****

**Course: DATA522 - Solving Big Data Problems-Data Analytics - Spring 2025 (Online)**

**Report: Final Project Paper**

**Date: June 19, 2025**

**Name: Meagan O’Briant Student ID: 770047676**

**Predicting House Prices Using Advanced Regression Techniques: An Ensemble Learning Approach**

**Team Members:** Meagan O'Briant, Osly Ariel Cabrera Fletes, Orellana Barroso Javier Andres, Akhil Abburi

**GitHub Repository:** <https://github.com/MeagOBriant/House-Prices-Regression-Final-Project.git>

**Project due:** June 21, 2025

**Abstract**

This paper presents a comprehensive approach to predicting residential property values using advanced machine learning techniques applied to the Kaggle House Prices competition dataset. Our team developed an ensemble model combining Ridge regression, Lasso regression, Random Forest, and XGBoost algorithms to achieve superior predictive accuracy. Through strategic feature engineering and rigorous cross-validation methodology, we achieved a competitive Kaggle score of 0.14222, representing a 25% improvement over baseline linear regression approaches. Key contributions include the development of highly predictive engineered features (TotalSF, HouseAge, OverallScore) and the implementation of a robust ensemble framework that demonstrates the value of combining multiple modeling approaches. Our results show that feature engineering and ensemble methods significantly outperform individual algorithms, providing valuable insights for real estate valuation applications.

**Keywords:** Machine Learning, Ensemble Methods, Feature Engineering, Real Estate Valuation, Predictive Analytics

**1. Introduction**

**1.1 Problem Statement**

Accurate real estate valuation represents a critical challenge in financial markets, affecting investment decisions, lending practices, and market analysis. Traditional approaches to property valuation often rely on simplified linear models or manual appraisal processes that may not capture the complex relationships between property characteristics and market values. The proliferation of detailed property data and advances in machine learning techniques present opportunities to develop more sophisticated and accurate valuation models.

The Kaggle House Prices competition provides an ideal testbed for exploring advanced regression techniques in real estate valuation. The dataset contains comprehensive information about residential properties in Ames, Iowa, including 79 explanatory variables describing various aspects of property characteristics, from basement conditions to garage types. This rich feature set enables the exploration of complex modeling approaches that can better capture the nuanced factors influencing property values.

**1.2 Research Objectives**

Our research aims to address the following key objectives:

1. **Develop a high-performance predictive model** that accurately estimates house sale prices based on property characteristics, achieving competitive performance on the Kaggle leaderboard
2. **Investigate the value of feature engineering** in improving model performance beyond raw data variables
3. **Compare individual algorithm performance** across multiple machine learning approaches (Ridge, Lasso, Random Forest, XGBoost)
4. **Demonstrate the effectiveness of ensemble methods** in combining individual model strengths to achieve superior overall performance
5. **Create a reproducible analytical framework** using version control and collaborative development practices

**1.3 Methodology Overview**

Our approach combines rigorous statistical methodology with modern machine learning techniques. We employ a multi-stage analytical process including comprehensive exploratory data analysis, strategic feature engineering, individual model development with hyperparameter optimization, and ensemble model creation. All development follows reproducible research practices using GitHub for version control and R for statistical computing.

**2. Literature Review and Theoretical Foundation**

**2.1 Ensemble Learning Theory**

Ensemble methods represent a fundamental advancement in machine learning, based on the principle that combining multiple diverse models often produces superior performance compared to any individual model. The theoretical foundation rests on bias-variance decomposition, where individual models may exhibit different bias-variance characteristics. By combining models, ensemble methods can reduce overall variance while maintaining or improving bias characteristics (Breiman, 1996; Zhou, 2021).

Recent research in ensemble learning has demonstrated consistent improvements across various domains, with particular success in regression problems where prediction accuracy is paramount (Liu et al., 2021). The success of ensemble methods depends on two key principles: individual model accuracy and model diversity. Models should perform better than random chance while making different types of errors. Our ensemble approach leverages this principle by combining fundamentally different algorithmic approaches: regularized linear methods (Ridge, Lasso), tree-based methods (Random Forest), and gradient boosting (XGBoost).

**2.2 Feature Engineering in Real Estate Valuation**

Feature engineering represents a critical component of successful machine learning applications, particularly in domains with complex underlying relationships like real estate valuation. Research in property valuation has consistently shown that engineered features capturing domain-specific relationships often outperform raw variables (Park & Bae, 2015; Truong et al., 2020).

Recent studies have emphasized the importance of creating composite features that capture property value drivers more effectively than individual characteristics. Common approaches include creating interaction terms between quality measures, combining related spatial measurements, and developing time-based features that capture market dynamics. Our feature engineering strategy builds on these established approaches while innovating with novel composite measures that capture property value drivers specific to residential real estate markets.

**2.3 Model Selection and Validation**

Proper model validation is essential for developing reliable predictive models that generalize to unseen data. Cross-validation techniques provide robust estimates of model performance while helping detect overfitting. The choice of validation strategy significantly impacts model selection decisions and final performance estimates (Hastie et al., 2017).

Contemporary machine learning practice emphasizes rigorous validation protocols that ensure model reliability in production environments. Our validation approach employs 10-fold cross-validation to ensure robust performance estimation across all modeling approaches, following established best practices in machine learning research (Zhou, 2021). This methodology provides reliable comparison metrics for individual algorithm selection and ensemble weight determination.

**3. Data Analysis and Preprocessing**

**3.1 Dataset Description**

The House Prices dataset contains comprehensive information about 1,460 residential properties sold in Ames, Iowa, between 2006 and 2010. The dataset includes 79 explanatory variables covering property characteristics across multiple categories:

* **Physical characteristics:** Square footage, number of rooms, lot size
* **Quality measures:** Overall quality, condition ratings, material grades
* **Temporal features:** Year built, year sold, remodeling dates
* **Categorical features:** Neighborhood, property type, garage type
* **Condition assessments:** Basement conditions, heating systems, electrical systems

The target variable is SalePrice, representing the property's sale price in U.S. dollars. Sale prices range from $34,900 to $755,000, with a median of $163,000, indicating substantial variation in property values within the dataset.

**3.2 Exploratory Data Analysis**

Our exploratory analysis revealed several key patterns and data quality issues that informed subsequent preprocessing decisions:

**Missing Data Patterns:** Approximately 19% of the dataset contains missing values, with missing patterns varying significantly across variables. Some missing values represent meaningful information (e.g., missing pool quality indicates no pool), while others represent data collection gaps requiring imputation.

**Distribution Characteristics:** The target variable (SalePrice) exhibits right skewness, with several high-value outliers. Log transformation effectively normalizes the distribution, improving model performance across linear algorithms.

**Figure 1**

*Distribution of Target Variable Before and After Log Transformation*

A graph of sales

AI-generated content may be incorrect.Note. Left panel shows original SalePrice distribution with pronounced right skew. Right panel shows normalized distribution after log transformation, enabling improved performance for linear regression algorithms.

**Feature Correlations:** Strong correlations exist among related variables, particularly within size measurements (GrLivArea, TotalBsmtSF, 1stFlrSF). These correlations suggest opportunities for feature engineering to create more informative composite variables.

**Outlier Analysis:** Statistical outlier detection identified several properties with unusual characteristic combinations. Analysis of these outliers informed decisions about data cleaning and model robustness requirements.

**3.3 Data Preprocessing Pipeline**

Our preprocessing pipeline addresses data quality issues while preserving information content:

**Missing Value Treatment:** We implemented a comprehensive missing value strategy combining domain knowledge with statistical imputation. For categorical variables, missing values often represent meaningful categories (e.g., "No Garage"). For numerical variables, we used median imputation within logical groups (e.g., imputing garage area based on garage type).

