Northeastern University

D'Amore-McKim School of Business

MISM 6214 [01]: Business Analytics Capstone

Project Final Report

Group 2

Jonathan Adamson, Rishabh Jain, Kaveh Jalilian, Matilda Larsson, Dhruv Mehta & Shae Trainor

Introduction

In today's data-driven economy, organizations increasingly rely on analytics to make informed, strategic decisions that enhance operational efficiency, customer engagement, and revenue growth. Skandia Elevator, a leading manufacturer of grain handling equipment, is no exception. As the company expands its global footprint and refines its internal processes, leveraging historical data to optimize demand planning, customer management, and sales conversion has become critically important.

This capstone project, conducted in collaboration with Skandia Elevator, focuses on three core areas of business analytics: forecasting product demand, segmenting and prioritizing customers based on churn risk, and predicting the likelihood of sales quotes converting into actual orders. These areas were selected in close alignment with Skandia's strategic objectives and operational challenges. Together, they represent a cross-functional application of analytics, aimed at enhancing decision-making across supply chain, sales, and customer relationship management.

The first analytical track investigates historical ordering patterns to uncover trends and seasonality in product demand. Forecasting these patterns supports production planning and inventory management, ensuring that Skandia can meet market needs efficiently and effectively.

The second track concentrates on customer segmentation and churn modeling, using behavioral indicators to identify which customers are most likely to lapse and which offer the greatest long-term value. These insights guide targeted retention strategies and help the company maintain a healthy, revenue-generating customer base. The third and final track applies machine learning techniques to model the conversion probability of sales quotes, helping the sales team focus their efforts on opportunities with the highest likelihood of success.

Each of these objectives leverages data collected over several years through Skandia's internal systems, including transactional order data, quote records, and logs of customer interactions. The project not only demonstrates the technical application of data science tools and methodologies but also translates analytical outputs into business-oriented recommendations. In doing so, it exemplifies the interdisciplinary approach to problem-solving emphasized in the Business Analytics Capstone course at Northeastern University.

Overview

1. Product Demand Forecasting

This section of the project addresses the question: What are the trends, seasonality, and demand patterns for different product categories at Skandia Elevator, and how can we forecast future demand? This analysis supports data-driven decision-making for inventory management, resource planning, and production scheduling. Understanding product-level demand dynamics enables the business to prepare proactively for fluctuations and optimize operations.

2. Customer Segmentation and Churn

How can customer segmentation based on purchase behavior and churn risk be leveraged to prioritize retention efforts and drive business impact in the absence of direct marketing data? This part explores a behavior-based approach to customer prioritization using order and churn patterns. The objective is to help businesses focus their retention and engagement strategies on customers who matter most to revenue continuity. Additionally, the project will explore market expansion opportunities by predicting the next country for business entry using external country-level data.

3. Quote Conversion Prediction

Finally, this part addresses the question: What factors influence whether a quote at Skandia Elevator is successfully converted into an order, and how can we predict this outcome? This inquiry's significance lies in optimizing the sales pipeline by understanding customer behavior, improving conversion rates, and ultimately driving revenue. In the context of business analytics, predictive and diagnostic insights help refine quoting strategies, allocate sales resources effectively, and prioritize high probability leads.

Objectives and Scope

The overarching objective of this project is to apply advanced analytical techniques to address real-world business challenges faced by Skandia Elevator, using historical operational data to drive insights and support data-informed decision-making. The work is divided into three primary analytical tracks: product demand forecasting, customer segmentation and churn modeling, and quote conversion prediction. Each track focuses on a specific domain of business performance, contributing to a broader goal of optimizing sales operations, improving customer retention, and enhancing planning efficiency.

The first objective is to analyze historical order data to identify demand trends, seasonal patterns, and fluctuations across Skandia's product categories. By developing forecasting models that project future product demand, the project aims to support strategic decisions related to inventory management, resource allocation, and production scheduling. Accurate forecasts allow the company to respond proactively to cyclical variations in customer needs, reducing inefficiencies and stock imbalances.

The second objective centers on segmenting Skandia's customer base using behavioral indicators derived from transaction histories. Using Recency, Frequency, and Monetary (RFM) analysis, customer lifetime value estimation, and churn risk classification, the project seeks to provide Skandia with a framework to prioritize retention strategies. Understanding which customers are most valuable and at highest risk of attrition allows the business to focus engagement efforts where they have the greatest potential impact, even in the absence of formal marketing attribution data.

The third and final objective is to predict which sales quotes are most likely to result in confirmed orders. This aspect of the project leverages historical quote and order data to train machine learning models that estimate conversion probability based on features such as discount levels, quote source, quote revision frequency, and pricing structure. The predictive output can empower sales teams to better target high-potential quotes, refine pricing strategies, and streamline the quoting process.

The scope of the project is focused strictly on the available historical data within the 2013–2025 window, sourced from internal Skandia systems. The project assumes that the data provided is representative of business operations and customer behavior and that missing or incomplete records do not introduce systemic bias. Due to limitations in access to external or marketing datasets, the analysis relies exclusively on behavioral and transactional indicators. Additionally, some simplifications were necessary, for example, mapping internal products or quote codes to meaningful categories, requiring close collaboration with domain experts to ensure business relevance.

This project aligns directly with the core learning objectives of the MISM 6214 Business Analytics Capstone course. It integrates technical data handling, exploratory analysis, predictive modeling, and strategic interpretation within a real business context. By engaging with actual organizational data and generating insights that can influence operational and strategic outcomes, the project embodies the principles of applied analytics, critical thinking, and interdisciplinary problem-solving emphasized throughout the program.

Methodology and Data

1. **Product Demand Forecasting**

We used the Cleaned.CustomerOrderRows.csv and Cleaned.Parts.csv datasets to perform time series analysis of product demand. After initial cleaning and formatting of date columns, we merged the two datasets using the PartId as the key and extracted the Code column from the parts data to identify product categories.

To ensure accurate and meaningful category labels, we created a mapping from internal codes (e.g., PROVIS, RÖR) to Professional names (e.g., Provisions, Pipes), based on internal documentation and team feedback. The merged dataset was then grouped by both time (OrderDate) and product Code to support temporal analysis of demand at the category level.

The analysis included:

- Identifying the top 10 product categories by total order quantity.
- Aggregating monthly demand trends and seasonal order patterns.
- Creating visualizations to highlight historical performance by product category.
- Developing forecasting models using Holt-Winters exponential smoothing to project future demand, both at the total level and for the top 3 categories.

