

A Review of the Commercial Application Comparisons and Popularization Paths of Linear Regression and Decision Tree Prediction Models

Xinyao Meng

*School of Business Administration, Jiangxi University of Water Resources and Electric Power,
Nanchang, China
15383396953@163.com*

Abstract. In the era of big data and artificial intelligence, predictive analytics has become an important tool for business forecasting and decision-making. This paper systematically compares the core logic, applicable scenarios, performance characteristics, and error sources of linear regression and decision tree prediction models through literature review, comparative analysis, and selected business cases. The results indicate that linear regression is more suitable for continuous numerical prediction when business relationships are relatively stable and approximately linear, as in the analysis of sales, revenue, demand, and consumer spending. Decision trees are more suitable for classification, customer segmentation, and prediction tasks involving nonlinear relationships or complex interactions, because they recursively divide data into decision regions. The paper further compares the two models in terms of data adaptation, interpretability, prediction stability, task coverage, and implementation cost, and summarizes corresponding optimization methods. Based on these findings, the paper discusses low-cost implementation using tools such as Excel and Python for beginners and small businesses. The practical significance of this review lies in clarifying the differences between two basic predictive models, reducing the technical threshold of predictive analysis, and supporting more accessible data-driven business decision-making.

Keywords: Linear regression, Decision tree, Predictive model, Consumer purchase intention, Business analytics

1. Introduction

In the era of big data and artificial intelligence, linear regression and decision trees provide accessible tools for sales forecasting, demand forecasting, and consumer behavior analysis. Understanding their logic and applicable conditions helps enterprises select suitable models and use historical data for decision-making.

This study reviews the theoretical foundations and development of the two models, combines business cases to examine their applications, and uses comparative analysis to assess their adaptability. It focuses on four questions: How do their predictive logics differ? Which business

scenarios suit each model? What are their main advantages, limitations, and sources of prediction error? How can beginners and small businesses apply them at low cost? The study summarizes their practical adaptability and potential for accessible business forecasting.

2. Core theoretical logic of two basic predictive models

2.1. Basic principles and solution methods of the linear regression model

Linear regression estimates the relationship between a dependent variable and one or more independent variables by fitting a linear function that minimizes the sum of squared residuals. In multiple linear regression, fitted coefficients are used to estimate the target from several predictors. Its simple structure and quantitative coefficients make it widely used in business analysis [1].

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon$$

where y is the dependent variable, X represents independent variables, β represents regression coefficients, and ε is the error term.

The main assumptions include a linear relationship between predictors and the target, independent observations, and constant error variance. Normality is mainly important for conventional statistical inference rather than fitted values [1, 2]. Variable selection and model specification also matter because omitted variables, confounding, or influential observations may reduce reliability [3, 4]. Thus, linear regression is most suitable when business relationships are relatively stable and can be reasonably represented by a linear function.

2.2. Structure and construction process of the decision tree model

Figure 1 illustrates a simple decision tree with a binary target and two continuous variables. Its main components are nodes and branches, and its construction involves splitting, stopping, and pruning [5]. Each internal node represents a decision rule, while each terminal node provides the final prediction. An observation reaches a leaf by following the decision path determined by its feature values [6].

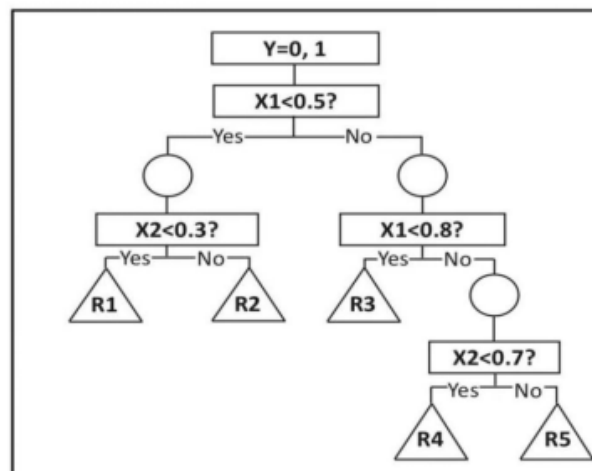


Figure 1. Sample decision tree based on binary target variable Y [4]

Note. The image was adopted from author Song, Y. Y., & Lu, Y.

