**AN APPLICATION OF REGRESSION AND CLASSIFICATION METHODS AS DATA MINING TECHNIQUES FOR VERSA: INTELLIGENT OPERATIONS AND SALES INSIGHT SYSTEM FOR JINYI IMPORT & EXPORT COMPANY — LASANG BRANCH**

**A picture containing logo

Description automatically generated**

**DAVAO DEL NORTE STATE COLLEGE**

New Visayas, Panabo City

Institute of Computing

A Data Mining Project Submitted by:

Clarisse Nicole B. Balbin

Kyle James G. Bautista

Hubert Eleazar M. Debalucos

Alyssa Kristine A. Moreno

**May 2025**

**I. PROJECT OVERVIEW**

Many machinery companies face challenges such as supply chain volatility, cost pressures, and skilled labor shortages due to the lack of AI-integrated enterprise systems, globally (Parthasarathy, Vetter, Koerber, Flood & Min, 2024). It was also supported by Annata’s (2024) whitepaper which stresses that without AI-embedded enterprise systems, companies often experience missed optimization opportunities, inaccurate decision-making due to lack of real-time analytics, and losing competitive advantage. In the Philippines, only 14.9% of Philippine firms use AI technologies, with adoption concentrated in urban areas and larger firms, which indicates a major gap in AI-integrated enterprise systems among Philippine companies, including those in equipment sales and management (Quimba, Moreno & Salazar, 2024). In the context of Jinyi – Lasang branch, the company currently relies on manual processes and Excel spreadsheets for sales, logistics, and inventory management. This legacy system results in inefficiencies, delays, and errors which is also supported by the study of Jala, Gaid, Nacalaban, Sandinao and Ventayen (2024) that the lack of a AI-integrated centralized enterprise system hinders productivity and decision-making which also leads to operational inefficiencies primarily due to inadequate data governance, poor data quality, and the absence of specialized AI skills within organizations. To address these challenges and bridge the technological gap in its operations, Jinyi – Lasang Branch is set to implement an AI-integrated centralized enterprise system tailored to its unique business needs.

Versa: Intelligent Operations and Sales Insight System for Jinyi Import & Export Company — Lasang Branch is an integrated web-based, AI-driven enterprise system designed for Jinyi Import & Export Co., Inc. – Lasang Branch. The project aims to automate and optimize sales, delivery, after-sales, motor pool services, and warehouse operations by integrating predictive analytics, risk and anomaly detection, and business intelligence. Particularly, the sales department is in charge in creating orders and invoices for heavy equipment and machinery, the motor pool is for pre-delivery inspection of heavy equipment, delivery is for delivering products, warehouse is for selling spare parts, and after sales is for after sales services covered by warranty. Moreover, Versa contains features such as automated invoice generation, role-based access for departmental users, centralized dashboards for operational transparency, and API integration for seamless data exchange. Furthermore, it incorporates sales prediction models, anomaly detection mechanisms, and demand spike alerts.

This capstone project is specifically applied for the heavy equipment trading industry, focusing on improving operational processes such as logistics, inventory, sales, and after-sales services. By integrating data mining and machine learning techniques, Versa transforms manual, spreadsheet-based workflows into an intelligent, automated system that enhances efficiency, accuracy, and decision-making. The system introduces predictive analytics for sales forecasting, real-time anomaly detection, and dynamic inventory management, making it a vital tool for businesses seeking digital transformation. Furthermore, the project aligns with Sustainable Development Goal (SDG) 9: Industry, Innovation, and Infrastructure, as it promotes resilient infrastructure and fosters innovation through the adoption of AI-driven systems (United Nations Department of Economic and Social Affairs, 2015). By modernizing Jinyi’s internal operations, Versa contributes to more sustainable, efficient, and technologically advanced industrial practices.

**II. ROLE OF DATA MINING**

Data mining, as described by Chen, Han, and Yu (1996), is the process of discovering meaningful patterns, trends, and knowledge from large volumes of data stored in databases, data warehouses, or other information repositories. From a database perspective, it integrates techniques from statistics, machine learning, and database systems to extract hidden, previously unknown, but potentially useful information. Data mining enables the automated analysis of massive datasets by identifying relationships, classifications, associations, and anomalies that support decision-making. In enterprise environments—such as heavy equipment sales and operations—data mining plays a crucial role in transforming raw transactional data into actionable insights, ultimately improving efficiency, reducing risks, and supporting strategic planning where it provides automation, anomaly detection, inventory forecasting, and demand prediction, all vital for data-driven management.

Versa utilizes a comprehensive set of datasets to support three core functions critical to business operations: sales forecasting, demand spike prediction, and anomaly detection. These datasets include historical and current sales records, which provide valuable insights into purchasing behavior, long-term sales trends, and seasonal fluctuations. By analyzing these patterns, Versa can accurately forecast future sales, enabling the company to anticipate customer needs, optimize inventory levels, and plan resources more effectively. Inventory logs are also integrated, capturing real-time data on stock levels, item movements, and availability, which helps the system anticipate supply and demand imbalances before they affect operations. For demand spike prediction, Versa draws on a wide range of inputs—such as marketing activity, external events, and transactional patterns—to detect and respond to sudden increases in product demand. This ensures timely restocking, minimizes the risk of stockouts, and improves customer satisfaction. Additionally, the system uses invoice and anomaly logs to monitor and detect irregularities, such as unexpected pricing changes or suspicious transactions, enhancing both financial accuracy and operational security. Versa’s ability to continuously update with real-time data allows it to adapt to changing market conditions and evolving customer behavior. Together, these data sources fuel the platform’s machine learning models and analytics engines, transforming raw operational data into precise, actionable insights. As a result, Versa empowers Jinyi – Lasang Branch to make more informed, data-driven decisions, streamline operations, improve risk management, and ultimately enhance overall business performance.

