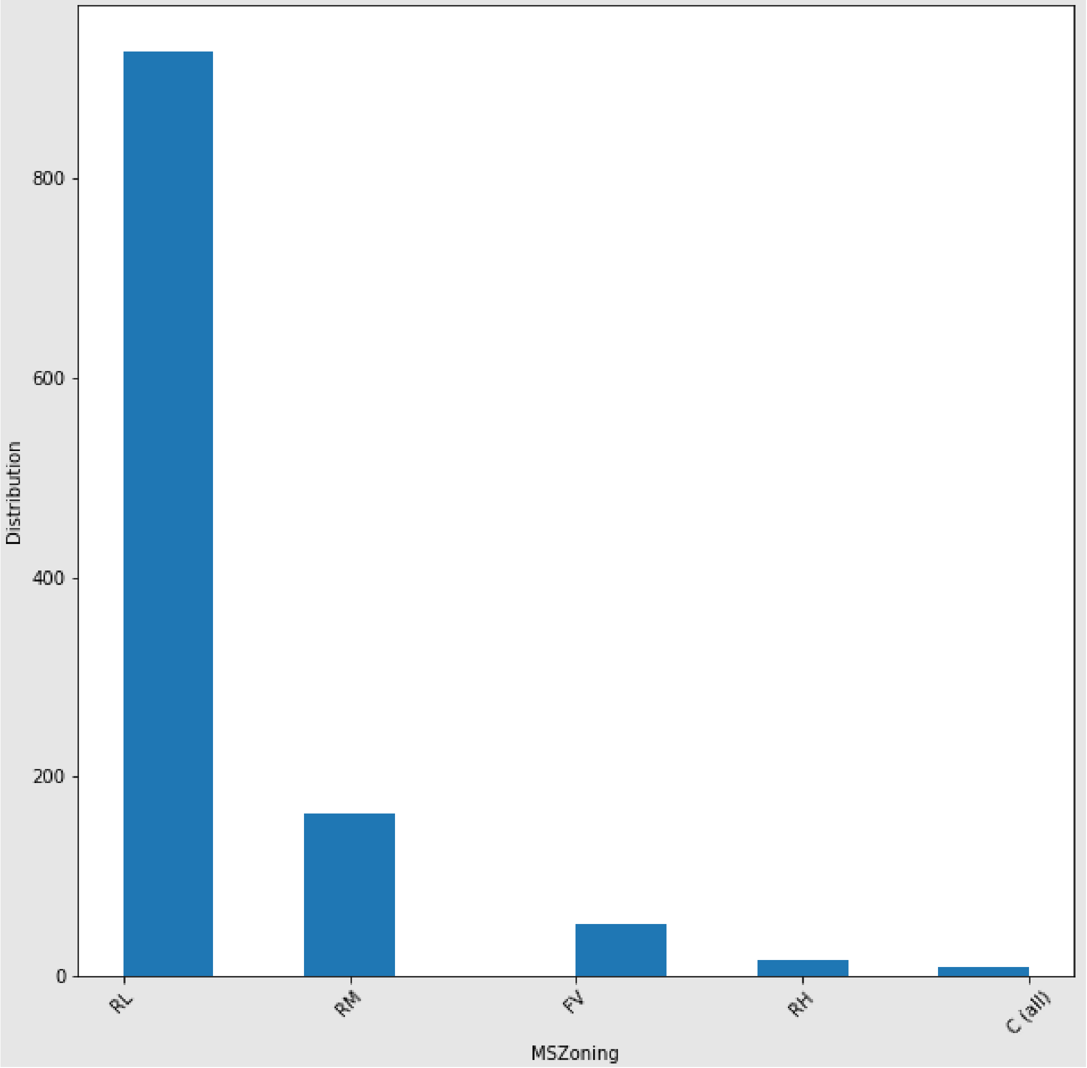
```
Model: KNeighborsRegressor(n_neighbors=3) random state 4 has r2 s
core-test 0.67 and r2 score -train 0.84 and Mean Absolute Percent
age Error is 0.15
Model: SVR() random state 0 has r2 score-test -0.03 and r2 score
-train -0.03 and Mean Absolute Percentage Error is 0.31
Model: SVR() random state 1 has r2 score-test -0.03 and r2 score
-train -0.05 and Mean Absolute Percentage Error is 0.34
Model: SVR() random state 2 has r2 score-test -0.02 and r2 score
-train -0.03 and Mean Absolute Percentage Error is 0.34
Model: SVR() random state 3 has r2 score-test -0.01 and r2 score
-train -0.03 and Mean Absolute Percentage Error is 0.32
Model: SVR() random state 4 has r2 score-test -0.06 and r2 score
-train -0.04 and Mean Absolute Percentage Error is 0.28
Model: GradientBoostingRegressor() random state 0 has r2 score-te
st 0.79 and r2 score -train 0.94 and Mean Absolute Percentage Err
or is 0.14
Model: GradientBoostingRegressor() random state 1 has r2 score-te
st 0.83 and r2 score -train 0.95 and Mean Absolute Percentage Err
or is 0.14
Model: GradientBoostingRegressor() random state 2 has r2 score-te
st 0.8 and r2 score -train 0.95 and Mean Absolute Percentage Erro
r is 0.16
Model: GradientBoostingRegressor() random state 3 has r2 score-te
st 0.75 and r2 score -train 0.95 and Mean Absolute Percentage Err
or is 0.16
Model: GradientBoostingRegressor() random state 4 has r2 score-te
st 0.78 and r2 score -train 0.95 and Mean Absolute Percentage Err
or is 0.13
Model: AdaBoostRegressor() random state 0 has r2 score-test 0.67
and r2 score -train 0.81 and Mean Absolute Percentage Error is 0.
19
Model: AdaBoostRegressor() random state 1 has r2 score-test 0.71
and r2 score -train 0.81 and Mean Absolute Percentage Error is 0.
2
Model: AdaBoostRegressor() random state 2 has r2 score-test 0.68
and r2 score -train 0.81 and Mean Absolute Percentage Error is 0.
22
Model: AdaBoostRegressor() random state 3 has r2 score-test 0.7 a
nd r2 score -train 0.83 and Mean Absolute Percentage Error is 0.1
8
Model: AdaBoostRegressor() random state 4 has r2 score-test 0.73
and r2 score -train 0.82 and Mean Absolute Percentage Error is 0.
15
```

