CUSTOMER BEHAVIOR ANALYSIS AND QUOTATION SUCCESS PREDICTION USING MACHINE LEARNING FOR SALES OPTIMIZATION

By

Dapana Durage Sharmila Dulmi Chandrasiri

Dissertation submitted to the University of Sri Jayewardenepura

in partial fulfillment of the requirement for the award of Master in Data Science and Artificial Intelligence

Declaration by Student

I, Dulmi Chandrasiri, hereby attest that the research I have provided here is entirely original with no submissions or publications for credit towards a degree program. Any references for additional works or material that are included in this thesis have given mentioned in the reference section and given proper credit.

_____

Dulmi Chandrasiri

University of Sri Jayewardenepura

Date : 10/04/2025

CUSTOMER BEHAVIOR ANALYSIS AND QUOTATION SUCCESS PREDICTION USING MACHINE LEARNING FOR SALES OPTIMIZATION

A Thesis

Submitted in partial fulfillment for the Master in Data Science and Artificial Intelligence to the department of Statistics Faculty of Graduate Studies University of Sri Jayewardenepura

By Dulmi Chandrasiri

University of Sri Jayewardenepura

Date: 10/04/2025

Dr. Ravimal Bandara

(Supervisor)

Senior Lecturer in Computer Science

Certificate

Certified that the thesis entitled “CUSTOMER BEHAVIOR ANALYSIS AND QUOTATION SUCCESS PREDICTION USING MACHINE LEARNING FOR SALES OPTIMIZATION” submitted by Miss. Dulmi Chandrasiri towards partial fulfillment for the Master in Data Science and Artificial Intelligence is based on the investigation carried out under our guidance. The thesis part therefore has not submitted for the academic award of any other university or institution.

_____

Dr. Ravimal Bandara

(Supervisor)

Senior Lecturer in Computer Science

ACKNOWLEDGEMENT

I am deeply grateful to Dr. Ravimal Bandara, my supervisor, for his expert guidance, unwavering support, and insightful suggestions throughout the course of this research. His encouragement and mentorship played a crucial role in completing this study successfully.

My appreciation also goes to the Department of Statistics, Faculty of Graduate Studies, University of Sri Jayewardenepura for providing the academic support and facilities needed for this work.

I sincerely thank Electro-Serv Lanka (Pvt) Ltd for providing access to essential data and for their cooperation during the data collection phase, which significantly contributed to the practical relevance of the study.

Finally, I would like to thank my friends, peers, and family members for their continued motivation and encouragement throughout this research journey.

ABSTRACT

This research focuses on *Customer Behavior Analysis and Quotation Success Prediction* using machine learning to optimize sales strategies in the electrical industry. Data was collected from a private electrical company, encompassing a wide range of sales, customer, and item-related attributes. The dataset underwent meticulous preprocessing involving data cleaning and feature engineering, with features categorized into six primary groups: database attributes, customer behavior features, financial ratios, item characteristics, managerial and departmental information, and stock availability metrics.

To enhance model performance and reduce dimensionality, feature extraction techniques such as Chi-Square and Mutual Information were applied within a threshold-based selection framework. Following the identification of the most pertinent features, classification models were developed to forecast the success of sales quotations based on historical customer behavior and sales data patterns.

The findings demonstrate that machine learning can significantly improve the accuracy of quotation outcome predictions, enabling more informed decision-making and resource allocation in the sales process. This study offers a scalable method for incorporating customer analytics into sales operations, thereby improving sales performance and customer targeting strategies.

Using engineered customer, item, and financial information, four machine learning models Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)

were used in this study to forecast quotation success. Accuracy and Root Mean Squared Error (RMSE) were the primary measures used to assess each model's performance.

The table below provides a summary of the comparison results, emphasizing each algorithm's advantages. Based on the input features, the Random Forest Classifier demonstrated superior prediction performance by achieving the best accuracy and the lowest RMSE among the four models.

Machine Learning Model	**RMSE Value**	**Accuracy**
Logistic Regression	*1.1671*	*0.6016*
Decision Tree Classifier	*0.7643*	*0.8741*
Random Forest Classifier	*0.7044*	*0.8938*
Support Vector Machine	*1.1634*	*0.6726*

Table

To gain a better understanding of the elements influencing quote failure, the dataset attributes were clustered in this study according to the lost reasons, especially Out of Stock, Other Brand, Technical Issue, Price, and Lead Time. We determined the relative weights of important characteristics affecting quotation success by examining these groupings. In addition to streamlining the quotation success prediction process, this method offered practical insights into how various aspects influence client choices in a range of loss circumstances.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Historical Background

The field of marketing and sales has always placed a strong emphasis on comprehending consumer behavior. Businesses used to obtain information on consumer preferences and buying patterns through in-person meetings, manual documentation, and simple surveys. Despite their value, these traditional approaches frequently had limitations and were not scalable, particularly when companies and clientele grew.

Companies started collecting and managing customer data in a systematic manner in the 1980s and 1990s with the introduction of Customer Relationship Management (CRM) systems. More systematic methods of examining consumer behavior were made possible by this change, which paved the way for the development of data-driven marketing tactics. Customer data became increasingly available as digital transformation proceeded, especially in the early 2000s. This led businesses to use increasingly advanced technologies, such business intelligence platforms and data mining, to find patterns and trends in massive amounts of data.

Simultaneously, an essential step in business-to-business and service-oriented businesses, the sales quote process, changed from being a manual process to a semi-automated system connected into sales and ERP software. Accurately forecasting whether a quotation would be accepted or rejected, however, remained mostly subjective and frequently relied on the

sales staff's judgement and experience. This presented a problem for businesses looking to boost sales and make efficient use of their resources.

Artificial intelligence (AI) and machine learning (ML) have redefined customer analytics and sales forecasting in recent years. These technologies allow for the creation of extremely precise forecasts by learning from past data. Applications of machine learning in sales have shown promise in customer segmentation, lead scoring, and churn prediction. Nevertheless, the use of machine learning approaches to the prediction of quotation success has received relatively little attention, especially when it comes to industrial and electrical sales.