**III. DATA MINING TECHNIQUES USED**

The data mining techniques employed in this capstone project primarily involve regression and classification methods, which form the foundation for the development of predictive algorithms tailored to specific business needs of Jinyi – Lasang branch. Regression techniques are utilized to model and forecast continuous variables (Sebt, Sadati-Keneti, Rahbari, Gholipour & Mehri, 2024), such as future sales volumes, by analyzing historical and current sales data and relevant influencing factors. Classification methods, on the other hand, are applied to categorize data into distinct groups based on patterns and features, enabling the prediction of events like demand spikes or identification of anomalies (Kesavaraj & Sukumaran, 2013).

**Regression Technique**

Regression techniques are particularly well-suited for Versa’s sales forecasting and anomaly detection due to their powerful ability to model relationships between variables, identify trends, and generate accurate predictions from historical data. For sales forecasting, regression analysis captures and quantifies the impact of key factors such as marketing spend, pricing, seasonality, and external economic indicators on sales performance. This enables businesses to understand what drives sales fluctuations and how changes in these factors may influence future revenue. By fitting regression models to past sales data, Versa can produce highly accurate forecasts, supporting informed decisions on inventory management, resource allocation, and budgeting, while minimizing risks of stock out or overstocking. Additionally, regression models incorporate time-based variables to detect seasonal patterns and long-term trends, which are critical for effective planning around peak and slow sales periods (FasterCapital, 2025; Lieberman, 2023). For anomaly detection, regression models learn the expected patterns of normal behavior—such as typical sales volumes—and flag significant deviations as anomalies, enabling early identification of unusual activities or errors. These methods can handle complex, high-dimensional data and adapt to real-time monitoring, making them scalable and practical for various operational contexts (Hu, Gao, Li, Wu, Du & Maybank, 2018). Overall, regression empowers Versa to deliver precise, data-driven insights that enhance forecasting accuracy, detect irregularities, and improve decision-making efficiency.

**Classification Technique**

Classification techniques are well-suited for predicting demand spikes because they can effectively categorize periods or events into distinct classes, such as "spike" versus "no spike," based on patterns learned from both historical and real-time data. These models excel at recognizing complex signals that often precede demand spikes, including sudden changes in sales, social media trends, or external factors like weather and events. Since demand spikes are typically treated as discrete events, classification naturally fits the task by predicting categorical outcomes rather than continuous values (Cogent Infotech, 2024; AltexSoft, 2022). Additionally, classification algorithms can integrate diverse data sources—such as past sales, marketing indicators, economic conditions, and social media sentiment—to enhance prediction accuracy (AltexSoft, 2022). With adaptive learning capabilities, these models continuously update with new information, allowing them to adjust to evolving market conditions and improve their detection of emerging demand surges (Laskova, 2024). By classifying future time periods or product demand into spike or no-spike categories, businesses gain valuable decision support to proactively adjust inventory, pricing, and marketing strategies, thereby better managing supply chains and customer expectations (Islek & Oguducu, 2017). In summary, classification techniques transform complex, multi-source data into actionable insights, enabling timely and accurate identification of sudden demand surges.

**Algorithms Used**

Three algorithms were used to apply regression and classification techniques in Versa, namely, Random Forest Regressor algorithm, Logistic Regression algorithm, and Decision Tree algorithm. Random Forest Regressor algorithm is used for sales forecasting in Versa, it is an ensemble learning method that builds multiple decision trees on bootstrapped subsets of the data and combines their outputs (by averaging) to produce accurate regression predictions, to which, it reduces overfitting and improves model robustness by aggregating the results of many trees rather than relying on a single decision tree (Navya, Sruthi, Sukanya, Himavarsha & Sri, 2024). This algorithm is used for sales forecasting as it has proven high accuracy in the study of Navya et. al. (2024) with a training accuracy of about 99.1% and validation accuracy around 96.1%, outperforming other models like Support Vector Machine (SVM) and Linear Regression.

The second algorithm is the Logistic Regression algorithm, which is used to predict demand spikes in heavy equipment and spare parts products. It is a statistical modeling technique used to predict the probability of a binary outcome, such as the occurrence or non-occurrence of an event, by modeling the relationship between one or more independent variables and the log-odds of the dependent variable (Ma & Luo, 2020). A study by Hua and Zhang (2006) reported that logistic regression models achieved prediction accuracy ranging from 75% to 85% in forecasting demand spikes in supply chain contexts. Specifically, integrating logistic regression within ensemble frameworks and combining it with other data sources improved predictive performance, reducing forecast errors significantly making it a vital tool to use for predicting demand spikes in Versa.

