PROFESSIONAL TRAINING REPORT

At

Sathyabama Institute of Science and Technology (Deemed to be University)

Submitted in partial fulfilment of the requirements for the award of Bachelor of EngineeringDegree in Computer Science and Engineering

By

PANKHURI SANTOSHI
REG. NO. 39110740

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI
SALAI,CHENNAI – 600119, TAMILNADU

APRIL 2022

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC (Established
under Section 3 of UGC Act, 1956) JEPPIAAR NAGAR,
RAJIV GANDHI SALAI
CHENNAI– 600119
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **PANKHURI SANTOSHI (Reg. No: 37110740)** who carried out the project entitled **"HOUSE PRICE PREDICTION**" under my supervision from March 2022 to April 2022.

Internal Guide

Mr Murali- BE., M.TECH

Assistant Professor

SUBMITTED FOR VIVA VOCE EXAMINATION HELD ON____________________

Internal Examiner **External Examiner**

DECLARATION

I, **PANKHURI SANTOSHI** hereby declare that the project report entitled **House Price Prediction** done by me under the guidance of **Mr. Murali BE., M.Tech** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

DATE:

PLACE: **SIGNATURE OF CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them. I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr. S. Vigneshwari, M.E., Ph.D., and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Mr. Murali BE., M.Tech** for his valuable guidance, suggestions, and constant encouragement paved way for thesuccessful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

ABSTRACT

The field of machine learning is introduced at a conceptual level. Ideas such as supervised and unsupervised as well as regression and classification are explained. The tradeoff between bias, variance, and model complexity is discussed as a central guiding idea of learning. Various types of models that machine learning can produce are introduced such as the neural network (feed-forward and recurrent), support vector machine, random forest, self-organizing map, and Bayesian network. Training a model is discussed next with its main ideas of splitting a dataset into training, testing, and validation sets as well as performing cross-validation. Assessing the goodness of the model is treated next alongside the essential role of the domain expert in keeping the project real.

In the following project we have tried implementing various machine learning algorithm in order to predict house prices according to various attributes. The method with the best accuracy is than deployed to predict the price according to the user’s need. Various attributes used in this project were location, bathroom, bedroom and square feet of the house.

.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Machine Learning:

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Recommendation engines are a common use case for machine learning. Other popular uses include fraud detection, spam filtering, malware threat detection, business process automation (BPA) and Predictive maintenance.

1.2 Importance of Machine Learning:

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

1.3 Types of Machine Learning:

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

a. Supervised learning: In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.
b. Unsupervised learning: This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.
c. Semi-supervised learning: This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

d. Reinforcement learning: Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

1.3.1 Supervised Machine Learning:

Supervised machine learning requires the data scientist to train the algorithm with both labeled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:

Binary classification: Dividing data into two categories.
Multi-class classification: Choosing between more than two types of answers.
Regression modeling: Predicting continuous values.
Ensembling: Combining the predictions of multiple machine learning models to produce an accurate prediction.

1.3.2 Unsupervised Machine Learning:

Unsupervised machine learning algorithms do not require data to be labeled. They sift through unlabeled data to look for patterns that can be used to group data points into subsets. Most types of deep learning, including neural networks, are unsupervised algorithms. Unsupervised learning algorithms are good for the following tasks:

Clustering: Splitting the dataset into groups based on similarity.
Anomaly detection: Identifying unusual data points in a data set.
Association mining: Identifying sets of items in a data set that frequently occur together.
Dimensionality reduction: Reducing the number of variables in a data set.

1.3.3 Semi-supervised Machine Learning:

Semi-supervised learning works by data scientists feeding a small amount of labeled training data to an algorithm. From this, the algorithm learns the dimensions of the data set, which it can then apply to new, unlabeled data. The performance of algorithms typically improves when they train on labeled data sets. But labeling data can be time consuming and expensive. Semi-supervised learning strikes a middle ground between the performance of supervised learning and the efficiency of unsupervised learning. Some areas where semi-supervised learning is used include:

Machine translation: Teaching algorithms to translate language based on less than a full dictionary of words.
Fraud detection: Identifying cases of fraud when you only have a few positive examples.
Labelling data: Algorithms trained on small data sets can learn to apply data labels to larger sets automatically.

1.3.4 Reinforcement Machine Learning:

Reinforcement learning works by programming an algorithm with a distinct goal and a prescribed set of rules for accomplishing that goal. Data scientists also program the algorithm to seek positive rewards -- which it receives when it performs an action that is beneficial toward the ultimate goal -- and avoid punishments -- which it receives when it performs an action that gets it farther away from its ultimate goal. Reinforcement learning is often used in areas such as:

Robotics: Robots can learn to perform tasks the physical world using this technique.
Video gameplay: Reinforcement learning has been used to teach bots to play a number of video games.
Resource management: Given finite resources and a defined goal, reinforcement learning can help enterprises plan out how to allocate resources.

1.4 Uses of Machine Learning:

Today, machine learning is used in a wide range of applications. Perhaps one of the most well-known examples of machine learning in action is the recommendation engine that powers Facebook's news feed.

Facebook uses machine learning to personalize how each member's feed is delivered. If a member frequently stops to read a particular group's posts, the recommendation engine will start to show more of that group's activity earlier in the feed.

Behind the scenes, the engine is attempting to reinforce known patterns in the member's online behavior. Should the member change patterns and fail to read posts from that group in the coming weeks, the news feed will adjust accordingly.

In addition to recommendation engines, other uses for machine learning include the following:

Customer relationship management. CRM software can use machine learning models to analyze email and prompt sales team members to respond to the most important messages first. More advanced systems can even recommend potentially effective responses.

Business intelligence. BI and analytics vendors use machine learning in their software to identify potentially important data points, patterns of data points and anomalies.

Human resource information systems. HRIS systems can use machine learning models to filter through applications and identify the best candidates for an open position.

Self-driving cars. Machine learning algorithms can even make it possible for a semi-autonomous car to recognize a partially visible object and alert the driver.

Virtual assistants. Smart assistants typically combine supervised and unsupervised machine learning models to interpret natural speech and supply context.

1.5 Pros and Cons of Machine Learning:

Machine learning has seen use cases ranging from predicting customer behavior to forming the operating system for self-driving cars.

When it comes to advantages, machine learning can help enterprises understand their customers at a deeper level. By collecting customer data and correlating it with behaviors over time, machine learning algorithms can learn associations and help teams tailor product development and marketing initiatives to customer demand.

Some companies use machine learning as a primary driver in their business models. Uber, for example, uses algorithms to match drivers with riders. Google uses machine learning to surface the ride advertisements in searches.

But machine learning comes with disadvantages. First and foremost, it can be expensive. Machine learning projects are typically driven by data scientists, who command high salaries. These projects also require software infrastructure that can be expensive.

There is also the problem of machine learning bias. Algorithms trained on data sets that exclude certain populations or contain errors can lead to inaccurate models of the world that, at best, fail and, at worst, are discriminatory. When an enterprise bases core business processes on biased models it can run into regulatory and reputational harm.

1.6 Choosing the Right Machine Learning Model:

The process of choosing the right machine learning model to solve a problem can be time consuming if not approached strategically.