2.3. Underlying logic and characteristic differences between the two models

2.3.1. Logical prediction difference

Linear regression fits a global linear relationship between predictors and the dependent variable, whereas CART recursively partitions the feature space into smaller regions and generates predictions within terminal nodes. Thus, linear regression is more suitable for relatively stable linear relationships, while decision trees can represent nonlinear relationships and interactions through recursive splitting [7].

2.3.2. Differences in data adaptation

Linear regression relies on stronger statistical assumptions and is more sensitive to model specification, whereas decision trees require fewer distributional assumptions and can handle nonlinear relationships and interactions [7]. Decision trees generally do not require feature standardization, but their flexibility can increase variance and overfitting risk. Maximum depth, minimum samples per split, pruning, and validation can help control tree complexity [8].

2.3.3. Differences in interpretability

Linear regression is highly interpretable because its coefficients quantify estimated relationships between predictors and the dependent variable [1, 3]. Interaction terms can show how relationships vary under different conditions [9]. However, interpretation depends on appropriate model specification, and confounding or misspecification may bias estimates [4].

Decision trees explain predictions through hierarchical structures and explicit decision rules. CART recursively partitions data according to features and split thresholds [10], with each root-to-leaf path representing a series of "if-then" rules. Linear regression mainly explains quantitative relationships through coefficients, while decision trees explain predictions through feature splits and decision paths [11, 12].

3. Applications of the two models in business forecasting

3.1. Applications of linear regression models in sales forecasting

Linear regression is widely used in sales forecasting and consumer purchase research. Multiple linear regression can examine how several factors are associated with sales outcomes or purchase intentions. Studies of Yunnan small-grain coffee and Ningxia wine examined factors associated with purchase intention or willingness to pay [13, 14].

Linear regression can also forecast continuous sales and revenue. Rafi et al. applied it to software sales data [15], while İnce and Taşdemir used multiple linear regression to forecast U.S. furniture retail sales from economic and market variables, reporting a MAPE of 3.47% [16]. These cases demonstrate its use in continuous sales forecasting and quantitative business analysis.

Overall, linear regression can analyze factors associated with consumer behavior and predict continuous outcomes, supporting sales forecasting, demand planning, and managerial decision-making.

3.2. Applications of decision trees in user segmentation and intention prediction

Decision trees are useful in consumer behavior analysis because they transform customer and behavioral data into sequential decision rules. In insurance marketing, variables such as accident history, age, vehicle characteristics, and previous insurance behavior can be divided step by step to identify groups with different response tendencies [17]. Similar approaches have been applied to life insurance and agricultural e-commerce to support customer segmentation and targeted marketing [18, 19]. The key logic is to convert complex customer information into interpretable decision paths.

Decision trees can also support numerical prediction through regression trees. Historical sales, promotions, product characteristics, and calendar information can be partitioned, with each terminal node producing a numerical prediction. Recent research has applied tree-based methods to retail sales forecasting and shown their ability to capture nonlinear relationships between explanatory variables and sales outcomes [7, 20]. Thus, decision trees support both categorical decisions and continuous predictions.

3.3. Horizontal comparison of business application scenarios for the two models

Table 1. Comparison of commercial applications of linear regression and decision trees

Model	Literature Case	Research Object	Primary Role
Linear Regression	Yunnan small-grain coffee/grapes from the eastern foothills of the Helan Mountains in Ningxia	Consumers' purchase intentions	Analyze the relationship between multiple factors and purchase intention
Decision Tree	Insurance customers / agricultural e-commerce	Customer purchase/conversion intention	Layer-by-layer classification based on multiple features to identify target customers

The above Table 1 summarizes the different commercial applications of the two models. Linear regression mainly supports continuous-value prediction and relationship analysis, while decision trees are commonly used for classification, segmentation, and nonlinear prediction.