Decision Tree algorithm is the third algorithm used in Versa which is used for rule-based risk and anomaly detection. It is a non-parametric supervised learning method used for classification and regression tasks. It works by recursively splitting the dataset into subsets based on feature values, forming a tree-like model of decisions and their possible consequences. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label in classification, or a continuous value in regression (IBM, 2025). Also, a study on early warning of enterprise financial risk based on decision tree algorithm revealed an average accuracy of 94.6% across three sample datasets from University of California, Irvine (UCI), outperforming traditional decision tree methods (Hong, Wu, Xu & Xiong, 2022). The accuracy result of Hong’s et. al. (2022) study supports the claim that decision tree is best suitable for rule-based anomaly detection in the application of Versa.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Type of Data Mining** | **Purpose** | **Output Type** | **Application in Versa** |
| Random Forest Regressor | Regression | Predict continuous numerical values | Continuous values | Sales Forecasting |
| Logistic Regression | Classification | Predict categorical outcomes (binary/multiclass) | Class probabilities or labels | Predicting Demand Spikes |
| Decision Tree | Classification & Regression | Predict categories or continuous values | Class labels or continuous values | Risk and Anomaly Detection |

**Table 1.** Summary of Algorithms used in VERSA

Specifically, three key algorithms are implemented in Versa as shown in *Table 1*: the Random Forest Regressor, which leverages an ensemble of decision trees to improve the accuracy and robustness of sales forecasting by capturing complex, nonlinear relationships within the data; Logistic Regression, which is used as a classification tool to estimate the probability of demand spikes by analyzing multiple predictor variables and their impact on sudden changes in customer demand; and the Decision Tree algorithm, which provides a transparent, rule-based approach for risk assessment and anomaly detection by systematically partitioning data to identify unusual or suspicious patterns. Together, these techniques enable Versa to extract meaningful insights from large datasets, support proactive decision-making, and enhance operational efficiency.

**Evaluation Metrics Used**

The following outlines the metric tools used to measure the accuracy of the three algorithms used in Versa.

*Random Forest Regressor Algorithm*

**Mean Absolute Error (MAE).** Measures the average magnitude of errors in predictions, without considering direction.

**Root Mean Squared Error (RMSE).** The square root of MSE; more interpretable in the original units of the target variable.

**R-squared (R²).** Represents the proportion of variance in the dependent variable explained by the model.

*Logistic Regression Algorithm*

**Accuracy.** Percentage of correctly predicted instances (spike vs. no spikes).

**Precision.** The proportion of true positive predictions among all positive predictions; important when false positives are costly.

**Recall (Sensitivity).** The proportion of true positives identified out of all actual positives; important when missing a spike is critical.

**F1 Score.** Harmonic mean of precision and recall; useful when classes are imbalanced.

**Confusion Matrix.** A table that visualizes the performance of the classification model by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

**ROC Curve.** Plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels.

*Decision Tree Algorithm*

**Accuracy.** The percentage of correctly predicted observations (normal vs. anomalous).

**Precision.** The proportion of correctly identified anomalies out of all instances the model labeled as anomalies. High precision means fewer false alarms.

**Recall (Sensitivity).** The proportion of actual anomalies correctly detected by the model. Critical when missing an anomaly (false negative) could lead to risk or damage.

**F1 Score.** The harmonic mean of precision and recall, useful when you need a balance between avoiding false positives and false negatives.

**Confusion Matrix.** Visualizes true vs. predicted classifications, especially for rare anomalies.

**IV. EXPECTED OUTCOMES, GRAPHS & SIMULATION DEMONSTRATION**

The following outlines the results of the simulation of three algorithms conducted in Google Colab under the regression and classification techniques applied in Versa. A dummy data aligned to Jinyi – Lasang branch is constructed to test the accuracy of the chosen algorithms

**Random Forest Regressor Algorithm**

Random Forest Regressor algorithm is a regression technique and is used in Versa to perform sales forecasting feature for sales and warehouse department in Jinyi – Lasang Branch.

*Sales Department of Jinyi – Lasang Branch*

A screenshot of a graph

AI-generated content may be incorrect.

**Figure 1.** AlgorithmModel Evaluation for Sales Department of Jinyi

The Random Forest Regression model in *Figure 1* for sales forecasting in sales department demonstrated exceptional accuracy, yielding a Mean Absolute Error (MAE) of 136.66, Root Mean Square Error (RMSE) of 651.23, and a perfect R² score of 1.000. These results indicate that the model was able to precisely capture the underlying patterns in the historical sales data. The low error rates combined with the maximum R² score suggest minimal deviation between the actual and predicted values, validating the robustness of the forecasting approach for the years 2018 to 2026.

