Predicting Housing Market Trends in Melbourne: Project Report

1. Project Overview

Objective

This project focuses on analyzing historical data to predict housing prices in Melbourne, Australia. By leveraging machine learning models, we aim to develop accurate predictions of median house prices, assisting stakeholders such as real estate agents, investors, and policymakers in understanding market trends.

Key Highlights

- Utilized historical transaction and property data.
- Explored both baseline and advanced machine learning models.
- Incorporated feature engineering to enhance model performance.

---

2. Methodology

Data Overview

- **Source**: City of Melbourne open datasets.
- **Features**: Transaction counts, property types, small area indicators, and historical price trends.
- **Target Variable**: Median housing price.

Workflow

1. **Data Cleaning**:
    - Handled missing values using median imputation based on property type.
    - Encoded categorical features using label encoding.
    - Dropped redundant or irrelevant features.
2. **Feature Engineering**:
    - Created rolling averages of median prices to capture historical trends.
    - Calculated transaction density as a proxy for market activity.
3. **Modeling**:
    - **Baseline**: Linear Regression using a simple train-test split.
    - **Advanced**: Tuned Random Forest Regressor with hyperparameter optimization via GridSearchCV.
4. **Evaluation Metrics**:
    - Root Mean Squared Error (RMSE).
    - R^2 Score.

---

3. Results

Model Comparison

Model	Test RMSE	Test R^2
Linear Regression	276,091	0.519
Random Forest (Tuned)	197,573	0.754

Key Insights

1. The tuned Random Forest model significantly outperformed the Linear Regression model:
   - **28% reduction** in RMSE.
   - Explained **75.4%** of the variance in housing prices, compared to 51.9% for Linear Regression.
2. Random Forest’s ability to capture nonlinear relationships and interactions between features contributed to its superior performance.

---

4. Visual Analysis

Residual Plot

- Residuals are centered around zero, indicating unbiased predictions.
- Some outliers exist, particularly for higher price ranges.

Feature Importance

- **Key Features**:
  1. Rolling average price: Most influential predictor.
  2. Transaction density: Highlights the significance of market activity.
  3. Sale year and property type: Contributed moderately to predictions.

---

5. Conclusion

Achievements

- Developed a robust machine learning model to predict housing prices in Melbourne.
- Enhanced accuracy by leveraging feature engineering and Random Forest's capability to handle complex data relationships.

Key Takeaways

- Historical price trends and transaction density are critical predictors of housing prices.
- Random Forest outperformed Linear Regression due to its ability to model nonlinearities and interactions.

Next Steps

1. **Expand Features**:
   - Include socioeconomic and environmental factors for richer insights.
2. **Address Outliers**:
   - Investigate and manage high-residual cases.
3. **Deployment**:
   - Implement the model in a dashboard or API for real-time predictions.
4. **Broader Applications**:
   - Apply the methodology to other regions for generalization.

---

6. Appendix

Data Sources

- City of Melbourne Open Data: House Prices by Small Area.

Tools and Libraries

- **Programming**: Python (Pandas, Scikit-learn, Matplotlib, Seaborn).
- **Modeling**: Linear Regression, Random Forest.
- **Environment**: Jupyter Notebook, GitHub for version control.

This report highlights the project's approach, key findings, and future potential. The Random Forest model, with its superior accuracy and robust insights, stands ready for deployment and further development.