**Machine Learning Based Advance Housing Price Prediction**

Rahul Kumar

09/08/2024

Abstract

This project explores the integration of advance ML techniques in Predicted the housing price increasing in futurewe will focus on creating Machine Learning Pipelines considering all the life cycle of a Data Science Projects. This will be important for professionals who have not worked with huge dataset. The main aim of this project is to predict the house price based on various features which we will discuss as we go ahead.

Advanced House Price Prediction Report

Table of Contents

1. Introduction
2. Data Collection
3. Exploratory Data Analysis (EDA)
4. Feature Engineering
5. Model Building
6. Model Evaluation
7. Conclusion
8. References

**1. Introduction**

**Project Overview**

The objective of this project is to develop a predictive model to estimate the price of residential houses. This is crucial for buyers, sellers, and real estate professionals to make informed decisions based on a variety of factors that influence house prices.

**Problem Statement**

Accurately predicting house prices is a challenging task due to the numerous factors that affect property values. Our goal is to create a model that provides reliable predictions by analyzing historical housing data and identifying significant features contributing to house prices.

**Objectives**

* **Data Analysis**: Understand the characteristics of the dataset and perform necessary cleaning.
* **Feature Engineering**: Create and select the most relevant features for the prediction model.
* **Modeling**: Build and evaluate different machine learning models for accuracy and efficiency.
* **Prediction**: Deploy the best-performing model for real-time price prediction.

**2. Data Collection**

**Data Sources**

The dataset used in this project is obtained from Kaggle's House Prices - Advanced Regression Techniques. It contains 79 explanatory variables describing various aspects of residential homes.

**Dataset Description**

* **Size**: 1460 samples
* **Features**: 81 columns, including target variable SalePrice.
* **Types**: Numerical, categorical, and ordinal features.

**Key Variables**

* **Lot Area**: Lot size in square feet.
* **OverallQual**: Overall material and finish quality.
* **YearBuilt**: Original construction date.
* **TotalBsmtSF**: Total square feet of basement area.
* **GrLivArea**: Above grade (ground) living area in square feet.
* **GarageCars**: Size of the garage in car capacity.
* **GarageArea**: Size of garage in square feet.

**3. Exploratory Data Analysis (EDA)**

**Data Cleaning**

* **Missing Values**: Identified and imputed missing values in various features such as GarageType, MasVnrArea, etc.
* **Outliers**: Detected and treated outliers in critical features like GrLivArea and TotalBsmtSF.
* **Encoding Categorical Variables**: Converted categorical variables into numerical using one-hot encoding and label encoding.

**Data Visualization**

Using various visualization techniques, we explored relationships and patterns within the data:

* **Correlation Heatmap**: Analyzed correlations between features and SalePrice.
* **Distribution Plots**: Visualized distributions of key numerical features like GrLivArea and SalePrice.
* **Boxplots**: Examined the effect of categorical variables like OverallQual on SalePrice.
* **Pair Plots**: Explored pairwise relationships between features.

**Findings**

* **Strong Correlations**: GrLivArea, OverallQual, and TotalBsmtSF show strong positive correlations with SalePrice.
* **Distribution**: SalePrice is right-skewed; a log transformation can normalize it.
* **Missing Values**: Features like PoolQC and MiscFeature have a significant number of missing values, handled appropriately.

**4. Feature Engineering**

**Feature Selection**

We used techniques such as **Recursive Feature Elimination (RFE)** and **Lasso Regression** to select the most impactful features.

**New Features**

* **Age of House**: YrSold - YearBuilt
* **Total Area**: Sum of all floor areas including basement and garage.
* **Quality Index**: Product of OverallQual and OverallCond.

**Transformations**

* **Normalization**: Applied log transformation to SalePrice for normal distribution.
* **Encoding**: Used label encoding for ordinal variables and one-hot encoding for nominal variables.

**Dimensionality Reduction**

**Principal Component Analysis (PCA)** was employed to reduce dimensionality, retaining 95% of variance.

**5. Model Building**

**Models Used**

We evaluated several machine learning algorithms, including:

* **Linear Regression**
* **Ridge Regression**
* **Lasso Regression**
* **Decision Trees**
* **Random Forests**
* **Gradient Boosting Machines (GBM)**
* **XGBoost**
* **LightGBM**
* **Neural Networks**

**Model Selection**

The models were selected based on their performance using cross-validation techniques.

**Hyperparameter Tuning**

Grid Search and Random Search techniques were utilized to optimize hyperparameters for each model.

**6. Model Evaluation**

**Evaluation Metrics**

* **Mean Absolute Error (MAE)**
* **Mean Squared Error (MSE)**
* **Root Mean Squared Error (RMSE)**
* **R² Score**

**Best Model**

The **LightGBM model** performed the best, achieving the highest R² Score and lowest RMSE. Its ability to handle complex interactions and large datasets contributed to its superior performance.

**7. Conclusion**

**Summary**

This project successfully developed an advanced predictive model for house prices using various machine learning techniques. The LightGBM model demonstrated the best performance in terms of accuracy and computational efficiency.

**Key Insights**

* Feature selection and engineering played a crucial role in improving model accuracy.
* Handling missing values and outliers significantly impacted model performance.
* Advanced algorithms like LightGBM and XGBoost proved highly effective for regression tasks in this domain.

**Future Work**

* Explore the integration of more external data sources, such as economic indicators, for even better predictions.
* Implement a real-time prediction system that can be deployed in a production environment.
* Investigate more complex deep learning architectures for potentially improved performance.

1. **References**
2. Kaggle. (n.d.). House Prices - Advanced Regression Techniques. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.
4. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer series in statistics New York.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).