**Name**: - Kumari Versha Prajapati

**Reg.No.**: - 11910904

**Course code:** - INT247

**Section:** -KM018

**Roll No.:** - 03

**Introduction:**

The study on land price trend is deemed to be significant to support the decisions in urban planning. The real estate system is an unstable stochastic process. Investor’s decisions are based on the market trends to reap maximum returns. Developers are interested to know the future trends for their decision making. In order to accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data revealed that the prices show a non- linear characteristic. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyses and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project. house prices will be predicted given explanatory variables that cover many aspects of residential houses. As continuous house prices, they will be predicted with various regression techniques Ridge Regression. We will also perform EDA and RMSE to improve the prediction accuracy. The goal of this project is to create a regression model and a classification model that are able to accurately estimate the price of the house given the features. The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and high-rise buildings. Investments in Real Estate Industry has grown significantly high over the years and we have noticed a non-uniform pattern in terms of land pricing. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors.

**Literature Reviews:**

Machine learning is a form of artificial intelligence which compose available computers with the efficiency to be trained without being veraciously programmed. Machine learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data. Machine learning algorithms are broadly classified into three divisions, namely; Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer.

After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike, supervised learning, no teacher is provided that means no training will be given to the machine.

Therefore, machine is restricted to find the hidden structure in unlabeled data by our-self. Real Estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly.

Changes in the real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic index.

It is an insistent demand by current real estate industry to establish an easy-operate and logical scientific prediction model. But the real estate price is a chronological sequence with a particular statistic relationship which is difficult to be expressed by a predetermined function or equation. To this day, literature about research on machine learning prediction of house prices in India is extremely limited.

This paper provides a review of the usage of existing machine learning algorithms on two extremely different datasets and tries to implement this prediction engine for real-life usage by users. The findings indicate that using different algorithms can drastically change accuracy. Also, a poor dataset can negatively affect the predictions. Furthermore, it provides sufficient proof of what algorithm is best suitable for this task.

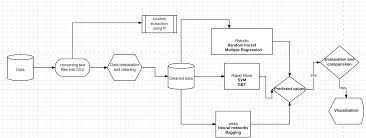
And this character makes it difficult to predict the real estate price. This kind of character can have a good effect to forecast the price of real estate. At present each framework may be moved towards innovation for the simplicity from claiming operations. The training framework will be moving towards e-taking. Individuals tend to move from the manual to robotized methodology. That primary goal of the this will be will anticipate that lodging cost with admiration to the plan of the clients. Those exhibit strategies may be A long procedure in which those customers necessities to contact the land operator.

The land operators give acceptable A suggestive on the lodging costs prediction. This strategy includes high hazard a direct result the land operator might furnish the bad data of the clients. They employments those straight relapse calculations should figure the cost. This analyses likewise utilized to foresee the best area for the clients to purchasing the houses.

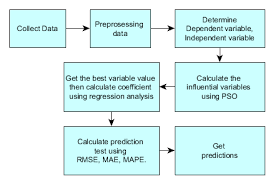
**Methodology:**

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted. The below document presents the implementation of price prediction project for the real estate markets and housing. Many algorithms are used here to effectively increase the accuracy percentage, various researchers have done this project and implemented the algorithms like Ridge Regression. These are considered as the base models and by the help of advanced data mining tools algorithms like a random forest, gradient boosted trees, multi-layer perceptron and ensemble learning models are used and prediction accuracy is attained in a higher rate. The results and evaluation of these models using the machine learning and advanced data mining tools like Weka, Rapid Miner etc.

The below passages describe about the methodology used in the real estate house price predictions and the architecture flow diagram is given.



