International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 6
ISSN: 2088-5334

# Leveraging Random Forest Algorithm for Enhanced Lead Conversion and Customer Retention

Xin Yi Tan [a], Siew Mooi Lim [a,*], Tong Ming Lim [a], Chi Wee Tan [a], Noor Aida Husaini [a]

[a] *Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia*
*Corresponding author: [*]siewmooi@tarc.edu.my*

***Abstract*—This paper presents the development, implementation, and evaluation of a Random Forest-based Social Lead Scoring Model to address key business challenges in lead generation, customer retention, and optimization of lead management processes. The main goal is to create a strong, data-driven tool to precisely estimate lead conversion probabilities to guide better marketing and sales strategy decision-making. The model uses social metrics and past lead data to estimate conversion probabilities for every lead. The Tkinter library created a user-friendly interface that allows straightforward usage for non-technical business professionals. This model includes two fundamental functions calculate_probability and predict_conversion– which offer practical and pragmatic insights. During the development phase, a thorough cross-valuation was conducted by the model trained on a large dataset, including several lead characteristics, to decrease overfitting and improve the model's predictive performance. Thus, the model scored 89.46%, which is higher than that of conventional lead-scoring techniques. However, there is still room for development, specifically in enhancing its predictive power and reducing overfitting risks on different datasets, although this model has great accuracy. The results highlight the need for data-driven strategies in raising conversion rates and show the possibilities of machine learning in lead management optimization. Including extra data sources and investigating advanced technologies such as deep learning should be conducted by future studies to improve model performance further. The increased accuracy in predictive analytics will give companies a competitive edge in their operations.**

***Keywords*—Machine learning; lead scoring model; social media; random-forest; prediction.**

*Manuscript received 11 Dec. 2023; revised 18 Aug. 2024; accepted 23 Nov. 2024. Date of publication 31 Dec. 2024.*


## I. INTRODUCTION

Lead generation and management are essential for organizational success in today's fast-paced corporate world. Since conventional approaches to lead scoring sometimes rely on subjective assessments and limited demographic information, it has brought about inefficiencies in resource allocation and missed income growth opportunities [1], [2], [3], [4]. To improve lead prioritizing and maximize marketing plans, companies that understand the importance of data-driven approaches have started to look into machine learning (ML) methods [5], [6], [7]. The creation and assessment of ML-based social lead scoring model that uses cutting-edge algorithms to examine social media data and faithfully forecast lead conversions are being investigated in this work [8], [9]. This model intends to provide a complete solution for companies by solving issues in lead generation, customer retention, and lead management [10], [11].

This project wants to build a model considering several elements, including social media behavior, interests, and demographics, to properly forecast lead scores [12], [13], [14]. The accuracy and dependability of the model in leading conversion will be guaranteed by rigorous evaluation using suitable metrics—accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) analysis [15]. Creating an easy-to-use application helps the model be implemented and accessible to various companies and digital marketing agencies.

The model provides various advantages and contributions. Companies not only easily find possible clients depending on their interests and behavior but also improve lead generation [16], [17], [18]. The problem of client retention is solved by spotting at-risk clients and supporting proactive policies meant to keep them [19], [20], [21]. Companies can now concentrate on high-quality leads and best utilize their resources [22], [23], [24] by simplifying lead management systems which guarantees consistent and accurate lead classification [25], [26]. Ultimately, accepting a data-driven approach helps companies acquire a competitive edge, enable

informed decisions, and maximize their marketing efficacy in the modern data-driven environment.

Traditional approaches often result in inefficiencies based on manual, rule-based assessments [27]. Although ML presents interesting solutions, there is still uncertainty about the best ML methods and evaluation criteria. This review emphasizes social lead scoring and the Random Forest (RF) method and looks at the development of lead scoring. Though common, traditional approaches have subjective biases and scaling problems. By contrast, RF, an ensemble learning approach, has become popular for its adaptability and accuracy [28], [29].

Evaluation criteria, including accuracy, precision, recall, and AUC-ROC curve analysis, provide a comprehensive understanding of model performance. Using ML methods, especially RF, has excellent potential to improve lead-scoring accuracy and efficiency in contemporary corporate operations. As shown in Table I, Benhaddou and Leray [30] have presented a lead-scoring model utilizing Bayesian networks (BNs), which are adept at managing uncertainty and drawing from expert knowledge. Employing techniques like parent divorce and NoisyOr, the model incorporates domain expertise while mitigating complexity. Parameter estimation utilizes elicitation, ranking, and analytic hierarchy process methods. Validation against available data shows promising precision and recall. Further enhancements are proposed, advocating expert interactions and a rapid, focused, and incremental learning approach to bolster accuracy and effectiveness.

In 2020, Nygård and Mezei [31] conducted a model performance comparison involving logistic regression (LR), decision tree (DT), RF, and neural network (NN) models. The RF model emerged as the best performer with 69% accuracy, 76% AUC, 69% sensitivity, and 69% specificity. Further analysis categorized data points based on purchase probability estimates from the RF model. Five groups were formed to examine activity levels across different purchase probability groups. This approach facilitates enhanced customer understanding and sales strategies, empowering sales teams to target leads more effectively

This study examines how coupons, especially for bars and restaurants, might boost sales and inspire customer repurchases. The study gathers data using in-vehicle surveys and employs data mining classification methods to ascertain coupon acceptance by customers. J48, Random Tree, and Random Forest decision tree algorithms are compared to predict restaurant and bar coupon acceptance. Although slower to develop, Random Forest performs best with 77% accuracy. Thus, it is perfect for companies looking for more exact insights to maximize coupon distribution and increase customer involvement and sales [32].

