ADVANCED REGRESSION BASED HOUSING PRICE PREDICTION MODEL

AN INTERNSHIP REPORT

Submitted by

ABHIJEET SENAPATI [Reg No: RA1811030010064]

Under the Guidance of

DR. S. PRABAKERAN
(Assistant Professor, Department of Networking and Communications)

In partial fulfillment of the requirements for the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING
with specialization in Cyber Security

DEPARTMENT OF NETWORKING AND COMMUNICATIONS
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR - 603203

MAY 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR–603 203

BONAFIDE CERTIFICATE

Certified that this B.Tech project report titled "**ADVANCED REGRESSION BASED HOUSING PRICE PREDICTION MODEL**" is the bonafide work of **ABHIJEET SENAPATI [Reg No: RA1811030010064]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

SIGNATURE	**SIGNATURE**
DR. S. PRABAKERAN **GUIDE** Assistant Professor Department of Networking and Communications SRM Institute of Science & Technology, KTR	DR. ANNAPURANI PANAIYAPPAN .K **HEAD OF THE DEPARTMENT** Professor Department of Networking and Communications SRM Institute of Science & Technology, KTR
Signature of the Internal Examiner	Signature of the External Examiner

Department of Network and Communications

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/Course **: B.Tech in Computer Science Engineering with spcl. in Cyber Security**

Student Name **: Abhijeet Senapati**

Registration Number **: RA1811030010064**

Title of Work **: Advanced regression based housing price prediction model**

I/ We here by certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references/ listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:
I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above. RA1811030010064
If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr T.V.Gopal**, for his invaluable support.

We wish to thank **Dr Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department**, Dr K. Annapurani Panaiyappan,** Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our program coordinators **Dr. M. B. Mukesh Krishnan, Program coordinator**, and Panel Head, **Dr. S. Prabakeran,** Assistant Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor**, Dr. M. Uma,** Associate Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Dr. S. Prabakeran,** Assistant Professor, Department of Networking and Communications, SRM Institute of Science and Technology, for providing me with an opportunity to pursue my project under his mentorship. He provided me with the freedom and support to explore the research topics of my interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Networking and Communications Department staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

ABHIJEET SENAPATI

ABSTRACT

With a growing civilization and ever-changing market demands, it is critical to be aware of industry trends. Today, the study's main focus is on predicting property values based on trends. It is critical for a person to comprehend the business trends in order for an individual to plan his or her budgetary demands. Real estate is an ever-expanding industry in an ever-expanding civilization. Understanding trends is critical for an investor to properly underwrite and expand his business throughput Sometimes clients are duped by the agent's set-up of a phoney market rate, and the real estate market suffers as a result. These days, the industry is less transparent. With availability of dataset, it is possible for a researcher to create a model that is highly accurate. The first goal is to create a comprehensive model that benefits everyone. This study employs a number of methodologies, including the variance influence factor, dimensionality reduction techniques, and data analysis. Outliers and missing data are examples of transformation strategies. There are several elements that influence the price of a home involves physical characteristics, location, and numerous more factors.At the moment, economic considerations were persuasive. All of these imply that house price forecasting is a new concept, regression research field that necessitates the understanding of artificial intelligence. This design is meant to help a client by reducing his fieldwork ,wrest control of his time and money.Machine Learning algorithms like Linear Regression,Lasso Regression, Random Forest, and Xgboost can be used to overcome this complex problem.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Project Overview:

The Project describes the process of creating a Housing Price Prediction Model. A model is built using Sklearn and Linear Regression techniques and the dataset to be used is bangalore housing price dataset. In the second component of the project, we will write a Python Flask Server program that uses saved model to serve http requests. In the third component, a website will be built using Html, CSS and Javascript that will allow the user to enter his/her details like square feet area, number of bedrooms, number of balconies, the locality etc and a call is made to the the python flask Server which will retrieve the predicted price. During the process of model creation, a wide variety of data science concepts have been used such as Data Loading, Data Cleaning, Outlier Detection and Removal, GridSearch cv for hyperparameter tuning, Feature Engineer- ing, Dimensionality Reduction, K Fold Cross Validation etc.

1.2 Real Estate Market:

Investment in property has become more popular in recent times. The Real Estate (Pu- tatunda (2019)) is one of the fastest growing industries but at the same time there are various factors which buyers are not aware of, making it less transparent. There are various parameters, on which the property price depends, like the area, locality, available amenities, number of bedrooms,

balcony etc. Other factors also include accessibility to public transport like metro, connectivity to national highways, schools, expressways, and health facilities around. Prediction of property price might become tricky when done manually. Also, the price of any property should not be considered based on national trends because the value tends to change from state to state and even in neigh- bouring cities of the same state. Sometimes, the Market values tend to rise or fall based on a particular situation, for example in case of natural disaster, the real estate prices may increase to 3 folds the original price and home buyers may find it tough to get a home of their choice. Hence many different types of prediction structures are developed for property prices. The aim of our system is to develop a website where the user can get the price of any property based on the inputs made as per their requirements. To build a model which predicts the price of any property based on several factors, various regres- sion techniques are used. Various regression models are taken in account like Linear Regression, Lasso and Ridge regression, Decision tree regression and Random Forest regression. Based on the accuracy of the models and the percentage error, a compara- tive study is done and the best performing model is taken for further evaluation. After getting the best performing model, we can use it for estimating the property prices. Our data set consists of various features like availability, area, number of bedrooms, balcony, society, area type and price.

1.3 Proposed work:

From the inquiring price and the general description, the broad and consistent real estate attributes are typically presented separately. As a result, these traits or features are separately described in a prepared, ordered manner, allowing them to be easily compared across a wide variety of potential homes. Despite

the fact that each property have their unique identifiers, like special view, balcony, parking area, garage type, the dealers can provide a compendious of features. House’s most essential attributes. Thus, the supplied real estate attributes may be assessed by potential purchasers, but owing to the vast diversity, it appears to be practically difficult to provide an automated review on all aspects or factors. This is also true in the opposite direction: home sellers must calculate the value of their property, qualities or facet in comparison with current market forfeit of comparable properties.

Because of the variety of qualities or the large number of features, estimating a reasonable market price is a difficult process. Aside from providing a description of the house's important qualities, the house portrayal also serves to pique the reader's attention, or to persuade them.

This problem statement from the user summons to forecast the final price of each house using seventy - nine explanatory factors characterizing (nearly) every characteristic of residential properties.

Tasks to be completed include:

Research and document the customer's requirements, Identify and describe the key metrics that must be tracked in order to solve the situation, Analysis of exploratory data, Innovative feature development, Selection of Features, Advanced regression techniques such as random forest and gradient boosting are available.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction:

Real Estate is one of the fastest growing industry and at the same time,it is the least transparent one.The market and demand for housing is growing every year due to the increase in population and wide migration of people from villages to cities or from one city to other city for various purpose.The growing unaffordability in real estate is one of the major problems in metropolitan cities around the globe.Property value tend to change over time but factors like recession and natural disasters can affect the price.Technology has become more dependent and we can get accurate predictions by using various techniques, for these applications the researchers have proposed various machine learning techniques.

2.2 Literature Review:

P. Durganjali et al. strategy work classification algorithms to anticipate property resale prices. In this research, Linear, Logistic, Random Forest, and XgBoost algorithms to forecast the resale price of a property utilising several classification techniques such as Logistic regression, Decision tree, Naive Bayes, and Random Forest. Physical attributes, location, and numerous economic variables impacting at the moment all influence the property sale

price. For various datasets, we utilise accuracy as a performance measure, and these algorithms are run and compared to discover the most acceptable approach for sellers to use as a reference when selecting resale price.

Ayush Varma et al. advocated utilising machine learning and neural networks to anticipate housing prices. Housing prices change on a daily basis which sometimes are overstated, when it should be based on evaluation. Main emphasis of the research is to estimate home prices using actual world data. The objective is to substructure the evaluations on every critical aspect which would be considered while estimating the price. The regression methods used in this procedure, and our decree will be based on a weighted mean of the many procedures provide most precise results.
Findings pointed out that this procedure produces the fewest errors and noteworthy amount of accuracy as compared against comma separate algorithms. We also propose using Google maps to acquire precise real-world values by employing real-time local facts.

Sifei Lu and colleagues suggested a hybrid regression approach for predicting property prices. With limited dataset and data characteristics, this study analyses innovative feature engineering, practical and ingenious data pre-processing technique. To forecast individual property prices, the research presents a combination of Lasso and Gradient boosting model.