- Key Metrics for success in solving problem under consideration

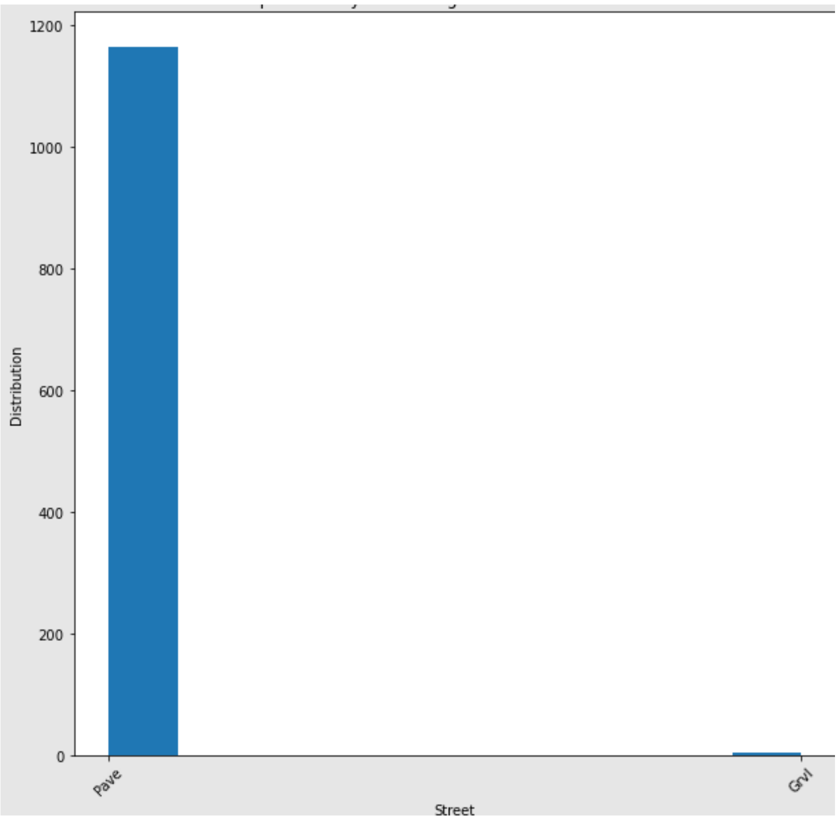
The 2 metrics used in this project were R2 score and Mean Absolute Percentage Error. An indicator of how much variance in a dependent variable is explained by one or more independent variables in a regression model is the R-Squared statistic.