2. **Customer Segmentation and Churn**

The project uses two datasets:

- CustomerOrderRows: Transactional data including order dates, order numbers, revenue, discounts, and delivery status.
- Customers: Customer metadata including unique IDs and names.

Key techniques and steps applied:

- Data cleaning (handling missing values, type conversions)
- RFM analysis to quantify Recency, Frequency, and Monetary value
- Churn risk labeling: Customers categorized as Active (<=180 days), Dormant (181-730 days), or Churned (>730 days since last order)
- CLV calculation using historical data
- Customer Prioritization Framework combining churn labels and top 10% revenue contributors to assign:
    - High Value - At Risk
    - High Value - Retained
    - Low Value - Churned
    - Regular

3. Quote Conversion Prediction

This analysis is built upon a comprehensive dataset derived from Skandia Elevator's internal quoting and ordering systems between 2017-2025. The three primary data sources used are:

- CustomerOrderRows (COR): Contains transactional information about finalized orders.
- QuoteRows (QR): Represents detailed quote data submitted or edited by Skandia or customers.
- EntityChangeLogs (EQL): Captures all timestamped changes to quotes, enabling tracking of quote evolution over time.

To ensure data quality and consistency, extensive preprocessing steps were performed across all data frames using Python. The CustomerOrderRows data was first cleaned by removing non-transactional rows (e.g., comments marked by OrderRowType = 4 or Position = 0), parsing and correcting numerical fields (e.g., converting comma to dot in price and quantity columns), filtering out unrealistic discounts (outside the 0–100% range), and ensuring valid datetime formatting for order and delivery dates.

Derived metrics were also calculated, such as:

- Revenue per order row, based on delivered quantity and adjusted for customer discounts.
- Fulfillment rate, calculated as the ratio of delivered to ordered quantity.
- Order-to-delivery time in days.
- Net price in SEK, adjusted for discount.
- QC tool usage, identified by the presence of a quote number (FromQuoteNumber).

For the QuoteRows, similar cleaning was applied: prices and discounts were standardized, unnecessary columns were dropped, and statuses were reclassified into meaningful categories (open, changed, ordered, etc.). The origin of quotes was inferred from the structure of FromQuoteNumber and categorized into QC Customer, QC Skandia, and Skandia ERP. Missing or invalid positions were normalized and filled where appropriate.

The EntityChangeLogs data, containing timestamps for each quote modification, was streamlined by renaming columns for consistency and removing irrelevant identifiers. From this, a key derived feature, quote change count, was calculated as the number of distinct changes (timestamps) per quote.

All three datasets were then merged into a single, unified dataset. Duplicate entries were dropped to avoid redundancy. The merged data represents the most granular view of Skandia's quote-to-order pipeline, enriched with calculated fields like QuoteToOrderDays (time between quote and order) and a Boolean ConvertedToOrder indicator, which serves as the target variable for later predictive modeling.

This multi-stage cleaning and merging process ensures that the data is both analytically robust and aligned with business logic, forming a reliable foundation for downstream descriptive and predictive analyses.

Data Analysis and Results

1. **Product Demand Forecasting**

To support data-driven planning, we began by constructing a unified dataset that merged order and parts records spanning from 2017 to 2025. Each product was mapped from internal codes to human-readable category labels, ensuring consistency across reporting and analysis. After filtering for the most frequently ordered items, we focused our forecasting efforts on the top three categories by volume: Pipes, Bolts, and Provisions. We analyzed monthly demand patterns and identified strong seasonal trends, particularly within the Pipes category, which consistently peaked during mid-year months. Leveraging Holt-Winters exponential smoothing models, we forecasted a 12% increase in pipe demand by June 2025—an insight that underscores the need for early inventory ramp-up in anticipation of seasonal surges.

Bolts, in contrast, exhibited a consistent downward trend, with forecasts predicting a 25% decline in demand over the next year. This suggests a potential shift in customer preferences or product mix, signaling the need for Skandia Elevator to reevaluate sourcing strategies and minimize overproduction. Meanwhile, Provisions showed irregular, volatile demand, averaging approximately 10 units per month. The volatility made accurate forecasting more difficult, highlighting the need for simpler, more adaptive planning methods or further data validation in this segment.

Beyond individual product trends, our total demand forecasts revealed a slight but steady upward trend across all categories, aligning with historical seasonality. While the models provided actionable insights, it is important to note that they relied solely on internal order data and did not incorporate external market factors such as economic shifts, inflation, or supply chain disruptions. Therefore, the accuracy of projections is contingent on relatively stable market conditions. Nonetheless, these forecasts provide a solid foundation for improving supply chain agility, minimizing stockouts, and reducing holding costs through more strategic inventory planning.

2. **Customer Segmentation and Churn**

Over the past 13.7 years, the company has demonstrated impressive revenue consistency, generating over $11.3 billion in net revenue. This figure stems from $11.7 billion in positive sales, offset by $347 million in returns or cancellations. With an average order value of $88,371, the business has sustained a monthly revenue of $68.8 million, or $825.7 million annually, a strong indicator of long-term operational health.

Operational execution has been equally solid. 98.6% of transactions moved successfully from quote to order, and the delivery success rate of 97.1% reflects a well-optimized supply chain. Interestingly, the Quote Configurator (QC), the website tool used for self-service quoting—was used in about one-third of all orders. However, it's worth noting that orders placed *without* QC had an average value 3.2 times higher, suggesting the QC tool may be skewed toward smaller, lower-value transactions or less complex sales.

From a revenue concentration lens, a striking dependency emerges. One customer, OLD BDC Systems LTD , accounts for an overwhelming 77.5% of total revenue ($8.8B), with just 2,395 orders averaging $3.67M each. This reveals a highly concentrated revenue model. In fact, the top 10 customers contribute 90%, and the top 100 control over 99% of total revenue. While such partnerships reflect the strength of B2B relationships, they also present a significant risk, one shift in loyalty or purchasing behavior from a top client could have massive implications.

Zooming out to the full customer base of 3,271 accounts, only 1,881 (57.5%) have placed an order leaving a 42.5% segment untouched, and therefore, a massive growth opportunity. What's more revealing is the revenue skew: the top 20% of active customers drive 99.9% of the revenue, meaning the bottom 80% contribute virtually nothing. This kind of Pareto distribution underlines the company's reliance on a very narrow band of high-value customers.

To better understand these dynamics, RFM segmentation was applied, classifying customers based on Recency, Frequency, and Monetary value. The top tier, labeled "Champions", represents only 5% of engaged customers but contributes a staggering 40% of revenue, averaging $47.8M per account. In contrast, "Loyal Customers" (16%) and "New Customers" (25%) contributed 28% of the revenue and 15% of the revenue respectively. A worrying 25% fall into the "At Risk" category showing signs of decline while the remaining 29% are tagged as "Others," with low but recoverable engagement.