2.3 Research Gap:

Though many regression techniques gave good prediction results but there were also drawbacks associated with these techniques.In one of the research works, the researchers used a very small dataset and this led to

poor accuracy, hence, the accuracy could have been increased if large datasets were implemented.There was also problem of overfit- ting, (Shinde and Gawande (2018)). The problem of overfitting could have been re- duced if trained with Support Vector Machine (SVM) with higher accuracy.In Some of the research works, the final predicted graph had presence of outliers due to noise in the dataset, The outliers could have been detected and removed using Outlier detection algorithms which were not implemented. Some researchers used one factor square feet area to estimate their prediction while the property price of any property depends on other factors like location, number of bedrooms, proximity to school, nearest metro etc.There were also researchers who used neural networks in their prediction but the neural network used in the implementation did not provide satisfactory results and this resulted in poor performance. In business context, the application can involve direct use of the appropriate algorithm for the prediction purpose by taking some inputs from the user and giving most justified value without taking price input from user and preventing any exceptions from occurring in the system.

To sum up some of the given gaps that can be used for further development of the RealEstate price prediction:

Real estate market depends on various features and we cannot get proper idea of the value of any property by considering one or two features. Other features also need to be taken care of. If the model is prone to over-fitting, then some techniques like outlier removal can be used to remove the unwanted noise from the data which could interfere in the predicted value. Systems should have an easy, smooth and intractable interface for the users to ease the task of understanding the predictions.

Table 2.1: Overview of Survey

PAPER	AUTHOR	DESCRIPTION	FINDINGS
ELECTRONIC GOVERNANCE OF HOUSING (2019) Lydia et al. (2019)	E. Laxmi Lydia, Gogineni Hima Bindu.	Only Linear Regression for Boston Dataset.Mean Square Error calculated to 30.4187%.	Presence of Outliers, due to presence of noise in dataset. Functions in few situationsto estimate Boston City information.
HOUSING PRICE PREDICTION (2019) Madhuri et al. (2019)	CH. Raga Madhuri, Anuradha G, M. Vani Pujitha	Multiple-Linear model, Lasso, Ridge, Gradient Boostingmodel. Gradient Boosting [Highest score 0.9177], Least of Elastic Net.	Mainly comparison of various algorithms, price calculated using one factor only-square feet area.
HOUSE PRICE PREDICTION, MELBOURNE CITY (2018) Phan (2018)	Danh Phan	Polynomial Regression, Support Vector Machine, Regression Tree and Neural Networks used. Regression Tree delivers as good result as Linear Regression.	Neural Networks seems not to work effectively with this dataset and there is overfitting issue.

CHAPTER 3

SYSTEM DESIGN

The system consists of mainly two sections. First section is of the machine learning model and the other section is of the user interface part which would host the website and the result will be displayed on the system. These sections combined together will form the resultant system as a whole.

3.1 Tools and Platform:

Anaconda, Jupyter Notebook, Pyspark, Microsoft Excel

3.2 Machine Learning Model:

In this section we have developed a predicting model which would estimate the price of the property based on various factors. Different kinds of algorithms have been used and the one with good score will be taken into evaluating the further system.

First step is to collect the data which we have collected from the Kaggle repository.

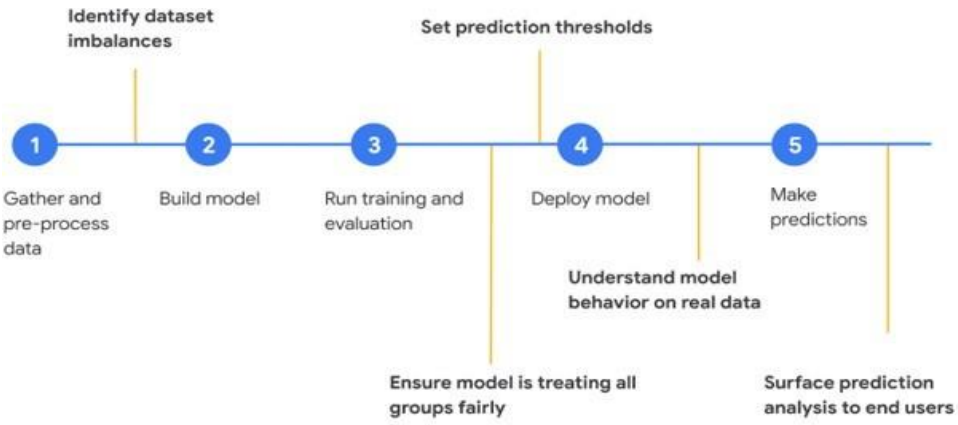

Fig 3.1 Machine Learning Systems

Then we have eliminated the imbalances like missing data or invalid form of data that was present in the data set. Outliers present in the data can also be eliminated at this stage. The mode and mean values were taken into account while dealing with the imbal- ances. For outliers, certain statistical methods were taken into account. Randomization of the data is also taken into account as it eliminates the consequences of the order in which data is collected. After removing these problems we have to build the predicting model with the pure data obtained. We check the score of the accuracy and then select the appropriate model into account. Training and testing of the data is done and the model is evaluated on the given parameters. After the training part the model is tested on the unseen or the test data. When this stage is completed then we deploy the model and start making the predictions for the price. In this first component we have made use of various libraries like NumPy and Pandas for the data preparation and cleaning. Mtplotlib was also taken into account which was used for the visualization part. Sklearn was used for the building of the predictive model.

3.3 User Interface Section:

After obtaining the predicted price from the machine learning predictor model, we need to surface that value to the user in the interface section. The concept of python flask server has been used to implement the interface part. It acts as an interface between the predicting model and the frontend.

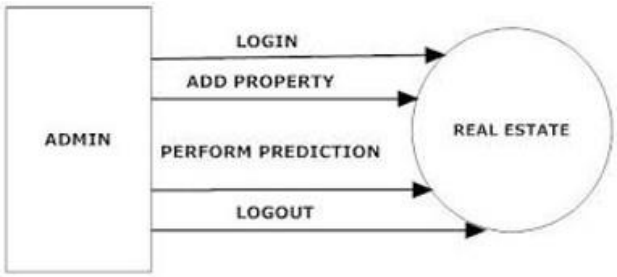

Fig 3.2 User Interface Section

The frontend is made of HTML, CSS and basic Javascript. In this part the user needs to input the values as per the requirement like number of rooms, area, location, etc. After this, the inputs are sent to the predicting model which gives the price based on these inputs and this value is transmitted via flask server.

3.4 ARCHITECTURE DIAGRAM:

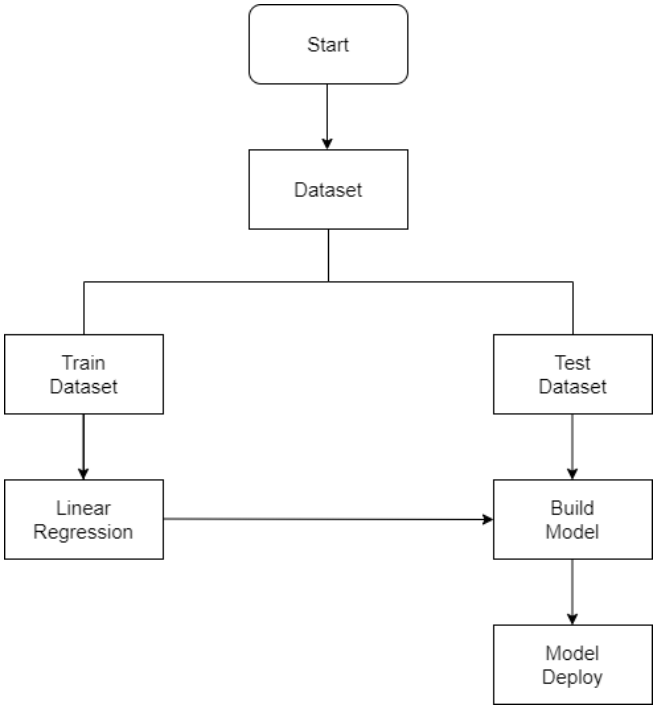

Fig 3.3 Architectural Diagram

Training and Testing Dataset:

We must keep the data when it has been extracted, according to our needs. The data will be output in a standard format, such as JSON or CSV, which may be saved in a database.

Linear Regression:

In Simple Linear Regression we have two continuous variables. One is predictor or independent variable and other is dependent variable. It has an equation of the from: y=v+rx, where y is the dependent variable, x is explanatory variable, r is the slope of the line and v is the intercept.

