Housing Values Prediction - Model Evaluation Summary

Dataset Description

- **Dataset**: Housing values in suburbs of Boston
- **Number of Instances**: 506
- **Number of Attributes**: 13 continuous attributes (including "class" attribute "MEDV"), 1 binary-valued attribute

Attribute Information

1. **CRIM**: per capita crime rate by town
2. **ZN**: proportion of residential land zoned for lots over 25,000 sq.ft.
3. **INDUS**: proportion of non-retail business acres per town
4. **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. **NOX**: nitric oxides concentration (parts per 10 million)
6. **RM**: average number of rooms per dwelling
7. **AGE**: proportion of owner-occupied units built prior to 1940
8. **DIS**: weighted distances to five Boston employment centres
9. **RAD**: index of accessibility to radial highways
10. **TAX**: full-value property-tax rate per $10,000
11. **PTRATIO**: pupil-teacher ratio by town
12. **B**: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. **LSTAT**: % lower status of the population
14. **MEDV**: Median value of owner-occupied homes in $1000's (target variable)

Model Evaluation Summary

1. Baseline Models

Linear Regression

- **Test MSE**: 36.744
- **Explanation**: Serves as a baseline model. Shows relatively high MSE indicating poor performance in capturing complex patterns.

Decision Tree Regressor

- **Test MSE**: 25.770
- **Explanation**: Better than Linear Regression but still has limitations due to overfitting.

2. Advanced Machine Learning Models

XGBoost

- **Test MSE**: 9.545
- **Explanation**: One of the best performing models. Shows significant improvement over baseline models.

LightGBM

- **Test MSE**: 13.264
- **Explanation**: Performs well but not as good as XGBoost.

CatBoost

- **Test MSE**: 10.139
- **Explanation**: Close to XGBoost in terms of performance.

3. Hyperparameter Tuning and PCA

XGBoost with PCA and Grid Search

- **Best Parameters**: {'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}
- **Best Cross-Validation Score**: 28.476
- **Test MSE**: 28.096
- **Explanation**: PCA did not significantly improve performance. Indicates potential loss of important features.

XGBoost with Regularization and Early Stopping

- **Best Parameters**: {'colsample_bytree': 0.8, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 1000, 'subsample': 0.9, 'reg_alpha': 0.1, 'reg_lambda': 0.1, 'early_stopping_rounds': 50}
- **Test MSE**: 8.392
- **Explanation**: Best overall performance. Regularization and early stopping effectively prevent overfitting.

4. Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT)

GCN

- **Test MSE**: 18.681
- **Explanation**: Shows potential but not as competitive as XGBoost.

GAT

- **Test MSE**: 50.364
- **Explanation**: Significantly higher MSE, indicating poor performance for this dataset.

5. ResNet and Transformer Combined Model

ResNet and Transformer

- **Test MSE**: 16.976
- **Explanation**: Shows promise with a reasonably low MSE. Can be further optimized.

6. Ensemble Model

Ensemble of Multiple Neural Networks

- **Test MSE**: 17.663
- **Explanation**: Combining multiple networks slightly improves performance but not significantly better than XGBoost.

7. Other Models

Ridge Regression

- **Mean MSE**: 66.070
- **Std**: 49.850

Random Forest Regressor

- **Mean MSE**: 31.193
- **Std**: 19.927

Support Vector Regressor (SVR)

- **Mean MSE**: 51.897
- **Std**: 22.457

New Model Evaluation Results

Random Forest Regressor

- **Cross-Validation Scores**: [4.70131959 5.37819324 2.53880605 3.16603768 5.10342101 5.61596865 2.89497432 2.79676277 3.53836665 3.91031402]
- **Mean**: 3.964
- **Standard Deviation**: 1.092
- **Test MSE**: 3.637
- **Explanation**: Shows strong performance with relatively low MSE.

XGBoost Regressor

- **Cross-Validation Scores**: [5.62262948 6.74368432 2.72090821 4.01210315 5.05194837 7.30687908 4.11483896 2.72731979 4.07106134 2.84766596]
- **Mean**: 4.522
- **Standard Deviation**: 1.552
- **Test MSE**: 3.778
- **Explanation**: Slightly higher MSE compared to Random Forest, but still demonstrates good performance.

Detailed Results Table

Model	Parameters	Cross-Validation MSE	Test MSE
Linear Regression	-	-	36.744
Decision Tree Regressor	-	-	25.770
XGBoost	-	-	9.545
LightGBM	-	-	13.264
CatBoost	-	-	10.139
XGBoost with PCA	{'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}	28.476	28.096
XGBoost with Regularization & Early Stopping	{'colsample_bytree': 0.8, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 1000, 'subsample': 0.9, 'reg_alpha': 0.1, 'reg_lambda': 0.1, 'early_stopping_rounds': 50}	-	8.392
GCN	-	-	18.681
GAT	-	-	50.364
ResNet and Transformer	-	-	16.976
Ensemble of Multiple Neural Networks	-	-	17.663
Ridge Regression	-	66.070	-
Random Forest Regressor	-	3.964	3.637
SVR	-	51.897	-
XGBoost Regressor	-	4.522	3.778

Key Insights

1. **Best Model**: XGBoost with regularization and early stopping demonstrated the best performance with the lowest Test MSE of 8.392. This suggests that it effectively captures the underlying patterns in the data without overfitting.

2. **PCA Impact**: The application of PCA did not significantly enhance the model's performance, indicating that the original features are more informative than the principal components.

3. **Regularization and Early Stopping**: These techniques proved crucial in improving model performance by preventing overfitting, as seen in the XGBoost model.

4. **Model Complexity**: More complex models like GCN and GAT did not perform as well as simpler gradient boosting methods, highlighting that complexity does not always equate to better performance for this dataset.

5. **Feature Engineering**: Including interaction terms and polynomial features can potentially improve model performance, but the impact varies based on the model used.

6. **Random Forest and XGBoost**: Both models showed strong performance with Random Forest achieving a slightly lower Test MSE than XGBoost in the final evaluation, making it a robust choice for this dataset.