

House Capital Prediction Using Machine Learning in Python

BY

Golam Ahmed Bone

ID: 203-15-14556

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Dewan Mamun Raza

Senior Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

27th JANUARY 2024

Approval

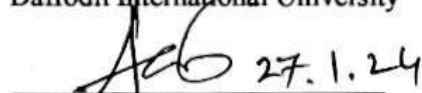
The Thesis "**House Capital Prediction Using Machine Learning in Python**" Submitted by Golam Ahmed Bone, ID:203-15-14556, To the Daffodil International University Department of Computer Science Engineering, has been accepted as sufficient for partially fulfilling the criteria for the degree of B.Sc. in Computer Science and Engineering (BSc) and has been accepted in terms of Style and Content. On January 27 2024 the Presentation Ended

BOARD OF EXAMINERS



Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



27.1.24

Dr. Arif Mahmud
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

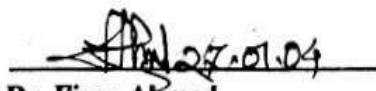
Internal Examiner



27/01/24

Mr. Shahadat Hossain
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



27.01.04

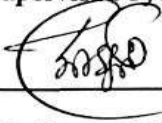
Dr. Firoz Ahmed
Professor
Information & Communication Engineering
Rajshahi University

External Examiner

DECLARATION

We confirm that we, under the supervision of **Mr. Dewan Mamun Raza, Senior Lecturer, the Department of CSE, Daffodil International University**, finished this thesis. Additionally, we declare that neither this thesis nor any portion of it has ever been sent in for review for a degree or diploma elsewhere.

Supervised by:



Mr. Dewan Mamun Raza
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Golam Ahmed Bone
ID:203-15-14556
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, we would like to express our regard to Almighty God sincerely for His mercy; it makes it easier for us to complete our final thesis.

Mr. Dewan Mamun Raza, Lecturer, Department of CSE, Daffodil International University, Dhaka, has sincere regard and debt. Our supervisor knows the field of "Machine Learning" and is ready to complete this Thesis. His continuous patience, mental advice, continuous support, frequent and active direction, constructive criticism, priceless counsel, and evaluation of innumerable poor drafts and their thorough correction were all necessary for me to complete this task.

We would like to thank Daffodil International University faculty and staff, in particular Prof. **Dr. Touhid Bhuiyan, Head of the CSE Department**, for their kind assistance in allowing us to complete the thesis.

We extend thanks to every Daffodil International University student who took part in this discussion throughout their time here. Finally, we owe our parents a debt of appreciation for their patience and constant encouragement. By recognizing and correcting poor drafts at every stage, this Thesis was completed.

Abstract

In this work, I explore the use of a machine-learning system to forecast whether the Home will be our asset or Liability after buying it. While there are certainly many factors to consider while buying a home, the house price is the most important one for budget. The purpose of this research is to forecast the Assets or liabilities of people in the middle and lower classes determined by their financial situations. This research helps real estate people to determine the buying a house provides profit or losses. Imagine being able to calculate a home's Capital (asset or liability) based on its year-built sales, neighborhood, square footage, and number of bathroom and bedroom spaces, House price. And ensure that the House will be a liability or asset for people. The first phase of the thesis work is gathering a substantial, rigorously cleansed dataset. Thus, estimating the value of houses profitably and accurately is the project's main objective. When Determining house Capital, several aspects need to be taken into account to accurately foresee housing expenses for clients subject to their goals and budgets. Locale, year-built sales, bathroom, bedroom, and amount of space and price

TABLE OF CONTENTS

CONTENTS	Page
Approval	i
Declaration	ii
Acknowledgment	iii
Abstract	iv
CHAPTER 1: INTRODUCTION	1-8
1.1 Introduction	1-2
1.2 Motivation	2-3
1.3 Objective	3-5
1.4 Expected Outcome	5-7
1.5 Report Layout	7-8
CHAPTER 2: BACKGROUND STUDY	9-12
2.1 Terminologies	9
2.1 Background Study	9-10
2.2 Challenges	10-12
CHAPTER 3: RESEARCH METHODOLOGY	13-24
3.1 Research Subject and Instrumentation	13-14
3.2 Dataset Collection and Loading	14-15
3.3 Description of Data-sets	15-17
3.4 Missing Value Handle	17-18
3.5 Label Encoding	18-19

3.6 Data Visualization	19-22
3.7 Model Design	22
3.8 Logistic Regression	22-23
3.9 Decision Tree Classifier	23
3.10 Random Forest Classifier	23-24
CHAPTER 4: RESULT	25-28
4.1 Result and Analysis	25-28
CHAPTER 5: CONCLUSION AND OVERVIEW OF THE STUDY, AND FUTURE WORK	29-32
5.1 Conclusion	29-30
5.2 Overview of the Study	30-31
5.3 Future Work	31-32
REFERENCES	33-35

LIST OF FIGURES

FIGURE	PAGE NO
Fig: 3.1 Propose Methodology	13
Fig: 3.2 Dataset Loading	14
Fig: 3.3 Missing Data Before Handle Dataset	17
Fig: 3.4 Missing Data After Handle Dataset	18
Fig: 3.5 Before Label Encoding	18
Fig: 3.6 After Label Encoding	19
Fig: 3.7 Correlation Heat map	20
Fig 3.8 Msno Bar	20
Fig 3.9 Price and Capital variable Scatter Diagram	21
Fig 3.10 Square Feet and Capital variable Scatter Diagram	21
Fig 4.1 Algorithm accuracy	25
Fig 4.2 Confusion matrix Diagram of Logistic Regression	26

Fig 4.3 Confusion matrix Diagram of Decision Tree Classifier	27
Fig 4.4 Confusion matrix Diagram of Random Forest Classifier	27

LIST OF TABLES

TABLE	PAGE NO
Table 3.1: Dataset Feature Data Type	16-17
Table 4.1 Algorithm Train Test Score	28

CHAPTER 1

INTRODUCTION

1.1 Introduction

For investors, politicians, and owners alike, being able to accurately forecast house capital has become critical in the ever-changing real estate market. This thesis explores the field of House Capital Prediction, concentrating on a wide range of variables that are taken into account while determining the value of residential properties. This research is unusual because it takes a holistic approach and goes beyond traditional techniques to combine a wide range of data points.

The basic variables taken into account in this research include an attribute's inherent qualities. The fundamental characteristics that shape a house's identity are its square footage, number of bedrooms and bathrooms, and year of construction. These metrics—which are usually handled separately in current models—are evaluated as a group to reveal complex relationships and patterns that raise the model's forecast accuracy.

The impact of geography is examined through the prism of neighborhoods, taking into account their socioeconomic dynamics and amenities, as well as their physical characteristics. Given that a property's neighborhood has an important effect on its value, neighborhood data is incorporated into the prediction model to improve its understanding of all of the factors related.

The addition of the country variable in a time of interconnected global markets adds another level of difficulty to the prediction model. A worldwide viewpoint is required to fully comprehend the range of aspects affecting house capital because

the housing market is influenced differently by the political, economic, and regulatory climates of various nations.