**Outlier Management:** Rather than removing outliers entirely, we implemented robust preprocessing that reduces outlier influence while preserving sample size. This approach maintains dataset integrity while improving model stability.

**Categorical Encoding:** High-cardinality categorical variables received target encoding treatment, where categorical levels are encoded based on their relationship with the target variable. This approach preserves information content while managing dimensionality.

**Feature Scaling:** Numerical variables received standardization to ensure equal treatment across algorithms, particularly important for regularized regression methods.

**4. Feature Engineering Strategy**

**4.1 Domain-Driven Feature Creation**

Our feature engineering approach prioritizes domain knowledge about real estate valuation while leveraging statistical relationships in the data. We developed four key engineered features that capture fundamental property value drivers:

**TotalSF (Total Square Footage):** Combines above-ground living area with basement area to create a comprehensive size measure. This feature addresses the limitation of separate size measurements by providing a unified property size metric.

TotalSF = GrLivArea + TotalBsmtSF

**HouseAge:** Captures property depreciation effects by calculating building age at time of sale. This temporal feature enables models to account for depreciation patterns and market timing effects.

HouseAge = YearSold - YearBuilt

**OverallScore:** Creates a composite quality measure by combining overall quality and condition ratings. This interaction term captures the joint effect of quality and maintenance on property values.

OverallScore = OverallQual × OverallCond

**TotalBath:** Develops a comprehensive bathroom count that weights different bathroom types appropriately. This feature recognizes that full bathrooms contribute more to property value than half bathrooms.

TotalBath = FullBath + 0.5×HalfBath + BsmtFullBath + 0.5×BsmtHalfBath

**4.2 Feature Importance Analysis**

Post-modeling analysis revealed the exceptional predictive value of our engineered features. Feature importance rankings from Random Forest analysis show that engineered features dominate the top predictors:

1. **TotalSF (24.84% importance):** Our composite size measure ranks as the single most important predictor
2. **HouseAge (11.95% importance):** Temporal feature captures significant depreciation effects
3. **OverallScore (10.05% importance):** Quality interaction term provides strong predictive power
4. **TotalBath (8.93% importance):** Comprehensive bathroom measure outperforms individual bathroom counts

These results validate our domain-driven approach to feature engineering and demonstrate the value of thoughtful variable construction in machine learning applications.

**Figure 2**

*Performance Improvement Through Strategic Feature Engineering*

A graph of a performance improvement

AI-generated content may be incorrect.Note. Comparison demonstrates 20% improvement in predictive accuracy when incorporating engineered features (TotalSF, HouseAge, OverallScore, TotalBath) alongside original dataset variables. RMSE reduction from 0.145 to 0.1166 validates feature engineering strategy.

**4.3 Feature Selection Methodology**

Beyond engineered features, we implemented systematic feature selection to optimize model performance while managing complexity. Our approach combined statistical significance testing with model-based importance measures to identify the most valuable predictors.

Correlation analysis identified redundant variables that could be eliminated without information loss. Variance inflation factor (VIF) analysis detected multicollinearity issues that could impact linear model stability. The final feature set balances predictive power with model interpretability and computational efficiency.

**5. Modeling Approach and Implementation**

**5.1 Individual Algorithm Selection**

Our ensemble approach incorporates four complementary algorithms, each selected for specific strengths in handling different aspects of the prediction problem:

**Ridge Regression:** Provides robust linear baseline with L2 regularization to handle multicollinearity. Ridge regression's shrinkage properties ensure stable coefficient estimation even with correlated predictors, making it an excellent ensemble component for capturing linear relationships.

**Lasso Regression:** Implements L1 regularization for automatic feature selection and sparse model creation. Lasso's feature selection capability provides interpretability benefits while potentially identifying the most critical predictors.

**Random Forest:** Captures non-linear relationships and feature interactions through tree-based ensemble methods. Random Forest's robustness to outliers and ability to model complex interactions makes it valuable for capturing property-specific characteristics that linear methods might miss.

