Analysis of United States Housing Prices

CA683 – Data Analytics and Data Mining Assignment

Satyam Ramawat (19210520)
MSc.in Computing,

School of Computing,
Dublin City University, Dublin, Ireland

Dalbir Singh Dhillon (19210611)
MSc.in Computing,

School of Computing,
Dublin City University, Dublin, Ireland

Ajit Kumar (19210438)
MSc.in Computing,

School of Computing,
Dublin City University, Dublin, Ireland

Rajneesh Sharma (19210497)
MSc.in Computing,

School of Computing,
Dublin City University, Dublin, Ireland

***Abstract*— The United States is a country that covers 50 states along with North America, Alaska in the northwest. United states also known as the world's market leader where most of the multi-national companies have its headquarters. More business leads to more population around the globe. In the U.S. many students come from different countries to pursue their higher studies and find a job opportunity, thus leads to housing crisis in the United States. With constrained dataset and information includes, a reasonable and composite data pre-processing, the inventive component designing technique is analyzed in this paper. This paper will focus on housing prices of the U.S., by the power of prediction with Regression modeling, considering the amenities.**

Keywords*— *Feature Engineering, Housing Prices, Prediction, Regression Modelling, Statistical Coefficient, Statistical correlation.

I. INTRODUCTION

The United States is the biggest country with a wider population growth in the world. It's known to be the world's best-developed country where it has various types of Multi-national companies and the world's finest universities and colleges. The high demand for the course and jobs in the U.S. leads to increment in immigrants. The increment in population raises the cost of living in the United States, where students and professionals are residing, which leads to the high unavailability of houses.

Unavailability of rented house made landlords demand more rent than usual. Where this trend has been followed by all over the United States. The continuous upward trend of the high house for rent along with higher price demand leads to housing crisis among the population.

House is charged differently with amenities like smoking allowance, pet allowance, car parking, and vice versa. Although these amenities are the basic requirement of the house due to crisis, rent varies from these features.

In this paper, the dataset of United States housing prices has been used. Developed and Evaluated the performance and the predictive power of a model trained and tested on collected data. Methods like regression and feature engineering have been followed in order to get a good fit. Used a well-fitted model in order to predict, how the monetary value of a house varies with the respect to different amenities located in the USA.

II. RELATED WORK

The paper "On the relation between local amenities and house price dynamics" discusses how the local amenities along with locality is correlated with housing prices. Authors Berecha et., had researched the correlation between volatility of housing prices and local amenities whereas, regression analysis has been done in volatility (by appreciating and depreciating market segmentation) and census location control with amenities. By this methodology, they have proved that property with high amenities experiences higher housing prices [1].

Paper by Stephen Law [2], finds that there is a strong link over street-based housing prices, where house prices street-based are far better than regional based. The author had used Multi-Level Regression model,

$$\underbrace{Y}_{Observed} = \underbrace{BX_i + \mu}_{Fixed} + \underbrace{\mu_i + \varepsilon_{ijk}}_{Random}$$

Where Y is the Observed, B implies coefficient for predictors, Xi is the predictor, u means MEAN, Ui shows the random effect of the local area, and Ei is the Error. Below Fig.1 shows how hierarchical multi-level model has been implemented in the domain of housing crisis.

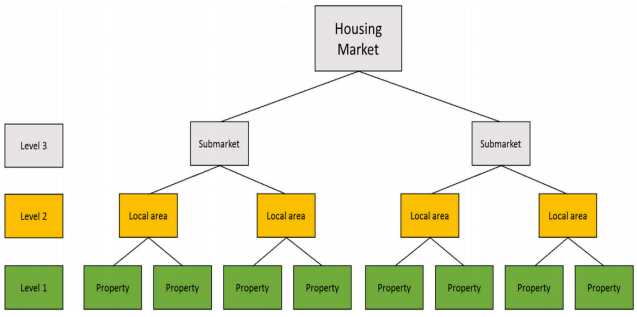

Fig. 1 Hierarchical multi-level regression model.

The author had used the dataset of Greater London to identify the housing crisis into three different levels as per Fig. 1. Level 1 is Property which derives from Level 2 which is Local Area, Level 2 derives from Level 3 that is Submarket, and whole multi-level hierarchy belongs to Housing Market. Additionally, the author had used a couple more datasets to predict and answer the question with prominent facts where datasets are, Regional sold price dataset which gives information about exchange prices between property buyer and seller and London street network is used to calculate the pedestrian street network for accessibility measures. In the end, the author had concluded

that street-based price is more accurate for prediction whereas regional-based prediction is less as compared to street-based by using the multi-level regression model.