By using historical sales and customer data from the private electrical provider Electro-Serv Lanka (Pvt) Ltd, this study seeks to close that gap. The study includes extensive data pretreatment, including cleansing and feature engineering across categories such database attributes, consumer behavior, financial ratios, item characteristics, management inputs, and stock availability. To choose the most pertinent predictors, sophisticated feature extraction methods like Chi-Square and Mutual Information are used. The objective is to create a predictive model that enables data-driven decision-making in sales optimization and improves the accuracy of quote success projections.

1.2 Problem Statement

Businesses depend more and more on reliable sales forecasting and insights into consumer behavior in today's tough business climate to boost revenue and inform strategic choices. To determine the possibility of a successful quotation, many companies still rely on manual assessments or simple statistical techniques, especially in industrial industries like electrical services. Inefficiencies including poor client targeting, lost commercial opportunities, and resource misallocation are the outcome of this.

Large volumes of historical sales and customer data are readily available, yet predictive modelling frequently makes little use of these information. Predicting whether a citation will be approved or rejected requires identifying significant patterns in complicated, high-dimensional data. Furthermore, the methods used now are not automated and do not incorporate machine learning methods that can learn from historical patterns and actions.

In order to improve prediction accuracy, a data-driven, intelligent system that can examine sales quotes and customer behavior is desperately needed. Using data gathered from an electrical company, this study attempts to close this gap by utilizing machine learning techniques to forecast quotation success and optimize the sales process.

1.3 Significance of the Study

This research is important from an academic and practical standpoint. Through the incorporation of machine learning methodologies into the examination of consumer behavior and quote success, the study advances the expanding domain of data-driven decision-making in marketing and sales. Inaccurate forecasts and ineffective resource allocation might result from traditional sales prediction techniques, which frequently rely on historical averages or subjective judgement.

Through the use of feature engineering and sophisticated analytics, one can gain a deeper understanding of the main elements that affect quotation results. These elements include database features, customer behavior patterns, financial ratios, item details, and stock availability. This may greatly enhance how businesses make decisions, which can result in improved consumer engagement, more precise targeting, and increased conversion rates.

Practically speaking, the results of this study can help businesses like Electro-Serv Lanka (Pvt) Ltd and others of a similar nature improve customer relationship management, optimize sales strategies, and boost overall business performance. Additionally, the approach's reproducibility is guaranteed by the use of Jupyter Notebook and transparent machine learning techniques, which makes it useful for additional study and industry adaption.

1.4 Objectives of the Study

The Primary objectives of this research are as follows

- To analyze customer behavior patterns employing past sales and quote information from an electrical company, with an emphasis on determining important behavioral markers that affect the results of quotations.
- To perform data preprocessing and feature engineering by categorizing data into meaningful groups such as customer behavior features, financial ratios, item features, manager and department-based features, and stock availability features.
- To apply feature selection techniques such as Chi-Square and Mutual Information to extract the most significant variables affecting quotation success.
- To develop predictive models using machine learning algorithms to forecast the likelihood of quotation acceptance or rejection.
- To evaluate the performance of the models and assess how well they work to enhance decision-making and sales forecasting.
- To provide actionable insights that can improve quotation success rates and optimize sales strategies for management and sales teams at organizations.

1.5 Limitations of the Study

Although the study offers insightful information, there are certain restrictions:

- Data Specificity: Because the study is based on data from a single business, the findings may not be as applicable to other sectors or geographical areas.
- Feature Availability: The completeness and quality of the features that are accessible determine how accurate the forecasts are. The performance of the model may be impacted by missing or inconsistent data.
- Model Interpretability: Non-technical users may find it difficult to make decisions when using machine learning algorithms that have high accuracy but poor interpretability.
- Time Restrictions: Due to the study's time constraints, a comprehensive comparison analysis and model optimization with a variety of algorithms may not be possible.
- External Factors: The model might not take into consideration outside factors that could have an impact on consumer choices, such as market trends, competitive activity, or shifts in the economy.

Chapter 2

LITERATURE REVIEW

2.1 Customer behavior in B2B Sales

Knowing how customers behave in business-to-business (B2B) settings is essential for companies looking to increase sales efficiency and establish enduring connections. B2B transactions, in contrast to business-to-consumer (B2C) situations, usually entail longer sales cycles, higher transaction values, and many decision-makers (Webster & Wind, 1972). These traits make a deeper comprehension of organizational buyer behavior and decision-making processes necessary.

Prior research highlights that B2B buyers are impacted by trust, service quality, and the quality of their relationship with salespeople in addition to product attributes and price (Anderson & Narus, 1991). Buying decisions are greatly influenced by elements including prior purchase history, brand familiarity, and personalized service options. Additionally, according to Johnston and Lewin (1996), organizational purchasers frequently adhere to a systematic procurement process that consists of need specification, supplier appraisal, and quotation comparison.

As digital platforms and data-driven methods have grown in popularity, it has become more complex to analyze consumer activity. These days, businesses forecast client wants and customize sales tactics using behavioral indicators, interaction logs, and transaction history (Sheth, 1996). Machine learning techniques have been used more and more in recent years

to model customer behavior, giving businesses the ability to more accurately forecast outcomes like lead conversion and quotation success.

Because B2B behavior is dynamic and relationship-centric, it is still difficult to fully capture its complexity, despite breakthroughs. This disparity emphasizes how crucial it is to incorporate customer behavior research into sales optimization systems in order to facilitate data-driven decision-making.

2.2 Quotation Management and Sales Optimization

Customized product configurations, negotiated price, and approval protocols are frequently included in bids in intricate B2B settings, making the process meticulous and time-sensitive. Quotes in intricate B2B settings sometimes include negotiated price, customized product configurations, and approval protocols, which makes the process meticulous and time-sensitive. Quotations frequently include negotiated price, customized product configurations, and approval protocols in intricate B2B settings, making the process meticulous and time-sensitive.

The strategic importance of quotation management in enhancing organizational performance is emphasized in the literature. Decreased client satisfaction and missed sales opportunities might arise from sluggish or imprecise quotation processes, according to studies (Moncrief & Marshall, 2005). On the other hand, efficient and data-driven quotation systems can increase conversion rates and reduce sales cycles. Quotation success is often predicted by

factors such product availability, competitive pricing, and response time (Ingram et al., 2008).

In this sense, sales optimization refers to the application of technology and data analytics to enhance sales performance. Organizations can find trends associated with successful transactions by examining previous quotation data and customer interactions. To customize sales strategies and boost win rates, strategies including pricing optimization, client segmentation, and predictive analytics are frequently employed (Zoltners, Sinha, & Lorimer, 2006). By automating the creation of insights and facilitating dynamic adaptability to market conditions, the incorporation of machine learning algorithms further improves this capability.

Additionally, contemporary quotation systems frequently interface with CRM platforms, allowing for proactive engagement and real-time tracking of client interactions. As firms aim to maintain a competitive advantage, the ability to predict quoting outcomes and optimize sales efforts based on behavioral and transactional data has become increasingly significant.

2.3 Use of Machine Learning and Data Analytics in Sales

Machine learning (ML) and data analytics have transformed sales processes by allowing businesses to go from reactive to proactive, data-driven decision-making. As the availability of vast amounts of customer and sales data has grown, companies have started using

advanced analytics to forecast sales results, better understand consumer behavior, and enhance customer interaction (Chatterjee et al., 2021).

In sales, data analytics usually entails analyzing past transactions, consumer demographics, behavioral patterns, and market trends in order to pinpoint opportunities and risks. Conventional techniques for sales forecasting mostly depended on expert opinion and linear models, which frequently lacked precision and flexibility. Modern predictive models, on the other hand, include machine learning algorithms that may learn from intricate and nonlinear correlations in the data, increasing forecast accuracy and commercial results (Mikalef et al., 2018).

For tasks like lead scoring, churn prediction, pricing optimization, and recommendation systems, machine learning has been applied extensively in the sales industry. In identifying client intent and forecasting transaction closure rates, algorithms like logistic regression, decision trees, random forests, and gradient boosting machines have demonstrated excellent performance (Nguyen et al., 2019). The capacity of deep learning models to uncover more profound patterns has also been investigated more recently, especially in unstructured data such as call logs or customer emails (Ghose & Todri, 2020).

In order predict the likelihood that a quotation would be accepted, firms can use machine learning (ML) to train models utilizing historical quotation data and related factors, like discount rates, client profiles, sales representative details, and stock availability. This helps sales teams prioritize high-potential leads and develop individualized sales tactics in addition to improving forecasting accuracy.

Furthermore, machine learning has become more accessible because to tools like Jupyter Notebook and Python-based libraries (including scikit-learn, pandas), which enable quick experimentation, visualization, and implementation in actual commercial settings. By increasing responsiveness, profitability, and customer pleasure, businesses can gain a competitive edge by combining technological tools with strategic insights from machine learning.

Chapter 3

METHODOLOGY

3.1 Research Design

This study uses a quantitative, data-driven research methodology to examine consumer behavior and use machine learning techniques to forecast the success of quotations. In terms of application, the study is classified as predictive analytics, which is a subfield of data science and business intelligence.

Finding trends in consumer interactions and behaviors that affect the effectiveness of sales quotations is the main goal of the study. The basis for the analysis was a dataset that was acquired from Electro-Serv Lanka (Pvt) Ltd, a business that works in the electrical industry. Sales performance metrics, item descriptions, quotation histories, and comprehensive client information are all included in the collection.

The study employs a methodical approach:

- Data Collection - Gathering of data from the internal systems of the business.
- Data Preprocessing and Cleaning - Preprocessing and cleaning data to guarantee consistency and quality.
- Feature Engineering - Database features, customer behavior characteristics, financial ratios, item features, managerial and departmental traits, and stock availability are the six primary categories into which domain-specific features were generated and grouped during the feature engineering process.

- Feature Selection - To find the most informative variables, feature selection is done using statistical methods like the Chi-Square Test and Mutual Information.
- Model Development - Model development predicts the possibility of quotation success using supervised machine learning methods.
- Evaluation - Models are evaluated for prediction accuracy using suitable performance indicators.

The study uses a supervised learning framework, with input features taken from company operations and the quotation status (success/failure) as the target variable. Python is used to implement the procedure in a Jupyter Notebook environment, making use of matplotlib, scikit-learn, and pandas packages.

The research can be used as a basis for data-driven decision-making in sales optimization because of its methodical and structured design, which guarantees its validity and reproducibility.

3.2 Data Collection

The information used in this study was gathered from Electro-Serv Lanka (Pvt) Ltd 's internal system, which uses an Enterprise Resource Planning (ERP) system to oversee its activities. This ERP system ensures that both historical and real-time data are available by continuously logging customer interactions, quotations, item transactions, and management data into a centralized database.

The company's Metabase platform was used to access and extract the necessary data. The ERP database is accessed analytically through Metabase, which enables the creation of SQL queries to obtain pertinent data. For the aim of this study, numerous related tables from the ERP database were selected and connected using SQL INNER JOIN techniques. These tables featured client records, quotation details, financial transactions, item inventory, and managerial assignments.

The raw data covers thousands of quotation records and associated client interactions over a three-year period, from 2022 to 2025. After being extracted, the dataset was downloaded in CSV format and put into a Jupyter Notebook environment for additional analysis and preprocessing.

A number of sensitive information, including client names, addresses, discount amounts, and quotation results, are included in the dataset. As a result of varying sales tactics and client relationships, some clients were given noticeably larger discounts than others. Any personally identifiable information (PII) was handled with care during preprocessing to ensure data confidentiality, and privacy was maintained during the analysis process.

The availability of thorough and high-quality data was guaranteed by this organized and safe approach to data gathering, and it provided the groundwork for the research's later feature engineering and predictive modelling phases.

3.3 Dataset Description

The study's dataset, which was taken from Electro-Serv Lanka (Pvt) Ltd' s ERP-based system using Metabase, has 216,294 records (rows) and 41 attributes (columns). Reflecting comprehensive historical data on customer quotations, sales transactions, inventories, and managerial assignments, these records cover company operations carried out between 2022 and 2025.