Step 1: Align the problem with potential data inputs that should be considered for the solution. This step requires help from data scientists and experts who have a deep understanding of the problem.

Step 2: Collect data, format it and label the data if necessary. This step is typically led by data scientists, with help from data wranglers.

Step 3: Chose which algorithm(s) to use and test to see how well they perform. This step is usually carried out by data scientists.

Step 4: Continue to fine tune outputs until they reach an acceptable level of accuracy. This step is usually carried out by data scientists with feedback from experts who have a deep understanding of the problem.

Explaining how a specific ML model works can be challenging when the model is complex. There are some vertical industries where data scientists have to use simple machine learning models because it's important for the business to explain how every decision was made. This is especially true in industries with heavy compliance burdens such as banking and insurance.

Complex models can produce accurate predictions, but explaining to a lay person how an output was determined can be difficult.

1.7 Future & Evolution of Machine Learning:

While machine learning algorithms have been around for decades, they've attained new popularity as artificial intelligence has grown in prominence. Deep learning models, in particular, power today's most advanced AI applications.

Machine learning platforms are among enterprise technology's most competitive realms, with most major vendors, including Amazon, Google, Microsoft, IBM and others, racing to sign customers up for platform services that cover the spectrum of machine learning activities, including data collection, data preparation, data classification, model building, training and application deployment.

As machine learning continues to increase in importance to business operations and AI becomes more practical in enterprise settings, the machine learning platform wars will only intensify.

Continued research into deep learning and AI is increasingly focused on developing more general applications. Today's AI models require extensive training in order to produce an algorithm that is highly optimized to perform one task. But some researchers are exploring ways to make models more flexible and are seeking techniques that allow a machine to apply context learned from one task to future, different tasks.

1642 - Blaise Pascal invents a mechanical machine that can add, subtract, multiply and divide.

1679 - Gottfried Wilhelm Leibniz devises the system of binary code.

1834 - Charles Babbage conceives the idea for a general all-purpose device that could be programmed with punched cards.

1842 - Ada Lovelace describes a sequence of operations for solving mathematical problems using Charles Babbage's theoretical punch-card machine and becomes the first programmer.

1847 - George Boole creates Boolean logic, a form of algebra in which all values can be reduced to the binary values of true or false.

1936 - English logician and cryptanalyst Alan Turing proposes a universal machine that could decipher and execute a set of instructions. His published proof is considered the basis of computer science.

1952 - Arthur Samuel creates a program to help an IBM computer get better at checkers the more it plays.

1959 - MADALINE becomes the first artificial neural network applied to a real-world problem: removing echoes from phone lines.

1985 - Terry Sejnowski's and Charles Rosenberg's artificial neural network taught itself how to correctly pronounce 20,000 words in one week.

1997 - IBM's Deep Blue beat chess grandmaster Garry Kasparov.

1999 - A CAD prototype intelligent workstation reviewed 22,000 mammograms and detected cancer 52% more accurately than radiologists did.

2006 - Computer scientist Geoffrey Hinton invents the term deep learning to describe neural net research.

2012 - An unsupervised neural network created by Google learned to recognize cats in YouTube videos with 74.8% accuracy.

2014 - A chatbot passes the Turing Test by convincing 33% of human judges that it was a Ukrainian teen named Eugene Goostman.

2014 - Google's AlphaGo defeats the human champion in Go, the most difficult board game in the world.

2016 - LipNet, DeepMind's artificial intelligence system, identifies lip-read words in video with an accuracy of 93.4%.

2019 - Amazon controls 70% of the market share for virtual assistants in the U.S.

1.8 House Price Prediction:

Investment is a business activity that most people are interested in this globalization era. There are several objects that are often used for investment, for example, gold, stocks and property. In particular, property investment has increased significantly since 2011, both on demand and property selling. One of the increasing of property demand is because of high population in India. Indian Central Bureau of Statistics states that in North India, 50% of the population is classified as a young population who have age approximately at 30 years old. The result of this census indicates that the younger generation will need a house or buy a house in the future. Based on preliminary research conducted, there are two standards of house price which are valid in buying and selling transaction of a house that is house price based on the developer (market selling price) and price based on Value of Selling Tax Object (NJOP). According to Lim, et al the fundamental problem for a developer is to determine the selling price of a house. In determining the price of home, the developer must calculate carefully and determine the appropriate method because property prices always increase continuously and almost never fall in the long term or short. There are several approaches that can be used to determine the price of the house, one of them is the prediction analysis. The first approach is a quantitative prediction. A quantitative approach is an approach that utilizes time-series data. The time-series approach is to look for the relationship between current prices and prevailing prices. The second approach is to use linear regression based on hedonic pricing. In linear regression, determining coefficients generally using the least square method, but it takes a long time to get the best formula. Particle swarm optimization (PSO) is proposed to find the coefficients aimed at obtaining optimal results. Some previous researches such as Marini and Walzack, show that PSO gets better results than other hybrid methods. There are several advantages of PSO, in the small search space PSO can do better solution search. Although the PSO global search is less than optimal, but on the optimization problem the value of the variable on the regression equation can find a maximum solution using PSO. This research aims to create a house price prediction model using regression and PSO to obtain optimal prediction results. PSO is used for selection of affect variables in house prediction, regression is used to determine the optimal coefficient in prediction. In this study, researchers wanted to know the performance of the developed model in time series data. Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location. This research is focused in Bangalore City, because Bangalore being the Silicon Valley of India attracts a lot of target youth thereby increasing the population who are in demand of property.

1.9 Factors Affecting House Prices:

1. Growth in the Economy:

Housing demand depends on revenue. With higher economic growth and growing wages, people can spend more on housing, improving application and boosting prices. In reality, housing demand is often seen as elastic in terms of income, leading to an increase in revenues for households. In a recession, reduced sales will also stop people from buying, and people who are losing their jobs will fall behind their mortgage payments and end up in their homes repossessed.

2. Unemployment:

The second important point that comes under the economic factors affecting housing market is related to economic growth. Very few people will have possible to afford a house as unemployment rises. But even fear of unemployment can stop people from entering the real estate market.

3. Interest Rates:

Interest rates influence the monthly payment value for mortgages. A high-interest rate era would increase mortgage costs and reduce the demand for a house to be purchased. In contrast to renting, high-interest rates make rental attractive. Homeowners with high adjustable mortgage rates have a more significant effect.

4. Customer Trust:

Confidence is an essential part when people are to take the risk of taking out a mortgage. Mainly house market expectations are significant. When people fear house prices will decrease, people will postpone purchasing.

5. Mortgage Availability:

Most banks are keen to lend mortgages during the boom years of 1996-2006. It enabled people to borrow large amounts of revenue (for example, five times the income). Additionally, minimal deposits that are 100% of mortgages are provided by banks. The flexibility of hypothecating meant that the housing demand grew as more people could buy now. Yet banks and construction companies have had trouble raising money to finance the financial markets since the 2007 credit crunch. Therefore, their borrowing conditions for a larger house purchase deposit have been improved. The supply of loans has been limited, and demand fell. Get some property documents required for home loan.