Alexander and William explore the aftereffect of property upgrades in wide-scale US geologies [3], the outcome shows that the cost could be expanded 15% in the focal regions of huge urban areas, while less distortionary impact outside of downtown zones or in the smaller part of urban communities. This paper acknowledges the housing price crisis consequences or the impact takes more place in the urban or popular area of the city whereas the houses outside the city would have a lesser ripple effect.

[4] In the paper "Geographically weighted regression with a non-Euclidean distance metric" by authors Binbin Lu et., had built Geographical Regression model in order study the London house prices, where road network, Euclidean distance, and travel time metrics also considered.

GWR makes a point-wise adjustment concerning a 'bump of impact': around every regression point where closer outcomes have more impact in evaluating the local set of coefficients than outcomes further away. From each data point in regression i, it also measures the internal relationships, whereas weighted least square has been used to estimate the separate set of regression coefficients. For this estimation matrix expression is,

$$\hat{\beta}_i = \left(X^T W_i X\right)^{-1} X^T W_i y$$

where, βi is the vector, with m+1 local regression coefficient (βi0… βin); X is the matrix of columns of ones for the intercept of the independent variable, Wi represents the diagonal matrix contains geographical weight values for each perceived data at regression point(i) and; y represents dependent variable vector.

The result shows that collaboration of Geographical Weighted Regression model with Non-Euclidean metric not only enhance model fit efficiency but also it leverages to provide more useful insight information about relationships of each variable with each other in the data of house price.

Overall, this paper concludes that road network distance, Euclidean distance, and travel time metrics are also a key feature to predict housing prices to get good fit regression model.

The paper "A hybrid Regression Technique for House Price Prediction" by authors Sifie Lu et., has been produced for Kaggle challenge where the author ranked among the top 1% out of all competitors. For prediction authors had proposed the hybrid lasso and gradient boost regression model to predict house price [5].

Authors had done creative feature engineering, where they have introduced many new variables by finding the correlation of sale price for each variable. Below Fig.2 describes the major difference between sale price among nearby areas.

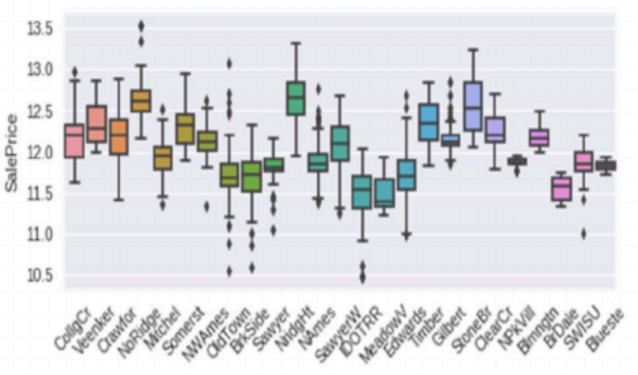

Fig. 2 Log Transformation of sale price

There are various Regression Algorithms available, but authors had focused on Ridge, Lasso, and Gradient Regression on training data, which lead to an outcome of the Hybrid Regression Model.

In the results authors mentioned that post-investigation few set of hybrid predictions over test data, they found 230 features are important from 280 features which affect house prices. The combination of Lasso 65% with Gradient 35% had achieved the best score. So, feature engineering plays a vital role to generate or identify the important feature according to the target variable.

From the below Fig.3, the combination of Lasso 65% and Gradient 35% generates the best results for test data is 0.11260.

Features	Hybrid Method	Score
230	0.65Ridge+0.35Xgb	0.11318
230	0.70Lasso+0.30Xgb	0.11294
230	0.65Lasso+0.35Xgb	**0.11260**
230	0.60Lasso+0.40Xgb	0.11277
230	0.3Ridge+0.35Lasso+0.35Xgb	0.11285
230	0.25Ridge+0.40Lasso+0.35Xgb	0.11283
280	0.65Ridge+0.35Xgb	0.11458
280	0.65Lasso+0.35Xgb	0.11539

Fig. 3, Hybrid Combination of Regression

From the regression, combination authors had introduced a new feature, which is, interest rate, population movement, an economic cycle that made sale price prediction more accurate.

Authors also discussed future work and mentioned, if there are features or generated from feature engineering like annual property tax, the crime rate of locality, cost of living, and marketing would help to get more accuracy in predictions. Also, the Random Forest algorithm may help to improve predictions accuracy.

III. Data Mining Methodology

For an efficient implementation with desired results, the research project must follow a data mining methodology that is best suited to solve the problem. There are multiple methodologies used for implementing data mining models based on the criteria, resources, and issues to be solved.

