Project and Data Management Plan

Comparison of Machine Learning vs Deep Learning Algorithms for House Price Prediction

Student Name: Shivakumar pole

Student Number: 21066186

Course: 7PAM2002-0509-2023-DATA SCIENCE PROJECT

Supervised by: Dr. William Cooper

Table of Contents

Chapter 1: Project Plan

Project Title

Comparison of Machine Learning vs Deep Learning Algorithms for House Price Prediction

Research Question

1. How do traditional ML algorithms perform compared to DL algorithms in the context of house price predictions?

2. Which features are most significant in influencing the predictions?

3. What impact does the volume of data have in performance of ML vs DL algorithms?

Objectives

➤ To review existing literature on the ML applications of real estate prices

➤ To analyse and process data obtained from open-source website without any ethical issues.

➤ To train multiple ML and DL models

➤ To design and tune DL models to increase its performance

To compare the results from ML and DL models using performance metrics Goyal, S. (2021) like Mean Squared Error and R squared (Goyal, 2021).

➤ To document the findings and provide a critical discussion about the results.

Background and Summary

Machine Learning and Deep Learning are two sides of the same coin. These statistical and mathematical techniques help the user to study past data and create an algorithm trained on the past data to predict future (*Deep learning vs. machine learning*, 2020). In this research these techniques will be tested on "House Price Prediction" problem (link). This research compares multiple Machine Learning and Deep Learning techniques like (Ahtesham, M., Bawany, N. Z. and Fatima, K. (2020))SVM, XGBoost, Multi-Layer Perceptron etc., (Ray, 2017) By evaluating multiple approaches, this research aims to understand which algorithm provides the best and most reliable predictions in the context of house prices prediction problem.

This project will encompass the results of the comparison between ML and DL algorithms. Along with the comparison of the algorithms the other part is to understand how the data works, which feature has the most impact, how the volume of the data affects the performance of these algorithms(Zhang, Z. 2019). To highlight the unique strengths and limitations of each approach while fine tuning these algorithms. This research will also include potential for future research and practical applications in real estate area.

List of References

1.Ahtesham, M., Bawany, N. Z. and Fatima, K. (2020) "House price prediction using machine learning algorithm - the case of Karachi city, Pakistan," in *2020 21st International Arab Conference on Information Technology (ACIT)*. IEEE, pp. 1–5.

https://ieeexplore.ieee.org/document/9300074

2.Chaurasia, A. and Haq, I. U. (2023) "Housing price prediction model using machine learning," in *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*. IEEE, pp. 497–500.

https://ieeexplore.ieee.org/document/10303359

3.Kiran, A. *et al.* (2023) "Modeling house price prediction model using XG boost and machine learning algorithms," in *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*. IEEE, pp. 1–9.

https://ieeexplore.ieee.org/document/10526190

4.The Danh Phan (2018) "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia," in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*. IEEE, pp. 35–42.

https://ieeexplore.ieee.org/document/8614000

Chapter 2: Task List and Project Time Line

Task	10-Jun	25-Jun	10-Jul	25-Jul	09-Aug	24-Aug	08-Sep
Data Collection and Project Plan	███						
Literature Review		███					
Data Pre-processing			███				
Machine Learning Model Development				███			
Deep Learning Model Development					███		
Comparison and Results						███	
FPR							

- Data Collection and Project Plan: Conduct initial data gathering and outline project steps.

   a)started data collection and project from the date :10/06/2024

- Literature Review: Review existing research and literature relevant to the project.

   a)literature review starts from 25/06/2024, it need sometime to understand regarding project.

- Data Pre-processing: Clean and prepare data for analysis and modelling.

   a)data pre-processing from 10th of July .

- Machine Learning Model Development: Develop and train machine learning models.

   a)developing and training machine learning models starts from 25/07/204

- Deep Learning Model Development: Develop and train deep learning models.

   a)after developing machine learning model, I'll start deep learning model that starts from 09/08/24

- Comparison and Results: Compare the performance of different models and summarize findings.

    a)comparison and output results on 24/08/2024

- FPR (Final Project Report): Compile the final project report with all results and conclusions.

Chapter 3: Data Management Plan

Summary of Dataset

It consists of a wide-ranging series of real estate transactions for residential properties (~ 600K observations thereof). Using data on house sales from a specific area and point in time, and detailed data on each individual property. It contains several attributes describing the 79 parcels and conditions under which has sold.

This database has the following main characteristics:

1. Selling Date and Price: For every record, there is a selling date and a deal price which is the primary dependent variable for all the analysis.
2. Property Details - Number of Bedrooms and Bathrooms, Size of Living Area, Size of Lot, Number of Stories, Waterfront (Y/N)
3. Additional Features: Additional important items include the condition and rating of the view, whether a basement exists and, if so, the floor size of the basement, the year the property was built and any improved or remodelling.
4. Spatial Context: This describes the street address, city, state, and postal code to which each property belongs.

A collection of data for a variety of cities in the state of Washington, USA the data contains the following attributes City related: Shoreline Seattle Kent Bellevue Redmond Year renovation data allows for a comparison of price per square foot over time for houses in the dataset.

Document control

GitHub Repository:

For Version control and Document control a combination of GitHub repository and Git.

Ethical requirements

- Does the data meet GDPR requirements? - Yes
- Does the project conform to UH ethical policies? – Yes
- Do you have permission to use the data for your proposed research project? - Yes
- Are you assured that the data was collected ethical (i.e. by the original people who gathered/collected/ collated/made the data)? - Yes

References:

5.Deep learning vs. machine learning (2020) *Zendesk*. Available at:
https://www.zendesk.com/in/blog/machine-learning-and-deep-learning/ (Accessed: June 8, 2024).

6.Goyal, S. (2021) *Evaluation metrics for regression models - analytics Vidhya - medium, Analytics Vidhya*.
Available at: https://medium.com/analytics-vidhya/evaluation-metrics-for-regression-models-
c91c65d73af (Accessed: June 8, 2024).

7.Ray, S. (2017) *Top 10 machine learning algorithms to use in 2024, Analytics Vidhya*. Available at:
https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ (Accessed: June
8, 2024).

8.Zhang, Z. (2019) *Understand data normalization in machine learning, Towards Data Science*. Available
at: https://towardsdatascience.com/understand-data-normalization-in-machine-learning-8ff3062101f0
(Accessed: June 8, 2024).

Dataset link : House price prediction (kaggle.com)

GitHub link: