Muhammed Yahya Avar 1921221004

BLM19432E INTRODUCTION TO DATA SCIENCE FINAL PROJECT

BOSTON HOUSING PRICE PREDICTION

Motivation

We plan to predict the housing prices in Boston. Boston has a competitive and dynamic real estate market, and accurate price prediction can help buyers, sellers, and real estate agents make informed decisions. Predicting housing prices is a challenging task due to various factors like location, size, number of rooms, accessibility to transportation etc. Our goal is to develop a machine learning model that can accurately predict the prices of houses in Boston using a dataset of housing prices and various features that can impact the prices.

Dataset

The Boston Housing Dataset is a popular and widely used dataset in the machine learning community. It contains information on various features of houses and their corresponding towns in Boston and their corresponding prices. The dataset includes 506 observations and 14 variables. The variables include features like the number of rooms, crime rate, pupil-teacher ratio, and more. The dataset is suitable for regression analysis, making it a good pick for our problem of predicting housing prices.

Method

In this project, we employed multiple regression-based machine learning techniques to predict housing prices. Regression is a widely used approach in data science for predicting continuous variables, making it suitable for our task of housing price prediction. We further applied various techniques such as Hyperparameter Tuning, Feature Engineering, and Evaluation methods to enhance and compare the performance of our models.

Experiments

- We first described our regression and evaluation methods.
- We started by reading the data and splitting it into training and test sets.
- We performed Hyperparameter Tuning for Decision Tree Regression and Random Forest Regression to optimize the models' hyperparameters.
- We implemented the three regression algorithms and evaluated their performance using techniques such as Root Mean Squared Error (RMSE), R-squared (R2) and Hypothesis Testing.
- We compared the algorithms based on their evaluation metrics and conducted hypothesis testing to assess the significance of their differences.
- We applied feature engineering to improve the Linear Regression model and compared its results with our old Linear Regression model.

Conclusion

Based on our findings,

- Tuned hyperparameters significantly improved the performance of the models.
- Random Forest Regression outperformed Linear Regression and Decision Tree Regression in terms of accuracy.
- Hypothesis testing indicated that there was no significant difference between the performance of Linear Regression and Decision Tree Regression.
- Feature engineering successfully enhanced the performance of the Linear Regression model.

Throughout the project, we encountered some difficulties. Hyperparameter turned to be performance costly. Understanding and implementing hypothesis testing posed challenges.

In conclusion, our project successfully compared regression algorithms, conducted hypothesis testing, and applied feature engineering for model improvement. Based on the evaluation results, Random Forest Regression emerged as the best-performing algorithm for predicting housing prices. The findings contribute to the understanding of different regression techniques and provide insights into their applicability in housing price prediction tasks.