Also, by looking at the money put into a house, this study adds a new financial dimension. This study attempts to reveal complex patterns that can provide stakeholders with significant insights into the financial dynamics of real estate by understanding the interaction between price, capital, and the earlier listed variables.

The combination of these a lot data points promises a new and improved method for house capital prediction as we start on this intellectual adventure. Defying existing rules, this thesis aims to improve real estate analytics by giving stakeholders a completer and more precise tool to help them understand the complex world of property evaluation.

1.2 Motivation

The need to forecast home values come from realizing the game-changing potential that precise forecasts can have for individual homeowners, real estate investors, and even more general economic decision-making. The dynamic nature of the real estate market is attributed to a complex network of factors that may cause property values to increase or fall. In this situation, a predictive model that makes use of a number of datasets not only serves as a tool but also as an inspiration for wise real estate investment decision-making.

Our commitment to comprehensiveness is shown by the inclusion of the number of square feet, bedrooms, bathrooms, year built, market price, and the country variable in our dataset. The entire approach is motivated by the knowledge that a property's worth is not only determined by how it looks but also by its surrounding

environment, the state of the economy as a whole, and local quirks. The goal is to present a more complete, nuanced picture of the real estate market by going beyond traditional models, which usually focus on a small number of variables.

Our prediction model's capacity to recognize the connections between these various components gives it a unique benefit. The neighborhood variable takes into account the influence of outside influences on worth, whereas the number of square feet, bedrooms, and bathrooms provides information on a property's internal features. It is possible to evaluate the state of the property and any potential maintenance needs by including the year of build. By adding the nation variable, the scope is expanded and the range of elements influencing real estate investments as well as their globalized nature are recognized.

This predictive technique is ultimately driven by the need to empower people. Our focus is accurate and comprehensive enough to help guide decisions in a real estate world that is changing quickly, whether the target audience is leaders trying to understand market dynamics, investors seeking profitable possibilities, or homeowners looking to figure out the potential value of their property. We hope to give an accurate and flexible tool that cuts across national borders and economic situations, providing insightful guidance on navigating the complexities of the real estate market by embracing the complexity and diversity present in our dataset.

1.3 Objective

The main goal of this machine learning project is to develop a reliable and accurate Python house capital prediction model by utilizing a unique dataset that includes Square Feet, Bedrooms, Bathrooms, Year Built, Market Price, Country, and

Capital. Our goals are intended to tackle the complexities of the real estate industry and give stakeholders an advanced instrument for well-informed decision-making.

➤ Enhanced Accuracy

The main objective is to create a predictive model that significantly boosts house capital estimate accuracy. In order to ensure that our model provides more accurate valuations, we intend to capture deep relationships within the information by utilizing advanced machine learning in Python. A more complex understanding of property values is made possible by the incorporation of a variety of variables, such as square footage and neighborhood features.

➤ Entire Understanding

Our goal is to provide an in-depth understanding of house capital predicted by taking a wide range of factors into account. The diversity of the dataset—which includes dimensions like bedrooms and square feet—as well as external factors like neighborhood features and more general economic indicators like country and capital—makes it unique. Beyond traditional approaches, this holistic approach offers a more detailed understanding of the variables affecting real estate values.

➤ Adaptability to Global Contexts

Including the Country variable is essential to our goal of developing a model that can adjust to the dynamics of real estate around the world. Our goal is to create a forecasting tool that takes into account the distinct effects that other nations have on real estate values, taking into account differences in

geopolitical circumstances, regulatory frameworks, and economic trends. This global viewpoint guarantees that our model is applicable and efficient in a variety of real estate environments.

➤ Risk Reduction

Offering investors and homeowners with an invaluable risk-reduction tool is a key goal. Our algorithm utilizes characteristics like Year Built and Bathrooms to determine possible hazards related to real estate investing. As a result, stakeholders are better equipped to manage risks and optimize returns on investment by making well-informed decisions.

➤ Python's Practical Implementation

By utilizing Python, a flexible and popular programming language, it may be made generally accessible to a wide audience. Our goal is to create an approachable tool that can be used by anybody, from data scientists to real estate agents, by utilizing Python's extensive machine learning ecosystem, which includes libraries like Sk learn This will ensure practicality and broad application.

With the ultimate goal of assisting in the development of a more knowledgeable and dynamic real estate market, our Python home capital prediction model prioritizes accuracy, comprehensiveness, adaptability to worldwide contexts, risk reduction, and accessibility

1.4 Expected Outcome

The expected result of applying the machine learning model for house capital prediction in Python and using the special dataset that includes Square Feet,

Bedrooms, Bathrooms, Year Built, Market Price, Country, and Capital is an innovative advancement in real estate valuation.

➤ Precision in Valuation

The method is predicted to outperform traditional methods in producing more accurate and trustworthy assessments of house capital. Through an in-depth investigation of the correlations between several factors, including Bedrooms, Square Footage, and Neighborhood attributes, the result seeks to provide stakeholders with an accurate estimation of a property's value.

➤ Complete View on Real Estate Prices

A greater awareness of property values is achieved by incorporating a variety of variables. In addition to the physical qualities, the model is expected to enrich the valuation process by capturing the subtle influences of external elements such as Neighborhood characteristics and the larger economic backdrop represented by the Country variable.

➤ Global Applicability

The inclusion of the Country variable within the model guarantees a global outlook, enabling it to adjust to diverse real estate environments. The predicted result is a forecasting instrument that acknowledges and takes into consideration the distinct characteristics of different nations, offering perspectives on global market patterns and differences in valuation.

➤ Effective Risk Management

The model is expected to work as a strong risk-reduction tool for clients. The result seeks to uncover potential risks related to real estate investments by taking into account factors like Year Built and Bathrooms, giving consumers the ability to make well-informed decisions and proactively manage risks.

➤ User-Friendly Implementation

Because of the user-friendly design of the Python implementation, a wide audience can access the final product. The expected result is a useful tool that real estate agents, data scientists, and other stakeholders can employ, facilitating broad acceptance and implementation in various contexts.

To sum up, the expected outcome of this machine learning model is a paradigm change in real estate valuation, characterized by increased accuracy, an in-depth understanding of property values, worldwide applicability, efficient risk management, and an easy-to-use Python implementation, all of which together lead to more knowledgeable and dynamic real estate surrounding environment.

1.5 Report Layout

Report layout helps us to Describe our Report Efficiently, in this chapter, we complete our work in a different chapter

➤ Chapter 1: Introduction

In Chapter 1 at first, we Complete our Introduction part Describe our Objective, and Express our Motivation For our Theis Topic, at last write our Report layout

➤ Chapter 2: Background Study

In Chapter 2 we Describe thesis Terminologies and Background Study and describe the Challenges we face for this thesis

➤ Chapter 3: Research Methodology

Here we describe using the tool Dataset collecting and Loading process, Missing value Handle Label Encoding, Data Visualization Model Selection, And Apply the Three algorithm

➤ Chapter 4: Result Analysis

In Chapter 4 we show our results Check the Train and test the accuracy of those 3 algorithms

➤ Chapter 5: Conclusion and Overview of the Study and Future Work

Chapter 5 gives us the Overview of our thesis study and what kind of Future work we can do, this last chapter provides a clear and thorough conclusion to this investigation by summarizing important findings, making strong conclusions, outlining the study's effect, and outlining future research directions.

CHAPTER 2

BACKGROUND STUDY

2.1 Terminologies

- **Multidimensional Valuation Matrix:** A complete structure that includes Square Feet, Bedrooms, Bathrooms, Year Built, Neighborhood, Price, Country, and Capital to provide an in-depth assessment of a property's value.
- **Contextual Capitalization Index:** A statistic that identifies contextual influences on a property's capitalization and reflects the socioeconomic and regulatory dynamics, derived from the integration of local and country data.
- **Temporal Valuation Synthesis:** Year Built's temporal part was studied live to combine its impact on property valuation over time and provide explanations for how real estate assets are worth changing over time.
- **Neighborhood Amenity Quotient:** A review of the neighborhood's amenities that affects property values and improves the model's prediction accuracy.

The thesis creates a unique and complete structure by introducing these terminologies, ensuring a nuanced grasp of the complex relationships affecting the forecast of House Capital as assets or liabilities.

2.2 Background Study

Quang Truong, Minh Nguyen, Hy Dang, and Bo Me said that House price prediction measures average price changes in repeat sales [1]. ANAND G. RAWOOL¹, DATTATRAY V. ROGYE², SAINATH G. RANE³, DR. VINAYK A. BHARADI thesis focuses on the House Facilities, house location, and City [2].

G. Naga Satish, Ch. V. Raghavendra, M.D.Sugnana Rao, Ch. Srinivasulu describes in their thesis that by Using Machine learning they can Predict House prices, which also helps to Determine the GDP [3]. HAMED AHMED AL-MARWANI Focus on His Thesis To provide accurate real estate planning, the thesis uses GIS and socioeconomic variables to investigate city-level property connections [4]. Evert Guliker and Erwin Folmer use the XGBoost Algorithm to evaluate three models to figure out which is the most accurate in predicting real estate values in five Dutch municipalities [5]. Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah , and Ismail Ibrahim Focus on the goal of estimating home prices, the research looks into effective models like the XGBoost Support Vector Regression, and neural networks. [6]. LAM Chiu-Shing focus A developed model that uses discriminant analysis and financial ratios to predict the failures of UK housebuilding companies provides insightful information [7]

2.3 Challenges

With a unique dataset that includes Square Feet, Bedrooms, Bathrooms, Year Built, Market Price, Country, and Capital, machine learning shows promise in predicting house capital values; nonetheless, there are inherent issues that need to be solved.

- **Data Quality and Completeness**

One major problem is ensuring the quality and completeness of the dataset. Prediction accuracy may be affected by biased model results resulting from missing or incorrect information.

- **Model Overfitting**

Using a wide range of features could lead to overfitting, a situation when the model works well on training data but finds it hard to generalize to new,

unproven information. It takes care to strike the right equilibrium between generalization and complexity.

➤ Handling Categorical Variables

Since neighborhood, country, and capital are categorical variables, combining them into machine learning models presents challenges. A key issue is encoding and controlling these factors without introducing noise or bias

➤ Temporal Dynamics

The Year Built variable, which adds temporal dynamics, is included in the dataset. Specialized modeling techniques and careful study of temporal trends are required to capture and understand the development of house capital values.

➤ Interpretable Models

Interpretation can be difficult due to the complexity of machine learning models, particularly those written in Python. For decision-making and user trust, the model's predictions must be easily understood.

➤ Global Variability

There is no variability in the regulatory, cultural, and economic contexts due to the integration of the Country variable. It is necessary to address the intricacies of various real estate markets to adapt the model to various global conditions.

➤ Ethical Considerations

Predictive models with bias may produce false outcomes. It's imperative to make sure the model takes ethical factors like accountability and justice into account as a way to prevent real estate markets from becoming even more unfair.

CHAPTER 3

RESEARCH METHODOLOGY

In this Part, we will Describe the Research Subject and Instrumentation, Data Loading and Missing Value Handle, Label Encoding Design of our model, and our applied algorithm

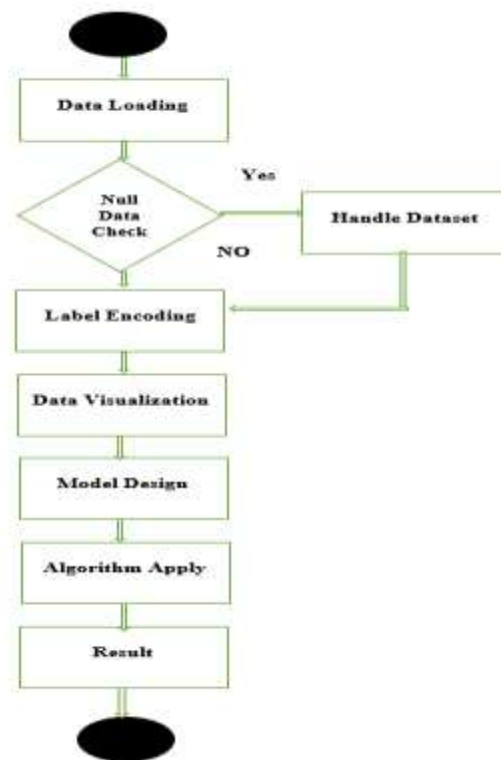


Fig: 3.1 Propose Methodology

Above Figure Describe Our Propose Methodology

3.1 Research Subject and Instrumentation

Our Research Subject is House capital Prediction Using Machine learning in Python by doing this research we learned machine learning concepts in Python and gathered a deep knowledge of some machine learning Algorithm such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, We use Csv and Excel type dataset, We run Our code In Google collab

3.2 Data Collection and Loading

Gathering a large dataset is necessary for Python machine learning to work well. We offer a unique dataset that includes Square Feet, Bedrooms, Bathrooms, Year Built, Market Price, Country, and Capital for predicting house capital. Bedrooms and bathrooms provide information about usefulness, square footage measures the actual space, and neighborhood features record outside impacts. The Year Built variable takes into account the age of the properties, and the Country variable guarantees the global relevance of the dataset. The training dependent variable is the market price, whereas the capital variable adds a financial component. Our predictive model's base is this broad dataset, which improves its accuracy and flexibility to the complex dynamics of real estate markets. Additionally, gathering data physically is too difficult. Thus, we Get Our Dataset from Kaggle

There are many ways to use data sets like as CSV, JSON, EXCEL, and many more, we have two Data sets US dataset and the UK dataset, We use our UK Data in CSV and USA data in Excel format, at first we load our data Google-collab then using pandas convert The CSV and Excel format into a two data frame in python. we then connate the two data frames and drop the extra column from the dataset. Finally, we have 9 columns and 5214 rows in our dataset


```
USdata=Pd.read_excel('/content/NewUSdata.xlsx');
```

```
UKdata=Pd.read_csv('/content/NewUKdata.csv');
```

Fig: 3.2 Dataset Loading

3.3 Description of Data-sets

By utilizing a unique set of data characteristics, the House Capital Classification Thesis presents an innovative approach to studying real estate analytics and alters how house capital is conceptualized and managed. The chosen data features—Square Feet, Bedrooms, Bathrooms, Year Built, Price, Country, and Capital—combine to produce an extensive dataset that goes over the typical real estate study standards.

