Predicting House Sale Prices: A Comprehensive Analysis and Model Comparison

Nasrin Sultana Nipa, Graduate Student
Department of Computer Science and Mathematics
Arkansas State University

SCAN ME

Abstract

The objective of our project is to accurately predict housing prices in urban areas using machine learning techniques.

Overall, we have conducted a comprehensive analysis involving data preprocessing, model training, hyperparameter tuning, and evaluation. We selected and trained three machine learning algorithms: Linear Regression, Random Forest, and XGBoost, to predict housing prices based on various features. We optimized the models through hyperparameter tuning and evaluated their performance using metrics such as RMSE and R-squared. Our aim was to develop accurate predictive models capable of estimating housing prices with precision.

In conclusion, our study successfully achieved its objective of predicting housing prices using machine learning techniques. We demonstrated the effectiveness of ensemble methods, particularly Gradient Boosting, in accurately predicting prices. Our findings provide valuable insights for stakeholders and policymakers involved in real estate valuation and urban planning.

Introduction

Machine learning techniques have emerged as valuable tools for leveraging diverse property features to make predictions. This study explores the effectiveness of combining various machine learning models using the Stacked Generalization technique to enhance predictive accuracy. Utilizing a comprehensive dataset from Kaggle, we assess model performance and aim to achieve more accurate predictions. Before delving into the methodology, we first examine the distribution of the dependent variable 'SalePrice' to understand the dataset's characteristics. This introduction provides context for our analysis and sets the stage for our approach.

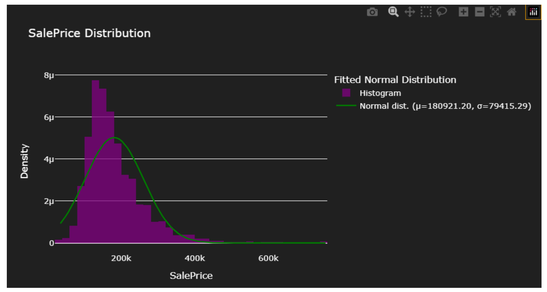

Figure 1: Fitted Normal Distribution for SalePrice vs. Density

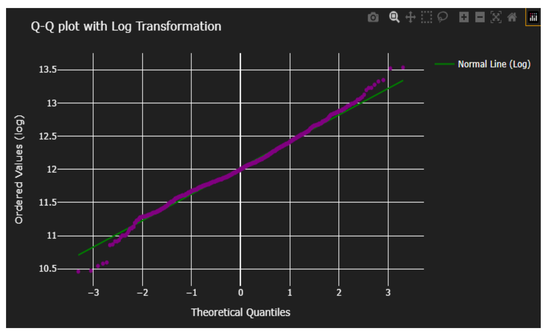

Figure 2: Q-Q plot with log transformation to make the best fit of line

Methods and Materials

START
1. DATA ACQUISITION:
   - Obtain housing dataset.
   - Ensure dataset permissions.
   - Download dataset.
2. DATA PREPROCESSING:
   - Handle missing values.
   - Clean data (remove duplicates, outliers).
   - Encode categorical variables.
   - Split data into train/test sets.
3. FEATURE ENGINEERING:
   - Create new features.
   - Combine related features.
   - Transform features (e.g., log transformation).
4. MODEL SELECTION:
   - Choose regression models (e.g., Random Forest).
   - Consider ensemble methods.
5. MODEL TRAINING:
   - Train models on training data.
   - Tune hyperparameters.
6. MODEL DEPLOYMENT:
   - Apply trained models to test data.
   - Make predictions.
7. MODEL EVALUATION:
   - Evaluate performance (RMSE, R^2, MAE).
   - Compare model performance.
   - Assess robustness (cross-validation).
8. REPORTING:
   - Document methodology.
   - Present findings.
   - Discuss limitations/future research.

END

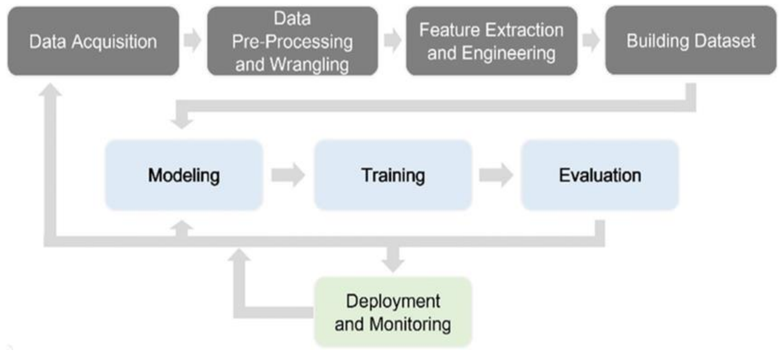

Figure 3: Work flow for predictive analysis

Results

Performance on Training Set	Performance on Test Set
Fitting 3 folds for each of 1 candidates, totalling 3 fits Best parameters for LinearRegression: {} Best RMSE for LinearRegression: 0.16395567878468517	Training and tuning LinearRegression as the meta-model... Fitting 3 folds for each of 1 candidates, totalling 3 fits Best parameters for LinearRegression: {} Best RMSE for LinearRegression: 0.1485080907637146
Fitting 3 folds for each of 27 candidates, totalling 81 fits Best parameters for RandomForest: {'max_depth': None, 'min_samples_split': 5, 'n_estimators': 500} Best RMSE for RandomForest: 0.15030523155118708	Training and tuning XGBoost as the meta-model... Fitting 3 folds for each of 27 candidates, totalling 81 fits Best parameters for XGBoost: {'final_estimator__learning_rate': 0.01, 'final_estimator__max_depth': 3, 'final_estimator__n_estimators': 500} Best RMSE for XGBoost: 0.13431634146082957
Fitting 3 folds for each of 27 candidates, totalling 81 fits Best parameters for XGBoost: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200} Best RMSE for XGBoost: 0.13782341817422314	LinearRegression's RMSE on test data: 0.1422618725378032 RandomForest's RMSE on test data: 0.15181965528816396 XGBoost's RMSE on test data: 0.14156248294694176

Discussion

Our analysis demonstrates the effectiveness of machine learning models in predicting housing prices, with XGBoost emerging as the top performer. Its ability to handle complex relationships and capture nonlinear patterns led to superior predictive accuracy compared to Linear Regression and Random Forest.

The implications extend to various stakeholders, enabling them to make informed decisions in the housing market. Future research should explore alternative techniques and data sources to further enhance predictive accuracy and provide deeper insights into market dynamics.

Conclusion

XGBoost's success highlights the potential of machine learning in real estate valuation. By leveraging advanced algorithms, stakeholders can enhance their forecasting capabilities and navigate the housing market more effectively. Further research can lead to even more accurate predictions and valuable insights for stakeholders.

Acknowledgement

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- researchgate.net/figure/The-basic-machine-learning-workflow-includes-data-acquisition-data-pre-processing-data_fig1_363282014
- DR. JASON L CAUSEY
  ASSOCIATE PROFESSOR OF BIOINFORMATICS
  Associate Director, Center for No-Boundary Thinking Division Lead, CNBT Division of Algorithms & Computational Methodology