Model Evaluation & Validation: Predicting Housing Prices

This presentation explores techniques for evaluating and validating machine learning models to predict housing prices. We'll examine data exploration, model development, performance analysis, and making predictions on real-world data.

Menna Mahmoud EL-Bagoury

Data Exploration

We begin by exploring the Boston housing dataset, containing 489 data points with 4 key features. Understanding the data distribution and relationships is crucial before modeling.

1. **Key Features**

   Room count (RM), neighborhood poverty (LSTAT), student-teacher ratio (PTRATIO), and median home value (MEDV).

2. **Data Preparation**

   Removed outliers and scaled the target variable to account for 35 years of inflation.

3. **Statistical Analysis**

   Calculated minimum, maximum, mean, median, and standard deviation of home prices.

Performance Metric: R-squared

We use the coefficient of determination (R-squared) to quantify model performance. It measures the proportion of variance in the target variable explained by the features.

R-squared Range

Values range from 0 to 1. Higher values indicate better predictive power.

Interpretation

R-squared of 0.40 means 40% of variance in Y is predictable from X.

Implementation

We use sklearn's r2_score function to calculate R-squared for our models.

Data Splitting: Train-Test

We split the data into training (80%) and testing (20%) subsets. This allows us to evaluate model performance on unseen data.

1. ## Shuffle Data

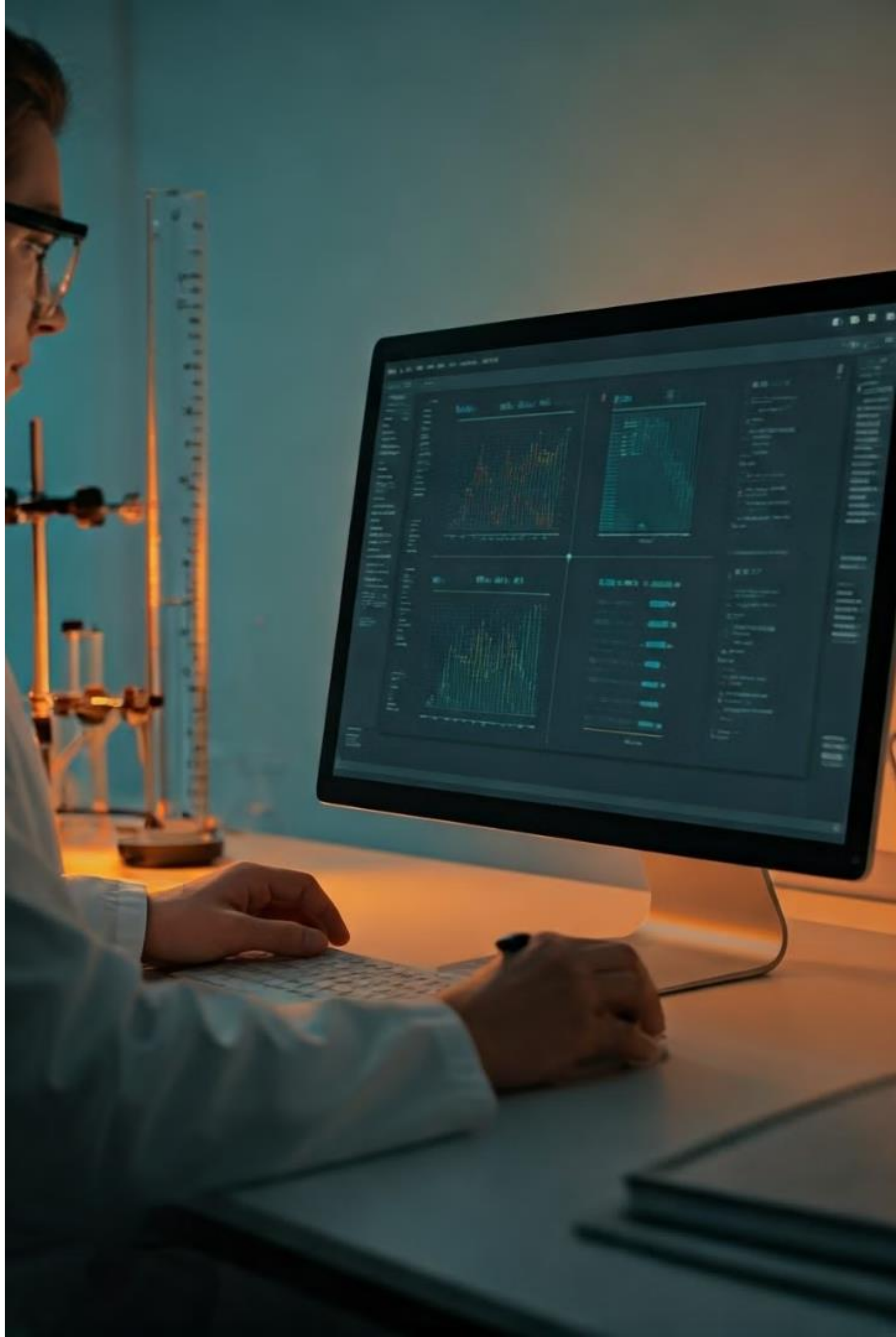
   Randomize data order to remove potential bias in the dataset.