Kumar and Hariharanath [33] have employed multiple regression analysis and the chi-square test to assess the relationship between implicit and explicit parameters and lead scores. Implicit parameters predict 85% of score variation, while explicit parameters predict 25%. All parameters demonstrate associations with lead categorization. The study aims to design a lead-scoring model for online education marketing that accurately reflects lead behavior and engagement. This model enhances lead qualification, increasing conversion chances by filtering out unqualified leads based on their scores [33].

Jadli et al. [34] developed a smart lead scoring system using ML algorithms, comparing RF and DT models. RF exhibits superior classification performance, while DT shows competitive results with shorter training times. Cross-validation confirms that RF and DT consistently outperform other models in performance and stability. However, when execution time is considered, DT becomes more optimal than RF. This study offers insights into selecting the most suitable algorithm for lead-scoring systems based on performance and efficiency.

In short, traditional lead scoring methods face inefficiencies due to manual, rule-based assessments, contrasting with the promising solutions offered by ML. Notably, RF and BNs have significant potential for enhancing accuracy and scalability. Studies emphasize the importance of selecting optimal algorithms, such as DT or RF, based on performance and efficiency. Adopting ML techniques holds immense promise for revolutionizing lead scoring in modern business practices.

TABLE I
COMPARISON OF LEAD SCORING MODELS FROM VARIOUS STUDIES

| Ref | Method | Main Findings | Focus |
|---|---|---|---|
| [30] | BNs with NoisyOr | Expert-based BNs. | Expert involvement in BNs for lead scoring. |
| [31] | LR, DT, RF, NN comparison | RF outperforms other models. | Comparison of lead scoring models' performance. |
| [32] | J48, Random Tree, RF | RF shows the best performance with an accuracy of 77%. | Evaluating the effectiveness of in-vehicle coupon recommendations. |
| [33] | Multiple regression, chi-square test | Implicit parameters predict 85%, explicit 25% variation. | Parameter-lead score relationship in the education market. |
| [34] | RF, DT comparison | RF superior, DT shorter training. | ML algorithms' performance comparison for lead scoring. |

## II. MATERIALS AND METHOD

### A. Description of Dataset

The Kaggle dataset, Lead Scoring Dataset, utilized for this study originates from X Education, an online education company. It contains information on leads generated through various marketing channels and contains 9240 entries. It encompasses 37 columns representing different attributes associated with each lead, as shown in Fig. 1 and Fig. 2.

Key variables, such as *Lead Origin* and *Lead Source*, detail how leads were identified and acquired, alongside the crucial *Converted* indicator denoting successful conversions. Additionally, demographic data, including *Country*, *City*, occupation, and insights into preferences and website interactions, enrich the dataset. *Lead Quality* and *Tags* provide further context on lead characteristics. This comprehensive dataset facilitates the development of predictive models to assign lead scores, aligning with X Education's goal of improving conversion rates by prioritizing engagement with high-potential leads.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
```

Fig. 1 DataFrame Overview of X Education from column 0 to 17

```
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Fig. 2 DataFrame Overview of X Education from columns 18 to 36

## *B. Exploratory Data Analysis*

*1) Duplication Check*: A duplication check was conducted to verify the uniqueness of customer IDs (Prospect ID and Lead Number) and ensure data integrity. No duplicate values were found, affirming the integrity of the data, as shown in Fig. 3.

```
# Check duplicates under Prospect ID

duplicate_prospect_ID = lead.duplicated(subset = 'Prospect ID')

print (sum(duplicate_prospect_ID) == 0)
```

```
True
```

```
# Check duplicates under Lead Number

duplicate_LeadNo = lead.duplicated(subset = 'Lead Number')
print(sum(duplicate_LeadNo) == 0)
```

```
True
```

Fig. 3 Checking for duplicates based on 'Prospect ID' and 'Lead Number'

*2) Data Cleaning*: To address missing or irrelevant values, select values across columns were replaced with NULL values, as shown in Fig. 4. Unnecessary columns, which account for over 45% of the total, were dropped to streamline the dataset, as shown in Fig. 5. Imputation strategies were employed to handle missing values in categorical variables. For example, NULL values in City were replaced with the most frequently occurring city name, Mumbai, as shown in Fig. 6.

```
# Converting 'Select' values to NaN.
lead = lead.replace('Select', np.nan)
lead.head()
```

Fig. 4 Converting *'Select'* values to NaN

```
# remove the columns >45% of missing values
# as they do not add on to the analysis
columns = lead.columns
columns_to_drop = []
for i in columns:
    if (100 * (lead[i].isnull().sum() / len(lead.index))) >= 45:
        columns_to_drop.append(i)

lead.drop(columns=columns_to_drop, inplace=True)

print("After : ", lead.shape)
```

```
After :  (9240, 28)
```

Fig. 5 Dropping columns with over 45% missing values

```
lead['City'].mode()
```

```
0    Mumbai
Name: City, dtype: object
```

```
# Mumbai - most frequently occuring value.
# Impute Null values with Mumbai

lead['City'] = lead['City'].replace(np.nan,'Mumbai')
```

Fig. 6 Imputing missing values in the city column with Mumbai

Exploratory data analysis on numerical variables was used to assess data distribution, detect outliers, and treat outliers to ensure data quality. Outliers in numerical features like *TotalVisits* and *Page Views Per Visit* were identified and capped to mitigate their impact on subsequent analyses, as shown in Fig. 7. These steps collectively ensured the dataset's readiness for further modeling and analysis.