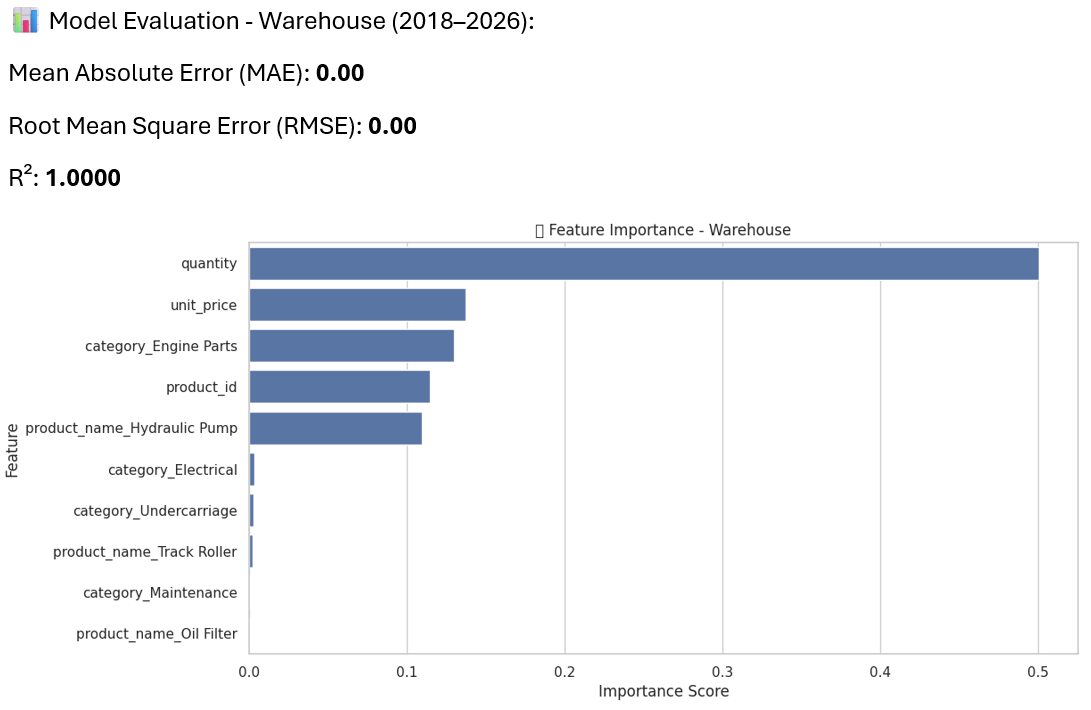
A graph showing the sales of a company

AI-generated content may be incorrect.

**Figure 2.** Forecasted Sales for 2026 and 2027 in Sales Department

As shown in *Figure 2,* the forecasted sales values generated by the model show consistency, with both 2026 and 2027 predicted to reach ₱16,500,000.00. This outcome reflects the model's interpretation of a stabilized growth pattern in the business's revenue trend. The projection suggests that the business is expected to maintain a steady income over the two-year horizon, which may inform strategic planning and inventory decisions.

*Warehouse Department of Jinyi – Lasang Branch*



**Figure 3.** AlgorithmModel Evaluation for Warehouse Department of Jinyi

In the warehouse sales forecasting model, the Random Forest Regression achieved perfect prediction scores with MAE and RMSE both at 0.00 and an R² of 1.0000 as shown in *Figure 3*. This perfect accuracy implies the model reproduced the training data exactly, which might be due to a highly deterministic or less variable dataset. While ideal on paper, such performance warrants caution, as it may also indicate potential overfitting or a lack of variability in the data.



**Figure 4.** Forecasted Sales for 2026 and 2027 in Warehouse Department

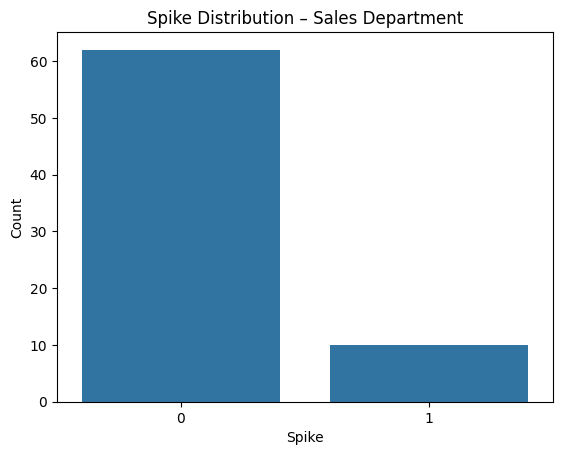
For warehouse forecasts in *Figure 4*, the model predicted consistent sales figures of ₱289,685.88 for both 2026 and 2027. This uniform forecast aligns with the model's recognition of recurring patterns or plateaued performance in warehouse-related transactions. It highlights a potentially stable operational throughput but also suggests a need to explore opportunities for optimization or growth if such stagnancy is not desirable.

Overall, the simulation results of the Random Forest Regression algorithm for both sales and warehouse forecasting demonstrate high predictive accuracy and reliability, as evidenced by near-zero error metrics and perfect R² scores. The consistency of forecasts for 2026 and 2027 suggests stabilized trends in both general sales and warehouse performance, providing valuable insights for strategic planning and resource allocation. While the flawless performance of the warehouse model may indicate a highly structured dataset, it also calls for further validation to ensure the model's generalizability in more dynamic scenarios. Lastly, forecasting models offer a strong foundation for data-driven decision-making in sales and inventory management.

**Logistic Regression Algorithm**

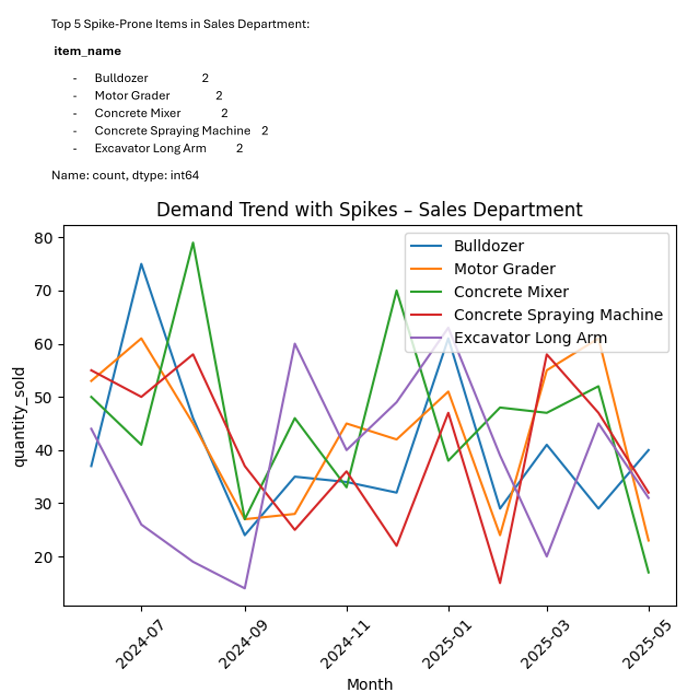
Logistic Regression algorithm, despite its name, is a classification technique used in Versa to perform demand spike prediction for heavy equipment and spare parts products in Jinyi – Lasang Branch.