Build Model:

The algorithms are Linear Regression, Random Forest Regression, and

XGBoost Regression. One we get the insights and understanding of data, which has been cleaned and anal- ysed, we proceed with applying various algorithms which fits our data and the best performing model is taken for further consideration. These algorithms are implemented with the help of the Sklearn or the SciKit-Learn library in python.

```
train_data = pd.read_csv('train.csv')
train_data.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoS
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	

5 rows × 81 columns

```
test_data = pd.read_csv('test.csv')
test_data.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeatu
0	1461	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	AllPub	...	120	0	NaN	MnPrv	Na
1	1462	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN	Gar
2	1463	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	MnPrv	Na
3	1464	60	RL	78.0	9978	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN	Na
4	1465	120	RL	43.0	5005	Pave	NaN	IR1	HLS	AllPub	...	144	0	NaN	NaN	Na

5 rows × 80 columns

```
train_data.shape
```

```
(1460, 81)
```

```
test_data.shape
```

```
(1459, 80)
```

Fig 3.4 Train Test Details

CHAPTER 4

IMPLEMENTATION

4.1 Conduct a data analysis that is exploratory in nature

Make a feature engineering

Work on feature development/selection

4.4) Linear regression, Lasso regression, Random Forest, XgBoost algorithm are used, and the results are compared.

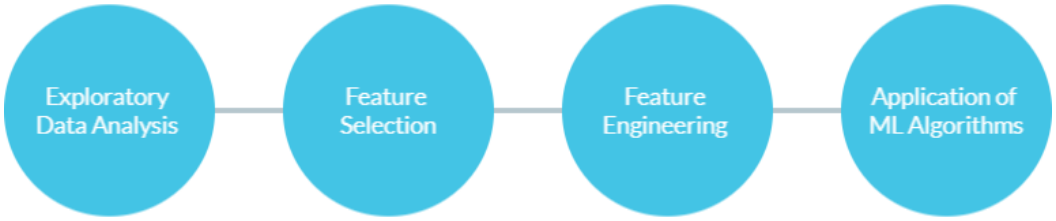

Fig 4.1 : Flowchart Representation

Data Analysis: Exploratory In Nature

Method for evaluating data sets for the purpose of of summarizing the main attributes, which often includes the usage of qualitative graphics with other data visualisation methods. Used for examination of what data can tell us without formal modelling or hypothesis testing. Key objectives are:

i)Make assumptions about what's causing the observed occurrences.

ii)Assess the assumptions that will be used to make statistical inferences.

iii)Assist in the selection of appropriate statistical tools and approaches.

iv)Provide a foundation for additional data collecting via surveys or experiments.

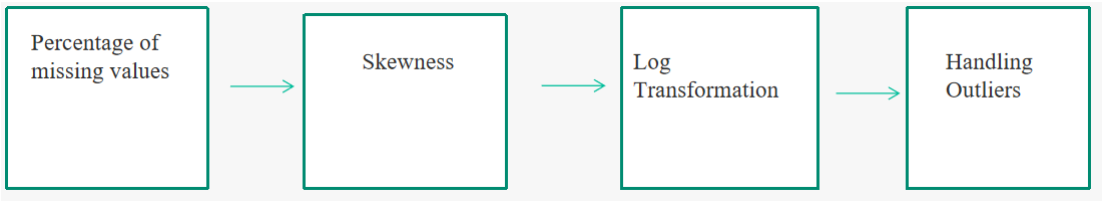

Fig 4.2: Tasks to be performed

Percentage of missing values:

Since there are many missing values a relationship between missing values and sales prices has been plotted.

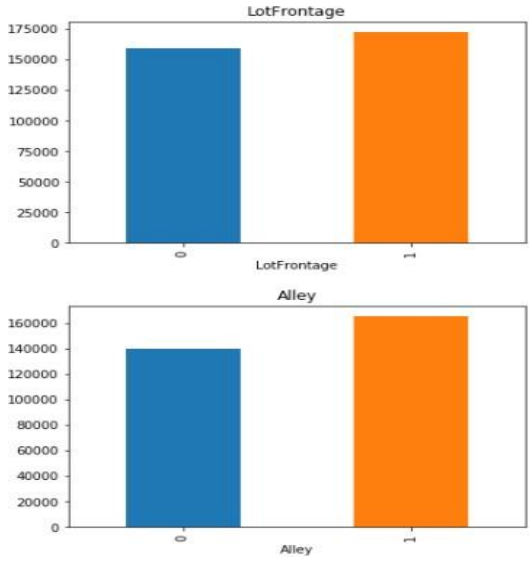

Fig 4.3 : Actual Vs Null Features

The relationship between the missing numbers and the dependent variable may be seen clearly here. The prices of the house have risen significantly as a result

of the missing information, and these inflated prices will impact the accuracy of the ML model. As a result, these nan values must be changed with something useful as part of the feature engineering process.

Skewness:

is a measure of the asymmetry of a real-valued random feature's probability distribution around its mean. To non continuous degrees, distributions will have positively or negatively skewed. The skewness of a normal distribution (bell curve) is equal to zero.

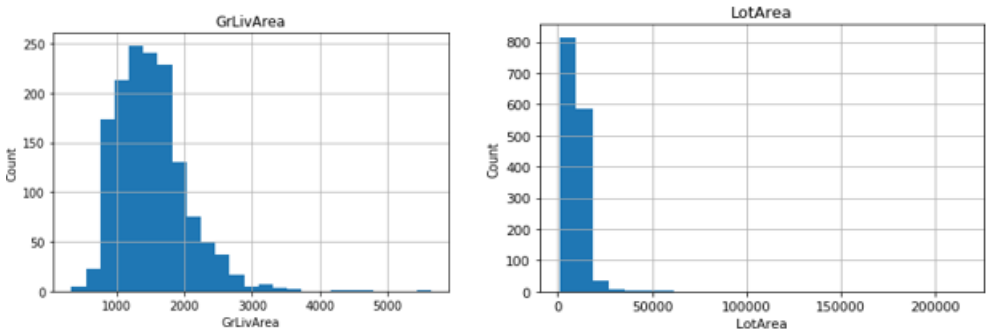

Fig4.4: Counts Vs Sales Price

Note: Because the data is skewed, a log transformation is required to normalise the data.

Log Transformation:

If continuous data does not follow the bell curve, we may log convert it to make it as "normal" as possible, improving the statistical analyses validity results. The log transformation, in other words, decreases or eliminates the skewness of our original data.

The data is typically log distributed, which implies the mean is zero and the

standard deviation is one.

```
## We will be using logarithmic transformation


for feature in continuous_feature:
    data=dataset.copy()
    if 0 in data[feature].unique():
        pass
    else:
        data[feature]=np.log(data[feature])
        data['SalePrice']=np.log(data['SalePrice'])
        plt.scatter(data[feature],data['SalePrice'])
        plt.xlabel(feature)
        plt.ylabel('SalesPrice')
        plt.title(feature)
        plt.show()
```

Fig 4.5: SalesPrice Vs Count

We get a positive correlation value after using the Log-Normal Distribution.

4.2 Handling Outliers:

They are values that deviate significantly from those of other data points, and they can pose complications in statistical techniques. Outliers might occur as a result of a data collecting error or simply as a sign of data volatility.

Outlier detection and management techniques include the following: The box plot, scatter plot, Zscore, and IQR (Inter-Quartile Range) were utilised to discover abnormalities in our dataset.

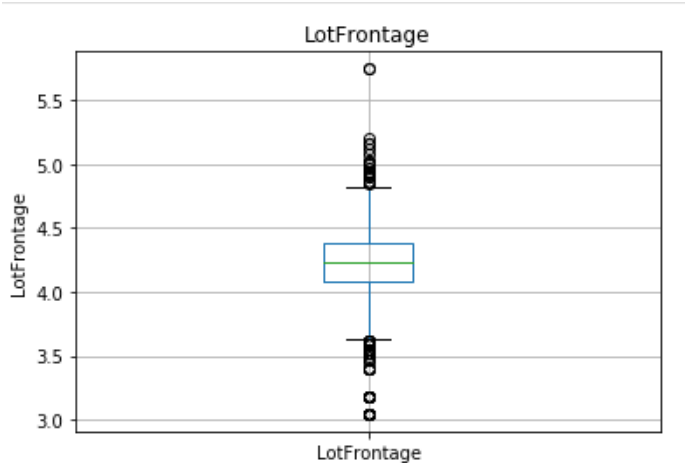

Fig :4.6 :Represents Zeroth Percentile,Fiftyth Percentile ,Hundreth Percentile Values

4.3 Feature Selection

Engineering of Features :

The act of selecting, modifying, and converting raw data features which can be used in the supervised learning which is called as feature engineering. It may be essential to create and train features so that machine learning model will perform effectively on new jobs.