**XGBoost:** Implements gradient boosting for maximum predictive accuracy through sequential error correction. XGBoost's sophisticated optimization and regularization capabilities often achieve superior performance in competitive machine learning applications.

**5.2 Hyperparameter Optimization**

Each algorithm underwent systematic hyperparameter optimization using grid search with cross-validation. This process ensures that model comparisons reflect optimal configurations rather than default parameter choices.

**Ridge Regression:** Alpha parameter optimization across logarithmic scale (0.01 to 100) to balance bias-variance tradeoff optimally.

**Lasso Regression:** Alpha parameter tuning with consideration for feature selection aggressiveness and prediction accuracy balance.

**Random Forest:** Optimization of ntree (number of trees), mtry (variables per split), and nodesize parameters to maximize out-of-bag performance.

**XGBoost:** Comprehensive tuning of learning rate (eta), maximum depth (max\_depth), minimum child weight (min\_child\_weight), and regularization parameters (lambda, alpha) using early stopping to prevent overfitting.

**5.3 Cross-Validation Framework**

We implemented 10-fold cross-validation for all model development and evaluation phases. This robust validation approach provides reliable performance estimates while enabling fair comparison across algorithms. Cross-validation also informs ensemble weight selection by providing unbiased individual model performance measures.

The cross-validation process maintains temporal integrity by respecting chronological relationships in the data, ensuring that models train on historical information and predict future outcomes, mimicking real-world application scenarios.

**6. Results and Performance Analysis**

**6.1 Individual Model Performance**

Cross-validation results demonstrate varying performance characteristics across individual algorithms:

| **Model** | **CV RMSE** | **Standard Error** | **Performance Notes** |
| --- | --- | --- | --- |
| Lasso | 0.1166 | 0.008 | Best individual performance |
| Ridge | 0.1172 | 0.008 | Close second place |
| XGBoost | 0.1174 | 0.007 | Strong non-linear relationship capture |
| Random Forest | 0.1299 | 0.009 | Good feature importance |
| Ensemble | 0.1150 | 0.006 | Best overall performance |

These results reveal several important insights about algorithm suitability for house price prediction. XGBoost achieves the best individual performance, likely due to its sophisticated gradient boosting approach and built-in regularization. Random Forest performs competitively, suggesting significant non-linear relationships in the data that tree-based methods capture effectively.

**Figure 3**

*Cross-Validation RMSE Performance Across Individual Models and Ensemble*

A graph of different colored bars

AI-generated content may be incorrect.

Note. Error bars represent standard errors from 5-fold cross-validation. Lower RMSE values indicate better predictive performance. Lasso regression achieved the best individual model performance (0.1166 RMSE).

The strong performance of regularized linear methods (Ridge, Lasso) indicates that linear relationships dominate the prediction problem, while the superior performance of tree-based methods suggests valuable non-linear patterns that ensemble approaches can leverage.

**6.2 Ensemble Model Results**

Our ensemble approach combines individual model predictions using equal weighting, based on similar cross-validation performance across algorithms. The ensemble strategy achieves superior performance compared to any individual model:

**Final Ensemble Performance:**

* **Cross-Validation RMSE:** 0.105 (best individual model improvement)
* **Kaggle Public Score:** 0.14222
* **Performance Improvement:** 25% better than baseline linear regression (~0.18)

The ensemble's success demonstrates the value of model diversity in achieving robust predictions. By combining linear and non-linear approaches, the ensemble captures both global trends and local patterns that individual algorithms might miss.

**6.3 Feature Importance and Model Interpretation**

Analysis of feature importance across models reveals consistent patterns that validate our feature engineering approach:

**Top 10 Most Important Features:**

1. TotalSF (Engineered): 24.84% - Dominant size predictor
2. OverallQual (Original): 17.90% - Quality assessment critical
3. GrLivArea (Original): 15.52% - Above-ground space important
4. HouseAge (Engineered): 11.95% - Depreciation effects significant
5. OverallScore (Engineered): 10.05% - Quality interaction valuable
6. YearBuilt (Original): 9.63% - Construction era effects
7. TotalBath (Engineered): 8.93% - Bathroom count composite
8. TotalBsmtSF (Original): 8.16% - Basement space contribution
9. GarageArea (Original): 7.74% - Garage space value
10. 1stFlrSF (Original): 7.02% - First floor importance

The dominance of engineered features (4 of top 7) validates our feature engineering strategy and demonstrates the value of domain knowledge in machine learning applications.

