A Comparative Study of Selected Machine Learning Methods for Predicting Housing Prices

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Kwanele N. Mnisi (18164049)
Department of Statistics: University of Pretoria

TOMORROW TRUST

Abstract

Buying or selling a house is one of the most significant financial transactions that most people undertake in their lifetime. For this reason, it is crucial to have an accurate understanding of the value of a property, whether you are a buyer looking to make a wise investment, or a seller seeking to maximize your profit. The study:

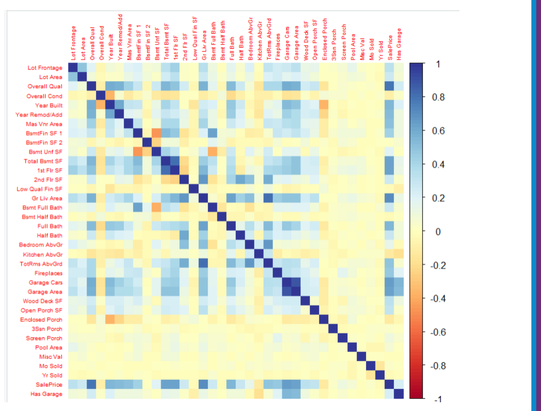
- Compares Random Forest, Support Vector Regression, and XGBoost, focusing on their predictive performance.
- It employs both internal (within techniques) and external (between techniques) comparisons.

The findings were that XGBoost performs the best, especially with feature selection. The study emphasizes the importance of feature selection in enhancing model performance.

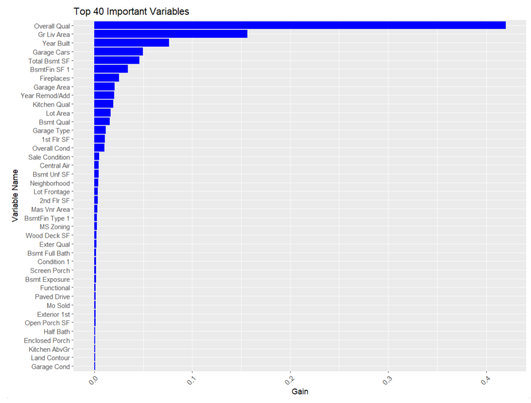
Theoretical Comparisons of the three selected techniques

Aspect	Random Forest	Support Vector Regression	XGBoost
Handling of missing data	Can handle missing data	Needs preprocessing for missing data	Can handle missing data
Handling of outliers	Relatively robust to outliers	Sensitive to outliers, especially when using the epsilon parameter in SVR. Preprocessing is often needed.	Robust to outliers due to ensemble learning and regularization. Extreme outliers may have limited impact.
Numerical Data	Can handle categorical features by encoding them. Works well with mixed data types.	Requires numerical data or extensive preprocessing (e.g., one-hot encoding).	Requires numerical data but can handle mixed data types with appropriate encoding.
Overfitting	Less prone to overfitting	May require regularization	Needs hyperparameter tuning for control
Computational complexity	Can be computationally expensive	Can be computationally expensive	Efficient and scalable
Used R packages	RandomForest	Kernlab, Caret	xgboost
Feature selection	Provides variable importance	Require separate feature selection	Provides variable importance

Correlation Plot



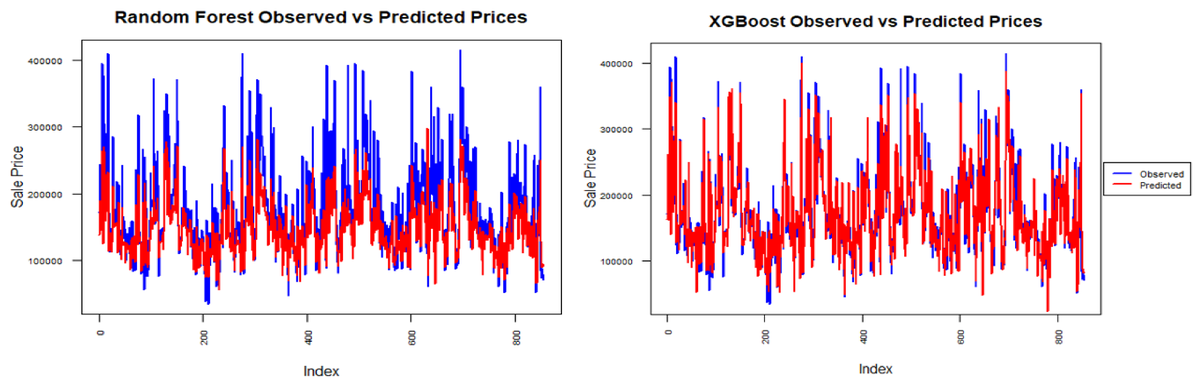
Top 40 Important Features



Performance Evaluation

	R^2	MSE	$RMSE$	MAE
Random Forest				
Full model	61%	1831068162	42790	29625
Reduced Model	62%	1776790450	42152	29317
Support Vector Regression				
Full model	89%	539307729	23223	15064
Reduced model	90%	466041744	21588	13945
XGBoost				
Full model	92%	413918676	20345	13115
Reduced model	92%	394419635	19860	13045

Performance Visuals of the low vs the high performing models



Conclusion

- Employing a 70:30 train-test data split on the Ames Housing dataset, this study demonstrated that selecting the top 40 features significantly enhances model performance, highlighting the importance of feature selection. Across all three techniques utilized for property sale price prediction, "Overall Quality" and "Ground Living Area" consistently emerged as the two most crucial features.
- However, the study acknowledges limitations due to the use of an outdated 2010 dataset, which lacks contemporary factors like transportation access and energy efficiency, along with essential economic indicators influencing housing prices. Nevertheless, it underscores the effectiveness of machine learning techniques, notably XGBoost, in accurately predicting house prices.

References

[1] A. Cutler, D. R. Cutler, and J. R. Stevens, *Random forests*, Ensemble Machine Learning: Methods and Applications, (2012), pp. 157–175.

[2] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, *Support vector regression machines*, Advances in Neural Information Processing Systems, 9 (1996), pp. 155–161.