Here the following process will be followed :

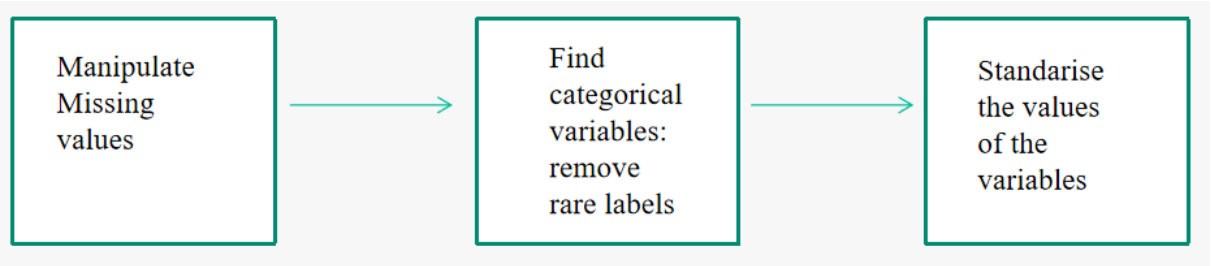

Fig 4.7: Feature Engineering flowchart

Manipulating missing NaN values:

Replacing the NaN values with mean, mode for numerical features and categorical features.

```
test_df['Utilities']=test_df['Utilities'].fillna(test_df['Utilities'].mode()[0])
test_df['Exterior1st']=test_df['Exterior1st'].fillna(test_df['Exterior1st'].mode()[0])
test_df['Exterior2nd']=test_df['Exterior2nd'].fillna(test_df['Exterior2nd'].mode()[0])
test_df['BsmtFinType1']=test_df['BsmtFinType1'].fillna(test_df['BsmtFinType1'].mode()[0])
test_df['BsmtFinSF1']=test_df['BsmtFinSF1'].fillna(test_df['BsmtFinSF1'].mean())
test_df['BsmtFinSF2']=test_df['BsmtFinSF2'].fillna(test_df['BsmtFinSF2'].mean())
test_df['BsmtUnfSF']=test_df['BsmtUnfSF'].fillna(test_df['BsmtUnfSF'].mean())
test_df['TotalBsmtSF']=test_df['TotalBsmtSF'].fillna(test_df['TotalBsmtSF'].mean())
test_df['BsmtFullBath']=test_df['BsmtFullBath'].fillna(test_df['BsmtFullBath'].mode()[0])
test_df['BsmtHalfBath']=test_df['BsmtHalfBath'].fillna(test_df['BsmtHalfBath'].mode()[0])
test_df['KitchenQual']=test_df['KitchenQual'].fillna(test_df['KitchenQual'].mode()[0])
test_df['Functional']=test_df['Functional'].fillna(test_df['Functional'].mode()[0])
test_df['GarageCars']=test_df['GarageCars'].fillna(test_df['GarageCars'].mean())
test_df['GarageArea']=test_df['GarageArea'].fillna(test_df['GarageArea'].mean())
test_df['SaleType']=test_df['SaleType'].fillna(test_df['SaleType'].mode()[0])
```

Fig 4.8 : Performing Mean/Mode for NaN values

Handling Rare Categorical Features:

Using one-hotencoding and creating a binary feature for each conceivable category is a popular strategy when working with categorical features.

The most common method is one hot encoding, which works well until your categorical variable has a huge number of values.

If any category in a particular feature, such as MsZoning, is present and has a weight of less than 1%, we may disregard it because it will have no substantial influence on the dataset.

88	89	50	Rare_var	4.663960	9.044286	Pave	Missing	IR1	Lvl	AllPub	Corner	Gtl	IDOTRR	Feedr	Rare_var	1Fam
89	90	20	RL	4.094345	8.995909	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam
90	91	20	RL	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam
91	92	20	RL	4.442651	9.047821	Pave	Missing	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam
92	93	30	RL	4.382027	9.500020	Pave	Grvl	IR1	HLS	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam
93	94	190	Rare_var	4.094345	8.881836	Pave	Missing	Reg	Lvl	AllPub	Corner	Gtl	OldTown	Norm	Norm	2fmCon

Fig 4.9 :The Rare_var can be dropped

Feature/Attribute Selection:

Is a method for standardizing the independent characteristics included in data over a defined range of values.

In pre-processing, data is utilised to manage drastically shifting magnitudes, values, or units. An algorithm will tend to weigh bigger values higher and smaller values lower if feature scaling is not done, nevertheless of the unit of the data.

The present dataset uses the MinMax scaler.

The values have been transformed to a range of ones and zeros.

```
feature_scale=[feature for feature in dataset.columns if feature not in ['Id','SalePrice']]

from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(dataset[feature_scale])
```

	Id	SalePrice	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	I
0	1	12.247694	0.235294	0.75	0.418208	0.366344	1.0	1.0	0.000000	0.333333	1.0	0.00	0.0	
1	2	12.109011	0.000000	0.75	0.495064	0.391317	1.0	1.0	0.000000	0.333333	1.0	0.50	0.0	
2	3	12.317167	0.235294	0.75	0.434909	0.422359	1.0	1.0	0.333333	0.333333	1.0	0.00	0.0	
3	4	11.849398	0.294118	0.75	0.388581	0.390295	1.0	1.0	0.333333	0.333333	1.0	0.25	0.0	
4	5	12.429216	0.235294	0.75	0.513123	0.468761	1.0	1.0	0.333333	0.333333	1.0	0.50	0.0	

Fig 4.10 : MinMax Scaler

CHAPTER 5

METHODOLOGY

The methodology consists of detailed description of the framework that is used in the project. It consists of a checklist that needs to be covered in order to achieve the ob- jectives. We have taken various data mining, machine learning and web-based concepts for achieving the goal. The best performing model would be used for further evaluation and it would be connected to the next component which would deal with the interface part.

5.1 Predicting Model:

We've experimented with several data mining and machine learning techniques. The steps below describe step-by-step tasks that must be completed:

5.1.1) Data collection

5.1.2) Data preprocessing

5.1.3) Missing value management

5.1.4) How to Deal with Outliers

5.1.5) Obtaining fictitious category traits

5.1.6) Algorithms in Practice

A brief explanation above the above-mentioned tasks are as follows:

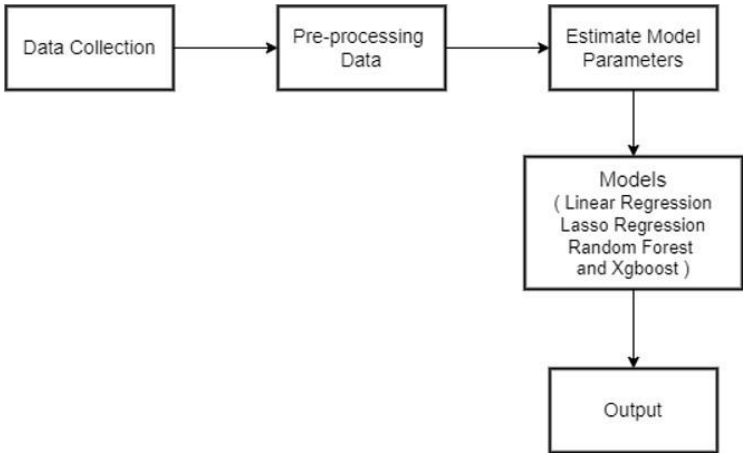

Fig 5.1 Data Flow Diagram

5.1.1 Data collection:

In our training data set, houses (observations) were accompanied with qualities (i.e. features, variables, or predictors) and the sales price for each property. The residences in our testing set all had the same characteristics, but the sales.Price was a target variable, it was not included. However, only those parameters were picked that are likely to have an impact on house prices.Area in square metres, for example. Overall quality, which considers the house's overall condition and finish, location, the year it was built, the swimming pool area, the house's selling year, and the price at which it was sold. A variety of different variables impact the selling price, which is a dependent variable.