Some of the widely used data mining methodologies are KDD, CRISP-DM, and SEMMA approach. For this project, we have used CRIPS-DM (Cross-Industry Standard Process for Data Mining) methodology. CRISP-DM framework offers a systematic approach to Data Mining project planning. The steps performed in the CRISP-DM approach are as follows:

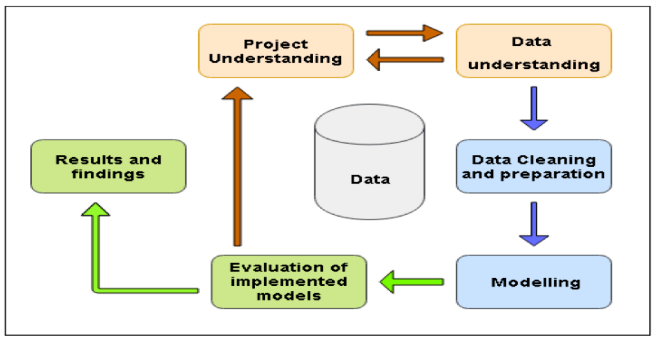

Fig. 4 CRISP-DM methodology followed in this project

A. *Project Understanding*

Prediction of housing rent considering various factors and amenities in the USA. Whereas hypotheses are, what is the base rent of a house with basic amenities? What are the factors that affect most to house rent?

B. *Data Understanding*

Data understanding is a key step in a data mining technique because to produce better performance, the attributes listed in the datasets need to be closely connected and co-related with each other. For our analysis, we have used the USA housing listing dataset which generated by Craigslist on 7th January 2020 and available at Kaggle [6]. Craigslist is the American housing platform where buyers and sellers interact with each other regarding the property deal. Below Fig. 5, describes the important details about the dataset where it has 384977 rows and 22 columns relatively. The dataset which we have used belongs to the category of Big Data, it belongs to Volume and Veracity from Big Four V's of Big Data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 384977 entries, 0 to 384976
Data columns (total 22 columns):
id                      384977 non-null int64
url                     384977 non-null object
region                  384977 non-null object
region_url              384977 non-null object
price                   384977 non-null int64
type                    384977 non-null object
sqfeet                  384977 non-null int64
beds                    384977 non-null int64
baths                   384977 non-null float64
cats_allowed            384977 non-null int64
dogs_allowed            384977 non-null int64
smoking_allowed         384977 non-null int64
wheelchair_access       384977 non-null int64
electric_vehicle_charge 384977 non-null int64
comes_furnished         384977 non-null int64
laundry_options         305951 non-null object
parking_options         244290 non-null object
image_url               384977 non-null object
description             384975 non-null object
lat                     383059 non-null float64
long                    383059 non-null float64
state                   384977 non-null object
dtypes: float64(3), int64(10), object(9)
memory usage: 64.6+ MB
```

Fig. 5 USA Housing Data Description

For further Data Analysis, we have used EDA functions from source code [8] to generate more details like unique values of non-categorical values, time series plot, and numeric EDA of data. Below Fig. 6 describes the statistics of simple distribution which include mean, Standard Deviation, and Quartiles of data.

Distribution of numeric data

	count	mean	std	min	25%	50%	75%	max
id	384977.0	7.040982e+09	8.800376e+06	7.003808e+09	7.035979e+09	7.043320e+09	7.048426e+09	7.051292e+09
price	384977.0	8.825722e+03	4.462200e+06	0.000000e+00	8.050000e+02	1.036000e+03	1.395000e+03	2.768307e+09
sqfeet	384977.0	1.059900e+03	1.915076e+04	0.000000e+00	7.500000e+02	9.490000e+02	1.150000e+03	8.388607e+06
beds	384977.0	1.905345e+00	3.494572e+00	0.000000e+00	1.000000e+00	2.000000e+00	2.000000e+00	1.100000e+03
baths	384977.0	1.480718e+00	6.180605e-01	0.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00	7.500000e+01
cats_allowed	384977.0	7.268902e-01	4.455574e-01	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
dogs_allowed	384977.0	7.079176e-01	4.547206e-01	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
smoking_allowed	384977.0	7.317710e-01	4.430381e-01	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
wheelchair_access	384977.0	8.211140e-02	2.745347e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
electric_vehicle_charge	384977.0	1.287090e-02	1.127177e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
comes_furnished	384977.0	4.812755e-02	2.140360e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
lat	383059.0	3.723349e+01	5.546171e+00	-4.353330e+01	3.345470e+01	3.764780e+01	4.113830e+01	1.020360e+02
long	383059.0	-9.270063e+01	1.653198e+01	-1.638940e+02	-1.007750e+02	-8.774510e+01	-8.117960e+01	1.726330e+02

Fig. 6 Distribution of Numeric Data

Let's see how much noise/null values are there in the dataset by using below function:

```
def noise(df):
  if len(df[df.isnull().any(axis=1)] != 0):
    print("\nPreview of data with null values:")
    missingno.matrix(df)
    plt.show()
noise(Your_Data_Frame)
```

From Fig. 7 we can see there are lots of null values in the column Laundry Options, Parking Options, and Latitude & Longitude. From Fig. 9 the price column of the dataset is skewed to right.

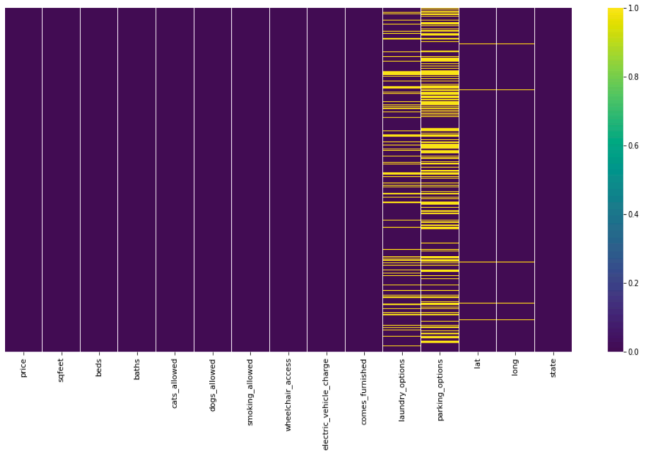

Fig. 7 Noise analysis of dataset

Fig. 8 shows the box plot of each column to showcase the outliers in the dataset whereas Fig. 8 shows the joint probability distribution. To check the probability distribution for random variables we have used joint probability distribution which looks for the relationship between two variables, where formal is:

$$\mathbf{f(x, y) = P(X=x, Y=y)}$$

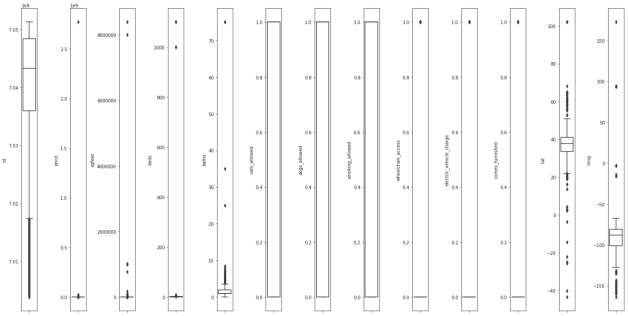

Fig. 8 Availability of Outliers in USA Housing Dataset

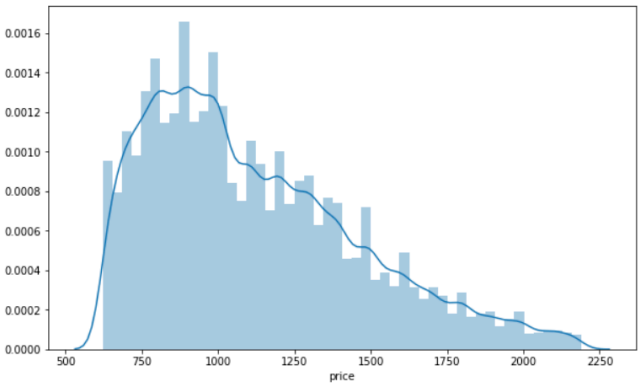

Fig. 9 Target variable price skewed to right

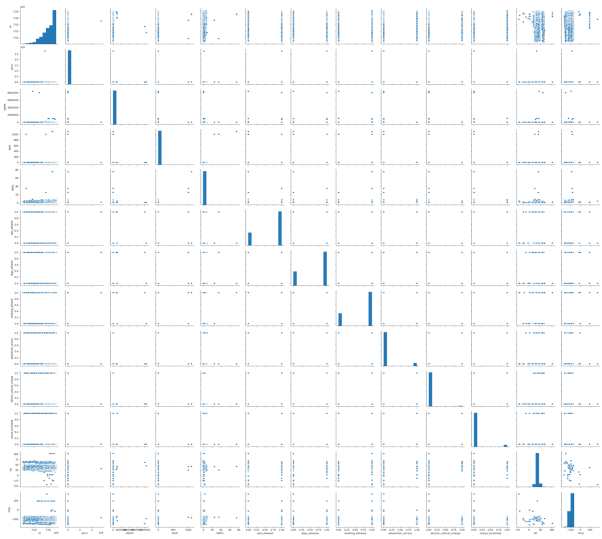

Fig. 10 To check pairwise joint probability distribution of numeric data

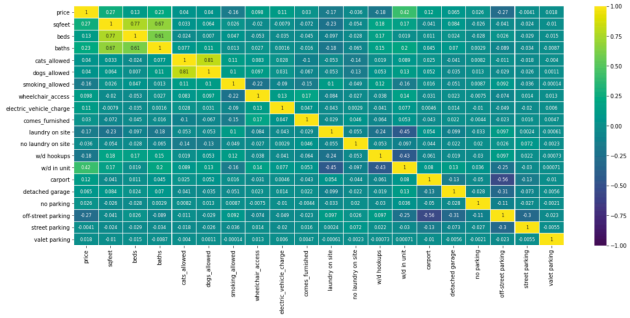

Fig.11 Correlation Matrix

C. *Data Cleaning And Preperation*

The data processing process includes all activities from the initial raw data to create the final data collection. Since we have scratched data from different sources there will be data irregularities like outliers, missing values, noisy data, and null values. Each cleaning task was done in Python (Jupyter notebook) using inbuilt packages.

The comprises of the column-like id, URL, region, region URL, price, type, square feet, beds, baths, cats allowed, dogs allowed, smoking allowed, wheelchair allowed, electric vehicles charge, furnished, laundry options, parking options, image URL, description, latitude, longitude, and state.

The first step is removing unnecessary columns and duplicates which do not influence the target variable i.e. price. The columns that have been removed are id, URL, region, region URL, type, description, image URL. We have used Pandas.Duplicated(), Pandas.Drop(). After performing operations, we only left with 318031 rows and 15 Columns.