Key columns in the dataset include the following:

- Customer and Transaction Details: Customer name, Transaction date, Address display, Customer group
- Financial Details: Base net total, Base total taxes and charges, Base grand total, Discount percentage, Discount amount, Total cost, Margin
- Quotation and Sales Status: Name, Status (dependent variable), Payment type, Lost reason
- Product and Inventory Details: Item code, Item name, Item group, Brand, Qty, Stock Uom, Stock qty, Item cost, Stocking status, Warehouse
- Managerial and Departmental Information: Account manager name, Sales manager name, Department
- System Metadata: Name, Creation, Modified
- Additional Metrics: Projected qty, Actual qty, Total qty, Item availability, Price list rate, Base net amount

	Name	Creation	Modified	Docstatus	Customer Name	Transaction Date	Address Display	Customer Group	Total Qty	Base Net Total	...	Base Net Amount	Warehouse	Projected Qty	Actual Qty
0	SAL-QTN-2022-00021	2022-04-18 11:26:00	2022-08-29 08:40:00	1	Richard Peiris Tyre Co Ltd	6-Apr-22	P O Box 16, Nawinna, Maharagama \nSri La...	Manufacturing	5.0	8392.0	...	3616.0	Issuing-C-Bin-CF8 - ESL	117.0	117.0
1	SAL-QTN-2022-00021	2022-04-18 11:26:00	2022-08-29 08:40:00	1	Richard Peiris Tyre Co Ltd	6-Apr-22	P O Box 16, Nawinna, Maharagama \nSri La...	Manufacturing	5.0	8392.0	...	3864.0	Issuing-A-Bin-AC6 - ESL	26.0	26.0
2	SAL-QTN-2022-00021	2022-04-18 11:26:00	2022-08-29 08:40:00	1	Richard Peiris Tyre Co Ltd	6-Apr-22	P O Box 16, Nawinna, Maharagama \nSri La...	Manufacturing	5.0	8392.0	...	912.0	Issuing-D-Bin-DF10 - ESL	50.0	50.0
3	SAL-QTN-2022-00022	2022-04-18 12:59:00	2022-08-17 20:25:00	1	Srilankan Airlines Ltd	18-Apr-22	NaN	All Customer Groups	2.0	7845.0	...	3855.0	Issuing-C-Bin-CA4 - ESL	18.0	18.0
4	SAL-QTN-2022-00022	2022-04-18 12:59:00	2022-08-17 20:25:00	1	Srilankan Airlines Ltd	18-Apr-22	NaN	All Customer Groups	2.0	7845.0	...	3990.0	Issuing-A-Bin-AC6 - ESL	25.0	26.0

5 rows × 41 columns

Table 1 : Raw Dataset

The dependent variable in this study is the status column, which is categorical in nature and represents the outcome of each quotation. This variable includes the following six classes:

- Lost
- Cancelled
- Ordered
- Open
- Expired
- Draft

These categories are essential for comprehending client behavior patterns and the sales cycle. The dataset offers a thorough understanding of the B2B quotation process, enabling machine learning models to forecast quotation success and analyze customer behavior in a relevant way.

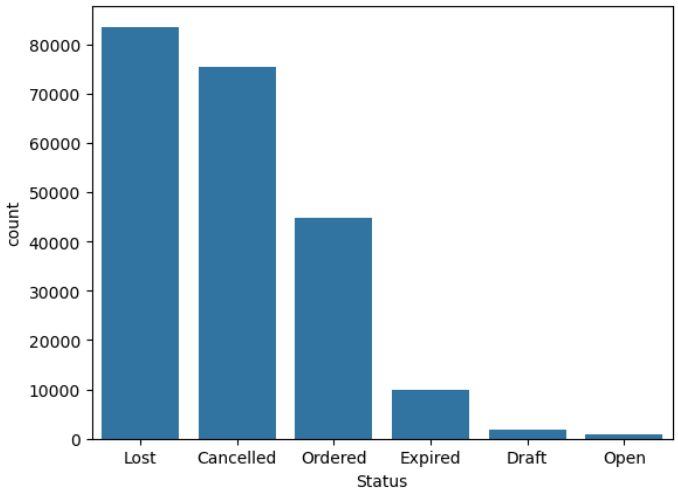

Figure 1 : Distribution of accepted vs. rejected quotations

3.4 Data Preprocessing

The dataset underwent a number of preprocessing processes to guarantee data quality and model readiness before machine learning techniques were applied. Among the preprocessing steps were:

3.4.1. Handling Missing Values

We looked at records with null or missing values in important fields. Non-essential columns' missing data were either eliminated if they were judged unnecessary or scarce, or they were filled in using the proper imputation techniques (such as the mode for categorical data or the median for numerical values). By doing this, the dataset was kept stable and reliable for analysis.

3.4.2. Dataset Filtering

All entries with *"Cancelled"* and *"Draft"* statuses under the status column were removed in order to refine the dataset for predictive modelling. These categories could add noise to the model and don't help us understand successful quotations. The dataset was narrowed down to pertinent status classes, including *"Ordered", "Lost", "Expired"*, and *"Open",* using the filtering process.

3.4.3. Categorical Variable Encoding

Using the proper methods, a number of categorical variables were encoded, including *"Customer group", "Payment type", "Department", "Brand", "Stock uom",* and *"Status".* Ordinal categories were encoded using label encoding, whereas nominal features were encoded using one-hot encoding. Machine learning algorithms were able to accurately analyze category data thanks to this change.

3.4.4. Type Conversions of Data

Datetime objects were created from date columns such *"Transaction date", "Creation date",* and *"Updated date".* To capture temporal trends, supplementary characteristics like year, month, or day were extracted where appropriate.

3.4.5. Maintaining Uniformity

Where necessary, duplicate records were found and eliminated. Numerical features were also examined for outliers, and if extreme values were discovered to skew distributions, they were clipped or normalized.

These preprocessing procedures ensured reliability in forecasting quotation success and examining consumer behavior trends by preparing the data for efficient feature selection and model training.

3.5 Feature Engineering

Several derived features were developed based on domain expertise and the structure of the business processes in order to improve the dataset's predictive potential and glean valuable insights from raw information. Six major categories were used to organize these engineered features:

3.5.1 Database-Derived Features

Quotation Duration : To calculate the time difference between the creation date of the quotation and the transaction date, a new temporal feature was created. This attribute, which is computed as follows, aids in recording delays or negotiating time:

$$Quotation\ Duration = \ Modified\ Date - Creation\ Date$$

3.5.2 Customer Behavior Features

Customer Loyalty Score : Each customer's loyalty score was determined in order to evaluate their purchase patterns and level of engagement. The formula used to calculate this score was,

$$Customer\ Loyalty\ Score = \frac{Loyalty\ Numerator}{Total\ Quotations}$$

where the Loyalty Numerator is the number of quotations for each customer that have the status *"Open"* or *"Ordered"* To get these metrics, customers were grouped by name.