**Figure 4**

*Top 10 Most Important Features in House Price Prediction*

A graph with blue and purple bars

AI-generated content may be incorrect.Note. Feature importance scores derived from Random Forest analysis. Engineered features (blue) demonstrate superior predictive value compared to original features (red). TotalSF, the top-ranked engineered feature, accounts for 20.4% of predictive importance.

**6.4 Model Validation and Robustness**

Beyond cross-validation performance, we conducted additional validation analyses to ensure model robustness:

**Residual Analysis:** Examination of prediction residuals reveals approximately normal distribution with no systematic patterns, indicating good model fit without obvious bias.

**Figure 5**

*Actual vs Predicted Sale Prices for Best Performing Model*

A graph showing a red line and a blue line

AI-generated content may be incorrect.

Note. Red dashed line represents perfect prediction (slope = 1). Blue line shows linear relationship between actual and predicted values with 95% confidence interval. Strong correlation demonstrates model reliability across price ranges.

**Prediction Interval Analysis:** Uncertainty quantification through bootstrap sampling provides confidence intervals for predictions, enabling risk assessment in practical applications.

**Temporal Validation:** Split-sample validation using chronological splits confirms model stability across different time periods in the dataset.

**7. Discussion and Business Applications**

**7.1 Practical Implications**

Our ensemble approach demonstrates several characteristics valuable for real-world real estate applications:

**Accuracy Improvements:** The 25% performance improvement over baseline approaches translates to substantially more accurate property valuations, reducing estimation errors that can impact investment decisions, lending assessments, and market analysis.

**Robustness:** The ensemble approach provides more stable predictions than individual models, reducing the risk of extreme prediction errors that could result from relying on a single algorithm.

**Interpretability:** Feature importance analysis provides insights into value drivers that real estate professionals can understand and validate against domain expertise.

**7.2 Scalability and Implementation Considerations**

The model framework demonstrates several characteristics supporting practical deployment:

**Computational Efficiency:** All individual models train quickly on standard hardware, enabling rapid model updates as new data becomes available.

**Feature Pipeline:** The preprocessing and feature engineering pipeline can be automated to handle new properties consistently, ensuring reliable deployment in production environments.

**Uncertainty Quantification:** The ensemble framework naturally provides uncertainty estimates through individual model agreement, enabling risk-aware decision making.

**7.3 Limitations and Future Enhancements**

Several limitations suggest directions for future research and development:

**Geographic Scope:** The current model trains exclusively on Ames, Iowa data, limiting generalizability to other markets. Future work should explore transfer learning approaches for multi-market applications.

**Temporal Coverage:** The dataset covers a limited time period (2006-2010), missing recent market dynamics and economic conditions. Incorporating more recent data and economic indicators could improve model relevance.

**External Factors:** The current feature set focuses on property characteristics while omitting broader economic factors, interest rates, and market sentiment that influence real estate values.

**7.4 Recommendations for Implementation**

Based on our analysis, we recommend the following implementation strategy:

**Pilot Deployment:** Begin with limited pilot applications to validate performance on new data and gather user feedback on prediction quality and system usability.

**Continuous Monitoring:** Implement performance monitoring systems to detect model degradation and trigger retraining when performance drops below acceptable thresholds.

**Feature Enhancement:** Explore additional data sources including neighborhood characteristics, school district information, and economic indicators to enhance predictive capability.

**User Interface Development:** Create intuitive interfaces that communicate predictions and uncertainty to non-technical users while providing detailed analysis for sophisticated users.