```
housing.shape
(318032, 15)
```

Fig. 12 Dataset after removing duplicates and unwanted columns

The second step to clean the data is detecting and removing the outliers. The first step involved is detecting the outliers in

various columns, which need to be removed. An outlier is an extremely high or extremely low-value value in the data box plotting method is used to detect the outliers. Box plotting has been done for square feet, price, beds, and baths column. The value which is above the upper fence is calculated by (Q3 + (1.5 * IQR)) and the lower fence is calculated by (Q1 – (1.5 * IQR)) are termed as outliers. The box plot shows the data points which have a significantly higher value corresponding to other values in the column as per Fig. 8. We encountered a problem with the PRICE feature in the USA housing dataset where the lower fence was not visible, so unable to identify the lower fence outliers. By considering [9], we came to the conclusion that house prices in the USA are at least 623 so these values need to be removed as these values are outliers. The values which are crossing the upper fence and lower fence of the box plot have been removed, by the below line of code. And, we have removed floating numbers from bathroom features, like 1.5, 2.5, and 3.5 as we have assumed that it doesn't make sense because the bathroom could be 1, 2, 3, or any integer but not float number.

remove = housing[housing['price'] > 2190]. index | housing [(housing['price'] < 623)]. index

Above the same line of code has been used for all other features in order to remove the outliers by changing housing['X'] and where values 0 has been removed because they have not considered as NULL nor NOT AVAILABLE in the columns Baths, Price and Sq. Feet as per assumption. we have also analyzed there were 13% outliers found in the data. After removing outliers 253046 rows x 15 columns.

According to this article [10], We have used KNN Imputer to fill missing values in Longitude, Latitude feature. KNN Imputer is an efficient algorithm where it fills the missing values by considering its neighbours values. So in in case of longitude and latitude the missing location has been filled with just next to nearest available location.

We have checked laundry options and parking options, where data are missing more than 40%. In parking option and laundry option there is an option of "No Parking" and "No Laundry" available which say there is no option for these features, so we are assuming that we cannot impute missing values with "NO" or "Not Available", thus we have removed the pertaining rows where values are missing but not removed the column since Parking Option and Laundry Option are significant fields in the data which highly related to target variable "Price".

```
[272] housing.shape

      (147258, 13)
```

Fig. 13 Records in a dataset after cleaning completion

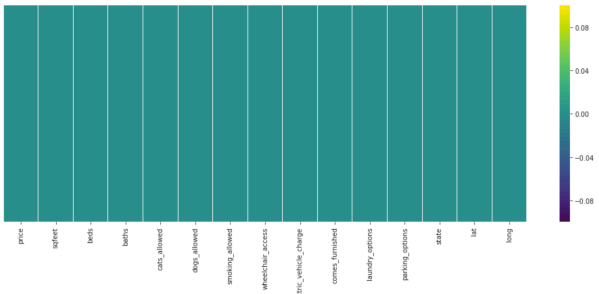

Fig. 13 Dataset after cleaning operations

For further pre-processing we have done one-hot encoding by using Pandas.get_dummies() method to convert categorical values into binary representation where each row has one feature with a value of 1, and the other features with value 0.Laundry Options and Parking Options has vast variety of categories as per given in Figure 14 and 15 and same has been converted to binary representation with the help of get_dummies() method which is helpful for machine learning algorithms to do a better job prediction.