2. ## Split Data

   Divide into 80% training and 20% testing subsets using train_test_split.

3. ## Assign Variables

   Store splits in X_train, X_test, y_train, and y_test variables.



Learning Curves

Learning curves visualize model performance as training set size increases. We examine curves for decision trees with different maximum depths.

1

Training Score

Measures how well the model fits the training data.

2

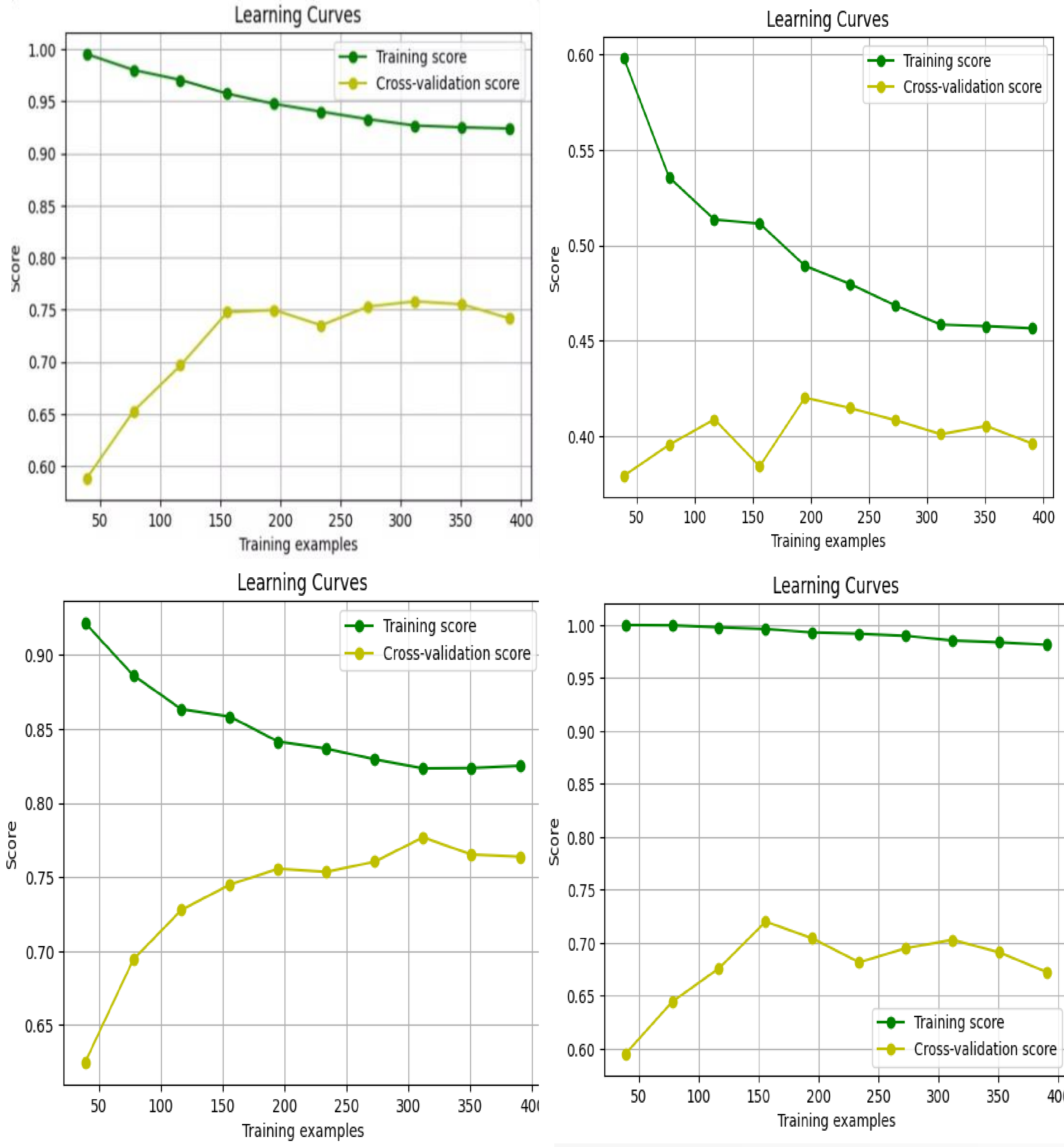
Cross-validation Score

Estimates model performance on unseen data.