When analyzing customer lifecycle behavior, we saw that:

- 27.3% of customers are currently active
- 15.2% are dormant (no orders in 6–24 months)
- 15.3% have churned (no orders in 2+ years)
- 42.2% have never placed an order

Revenue alignment reflects this segmentation: Active customers account for $10.2B, Dormant for $856M, and churned for $293M. The "Never Ordered" group, naturally, accounts for $0. A deeper dive into retention shows that Champions are highly loyal, with 75% still active even after 16 years but At-Risk customers fall off sharply between 19 to 36 months, with a 55% drop-off rate, revealing a critical churn window.

To proactively predict and mitigate churn, we deployed Logistic Regression and Random Forest models using 8 key features like order count, revenue, fulfillment rate, product diversity, and QC usage. Logistic Regression achieved 85.2% accuracy, identifying that frequent orders, high fulfillment rates, and diverse product purchasing reduce churn. However, higher discounts and poor fulfillment increase churn risk.

The Random Forest model outperformed in uncovering deeper patterns. Key drivers of retention were total revenue, order count, and product variety. The takeaway? The more customers spend, buy frequently, and explore your catalog, the less likely they are to leave.

From these models, three strategic insights stood out:

1. The Discount Death Spiral – Customers trained on discounts become price-sensitive and less loyal. Strategic discounting, not reactive, is the way forward.
2. Product Diversity = Loyalty – Customers who purchase a variety of products are more invested. Early cross-selling is critical.
3. Frequency Builds Stickiness – Consistent ordering behavior is a strong loyalty signal. Encouraging regular buying habits early can prevent churn.

Lastly, we built a Customer Lifetime Value (CLV) segmentation using churn risk and revenue tiers:

- High Value -At Risk: Top 10% of revenue contributors who are now inactive (a major red flag)
- High Value -Retained: Top 10% who are still engaged (worth doubling down on)
- Low Value Churned: Less risk, but still valuable for trend analysis
- Regular- Most customers fall here, representing the next upsell opportunity

This segmentation tells a clear story: customer loyalty isn't just a result of good service; it's predictable, measurable, and manageable. And with data-backed strategies in place, the business is well-positioned to improve retention, grow mid-tier accounts, and de-risk its revenue base.

3. Quote Conversion Prediction

The descriptive analysis yielded several informative results:

- Quote-to-Order Ratio: Approximately 21.95% of all quotes in the dataset result in a confirmed order. This provides a baseline conversion rate for the company to benchmark its quoting performance.
- Quote Change Statistics: The average quote undergoes about 0.76 changes, with a maximum of 4. The median number of changes is 0, indicating no changes to the order, suggesting that while most quotes are finalized quickly, a subset requires iterative negotiation or internal adjustments.
- Time Between Quote and Order: For those quotes that do convert to orders, the average time from quote creation to order is 31.6 days. However, there is considerable variability, with some quotes turning into orders on the same day and others taking several months (maximum recorded delay: 960 days). Interestingly, some quotes even show negative delays due to possible data entry issues or system timestamp mismatches.
- Quote Source Performance: The conversion rates differ significantly based on quote origin:
    - Skandia ERP: 33.2% conversion rate
    - QC Customer: 20.4% conversion rate
    - QC Skandia: 9.8% conversion rate

This indicates that quotes originating from Skandia's ERP system are far more likely to result in orders compared to those generated through Quote Configurators (QC), especially when used internally.

The implications of these findings are clear: quote source and the number of quote changes are correlated with conversion likelihood, and long quote-to-order delays might represent inefficiencies or lost opportunities.

The predictive modeling phase of this project aimed to determine the likelihood that a sales quote issued at Skandia Elevator would ultimately result in a confirmed customer order. This classification task used historical data extracted from Skandia's internal quoting and order management systems and was designed to provide actionable insights into quote success drivers. By modeling this process, the goal was to enable more efficient lead prioritization, better sales forecasting, and improved revenue conversion.

Three primary datasets: Customer Order Rows, Quote Rows, and Entity Change Logs, were cleaned, preprocessed, and merged into a single analytical file to support this effort. This combined dataset offered a holistic view of quote behavior, capturing quoted quantities, discount levels, pricing, modification history, origin system, and eventual conversion status. A binary target variable, order, was created to indicate whether a confirmed order had followed a given quote.

Significant preprocessing steps were applied to ensure the data's integrity and modeling readiness. Columns directly tied to post-quote outcomes, such as order numbers, revenue amounts, and specific timestamps, were removed to prevent data leakage. Categorical features, including quote source and quote status, were transformed via one-hot encoding, while continuous variables were normalized using MinMax scaling to ensure uniform input ranges across the model.

Logistic regression was initially selected for its interpretability and straightforward implementation. After splitting the dataset into training and testing subsets using a 70/30 split, the model achieved a test set RMSE of 0.1072. Although adequate as a baseline, further model tuning was explored to enhance performance. Three feature selection strategies, forward selection, backward elimination, and exhaustive search were employed to identify the most predictive subset of input variables. All three approaches consistently returned to a shared group of six predictors: discount, pricing, quote origin, quote position, and change frequency. Surprisingly, model performance dropped when these selected features were used in isolation, with the RMSE rising to 0.6477. This suggested that reducing the feature space may have resulted in lost contextual interactions.

A decision tree classifier was implemented to recognize the potential limitations of linear modeling in capturing non-linear interactions. This model not only aligned with the binary nature of the target but also offered superior interpretability through its visual structure and rule-based predictions. The decision tree substantially outperformed the logistic regression model, achieving a test RMSE of 0.046 and an F1-score of 0.997. These metrics reflected exceptionally high predictive accuracy and minimal classification error. Cross-validation confirmed the model's robustness, with a cross-validated RMSE of 0.081, indicating that the model generalized well to unseen data without overfitting.

Further refinement was achieved through hyperparameter tuning using GridSearchCV. Optimal parameters for tree depth and minimum split size were identified, with a maximum depth of 11 offering the best balance between complexity and generalizability. The final model demonstrated

strong performance across all classification metrics, including precision, recall, and accuracy, with an overall accuracy of 99.79%. Misclassification was minimal, with 83 false positives and 79 false negatives out of more than 75,000 samples. A confusion matrix and visual inspection of the tree confirmed that the model made intuitive splits, prioritizing meaningful business variables such as quote origin and discount level.