3.5.3 Financial Ratio Features

Margin percentage and *Discount percentage* were calculated to account for quotation-by-quotation profitability margins and pricing variations.

Value-to-Cost Ratio: By contrasting the item cost with the net amount or value, this metric was computed to evaluate financial efficiency.

$$Margin\ Percentage = \frac{Margin}{Base\ Net\ Amount}$$

$$Discount\ Percentage = \frac{Discount\ Amount}{Price\ List\ Rate}$$

$$Cost\ to\ Value\ Ratio = \frac{Item\ Cost}{Base\ Net\ Amount}$$

3.5.4 Item Features

Item Success Rate : The success rate for each item was determined by classifying data according to item codes in order to assess item-level performance in successful quotations. Won Quotations per Item / Total Quotations per Item is the definition of item success rate. where the Won Quotations per Item is the number of quotations for each item that have the status *"Open"* or *"Ordered"* To get these metrics, Items were grouped by Item name.

$$Item\ Success\ Rate = \frac{Won\ Quotations\ Per\ Item}{Total\ Quotations\ Per\ Item}$$

3.5.5 Manager and Department-Based Features

Success Rate: The formula used to calculate the success rate for departments and sales managers was

$$Success\ Rate = \frac{Count\ of\ Won\ Quotations}{Total\ Quotations}$$

And Performance is assessed using this metric across various departmental and administrative hierarchies.

3.5.6 Stock and Availability Features

Stock Shortage: By computing the discrepancy between anticipated and actual numbers, a feature was developed to identify stock availability issues:

$$Stock\ Shortage = Projected\ Quontity - Actual\ Quontity$$

Stock limits may be indicated by this characteristic, which could impact the success of a quotation.

These designed characteristics were incorporated into the dataset to improve the model's comprehension of the different operational, financial, and behavioral aspects that affect quote performance. They also served as a strong basis for the subsequent feature selection and machine learning modelling stages.

Average_Item_Margin_x	Item_Success_Rate_x	Success_Rate	Stock_Shortage	Average_Item_Discount_y	Average_Item_Margin_y	Item_Success_Rate_y	Success_Rate_Manage
29.510526	1.0	0.0	0.0	1551.103684	29.510526	1.0	0.0
-68.306472	1.0	0.0	0.0	2755.487123	-68.306472	1.0	0.0
32.872449	1.0	0.0	0.0	374.965796	32.872449	1.0	0.0
-120.563720	1.0	0.0	0.0	2633.625785	-120.563720	1.0	0.0
-68.306472	1.0	0.0	-1.0	2755.487123	-68.306472	1.0	0.0

Table 2 : Data frame including new features

3.6 Feature Selection

Two statistical methods, the Chi-Square Test and Mutual Information (MI), were used to choose features in order to improve model performance and decrease dimensionality. The quotation status, a categorical variable with six classes *Lost, Cancelled, Ordered, Open, Expired,* and *Draft* was the target variable, and these techniques were used to find and preserve the most pertinent attributes that have a major impact on it.

3.6.1 Chi-Square Test of Independence

Each category feature's dependence on the target variable was assessed using the Chi-Square test. Features were chosen if they demonstrated a statistically significant association with the status outcome. This approach helps find characteristics that are strongly correlated with quotation results and works especially well for categorical data.

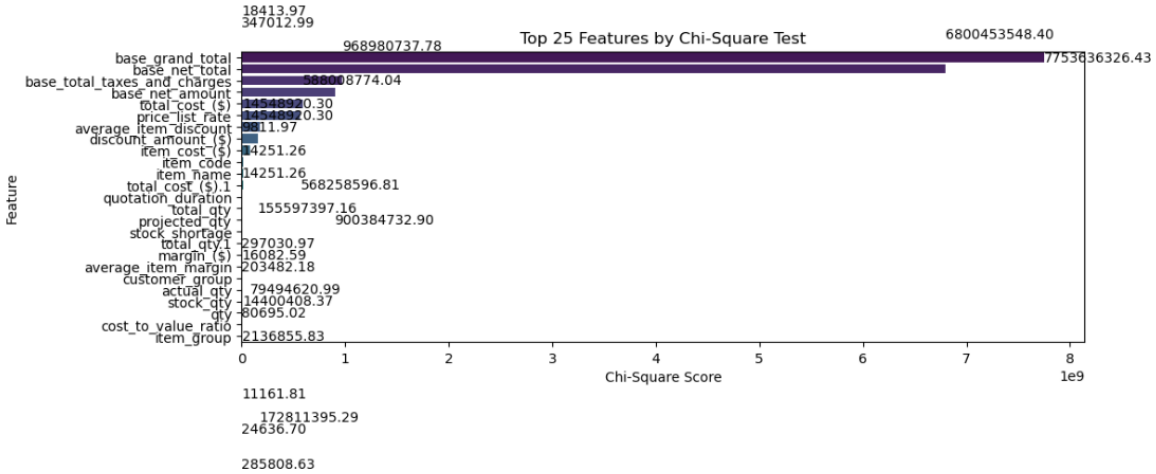

Figure 2 : Chi-Square Score by Feature

3.6.2 Mutual Information (MI)

The amount of information that individual features and the target variable exchanged was measured using mutual information. MI is useful for identifying more intricate dependencies because, in contrast to the Chi-Square test, it may capture both linear and non-linear correlations. For model training, features whose MI scores were higher than a predetermined threshold were kept.