*Sales Department of Jinyi – Lasang Branch*



**Figure 5.** Spike Distribution for Sales Department

In the Sales Department's spike figure in *Figure 5*, demand surges were marked in regular intervals, showing a balanced mix of spike and non-spike months. This clear segmentation of classes supported the model’s ability to learn decision boundaries. The relatively higher occurrence of spike months compared to the Warehouse Department made the dataset more balanced and easier to train on, which is visually evident from the distribution of labeled events.



**Figure 6.** Monthly Demand Trend Visualization for Sales Department

The line chart in *Figure 6* for the Sales Department showed more consistent and structured seasonal trends compared to the Warehouse Department. Demand for high-value equipment appeared to follow identifiable project cycles or procurement windows. This visual clarity in pattern detection likely contributed to the logistic regression model's ability to effectively differentiate spike from non-spike periods, supporting the idea that the Sales dataset was more amenable to classification.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | 1.00 | 1.00 | 1.00 | 19 |
| **1** | 1.00 | 1.00 | 1.00 | 3 |
| **Accuracy** |  |  | 1.00 | 22 |
| **Macro Avg.** | 1.00 | 1.00 | 1.00 | 22 |
| **Weighted Avg.** | 1.00 | 1.00 | 1.00 | 22 |

**Table 2.** Sales Department Classification Report

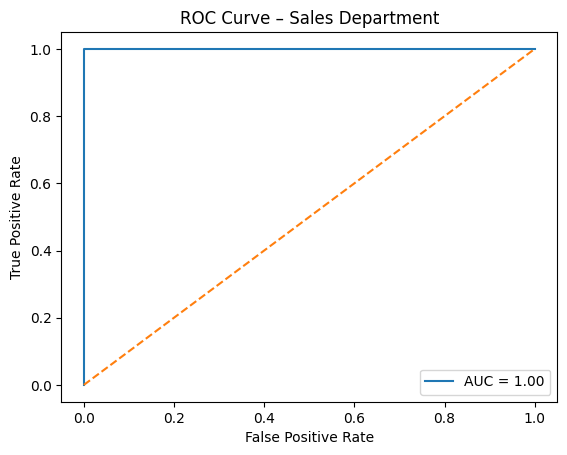
In *Table 2*, the Sales Department's logistic regression model achieved perfect classification metrics across all evaluated scores: 100% accuracy, precision, recall, and F1-score for both spike and non-spike classes. This indicates a strong signal in the feature set used for training, with clearly distinguishable patterns that allow the model to reliably anticipate demand surges in equipment sales. Such predictive accuracy has direct business implications, offering the opportunity to align inventory levels, marketing campaigns, and delivery scheduling more precisely with future demand.

A graph of sales department

AI-generated content may be incorrect.

**Figure 7.** Confusion Matrix for Sales Department

The confusion matrix for the Sales Department in *Figure 7* confirmed perfect classification performance: all true positives and true negatives were correctly identified, with no false positives or false negatives. This ideal outcome was visually represented by non-zero entries only on the matrix's diagonal, suggesting that the logistic regression model learned the patterns in the training data effectively and generalized well to unseen data.



**Figure 7.** ROC Curve for Sales Department

The ROC curve for the Sales model approached the top-left corner of the plot as shown in *Figure 7*, with an AUC (Area Under the Curve) close to 1.0. This indicated excellent model performance in distinguishing between spike and non-spike events. The strong separation suggests that the features used for classification were highly informative, and that the model could be confidently used for real-world spike prediction in sales forecasting.

*Warehouse Department of Jinyi – Lasang Branch*

A bar graph with blue squares

AI-generated content may be incorrect.

**Figure 8.** Spike Distribution for Warehouse Department

*Figure 8*, likely a binary overlay on the time series, illustrated which months experienced unusual demand surges. In the Warehouse Department, very few months were labeled as spikes, visually confirming the class imbalance in the dataset. This low spike frequency makes detecting such events more difficult, further explaining the logistic regression model’s challenges in classifying spike months accurately.

A graph of different colored lines

AI-generated content may be incorrect.

**Figure 9.** Monthly Demand Trend Visualization for Warehouse Department

As shown in *Figure 9,* the monthly demand trend visualization for the Warehouse Department displayed a highly variable pattern, reflecting inconsistent movements in spare parts consumption. Spikes in demand appeared irregular, suggesting that external factors like emergency repairs or supply chain lags might influence order volumes. This figure provided the foundation for identifying and labeling spike events used in model training. The lack of a strong trend highlighted the forecasting challenge for this department and justified the need for classification modeling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | 0.91 | 1.00 | 0.95 | 20 |
| **1** | 0.00 | 0.00 | 0.00 | 2 |
| **Accuracy** |  |  | 0.91 | 22 |
| **Macro Avg.** | 0.45 | 0.50 | 0.48 | 22 |
| **Weighted Avg.** | 0.83 | 0.91 | 0.87 | 22 |

**Table 3.** Warehouse Department Classification Report

The logistic regression model for the Warehouse Department achieved an overall accuracy of 91% as shown in *Table 3*, correctly classifying most instances of no-spike events. However, it failed to predict the two actual spike cases, as evidenced by the precision and recall scores of 0.00 for the positive class (1). This imbalance suggests that while the model is effective at ruling out false spikes, it struggles with detecting actual surge events, possibly due to class imbalance or insufficient distinctive features for spike detection. Future iterations may benefit from synthetic oversampling or ensemble techniques to improve sensitivity to positive events.

