Steve Mathew

CMPSC 445

March 6, 2024

Individual Project Report

**Predicting House Prices in Philadelphia**

**Introduction**

This project embarked on the development of a predictive model for accurately forecasting house prices in Philadelphia, leveraging the capabilities of machine learning (ML) algorithms. With the real estate market's dynamic nature, stakeholders such as buyers, sellers, and investors face significant challenges in making informed decisions. The model developed in this project aims to demystify price determinants and predict outcomes with high precision, thereby serving as a crucial tool for real estate decision-making.

**Data Collection**

The project utilized a comprehensive dataset from the Real Estate Transfers section of OpenDataPhilly, available at https://opendataphilly.org/dataset/real-estate-transfers. This dataset includes detailed records of property transactions within Philadelphia, capturing essential attributes like property location, size, and other pertinent characteristics that influence house prices. The richness and diversity of this dataset provide a solid foundation for building a predictive model reflective of the real estate market's complexities.

**Data Preprocessing**

Data preprocessing played a pivotal role in preparing the dataset for effective model training. Key steps undertaken include:

Numeric Conversion: The “SalePrice” column contained non-numeric characters, necessitating its conversion to a numeric format to facilitate mathematical operations.

Datetime Conversion: The “SaleDate” was transformed into a datetime format, enabling temporal analyses and the calculation of property ages.

Outlier Removal: Utilizing the Interquartile Range (IQR) method, outliers in critical columns were identified and removed, thereby reducing potential skewness and improving model accuracy.

**Feature Engineering**

In anticipation of enhancing the model's predictive capabilities, several features were engineered:

* Property Age: This feature reflects the age of a property at the time of sale, providing insights into its depreciation or appreciation over time.
* Size per Bedroom/Bathroom: These features offer perspectives on space utility and efficiency, factors often considered in property valuation.
* SqftAge Interaction: An innovative feature capturing the interaction between a property's size and age, hypothesizing a compounded effect on the property's market value.

**Model Development**

The project explored four distinct ML algorithms to ascertain the most effective model for predicting house prices:

1. Linear Regression: A baseline model for continuous outcome prediction, providing a straightforward interpretation of feature impacts.
2. Decision Tree Regressor: A model capable of capturing non-linear patterns through hierarchical decision rules.
3. Random Forest Regressor: An ensemble method enhancing prediction accuracy by aggregating multiple decision trees.
4. Gradient Boosting Regressor: An advanced ensemble technique focusing on sequentially correcting prediction errors.

**Model Evaluation**

The models were evaluated based on their Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) on the test dataset. The results are as follows:

- Linear Regression

- MAE: 4470.36

- RMSE: 5409.62

- Decision Tree

- MAE: 1089.71

- RMSE: 2441.40

- Random Forest

- MAE: 1477.90

- RMSE: 2645.19

- Gradient Boosting

- MAE: 1790.46

- RMSE: 3850.70

The Decision Tree Regressor outperformed the other models, achieving the lowest MAE and RMSE, indicating its superior accuracy in predicting house prices in Philadelphia.

**Discussion**

The Decision Tree Regressor's superior performance is attributed to its ability to model complex, non-linear relationships within the data. This model's success underscores the importance of feature engineering and preprocessing in enhancing predictive accuracy. However, the process encountered challenges such as handling outliers and optimally selecting features, addressed through meticulous data preprocessing and innovative feature engineering.

**Future Considerations**

To improve the model further, I would integrate:

- Parameter Tuning: Apply grid search or random search to fine-tune the Decision Tree Regressor's hyperparameters.

- Additional Data Collection: Expanding the dataset to include more recent transactions or additional variables could offer deeper insights.

- Advanced Feature Engineering: Exploring more complex features and interactions could capture additional nuances affecting house prices.

**Conclusion**

This project demonstrates machine learning's potential in transforming real estate market analyses. The Decision Tree Regressor, with its promising results, offers a glimpse into the future of real estate pricing models. By continuously refining the model and incorporating more complex data, we can significantly enhance its utility for stakeholders in the Philadelphia real estate market, ensuring more accurate and informed decision-making.