6. Offering:

A supply shortage drives prices up. Over-supply could lead to a fall in prices.

7. Effectiveness/House Income Rates:

The price-to-earnings ratio impacts demand. For house prices rising about wages, you would expect fewer people to afford. For example, the house price-to-income ratio increased to 5 in 2007. Homes were relatively expensive at this level, and we saw a correction with the drop in house prices.

8. Home Sales Economy Mirror

Home sales are usually directly related to the stability of an economy and economic growth and decrease. When the economic growth slowdown, cash supplies get limited thoroughly. Because capital is difficult to buy, the housing market will be less available to home buyers. Housing inventories increase and take longer to sell, as stringent credit standards make fewer buyers available. A higher consumer supply combined with lower demand usually leads to lower prices. Also, read about the real estate trends in Kerala.

9. Sales of Household Cash Supply

The supply of money is vital to its overall health, and in general, for the sustainability of the housing market. The availability of funds in an economy that too when money is hard to receive, sales of home will dry up. Once cash is also easy to buy, too many investors enter the housing market and price rises for some time until the inevitable market correction or even crash happens. Home buildings and home sales should ideally align with economic activities, but this is not the case sometimes. Also read about the real estate portals.

10. Closing Reflects a Market Crash

Across different economies, housing markets operate differently. The housing market is usually healthy during a strong economy. Then fewer people buy as interest rates rise. There may be an increase in foreclosures when people default on their loans, which usually happens with adjustable mortgages when the prices rise.

11. Home Sales Financial Slowdown

After an economy slows, the housing markets can be impacted. Slowdowns in the economy impact housing markets as housing-related activity decreases and overall economic demand slows. When economic reforms start and housing prices reflect the willingness of consumers to pay, the economic cycle breaks down.

Other Areas and Real Estate Development

It can be said that the Real estate market in India will grow based on our speculations. So many factors motivate the buyer of India. The first time, home-seekers will be the general and once adequately regulated, we can certainly take advantage of the factors and participate in India’s economy. It is also true to say that it will only allow other sectors to expand if it is appropriately regulated.

Retail firms are also profoundly affected by the increased demand for property markets, such as housing, cement, finance companies. Also, there is a significant difference in pricing between rural India and Urban India. The lack of jobs in rural areas was also of considerable importance. Increasingly more people are moving to cities, engaging in increasing housing requirements and other real estate requirements.

The government of India should invest heavily in creating rural employability. Attractive housing schemes can also help generate higher demands. Low bank interest rates and adverse economic conditions can to deter some prospective buyers. You can also increase the supply to reduce the cost of smarter games. Checkout some of the best ways to increase home value

Demographic Factors

Apart from the economic factors affecting housing market, there are some demographic factors too. The number of households in India. has grown, and there are also more individuals living alone. Some of the reasons for increased house demand in India are:

1. Life expectancy increased for the elderly
2. Divorce rates rise
3. As per now, children leave their homes in early years itself
4. Increase in Marriages
5. They are dreaming to be more Independent.

Factors that have Long-Term Effects on Housing Supply:

- Disposal of permission for planning
- Opportunity costs for construction companies as other projects have better returns
- New houses are not ideal for staying in
- Building new homes efficiency
- A rise in construction costs would move demand to the left

Unusual Factors Affecting the Demand for Property in India, including:

- The attractive dealer’s supply rate
- Increased profits
- Fast-moving nuclear families
- Mortgage availability
- Confidences for customers
- Lower interest rates
- Increased cost-effectiveness

Economic Factors Influencing the Quality of a License Test

As we discussed the economic factors affecting the housing market, now see what all comes during the license test

1. Advance

All property values are generated by predicting the potential benefits of the land. The value of land is now getting increased day by day. If you buy a home now, after many years you will have a good value to that home. The physical, political, economic and social changes all have an impact on the value of the land. Environmental changes such as climate or pollution may include physical factors. Financial problems can alter job rates in a given area. Social factors such as baby boomers' aging were problems. Any or all of these and others may affect property values.

2. Balance

A balance can be found in any given area between land value and building value. When the balance is retained, gross property values and constructor income are maximized in new homes. In most cases, for instance, a house that costs ₹ 71,30,650.00 in a settled land costing ₹ 3,56,53,750.00 will never be built. The balance must generally be similar to the balance in the area.

3. Compliance

Value is developed and maintained in similar situations. You don't want to build an office building across the street from your house because you live in a neighbourhood that includes single-family homes. Your house's value would likely be affected negatively by this inconsistent land use.

4. Change

All of which affect property value are physical, political, financial, and social changes. Environmental changes such as weather or pollution can involve physical factors. Economic problems can change job rates in a region. Social factors such as baby boomers aging were problems. Everyone or everyone else can have an impact on property values.

5. Competition

Competition shows that the supply side is trying to meet the demand side on the real estate until demand is met. A developer may see the need in a specific location for a new office building.

6. Extinct

Real estate is affected by all that happens around it as it stays in a fixed location. The gas station on the street, school quality, factory closing in town, mortgage interest rates, etc. has an impact on the value of homes.

7. Return Increases and Decreases

Increasing and decreasing returns are associated with the addition of improvements in a property. Increasing returns arise when an upgrade gives the property more value than its price. You get more than one dollar back from spent cent. Returns decrease when the cost is increasing by an increase.

8. Surplus Productivity

The main difference between cost and sale price is excess productivity after the contractor assembles property, workforce, resources, and teamwork required to create and then sell the house. This term is used by economists to mean money.

CHAPTER 2: AIM AND SCOPE OF PROJECT

2.1 AIM

The aim of this project is to implement the machine learning algorithm with the best accuracy to predict house prices based on features such as super built-up area, number of bedrooms, bathrooms as well as area.

2.2 SCOPE OF PROJECT

Real Estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. According to Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic index suggests that every single organization in today's real estate business is operating fruitfully to achieve a competitive edge over alternative competitor. There is a need to simplify the process for a normal human being while providing the best results proposed to use machine learning and artificial intelligence techniques to develop an algorithm that can predict housing prices based on certain input features. The business application of this algorithm is that classified websites can directly use this algorithm to predict prices of new properties that are going to be listed by taking some input variables and predicting the correct and justified price i.e., avoid taking price inputs from customers and thus not letting any error creeping in the system used Google Colab/Jupiter IDE. Jupiter IDE is an open-source web app that helps us to share as well create documents that have LiveCode, visualizations, equations, and text that narrates. It contains tools for data cleaning, data transformation, simulation of numeric values, modeling using statistics, visualization of data, and machine learning tools designed a system that will help people to know close to the precise price of real estate. User can give their requirements according to which they will get the prices of the desired houses User can also get the sample plan of the house to get a reference for houses. In Housing value of the Boston suburb is analyzed and forecast by SVM, LSSVM, and PLS methods and the corresponding charactristics. After getting rid of the missing samples from the original data set, 400 samples are treated as training data and 52 samples are treated as test data. Housing value of the training data. As per the findings, the best accuracy was provided by the Random Forest Regressor followed by the Decision Tree Regressor. A similar result is generated by the Ridge and Linear Regression with a very slight reduction in Lasso. Across all groups of feature selections, there is no extreme difference between all regardless of strong or weak groups. It gives a good sign that the buying prices can be solely used for predicting the selling prices without considering other features to disseminate model over-fitting. Additionally, a reduction in accuracy is apparent in the very weak features group. The same pattern of results is visible on the Root Square Mean Error (RMSE) for all feature selections. observed that their data set took more than one day to prepare. As opposed to performing the computations sequentially, we might utilize various processors and parallel the computations involved, which might possibly decrease the preparation time Furthermore prediction period. Include All the more functionalities under the model, we can give choices for clients with select a district alternately locale should produce those high-temperature maps, as opposed to entering in the list. used a data set of 100 houses with several parameters. We have used 50 percent of the data set to train the machine and 50 percent to test the machine. The results are truly accurate. And we have tested it with different parameters also. Not using PSO makes it easier to train machines with complex problems and hence regression is used. Experimented with the most fundamental machine learning algorithms like decision tree classifier, decision tree regression, and multiple linear regression. Work is implemented using the Scikit-Learn machine learning tool. This work helps the users to predict the availability of houses in the city and also to predict the prices of the houses used machine learning algorithms to predict house prices. We have mentioned the step-by-step procedure to analyze the dataset. These feature sets were then given as an input to four algorithms and a CSV file was generated consisting of predicted house prices. expressed that there is a need to use a mix of these models a linear model gives a high bias (underfit) whereas a high model complexity-based model gives a high variance (overfit). The outcome of this study can be used in the annual revision of the guideline value of land which may add more revenue to the State Government while this transaction is made, concludes that by conducting this experiment with various machine learning algorithms it’s been clear that random forest and gradient boosted trees are performing better with more accuracy percentage and with fewer error values. when this experiment is compared with the label and to the result achieved these algorithms predict well.

There is always a lot of enthusiasm while purchasing a residential property, especially in case of first-time buyers. An investment in a house is the biggest financial commitment for most individuals. Earlier, most people used to buy a residential property only after earning and accumulating funds for many years. However, in the recent times, aggressive financing by banks, increased disposable incomes and tax benefits have made it easy for much younger people to purchase property.

Although the availability of housing loan schemes has made it easy for many to buy property, it still requires a lot of understanding of various financial aspects, and accumulation of funds, to acquire a property and avoid getting into any sort of financial issues.

It's always advisable to prepare a budget to buy the house. A budget helps in segmenting the available options and in zeroing down on to the right property. As a thumb rule, you should consider purchasing a property that costs around six times your stable annual household income.

It is also important to consider the long-term utility of the property and balance it with financial capability. One should keep in mind that an investment in property cannot be frequently altered.

Easy access to housing loan schemes has made it easy for many to buy property. However, you still need to arrange a sizeable amount to finance the upfront advance payment, margin money payment, registration charges and brokerage as applicable, and furnishing costs. All these costs come up to 25-35 percent of the total property value. Therefore, it is advisable to start planning and arranging the required funds at least six months to one year in advance. These are some ways to generate the margin money needed: Liquidate the accumulated small savings Liquidate a part of the medium and long-term savings Sell some long-term assets such as gold smaller property etc. Withdraw or borrow money against long-term savings such as Provident Fund Prepay small and short-term loans such as car loan, soft loan, consumer loan, credit card loan etc.

It is very important to choose the right home loan scheme. There are various factors such as the bank, tenure, amount, floating or fixed rate etc that need to be thought of while choosing the appropriate home loan scheme. It's advisable to invest by comparing notes with those who have already taken a home loan in addition to exploring the various ongoing schemes offered by banks in the current market conditions.

CHAPTER 3: EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHM USED

3.1 Platform- JUPYTER NOTEBOOK

Project Jupyter like the planets a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando Pérez and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R, and also a homage to Galileo's notebooks recording the discovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS.

In 2014, Fernando Pérez announced a spin-off project from IPython called Project Jupyter. IPython continues to exist as a Python shell and a kernel for Jupyter, while the notebook and other language-agnostic parts of IPython moved under the Jupyter name. Jupyter is language agnostic and it supports execution environments (aka kernels) in several dozen languages among which are Julia, R, Haskell, Ruby, and of course Python (via the IPython kernel).

In 2015, GitHub and the Jupyter Project announced native rendering of Jupyter notebooks file format (.ipynb files) on the GitHub platform.

Project Jupyter's operating philosophy is to support interactive data science and scientific computing across all programming languages via the development of open-source software. According to the Project Jupyter website, "Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license".

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating notebook documents.

A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension.

Jupyter notebooks are built upon a number of popular open-source libraries:

- IPython
- ZeroMQ
- Tornado
- jQuery
- Bootstrap (front-end framework)
- MathJax

Jupyter Notebook can connect to many kernels to allow programming in different languages. A Jupyter kernel is a program responsible for handling various types of requests (code execution, code completions, inspection), and providing a reply. Kernels talk to the other components of Jupyter using ZeroMQ, and thus can be on the same or remote machines. Unlike many other Notebook-like interfaces, in Jupyter, kernels are not aware that they are attached to a specific document, and can be connected to many clients at once. Usually, kernels allow execution of only a single language, but there are a couple of exceptions. By default Jupyter Notebook ships with the IPython kernel. As of the 2.3 release (October 2014), there are 49 Jupyter-compatible kernels for many programming languages, including Python, R, Julia and Haskell.

A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell. To simplify visualisation of Jupyter notebook documents on the web, the nbconvert libraryis provided as a service through NbViewer which can take a URL to any publicly available notebook document, convert it to HTML on the fly and display it to the user.

The notebook interface was added to IPython in the 0.12 release (December 2011), renamed to Jupyter notebook in 2015 (IPython 4.0 is Jupyter 1.0). Jupyter Notebook is similar to the notebook interface of other programs such as Maple, Mathematica, and SageMath, a computational interface style that originated with Mathematica in the 1980s. Jupyter interest overtook the popularity of the Mathematica notebook interface in early 2018.

JupyterLab is a newer user interface for Project Jupyter. It offers the building blocks of the classic Jupyter Notebook (notebook, terminal, text editor, file browser, rich outputs, etc.) in a flexible user interface. The first stable release was announced on February 20, 2018.

- JupyterHub is a multi-user server for Jupyter Notebooks. It is designed to support many users by spawning, managing, and proxying many singular Jupyter Notebook servers.[citation needed] While JupyterHub requires managing servers, third-party services like Jupyo[22] provide an alternative to JupyterHub by hosting and managing multi-user Jupyter notebooks in the cloud.
- Jupyter Book is an open-source project for building books and documents from computational material. It allows the user to construct the content in a mixture of Markdown, an extended version of Markdown called MyST, Maths & Equations using MathJax, Jupyter Notebooks, reStructuredText, the output of running Jupyter Notebooks at build time. Multiple output formats can be produced (currently single files, multipage HTML web pages and PDF files).
- nbgrader is a tool for creating and grading (marking) assignments in Jupyter notebooks. It allows the instructor to create assignments that include coding exercises in Python or any other supported kernel, and text responses. The submitted assignments can be automatically marked, manually scored or a mixture of both.

