CS6220 - Data Mining Techniques

Project Proposal

Project Title: Predicting Boston Housing Prices using Regression Analysis

Team : Lemon Grass

By:

Deep Rahul Shah
Orijeet Mukherjee
Pravin Anand Pawar
Preksha Patil

Problem Statement

The problem this project aims to address is the prediction of housing prices in the Boston area using historical data. The goal is to understand the impact of various factors such as crime rate, number of rooms, and socioeconomic status on the median value of homes. By developing predictive models, we seek to determine key contributors to property prices, best performing ML model for this task, and provide insights that can help in understanding housing affordability and market trends, while also considering the ethical issues linked to the dataset.

Approach

- **Exploratory Data Analysis (EDA):** Conduct an in-depth analysis of the Boston Housing dataset to understand data distributions, relationships between features, and identify potential outliers or missing values. Visualize key findings and gain insights into important features.
- **Feature Importance Analysis:** Use techniques such as feature importance and SHAP (SHapley Additive exPlanations) to determine the contribution of individual features. Additionally, combine this analysis with PCA to reduce the dimensionality of the model input vector.
- **Model Development:** Compare four machine learning algorithms, including Linear Regression, KNN, Random Forest, and Neural Network. Train and evaluate each model using metrics such as MAE, RMSE and R^2.

Expected Deliverables:

- Complete project documentation, including source code, detailed report (with analysis, methodology, experiments, plots, etc), PowerPoint presentation, and a recorded video presentation.
- A comprehensive analysis of the factors affecting housing prices.
- All four fine-tuned ML models each capable of accurately predicting housing prices, and comparative analysis to identify the best-suited ML algorithm for predicting housing prices.
- Insights into the ethical implications of using socioeconomic data in predictive modeling.