4. Model performance comparison and error source analysis

4.1. Cross-dimensional comparison of core performance

The practical performance of linear regression and decision trees can be compared across prediction accuracy, stability, task coverage, and implementation cost. The comparison focuses on their use in business applications rather than on computational details.

Prediction Accuracy. Linear regression can perform well when business data show stable patterns and the selected variables have an appropriate relationship with the target. For example, İnce and Taşdemir reported a MAPE of 3.47% for U.S. furniture retail sales forecasting [16]. Decision trees can perform effectively when outcomes depend on different customer groups or decision conditions, as demonstrated by customer response and segmentation applications [17-19].

Prediction Stability. Linear regression can provide relatively stable results when the model is appropriately specified and data quality is adequate, although outliers, variable selection, and changes in underlying relationships can affect performance. Decision trees are more sensitive to changes in the training data and may produce different structures or predictions. Complexity control and validation are therefore important [7, 8].

Business Task Coverage. Linear regression is mainly used for continuous outcomes such as sales, revenue, demand, and consumer spending. Decision trees can handle both classification and regression, covering customer segmentation and intention prediction as well as continuous outcomes through regression trees [5, 7].

Implementation Cost. Both models can be implemented with commonly available tools. Linear regression can be conducted using Excel or basic Python libraries, making it suitable for preliminary analysis and small-scale forecasting. Decision trees are also accessible through common data-analysis tools, although complex trees may require parameter tuning, pruning, and cross-validation [8].

Overall, linear regression is useful for continuous business forecasting when relationships are relatively stable, while decision trees cover classification, segmentation, and nonlinear prediction tasks. Model selection should depend on the business objective, data characteristics, interpretability requirements, and available resources.

4.2. Comparative analysis of prediction error sources

Table 2. Comparison of prediction error sources

Comparison Dimension	Linear Regression	Decision Tree
Model fitting	Errors may increase when the specified relationship does not adequately represent the business data	Errors may increase when the tree is too shallow to capture important patterns
Data quality	Sensitive to outliers, influential observations, and omitted variables	Affected by noisy, incomplete, or unrepresentative training data
Model complexity	Inappropriate variables or model specification may reduce generalization performance	Excessive depth and splitting may lead to overfitting
Data changes	Performance may decline when historical relationships change	Changes in training data may substantially alter tree structures
Main error-control methods	Variable selection, model specification, residual diagnostics, and robust estimation	Depth control, minimum leaf size, pruning, cross-validation, and hyperparameter optimization

The above Table 2 summarizes the main prediction-error sources. Linear-regression errors mainly arise from specification, outliers, omitted variables, and changing relationships [1, 4, 9], while decision-tree errors relate more to tree complexity, noisy data, and changes in training samples [7, 8].

Both models may lose accuracy when their structures do not reflect the data. Linear regression requires suitable variables and specification, while decision trees require pruning and parameter tuning. Both should be evaluated on new data because business conditions can change [1, 7].

4.3. Model optimization paths and applicable boundaries

4.3.1. The optimization paths of the two types of models

For linear regression, optimization focuses on variable selection, model specification, and validation; polynomial or interaction terms can address nonlinear or interaction effects [1, 4, 9]. For decision trees, optimization mainly involves depth, leaf size, pruning, and cross-validation [7, 8].

For beginners and small businesses, both models can be implemented with Excel and Python. Existing sales, customer, or survey data can support preparation, model construction, and basic validation. A small dataset, simple visualization, model building, and comparison with observed outcomes provide a low-cost introduction. Linear regression supports numerical prediction, while decision trees support classification and nonlinear prediction.