The Jupyter Notebook has become a popular user interface for cloud computing, and major cloud providers have adopted the Jupyter Notebook or derivative tools as a frontend interface for cloud users.

Examples

Amazon's SageMaker Notebooks, Google's Colaboratory and Microsoft's Azure Notebook.

Google Colaboratory (also known as Colab) is a free Jupyter notebook environment that runs in the cloud and stores its notebooks on Google Drive. Colab was originally an internal Google project; an attempt was made to open source all the code and work more directly upstream, leading to the development of the "Open in Colab" Google Chrome extension, but this eventually ended, and Colab development continued internally. As of October 2019, the Colaboratory UI only allows for the creation of notebooks with Python 2 and Python 3 kernels; however, an existing notebook whose kernelspec is IR or Swift will also work, since both R and Swift are installed in the container. Julia language can also work on Colab (with e.g., Python and GPUs) Google's tensor processing units also work with Julia on Colab.

3.2 Machine Learning Algorithms

1. Linear Regression

To understand the working functionality of this algorithm, imagine how you would arrange random logs of wood in increasing order of their weight. There is a catch; however – you cannot weigh each log. You have to guess its weight just by looking at the height and girth of the log (visual analysis) and arrange them using a combination of these visible parameters. This is what linear regression in machine learning is like.

In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and represented by a linear equation Y= a *X + b.

In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

The coefficients a & b are derived by minimizing the sum of the squared difference of distance between data points and the regression line.

2. Logistic Regression

Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression.

These methods listed below are often used to help improve logistic regression models:

- include interaction terms
- eliminate features
- regularize techniques
- use a non-linear model

3. Decision Tree

Decision Tree algorithm in machine learning is one of the most popular algorithm in use today; this is a supervised learning algorithm that is used for classifying problems. It works well classifying for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

4. SVM (Support Vector Machine) Algorithm

SVM algorithm is a method of classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data. Lines called classifiers can be used to split the data and plot them on a graph.

5. Naive Bayes Algorithm

A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Even if these features are related to each other, a Naive Bayes classifier would consider all of these properties independently when calculating the probability of a particular outcome.

A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.

6. KNN (K- Nearest Neighbors) Algorithm

This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement.

KNN can be easily understood by comparing it to real life. For example, if you want information about a person, it makes sense to talk to his or her friends and colleagues!

Things to consider before selecting K Nearest Neighbours Algorithm:

- KNN is computationally expensive
- Variables should be normalized, or else higher range variables can bias the algorithm
- Data still needs to be pre-processed.

7. K-Means

It is an unsupervised learning algorithm that solves clustering problems. Data sets are classified into a particular number of clusters (let's call that number K) in such a way that all the data points within a cluster are homogenous and heterogeneous from the data in other clusters.

How K-means forms clusters:

- The K-means algorithm picks k number of points, called centroids, for each cluster.
- Each data point forms a cluster with the closest centroids, i.e., K clusters.
- It now creates new centroids based on the existing cluster members.
- With these new centroids, the closest distance for each data point is determined. This process is repeated until the centroids do not change.

8. Random Forest Algorithm

A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is planted & grown as follows:

- If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M, and the best split on this m is used to split the node. The value of m is held constant during this process.
- Each tree is grown to the most substantial extent possible. There is no pruning.

9. Dimensionality Reduction Algorithms

In today's world, vast amounts of data are being stored and analyzed by corporates, government agencies, and research organizations. As a data scientist, you know that this raw data contains a lot of information - the challenge is in identifying significant patterns and variables.

Dimensionality reduction algorithms like Decision Tree, Factor Analysis, Missing Value Ratio, and Random Forest can help you find relevant details.

These are boosting algorithms used when massive loads of data have to be handled to make predictions with high accuracy. Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness.

In short, it combines multiple weak or average predictors to build a strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix. These are the most preferred machine learning algorithms today. Use them, along with Python and R Codes, to achieve accurate outcomes.

10. Gradient Boosting Algorithm and AdaBoosting Algorithm

These are boosting algorithms used when massive loads of data have to be handled to make predictions with high accuracy. Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness.

In short, it combines multiple weak or average predictors to build a strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix. These are the most preferred machine learning algorithms today. Use them, along with Python and R Codes, to achieve accurate outcomes.

3.3 HARDWARE/ SOFTWARE REQUIRED

HARDWARE:

- Processor – Intel Xeon E2630 v4 – 10 core processor, 2.2 GHz with Turboboost upto 3.1 GHz. 25 MB Cache
- Motherboard – ASRock EPC612D8A
- RAM – 128 GB DDR4 2133 MHz
- 2 TB Hard Disk (7200 RPM) + 512 GB SSD
- GPU – NVidia TitanX Pascal (12 GB VRAM)
- Intel Heatsink to keep temperature under control
- Storm Trooper Cabinet

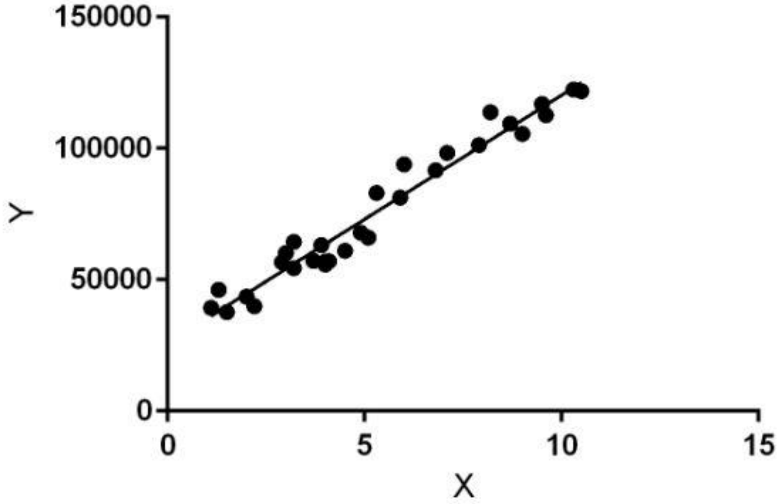
SOFTWARE:

- Anaconda Navigator
- Jupyter Notebook

3.4 IMPLEMENTATION

In this project we have used 3 machine leaning models and, then implemented the one with best accuracy.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression:

$$y = \theta_1 + \theta_2 . x$$

While training the model we are given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.

θ_1: intercept

θ_2: coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Lasso Regression is also another linear model derived from Linear Regression which shares the same hypothetical function for prediction. The cost function of Linear Regression is represented by J.

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.8}$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \tag{6.9}$$

Here, m is the total number of training examples in the dataset.

$h(x^{(i)})$ represents the hypothetical function for prediction.

$y^{(i)}$ represents the value of target variable for ith training example.

Linear Regression model considers all the features equally relevant for prediction. When there are many features in the dataset and even some of them are not relevant for the predictive model. This makes the model more complex with a too inaccurate prediction on the test set (or overfitting). Such a model with high variance does not generalize on the new data. So, Lasso Regression comes for the rescue. It introduced an L1 penalty (or equal to the absolute value of the magnitude of weights) in the cost function of Linear Regression. The modified cost function for Lasso Regression is given below.asso Regression performs both, variable selection and regularization too.