```
housing['parking_options'].value_counts()

off-street parking    128216
attached garage        40094
carport                38871
detached garage        16852
street parking         15908
no parking              3176
valet parking            146
Name: parking_options, dtype: int64
```

Fig. 14 Categories available in Parking Options

```
housing['laundry_options'].value_counts()

w/d in unit           131783
w/d hookups            75568
laundry on site        58873
laundry in bldg        36103
no laundry on site      3624
Name: laundry_options, dtype: int64
```

Fig. 15 Categories available in Parking Options

laundry

	laundry on site	no laundry on site	w/d hookups	w/d in unit
7	0	0	1	0
9	0	0	1	0
10	0	0	1	0
17	1	0	0	0
18	1	0	0	0
...	...	...	...	...
253016	0	0	0	0
253019	0	0	0	0
253020	0	0	0	0
253029	0	0	0	0
253030	0	0	0	0

95405 rows × 4 columns

Fig. 16 One-Hot coding of Laundry Option variable

parking

	carport	detached garage	no parking	off-street parking	street parking	valet parking
7	0	0	0	1	0	0
9	0	0	0	1	0	0
10	0	0	0	1	0	0
17	0	0	0	0	1	0
18	0	0	0	0	1	0
...	...	...	...	...	...	...
253016	0	0	0	1	0	0
253019	0	0	0	1	0	0
253020	0	0	0	1	0	0
253029	0	0	0	1	0	0
253030	0	0	0	1	0	0

95405 rows × 6 columns

Fig. 17 One-Hot coding of Parking Option variable

The Figure 16 and 17 shows the how Laundry option variable and Parking option variable categories has been split into binary data. We have done this to make machine learning model to understand the data. The price column was skewed to right. Thus, a Log Transformation of Price has been generated and same have been used in the model later [12] and same can be seen in fig 18.

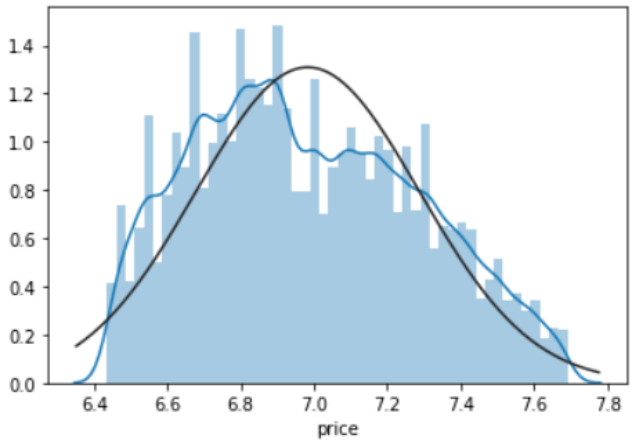

Fig. 18 Price after applying Log Transformation

Many machine learning algorithms perform best when features are on a scale that is comparatively similar and close to normal distribution. In this research StandardScaler is being used to standardizes the feature as it is beneficial for the regression type algorithm [20].

Before continuing with the modelling process dataset has been splitted into training and testing data. Training data holds up to 75% of the data to classify the trends concealed, while testing data carries 25% of the data.

D. Modelling

We need to define the purpose of this research to continue with the modelling process. The aim is to predict the house rent and various factors affecting the rent of the house in the USA. Thus, the research consists of different Supervised Machine Learning Regression algorithms like Linear Regression, Support vector machine, Random Forest Regressor, gradient booster and decision tree.

1) Linear Regression

Multiple linear regression is advanced version of simple linear regression which works with the multiple independent variable or multiple independent features to predict the dependent variable [13]. The equation of regression is nearly identical to the basic equation of regression, with only more variables.

$$Y'i = b0 + b1X1i + b2X2i$$

2) Random Forest Regressor

Random forest is a Supervised Learning algorithm that uses classification and regression techniques for ensemble learning. The key method is it builds lots of decision tree based on random data collection and random variables collection and it provides the dependent variable class based on other trees. The key benefit of using this algorithm on my dataset is that it manages the missing values so it can preserve the precision of the missing data, and the risk of overfitting the model is small, so when applying to the large-level dataset we can except high dimensionality[14].

3) Support Vector Machine

Support vector machine regression is derived from the classification algorithm known as the support vector machine. SVM find a hyperplane in N dimensions space that distinctly classifies the data points. It aims to minimize the upper limit of the generalization error, rather than the empirical error.Here we are using the same methods instead of separating the data it produces a hyperplane that is close to most of the points. So, the rent can be predicted [15].

4) Gradient Boosting

Gradient boosting can be used for both the regression and classification problem.it is a technique used for producing regression model consisting of collection of regressor [16].

5) Decision Tree Regressor