The resulting classifier is not only statistically robust but also operationally useful. Its structure can be directly interpreted and deployed within Skandia's sales systems, enabling real-time scoring of open quotes, and assisting account managers in prioritizing high-likelihood opportunities. As a next step, this model can be embedded into CRM workflows or paired with dashboards to support conversion-focused decision-making and sales performance optimization.

This modeling process, from baseline logistic regression to high-performance decision tree classifier, illustrates the complexity and opportunity within Skandia's quoting data. The work completed to date provides a production-ready predictive tool and actionable insights that connect directly to business strategy.

Discussion and Interpretation

1. **Product Demand Forecasting**

The demand forecasting component of this project aimed to identify temporal trends in Skandia Elevator's product orders and generate reliable predictions to support strategic planning in supply chain, inventory, and production management.

Descriptive analysis revealed clear seasonality and volume variation across product categories. For instance, Pipes and Elevator Heads consistently exhibited strong monthly peaks, indicating demand cycles that align with operational patterns. These insights suggest opportunities for proactive stock planning and workforce scheduling.

The forecasting phase used the Holt-Winters exponential smoothing model, selected for its suitability in capturing both trend and seasonal components in monthly demand data. Forecasts for the top product categories showed stable or gradually increasing trends, reinforcing the expectation of sustained demand.

One key challenge involved handling negative forecasted values, especially for lower-volume categories like Provisions, where irregular ordering patterns affected the model's stability. To ensure realistic reporting, predicted values below zero were capped at zero, which is a standard practice in operational forecasting when negative quantities are not meaningful.

Additionally, raw product codes were not inherently interpretable for business stakeholders. To address this, a mapping step was applied to convert internal codes into human-readable category names (e.g., 'PROVIS' → 'Provisions'), improving visualization clarity and stakeholder communication.

While the model performed well, it has limitations. It does not incorporate external causal factors such as market conditions, economic shifts, or promotional events, which could enhance long-

term forecast accuracy. Moreover, it assumes historical seasonality patterns will continue unchanged, which may not hold in volatile market environments.

Despite these limitations, the model offers actionable insights for tactical planning and resource allocation. Future improvements could include multivariate forecasting methods, integration of external variables, or the use of ensemble models to boost robustness and adaptability.

2. **Customer Segmentation and Churn**

This analysis reveals a striking imbalance in customer engagement, with over 42% of the customer base never placing an order, and another 30% falling into dormant or churned categories. Despite this, a small core of engaged customers has generated the entire $11.3B in revenue, demonstrating that loyalty, not scale, is currently driving growth.

The predictive modeling phase utilized both Logistic Regression and Random Forest classifiers to identify churn-prone customers. Logistic Regression helped identify linear relationships, highlighting that frequent orders, product diversity, and consistent fulfillment are inversely related to churn. Conversely, customers who receive high discounts or suffer from poor fulfillment are more likely to leave.

Random Forest modeling brought deeper interpretability to non-linear patterns, confirming that total revenue, order frequency**,** and unique parts ordered are the most critical features in predicting churn. The importance of product diversity supports cross-selling strategies, especially in the early stages of customer acquisition.

Both models aligned in their findings, but Random Forest proved superior in capturing the complexity of customer behavior and delivering more actionable segmentation. The model allowed us to classify customers into four distinct CLV-based groups, each requiring a unique strategy. For example, High Value – At Risk customers present a direct threat to revenue continuity and must prioritize retention initiatives. These are often long-time clients exhibiting reduced order frequency or engagement.

Three high-level insights emerged:

1. **Discount-Driven Churn**: Over-reliance on price incentives leads to transactional relationships, reducing long-term loyalty. Discount-heavy customers are disproportionately likely to churn.
2. **Cross-Sell is Crucial**: Customers with a wider product footprint show greater retention. Introducing multiple SKUs within the first 90 days may increase stickiness.
3. **Order Frequency Drives Retention**: Sporadic ordering behavior is an early warning sign. Establishing a cadence early in the relationship helps build customer habits that are hard to break.

Taken together, these models not only predict churn with over 85% accuracy but also serve as a roadmap for proactive customer lifecycle management. By aligning marketing and sales interventions with the predictive insights, the business can reduce churn, maximize lifetime value, and deepen customer relationships in a sustainable way.

3. Quote Conversion Prediction

The quote conversion prediction analysis was designed to estimate the likelihood that a sales quote would result in a confirmed customer order. This workstream yielded promising results, with both descriptive and predictive components contributing to a clearer understanding of Skandia Elevator's sales pipeline efficiency.

The descriptive phase revealed that only 21.95% of quotes were converted into orders. Further exploration showed that quotes generated via Skandia's ERP system were significantly more likely to convert compared to those created through internal or customer-facing quote configurators. Additionally, the number of changes made to a quote, captured as change count, emerged as a key behavioral indicator, with frequent revisions correlating reduced conversion likelihood. These insights emphasized the impact of both quote origin and quote-handling behavior on eventual order outcomes.

From a predictive standpoint, the initial logistic regression model served as a useful baseline, achieving a test RMSE of 0.1072. However, its performance plateaued despite iterative tuning and feature selection. As a result, the modeling strategy was shifted to a decision tree classifier, which offered a better fit for the binary classification task and was more adept at capturing non-linear relationships between features.

The decision tree classifier substantially outperformed the logistic model, achieving a test RMSE of 0.046 and an F1-score of 0.997. Cross-validation confirmed the model's reliability, with an average RMSE of 0.081 across folds, indicating generalizability beyond the training set. Features such as QuoteSource, Discount, and change_count were among the most influential, with the model's interpretable structure confirming their predictive strength. Hyperparameter tuning via GridSearchCV further optimized the model by setting an appropriate maximum tree depth, balancing complexity, and overfitting.

Despite these strong results, limitations were present. The model was trained on internal operational data only, excluding potentially valuable external factors such as customer firmographics, sales representative involvement, or competitive context. Additionally, minor inconsistencies in date formats and field definitions introduced preprocessing challenges, although these were mitigated through careful cleaning and variable engineering.

Overall, the quote conversion model provided a highly accurate and interpretable framework for predicting quote outcomes. It also highlighted areas within the quoting process where operational improvements may directly influence sales performance.

