Arvind Vinod
Stubbs
COE 379L

Project 2 Report

The goal of this project was to predict whether houses in California are priced above the median value using a dataset with 8 independent features and a binary target variable (price_above_median). The project involved multiple exploratory data analysis methods, data preprocessing, and employing several supervised learning techniques, including K-Nearest Neighbors (KNN), Decision Trees, Random Forest, and AdaBoost, to build classification models. This report summarizes the techniques I used, the optimization strategies I applied, the performance of each model, and my final recommendation for the best model to use for this dataset.

The raw data was given in a CSV file, which was then loaded into a Pandas DataFrame.Using the shape function, one can see that the dataset contains 20,640 rows and 9 columns, making it moderately sized for machine learning purposes. The dataset includes features such as median income, house age, average rooms, average bedrooms, population, average occupancy, latitude, and longitude (Longitude). The target variable is price_above_median, which indicates whether a house is priced above the median value. I began by checking for duplicate rows using the duplicated function and found no duplicates. Next, I examined the data types of each column and found that most columns were already in numerical formats, but I converted HouseAge and Population from floats to integers since they originally contained decimal values ending in .0 so it made more sense for them to be ints. The mean and max values of median income revealed significant wealth disparities across neighborhoods, with some areas having incomes as high as $150,000. This likely reflects affluent regions like the Bay Area and Los Angeles. The average bedrooms column had a minimum value of 0.33, which seemed unusual but could represent shared living spaces or studio apartments. The Population column showed densely populated neighborhoods, with a maximum value of 35,682 residents per block. Latitude and longitude values confirmed that the dataset covers the entire state of California. A few histograms and box plots were created using Seaborn. This revealed that most houses are between 15 and 50 years old, with noticeable spikes at 15, 25, 35, and 50 years. This suggests that certain decades saw

higher construction activity. Median income showed a peak around 35,000 but following a relatively Gaussian distribution from $0 to $80,000. Box plots for Latitude and Longitude confirmed that the data spans the geographical boundaries of California.

Next, I trained several models to predict whether a house is priced above the median value. I started with K-Nearest Neighbors (KNN). Since KNN is prone to failure due to the scale of the data, I standardized the features to ensure all features contributed equally to the distance calculation. I first ran a baseline then to optimize the model, I used GridSearchCV to tune hyperparameters like n_neighbors and weights. However, the tuned model did not significantly do better than the baseline, achieving an accuracy of 0.84 compared to the baseline's 0.83. The next model was Decision Trees. Like before, first came a baseline model and then I used GridSearchCV to tune hyperparameters. The tuned model achieved an accuracy of 0.85, a slight improvement over the baseline's 0.83. At this point I was questioning why I was even tuning the model. Interestingly, the training accuracy was significantly higher than the test accuracy, pointing to overfitting. Then I implemented Random Forest, a method that combines multiple decision trees to improve generalization. I tuned n_estimators, max_depth, and min_samples_leaf using GridSearchCV. The tuned model achieved the highest accuracy of 0.90. This made it the best-performing model for this project. Lastly, I did AdaBoost with a baseline. I then tuned n_estimators and learning_rate to optimize performance and the tuned model achieved an accuracy of 0.88, which was slightly lower than Random Forest but still competitive.

Moving on we can measure the performance of all models on the test data using accuracy, recall, precision, F1-score, and confusion matrices. The KNN model achieved an accuracy of 0.83 and an F1-score of 0.83. The confusion matrix showed that the model misclassified a significant number of instances, particularly false positives and false negatives. This shows that KNN struggled to capture some underlying patterns in the data. The Decision Tree model did slightly better than KNN, with an accuracy of 0.85 and an F1-score of 0.85. The confusion matrix showed less misclassifications compared to KNN, but the model still struggled with overfitting, especially seen with the gap between training and test performance. The Random Forest model achieved the best performance, with an accuracy of 0.90 and an F1-score of 0.90. The confusion

matrix showed the fewest misclassification which demonstrated the model's ability to generalize well. The AdaBoost model also performed well and was competitive, however, it fell slightly short of Random Forest in terms of overall performance.

For this dataset, I believe the F1-score is the most important metric. The dataset is balanced, meaning accuracy alone might not fully capture the model’s performance. The F1-score, which balances precision and recall, is necessary when the costs of false positives are similar. A high F1-score indicates that the model performs well in both identifying true positives and minimizing false positives and false negatives. This balance is essential for making reliable predictions in real-world scenarios.

To sum up, the training and evaluation of multiple classification models to predict whether houses in California are priced above the median value was performed in this project. The Random Forest (Tuned) model emerged as the best-performing model, achieving an F1-score of 0.90. Its ability to generalize well and handle complex relationships in the data makes it the ideal choice for this problem. Moving forward, I recommend using this model for predicting house prices above the median in California. Further improvements could involve collecting more data or experimenting with advanced ensemble methods.