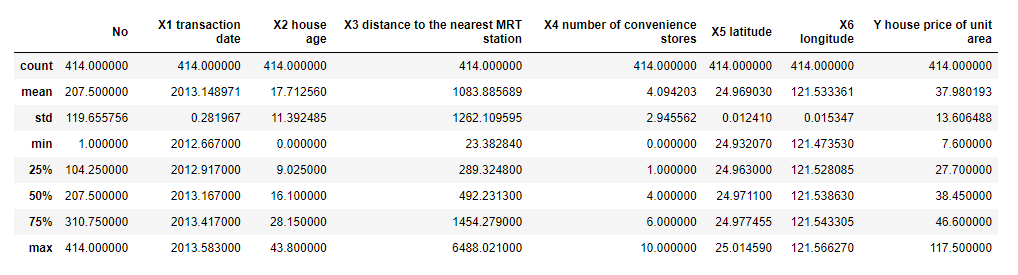
This study has been organized through theoretical research and practical implementation of regression algorithms. Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.

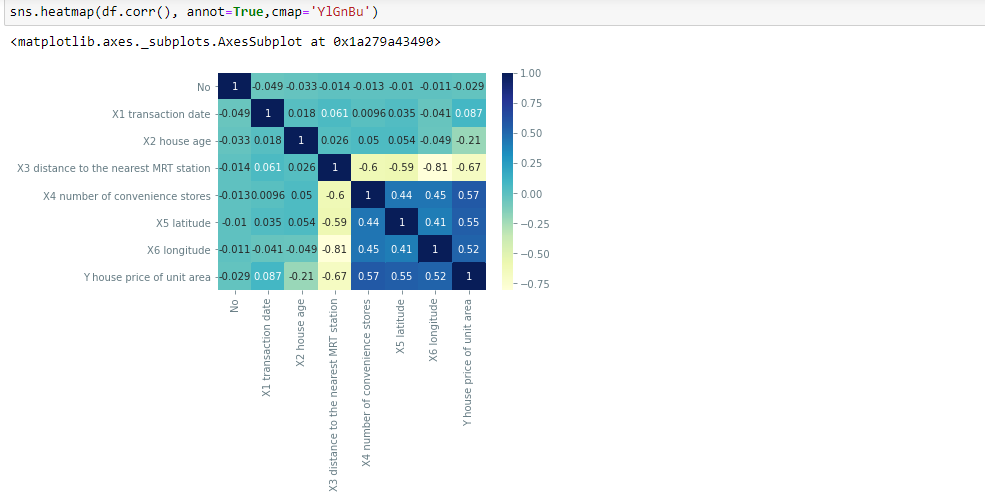


After computing the predicted values for our training data set by training our models in the form of supervised learning method with Python, we obtain **θ=(XTX)−1(XTy)θ=XTX−1XTy** that minimizes the cost value for the training set. We then plug in the coefficients (or weights) into our models, using the test data to see whether our estimations are still accurate for the test data, as evaluated by the mean square error (*MSE*), root-mean-square error (*RMSE*), mean absolute percentage error (*MAPE*) and the coefficient of determination R2R2. These three-performance metrics range between 0 and ∞, and each of which indicates that the fit is perfect when it is estimated to be 0.

**Dataset Review:**

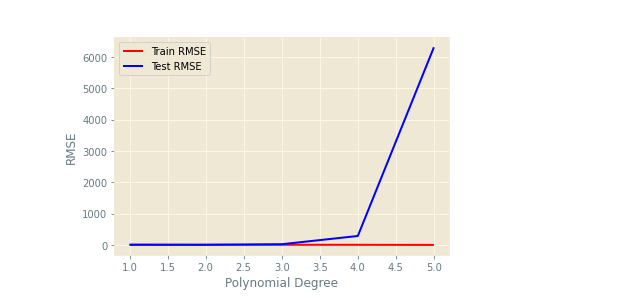
A dataset in machine learning is, quite simply, a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes. I have used one dataset in this paper where various existing machine learning algorithms are applied to the datasets for predicting prices. The first dataset is from the Kaggle which concerns price prediction. This dataset compromise of transition date, house age, longitude, latitude etc. As this paper uses machine learning for price prediction, attribute variables are used to predict the label/price. The following table shows the set of attribute variables to develop the prediction model. This study uses 7 attributes as independent variables for predicting house prices.





**Plot the Polynomial degree VS RMSE**

The graph is between the train and test of model with respect of root mean squared error



**The dataset consists of features in various formats. It has numerical data such as prices and numbers. In order to make this data with different format usable for our algorithms, categorical data was converted into separated indicator data, which expands the number of features in this dataset. We splitted our dataset into training and testing set with a roughly 70/30 split, with 1000 training examples and 460 testing examples. Besides, there were some features that had values of N/A; we replaced them with the mean of their columns so that they don’t influence the distribution.**

**we would need to perform multi-class classification to predict house prices into these seven buckets. The performance of each model can be characterized by accuracy rate, which is the number of test examples correctly classified over the number of total examples.**

**Algorithm:**

Regression analysis is widely used for forecasting. Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. If more independent variables are added, it is able to determine an estimating equation that describes the relationship with greater accuracy. Multiple regressions look at each independent variable and test whether it contributes significantly to the way the regression describes the data.

**Ridge Regression-**

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering.

Ridge regression is the method used for the analysis of multicollinearity in multiple regression data. It is most suitable when a data set contains a higher number of predictor variables than the number of observations. The second-best scenario is when multicollinearity is experienced in a set.

Regression Analysis > Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

**Advantages:** The estimates of the unknown parameters obtained from linear least squares regression are the optimal. Estimates from a broad class of possible parameter estimates under the usual assumptions are used for process modelling. It uses data very efficiently. Good results can be obtained with relatively small data sets.

**Disadvantages:** The outputs of regression can lie outside of the range [0,1]. It has limitations in the shapes that linear models can assume over long ranges. The extrapolation properties will be possibly poor. It is very sensitive to outliers. It often gives optimal estimates of the unknown parameters.

**Linear Regression-**

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation solves for) is called the dependent variable. Regression analysis allows you to understand the strength of relationships between variables. Using statistical measurements like R-squared / adjusted R-squared, regression analysis can tell you how much of the total variability in the data is explained by your model.

**Advantage**: A linear model can include more than one predictor as long as the predictors are additive. the best fit line is the line with minimum error from all the points, it has high efficiency but sometimes this high efficiency created.

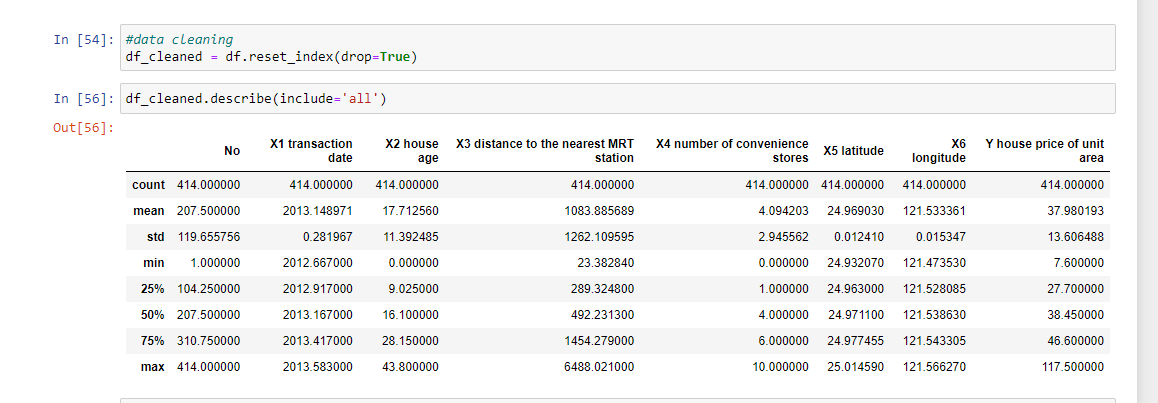
**Disadvantage**: Linear Regression Is Limited to Linear Relationships. Linear Regression Only Looks at the Mean of the Dependent Variable. Linear Regression Is Sensitive to Outliers. Data Must Be Independent

**Accuracy Calculation And visualization:**

The indicators I am used to evaluating the performance accuracy is the mean absolute error which is the difference between the predicted value and the actual value. After prediction the results data will be loaded into the tableau so it can be clearly visualized and it can also be used for the future works.