R-squared is typically understood in the context of investing as the proportion of a fund's or security's movements that can be accounted for by changes in a benchmark index. An R-squared of 100% indicates that changes in the index (or the independent variable(s) of interest) perfectly explain changes in the security (or other dependent variables).

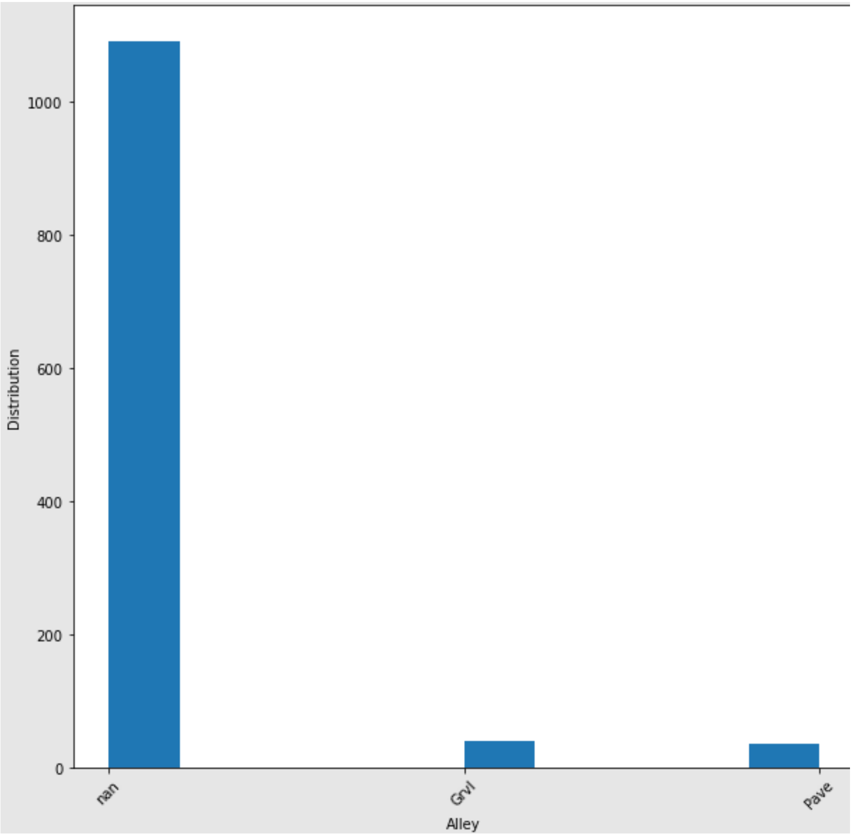
- Visualizations
  - MSzoning denotes the general classification of the sale. Majority of the data belong to the Residential Low-Density category followed by Residential Medium Density.



- Types of road access to the property has 2 values- Pave and Gravel, most of the data points have Pave type of road access.



- Many data points are null in type of alley access to property.



- LotShape is the general shape of the property. 740 houses have regular shape followed by 390 houses have slightly irregular shape.
- 1020 houses are almost levelled, followed by 30 houses are Banked - Quick and significant rise from street grade to building.
- All public utilities are available in the houses.
- 850 houses have Inside LotConfig followed by 210 houses having Corner LotConfig, 60 houses with CullDSac and 30 houses having FR2 LotConfig.
- 1100 houses have Gentle Sloping followed by 40 houses having Moderate Sloping and 20 houses having severe sloping.
- 230 houses have their neighbourhood in NorthPark Villa, followed by Sawyer, Edwards.
- With respect to Proximity to various conditions, 1000 houses are normal followed by 50 houses are adjacent to feeders,40 houses adjacent to arterial street.
- When it comes to type of dwelling, 980 houses are of single-family detached category followed by 100 Townhouse end unit.
- 580 houses are single storey, followed by 350 houses second storey, 130 houses are One and one-half story: 2nd level finished.
- Most of the houses have Gable type of roof followed by hip.
- Majority of the houses have roof type of Composite Shingle.
- Almost 570 houses have Exterior covering on house as Vinyl Siding followed by Plywood, Wood Siding, Brick Face and Cement Board.
- If there is more than one exterior covering, majority of the houses use Cement board followed by Plywood.
- 690 houses do not have Masonry Veneer type, 340 houses have brick face type followed by 85 houses with stone type.
- 730 houses have average exterior quality followed by 370 houses having good exterior quality.

