Automated Customer Segmentation AI-Powered Lead Scoring for Edtech

Abstract

EdTech companies collect vast amounts of data, such as browsing behavior, email engagement, and other contact details, which can be leveraged through predictive analytics to estimate a lead's purchase probability. This study investigates the use of machine learning for prospect scoring using a dataset of approximately 9,000 educational lead records. The objective is to enhance lead conversion rates by predicting the likelihood of conversion using historical behavioral data and engagement metrics. The problem is approached as a binary classification task, where supervised learning algorithms such as logistic regression, decision tree, and ensemble methods like random forest are applied. Purchase timestamps are used to define activity windows for converted leads, ensuring fair data representation. The models are evaluated using accuracy, precision, recall, and ROC-AUC. Among them, logistic regression achieved the highest accuracy and interpretability, while random forest provided valuable insights through feature importance analysis. The results demonstrate that machine learning-driven lead scoring can effectively prioritize high-potential leads, optimize marketing and sales strategies, and offer actionable business insights through visual analytics for decision makers.

Keywords

Lead Scoring, Machine Learning, Customer Identification, EdTech, Predictive Analytics

1. Introduction

In today's competitive market, effective customer acquisition is vital, and lead scoring plays a key role by helping businesses prioritize potential customers based on their engagement behaviors, such as website visits and email interactions [1]. Traditionally, this has been a manual process, assigning importance to each customer activity to rank leads. However, manual methods are often limited in scale and accuracy.

This article explores how machine learning can automate and enhance lead scoring in the B2C sector. Using real-world data, various models are developed and evaluated to overcome data preparation challenges and improve prediction accuracy. With a historical conversion rate of 30–40%, the goal is to help businesses target high-potential leads and increase conversions to 80%. Visual analytics are also applied to reveal actionable insights, supporting smarter decision-making and improving overall marketing efficiency through data-driven strategies.

2. Background

In the digital age, businesses generate vast amounts of data [2], leading to a shift toward data-driven decision-making [3], especially in marketing and CRM. Relationship marketing, which focuses on creating value through ongoing collaboration with customers [4], relies heavily on digital data to stay competitive [5]. By analyzing interactions from digital channels, companies can better identify and convert leads.

Integrating business analytics and machine learning into CRM enhances customer tracking and lead scoring efficiency [6]. Traditional intuition-based methods are now being replaced by automated systems [7] that detect user behavior trends to predict conversions [8]. Despite available tools, practical research on applying automation across the B2C sales funnel remains limited, underscoring the need for further study [8].

2.1 Manual Lead Scoring

Before diving into automated approaches, it's important to understand the conventional method widely used in industry manual lead scoring. As noted by Marion [9], this approach presents several critical limitations. One of the main concerns is that manual lead scoring lacks a foundation in statistical evidence, often relying on subjective assessments rather than data-driven insights. Typically, it uses a variety of demographic, behavioral, or firmographic Since this method usually depends on a scoring matrix, businesses must regularly revise and update it to stay aligned with changing market conditions a process that can be both labor-intensive and inefficient. Marion [9] highlights these issues through an experiment involving 800 leads evaluated using manual scoring. The results showed no significant difference between leads tagged as "ready for sales" and a random group of unscored leads. The study emphasizes that without a solid understanding of statistics, accurately assigning weights to lead behaviors is nearly impossible. Additionally, the manual process demands continuous adjustments, which consumes time that could be better allocated elsewhere. Bohlin [10] also

Activity	Points
Form/Landing Page Submission	+5
Submitted "Contact Me" Form	+25
Received an Email	0
Email Open	+1
Email Clickthrough	+3
Registered for Webinar (Optional)	+3
Attended Webinar	+10
Downloaded a Document	+5
Visited a Landing Page	+2
Unsubscribed from Newsletter	-2
Watched a Demo	+8
Contact is a CXO	+5
Visited Trade Show Booth	+3
Visited Pricing Page	+10

Table 1: Example manual lead scoring matrix [9]

critiques this approach, arguing that even when assumptions are used to develop rules and weights, manual lead scoring remains suboptimal.

2.2 Components of Lead Scoring

Lead scoring is a key part of CRM that assigns numerical values to prospects, helping prioritize leads based on conversion likelihood [11]. Higher scores guide leads to sales, while lower ones may enter nurturing workflows [12]. The model's success depends on selecting relevant variables, including implicit behavioral data and explicit user information [12]. Leading firms often use behavioral inputs and complex models for better performance [12].

As a predictive analytics method, lead scoring uses statistical tools to forecast outcomes [13]. Predictive marketing builds on this by personalizing customer journeys through data insights and lower computing costs [13]. Machine learning—particularly supervised learning—is commonly applied to predict lead conversions from historical data [14]. Bayesian networks [11] and modern ML approaches [15] enhance sales efficiency, even in limited data scenarios.

2.3 Machine Learning Applications in Customer Relationship Management

Machine learning has many effective applications in customer relationship management (CRM), enhancing decisions across the customer journey [16]. Key techniques include classification, clustering, regression, forecasting, and visualization, using algorithms like decision trees, KNN, genetic algorithms, neural networks, and logistic regression. These approaches support tasks such as lead scoring, segmentation, and behavior prediction.