The Decision tree is an important statistical pattern recognition tool. The methodology is useful in analysing the relationship between house prices and characteristics of housing, defining important determinants of house prices, and forecasting house prices [17].

E. Evaluation

Once the data mining models are implemented, we need to evaluate the results. Different Supervised Machine Learning Regression algorithms like Linear Regression, Support vector machine, Random Forest Regressor, gradient booster, and decision tree are evaluated based on MAE (Mean absolute error) is the mean of the absolute value of errors, MSE (Mean square error) is the mean of the squared errors, RMSE (root mean square error) is the square root of the mean of the squared errors and R square score. Both the root mean square error (RMSE) and the mean absolute error (MAE) are regularly employed in model evaluation studies

[18]. we have used these metrics because MAE is the easiest to understand and it calculates the average error.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MSE measures the average squared error of our predictions. For each point, it calculates the square difference between the predictions and the target and then averages those values and it punishes the larger errors.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

RMSE is the square root of MSE. The square root is introduced to make the scale of the errors to be the same as the scale of targets.it is popular to measure the error rate of a regression model.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R2 is usually defined as the square of the correlation coefficient between observed and predicted values in a regression. R square summarizes the explanatory power of the regression model and is computed from the sum of squares terms [19].

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

All the above-mentioned metrics are very essential as they are used to calculate the error value and help to determine how well various algorithm predicted the rent of the house.

F. Results And Findings

All five different Supervised Machine Learning Regression algorithms like Linear Regression, Support vector machine, Random Forest Regressor, gradient booster, and decision tree are trained and tested against the dataset. The results of random forest and decision tree shows the best result with an accuracy of 64 and 57 percent respectively. In Random forest regressor, we have kept the number of the estimator as 100 we have tried with a different number of estimators like 300or 500 but accuracy was the same it was not changing because when the number of trees grows, forest efficiency is not necessarily any higher than previous forests, so doubling the number of trees is meaningless[21]. The below table displays the comparison of different models along with their MAE, MSE, RMSE, and accuracy.

Table 1. Comparison of individual models

Model	MAE	MSE	RMSE	Accuracy
Random Forest	0.37	0.37	0.61	0.64
Decision Tree	0.37	0.43	0.65	0.57
Gradient Boosting	0.58	0.56	0.75	0.44
Linear Regression	0.68	0.72	0.84	0.29
Support Vector Machine	0.67	0.72	0.85	0.28

Fig.19 shows the graph of actual vs predicted value of random forest algorithm.

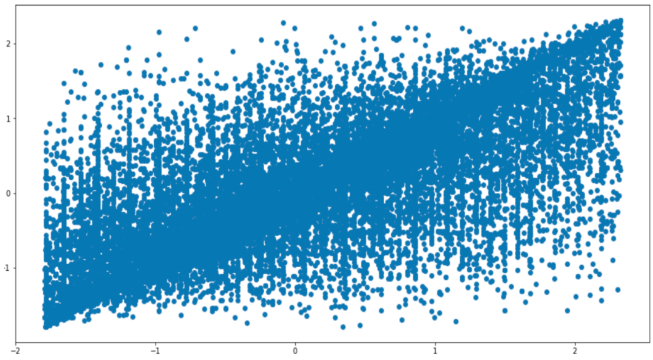

Fig 19. Random forest Predicted vs Actual value

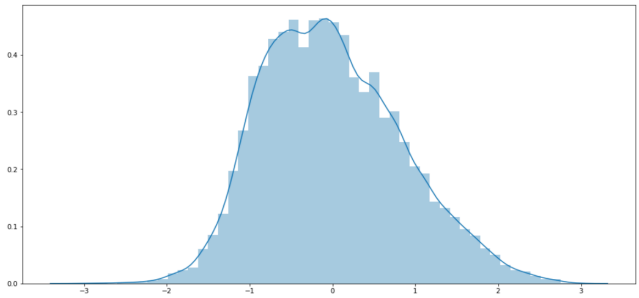

Fig 20. Residual Graph

Fig 20. Show residual graph which is the difference between the actual value and predicted value and it shows that the error is normally distributed.

The house rent is affected by various factors like w/d unit, square feet, electrical vehicle charger laundry on onsite, etc and the same can be seen in below table 2.

Table 2. Factors affecting house rent

Sr. No.	Features	Coefficient
1.	w/d unit	30.86
2.	Sq. feet	28.5
3.	Electric vehicle charger	6.3
4.	laundry on site	5.6
5.	Baths	3.9
6.	Wheelchair access	1.2
7.	Valet parking	1.1

Interpreting the coefficient which means that one unit increase in w/d unit increases the house rent by $30.86 per month, one unit increase in sq. feet increase the rent by $28.5 and one unit increase in electric vehicle charger increases the house rent by $6.3 etc. Thus, all the features shown in table 2 led to the increase of the rent per month.

IV. CONCLUSION AND FUTURE WORK

