House Price Prediction

Introduction

The objective of this study is to develop a predictive model for classifying whether a house in California is priced above the median value. To achieve this, several supervised learning models were employed, including K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, Support Vector Machine (SVM), and Gradient Boosting Classifier. Each of these models was trained and evaluated on a dataset containing eight independent variables describing house characteristics, population density, and geographic location. This report provides a detailed analysis of the methodologies employed, optimization techniques used, comparative model performance, and recommendations for the most suitable model for this classification task.

Methodology

A stratified 80-20 train-test split was implemented to ensure a balanced distribution of the target variable across training and test sets. Given the varying scales of the features, standardization using StandardScaler() was applied to normalize all numerical variables, improving the performance of models sensitive to feature scaling, such as KNN and AdaBoost.

Six classification algorithms were utilized:

- K-Nearest Neighbors (KNN): A distance-based method relying on majority voting among the k-nearest data points.
- Decision Tree Classifier: A model that iteratively splits data based on feature importance, maximizing class separation at each node.
- Random Forest Classifier: An ensemble method aggregating multiple decision trees to reduce variance and improve predictive stability.
- AdaBoost Classifier: A boosting technique that sequentially trains weak classifiers, assigning greater weight to misclassified instances to enhance predictive accuracy.
- Support Vector Machine (SVM): A classification model that finds the optimal hyperplane that maximizes class separation.
- Gradient Boosting Classifier: An advanced ensemble method that builds trees sequentially to correct errors of previous iterations.

To optimize model performance, hyperparameter tuning was conducted using grid search (GridSearchCV), particularly for Random Forest and AdaBoost. Parameters such as the number of estimators (n_estimators), tree depth (max_depth), and learning rate (learning_rate) were optimized to maximize classification accuracy and generalizability. Model performance was assessed using accuracy, precision, recall, and F1-score, with confusion matrices generated to further analyze classification errors.

Results and Discussion

A comparative evaluation of the models revealed significant differences in their predictive capabilities. The Random Forest Classifier demonstrated the highest accuracy (89%), outperforming all other models in terms of precision, recall, and F1-score. Gradient Boosting followed with an accuracy of 88%, excelling in recall, indicating its effectiveness in correctly identifying houses priced above the median. The Support Vector Machine (SVM) achieved an accuracy of 85%, followed closely by AdaBoost at 85%, both of which performed reasonably well. The Decision Tree Classifier achieved an accuracy of 84%, though it exhibited a tendency to overfit the training data. Lastly, K-Nearest Neighbors yielded the lowest performance, with an accuracy of 82%, likely due to its sensitivity to feature scaling and neighborhood size selection.

Model Performance Comparison

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)
K-Nearest Neighbors	82%	82%	83%	82%
Decision Tree	84%	85%	83%	84%
Random Forest	89%	89%	89%	89%
AdaBoost	85%	85%	85%	85%
Support Vector Machine	85%	86%	85%	85%
Gradient Boosting	88%	88%	87%	88%

The Random Forest model exhibited superior classification performance across all key metrics, making it the most robust and generalizable approach for this problem. While Gradient Boosting performed competitively, it was computationally more expensive and required extensive tuning. The Decision Tree model, despite reasonable accuracy, was prone to overfitting, limiting its reliability in real-world applications. KNN, while conceptually straightforward, was outperformed by ensemble-based models that leveraged feature interactions more effectively.

Recommended Model

Based on the empirical results, the Random Forest Classifier is the most suitable model for this classification task. It provides the highest classification accuracy (89%) while maintaining a balanced trade-off between precision and recall, which is essential for mitigating both false positives and false negatives. Unlike a single decision tree, which is highly susceptible to overfitting, Random Forest employs an ensemble of trees, thereby reducing variance and

improving generalization. Its capability to handle feature interactions and scale effectively with large datasets further strengthens its applicability in house price prediction.

While Gradient Boosting demonstrated strong performance, it required substantial hyperparameter tuning and was computationally more demanding. Decision Trees, though interpretable, lacked generalizability due to overfitting tendencies. KNN's reliance on local feature distances rendered it less effective in this dataset. Given these factors, Random Forest offers the best combination of accuracy, stability, and computational efficiency, making it the optimal choice for predicting whether a house price exceeds the median value.

Evaluation Metrics: Importance of F1-Score

For this classification task, the F1-score is the most appropriate evaluation metric as it balances precision and recall, ensuring a minimization of both false positives and false negatives.

- Precision is critical to prevent overestimating house prices, which could mislead stakeholders in real estate valuation. Misclassifying a house as above the median price when it is not could result in inflated appraisals or misaligned investment decisions.
- Recall is equally important, as failing to correctly classify high-priced houses could result in missed investment opportunities or inaccurate market assessments.

Since misclassification carries implications in both directions, the F1-score provides a comprehensive measure of model effectiveness, ensuring that both false positives and false negatives are controlled. Among all models, Random Forest achieved the highest F1-score, confirming its suitability as the preferred model.

Conclusion

This study conducted a comparative evaluation of six supervised learning models for predicting whether a house price in California exceeds the median value. Through extensive analysis, Random Forest emerged as the most effective model, outperforming alternatives in classification accuracy, robustness, and generalization capability. Its ensemble approach mitigates overfitting and enhances predictive stability, making it a reliable choice for real estate pricing applications.

Future work could explore feature engineering techniques, incorporating additional geographic and economic indicators to improve model performance further. Additionally, alternative ensemble methods such as XGBoost, or the inclusion of Neural Networks or Transformers could be explored to determine if further performance gains can be achieved. Nevertheless, based on the current findings, Random Forest remains the most practical and effective model for this classification task.