Predictions are made using the test.csv data once we've trained a model using the train.csv data. We'll just go over our main findings, which will help us with feature engineering. We have categorical qualities that will need to be transformed into dummy variables.

5.1.2 Data Preprocessing:

Is a technique for turning unstructured, challenging data into systematic, understandable data. It is the process of searching a dataset for missing or redundant data. Any observations that contain NaN are eliminated, and the full data set is examined for NaN. The dataset gets more consistent as a result. Before we can apply any model to our dataset, we must first define its attributes. As a result, we must study our data set and look into the various factors, as well as their correlations. We can also see if there are any outliers in our data. Outliers are anomalies in a data collection that must be deleted due to some form of experimental error.

```
#missing data
total = train_data.isnull().sum().sort_values(ascending=False)
percent = (train_data.isnull().sum()/train_data.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
```

```
print(total)
```

```
PoolQC         1453
MiscFeature    1406
Alley          1369
Fence          1179
FireplaceQu     690
               ...
ExterQual         0
Exterior2nd       0
Exterior1st       0
RoofMatl          0
SalePrice         0
```

Fig 5.2: Performing Data Pre-Processing

5.1.3 Missing value management:

Many residences in the dataset have a value of zero.They don't have a garage, according to Garage Area. More features will be changed later to fit this idea.

There are a few exceptions as well. By isolating our estimated regression line from the real population regression line, outliers might distort a regression model.

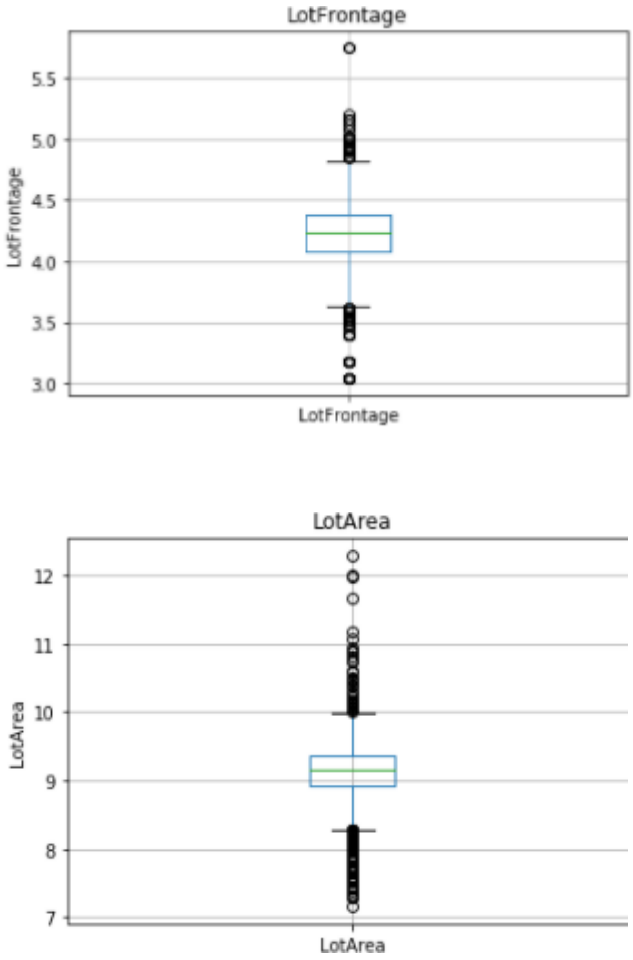

Fig 5.3 : Represents the examples outliers in the given Data Set

5.1.4 How to Deal with Outliers

We discovered that while most NA/NaN values for different variables related to an actual class, certain NA/NaN values genuinely reflected missing data. Three properties with NA/NaN values, for example the variable Pool Area should likewise have a non-zero value. Although these three homes are

expected to have pools, their quality was not examined or recorded in the data set. Our technique was to determine the median Pool Area for each class, then imputation of the absent classes depending on the mean of Pool Quality class.

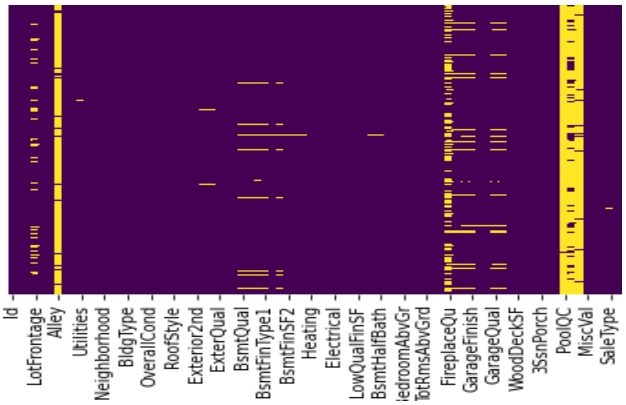

Fig 5.4: Represents null values (Yellow line indicates the null values)

5.1.5 Obtaining fictitious category traits:

Because the majority of existing machine learning algorithms can only take numbers (floats or integers) as inputs, we must encode these characteristics numerically if we are to use them in our models. The method we apply will differ based on the nature of the feature and the model we choose. Ordinal and categorical variables are two types of non-numerical variables. Dummy or one-hot encoding is the most common method for dealing with categorical variables. The following is how we apply dummy encoding to this:

5.1.6 Application of Algorithms:

Once the data is clean and we've gained insights into it, we'll need to find an acceptable machine learning model that suits our dataset. We used three strategies to predict the dependent variable in our dataset. We are training the

algorithms we picked to predict continuous values, despite the fact that they are generally used as classifiers. The algorithms are Linear Regression, Random Forest Regression, and XGBoost Regression. One we get the insights and understanding of data, which has been cleaned and anal- ysed, we proceed with applying various algorithms which fits our data and the best performing model is taken for further consideration. These algorithms are implemented with the help of the Sklearn or the SciKit-Learn library in python. The use of Pandas module is also done. It helps in reading and manipulating the data. Most of the data is in the Comma Seperated Values (CSV) form which is less understandable hence making it difficult to directly use that data in the model. It converts the data into human readable table format. This makes it easy to interpret the numerical values which can be used for the computational purpose. Numpy, which is a library in python, is also used for the manipulation of the tables which contains the numerical data. It has various functions which help in working with matrices and algebra.It includes multidimensional array and a group of routines with which these arrays are processed. Using all these tools we would be implementing the algorithms for prediction of the property prices. Following algorithms are implemented:

Linear Regression:

Linear Regression, (Ghosalkar and Dhage (2018)) is one of the conventional algorithms of machine learning. Here relationship is established between independent and dependent variable. The model is called Simple Linear Regression if there is one feature and Multiple Linear Regression when there are more predictor fea- tures. The main idea is to obtain the best fitting line or the line of best fit for our data in which the error margin is as low as possible.
In Simple Linear Regression we have two continuous variables. One is predictor or independent variable and other is dependent variable. It has an

equation of the from: y=v+rx, where y is the dependent variable, x is explanatory variable, r is the slope of the line and v is the intercept. Simple Linear Regression can be used to determine how strong the relationship exists between two features. It is also used to determine the value of any dependent feature based on any independent feature at a given point. This algorithm assumes certain factors for our dataset. The margin of error for the predicted outcome does not change to a large extent which respect to the independent features. This means that the variance is homogeneous in nature. The observations are collected with the help of some valid statistical methods by preventing any underlying relation- ship between features.

Normal distribution is usually followed by the data. As the name suggest, the relation between the dependent and independent features is linear which means the line of best fit is a straight line instead of a curve which passes through as may points of data follow- ing a given pattern. Linear regression makes use of the MSE, mean square error value to determine the error margin of the model. In this we measure the distance of observed value and predicted value from each of the independent feature. Then we square each of these distance values. After this step, mean of the values of distance is calculated.

In Multiple Linear Regression, (Manasa et al. (2020)) there are several explanatory features to depict the result of any outcome. This means that it is a type of linear regres- sion where two or more independent features establish a relationship with a dependent feature. The relationship can be linear or non-linear based on the relationship. Here the line of best fit passes through a multi-dimensional area of the given data points. This is widely used in financial and economical related features hence suitable for our property prediction. It can also be implemented to forecast or predict the impact of variations or changes. The equation is in the form of: yi=b0+b1.xi1+b2.xi2+. + bp where xi is predictor variable, yi is dependent variable b0=y-intercept, bp =slope of each predictor. This technique helps us to determine the variation in our model and

how each indepen- dent feature contributed in the evaluation of the variance. There should not be a high correlation between the independent features. High correlation means one feature can be used to determine another feature.