In Fig. 8, extreme values beyond the 1st and 99th percentiles have been excluded. This approach mitigates the influence of these extremes on statistical analyses and modeling. Consequently, the dataset's row count has dropped from 9103 to 9020 while the number of columns has stayed constant at 14.

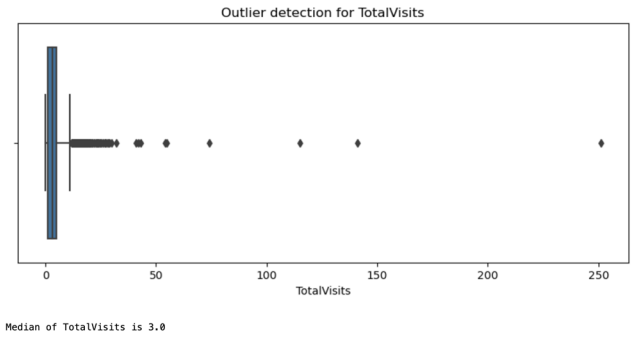


Fig. 7 Boxplot Detection of Outliers in *TotalVisits*

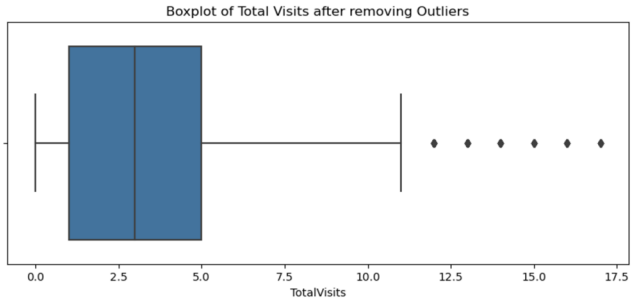


Fig. 8 Boxplot of *TotalVisits* after removing outliers

## C. Modelling

*1) Data Preparation:* The modeling stage mostly concentrated on applying the RF algorithm for predictive modeling on the lead dataset [35]. Data preparation comprised the first steps in which dummy variables—shown in Fig. 9 and Fig. 10—helpfully manage categorical data. The model might effectively capture categorical information through binary representations of categorical variables without adding ordinality or bias. This change enabled compatibility with numerical-input-requiring ML techniques.

```
# getting dummies and dropping the first column
# and adding the results to the master dataframe
dummy = pd.get_dummies(lead[['Lead Origin',
                             'What is your current occupation',
                             'City']], drop_first=True)

dummy = dummy.astype(int)

lead = pd.concat([lead,dummy], axis = 1)
```

Fig. 9 Generating Dummy Variables: One-Hot Encoding Categorical Variables

```
<class 'pandas.core.frame.DataFrame'>
Index: 8953 entries, 0 to 9239
Data columns (total 57 columns):
 #   Column                                              Non-Null Count  Dtype
---  ------                                              --------------  -----
 0   Converted                                           8953 non-null   int64
 1   TotalVisits                                         8953 non-null   float64
 2   Total Time Spent on Website                         8953 non-null   int64
 3   Page Views Per Visit                                8953 non-null   float64
 4   Lead Origin_Landing Page Submission                 8953 non-null   int32
 5   Lead Origin_Lead Add Form                           8953 non-null   int32
 6   Lead Origin_Lead Import                             8953 non-null   int32
 7   What is your current occupation_Housewife           8953 non-null   int32
 8   What is your current occupation_Other               8953 non-null   int32
 9   What is your current occupation_Student             8953 non-null   int32
 10  What is your current occupation_Unemployed          8953 non-null   int32
 11  What is your current occupation_Working Professional 8953 non-null  int32
 12  City_Other Cities                                   8953 non-null   int32
 13  City_Other Cities of Maharashtra                    8953 non-null   int32
 14  City_Other Metro Cities                             8953 non-null   int32
 15  City_Thane & Outskirts                              8953 non-null   int32
 16  City_Tier II Cities                                 8953 non-null   int32
 17  Specialization_Banking, Investment And Insurance    8953 non-null   int32
 18  Specialization_Business Administration              8953 non-null   int32
 19  Specialization_E-Business                           8953 non-null   int32
...
 55  Tags_Ringing                                        8953 non-null   int32
 56  Tags_Will revert after reading the email            8953 non-null   int32
dtypes: float64(2), int32(53), int64(2)
memory usage: 2.2 MB
```

Fig. 10 Final Dataset with One-Hot Encoded Variables

*2) Hyperparameter Tuning for RandomForestClassifier:*

Following the first data preparation, predictive models were developed using a Random Forest Classifier (RFC). Our dataset found the RF model to be the best fit since it could handle categorical data and resist feature scaling. Later assessments using metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared ($R^2$), and accuracy scores found the model's efficacy in spotting data correlations and generating accurate forecasts.

Key hyperparameters, including n_estimators (number of trees), max_depth (maximum depth of trees), and min_samples_split (minimum samples required to split a node), were tuned to increase predictive accuracy and lower error rates. The data was split, with 70% used for training and 30% for testing [36], [37]. Standardizing numerical features by feature scaling guarantees a constant scale over the dataset.