**Cleaning of dataset:**

Data cleaning is a process by which inaccurate, poorly formatted, or otherwise messy data is organized and corrected. For example, if you conduct a survey and ask people for their phone numbers, people may enter their numbers in different formats.



**Duplicates & NaNs**: I started by removing duplicates from the data, checked for missing or NaN (not a number) values. It’s important to check for NaNs (and not just because it’s socially moral) because these cause errors in the machine learning models.

**Categorical Features**: There are a lot of categorical variables that are marked as N/A when a feature of the house is nonexistent. For example, when no alley is present. I identified all the cases where this was happening across the training and test data and replaced the N/As with something more descriptive. N/As can cause errors with machine learning later down the line so get rid of them.

The problem here is that the machine learning algorithm could interpret the magnitude of the number to be important rather than just interpreting it as different categories of a feature. To solve the problem, I reverse engineered the categories and recoded them.

**Training and Testing of data:**

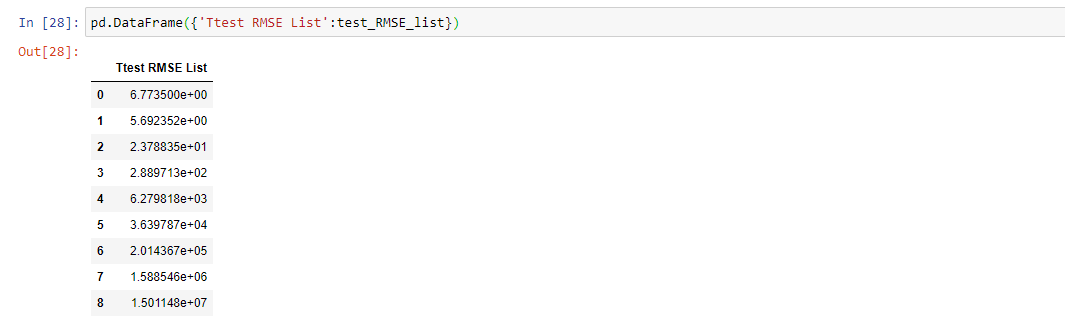
Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy.