When multicollinearity is present in the independent features then some problems might arise to determine which variable contributed to the value of variance in dependent fea- ture. This model assumes that the margin of error is constant at each point of linear system. When the analysis is done then the residuals should be plotted against the pre- dicted result to make sure that the points are evenly distributed across all independent features. The model also demands that the observations should not depend on other observed values. When the residuals are distributed normally then a condition of multi- variate normality arises. The linear and non-linear models determine the outcome using two or more features. The non-linear one is complicated to execute as it is based on the assumptions derived from the errors obtained.

For Linear Regression Model, we obtained an accuracy of 85.5 percent in our model.

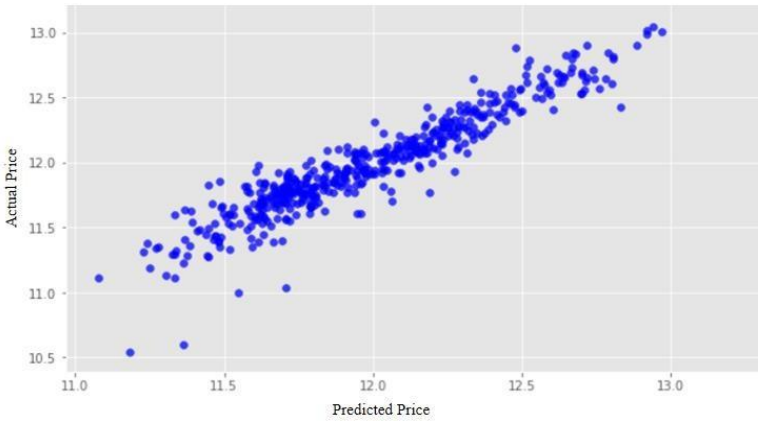

Fig 5.5 Feature Representation after linear regression

$$Cost\ Function\ For\ Linear\ Regression\ = 1/2m \sum_{i=1}^{n} (y - y\hat{})^2$$

Lasso Regression:

Lasso regression, (Lu et al. (2017)) also known as the Least Absolute Shrinkage and Selection OperatorLeast Absolute Shrinkage and Selection (LASSO) is a form of LR technique including the regularization function. In this model the absolute value summation of the magnitude of coefficients is taken into the consid- eration. It is basically a form of Linear Regression which makes implementation of the shrinkage technique. The concept of shrinkage implies that the value of any data point is shrunk to any central node which is generally the mean value.

This regression technique performs the L1 regularization method which involves a penalty term which is equivalent to the absolute value of coefficients' magnitude. When we implement the regularization then we can overcome the problem of over fitting. The accuracy of the model is compromised if we overlook the over fitting of the data in which the training data is more trained and this results in model giving poor results when served with data other than the training set.

A loss function is taken into account while fitting the model which is known as the sum of squares. The aim is to reduce the loss function to the maximum extent by choos- ing such coefficients in the equation. If the coefficients are not chosen properly then some unwanted data might also get involved in the training data set. When cases where there wrongly chosen coefficients arise then we try to shrink or regularize these values as close to zero. This type of algorithm is mainly used when we observe huge multi- collinearity in our features, which means there is large correlation between the predictor features where one factor can be used to estimate the value of other. This value of cor- relation is estimated with the help of the correlation coefficient.

When this coefficient has value equal to +1 then we say that the features have a strong and positive relation. When it has a value equal to -1 then we say that the predictors have a strong but negative relation. When this value is equal to zero then we can say that the predictors have no relation between them. Lasso Regression has the advantage that it has a very strong in-built capability of the selection of features which becomes very helpful in several situations. At the same time if the relationship between the pre- dictor and the target feature is not linear in nature then it might become complicated to implement this model in a non-linear relationship. Also this type of regularization can cause the data getting sparse or spread out which might also lead to loss of some coefficients. This technique works if the data has been scaled beforehand with the help of various scaling and standardizing techniques.

The equation for this model is given below:

Residual Sum Of Squares + λ * (Total Sum of the magnitude of coefficients absolute value). The λ is the value of shrinkage. When this value is equal to 0 then it indicates that it is almost equal to the linear regression in which all factors have been considered and residual sum of squares would be used to develop the predictive system. When this value comes out to be ∞ then it means that some no feature is left and all of them have been discarded. As this value reaches the proximity of ∞ then we can assume that some features are being discarded by the model. With the increase in the shrinkage value λ the bias also increases. The decrease in the value of λ leads to increase in the variance. Bias can be defined as the amount by which the prediction of our model differs from the target feature when we compare it to the training set. Model selection can be used to introduce the bias. In order to learn fast, linear model have high value of bias. On the other hand variance indicates how much the value of the target will deviate if the training set is altered. Variance can also lead to the over fitting of data where minute variations in the training data can get highlighted. Increasing the variance will decrease the bias and vice versa. The

correct balance has to be determined while evaluating both these values. In supervised algorithms of machine learning we aim to achieve low vari- ance and bias value in order to get good prediction results.

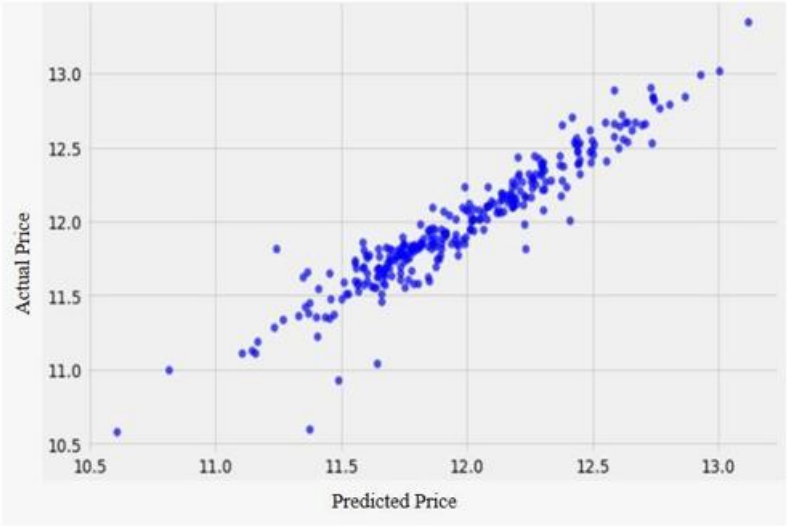

Fig 5.6 Feature Represenatation after lasso regression

$$Cost\ Function\ For\ Lasso\ Regression\ =\ 1/2m \sum_{i=1}^{n} (y - y\hat{})^2 +\ \alpha(slope)^2$$

Decision Tree Regression:

The Decision tree regression, (Navlani (2018)) comes under one of the supervised machine learning models. It is usually used for categori- cal values and continuous variables of output. As clear from the name of the model, it uses a tree form of structure for the development of classification and regression-based models. It divides the given dataset further into smaller subsets and hence simultane- ously through association method, decision tree is developed incrementally. It trains the model in the structure of a tree like model and yields a continuous meaningful value as an outcome.

Decision tree can be considered as the model based on predictions which depict the target feature with the help of some binary rules. It forms a simple model

which is comprised of leaves, nodes and branches. Root nodes represent the entire structure of the data which gets further split into smaller subsets which are homogenous in nature. Splitting is the process of breaking down the given node into two or more subsets.

Decision node is the one which can be defined as a sub-node which further gets divided into sub-nodes. Leaf nodes or terminal nodes are the one which cannot be divided fur- ther. Pruning is defined as reverse splitting process in which we remove the branches of any tree. It can be done using few ways like limiting the height of our tree, ignoring or removing the leaves which have few branches or by setting a limit on the leaf nodes number. The nodes which get divided into sub nodes are called the parent nodes and the ones which we get as the result of the splitting is known as the child node.

To determine the purity of the split we can make use of various features. Gini Impurity is defined as a measure to depict the purity of the split. It is a value which lies in the range of 0 to 0.5 where the value of 0 determines that the split is pure which means 100% lies in the same set or class. A value of 0.5 means impure split where 50% value lies in one class and 50% in other class. This makes the split difficult.

$$\text{Gini value} = 2 \times pr \times (1\text{-}pr).$$

Here “pr” is the probability or percentage of true events and (1-pr) is the probability of false events. Another measure of the purity of split is Entropy. The value of entropy lies in the range of 0 to 1. Here the value of 0 signifies that the split is pure and the data belongs to the same class whereas the value of 1 defines that the split is impure and 50% data is present in one class and remaining half is present in the other. Entropy value =

-[pr log2 pr + (1-pr) log2 (1-pr)], where Probability (pr) is the probability of positive event and (1-pr) is the probability of negative event. The parameter of

“criterion” is used to depict how the impurity of the splitting procedure. Generally, Ginni value is used for the criterion part but it works effectively in case of categorical data. Our data is continuous in nature hence we cannot use the ginni impurity value for this purpose. We have made use of the Mean Square Error (MSE) value for this implementation.