3

Convergence

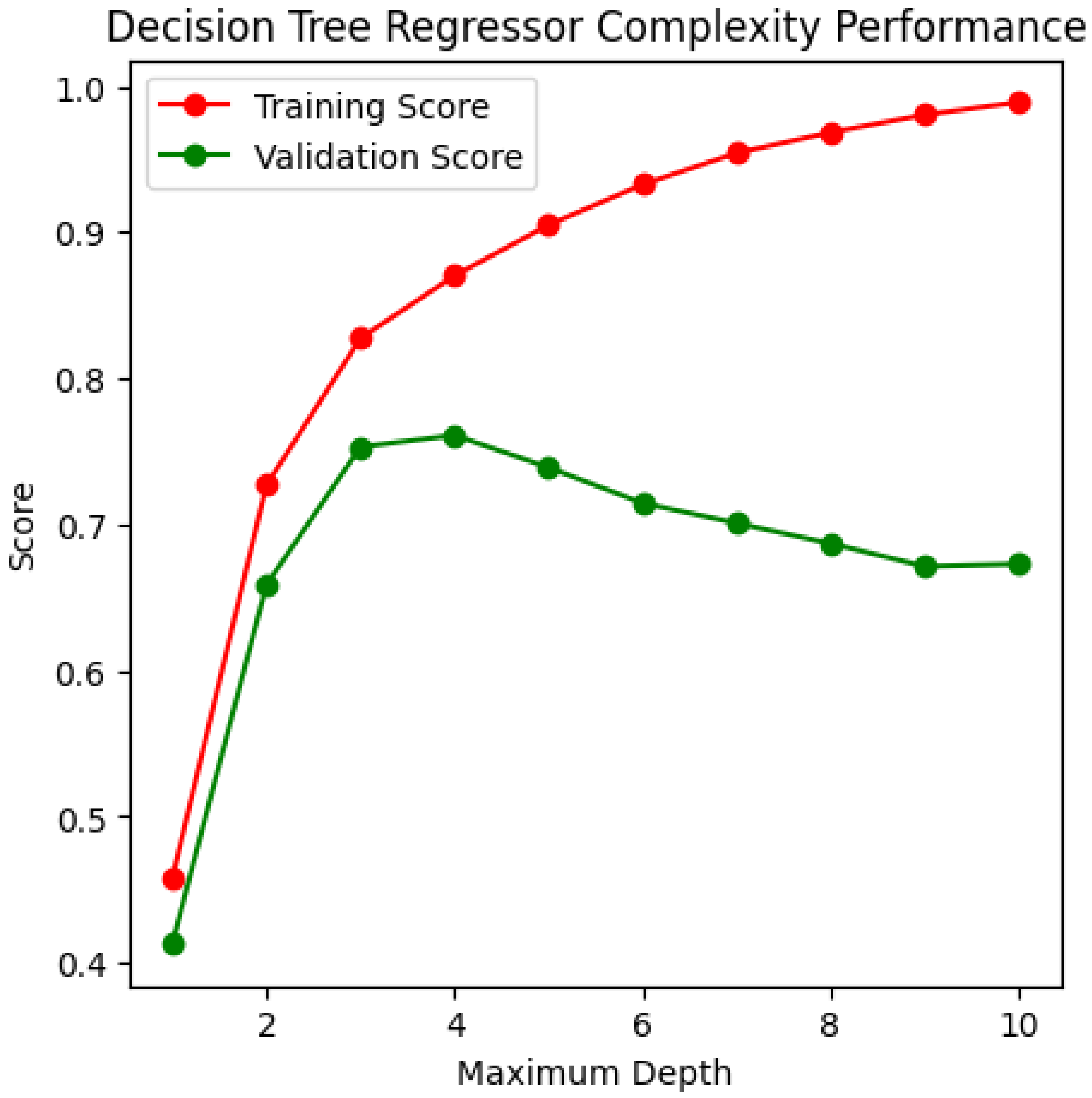
Ideal curves converge at a high score as training size increases.