Decision Tree Classification Algorithm

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- The decisions or the test are performed on the basis of features of the given dataset.

- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

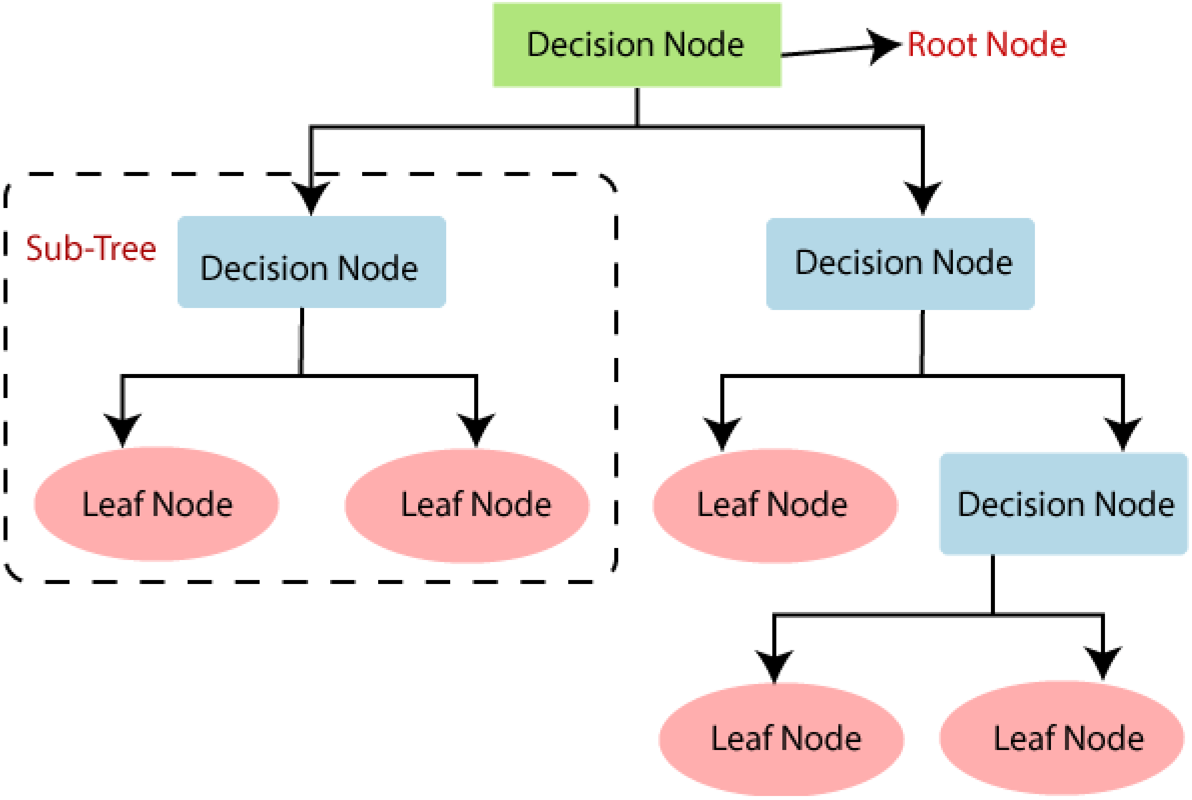
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

- Below diagram explains the general structure of a decision tree:

Note: A decision tree can contain categorical data (YES/NO) as well as numeric data.

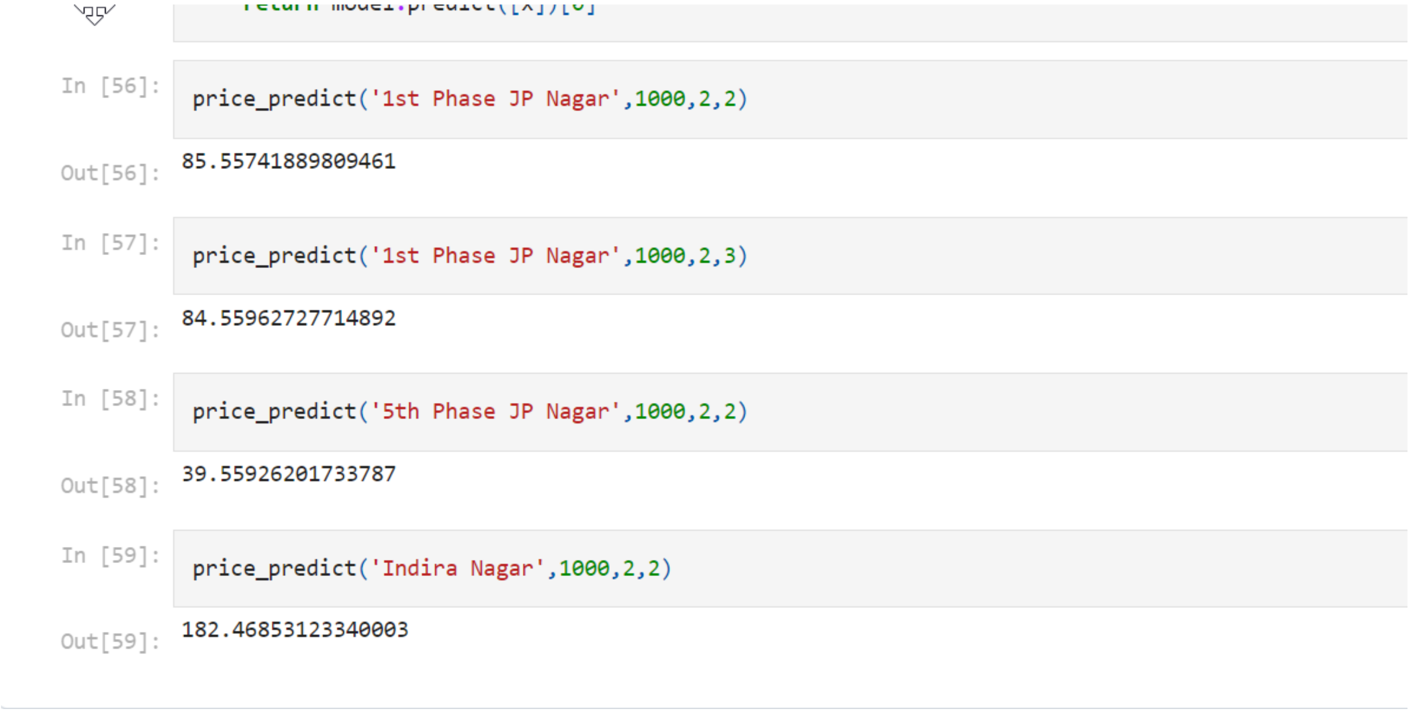


CHAPTER 4: RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

Link to the Project:

https://github.com/pankhuri2910/House-price-prediction

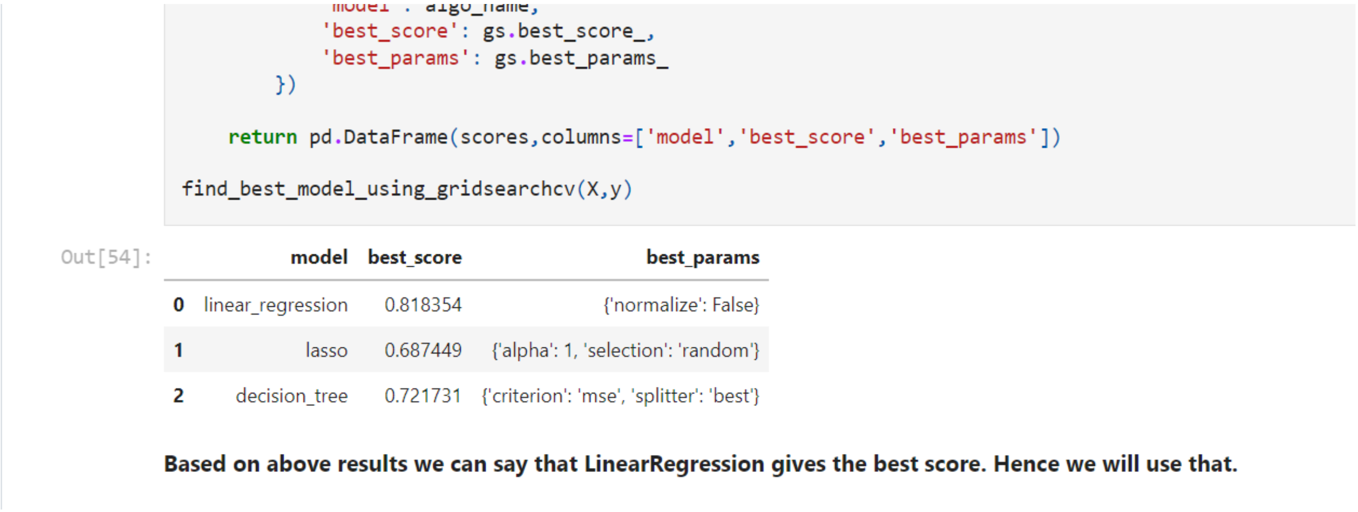
4.1 RESULT:



```
        return model.predict([x])[0]

In [56]: price_predict('1st Phase JP Nagar',1000,2,2)

Out[56]: 85.55741889809461

In [57]: price_predict('1st Phase JP Nagar',1000,2,3)

Out[57]: 84.55962727714892

In [58]: price_predict('5th Phase JP Nagar',1000,2,2)

Out[58]: 39.55926201733787

In [59]: price_predict('Indira Nagar',1000,2,2)

Out[59]: 182.46853123340003
```

4.2 Performance Analysis:

GridsearchCV was used to determinethe best approach towards the problem and following were the result which concluded in us using linear regression. The buzzwords in the field of Data Science such as Machine Learning, Artificial Intelligence and Deep Learning are appearing at the maximum number of places on the Internet in recent time. Everyone wants to try out different models of Machine Learning and Deep Learning and achieve the best results possible. There are some computational limits for some of the models. To get the best model in Machine Learning, there is something known as Hyperparameter Tuning.
Hyperparameter Tuning is basically getting the best set of parameters selected for a model. There are 2 common approaches to this: GridSearchCV and RandomizedSearchCV.
GridSearchCV is basically considering all the combinations of the candidates in finding the best parameters. This would in turn take a very long time when there are a greater number of parameter and their values to tune. There is an approach by which we can fasten this process. This is the main thing that occupies most of the time in Machine Learning. Before diving into the approach part let us skim through the basics of GridSearchCV and parallel computing concepts. GridSearchCV is a technique to search through the best parameter values from the given set of the grid of parameters. It is basically a cross-validation method. the model and the parameters are required to be fed in. Best parameter values are extracted and then the predictions are made.



```
            'model': algo_name,
            'best_score': gs.best_score_,
            'best_params': gs.best_params_
        })

    return pd.DataFrame(scores,columns=['model','best_score','best_params'])

find_best_model_using_gridsearchcv(X,y)
```

Out[54]:

	model	best_score	best_params
0	linear_regression	0.818354	{'normalize': False}
1	lasso	0.687449	{'alpha': 1, 'selection': 'random'}
2	decision_tree	0.721731	{'criterion': 'mse', 'splitter': 'best'}

Based on above results we can say that LinearRegression gives the best score. Hence we will use that.

CHAPTER 5: SUMMARY AND CONCLUSION

Machine learning is a branch of AI. Other tools for reaching AI include rule-based engines, evolutionary algorithms, and Bayesian statistics. While many early AI programs, like IBM's Deep Blue, which defeated Garry Kasparov in chess in 1997, were rule-based and dependent on human programming, machine learning is a tool through which computers have the ability to teach themselves, and set their own rules. In 2016, Google's DeepMind beat the world champion in Go by using machine learning–training itself on a large data set of expert moves.

There are several kinds of machine learning:

- In supervised learning, the "trainer" will present the computer with certain rules that connect an input (an object's feature, like "smooth," for example) with an output (the object itself, like a marble).
- In unsupervised learning, the computer is given inputs and is left alone to discover patterns.
- In reinforcement learning, a computer system receives input continuously (in the case of a driverless car receiving input about the road, for example) and constantly is improving.

A massive amount of data is required to train algorithms for machine learning. First, the "training data" must be labeled (e.g., a GPS location attached to a photo). Then it is "classified." This happens when features of the object in question are labeled and put into the system with a set of rules that lead to a prediction. For example, "red" and "round" are inputs into the system that leads to the output: Apple. Similarly, a learning algorithm could be left alone to create its own rules that will apply when it is provided with a large set of the object–like a group of apples, and the machine figures out that they have properties like "round" and "red" in common.

Many cases of machine learning involve "deep learning," a subset of ML that uses algorithms that are layered, and form a network to process information and reach predictions. What distinguishes deep learning is the fact that the system can learn on its own, without human training.

When did machine learning become popular?

Machine learning was popular in the 1990s, and has seen a recent resurgence. Here are some timeline highlights.

- 2011: Google Brain was created, which was a deep neural network that could identify and categorize objects.

- 2014: Facebook's DeepFace algorithm was introduced. The algorithm could recognize people from a set of photos.
- 2015: Amazon launched its machine learning platform, and Microsoft offered a Distributed Machine Learning Toolkit.
- 2016: Google's DeepMind program "AlphaGo" beat the world champion, Lee Sedol, at the complex game of Go.
- 2017: Google announced that its machine learning tools can recognize objects in photos and understand speech better than humans.
- 2018: Alphabet subsidiary Waymo launched the ML-powered self-driving ride hailing service in Phoenix, AZ.
- 2020: Machine learning algorithms are brought into play against the COVID-19 pandemic, helping to speed vaccine research and improve the ability to track the virus' spread.

Why does machine learning matter?

Aside from the tremendous power machine learning has to beat humans at games like Jeopardy, chess, and Go, machine learning has many practical applications. Machine learning tools are used to translate messages on Facebook, spot faces from photos, and find locations around the globe that have certain geographic features. IBM Watson is used to help doctors make cancer treatment decisions. Driverless cars use machine learning to gather information from the environment. Machine learning is also central to fraud prevention. Unsupervised machine learning, combined with human experts, has been proven to be very accurate in detecting cybersecurity threats, for example.