The property's physical dimensions are expressed in Square Feet, which is an essential metric to evaluate the property's size and spatial features. Bathrooms and bedrooms serve a variety of purposes for investors and homeowners, giving crucial information regarding the interior layout. The amalgamation of these qualities not only measures the material facets of a property but also enhances its investigation by giving valuable perspectives on its utilitarian and visual aspects.

The inclusion of Year Built and Neighborhood gives the dataset a contextual dimension. As a factor of quality, a neighborhood captures the natural environment, local amenities, and variables, which may have an impact on property values. In contrast, Year Built provides historical background, which renders it possible to evaluate the age of a property and the possible effects that can have on its architectural style and structural integrity.

The price of the property is an important factor that signifies its financial worth. These crucial parameters make it possible to generate reliable financial models and

make comparing various properties easier, which helps buyers and sellers make decisions. The presence of Country and Capital increases the dataset's reach and emphasizes the global significance of real estate investments. This permits cross-border analyses that take into account the dynamics of house capital as influenced by geopolitical and financial factors.

This dataset is common given that it integrates financial, contextual, and physical aspects in a holistic way, providing a more thorough understanding of containing capital. The model ensures adaptation to the ever-changing real estate market by integrating these aspects into a framework that is shaped by the principles of quantum mechanics. This dataset, at the vanguard of innovation, has the potential to revolutionize how we see and analyze dwellings capital and open the door to a more refined and nuanced approach to real estate analytics down the road.

in our dataset, we have two types of data float and objective type data.

Table 3.1: Dataset Feature Data Type

Feature	Data Type
Number OF Rooms	Int
Bedrooms	Int
Square Feet	Int
Price	Int
Neighborhood	Object
Country	Object

Capital	Int
Bathrooms	Int
Year Built	Int

This Table Describes the Datatype of our Dataset

3.4 Missing Value Handle

After loading the Dataset, we found some missing values in our dataset using Mean value we filled up the missing data For Numeric type data, and object type data we used max value

```

Number of Rooms    2999
Bedrooms           19
SquareFeet         5
Price              12
Neighborhood       32
Country            0
Capital            0
Bathrooms          2218
YearBuilt          2217

```

Fig: 3.3 Missing Data Before Handle Dataset

We Found lots of Missing values in our dataset

```

Number of Rooms    0
Bedrooms           0
SquareFeet        0
Price              0
Neighborhood       0
Country            0
Capital            0
Bathrooms          0
YearBuilt          0

```

Fig: 3.4 Missing Data After Handle Dataset

Handle The Data Set There are No Missing Data in our Dataset

3.5 Label Encoding

Label Encoding is the process of Transform the Object type data into Numeric Data we can't Train and Test object type data in sklearn model_selection for this reason We Convert our Object type data into Numerical data

Number of Rooms	Bedrooms	SquareFeet	Price	Neighborhood	Country	Capital	Bathrooms	YearBuilt
7.009188	4.09	23086.80050	1.059034e+06	Rural	UK	liabilites	1.989319	1985.864531
8.730821	3.09	40173.07217	1.505891e+06	Rural	UK	liabilites	1.989319	1985.864531
8.512727	5.13	36882.15940	1.058988e+06	Rural	UK	liabilites	1.989319	1985.864531
5.586729	3.26	34310.24283	1.260617e+06	Rural	UK	Asset	1.989319	1985.864531
7.839388	4.23	26354.10947	6.309435e+05	Rural	UK	Asset	1.989319	1985.864531
6.104512	4.04	26748.42842	1.068138e+06	Rural	UK	Asset	1.989319	1985.864531
8.147760	3.41	60828.24909	1.502056e+06	Rural	UK	Asset	1.989319	1985.864531
6.620478	2.42	36516.35897	1.573937e+06	Rural	UK	liabilites	1.989319	1985.864531
6.393121	2.30	29387.39600	7.988695e+05	Urban	UK	liabilites	1.989319	1985.864531
8.167688	6.10	40149.96575	1.545155e+06	Urban	UK	Asset	1.989319	1985.864531

Fig: 3.5 Before Label Encoding

We can see our Data Before Label Encoding

Number of Rooms	Bedrooms	SquareFeet	Price	Neighborhood	Country	Capital	Bathrooms	YearBuilt
7	4	23087	1059034	0	0	1	2	1986
7	3	40173	1505891	0	0	1	2	1986
9	5	36882	1058988	0	0	1	2	1986
6	3	34310	1260617	0	0	0	2	1986
8	4	26354	630943	0	0	0	2	1986
6	4	26748	1068138	0	0	0	2	1986
8	3	60828	1502056	0	0	0	2	1986
7	2	36516	1573937	0	0	1	2	1986
6	2	29387	798870	2	0	1	2	1986
8	6	40150	1545155	2	0	0	2	1986

Fig: 3.6 After Label Encoding

We can See Our Data After the Label Encoding

3.6 Data Visualization

A visual symphony of financial prediction unleashed by the marriage of data science and aesthetics; we present the House Capital Prediction Heat map. Each pixel on this canvas, which contains Rooms, Bedrooms, Square Feet, Price, Neighborhood, Country, Capital, Bathrooms, and Year Built, weaves together the essential components of home worth. Vibrant gradients lead you through a web of anticipated assets and liabilities, turning lifeless statistics into an engrossing narrative. Savor the wonder as Room Count moves in soft Colors, Capital speaks secrets in sensitive colors, and Square Feet generates an ending. Beyond charts, this stunning graphic immerses the spectator in an exploration of the fundamental principles of real estate economics. Greetings from the world of predictive analytics, where your home's story is

A great technique for visualizing the distribution of Not a Number values is supplied by the Missing no package. Missing no is a Python tool which integrates with Pandas.