- 1030 houses have average condition of the exterior material followed by 100 houses having a good condition.
- 530 houses have Foundation type of Cinder block followed by 520 houses have poured concrete.
- Around 530 houses have a basement height of 80-89 inches, followed by 400 houses with basement height of 90-99 inches.
- Majority of the houses have an average basement condition.
- With respect to walkout, 750 houses do not have exposure followed by 160 houses have average exposure.
- Almost 340 houses have unfinished basement area followed by 325 houses have good living quarters.
- 1140 houses have Gas forced warm air furnace type of heating.
- 570 houses have excellent heating condition followed by 340 houses having average heating condition.
- Most of the houses have central air conditioning.
- Majority of the houses have Standard Circuit Breakers & Romex type of Electrical Systems followed by Fuse Box over 60 AMP and all Romex wiring (Average)
- 570 houses have Average Kitchen Quality followed by 460 houses having Good Kitchen Quality.
- Most of the houses have typical functionality followed by Minor 1,2 deductions.
- 550 houses do not have a fireplace followed by 300 houses having good quality fireplace.
- 680 houses have attached garage followed by 310 houses have detached and 60 houses have built-in garage.
- 450 houses have unfinished garage followed by 340 houses have rough finished garage.
- Most of the garages are of average quality.
- Several houses have a paved driveway.
- Majority of the houses do not have a pool.
- 930 houses do not have a fence.
- 1000 houses are of conventional sale type followed by homes which were just constructed and sold.

- 990 houses had a normal sale.
- MSSubclass is the type of dwelling involved in the sale, it is a highly skewed feature having distribution from 0 to 200 having 4 peaks.
- LotFrontage is a highly skewed feature where most of the data distribution is between 20-100. LotFrontage determines the Linear feet of street connected to property.
- Lot Area is distributed between 0-1,60,000 where in majorly the data is distributed between 0-150000.
- OverallQual and Condition has majority data points from 4-8.
- 2000-2010 had a peak of houses being built as well as houses being remodelled.
- Most of the Basement Type1 Finished square feet is ranging from 0 till 1500 square feet.
- Total Basement Square feet is ranging from 200 -3000 with a peak from 0-100.
- Majority of the houses have first floor square feet from 500-1800 square feet.
- Second floor square feet is ranging from 0-1250 square feet.
- Above grade (ground) living area square feet is ranging from 800-2500.
- Most of the house do not have Basement full bathrooms, few houses have 1 full bathroom in basement.
- Majority of the houses have 2 full bathrooms followed by 1 and 3.
- Houses have bedroom of rating above 3 followed by 2 and 4.
- Many houses have room above grade 6 followed by 7 and 5.
- Most of the garages were built between 1950-1970
- Most of the garages constructed can accommodate 2 cars.
- The garages constructed have an area from 180-1000 square feet.
- Most of the houses were sold between the months of June to July followed by May.

  - Maximum number of houses were sold in 2007 followed by 2009, 2006 and 2008. 2010 has seen the least number of houses being sold.
  - Most of the data points for Sale Price is from 100000 to 250000.

- Interpretation of the Results

From the results of the modelling, the efficient model is Linear Regression Model for random state-2, test-size-0.20 where the r2 score obtained for training data is 0.80 and test data is 0.81 and mean absolute percentage error is 0.15.

CONCLUSION

- Learning Outcomes of the Study in respect of Data Science

In the given dataset the main features which affect the Sale price of the house are -OverallQual, GrLivArea, GarageCars, GarageArea, FullBath, TotalBsmtSF, TotRmsAbvGrd, YearBuilt, 1stFlrSF, ExterQual_Gd, Foundation_PConc, YearRemodAdd.

During the model development, the algorithms which have high r2 score for train data are Decision Tree Regressor, Random Forest Regressor, Gradient boosting Regressor followed by Linear Regression and Adaboost Regressor. But Decision Tree, Random Forest, KNeighbours, Adaboost,Gradient boosting regressor have overfitting problem which is reduced to a large extent in Linear Regression.

- Limitations of this work and Scope for Future Work

There are 82 columns in the training dataset wherein there are columns which can be consolidated, for example-Garage Quality and Garage condition, Basement quality and Basement Condition.

The linear regression model for predicting sale price of houses was hyper tuned which gave an increase in 0.02 in r2 score of the test data.