XG Boost:

It is a famous boosting method. In gradient boosting, each prediction can om it’s own correct thr prediction of the one before it. Rather of adjusting, weights of the trainings samples, each predictor will be trained using the left-out errors of the predecessors. The trees are built in a boosting order, with each succeeding tree striving to minimise the faults of the one before it.
Each tree builds on its forefathers' wisdom and corrects any flaws that remain. Thus the next set of tree in the progression will use a refurbished set of residuals to learn from. Basic learners in boosting are fragile learners with a bias and predictive capability that is just slightly better than random guess. Fragile learners provide important information for the prediction, enabling the boosting to constructively combine weak learners in order to create a powerful learner. Ultimately strong learner will reduce both the bias, variation.
This model received an accuracy of 90.42%

5.1.7 Evaluating The Model:

After selecting the model for the evaluation, we need to test that model against the val-ues of the unseen data or the test data. This unseen dataset is basically used to depict the performance of the model when it is put into application in the real world. The training data is the one on

which the model is first evaluated and this set is consid-ered as the seen data as the model has the experience of working with this data. In case of over-training, the model develops a great understanding of the train data but when some real-world data or the unseen data or the test data is given to the model then it fails to give good accuracy. This is because the model develops the extra capacity to predict only from the training data and anything beyond that set would not be able to get handled by the model.

Usually, the training set and the testing set is split in the ratio of 7:3 or 70% training dataset and remaining 30% for the testing data set. This ratio can be increased or decrease as per use but keeping a very small set of data for the testing part may leave the model as under-trained and when the test data would be exposed to the model then it might fail to give optimum results. The training set should also not be very large as the problem of over-fitting may arise. There should be sufficient data that can be used in order to test that how the model performs when it is exposed to the unseen data.

CHAPTER 6

INTERNSHIP PROJECT DELIVERABLE

Findings of the study:

The challenge is to create a hypothesis function that can predict the end value based on the data received.

Examine the prediction on the data's testing section. The information provided here pertains to the property pricing and the things that go along with it. As a result, developing a ML model that can learn data attributes and accurately anticipate pricing is a difficult undertaking.

The dataset's characteristics are as follows: it has 78 explanatory variables, commonly known as features or characteristics variables. The target value is the last variable; in this instance, it is termed Sale Price, (actual price of the home) When the ML model predicts the price, it will be compared to the real value, as well as the mean error is determined, giving the model's accuracy rate. The data set might include data about the residences' different details. Explanatory factors are used to describe (almost) every element of residential houses.

Examine all of the house's qualities that are relevant to the dependent aim. When analysing and visualising the data, look for any missing values and fill them in thorugh the median values for the specific attribute. Use a single hot encoder or label encoder to transform category data to numerical data.Using help of a heat map and a correlation matrix created with Python's Seaborn, find

the correct characteristics. Choose the qualities that are most similar to the genuine dependencies of the label target. Before using the regression function on the data, divide it through two portions. The training data comes first, followed by the testing data. Apply machine learning to the training part of the data using the sklearn package on the Python platform.

After examining the data, it was discovered many missing values are present in the data, This will make the data set as noisy data, and as a consequence.

CHAPTER 7

RESULTS & DISCUSSION

7.1 Results:

To forecast the response variable in our dataset, we used four techniques. Despite the fact that methods we chose are mostly employed as classifiers, here we have trained them to predict continuous valuess. Linear Regression, Lasso Regression, Random Forest, and Xgboost are the four algorithms used.

These algorithms were created using the Python SciKit-learn Library.The projected results of these algorithms were stored in a comma break up value file.Code produced this file during the run time. Outliers may be particularly sensitive to this model.

The following table represents various algorithms used with their accuracy.

ALGORITHM	TESTING ACCURACY SCORE
LINEAR REGRESSION	85.15 %
LASSO REGRESSION	92.27 %
RANDOM FOREST REGRESSION	97.69 %
XGBoost	90.42 %

Fig 7.1 : Accuracy of Xgboost

The accuracy of linear regression approaches is 85.15 percent.
On the identical training and testing data, the lasso regression approach produces a result of around
92.27 percent, which is somewhat superior to linear regression. Gradient boosting when applied to data, it produces a higher level of perfection of about 90.42 percent. The accuracy of random forest was 97.69 percent. As a consequence, when all four regression procedures are compared, the Random Forest regression methodology produces the best results. Random forest regression outperforms the pair and is hence we have used superior advanced regression approach for this data set.

7.2 Future Enhancement:

By training the model with new datasets, it may be further adjusted. This will increase the accuracy of the client's results while minimising or eliminating variations in the predicted housing values.
Performing in-depth exploratory data analysis to identify and minimise data abnormalities if they are there. Treatment of outliers as much as possible instead of dropping them or removing from the dataset.

7.3 Conclusion

It was feasible to judge the predicted efficiency of a number of house sales using a variety of analytical and graphical techniques. The models also helped define which dwelling attributes were most strongly linked to price and may explain the majority of the price difference. Furthermore, by accounting for the

influence of various regression techniques, which were capable to improve the precision of forecast of the models. Linear regression, Lasso regression, random forests, and gradient boosting for all attributes were all employed in this investigation. As a performance metric criterion, regression models tested and quantified. The significance of each selector in illuminating cost variance in a certain sample of home attributes was another major purpose of this research. Overall, the findings provide useful information on the causes of numerous aspects on property values, as well as their analyses.

REFERENCES

[1] P. Durganjali and M. Vani Pujtha, "House Sales Price Forecasting Using Classification Algorithms," ICSSS 2019, Chennai, India, pp.4-9,

[2] Ayush Sharma, Rohit Nagpal, "Prediction of house price using Neura1 network," 2017 IEEE, 2018 Second International Conference on Inventive Science & technology Coimbato

[3] SifeiLu, Zengiang Liu, Zhenog Qinn, "price prediction of house using linear regression," 2018 IEEE, 2018 IEEE International Conference on Engineering & communication, Singapore.

[4] Forrest E. Huffman and Paul K. Asabere In: Journal of Realtor Business and Finance 6 (1993), pp. 167–174, "Price Exemptions, Time of something like the Marketplace, and the Absolute Sale Price of Homes."

[5] Spatiol Depindence, "Houseing Unserved and House cost estimation," The Institute of RealEstate Macroeconomics, 148-1771, 2008. Robert C. Bourassa, Erika Cantoni, Martin Edward Ferdinand Hoesli.

[6] "House Price Prediction Using Machine Learning," VolYajurveda Chouthai, Mohamed Atuar Ranegila, Sanved Ammate, Pryag Ahikaari, and Vijay Kukre. Vol 28, 67-76, 2016

[7] Ghosalkar, N. N. and Dhage, S. N. (2018). “Real estate value prediction using lin- ear regression.” 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). 1–5.

[8] Hong, J., Choi, H., and Kim, W.-s. (2020). “A house price valuation based on the random forest approach: the mass appraisal of residential property in south korea.” International Journal of Strategic Property Management, 24(3), 140–152.

[9] Limsombunchai, V. (2004). “House price prediction: hedonic price model

vs. artificial neural network." New Zealand agricultural and resource economics society conference. 25–26.

[10] Lu, S., Li, Z., Qin, Z., Yang, X., and Goh, R. S. M. (2017). "A hybrid regression tech- nique for house prices prediction." 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE. 319–323.

[11] Lydia, E. L., Bindu, G. H., Sirisham, A., and Kiran, P. P. (2019). "Electronic gover- nance of housing price using boston dataset implementing through deep learning mech- anism." International Journal of Recent Technology and Engineering (IJRTE) ISSN, 2277–3878.

[12] Madhuri, C. R., Anuradha, G., and Pujitha, M. V. (2019). "House price prediction using regression techniques: A comparative study." 2019 International Conference on Smart Structures and Systems (ICSSS), IEEE. 1–5.

[13] Manasa, J., Gupta, R., and Narahari, N. S. (2020). "Machine learning based predict- ing house prices using regression techniques." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). 624–630.