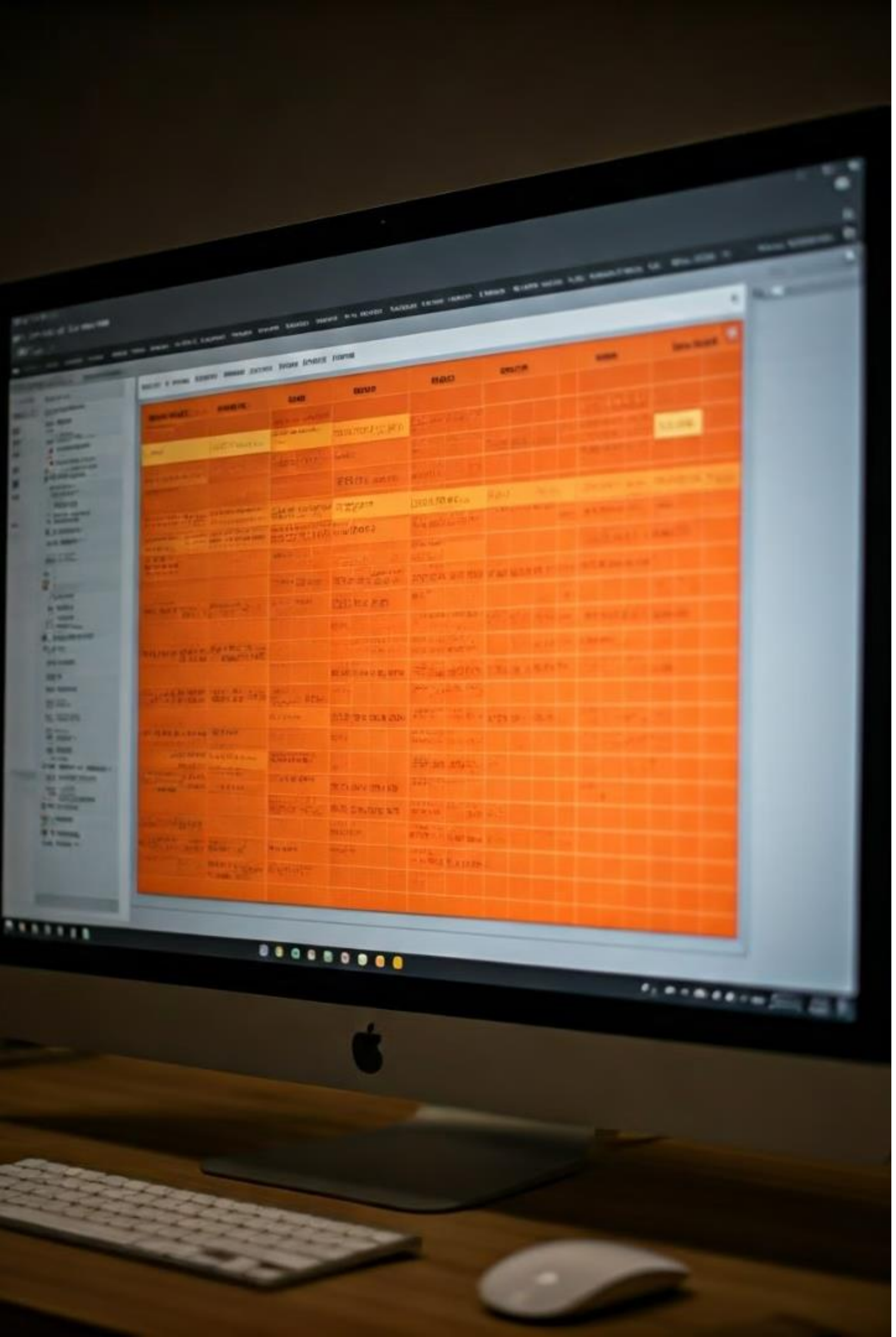


Bias-Variance Tradeoff

We analyze the bias-variance tradeoff using complexity curves. These show model performance as the maximum depth increases.

Low Depth	High Bias	Underfitting
High Depth	High Variance	Overfitting
Optimal Depth	Balance	Best Generalization





Grid Search Optimization

We use grid search to find the optimal maximum depth for our decision tree model. This technique systematically tests various hyperparameter combinations.

Parameter Grid

Define a range of max_depth values from 1 to 10.

Cross-validation

Use ShuffleSplit with 10 splits and 20% test size.

Scoring Function

Use R-squared metric to evaluate model performance.

Best Estimator

Select the model with the highest cross-validation score.

Optimal Model Selection

After grid search, we found the optimal maximum depth for our decision tree model. This balances model complexity and generalization ability.

Optimal Depth

The best max_depth value was determined to be 4.

Balanced Model

This depth provides a good tradeoff between bias and variance.

Generalization

The model should perform well on unseen data.

Making Predictions

We use our optimized model to predict housing prices for three hypothetical clients. Each client's home has different features.

Client	Rooms	Poverty %	Student-Teacher Ratio	Predicted Price
1	5	17%	15-to-1	$403,025
2	4	32%	22-to-1	$237,479
3	8	3%	12-to-1	$931,636

Conclusion and Next Steps

Our model successfully predicts housing prices based on key features. However, there's always room for improvement in machine learning projects.

1 Model Performance

The decision tree model provides reasonable predictions for different housing scenarios.

2 Limitations

The model may not capture all factors influencing housing prices.

3 Future Work

Consider ensemble methods or additional features to enhance prediction accuracy.

Any Questions

Thanks