Internship Final Report

Student Name: Amith S Bhoomkar
University: Mysore University School of Engineering
Major: Artificial Intelligence and Machine Learning
Internship Duration: March 1st, 2025 - February 31st, 2025
Company: ShadowFox
Domain: AIML
Mentor: Mr. Hariharan
Coordinator: Mr. Aakash

Objectives

The main objective of this project is to predict house prices based on various features using machine learning algorithms. The specific objectives are:

- **Data Preprocessing**: Clean the dataset by handling missing values and outliers.
- **Feature Selection**: Identify and select the features (independent variables) that will be used to predict house prices.
- **Model Development**: Train and test multiple machine learning models (Linear Regression, Decision Tree, and Gradient Boosting) to find the best-performing model for predicting housing prices.
- **Model Evaluation**: Evaluate model performance using Mean Squared Error (MSE) and R-squared values. Additionally, perform cross-validation to assess model generalizability.
- **Hyperparameter Tuning**: Improve model performance by fine-tuning hyperparameters using GridSearchCV, particularly for the Gradient Boosting model.

Tasks and Responsibilities

Data Loading: Load the housing dataset (CSV file) into a pandas DataFrame for analysis.

Data Preprocessing:

- Identify and handle missing values by filling with the mean of the respective columns.
- Detect and remove outliers using the Interquartile Range (IQR) method.

Feature Selection:

- Separate features (independent variables) from the target variable (house prices).

Model Selection and Training:

- Split the data into training and testing sets (80% training, 20% testing).
- Apply feature scaling using StandardScaler to standardize the feature set for training.
- Train multiple models: Linear Regression, Decision Tree, and Gradient Boosting Regressor.

Model Evaluation:

- Evaluate each model using Mean Squared Error (MSE) and R-squared scores.
- Perform cross-validation to evaluate the consistency of the models.

Hyperparameter Tuning:

- Use GridSearchCV to tune hyperparameters for the Gradient Boosting model.
- Evaluate the best hyperparameter combination.

Visualization:

- Visualize the relationship between actual and predicted house prices for the best model.

Responsibilities

- **Data Handling**: Ensure that the dataset is properly loaded, cleaned, and preprocessed.
- **Model Development**: Implement different machine learning algorithms and evaluate their performance.
- **Hyperparameter Optimization**: Apply GridSearchCV to find the optimal parameters for the Gradient Boosting model.
- **Visualization**: Plot the results of the models and make the comparison clear for insights.

Learning Outcomes

Data Preprocessing Techniques:

- Handling missing data with appropriate methods such as filling with mean values.
- Using boxplots for detecting and removing outliers based on the IQR method.

Model Training and Evaluation:

- Understanding and implementing different machine learning algorithms for regression tasks.
- Learning how to evaluate models using different metrics such as Mean Squared Error (MSE) and R-squared.

Model Tuning:

- Gained practical experience in hyperparameter optimization with GridSearchCV to improve model performance.

Visualization:

- Created effective visualizations to compare the actual and predicted values, aiding in the interpretation of model performance.

Cross-Validation:

- Improved understanding of model stability and generalizability by applying cross-validation.

Challenges and Solutions

Handling Missing Values:

- Missing data can lead to inaccurate model predictions if not handled properly.

Solution:

- Used the mean imputation method to fill missing values, ensuring that the dataset remained consistent and complete.

Outliers:

- Outliers can significantly affect model performance, particularly for regression models.

Solution:

- Detected outliers using boxplots and removed them using the IQR method to enhance the quality of the data and reduce model bias.

Feature Scaling:

- Some models, like Gradient Boosting, are sensitive to the scale of input features, which can impact their performance.

Solution:

- Applied StandardScaler to normalize the features and ensure consistent scaling across the dataset.

Model Selection and Performance:

- Different models have varying levels of performance depending on the data.

Solution:

- Trained multiple models (Linear Regression, Decision Tree, Gradient Boosting) and evaluated their performance using MSE and R-squared.

Hyperparameter Tuning:

- Fine-tuning models with optimal parameters can be computationally expensive and time-consuming.

Solution:

- Used GridSearchCV to automate the hyperparameter tuning process for the Gradient Boosting model, saving time and effort while maximizing performance.

Conclusion

This project successfully predicted house prices using machine learning algorithms, and it highlighted several key aspects of data science, including preprocessing, model selection, evaluation, and optimization. By comparing different models, it was found that Gradient Boosting outperformed Linear Regression and Decision Tree, especially after hyperparameter tuning with GridSearchCV.

The overall process improved the understanding of how to prepare data for machine learning, choose appropriate models, and fine-tune their performance. Challenges such as missing values,

outliers, and model overfitting were effectively addressed using standard data science techniques. This project serves as a foundation for future improvements and experimentation with other models or techniques to improve the prediction accuracy further.

Acknowledgments

I would like to express my heartfelt gratitude to ShadowFox, particularly my mentor, Mr. Hariharan, and coordinator, Mr. Aakash, for their continuous guidance and support throughout my internship. I am also thankful to Amrita Vishwa Vidyapeetham for offering me this valuable internship opportunity, which has played a significant role in both my personal and professional growth.

This report represents the fusion of academic knowledge with the practical skills I have acquired during the internship, emphasizing my journey of learning, growth, and development in the field of data science.