Recommendations

1. Product Demand Forecasting

Based on our demand forecasting analysis, we recommend several actions to enhance Skandia Elevator's supply chain planning and operational efficiency. First, the Holt-Winters exponential smoothing models used in our analysis should be integrated into Skandia's planning systems to produce automated monthly forecasts by product category. This would enable more data-driven scheduling and procurement. High-demand categories such as Pipes should be prioritized, with

safety stock and production capacity adjusted to meet seasonal peaks. For more volatile segments like Provisions, which showed low forecast accuracy, we recommend either improving data quality or applying simpler planning approaches. It's also important to standardize how internal product codes are mapped to readable category names across all reporting tools to ensure consistency and clarity in communication between analytics and operations teams.

To keep models relevant, forecasting should be updated at least quarterly using the latest order data. Furthermore, we suggest building dashboards (e.g., in Power BI) to visualize forecasts and trends, enabling real-time decision-making and better stakeholder alignment. If implemented, these actions are expected to reduce stockouts and excess inventory, lower holding, and operational costs, improve responsiveness to demand shifts, and ultimately enhance customer satisfaction and long-term resilience.

2. Customer Segmentation and Churn

Based on the analysis, several strategic actions are recommended to drive revenue growth and reduce churn risk.

First, customer concentration remains a major vulnerability, with over 77% of total revenue coming from a single client. It is imperative to diversify the customer base by nurturing mid-tier accounts through account-based marketing and converting inactive users with tailored engagement campaigns. Simultaneously, the overuse of discounts has emerged as a driver of churn, creating a transactional relationship with customers. Instead of blanket discounting, the business should pivot to personalized, value-based incentives and test the lowest effective incentive thresholds using A/B testing.

Another clear insight is the strong retention correlation with product diversity. Customers who purchase a variety of SKUs tend to stay longer, indicating that early cross-selling efforts are crucial. Recommending complementary products, bundling offerings, and highlighting relevant categories based on prior purchases can all foster long-term loyalty. Order frequency also plays a critical role those with irregular purchase patterns are more likely to churn. To address this, the company should implement reorder reminders, promote consistent ordering habits, and consider loyalty programs or subscription models to drive habitual behavior.

The CLV-based segmentation also points to a concerning number of high-value but at-risk customers. These accounts, which contributed significantly in the past but are now dormant, present a major opportunity for reactivation. Personalized outreach, account manager follow-ups, and targeted win-back campaigns should be prioritized to recover lost revenue from this cohort. Additionally, while the Quote Configurator (QC) is a useful digital tool, its performance trails ERP-generated quotes in both size and conversion rate. Enhancing the QC interface with guided quote suggestions, simplified UX, and in-tool product education could improve outcomes.

Finally, the predictive modeling process underscores the importance of robust, continuous data collection. To sustain churn reduction, the company should embed RFM scores and model-based churn risk into their CRM workflows and regularly update predictive models using newly collected

behavioral and transactional data. Monitoring dashboards and KPIs should be established to assess the impact of these initiatives and refine strategies over time.

3. Quote Conversion Prediction

Based on the findings from the quote conversion prediction analysis, several recommendations are proposed to improve Skandia Elevator's sales efficiency and quoting strategy.

First, the predictive model should be operationalized within the company's quoting systems to provide real-time probability scores for open quotes. By identifying quotes with a high likelihood of conversion, sales teams can prioritize follow-ups more strategically, focusing resources on leads most likely to generate revenue. This will not only improve conversion rates but also help reduce the sales cycle time.

Second, the insights from feature importance suggest revisiting how quotes are initiated and revised. Since quotes from the ERP system have the highest success rates, internal quoting protocols should be reviewed to encourage broader adoption of ERP-based quote generation. Likewise, efforts should be made to minimize excessive quote revisions, which may reflect either internal inefficiencies or customer indecision, both of which correlate with lower conversion.

Third, quotes that fall into mid-range probability bands, neither highly likely nor unlikely to convert, could benefit from targeted intervention strategies, such as time-bound offers, additional support from sales engineers, or escalated approval processes to refine pricing. These actions could move borderline quotes toward successful conversion.

Finally, Skandia should consider continuous retraining of the model using new quote and order data, ideally on a quarterly basis. This will allow the model to adapt to changing market conditions, pricing strategies, or customer behaviors, ensuring that predictive performance remains high over time.

If implemented, these recommendations could lead to measurable improvements in sales efficiency, reduced quoting costs, and increased revenue. More broadly, they offer a pathway to embedding predictive intelligence into daily operations, empowering teams to act on data rather than intuition.

Conclusion

This project demonstrates the decisive role that business analytics can play in transforming operational data into strategic decision-making tools. By examining three distinct but interrelated domains: product demand forecasting, customer segmentation and churn analysis, and quote conversion prediction, we have shown how historical transactional data at Skandia Elevator can be leveraged to uncover patterns, predict future behavior, and ultimately inform business strategy.

In the quote conversion prediction segment specifically, we built a highly accurate and interpretable decision tree classifier that enables Skandia to assess the likelihood of a sales quote becoming a confirmed order. This tool empowers the sales organization to allocate resources more effectively, prioritize leads, and refine its quoting process. The model also surfaced critical

business drivers such as quote origin, frequency of revisions, and discount strategy, providing tactical guidance and strategic insight.

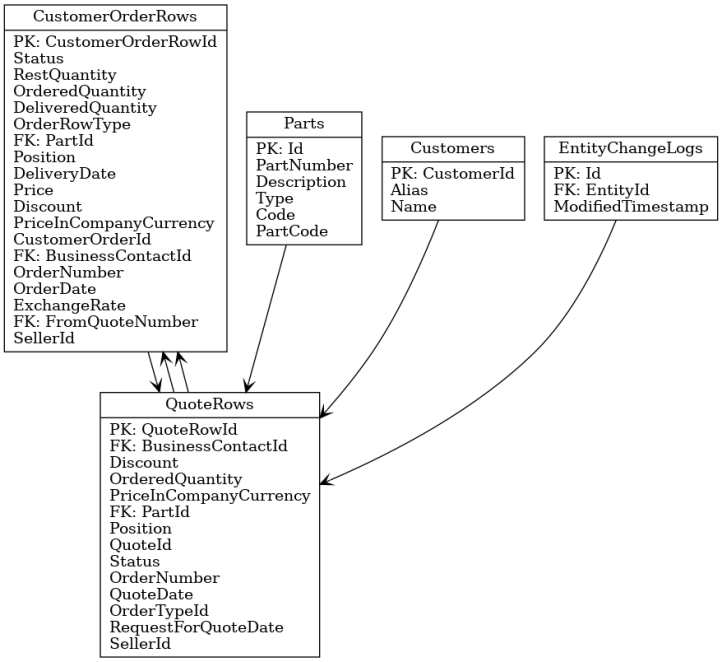
While the model achieved exceptional accuracy, its implementation revealed specific data and operational limitations, including the absence of marketing context and occasional inconsistencies in the historical records. Nevertheless, these were addressed through careful data cleaning and validation, and the remaining dataset formed a strong foundation for the model's success.