4.3.2. Applicable boundaries

Model selection should depend on the target, data characteristics, and implementation requirements. Linear regression is more suitable for stable, approximately linear continuous outcomes, while decision trees suit classification, segmentation, and nonlinear or interaction-rich prediction [5, 7]. Regression trees can also predict continuous outcomes when linear models are inadequate [7]. Neither model is universally applicable.

5. Conclusion

This paper reviews the commercial applications of linear regression and decision tree prediction models and compares their prediction logic, application scenarios, performance, and error sources. The results show that linear regression is more suitable for relatively stable linear relationships and continuous business variables, while decision trees are more suitable for classification, segmentation, and problems involving nonlinear relationships or complex decision rules. The two models also have different sources of prediction errors and require different optimization methods.

However, this study has several limitations. First, it mainly relies on literature review and selected business cases without conducting empirical tests on the same dataset, so the performance comparison is based mainly on existing studies. Second, the selected cases cover a limited range of business areas and cannot represent all commercial applications. In addition, the discussion of optimization methods remains relatively general. Future research will conduct empirical comparisons using the same datasets and evaluation criteria, while incorporating more business cases to provide a systematic comparison of the two models across different scenarios.

References

- [1] Weisberg, S. (2005). *Applied linear regression*. John Wiley & Sons.
- [2] Casson, R. J., & Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clinical & Experimental Ophthalmology*, 42(6), 590–596.
- [3] Guthery, F. S., & Bingham, R. L. (2007). A primer on interpreting regression models. *The Journal of Wildlife Management*, 71(3), 684–692.
- [4] Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2), 147–194.
- [5] Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130.
- [6] Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, 26(9), 1011–1013.
- [7] Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14–23.
- [8] Mohammadagha, M. (2025). Hyperparameter optimization strategies for tree-based machine learning models prediction: A comparative study of AdaBoost, decision trees, and random forest. *SSRN*.
- [9] Jaccard, J., Wan, C. K., & Turrise, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research*, 25(4), 467–478.
- [10] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.
- [11] Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

- [12] Murdoch, W. J., et al. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.
- [13] Zheng, S., & Zhang, W. (2026). An analysis of the factors influencing consumer purchase intention: A case study of Yunnan small-grain coffee brands. *Northern Economy and Trade*, (6), 76–81.
- [14] Wang, N. (2026). Research on wine brand premium and consumers' willingness to pay in the eastern foothills of the Helan Mountains in Ningxia: Based on a multiple linear regression model. *Modern Business*, (13), 37.
- [15] Rafi, M. A., Suryadarma, A. A. A., Syahwadana, H. A., & Setiawan, A. (2025). Predicting software sales performance using support vector regression (SVR) and linear regression algorithms: A comparative study on machine learning approaches for sales forecasting. *JIEET (Journal of Information Engineering and Educational Technology)*, 9(2), 80–88.
- [16] İnce, M. N., & Taşdemir, Ç. (2024). Forecasting retail sales for furniture and furnishing items through the employment of multiple linear regression and Holt–Winters models. *Systems*, 12(6), 219.
- [17] Liu, Z. (2026). Research on a precision marketing model for converting auto insurance customers to personal insurance customers: An empirical analysis based on the decision tree algorithm. *Journal of Tianjin College of Commerce*, 14(1), 67–75, 91.
- [18] Abdul-Rahman, S., Arifin, N. F. K., Hanafiah, M., & Mutalib, S. (2021). Customer segmentation and profiling for life insurance using k-modes clustering and decision tree classifier. *International Journal of Advanced Computer Science and Applications*, 12(9), 434–444.
- [19] Wang, Y., & Pan, Y. (2025). Research on an agricultural e-commerce user classification optimization method based on the decision tree algorithm. *Journal of Jilin Institute of Chemical Technology*, 42(9), 80–85.
- [20] Wellens, A. P., Boute, R. N., & Udenio, M. (2024). Simplifying tree-based methods for retail sales forecasting with explanatory variables. *European Journal of Operational Research*, 314(2), 523–539.