Ira Narang, in2933

Project 2: House Prices Classification

1. Introduction

This study uses machine learning classification models to predict whether a house in California is priced above the median. A dataset containing housing attributes such as median income, house age, and average number of rooms was used. The objective was to apply multiple classification techniques, optimize their performance, and determine the most suitable model based on evaluation metrics.

2. Techniques Used for Model Training

The following supervised classification techniques were applied to predict house price categories:

- **K-Nearest Neighbors (KNN):** A distance-based algorithm that classifies a new data point by assigning it the most common label among its k-nearest neighbors in feature space. It is simple and effective but can be sensitive to feature scaling and the choice of k. I used two different k values to show the importance of the chosen value.
- **Decision Tree Classifier:** hierarchical model that splits the dataset based on feature thresholds to create decision rules.
- **Random Forest Classifier:** An ensemble method that constructs multiple decision trees and averages their predictions to enhance model stability and reduce overfitting.
- **AdaBoost Classifier:** A boosting algorithm that sequentially trains weak classifiers, adjusting their influence based on misclassification errors. This iterative process improves predictive accuracy and model robustness.

Each model was trained on the dataset following proper preprocessing steps such as data standardization and train-test splitting while maintaining class proportions.

3. Techniques for Optimizing Model Performance

To enhance model performance, we employed data standardization, hyperparameter tuning, and cross-validation. Standardization was particularly important for distance-based models like KNN, ensuring that features contributed equally. Hyperparameter tuning was conducted for each model: KNN involved selecting the optimal k-value, Decision Tree required adjusting depth and splitting criteria, Random Forest was optimized by adjusting the number of trees and feature selection methods, and AdaBoost was improved by tweaking the number of weak learners and learning rate.

4. Model Performance Comparison

After training and tuning the models, they were evaluated using key classification metrics: accuracy, precision, recall, and F1-score. Below is a summary of their performance:

Training Data:

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (optimized)	1.00	1.00	1.00	1.00
Decision Tree	1.00	1.00	1.00	1.00
Random Forest	1.00	0.99	1.00	1.00
AdaBoost	0.89	0.89	0.89	0.89

Test Data:

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (optimized)	0.60	0.60	0.60	0.60
Decision Tree	0.83	0.83	0.84	0.84
Random Forest	0.89	0.89	0.89	0.89
AdaBoost	0.88	0.87	0.88	0.88

From these results, it is clear that KNN and Decision Tree exhibit overfitting, achieving perfect accuracy on the training data but experiencing a significant drop in test performance. Random Forest and AdaBoost provided the most balanced results, maintaining high performance on both datasets. Random Forest slightly outperformed AdaBoost on the test data, but both models showed strong generalization.

5. Confusion Matrices

To analyze misclassification patterns, the confusion matrices for each model were examined. These matrices reveal how well each classifier differentiates between price categories and highlight potential areas for improvement.

6. Recommended Model

Considering all evaluation metrics, Random Forest emerged as the best-performing model. It demonstrated superior predictive capability with a high test accuracy of 89%, indicating strong generalization. Unlike KNN and Decision Tree, which exhibited overfitting, Random Forest maintained a balance between training and test performance. Its ensemble learning approach, which combines multiple decision trees, enhances stability and mitigates overfitting. Furthermore, the model's high precision and recall ensured minimal classification errors, making it the most reliable option for predicting house prices above the median.

7. Most Important Metric

For this classification task, F1-score was identified as the most critical metric. Since misclassifying high-priced houses could have significant implications, balancing precision and recall was essential. A high F1-score ensures that both false positives and false negatives are minimized, enhancing model reliability. Precision is important to prevent overvaluation of lower-priced houses, while recall ensures that genuinely high-priced houses are correctly classified. By balancing these aspects, F1-score provides a comprehensive evaluation of the model's effectiveness.

8. Conclusion

This project successfully applied various classification techniques to predict house prices above the median. By performing exploratory data analysis, optimizing models, and evaluating performance metrics, we determined that Random Forest is the most suitable model for this task. Its ensemble learning strategy allowed for improved generalization, making it a robust and reliable choice. The findings from this study suggest that machine learning can be a powerful tool for real estate price prediction.

Beyond model selection, this project provided valuable insights into the importance of feature engineering, model evaluation, and optimization techniques. The use of multiple models allowed for a comprehensive comparison of different classification strategies, highlighting the trade-offs between interpretability, accuracy, and generalization. Future extensions of this work could include exploring additional features, applying more complex ensemble methods, or testing the model on different geographical regions. These enhancements could further improve predictive performance and provide deeper insights into the housing market dynamics.