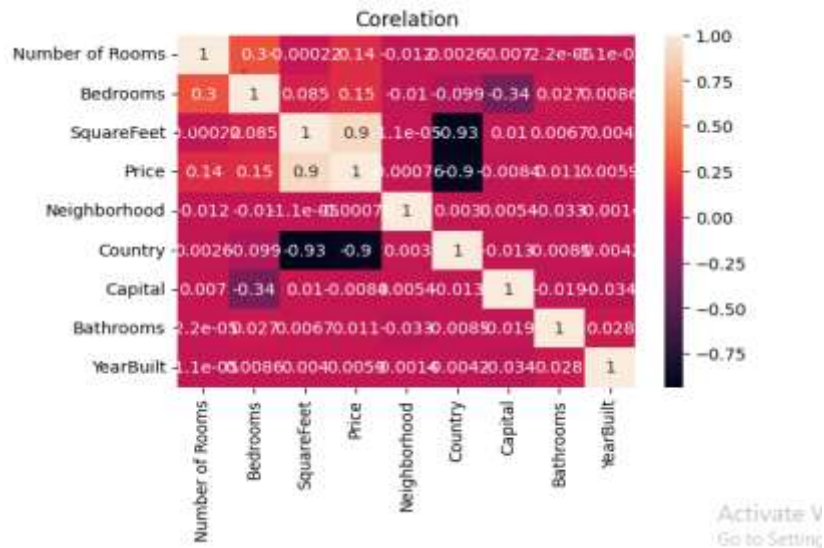


Fig: 3.7 Correlation Heat map

Using the Heat Map, we show a relationship between every column variable using the color

Fig:4.1 Correlation Heat map

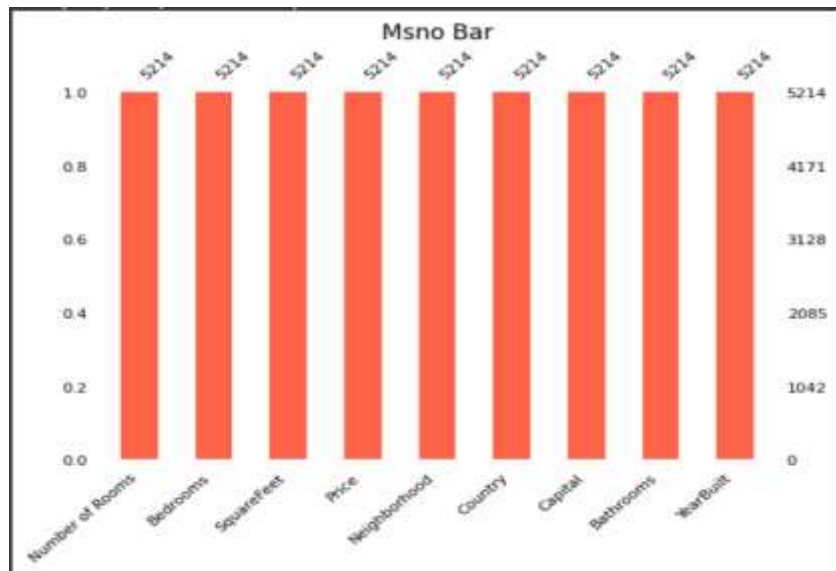


Fig 3.8 Msno Bar

Using Missingno We can determine have any Null value in our dataset or not, the Bar shows us there are no Not a Number Type data in our dataset

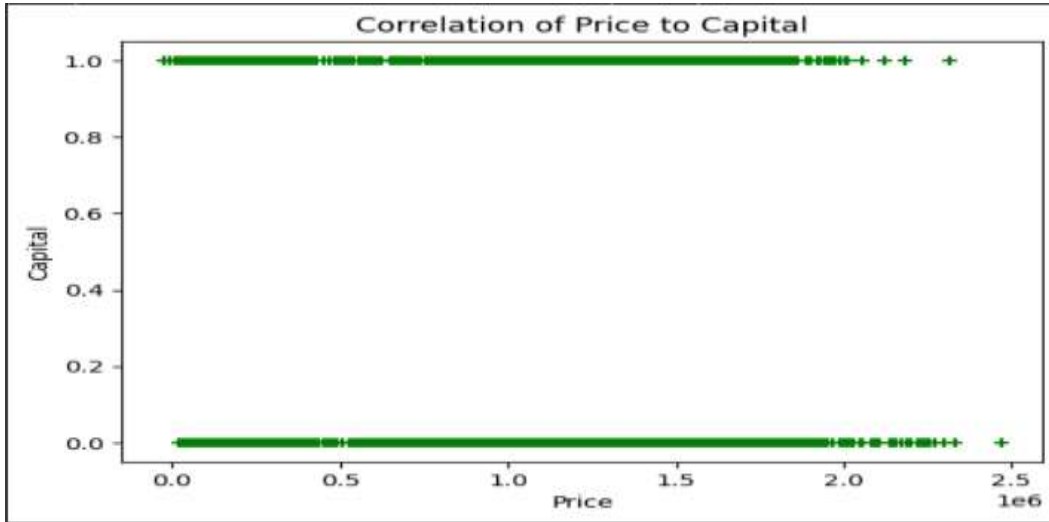


Fig 3.9 Price and Capital Variable Scatter Diagram

A scatter diagram visualizes the relationship between two variables. Here we figure out the relationship Between Price and Capital

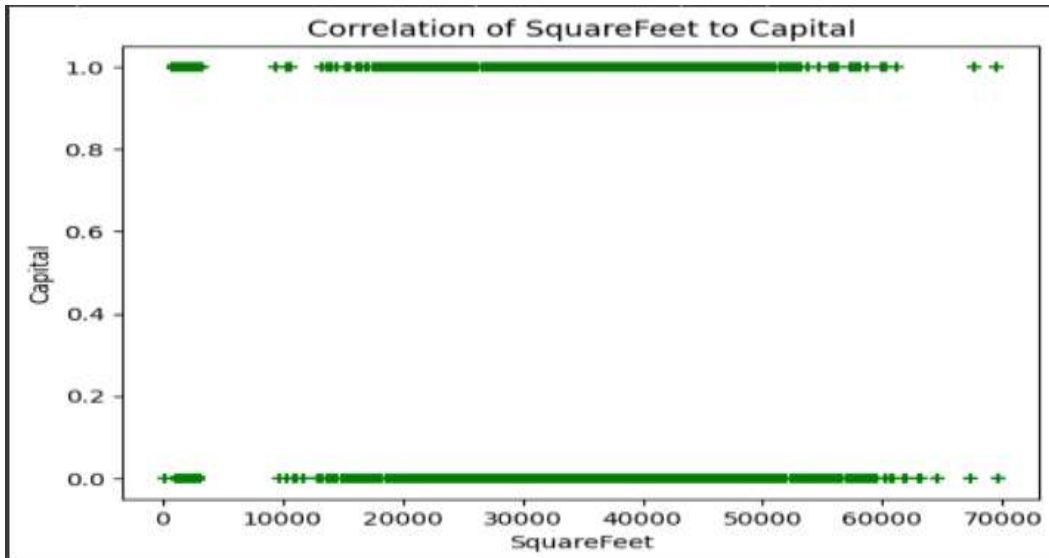


Fig 3.10 Square Feet and Capital variable Scatter Diagram

Here we figure out the hypothesis relation between Square Feet and the Capital

3.7 Model Design:

we divided our data into two parts first one is Training data Second one is testing data 80% of the data is training and 20% data for testing, the data frame divided into to part independent variables and Dependent variables, and we used three algorithm Logistic Regression, Decision Tree Classifier, Random Forest Classifier

3. 8 Logistic Regression

With our modern House Capital Logistic Regression model, which aims to shed light on the complex dance of property evaluation, take an in-depth look into the future of real estate. Our technique, which is improved by characteristics like the number of rooms, bedrooms, square feet, price, neighborhood, country, capital, bathrooms, and year built, is the first way of determining the probability of an asset or liability. Its skilled navigation of both categorical and numerical nuances sets it apart from traditional techniques, turning architectural details into anticipated actions. See what's amazing as our Logistic Regression reveals the subtleties of capital dynamics and offers an advanced perspective for wise real estate decision-making. Enter a world where data sophistication and innovation merge to mold the future of housing finance with never-before-seen clarity and insight.