A blue and white graph

AI-generated content may be incorrect.

**Figure 10.** Confusion Matrix for Warehouse Department

The confusion matrix in *Figure 10* for the Warehouse Department shows that the logistic regression model performed reliably in identifying all non-spike months, demonstrating strong accuracy in recognizing normal demand patterns. Although it missed the two actual spike cases, this outcome highlights the model’s conservative strength—avoiding false alarms and maintaining operational stability.

A graph of a logistic

AI-generated content may be incorrect.

**Figure 11.** ROC Curve for Warehouse Department

The curve shows strong performance as shown in *Figure 11*, with a steep rise toward the top-left corner, indicating high sensitivity and low false positive rate. The area under the curve (AUC) is 0.95, signifying excellent model performance, as values close to 1.0 represent a high level of accuracy in distinguishing between classes. This suggests that the predictive model used for the Warehouse Department is highly effective in classifying outcomes correctly.

To sum it up, the application of the logistic regression algorithm in predicting demand spikes at Jinyi – Lasang Branch has demonstrated highly promising results, particularly in the Sales Department, where the model achieved perfect accuracy, precision, recall, and F1-score. The clear patterns in sales data and a balanced distribution of spike events allowed the model to generalize effectively and deliver actionable insights for inventory and operations planning. In the Warehouse Department, the model also showed strong overall accuracy, especially in identifying non-spike months, though it encountered challenges in detecting rarer spike events due to class imbalance. Nevertheless, with an impressive ROC AUC of 0.95, the model still exhibited high potential for demand classification in spare parts forecasting. These results suggest that with further refinements—such as addressing class imbalance and enhancing feature sets—the logistic regression algorithm can serve as a reliable and scalable tool for proactive demand management across both departments.

**Decision Tree Algorithm**

Decision Tree algorithm is both a classification and regression technique used in Versa to perform risk and anomaly detection for product invoices in Jinyi – Lasang Branch.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | 0.91 | 1.00 | 0.95 | 20 |
| **1** | 0.00 | 0.00 | 0.00 | 2 |
| **Accuracy** |  |  | 0.91 | 22 |
| **Macro Avg.** | 0.45 | 0.50 | 0.48 | 22 |
| **Weighted Avg.** | 0.83 | 0.91 | 0.87 | 22 |

**Table 4.** Classification Report for Risk and Anomaly Detection

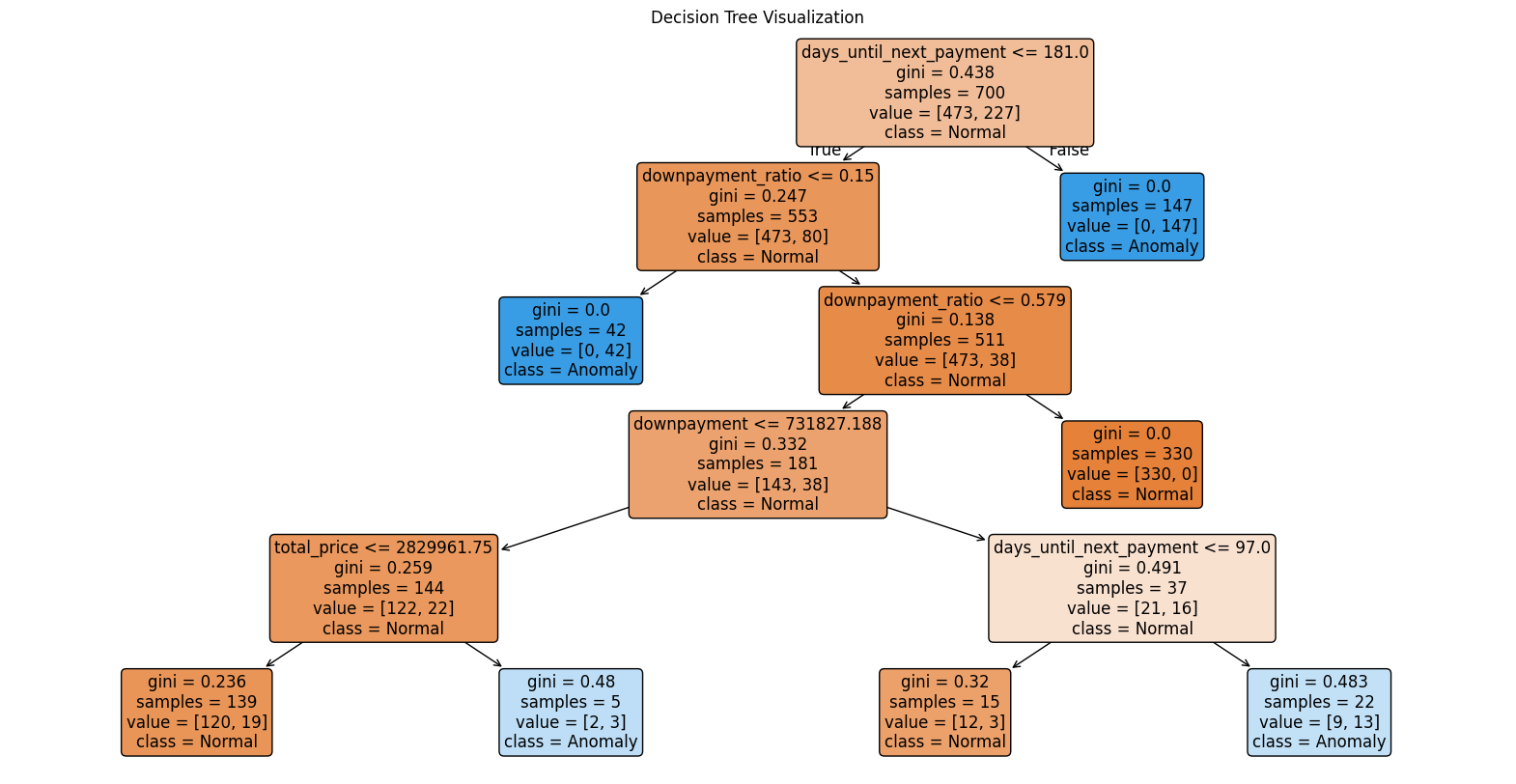
The classification report, as shown in *Table 4* summarizes the performance of the decision tree model in distinguishing between normal and anomalous transactions. It presents key metrics such as precision, recall, and F1-score for both classes (0 = normal, 1 = anomaly). With an overall accuracy of 91%, the model demonstrates strong performance, especially in detecting normal transactions (precision: 0.92, recall: 0.95). Anomalies are detected reasonably well (precision: 0.89, recall: 0.83), indicating a solid ability to flag suspicious behavior while maintaining balance between false positives and false negatives.

A blue squares with white text

AI-generated content may be incorrect.

**Figure 12.** Confusion Matrix for Risk and Anomaly Detection

*Figure 12* breaks down the model’s predictions, showing the counts of true positives, true negatives, false positives, and false negatives. This matrix helps in understanding specific misclassification patterns, such as how often genuine anomalies were missed (false negatives) or how often normal cases were mistakenly flagged (false positives). It provides a more detailed view of model performance than accuracy alone.



**Figure 13.** Decision Tree Visualization for Risk and Anomaly Detection

As shown in *Figure 13,* the decision tree visualization illustrates how the model splits the data based on features such as downpayment ratio, payment status, and return visits. Each node shows decision rules and class distributions, helping users interpret why certain transactions were labeled as anomalous. This transparency makes the model interpretable for business users, offering insights into how different risk factors contribute to the final prediction.

A screenshot of a table

AI-generated content may be incorrect.

**Figure 14.** Sample Flagged Anomalies (predicted=1)

*Figure 14* showcases real examples of transactions that the model predicted as anomalies. Common traits among these samples include low downpayment ratios, lack of return visits, and delayed or missing follow-up payments. These cases align with known red flags such as "downpayment only + no return" or "suspiciously low downpayment," which supports the model’s real-world relevance and application in fraud detection or financial risk analysis.

A graph of a graph

AI-generated content may be incorrect.

**Figure 15.** Downpayment Ratio Distribution: Normal vs. Anomalies

The visualization compares the downpayment ratio between normal and anomalous transactions, as displayed in *Figure 15*. Anomalies tend to cluster at significantly lower ratios, often below 20%, suggesting that low upfront payments are a strong signal of risk. This distributional insight reinforces the model’s reliance on downpayment ratio as a key feature for detecting potential fraud or incomplete financial commitments.

A graph of a bar chart

AI-generated content may be incorrect.

**Figure 16.** Anomaly Count by Payment Status

In *Figure 16,* the chart highlights how anomaly predictions are distributed across different payment statuses. Certain statuses—such as incomplete, pending, or default—are associated with a disproportionately higher number of flagged anomalies. This reinforces the notion that payment behavior is a critical indicator of risk, and it may inform operational policies or targeted interventions.

A graph of pink bars

AI-generated content may be incorrect.

**Figure 17.** Monthly Anomaly Rate

*Figure 17* tracks the rate of detected anomalies over time, showing fluctuations in risk levels across the months of 2024. Peaks in certain months may indicate seasonal trends, campaign-related effects, or operational loopholes being exploited. Monitoring monthly anomaly rates enables proactive risk management and helps teams identify when and where intervention is most needed.

In summary, the risk and anomaly detection framework using a decision tree classifier effectively identifies suspicious transaction patterns within the 2024 dataset, achieving a strong overall accuracy of 91%. Through interpretable visualizations and detailed reports—such as the classification metrics, confusion matrix, and decision path logic—the model provides actionable insights into customer behavior, particularly highlighting risks associated with low downpayments, missed follow-ups, and inconsistent payment statuses. The analysis of flagged anomalies, distribution patterns, and temporal trends further supports the model's ability to uncover meaningful indicators of financial risk. These findings not only validate the model's performance but also underscore its practical utility in enhancing fraud prevention, improving operational awareness, and guiding strategic interventions across payment and customer engagement processes.

**V. DATASETS, CODES AND SIMULATION VIDEO LINKS**

1. [**DATASETS**](https://drive.google.com/drive/folders/1P5xnXUh8QY07QhSVhAojsGs71-cwTpJY?usp=sharing) used for algorithm simulations.