We investigated the accuracy of our RFC across several values of n_estimators, max_depth, and min_samples_split using a validation curve analysis, as depicted in Fig. 11, Fig. 12, and Fig. 13, so optimizing the performance of our RFC. This process sought to maximize classifier accuracy using an ideal balance free from overfitting or underfitting. We improved the model's predictive performance by determining the most sensible hyperparameter values.

The RFC model (rfc2) was refined for best performance with designated parameter values. For a sizable collection of decision trees, n_estimators was set at 1000—a min_samples_split of 8 controlled overfitting by requiring a minimum number of samples in nodes. With max_depth at 15, the depth of decision trees was limited to manage complexity. Additionally, random_state was fixed at 0 for result reproducibility, as shown in Fig. 14. These choices aimed to balance model complexity and predictive accuracy, enhancing overall effectiveness.

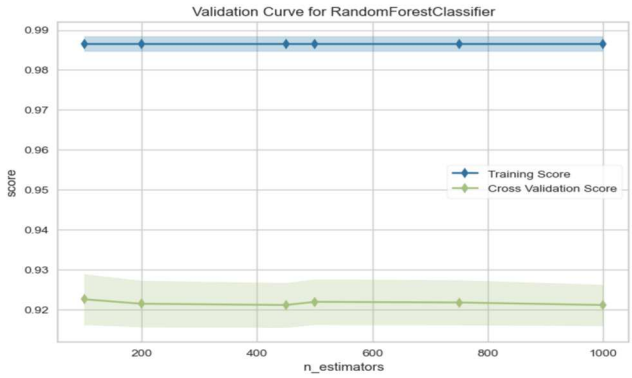


Fig. 11 Validation Curve for *n_estimators*

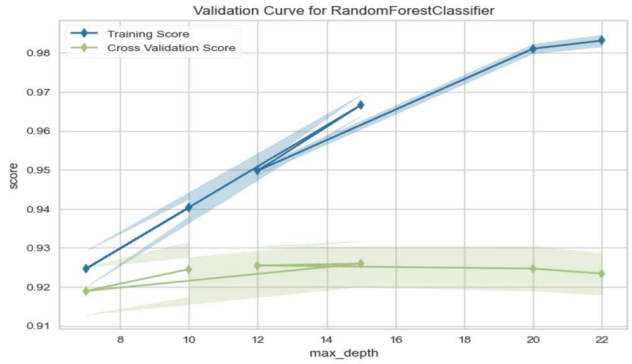


Fig. 12 Validation Curve for *max_depth*

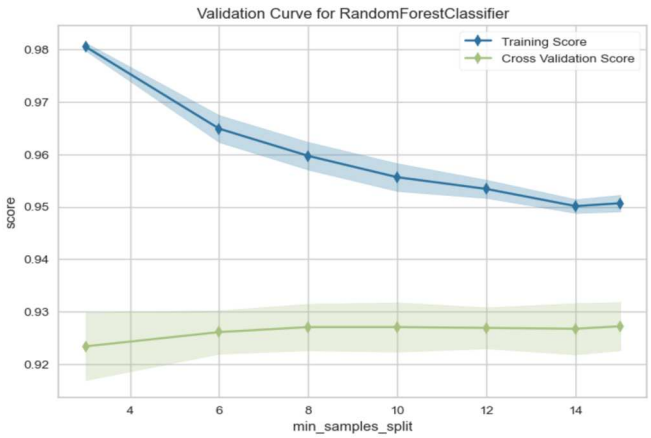


Fig. 13 Validation Curve for *min_samples_split*

```
rfc2 = RandomForestClassifier(n_estimators = 1000,
                              min_samples_split = 8,
                              max_depth = 15,
                              random_state = 0)

rfc2.fit(X_train, y_train)

pred2 = rfc2.predict(X_test)

accuracy_score(pred2, y_test)
```

```
0.8931496649292628
```

Fig. 14 Training Random Forest with Optimized Parameters

*3) Analysis*: In Fig. 15, feature importance analysis provided insights into the significant predictors influencing lead conversion. By identifying and prioritizing influential features like Total Time Spent on the Website and other specific tags, the model could better understand customer behavior and preferences, leading to more accurate predictions.