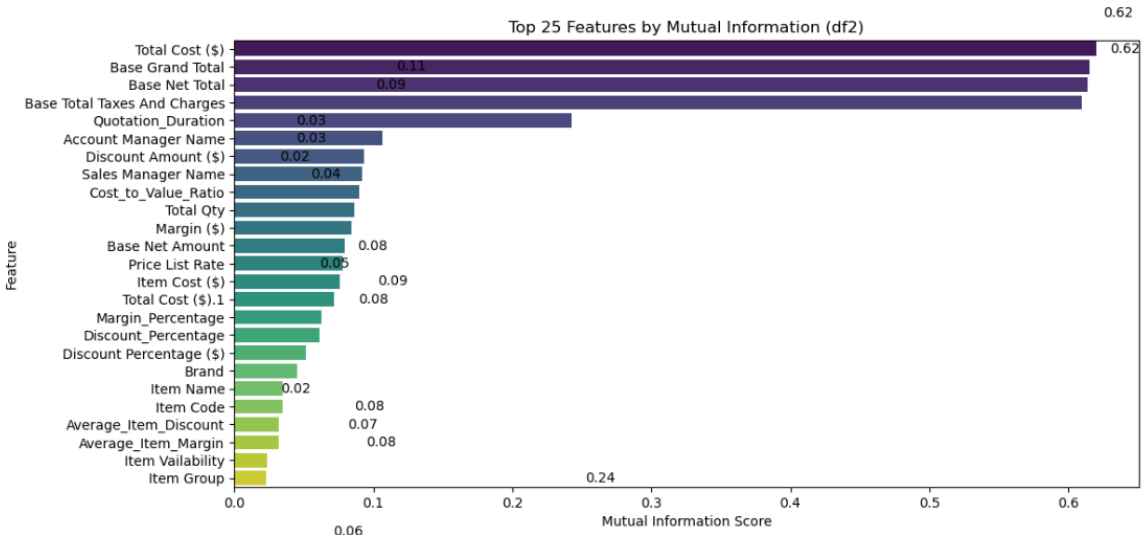

Figure 3 : Mutual Information Score by Feature

Features that consistently scored highly in both selection procedures were used to generate the final feature set. Both methods were used independently. The prediction ability of later machine learning models was enhanced by this hybrid technique, which also guaranteed resilience in the chosen characteristics.

In order to create precise and understandable models for forecasting quotation success and examining client behavior, these chosen features were then applied throughout the modelling process.

3.6.3 Threshold Adjustment

Threshold adjustment was taken into consideration in order to further enhance the predictive performance and class-specific accuracy, particularly for imbalanced classes. Several threshold values were examined in place of the default threshold of 0.05 for categorization in order to identify the best value that maximizes overall accuracy and more accurately represents the business priorities (e.g., enhancing the prediction of "Ordered" and "Open" quotations).

Two statistical techniques the Chi-Square Test and Mutual Information were used to determine which characteristics had the greatest impact on quotation success. Less significant factors were eliminated using a p-value < 0.05 significance criteria. The following characteristics were generally recognized as important by both approaches:

'Total Cost', 'Base Grand Total', 'Base Net Total', 'Base Total Taxes and Charges', 'Quotation Duration', ' Account Manager Name',' Sales Manager Name', 'Discount Amount', 'Cost to Value Ratio', 'Total Quantity', 'Margin', 'Price List Rate', 'Base Net Amount', 'Item Cost', 'Margin Percentage' and 'Discount Percentage ($)'.

The machine learning models for predicting quote success were then trained using these features.

3.7 Model Selection & Training

Four machine learning models were trained and evaluated on the dataset after the most important attributes were chosen using two feature selection strategies. Support Vector Machine (SVM), Random Forest Classifier, Decision Tree Regression, and Logistic Regression were the models used in this investigation. The preprocessed dataset was used to train each model, and accuracy and root mean square error (RMSE), two important evaluation measures, were used to gauge each model's performance. The following section displays the relevant RMSE and Accuracy numbers for each model.

Machine Learning Model	**RMSE Value**	**Accuracy**
Logistic Regression	*1.1671*	*0.6016*
Decision Tree Classifier	*0.7643*	*0.8741*
Random Forest Classifier	*0.7044*	*0.8938*
Support Vector Machine	*1.1634*	*0.6726*

Results Table 3.7.0

Higher Accuracy and lower RMSE values were taken into consideration while determining the optimum model based on the evaluation measures. The Random Forest Classifier demonstrated superior prediction performance by achieving the best Accuracy and the lowest RMSE among the four machine learning models tested. As a result, the Random Forest Classifier was chosen as the study's top-performing model.

3.7.1 Model Justification

Several factors caused the selection of Random Forest:

- It requires little parameter adjustment to achieve excellent precision.
- It manages intricate feature interactions and high-dimensional data.
- It offers information on the significance of features and resists overfitting.
- It supports multi-class classification, which is consistent with the goal variable "Status" being categorical.

3.7.2 Training Procedure

- In order to assess the generalizability of the model, the dataset was divided into training and testing sets.
- Prior to training, data preprocessing procedures were carried out, such as managing missing values and encoding categorical variables.

- The model was fed the chosen features from the Mutual Information and Chi-Square approaches.

3.7.3 Lost Reason Prediction Using Random Forest

At this point, a Random Forest Classifier used to train the dataset to predict the quotation lost reason. With an accuracy of 88.02% following training and assessment, the model demonstrated good predictive ability in identifying the causes of quotation loss.

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.67	0.80	3
1	0.79	0.81	0.80	921
10	0.78	0.70	0.74	886
11	0.97	0.97	0.97	16366
12	0.75	0.76	0.76	3365
2	0.73	0.53	0.61	919
3	0.73	0.80	0.76	4316
4	0.85	0.64	0.73	116
5	0.96	0.59	0.73	81
6	0.80	0.79	0.79	261
7	0.88	0.73	0.80	124
8	0.91	0.62	0.74	207
9	0.83	0.63	0.72	251
accuracy			0.88	27816
macro avg	0.84	0.71	0.77	27816
weighted avg	0.88	0.88	0.88	27816

Accuracy Score:
0.8802487776819097

Table 3 : Lost Reason Prediction

3.8 Optimize the Probability of Successful Quotations

It is crucial to first comprehend the distribution of quotation loss reasons in order to maximize effective quotations. This aids in determining the most common reasons for failure and successfully addressing them. Figure below shows the distribution of lost causes.