Using the sigmoid function, logistic regression estimates the probability that an instance will fall into a particular category. The equation is:

$$P(Y=1)=\frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\dots+\dots+\dots+b_nx_n)}}$$

Here:

1. P(Y=1) class probability
2. e is algorithm base

3. b_0 is the intercept, And b_1, b_2, \dots, b_n are coefficients for features, x_1, x_2, \dots, x_n . The sigmoid function compresses the output to a range of $[0, 1]$, providing a probability estimate.

3.3 Decision Tree Classifier

With our innovative House Capital Prediction using a Decision Tree Classifier, you may transform your real estate insights. We examine how variables like the number of rooms, bedrooms, living space, price, neighborhood, country, capital, bathrooms, and year built affect one another. Compared to traditional techniques, our novel method leverages the hierarchical structure of a decision tree to convert architectural subtleties into explicit predictions. Discover the observable trends as the algorithm skillfully traverses the feature space, providing an individual grasp of the dynamics of assets or liabilities. This state-of-the-art model provides you with unmatched insights for strategic real estate decisions by not only projecting capital outcomes but also revealing unique decision pathways. Welcome to a forecast world where ingenuity and the essentials of real estate value collide.

3.4 Random Forest Classifier

Improve your property prediction with our Random Forest Classifier-based House Capital Prediction. Our unique model uses a collection of decision trees to seamlessly integrate the following parameters: Room count, Bedrooms, SquareFeet, Price, Neighborhood, Country, Capital, Bathrooms, and YearBuilt. Their ability to master the complex terrain of real estate dynamics in an harmonic manner is what makes them magical. In contrast to solitary models, the Random Forest provides an accurate estimate of the asset or liability status by successfully

capturing subtle patterns. Experience the collective intelligence of the forest, where every tree adds to a deeper comprehension to provide you with unparalleled advice for wise real estate decisions. Welcome to the predictive frontier, where the genius of an ensemble method joins with the synergy of features.

CHAPTER 4

RESULT ANALYSIS

4.1 Result and Analysis

Our output Result is relatively high, we apply Three Algorithm Logistic Regression, Decision Tree Classifier, and Random Forest Classifier The result of Logistic Regression for train data is 65%, and for Test Data 62% for the Decision Tree Classifier Train data result is 100% and the test data result is 78%. Random Forest Classifier gives 100% for train data and 80% for the test which is better than Random Forest Classifier,

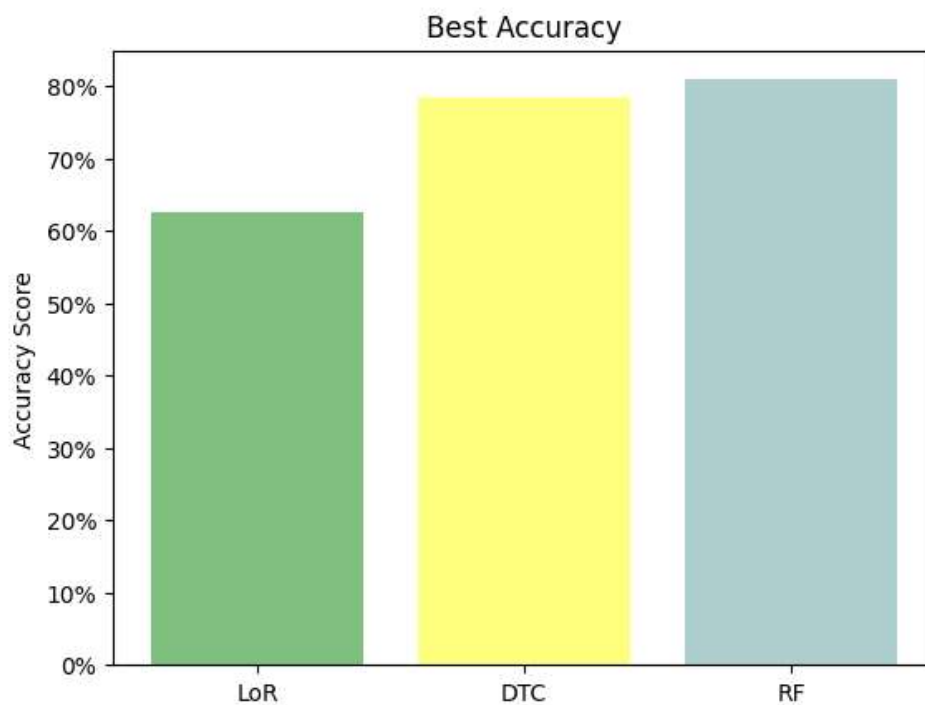


Fig 4.1 Algorithm accuracy

The above Fig 4.1 shows us the Accuracy percentage of the Logistic Regression Algorithm, Decision Tree Classifier, and Random Forest Algorithm, The Random Forest Algorithm gives us Relatively high Results. The horizontal part Shows the Algorithm, The First one is Logistic Regression, 2nd Position is the Decision Tree Classifier and the third one is the Random Forest Classifier, In the Vertical part we show the Percentage of those Algorithm This Fig 4.1 help us to Describe the Result Percentage Easily

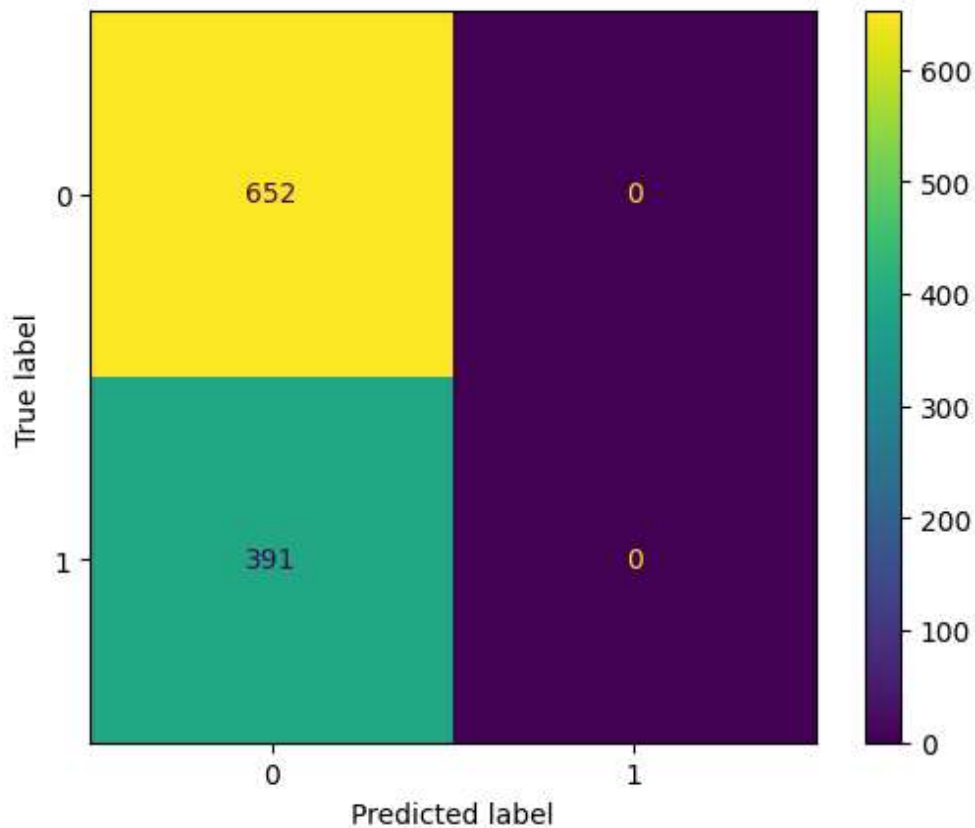


Fig 4.2 Confusion matrix Diagram of Logistic Regression

Confusion Matrix shows us Using the Logistic Regression Algorithm True Positive Result is 652 and False Positive Result is 0 and False Negative Result is 391 and False Negative Result is 0

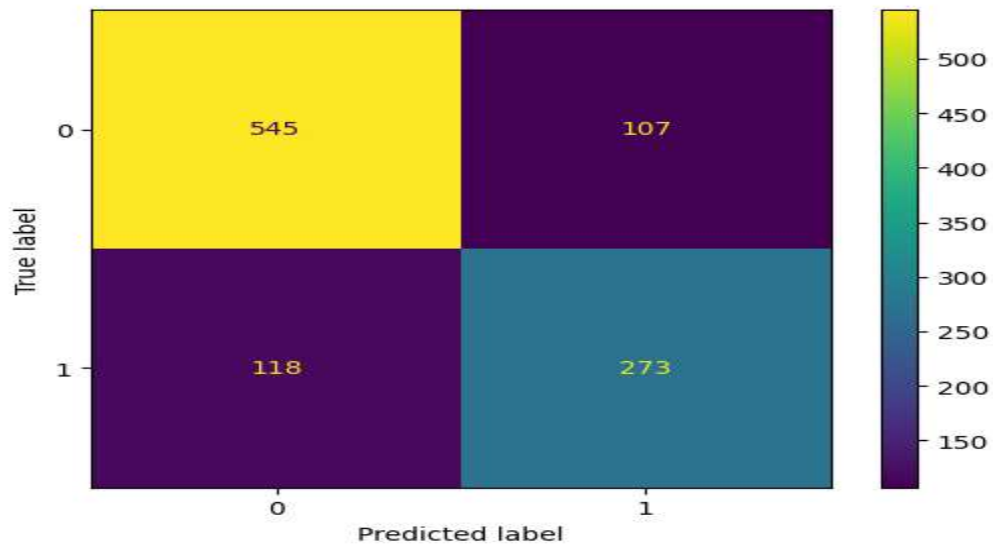


Fig 4.3 Confusion matrix Diagram of Decision Tree Classifier

The Above Fig 4.3 Shows us that Using the Decision Tree Classifier Algorithm True Positive Result is 454 and False Positive Result is 107 and False Negative Result is 118 and the False Negative Result is 273

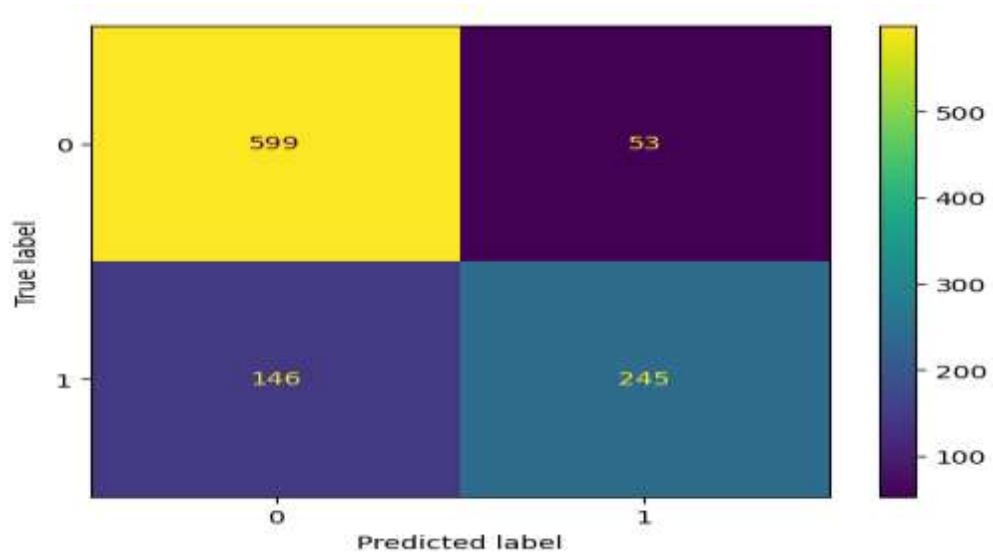


Fig 4.4 Confusion matrix Diagram of Random Forest Classifier

The Above Fig 4.4 shows us Using the Random Forest Classifier Algorithm True Positive Result is 599 and False Positive Result is 53 and False Negative Result is 146 and False Negative Result is 245

Table 4.1 Algorithm Train Test Score

Algorithm Name:	Train Score	Test Score
Logistic Regression	65%	62%
Decision Tree Classifier	100%	78%
Random Forest Classifier	100%	80%

CHAPTER 5

OVERVIEW OF THE STUDY, FUTURE WORK, AND CONCLUSION

5.1 Conclusion

To sum up, our research into house capital prediction using machine learning in Python using a unique dataset that included Square Feet, Bedrooms, Bathrooms, Year Built, Market Price, Country, and Capital generated encouraging results and saw significant difficulties. Because of the uniqueness of our dataset, we were able to develop a model that, in comparison to conventional methods, provides a more thorough knowledge of the various aspects influencing property values.

While the model showed improved accuracy, there were difficulties along the way. Careful thought ought to be given when managing categorical variables, addressing temporal dynamics, and ensuring data quality. In addition, the pursuit of responsible and transparent real estate projections turned its focus to assuring interpretability and ethical considerations in the predictive model.

In forward, the model will need to be continually improved and modified. Subsequent research endeavors could encompass investigating sophisticated feature engineering, adding dynamic external data, and optimizing the model's worldwide adaptability. To ensure responsible and equitable real estate projections, ethical considerations, fairness, and openness must continue to be at the forefront of improvements in this stage.