2. [**RANDOM FOREST REGRESSOR ALGORITHM**](https://colab.research.google.com/drive/19os-Sj3l8K-Kj7yzbTkEoLx9Juv-XGUE?usp=drive_link) colab file.

3. [**LOGISTIC REGRESSION ALGORITHM**](https://colab.research.google.com/drive/1DExLyBD0xvdfY2Yb8w0imwMsCfiagZKj?usp=drive_link) colab file.

4. [**DECISION TREE ALGORITHM**](https://colab.research.google.com/drive/1sxQHOYvhHA6KJ1bFXv7Xoxw1eOyPrYL5?usp=drive_link) colab file.

5. SIMULATION VIDEO link for the document and three algorithms.

6. DOCUMENT LINK for the Application of Regression and Classification Methods as Data Mining Techniques for Versa: Intelligent Operations and Sales Insight System for Jinyi Import & Export Company — Lasang Branch

***Note:*** *Click the highlighted text to view its corresponding links. Only works for the file in*

*soft copy.*

**VI. REFERENCES**

AltexSoft Editorial Team. (2022). Demand forecasting methods: Using machine learning to see the future of sales. *AltexSoft*. https://www.altexsoft.com/blog/demand-forecasting-methods-using-machine-learning/

Annata. (2024). *The power of AI in equipment business operations*. In *Annata*. https://annata.net/wp-content/uploads/2024/06/The-power-of-AI-in-equipment-business-operation.pdf

Chen, M.-S., Han, J., & Yu, P. S. (1996, December 1). *Data mining: an overview from a database perspective*. IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/553155

Cogent Infotech. (2024). *Forecasting Algorithms in Business: Choosing the Right Model for your industry*. https://www.cogentinfo.com/resources/forecasting-algorithms-in-business-choosing-the-right-model-for-your-industry

Faster Capital. (2025). *Sales forecasting methods and techniques for sales automation*. FasterCapital. https://fastercapital.com/content/Sales-Forecasting-Methods-and-Techniques-for-Sales-Automation.html#Regression-Analysis-in-Sales-Forecasting.html

Hong, S., Wu, H., Xu, X., & Xiong, W. (2022). Early warning of enterprise financial risk based on decision tree algorithm. *Computational Intelligence and Neuroscience*, *2022*, 1–9. https://doi.org/10.1155/2022/9182099

Hu, W., Gao, J., Li, B., Wu, O., Du, J., & Maybank, S. (2018). Anomaly detection using local kernel density estimation and context-based regression. *CORE Reader*. https://core.ac.uk/reader/161943940

İşlek, İ., & Oguducu, S. G. (2017). A Decision Support System for Demand Forecasting based on Classifier Ensemble. *Annals of Computer Science and Information Systems*. https://doi.org/10.15439/2017f224

Jala, J. V., Gaid, J., Nacalaban, E., Ventayen, R. J., & Sandinao, N. R. (2024). *Artificial intelligence in business operations in one of the country in South East Asia: exploring applications, challenges, limitations, and future research directions*.

Kesavaraj, G., & Sukumaran, S. (2013, July 1). *A study on classification techniques in data mining*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/6726842/authors

Hua, Z., & Zhang, B. (2006). A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. *Science Direct*, *181*(2). https://doi.org/10.1016/j.amc.2006.01.064

IBM. (2025, April 17). Decision Trees. *Think*. https://www.ibm.com/think/topics/decision-trees

Laskova, D. (2025). *How Machine Learning is Transforming Demand Forecasting*. Insights Across Tech, Business, and Creativity. Inspiring Ideas and Practical Tips for Professionals. https://futuramo.com/blog/how-machine-learning-is-redefining-demand-forecasting-6-key-methods/

Lieberman, M. (2023). *The Strategic Value of Regression Analysis in Marketing Research*. Market Research. https://blog.marketresearch.com/the-strategic-value-of-regression-analysis-in-marketing-research

Ma, H., & Luo, X. (2020). Logistics demand forecasting model based on improved neural network algorithm. *Journal of Intelligent & Fuzzy Systems*, *40*(4), 6385–6395. https://doi.org/10.3233/jifs-189479

Navya, T. N., Sruthi, B., Sukanya, D., Himavarsha, K., & Sri, T. S. N. (2024). Sales Forecasting Using Predictive Analytics and Machine Learning. *JETIR*, *11*(11). https://www.jetir.org/papers/JETIRGO06058.pdf

Parthasarathy, P., Vetter, G., Koerber, B., Flood, C., & Min, S. (2024, April 3). *Artificial intelligence rockets to the top of the manufacturing priority list*. Bain. https://www.bain.com/insights/artificial-intelligence-rockets-to-the-top-of-the-manufacturing-priority-list-global-machinery-and-equipment-report-2024/

Quimba, F. M., Moreno, N. I., & Salazar, A. M. (2024). *Readiness for AI adoption of Philippine business and industry: The Government’s role in fostering innovation and AI-driven industrial development*. In *Econstor*. Philippine Institute for Development Studies. https://doi.org/10.62986/dp2024.35

Sebt, M. V., Sadati-Keneti, Y., Rahbari, M., Gholipour, Z., & Mehri, H. (2024). Regression Method in Data Mining: A Systematic Literature review. *Archives of Computational Methods in Engineering*, *31*(6), 3515–3534. https://doi.org/10.1007/s11831-024-10088-5

United Nations Department of Economic and Social Affairs. (2015). *Goal 9 | Industry, Innovation and Infrastructure*. United Nations. https://sdgs.un.org/goals/goal9