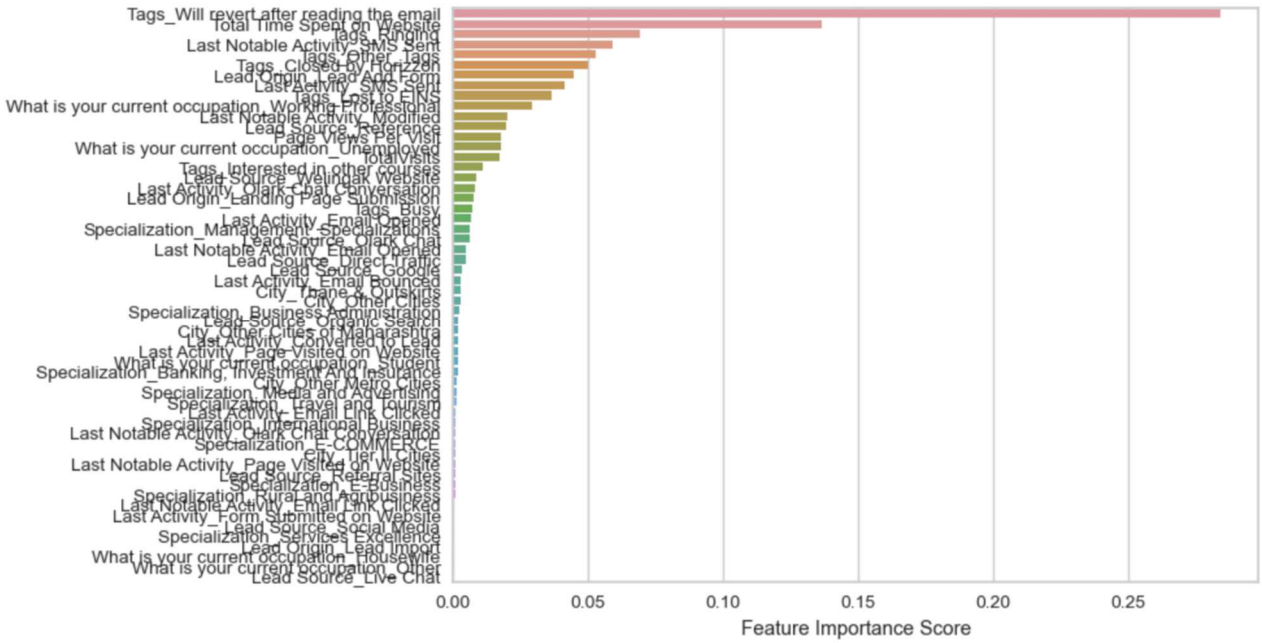


Fig. 15 Feature Importance Score on Dataset's Variables

| | Features | VIF |
|---|---|---|
| 9 | Last Notable Activity_SMS Sent | 6.41 |
| 7 | Last Activity_SMS Sent | 6.06 |
| 0 | Converted | 5.32 |
| 4 | What is your current occupation_Unemployed | 5.16 |
| 3 | Lead Origin_Lead Add Form | 4.28 |
| 6 | Lead Source_Reference | 4.04 |
| 14 | Tags_Will revert after reading the email | 3.30 |
| 2 | Page Views Per Visit | 3.10 |
| 1 | Total Time Spent on Website | 2.55 |
| 8 | Last Notable Activity_Modified | 2.23 |
| 5 | What is your current occupation_Working Profes... | 1.71 |
| 10 | Tags_Closed by Horizzon | 1.48 |
| 13 | Tags_Ringing | 1.47 |
| 12 | Tags_Other_Tags | 1.36 |
| 11 | Tags_Lost to EINS | 1.23 |

Fig. 16 Informative Features with VIF Scores

Additionally, Variance Inflation Factor (VIF) analysis was employed to address multicollinearity issues, ensuring the reliability of coefficient estimates in the regression model, as shown in Fig. 16 and Fig. 17. Predictions on both training and testing datasets demonstrated the model's high accuracy and performance in classifying leads into converted and unconverted categories.

```
features_to_remove = vif.loc[
    vif['VIF'] > 5,'Features'
    ].values
features_to_remove = list(features_to_remove)
print(features_to_remove)
```

Fig. 17 Identifying Features for Removal with VIF scores > 5

In short, the RF algorithm proved a powerful tool for predictive modeling in lead conversion analysis. Its ability to handle categorical data, feature scaling robustness, and hyperparameter tuning flexibility contributed to its success in accurately predicting lead outcomes.

Our lead conversion prediction system incorporates two crucial functions, *calculate_probability* and *predict_conversion*, seamlessly integrated into a user-friendly interface using the Tkinter library. The

calculate_probability(user_input) function runs in Fig. 18 under evaluation of the likelihood of lead conversion depending on user-provided input. It makes use of a dictionary, user input, with feature-value pairs. This function computes a weighted sum of particular features by using feature importance scores obtained from our ML model, so offering an estimate of the probability of lead conversion.

Function 1: Calculate Probability of Lead Conversion
Function : calculate_probability(user_input)
Input : user_input
Output : Probability
1. Initialize weighted_sum to 0
2. For every feature, value pair found in user-input:
   Add the weighted sum by first computing the product of the feature value and its importance score.
3. Calculate the return weighted sum as the lead conversion probability.

Fig. 18 *calculate_probability()* in pseudocode

Predict_conversion (), meanwhile, as shown in Fig. 19, Predict_conversion () coordinates the prediction process inside our Tkinter interface. It compiles user inputs, cleans them, and then uses calculate_probability to ascertain the conversion probability. This ability then produces practical suggestions depending on the likelihood of the outcome. It also shows prediction results, including conversion probability and recommendation, in a new Tkinter window, visualizing the relevance of particular features.

Function 2: Predict Lead Conversion Rate and Recommendation
Function: predict_conversion()
Input: None
Output: Tkinter window with prediction results and recommendations
1. Get user inputs via the Tkinter interface for lead characteristics.
2. Prepare the user inputs, such as a dictionary (user_input) with feature-value pairs results.
3. Calculate the lead conversion probability using the calculate_probability(user_input) function.
4. Generate recommendations based on the probability of the outcome
5. Visualize feature importance scores and prediction results in a Tkinter window
6. Display the prediction results, including conversion probability and recommendations, in the Tkinter window

Fig. 19 *predict_conversion()* in pseudocode

These functions are critical components of our lead conversion prediction system, empowering users with an interactive platform for inputting data, receiving predictions, and obtaining actionable insights. Their integration into a user-friendly Tkinter interface enhances usability and facilitates seamless interaction with our prediction model.

## III. Results and Discussion

The model's performance metrics, particularly in the training set, show that the model has learned the training data well and can make precise predictions. As shown in Table II, the accuracy of the train set is about 97.26%. However, a slight decline in these metrics on the test data, which is 89.46%, signals a potential need for model refinement to enhance generalization.