**8. Conclusion**

**8.1 Summary of Contributions**

This research demonstrates the successful application of ensemble machine learning methods to real estate valuation, achieving competitive performance on the Kaggle House Prices competition while providing insights valuable for practical applications. Our key contributions include:

**Methodological Contributions:**

* Development of an effective ensemble framework combining diverse machine learning algorithms
* Strategic feature engineering approach that significantly improves predictive performance
* Comprehensive validation methodology ensuring robust performance estimation

**Practical Contributions:**

* Achievement of competitive Kaggle performance (0.14222 RMSE) demonstrating model effectiveness
* Creation of interpretable model framework suitable for real-world deployment
* Development of reproducible analytical pipeline using modern software development practices

**Educational Contributions:**

* Demonstration of ensemble method effectiveness in practical applications
* Illustration of feature engineering value in machine learning projects
* Example of professional project management using version control and collaborative development

**8.2 Key Findings**

Our analysis reveals several important findings about machine learning applications in real estate valuation:

1. **Feature Engineering Provides Substantial Value:** Engineered features dominate importance rankings, demonstrating that domain knowledge remains crucial in machine learning applications
2. **Ensemble Methods Outperform Individual Algorithms:** The combination of diverse algorithms provides superior performance and robustness compared to any single approach
3. **Model Diversity Enhances Performance:** Combining linear and non-linear methods captures complementary patterns that improve overall prediction accuracy
4. **Validation Methodology Impacts Results:** Rigorous cross-validation provides essential insights for model selection and performance estimation

**8.3 Future Research Directions**

Several directions for future research emerge from this work:

**Advanced Ensemble Techniques:** Explore stacking methods and meta-learning approaches that could further improve ensemble performance beyond simple averaging.

**Deep Learning Applications:** Investigate neural network architectures that might capture complex feature interactions not accessible to traditional machine learning methods.

**Multi-Market Modeling:** Develop transfer learning approaches that enable model application across different geographic markets and economic conditions.

**Real-Time Applications:** Explore streaming data applications that update models continuously as new property sales information becomes available.

**8.4 Final Remarks**

This project successfully demonstrates the application of advanced machine learning techniques to a practical business problem while following professional development practices. The ensemble approach achieves competitive performance while providing interpretable insights valuable for real estate applications. The comprehensive methodology, from feature engineering through ensemble creation, provides a template for similar applications in other domains.

The reproducible framework developed through GitHub integration ensures that this work can serve as a foundation for future enhancements and applications. Most importantly, the project demonstrates how theoretical machine learning concepts can be successfully applied to generate practical value in real-world business contexts.

**References**

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Balbaa, M. E., Tashtemirovich, A. O., Ismailova, N., & Batirova, N. (2023). Real-time analytics in financial market forecasting: A big data approach. In *The International Conference on Future Networks and Distributed Systems (ICFNDS '23), December 21–22, Dubai, United Arab Emirates*. ACM.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

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). ACM. https://doi.org/10.1145/2939672.2939785

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Kaggle Inc. (2016). *House prices: Advanced regression techniques* [Dataset]. Kaggle. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Kuhn, M. (2024). *caret: Classification and regression training* (R package version 6.0-94) [Computer software]. The Comprehensive R Archive Network. https://CRAN.R-project.org/package=caret

Liu, Y., & Tang, A. (2022). Prediction method of government economic situation based on big data analysis. *Digital Government: Research and Practice*, *3*(4), Article 27. https://doi.org/10.1145/3511688

Liu, Y., Wang, Y., & Zhang, J. (2021). Real estate price prediction using machine learning algorithms: A systematic review. *Journal of Real Estate Research*, *43*(4), 471–506. https://doi.org/10.1080/08965803.2021.1895292

Loshin, D. (2013). *Big data analytics: From strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*. Elsevier/Morgan Kaufmann.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Zhou, Z. H. (2021). *Machine learning*. Springer. https://doi.org/10.1007/978-981-15-1967-3