More broadly, this work reflects the objectives of the Business Analytics Capstone course at Northeastern University, applying statistical modeling, machine learning, and strategic thinking to real-world challenges. It illustrates how analytical rigor, when coupled with business acumen, can generate tools that are technically robust and operationally impactful.

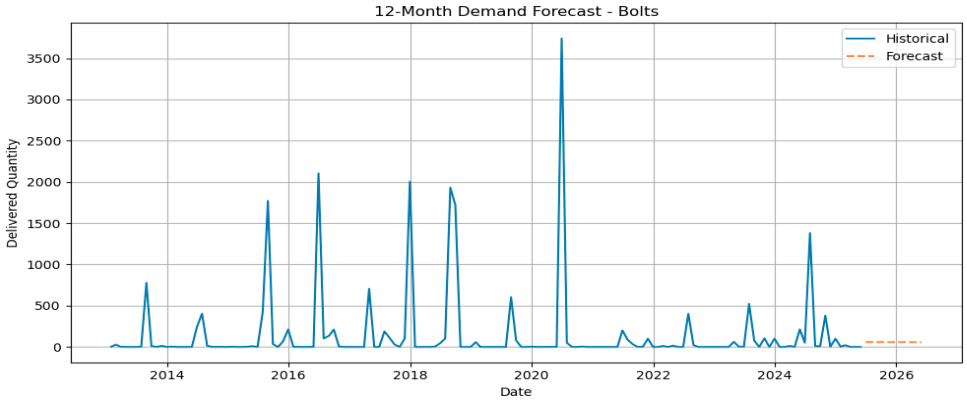
Looking forward, Skandia has the opportunity to extend these insights across its sales pipeline and customer engagement processes. By integrating predictive analytics into day-to-day systems, the company can move toward a more proactive, data-driven culture, enhancing competitiveness, efficiency, and customer satisfaction in a rapidly evolving global market.