**Correlation Matrix:**

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

**Root Mean Squared Error or RMSE:**

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

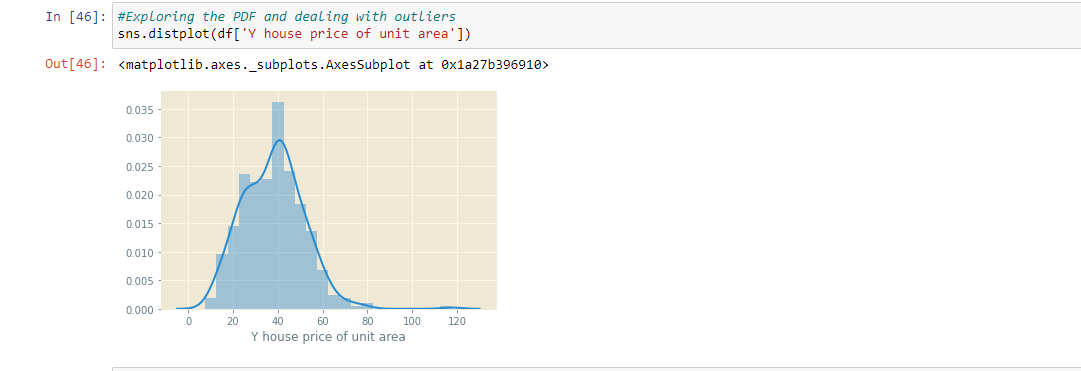


The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

**Exploratory Data Analysis (EDA)**

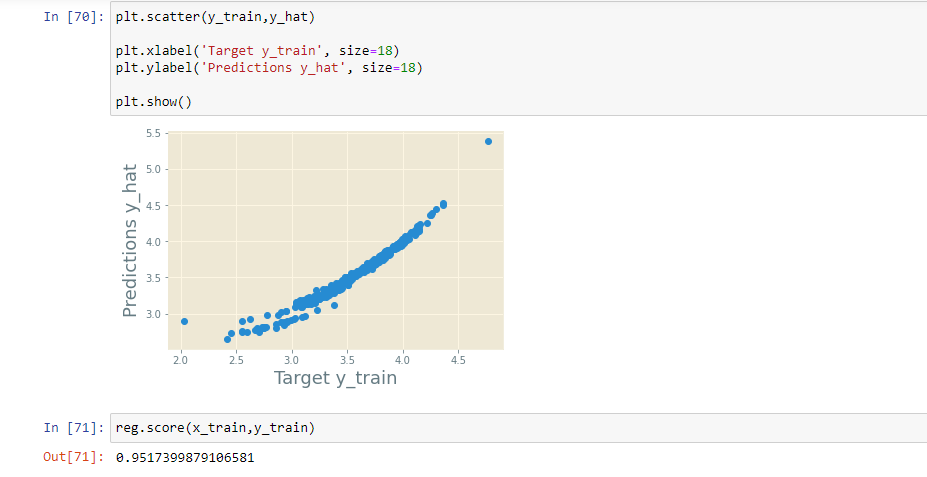
This is where our data visualization journey often begins. The purpose of EDA in machine learning I explore the quality of our data. A question to keep in mind is; are there any strange patterns that leave us scratching our heads?

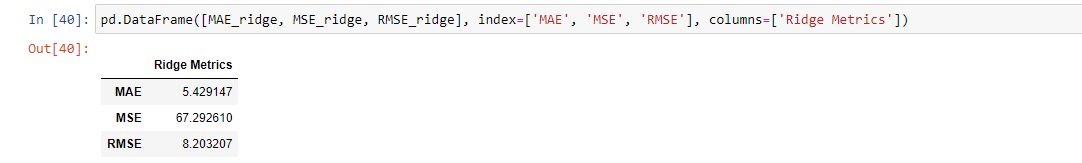
**Labels**: I plotted sales price on a histogram. The distribution of sale prices is right skewed, something that is expected. In your neighborhood it might not be unusual to see a few houses that are relatively expensive. Here I perform my first bit of feature engineering (told you the process was messy). I’ll apply a log transform to sales price to compress outliers making the distribution normal. Outliers can have devastating effects on models that use loss functions minimizing squared error. Instead of removing outliers try applying a transformation.



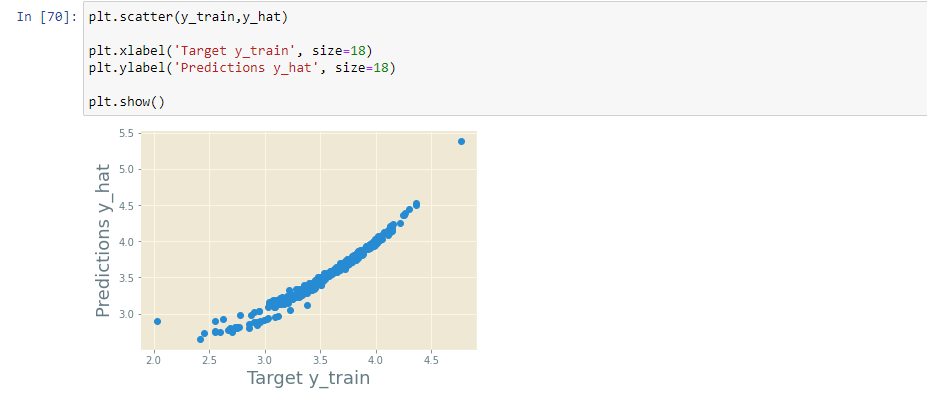
**Result:**

Overall, it shows that- There are different models used such as linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