TABLE II
Comparison of Train and Test Sets in RF Algorithms

| Dataset | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Train | 97.26% | 93.67% | 99.48% |
| Test | 89.46% | 82.78% | 93.38% |

Otherwise, the feature importance analysis sheds light on critical predictors, such as *Tags_Will revert after reading the email* and *Total Time Spent on Website*, underscoring their significant impact on lead conversion, as shown in Fig. 20. These results highlight how well the RF algorithm detects complicated relationships in the data and points up important elements causing lead conversions.

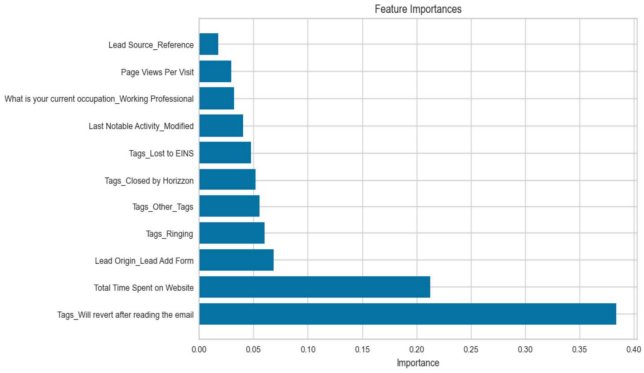


Fig. 20 Feature Importance Score After Modelling

The noted disparity in the RF model's performance on the test and training sets begs questions regarding possible overfitting. This discrepancy implies that the model might have become unduly sensitive to the subtleties of the training data, so restricting its capacity to extend to unspoken facts broadly. A few methods could solve this problem: investigating, including regularity, lowering model complexity, and obtaining more varied data. Modern methods for managing deviations, anomalies, selection techniques, or feature engineering can strengthen the model's generalizing and resilience powers. The RF model's performance could be raised to guarantee dependability and accurate forecasts for practical situations when the model is improved, and possible overfitting resilience is tackled via several approaches.

## IV. Conclusion

In conclusion, the creation and assessment of the RF-based Social Lead Scoring Model made a breakthrough in tackling important corporate concerns related to lead management, customer retention, and lead generation. Due to the project's extraordinary accuracy score of 89.46% attained during a test of its capacity to estimate lead conversions precisely, this project is deemed successful. This level highlights the model's effectiveness in lead conversions and viable application in real-world scenarios.

Although the model shows encouraging outcomes, it also emphasizes improving its functionality for practical uses. The obvious inconsistency in the model's performance during the test and training datasets raises the question of possible overfit. To solve this problem, future model versions can use techniques such as regularizing, lowering model complexity,

and obtaining more varied data. Hence, the robustness of the model and generalizing powers can be strengthened by sophisticated approaches for managing disparities or outliers and feature engineering or selection techniques.

Companies will be provided with a competitive edge in their particular sectors as the knowledge provided by this project contributes to strategic decisions. ML methods, such as RF algorithms, let companies make data-driven choices, improving their lead generation, lead management, and customer retention processes [38], [39], [40]. Moreover, the evolution of intuitive interfaces, including those produced with the Tkinter library, guarantees that these realizations are easily reachable and applicable to interests all around the company.

Predicting lead conversion probability uses a graphical user interface (GUI) shown in Fig. 21 and Fig. 22. Demonstrations, using sample inputs, confirm the system's functionality and provide forecasts derived from the given data.

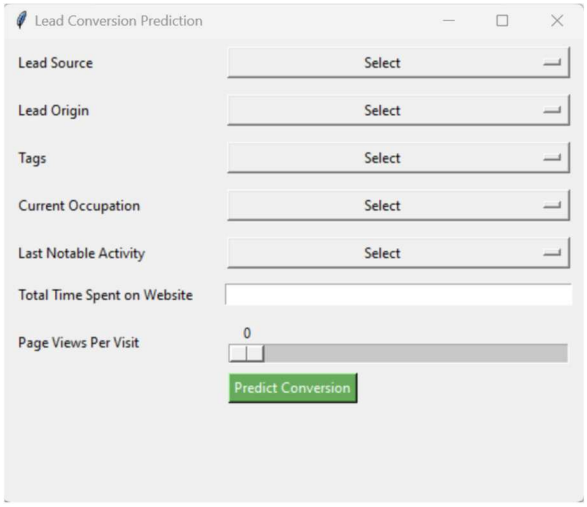


Fig. 21 User Interface

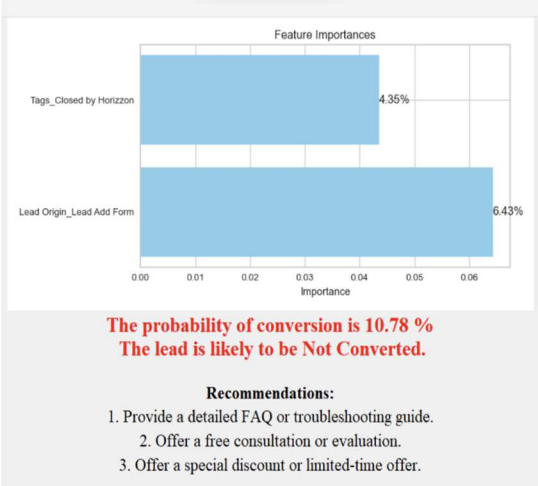


Fig. 22 Results

The RF-based Social Lead Scoring Model is very valuable for businesses as companies can use it to raise their lead conversion prediction capacity. By leveraging the predictive capability of ML, businesses can fully maximize their marketing plans, give top priority to high-quality leads, and finally stimulate income growth and profitability [41]. In terms of sustainable business models and innovation, this project paves the way for future developments in lead scoring and predictive analytics [42]. It highlights the significance of data-driven decision-making in contemporary corporate environments.