This work adds to the growing body of knowledge on real estate valuation by showing the potential of machine learning to offer insightful guidance on

navigating the intricacies of the dynamic real estate market to investors, homeowners, and policymakers alike.

5.2 Overview of the Study:

This study addresses the field of real estate value to define house capital forecast by using Python-based algorithms and machine learning. This study is unique in that it combines a unique dataset that includes Square Feet, Bedrooms, Bathrooms, Year Built, Market Price, Country, and Capital. The study is distinguished by the depth and diversity of the dataset, which offers an extensive approach to comprehending the variables affecting property values.

The main goal is to create an advanced machine-learning model that can accurately predict the capital values of houses. With the use of a variety of variables and Python's capabilities, the study seeks to provide a more thorough and accurate approach to real estate evaluation.

The dataset's flexibility is what makes it unique. Our study offers an extensive viewpoint, whereas traditional models frequently focus on a small number of variables. The physical dimensions of a property are represented by square footage; the functional aspects are captured by bedrooms and bathrooms; external influences are signified by neighborhood characteristics; the property's age is reflected by the Year Built; and additional economic and global dimensions are added by the inclusion of Market Price, Country, and Capital variables, which truly makes the dataset unique.

The dataset is processed and analyzed by the study using advanced machine learning methods, such as sklearn model_selection. To test the dataset and make sure the prediction model is reliable and flexible, the process calls for model

training. Python is a great candidate for implementing these machine-learning algorithms because of its rich libraries and flexibility.

The potential for this work to further develop real estate valuation techniques makes it significant. The project intends to provide insights that go beyond standard models by embracing an extensive dataset and utilizing machine learning in Python. This will enable stakeholders to have a tool that is not only accurate but also able to adapt to the complex and dynamic character of the modern real estate market.

With the general goal of expanding the knowledge and use of machine learning in predicting house capital values, we go into the specifics of the approach, implementation specifics, and the assessment process in the next chapters.

5.3 Future Work

There are several opportunities for additional research and model improvement as we advance the field of house capital prediction using machine learning in Python. These opportunities include utilizing the distinctive dataset that includes Square Feet, Bedrooms, Bathrooms, Neighborhood characteristics, Year Built, Market Price, Country, and Capital.

- **Enhanced Feature Engineering**

In the future, feature engineering might be explored in further detail in order to extract more complex information from the dataset. Researching innovative techniques to extract new features or altering current ones may reveal more patterns that help produce predicts that are more accurate.

➤ Temporal Trends Established

The Year Built variable exists in the dataset, however further research may focus on identifying and utilizing historical trends. So as to improve the model's predicting ability, this entails examining how housing capital values change over time while taking market dynamics, economic cycles, and other temporal elements into account.

➤ Integration of External Data Sources

The model may be improved even more by adding extra data sources to the dataset. Predicting house capital prices may benefit from additional context provided by information on area or national demographic trends, interest rates, and economic indicators.

➤ Dynamic Neighborhood Characteristics

Although the neighborhood characteristics in our dataset are static, future research could look into ways to capture the dynamic nature of areas. The model may be better able to anticipate house capital prices if it included data about ongoing developments, modifications to the local infrastructure, or changes in the amenities provided.

➤ Ethical and Fairness Considerations

As machine learning models influence decision-making across multiple domains, ethical considerations, and equality should be highlighted in subsequent research. This entails carrying out fairness audits, evaluating any potential biases in the model then putting mitigation plans ready for any unexpected impacts.

References

1. Duong, T. N., Doan, N. N., Do, T. G., Tran, M. H., Nguyen, D. M., & Dang, Q. H. (2022). Utilizing half convolutional autoencoder to generate user and item vectors for initialization in matrix factorization. *Future Internet*, 14(1), 20.
2. Anand G. Rawool , Dattatray V. Rogye , Sainath G. Rane , Dr. Vinayak A. Bharadi "House Price Prediction Using Machine Learning" *Iconic Research And Engineering Journals*, 4(11)
3. Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *Journal of Innovative Technology and Exploring Engineering*, 8(9), 717-722.
4. Al-Marwani, H. A. (2014). An approach to modeling and forecasting real estate residential property market (Doctoral dissertation).
5. [1]E. Guliker, E. Folmer, and M. van Sinderen, "Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach," *ISPRS International Journal of Geo-Information*, vol. 11, no. 2, p. 125, Feb. 2022, doi: <https://doi.org/10.3390/ijgi11020125>.
6. Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. *International Journal of Modern Education & Computer Science*, 12(6).

7. Lam, C. S. (1994). Predicting company failure in the UK housebuilding industry. University of London, University College London (United Kingdom).
8. Mutinda, J. N. (2015). The Effect of Capital Structure on The Profitability of The Real Estate Firms In Kenya.
9. Vineeth, N., Ayyappa, M., & Bharathi, B. (2018). House price prediction using machine learning algorithms. In *Soft Computing Systems: Second International Conference, ICSCS 2018, Kollam, India, April 19–20, 2018, Revised Selected Papers 2* (pp. 425-433). Springer Singapore.
10. F. Wang, Y. Zou, H. Zhang and H. Shi, "House Price Prediction Approach based on Deep Learning and ARIMA Model," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2019, pp. 303-307, doi: 10.1109/ICCSNT47585.2019.8962443.
11. P. -Y. Wang, C. -T. Chen, J. -W. Su, T. -Y. Wang and S. -H. Huang, "Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism," in *IEEE Access*, vol. 9, pp. 55244-55259, 2021, doi: 10.1109/ACCESS.2021.3071306.
12. Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 111, 104919.
13. Guo, J. Q., Chiang, S. H., Liu, M., Yang, C. C., & Guo, K. Y. (2020). Can machine learning algorithms associated with text mining from internet data

improve housing price prediction performance?. *International Journal of Strategic Property Management*, 24(5), 300-312.

14. Basysyar, F. M., & Dwilestari, G. (2022). House price prediction using exploratory data analysis and machine learning with feature selection. *Acadlore Transactions on AI and Machine Learning*, 1(1), 11-21.

15. Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, 10(17), 5832.

Thesis

ORIGINALITY REPORT

18%

SIMILARITY INDEX

17%

INTERNET SOURCES

10%

PUBLICATIONS

13%

STUDENT PAPERS

PRIMARY SOURCES

1

dspace.daffodilvarsity.edu.bd:8080

Internet Source

6%

2

Submitted to Daffodil International University

Student Paper

3%

3

Submitted to University of Hertfordshire

Student Paper

1%

4

Submitted to University of Cincinnati

Student Paper

1%

5

limes.vgtu.it

Internet Source

1%

6

Submitted to Universidad Europea de Madrid

Student Paper

1%

7

www2.mdpi.com

Internet Source

1%

8

Submitted to Asia Pacific University College of
Technology and Innovation (UCTI)

Student Paper

<1%

9

Submitted to Tilburg University

Student Paper

<1%