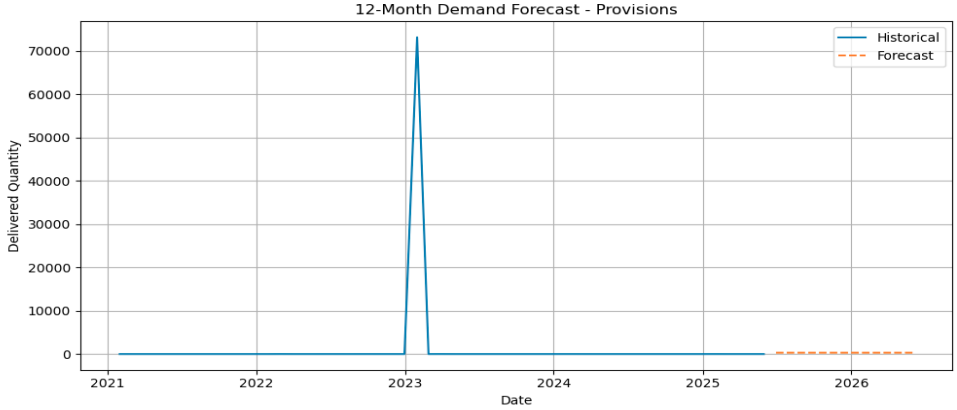
Appendix

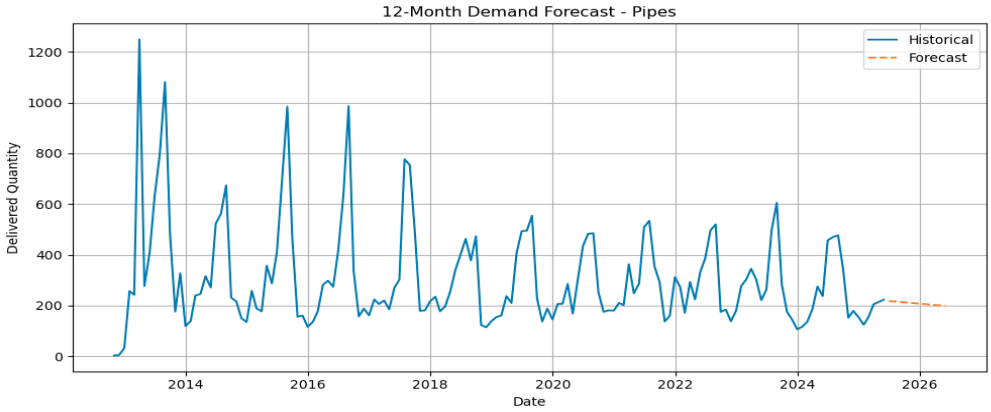
Entity Relationship Diagram

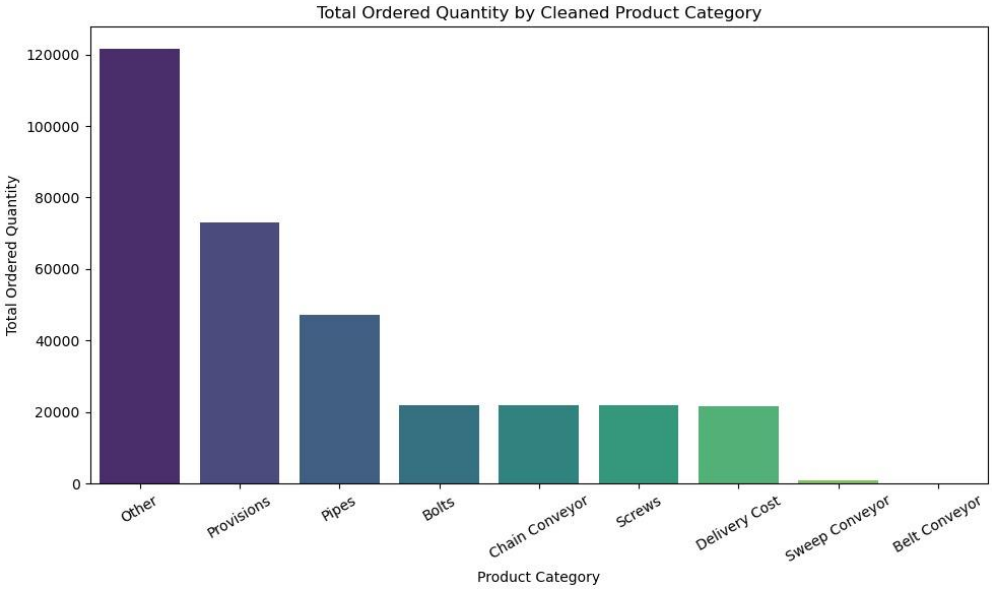


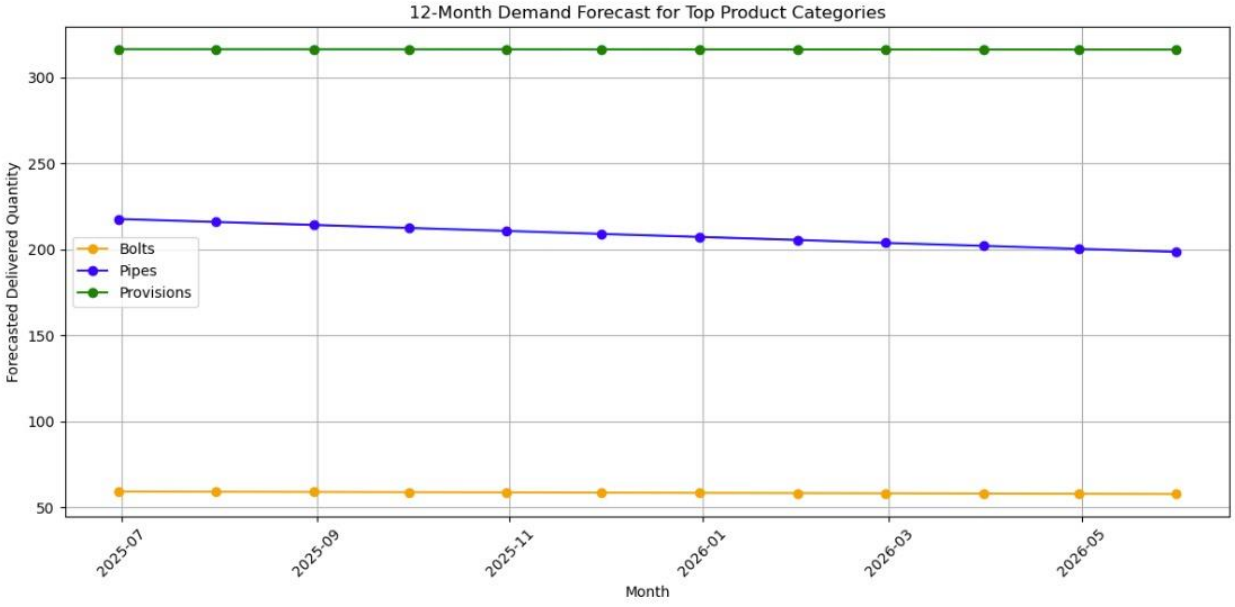
1. Product Demand Forecasting



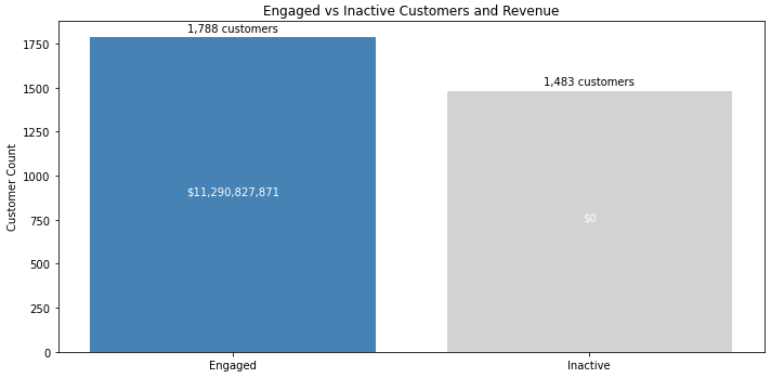


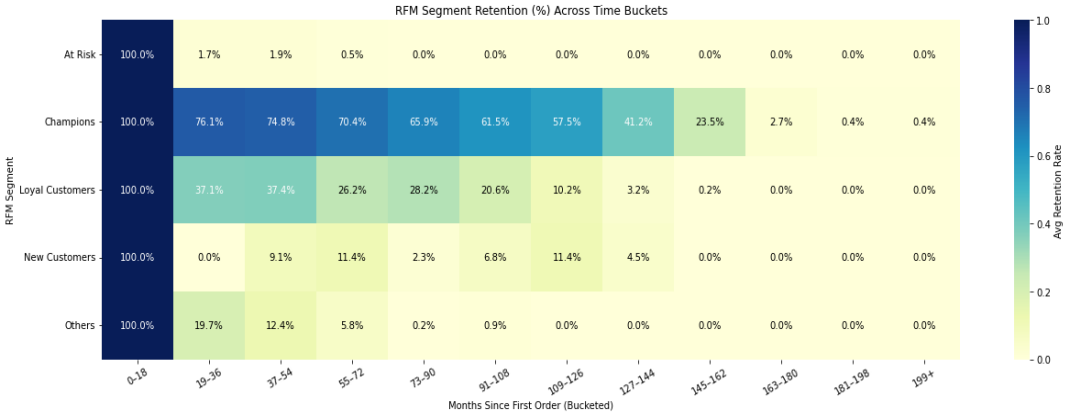

12-Month Demand Forecast - Pipes
Historical
Forecast
Delivered Quantity
Date
0
200
400
600
800
1000
1200
2014
2016
2018
2020
2022
2024
2026

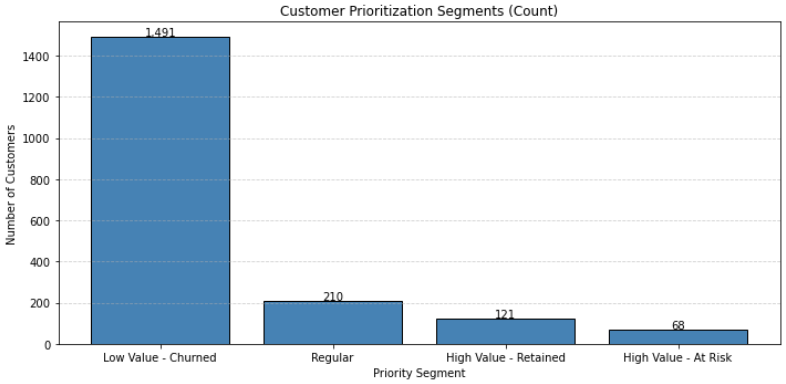

Total Ordered Quantity by Cleaned Product Category
Total Ordered Quantity
Product Category
0
20000
40000
60000
80000
100000
120000
Other
Provisions
Pipes
Bolts
Chain Conveyor
Screws
Delivery Cost
Sweep Conveyor
Belt Conveyor