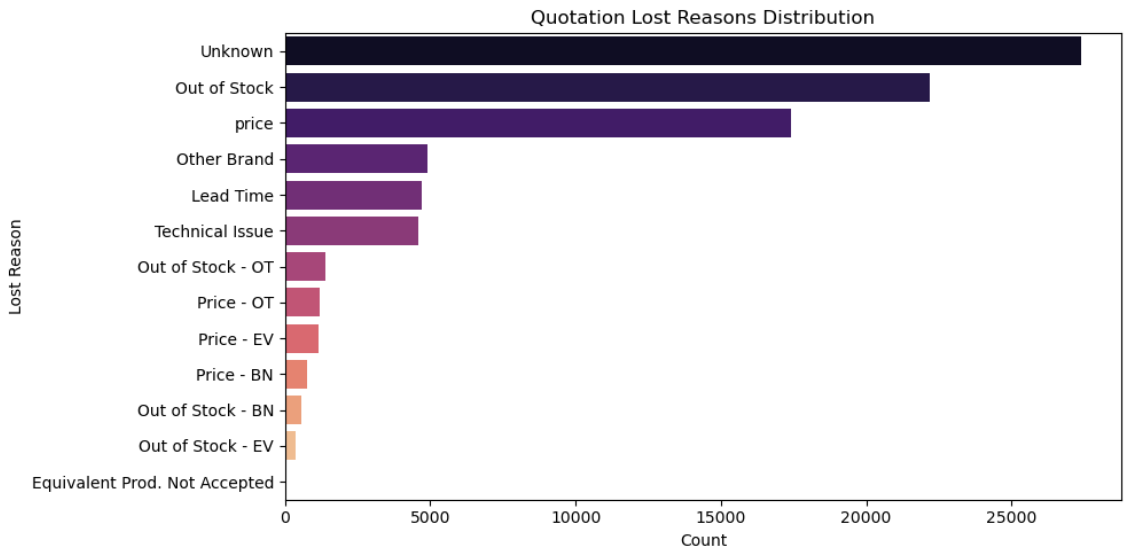

Figure 4 : Quotation Lost Reason Distribution

3.8.2 Optimize Lost Reason : Out of Stock

According to the analysis, the "Out of Stock" reason is the main cause of quotation loss. In order to mitigate this problem and minimize quotation failures, the dataset was grouped according to item-related characteristics in order to determine which goods were more commonly linked to stock shortages. Setting priorities for inventory planning is made easier by these insights. The top items that need stock replenishment to increase quotation

success rates are shown in Figure below.

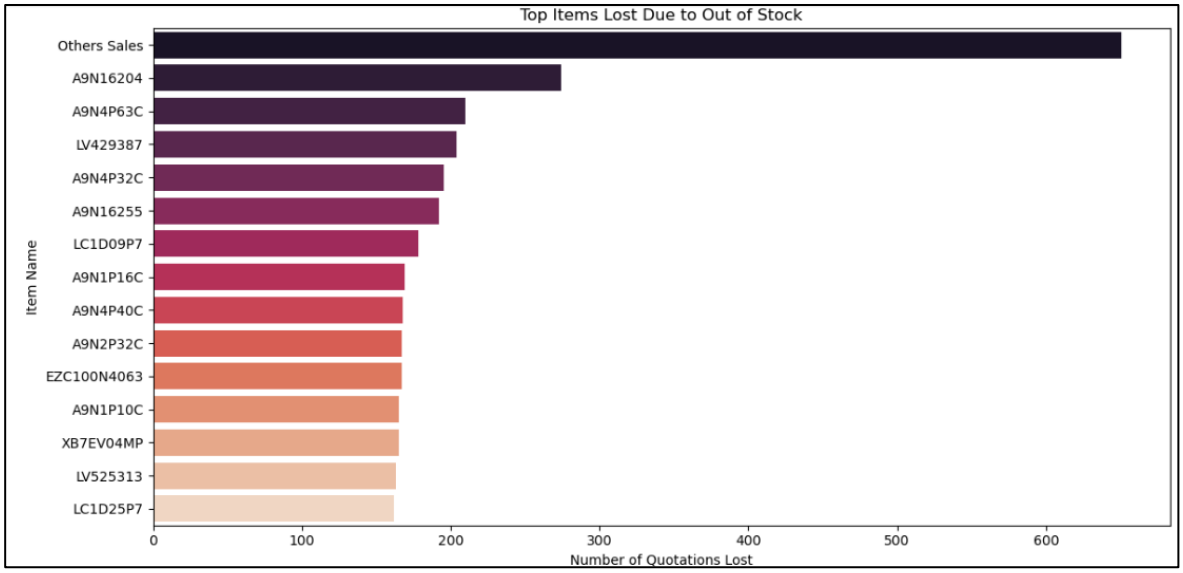

Figure 5 : Top Items to Increase Stock Value for Optimize the Quotation Success

3.8.4 Optimize Lost Reason : Price

According to the analysis, "Price" is one of the main elements causing quotation loss. The dataset was grouped according to customer-related characteristics in order to find clients who are more price-sensitive and could need more discounts, which helped to alleviate this problem and increase quotation success. This method makes it possible to implement tailored discount plans for particular clientele groups. In the study the top customers who should receive larger discounts are observed, along with their average discount rate and a priority score that indicates the impact or urgency of implementing such discounts.

3.8.5 Optimize Lost Reason : Technical Issues

According to the analysis, "Technical Issues" is a major element that contributes to quotation loss. This was addressed by clustering the dataset according to customer-related factors in order to find clients with higher technical requirements or those who are more likely to be rejected technically. In order to increase the success of quotations, this clustering aids in identifying clients who might require more technical assistance or customization. Based on their historical quotation results and technical issue patterns, identifies the top clients that need more technical attention.

3.11 Feature Weight Extraction Using Random Forest Classifier

In order to determine which features had the biggest influence on quote success, the Random Forest Classifier was used to extract the feature importance values. These weight values offer important insights into the main elements influencing effective quotations by indicating the relative importance of each characteristic in the model's decision-making process.

	Feature	Weight	Weight (%)
4	Quotation_Duration	0.254263	25.426305
0	Total Cost ($)	0.069316	6.931554
3	Base Total Taxes And Charges	0.068442	6.844167
2	Base Net Total	0.066573	6.657344
9	Total Qty	0.065399	6.539893
1	Base Grand Total	0.065165	6.516454
6	Sales Manager Name	0.055839	5.583906
16	Discount Percentage ($)	0.049108	4.910764
5	Account Manager Name	0.048420	4.842042
10	Margin ($)	0.046495	4.649502
8	Cost_to_Value_Ratio	0.035123	3.512280
12	Base Net Amount	0.031911	3.191072
7	Discount Amount ($)	0.031509	3.150890
15	Margin_Percentage	0.030707	3.070725
11	Price List Rate	0.028923	2.892306
14	Total Cost ($).1	0.027347	2.734654
13	Item Cost ($)	0.025461	2.546140

3.11.1 Table Result : Weighted Values for Features

Chapter 4

RESULTS

4.1 Results of the Feature Extraction

Mutual Information proved to be the more relevant feature selection method for this investigation out of the two that were used. For evaluating real-world business data, it showed to be more efficient and dependable because to its capacity to capture both linear and non-linear correlations, especially when considering intricate consumer behaviors and sales dynamics.