The objective of this paper was to forecast the rent increased per month in the USA and various factors which affects the rent of the house in the USA which has been addressed successfully using machine learning algorithms.

The model was evaluated based on the data used for modeling. We have used Supervised Machine Learning Regression algorithms to predict the rise of housing rent in the USA. The Random Forest regressor outperformed all the

individual model performance with better accuracy and fewer errors and the same can be seen in Fig. 21.

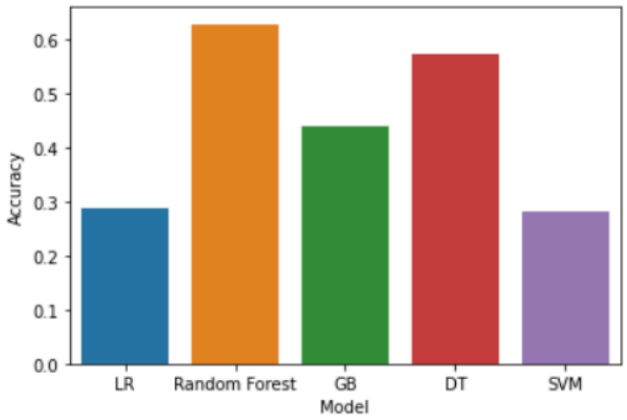

Fig. 21 Accuracy of different models.

Future works of this involve the prediction of housing rent per month state and region wise as we have features like state and longitude and latitudes of the places where the house is located depict the region so these features can be used for region and state-wise house price prediction. As well as improving the performance of all the different models by hyperparameter optimization also comes under the scope of future work.

REFERENCES

[1] Beracha, Eli and Gilbert, Ben and Kjorstad, Tyler and Womack, Kiplan S., On the Relation between Local Amenities and House Price Dynamics (June 27, 2016). Real Estate Economics (Forthcoming). Available at SSRN: https://ssrn.com/abstract=2801415

[2] Law, Stephen. (2017). Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London. Cities. 60. 166-179. 10.1016/j.cities.2016.08.008.

[3] Alexander N. Bogin & William M. Doerner, 2017. "Property Renovations and Their Impact on House Price Index Construction," FHFA Staff Working Papers 17-02, Federal Housing Finance Agency.

[4] Binbin Lu, Martin Charlton, Paul Harris & A. Stewart Fotheringham (2014) Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data, International Journal of Geographical Information Science, 28:4, 660-681, DOI: 10.1080/13658816.2013.865739

[5] Lu, Sifei & Li, Zengxiang & Qin, Zheng & Yang, Xulei & Goh, Rick. (2017). A hybrid regression technique for house prices prediction. 319-323. 10.1109/IEEM.2017.8289904. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[6] https://www.kaggle.com/austinreese/usa-housing-listings

[7] https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9

[8] https://gist.github.com/jiahao87/c97214065f996b76ab8fe4ca1964b2b5

[9] https://www.cbsnews.com/media/top-10-cheapest-us-cities-to-rent-an-apartment/

[10] Beretta, Lorenzo, and Alessandro Santaniello. "Nearest neighbor imputation algorithms: a critical evaluation." *BMC medical informatics and decision making* vol. 16 Suppl 3,Suppl 3 74. 25 Jul. 2016, doi:10.1186/s12911-016-0318-z

[11] https://towardsdatascience.com/the-tale-of-missing-values-in-python-c96beb0e8a9d

[12] https://medium.com/@ODSC/transforming-skewed-data-for-machine-learning-90e6cc364b0

[13] N. Bhagat, A. Mohokar, and S. Mane, "House Price Forecasting using Data Mining," Int. J. Comput. Appl., vol. 152, no. 2, pp. 23–26, 2016.

[14] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," Expert Syst. Appl., vol. 42, no. 1, pp. 259–268, 2015.

[15] T. B. Trafalis and H. Ince, "Support Vector Machine for Regression and Applications to Financial Forecasting. Machine Learning View project Kernel Methods View project SUPPORT VECTOR MACHINE FOR REGRESSION AND APPLICATIONS TO FINANCIAL FORECASTING," Proc. IEEE-INNS-ENNS Int. Jt. Conf. Neural Networks. IJCNN 2000. Neural Comput. New Challenges Perspect. New Millenn., no. x, pp. 348–353 vol.6, 2000.

[16] Y. Ganjisaffar, R. Caruana, and C. V. Lopes, "Bagging gradient-boosted trees for high precision, low variance ranking models," SIGIR'11 - Proc. 34th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., no. c, pp. 85–94, 2011.

[17] G. Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: A decision tree approach," *Urban Stud.*, vol. 43, no. 12, pp. 2301–2316, 2006.

[18] [1] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," Geosci. Model Dev., vol. 7, no. 3, pp. 1247–1250, 2014.

[19] D. L. J. Alexander, A. Tropsha, and D. A. Winkler, "Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models," J. Chem. Inf. Model., vol. 55, no. 7, pp. 1316–1322, 2015.

[20] https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

[21] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7376 LNAI, pp. 154–168, 2012.

[22] G. Gao, Z. Bao, J. Cao, A. K. Qin, T. Sellis, and Z. Wu, "Location-Centered House Price Prediction : A Multi-Task Learning Approach," pp. 1–14.