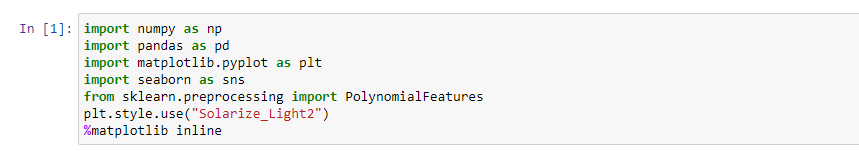




**Linear Regression** is the algorithm that is used for predicting House prices among various other algorithms. It is an algorithm of supervised machine learning in which the predicted output is continuous with having a constant slope. It is used to predict the values in a continuous range instead of classifying the values in the categories. Linear regression is used for performing different tasks like house price prediction. More expensive houses lose value when they are near smaller, less valuable homes. It shows the linear regression



By conducting this experiment with various machine learning algorithms, it’s been clear that ridge regression and linear regression are performing better with more accuracy percentage and with less error values. when this experiment is compared with the and to the result achieved these algorithms predicts well. This project has been done with the big data and its related technologies.



These are the libraries used in this project.

**The system is apt enough in training itself and in predicting the prices from the raw data provided to it. After going through several research papers and numerous blogs and articles, a set of algorithms were selected which were suitable in applying on both the datasets of the model. After multiple testing and training sessions. The system was potent enough for Predicting the prices of different houses with various features and was able to handle large sums of data. The system is quite user-friendly and time-saving.**We have defined several models with various features and various model complexities. There is a need to use a mix of these models a linear model gives a high bias (under fit) whereas a high model complexity-based model gives a high variance (overfit). Data Scientist tends to overfit their models which can be reduced by ridge regression and LASSO The study reveals that economic factors influence land price more than the social factors. The interaction of the selected factors (X) on land price (Y) is analyzed. It is found that four factors viz. GLV (84%), silver price per gram (92%), population (86%) and cost of crude oil (88%) have more positive effect on land price. The outcome of this study can be used in annual revision of guideline value of land which may add more revenue to the State Government while land transaction is made. This study will support the policy makers to relook the movement of the identified factors to have control on rise in the land price and stabilize it. Since there is a greater need for good long-term data analysis about land price, general land market behavior and spatial development, the results produced in this research may be of great use for Government and non-Government agencies which involve in land administration.

**CONCLUSION:**

We have defined several models with various features and various model complexities. There is a need to use a mix of these models a linear model gives a high bias (under fit) whereas a high model complexity-based model gives a high variance (overfit). Data Scientist tends to overfit their models which can be reduced by ridge regression and LASSO The study reveals that economic factors influence land price more than the social factors. The interaction of the selected factors (X) on land price (Y) is analyzed. It is found that four factors viz. GLV (84%), silver price per gram (92%), population (86%) and cost of crude oil (88%) have more positive effect on land price. The outcome of this study can be used in annual revision of guideline value of land which may add more revenue to the State Government while land transaction is made. This study will support the policy makers to relook the movement of the identified factors to have control on rise in the land price and stabilize it. Since there is a greater need for good long-term data analysis about land price, general land market behavior and spatial development, the results produced in this research may be of great use for Government and non-Government agencies which involve in land administration.

**GITHUB LINK: - https://github.com/VershaPrajapati/Real-Estate-Price-Prediction**