[14] Navlani, A. (2018). "Decision tree classification in python." Data Camp.

[15] Phan, T. D. (2018). "Housing price prediction using machine learning algorithms: The case of melbourne city, australia." 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), IEEE. 35–42.

[16] Putatunda, S. (2019). "Proptech for proactive pricing of houses in classified advertise- ments in the indian real estate market." arXiv preprint arXiv:1904.05328.

[17] Shinde, N. and Gawande, K. (2018). "Valuation of house prices using predictive tech- niques." International Journal of Advances in Electronics and Computer Science, ISSN, 2393–2835.

APPENDIX 1

PYTHON CODE SNIPPETS

```
feature_sel_model = SelectFromModel(Lasso(alpha=0.005, random_state=0)) # remember to set the seed, the random state in this function
feature_sel_model.fit(X_train, y_train)
```

```
SelectFromModel(estimator=Lasso(alpha=0.005, copy_X=True, fit_intercept=True, max_iter=1000,
   normalize=False, positive=False, precompute=False, random_state=0,
   selection='cyclic', tol=0.0001, warm_start=False),
        max_features=None, norm_order=1, prefit=False, threshold=None)
```

```
y_train_pred = lasso.predict(X_train)
print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))
```

```
0.9227862283066279
```

Accuracy of lasso regression

```
from sklearn.model_selection import train_test_split
X = train_data.drop(['Id','SalePrice'], axis=1)
y = train_data['SalePrice']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=65)
```

```
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)
```

```
(1095, 34) (1095,)
(365, 34) (365,)
```

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(X_train, y_train)
print(lr.score(X_test, y_test))
```

```
0.8515537848984642
```

Accuracy of linear regression

```
import xgboost as xgb

xgbr = xgb.XGBRegressor()
params = {'learning_rate': [0.10,0.14,0.15,0.16, 0.2], 'max_depth': [1,2,3,5] }

xgbr_reg = GridSearchCV(xgbr, params, cv = 10, n_jobs =1)
xgbr_reg.fit(X_train_xgb,Y_train_xgb)

print("Best params:{}".format(xgbr_reg.best_params_))

best_x = xgbr_reg.best_estimator_
y_train_pred_x = best_x.predict(X_train_xgb)
y_val_pred_x = best_x.predict(X_test_xgb)
```

```
Best params:{'learning_rate': 0.15, 'max_depth': 3}
```

```
model=XGBRegressor(learning_rate=0.1)
model.fit(X_train,y_train)
print(f"Train score : {model.score(X_train,y_train)}")
print(f"Validation score : {model.score(X_val,y_val)}")
```

```
Train score : 0.9042146806102414
```

Accuracy of xgb regression

APPENDIX 2

PLAGIARISM REPORT

abiv3

ORIGINALITY REPORT

6%	2%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Central Florida Student Paper	2%
2	Www.Ijeast.Com Internet Source	1%
3	github.com Internet Source	<1%
4	Shiv Shankar Prasad Shukla, Samir Kumar Pandey, Ujjwal Bharadwaj, Anil Kumar Yadav. "Chapter 60 Assessment of Real House Price Using Machine Learning", Springer Science and Business Media LLC, 2021 Publication	<1%
5	Suthagar S, Snegha C, Sureka M, Velmurugan S. "Analysis of Breast Cancer Classification using Various Algorithms", 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022 Publication	<1%
6	Submitted to RMIT University Student Paper	<1%

7	norma.ncirl.ie Internet Source	<1 %
8	Submitted to Kingston University Student Paper	<1 %
9	Giulianella Coletti, Romano Scozzafava, Barbara Vantaggi. "Chapter 15 Weak Implication in Terms of Conditional Uncertainty Measures", Springer Nature, 2007 Publication	<1 %
10	Submitted to Queen's University of Belfast Student Paper	<1 %
11	Submitted to South Bank University Student Paper	<1 %
12	"Soft Computing and Signal Processing", Springer Science and Business Media LLC, 2022 Publication	<1 %
13	Submitted to Delhi Metropolitan Education Student Paper	<1 %
14	Submitted to University College London Student Paper	<1 %
15	Submitted to University of Teesside Student Paper	<1 %
16	I-ling Yen. "A Unified Framework for Defect Data Analysis Using the MBR Technique",	<1 %

2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 06), 11/2006
Publication

17	ebin.pub Internet Source	<1 %
18	link.springer.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

INTERNSHIP LETTER

Mu Sigma

25 October 2021

Dear Abhijeet,

Following the offer, dated 8th September 2021, the Mu Sigma Leadership Team recently had their annual planning exercise. During this exercise, we have made significant revisions in our offer to you. We are certain that the revised terms are more exciting and inspires you to embark on a journey with Mu Sigma.

Summary of the new structure

1. We have reduced the penalty to INR 10 lakhs from 15 lakhs should you decide to terminate your employment services with Mu Sigma prior to the 42nd month of joining. The penalty for terminating the employment services post 42nd month to 54th month shall remain the same at INR 7.5 Lakhs. Kindly note that opportunity ecosystem that this program instantiates is valued at **INR 10 Lakhs**.

2. In addition, we are providing **guaranteed year end bonus** rather the earlier stated variable bonus, increasing certainty in the overall compensation

While you join Mu Sigma as a Trainee Decision Scientist, the journey of becoming a professional Decision Sciences Leader extends beyond 4 years. Mu Sigma will provide you the canvas to learn, experiment and create meaningful impact by solving real world problems. You will become an Apprentice Leader after 3 years of training as a Decision Scientist, after which you will work directly with leaders at Fortune 500 companies, helping them solve problems and make impactful decisions.

We would appreciate if you could review the offer carefully and, submit your acceptance no later than **October 30th 2021**. If you have any queries related to the offer prior to the deadline, you can write to us with your queries at campus_india@mu-sigma.com

With best regards,

Deepa S. Mahesh

Deepa S Mahesh
Director

Mu Sigma Business Solutions Private Limited
Registered Office: Aviator Building, Level 14, Ascendas – ITPL SEZ Zone Whitefield Road, Bangalore, Karnataka - KA – INDIA – 560066
Tel: +91 80 7154 8000 | Fax: +91 80 7154 8060 | Email: info@mu-sigma.com | Website: www.mu-sigma.com
CIN: U74140KA2005PTC036309

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY (Deemed to be University u/s 3 of UGC Act, 1956)		
Office of Controller of Examinations		
REPORT FOR PLAGIARISM CHECK ON THE SYNOPSIS/THESIS/DISSERTATION/PROJECT REPORTS		
1	Name of the Candidate **(IN BLOCK LETTERS)**	ABHIJEET SENAPATI
2	Address of the Candidate	SRMIST, KATTANKULATHUR **Mobile Number : 9896297313, 9205144564**
3	Registration Number	RA1811030010064
4	Date of Birth	29/12/2000
5	Department	DEPARTMENT OF NETWORK AND COMMUNICATION
6	Faculty	DR. S. PRABAKERAN
7	Title of the Synopsis/ Thesis/ Dissertation/Project	ADVANCED REGRESSION BASED HOUSING PRICE PREDICTION MODEL
8	Name and address of the Supervisor / Guide	Assistant Professor School of Computing - Department of Networking and Communications SRM Institute of Science & Technology (SRMIST) Kattankulathur, Chengalpattu District - 603 203 Tamil Nadu, India **Mail ID : prabakes@srmist.edu.in** **Mobile Number : 9042394880**
9	Name and address of the Co-Supervisor / Co- Guide (if any)	**Mail ID :** **Mobile Number :**
10	Software Used	TURNITIN
11	Date of Verfication	12/05/2022

12	Plagiarism Details: (to attach the final report)			
Chapter	Title of the Chapter	Percentage of similarity index (including self citation)	Percentage of similarity index (Excluding self citation)	% of plagarism after excluding Quotes, Bibliography, etc.,
1	Introduction	4%	<1%	<1%
2	Literature Survey	1.1%	<1%	<1%
3	System Analysis	2%	<1%	<1%
4	System Design	<1%	<1%	<1%
5	Results and discussion	<1%	<1%	<1%
6				
7				
8				
9				
10				
Appendices		6%	3%	1%

I / We declare that the above information have been verified and found true to the best of my / our knowledge.

Signature of the Candidate	Signature of the Supervisor / Guide
Signature of the Co-Supervisor/Co-Guide	Signature of the HOD

Date : 12/05/2022