## REFERENCES


[1] H.-P. Fu and T.-S. Chang, "An analysis of the factors affecting the adoption of cloud consumer relationship management in the machinery industry in Taiwan," *Information Development*, vol. 32, no. 5, pp. 1741–1756, Jul. 2016, doi: 10.1177/0266666915623318.

[2] M. Wu, P. Andreev, and M. Benyoucef, "The state of lead scoring models and their impact on sales performance," *Information Technology and Management*, vol. 25, no. 1, pp. 69–98, Feb. 2023, doi: 10.1007/s10799-023-00388-w.

[3] A. Powell, C. H. Noble, S. M. Noble, and S. Han, "Man vs machine," *European Journal of Marketing*, vol. 52, no. 3/4, pp. 725–757, Feb. 2018, doi: 10.1108/ejm-10-2015-0750.

[4] J. Bryan and P. Moriano, "Graph-based machine learning improves just-in-time defect prediction," *PLoS ONE*, vol. 18, no. 4, p. e0284077, Apr. 2023, doi: 10.1371/journal.pone.0284077.

[5] P. Kubiak and S. Rass, "An Overview of Data-Driven Techniques for IT-Service-Management," *IEEE Access*, vol. 6, pp. 63664–63688, 2018, doi: 10.1109/access.2018.2875975.

[6] V. Guerola-Navarro, H. Gil-Gomez, R. Oltra-Badenes, and P. Soto-Acosta, "Customer relationship management and its impact on entrepreneurial marketing: a literature review," *International Entrepreneurship and Management Journal*, vol. 20, no. 2, pp. 507–547, Jun. 2022, doi: 10.1007/s11365-022-00800-x.

[7] J. R. Dias and N. Antonio, "Predicting customer churn using machine learning: A case study in the software industry," *Journal of Marketing Analytics*, Dec. 2023, doi: 10.1057/s41270-023-00269-9.

[8] K. Zahoor, N. Z. Bawany, and S. Hamid, "Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning," *2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1–6, Nov. 2020, doi: 10.1109/acit50332.2020.9300098.

[9] R. Katarya, A. Gautam, S. P. Bandgar, and D. Koli, "Analyzing Customer Sentiments Using Machine Learning Techniques to Improve Business Performance," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 182–186, Dec. 2020, doi:10.1109/icacccn51052.2020.9362895.

[10] M. Rahman et al., "Revolutionizing Consumer Power Management: Unveiling Power Grid Feasibility Analysis Using Machine Learning," *2023 10th IEEE International Conference on Power Systems (ICPS)*, pp. 1–6, Dec. 2023, doi: 10.1109/icps60393.2023.10428886.

[11] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/access.2019.2914999.

[12] D. Kilroy, G. Healy, and S. Caton, "Using Machine Learning to Improve Lead Times in the Identification of Emerging Customer Needs," *IEEE Access*, vol. 10, pp. 37774–37795, 2022, doi:10.1109/access.2022.3165043.

[13] B. S. A. S. Rajita, P. Tarigopula, P. Ramineni, A. Sharma, and S. Panda, "Application of Evolutionary Algorithms in Social Networks: A Comparative Machine Learning Perspective," *New Generation Computing*, vol. 41, no. 2, pp. 401–444, Apr. 2023, doi:10.1007/s00354-023-00215-4.

[14] D. Arkhipova and M. Janssen, "AI recommendations' impact on individual and social practices of Generation Z on social media: a comparative analysis between Estonia, Italy, and the Netherlands," *Semiotica*, Mar. 2024, doi: 10.1515/sem-2023-0089.

[15] Y. Suh, "Machine learning based customer churn prediction in home appliance rental business," *Journal of Big Data*, vol. 10, no. 1, Apr. 2023, doi: 10.1186/s40537-023-00721-8.

[16] S. Zulaikha, H. Mohamed, M. Kurniawati, S. Rusgianto, And S. A. Rusmita, "Customer Predictive Analytics Using Artificial Intelligence," *The Singapore Economic Review*, pp. 1–12, Aug. 2020, doi: 10.1142/s0217590820480021.

[17] Y. Dai and T. Wang, "Prediction of customer engagement behaviour response to marketing posts based on machine learning," *Connection*

*Science*, vol. 33, no. 4, pp. 891–910, Apr. 2021, doi:10.1080/09540091.2021.1912710.

[18] S. Koli, R. Singh, R. Mishra, and P. Badhani, "Imperative role of customer segmentation technique for customer retention using machine learning techniques," *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 243–248, Jan. 2023, doi: 10.1109/aisc56616.2023.10085487.

[19] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," *Eighth International Conference on Digital Information Management (ICDIM 2013)*, pp. 131–136, Sep. 2013, doi:10.1109/icdim.2013.6693977.

[20] Ş. Dönmez, "Machine Learning-Based Merchant Churn Prediction in Banking," *2023 8th International Conference on Computer Science and Engineering (UBMK)*, pp. 503–508, Sep. 2023, doi: 10.1109/ubmk59864.2023.10286753.

[21] I. Jahan and T. Farah Sanam, "An Improved Machine Learning Based Customer Churn Prediction for Insight and Recommendation in E-commerce," *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, Dec. 2022, doi:10.1109/iccit57492.2022.10054771.