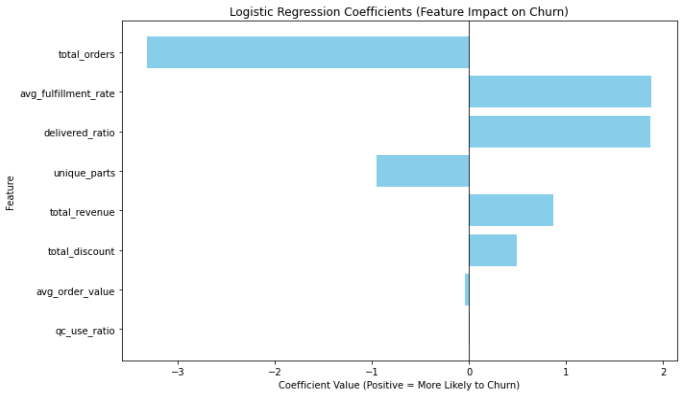

12-Month Demand Forecast for Top Product Categories
Forecasted Delivered Quantity
Month
Bolts
Pipes
Provisions
50
100
150
200
250
300
2025-07
2025-09
2025-11
2026-01
2026-03
2026-05

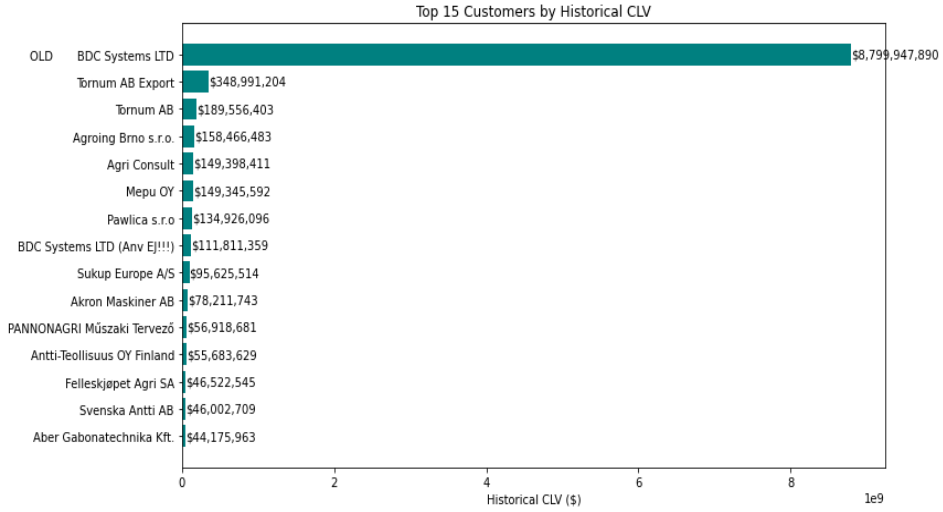
2. Customer Segmentation and Churn

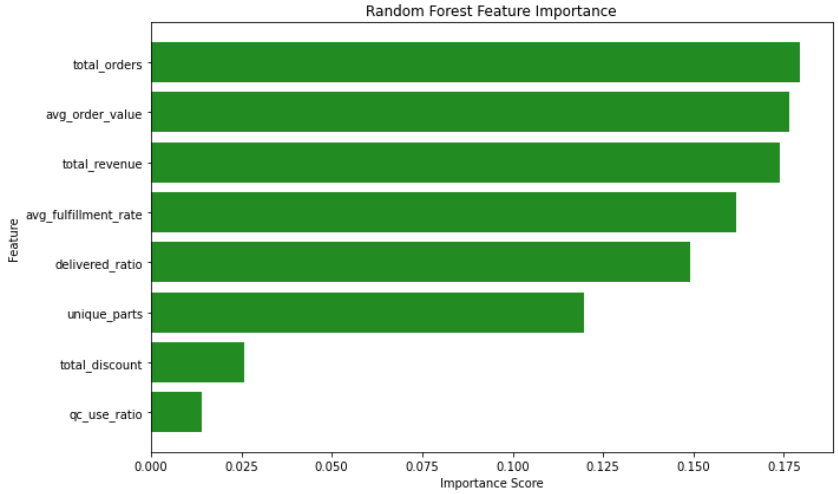


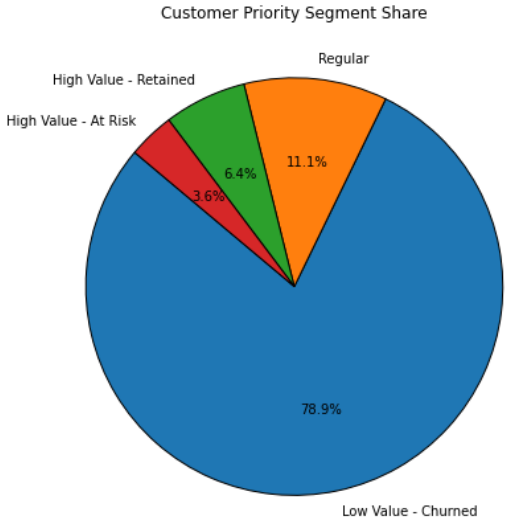







Top 15 Customers by Historical CLV
OLD BDC Systems LTD
Tornum AB Export
Tornum AB
Agroing Brno s.r.o.
Agri Consult
Mepu OY
Pawlica s.r.o
BDC Systems LTD (Anv EJ!!!)
Sukup Europe A/S
Akron Maskiner AB
PANNONAGRI Műszaki Tervező
Antti-Teollisuus OY Finland
Felleskjøpet Agri SA
Svenska Antti AB
Aber Gabonatechnika Kft.
$8,799,947,890
$348,991,204
$189,556,403
$158,466,483
$149,398,411
$149,345,592
$134,926,096
$111,811,359
$95,625,514
$78,211,743
$56,918,681
$55,683,629
$46,522,545
$46,002,709
$44,175,963
0
2
4
6
8
1e9
Historical CLV ($)


Random Forest Feature Importance
total_orders
avg_order_value
total_revenue
avg_fulfillment_rate
delivered_ratio
unique_parts
total_discount
qc_use_ratio
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance Score


Customer Priority Segment Share
Regular
High Value - Retained
High Value - At Risk
11.1%
6.4%
3.6%
78.9%
Low Value - Churned

3. Quote Conversion Prediction

Quote To Order Ratio	Total Quotes	Total Orders
0.22	37450	8234