While there are many potential benefits of AI, there are also concerns about its usage. Many worry that AI (like automation) will put human jobs at risk. And whether or not AI replaces humans at work, it will definitely shift the kinds of jobs that are necessary. Machine learning's requirement for labeled data, for example, has meant a huge need for humans to manually do the labeling.

As machine learning and AI in the workplace have evolved, many of its applications have centered on assisting workers rather than replacing them outright. This was especially true during the COVID-19 pandemic, which forced many companies to send large portions of their workforce home to work remotely, leading to AI bots and machine learning supplementing humans to take care of mundane tasks.

There are several institutions dedicated to exploring the impact of artificial intelligence. Here are a few (culled from our Twitter list of AI insiders).

- The Future of Life Institute brings together some of the greatest minds–from the co-founder of Skype to professors at Harvard and MIT–to explore some of the big questions about our future with machines. This Cambridge-based institute also has a stellar lineup on its scientific advisory board, from Nick Bostrom to Elon Musk to Morgan Freeman.
- The Future of Humanity Institute at Oxford is one of the premier sites for cutting-edge academic research. The FHI Twitter feed is a wonderful place for content on the latest in AI, and the many retweets by the account are also useful in finding other Twitter users who are working on the latest in artificial intelligence.
- The Machine Intelligence Research Institute at Berkeley is an excellent resource for the latest academic work in artificial intelligence. MIRI exists, according to Twitter, not only to investigate AI, but also to "ensure that the creation of smarter-than-human intelligence has a positive impact."

Additional resources

- IBM Watson CTO: The 3 ethical principles AI needs to embrace (TechRepublic)
- Forrester: Automation could lead to another jobless recovery (TechRepublic)
- Machine learning helps science tackle Alzheimer's (CBS News)

Which industries use machine learning?

Just about any organization that wants to capitalize on its data to gain insights, improve relationships with customers, increase sales, or be competitive at a specific task will rely on machine learning. It has applications in government, business, education–virtually anyone who wants to make predictions, and has a large enough data set, can use machine learning to achieve their goals.

Along with analytics, machine learning can be used to supplement human workers by taking on mundane tasks and freeing them to do more meaningful, innovative, and productive work. Like with analytics, and business that has employees dealing with repetitive, high-volume tasks can benefit from machine learning.

How do businesses use machine learning?

2017 was a huge year for growth in the capabilities of machine learning, and 2018 set the stage for explosive growth that, by early 2020, found that 85% of businesses were using some form of AI in their deployed applications.

One of the things that may be holding that growth back, Deloitte said, is confusion–just what is machine learning capable of doing for businesses?

There are numerous examples of how businesses are leveraging machine learning, and all of it breaks down to the same basic thing: Processing massive amounts of data to draw conclusions much faster than a team of data scientists ever could.

Some examples of business uses of machine learning include:

- Alphabet-owned security firm Chronicle is using machine learning to identify cyberthreats and minimize the damage they can cause.
- Airbus Defense &amp; Space is using ML-based image recognition technology to decrease the error rate of cloud recognition in satellite images.
- Global Fishing Watch is fighting overfishing by monitoring the GPS coordinates of fishing vessels, which has enabled them to monitor the whole ocean at once.
- Insurance firm AXA raised accident prediction accuracy by 78% by using machine learning to build accurate driver risk profiles.
- Japanese food safety company Kewpie has automated detection of defective potato cubes so that workers don't have to spend hours watching for them.
- Yelp uses deep learning to classify photos people take of businesses by certain tags.
- MIT's OptiVax can develop and test peptide vaccines for COVID-19 and other diseases in a completely virtual environment with variables including geographic coverage, population data, and more.

Any business that deals with big data analysis can use machine learning technology to speed up the process and put humans to better use, and the particulars can vary greatly from industry to industry.

AI applications don't come first–they're tools used to solve business problems, and should be seen as such. Finding the proper application for machine learning technology involves asking the right questions, or being faced with a massive wall of data that would be impossible for a human to process.

What are the security and ethical concerns about machine learning?

There are a number of concerns about using machine learning and AI, including the security of cloud-hosted data and the ethical considerations of self-driving cars.

From a security perspective, there are always concerns about the theft of large amounts of data, but security fears go beyond how to lock down data repositories.

Security professionals are nearly universally concerned about the potential of AI to bypass antimalware software and other security measures, and they're right to be worried: Artificial intelligence software has been developed that can modify malware to bypass AI-powered antimalware platforms.

Several tech leaders, like Elon Musk, Stephen Hawking, and Bill Gates, have expressed worries about how AI may be misused, and the importance of creating ethical AI. Evidenced by the disaster of Microsoft's racist chatbot, Tay, AI can go wrong if left unmonitored.

Ethical concerns abound in the machine learning world as well; one example is a self-driving vehicle adaptation of the trolley problem thought experiment. In short, when a self-driving vehicle is presented with a choice between killing its occupants or a pedestrian, which is the right choice to make? There's no clear answer with philosophical problems like this one–no matter how the machine is programmed, it has to make a moral judgement about the value of human lives.

Deep fake videos, which realistically replace one person's face and/or voice with someone else's based on photos and other recordings, have the potential to upset elections, insert unwilling people into pornography, and otherwise insert individuals into situtations they aren't okay with. The far-reaching effects of this machine learning-powered tool could be devastating.

Along with whether giving learning machines the ability to make moral decisions is correct, or whether access to certain ML tools is socially dangerous, there are issues of the other major human cost likely to come with machine learning: Job loss.

If the AI revolution is truly the next major shift in the world, there are a lot of jobs that will cease to exist, and it isn't necessarily the ones you'd think. While many low-skilled jobs are definitely at risk of being eliminated, so are jobs that require a high degree of training but are based on simple concepts like pattern recognition.

Radiologists, pathologists, oncologists, and other similar professions are all based on finding and diagnosing irregularities, something that machine learning is particularly suited to do.

There's also the ethical concern of barrier to entry–while machine learning software itself isn't expensive, only the largest enterprises in the world have the vast stores of data necessary to properly train learning machines to provide reliable results.

As time goes on, some experts predict that it's going to become more difficult for smaller firms to make an impact, making machine learning primarily a game for the largest, wealthiest companies.

What machine learning tools are available?

There are many online resources about machine learning. To get an overview of how to create a machine learning system, check out this series of YouTube videos by Google Developer. There are also classes on machine learning from Coursera and many other institutions.

And to integrate machine learning into your organization, you can use resources like Microsoft's Azure, Google Cloud Machine Learning, Amazon Machine Learning, IBM Watson, and free platforms like Scikit.

REFERENCE

1. https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data
2. https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1
3. https://github.com/
4. https://www.wikipedia.org
5. https://www.javatpoint.com
6. https://www.geeksforgeeks.org
7. https://www.tutorialspoint.com
8. https://www.wikihow.com
9. https://stackoverflow.com
10. https://www.w3schools.com