[22] A. H M, B. T, S. Tanisha, S. B, and C. C. Shanuja, "Customer Churn Prediction Using Synthetic Minority Oversampling Technique," *2023 4th International Conference on Communication, Computing and Industry 6.0 (C216)*, pp. 01–05, Dec. 2023, doi:10.1109/c2i659362.2023.10430989.

[23] N. Ali et al., "Fusion-Based Supply Chain Collaboration Using Machine Learning Techniques," *Intelligent Automation & Soft Computing*, vol. 31, no. 3, pp. 1671–1687, 2022, doi:10.32604/iasc.2022.019892.

[24] P. Thamaraiselvi, J. Masih, P. Giri, J. Sridevi, I. A. Karim Shaikh, and M. V. R. Prasad, "Analysis of Social Media Marketing Impact on Customer Behaviour using AI & Machine Learning," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pp. 1–6, Apr. 2024, doi:10.1109/iconstem60960.2024.10568805.

[25] R. Vadakattu, B. Panda, S. Narayan, and H. Godhia, "Enterprise subscription churn prediction," *2015 IEEE International Conference on Big Data (Big Data)*, pp. 1317–1321, Oct. 2015, doi:10.1109/bigdata.2015.7363888.

[26] M. Cowan, T. Moreau, T. Chen, J. Bornholt, and L. Ceze, "Automatic generation of high-performance quantized machine learning kernels," *Proceedings of the 18th ACM/IEEE International Symposium on Code Generation and Optimization*, pp. 305–316, Feb. 2020, doi:10.1145/3368826.3377912.

[27] D. S. K. Nayak, S. P. Routray, S. Sahooo, S. K. Sahoo, and T. Swarnkar, "A Comparative Study using Next Generation Sequencing Data and Machine Learning Approach for Crohn's Disease (CD) Identification," *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*, pp. 17–21, Aug. 2022, doi:10.1109/mlcss57186.2022.00012.

[28] C. Paramasivan, D. Paul Dhinakaran, S. S. P. Selvam, S. Mukherjee, A. P. Juliet, and S. R. Devi, "Comparing Supervised and Unsupervised Learning Technologies for Customer Segmentation in Marketing," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pp. 1–6, Apr. 2024, doi:10.1109/iconstem60960.2024.10568702.

[29] A. M. Carrington et al., "Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 329–341, Jan. 2023, doi:10.1109/tpami.2022.3145392.

[30] Y. Benhaddou and P. Leray, "Customer Relationship Management and Small Data — Application of Bayesian Network Elicitation Techniques for Building a Lead Scoring Model," *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pp. 251–255, Oct. 2017, doi: 10.1109/aiccsa.2017.51.

[31] R. Nygård and J. Mezei, "Automating Lead Scoring with Machine Learning: An Experimental Study," *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020, doi:10.24251/hicss.2020.177.

[32] D. R. Hermawan, M. Fahrio Ghanial Fatihah, L. Kurniawati, and A. Helen, "Comparative Study of J48 Decision Tree Classification Algorithm, Random Tree, and Random Forest on In-Vehicle CouponRecommendation Data," *2021 International Conference on Artificial Intelligence and Big Data Analytics*, pp. 1–6, Oct. 2021, doi:10.1109/icaibda53487.2021.9689701.

[33] G. N. Kumar and K. Hariharanath, "Designing a Lead Score Model for Digital Marketing Firms in Education Vertical in India," *Indian Journal of Science and Technology*, vol. 14, no. 16, pp. 1302–1309, Apr. 2021, doi: 10.17485/ijst/v14i16.290.

[34] A. Jadli, M. Hamim, M. Hain, and A. Hasbaoui, "Toward A Smart Lead Scoring System Using Machine Learning," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 2, pp. 433–443, Apr. 2022, doi: 10.21817/indjcse/2022/v13i2/221302098.

[35] R. L. Marchese Robinson, A. Palczewska, J. Palczewski, and N. Kidley, "Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets," *Journal of Chemical Information and Modeling*, vol. 57, no. 8, pp. 1773–1792, Aug. 2017, doi: 10.1021/acs.jcim.6b00753.

[36] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 994–998, Feb. 2021, doi: 10.1109/icicv50876.2021.9388482.

[37] J. B. G. Brito et al., "A framework to improve churn prediction performance in retail banking," *Financial Innovation*, vol. 10, no. 1, Jan. 2024, doi: 10.1186/s40854-023-00558-3.

[38] A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/access.2024.3402092.

[39] A. Raj and D. Vetrithangam, "Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review," *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 139–144, Apr. 2023, doi: 10.1109/cises58720.2023.10183530.

[40] P. Pulkundwar, K. Rudani, O. Rane, C. Shah, and S. Virnodkar, "A Comparison of Machine Learning Algorithms for Customer Churn Prediction," *2023 6th International Conference on Advances in Science and Technology (ICAST)*, Dec. 2023, doi:10.1109/icast59062.2023.10455051.

[41] T. Sunksuwan et al., "Utilizing Machine Learning to Enhance Profit Generation from Technical Trading Rules in Portfolio Construction within the Stock Exchange of Thailand," *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1189–1193, Dec. 2023, doi: 10.1109/icmla58977.2023.00178.

[42] M. A. Khan et al., "Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning," *IEEE Access*, vol. 8, pp. 116013–116023, 2020, doi:10.1109/access.2020.3003790.