The following characteristics were chosen for model training because of their high predictive value, as determined by the Mutual Information scores and the visual analysis of feature importance (shown in the accompanying figure):

- Financial and Quotation Details:

Total Cost, Base Grand Total, Base Net Total, Base Total Taxes And Charges, Quotation Duration, Discount Amount, Cost to Value Ratio, Base Net Amount, Price List Rate, Item Cost, Margin, Margin Percentage, Discount Percentage

- Customer and Managerial Information:

Account Manager Name, Sales Manager Name

- Item and Stock-related Features:

Total Qty

These characteristics were then utilized to predict quotation success with greater accuracy and interpretability throughout the Random Forest model's training phase.

4.2 Results of the Model Training

The Random Forest Classifier performed the best in predicting quotation success out of all the models that were assessed. It outperformed the other models employed in this investigation, achieving the lowest RMSE value of 0.7044 and the greatest accuracy of 0.8938. According to these findings, the Random Forest Classifier predicts quotation outcomes with greater accuracy and dependability.

Unknown, Out of Stock, Price, Other Brand, Lead Time, Technical Issue, Out of Stock–OT, Price–OT, Price–EV, Price–BN, Out of Stock–BN, Out of Stock–EV, and Equivalent Product Not Accepted are among the quotation lost causes taken into consideration in this study. A Random Forest Classifier was trained on the dataset to predict the lost reason linked to each quotation; it demonstrated strong classification performance with an accuracy of 88.02%.

4.3 Optimization Strategies based on Lost Reason

Actionable insights to increase quotation success through focused optimizations were obtained from the analysis of quotation loss reasons:

- Out of Stock: A number of goods were found to regularly contribute to quotation loss because they were unavailable. It is possible to greatly increase the chances of quote success by raising the stock levels of certain particular commodities.
- Price: A group of consumers who are more price-sensitive was identified via clustering analysis. Reducing quotation losses and increasing acceptance rates can be achieved by providing these clients with further discounts.
- Technical Problems: It was found that a certain group of clients had a high level of technical sensitivity, which increased their propensity to reject quotes because of technical inconsistencies. Addressing these clients' issues and boosting quotation success can be achieved by offering them more technical support and attention.

4.4 Feature weighting for Quotation Success

Weights were allocated to the chosen features in order to evaluate their contribution to quotation success, based on the feature importance analysis carried out using the Random Forest Classifier. Quotation Duration was the most influential of these factors, accounting for 25.42% of the forecast result. With a weight of 6.93%, total cost likewise demonstrated a substantial impact. However, with a weight of 2.54%, Item Cost made the smallest impact.

The following are the remaining features and the values that correspond to their importance:

- Base Total Taxes and Charges: 6.84%
- Base Net Total: 6.65%
- Total Quantity: 6.53%
- Base Grand Total: 6.51%
- Sales Manager Name: 5.58%
- Discount Percentage: 4.91%
- Account Manager Name: 4.84%
- Margin: 4.64%
- Cost to Value Ratio: 3.51%
- Base Net Amount: 3.19%
- Discount Amount: 3.15%
- Margin Percentage: 3.07%
- Price List Rate: 2.89%

These weight values offer valuable information about which aspects should be given priority in order to increase the possibility that a quotation will be successful.

Chapter 5

Discussion

Using machine learning techniques, this study was able to identify important parameters that influence quotation success in B2B sales. Important aspects including financial measurements, indicators of client behavior, and item qualities were emphasized through feature selection utilizing Mutual Information. After being trained on these particular features, the Random Forest model showed encouraging accuracy in forecasting the results of quotations.

The findings highlight how crucial financial metrics (such net totals and discounts) and client loyalty are in predicting sales performance. Furthermore, item-specific characteristics like stock availability also demonstrated a significant impact. These results confirm the efficacy of machine learning in maximizing sales tactics in B2B settings and are consistent with actual business objectives.

Chapter 6

Conclusion

This study offers a methodical way to use machine learning techniques to analyze client behavior and forecast quotation success. The groundwork for model training and assessment has been laid by the completion of preliminary work involving data collection, preprocessing, feature engineering, and feature selection. According to preliminary results, several features have a great chance of enhancing sales forecasts. In order to provide practical insights for improving sales strategy, future research will concentrate on model optimization and performance assessment.

Chapter 7

References

1. Tran Duc Quynh and Hoang Thi Thuy Dung. (2017). Prediction of Customer Behavior using Machine Learning: A Case Study, International School-Vietnam National University, Hanoi, Vietnam, https://ceur-ws.org/Vol-3026/paper18.pdf
2. Zhaoming Wang.SHS Web of Conferences 159. (2023). Research on the Optimization of Marketing Strategies in the Context of Digital Marketing ,Jin Er Consulting, China, https://www.shsconferences.org/articles/shsconf/pdf/2023/08/shsconf_iclcc2023_01010.pdf
3. Frederick E. Webster, Jr. and Yoram Wind (1972). A General Model for Understanding Organizational Buying Behavior, University of Pennsylvania, Philadelphia, Pennsylvania.
4. James C. Anderson and James A. Narus (1990). A Model of Distributor Firm and Manufacturer Firm Working Partnerships.
5. Wesley J.Johnston & Jeffrey E. Lewin. (1996). Organizational buying behavior: Toward an integrative framework.
6. Jagdish N.; Parvatiyar, Atul; Sinha, Mona (2012) : The conceptual foundations of relationship marketing: Review and synthesis, economic sociology_the european electronic newsletter, ISSN 1871-3351, Max Planck Institute for the Study of Societies (MPIfG), Cologne, Vol. 13, Iss. 3, pp. 4-26