Real-world use cases highlight their impact. A study in [17] built a loyalty prediction framework using random forest, logistic regression, and neural networks, with random forest showing strong accuracy and AUC. Another case in [18] combined a genetic algorithm and neural network for direct marketing, using feature selection to improve interpretability and profit-focused decision-making.

3. Literature Review

Traditional lead scoring methods often rely on manual rules that can be biased and subjective. To overcome this, predictive lead scoring leverages historical data and machine learning to identify traits linked to successful conversions. As Syam and Sharma [1] suggest, AI is transforming marketing decisions, with models like logistic regression and decision trees widely used to assess leads based on demographics and behavior. Chorianopoulos [6] and Duncan and Elkan [7] stress the value of analytics and probabilistic modeling in improving CRM and lead prioritization. Behavioral scoring, focusing on user interactions like site

visits, is especially effective in fast-changing sectors like EdTech. Järvinen and Taiminen [8] also highlight real-time tools that enhance automated B2B marketing.

In EdTech, AI can analyze user behavior to identify high-potential leads, streamlining marketing efforts. Research shows supervised learning models perform well when trained on features like traffic source, activity frequency, and geography. Marion [9] and Bohlin [10] argue that manual scoring is outdated, advocating for automation. Models are typically evaluated using ROC-AUC to measure classification accuracy [14]. Frameworks like Demandbase allow for real-time scoring by combining historical and current data, helping businesses focus on quality leads, improve conversions, and reduce wasted effort.

4. Methodology

The methodology adopted for this lead scoring project is structured around a clear, step-by-step machine learning workflow. It begins with gathering lead-related data, such as user activity and source of origin, followed by a thorough preparation phase. This includes managing missing entries, transforming categorical attributes through encoding, and normalizing numerical features to ensure uniformity. These preprocessing steps help prepare the dataset for reliable and effective model training. By refining the input data, the system becomes more capable of identifying meaningful patterns related to lead conversion behavior.

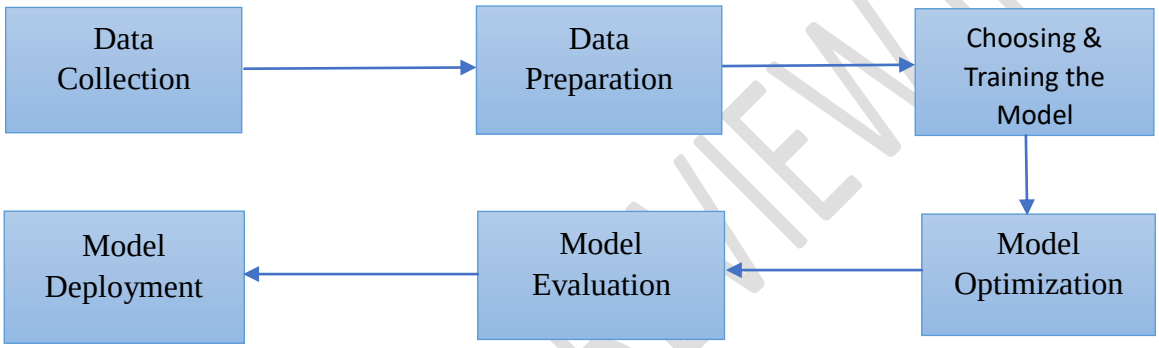

Figure 1: Proposed Methodology

Following data preparation, the project focuses on building and evaluating predictive models. Various algorithms like Logistic Regression, Decision Trees, and Random Forests are trained and compared using metrics such as AUC-ROC and lift curves. After evaluating the models, Logistic Regression is selected for its performance and ease of interpretation. The model helps prioritize leads by assigning scores, allowing the business to focus on those most likely to convert. This data-driven prioritization is expected to significantly improve the efficiency of the sales team and boost overall lead conversion rates.

4.1 Dataset

The dataset analyzed in this study contains 9,241 records, each representing a potential lead for an EdTech platform. It features 36 attributes that provide a well-rounded view of user behavior, demographics, and interactions. Key variables include lead source (e.g., Google, Facebook), lead origin (e.g., API, Landing Page Submission), and last activity (e.g., Email Opened, SMS Sent). Initially, the data is in an unstructured format. The dataset also includes numeric data such as total time spent on the website and page views per visit, offering insights into engagement levels. The target variable, lead conversion status, enables binary classification modeling. These varied data points make the dataset ideal for training machine learning models aimed at predicting conversion likelihood.

Collected from real operational data, the dataset reflects actual customer behavior, making it valuable for real-world applications like lead scoring, sales prioritization, and personalized marketing. Features such as academic specialization, course preferences, and city information add context that can enhance predictive accuracy. The presence of missing values and outliers introduces opportunities for data preprocessing, including imputation and outlier handling. With around 90% of the data usable for training and 10% for testing, the dataset is well-structured to support robust model development and performance evaluation, ultimately helping EdTech businesses improve conversion strategies and customer engagement.

4.2 Data Preprocessing

To ensure consistency and quality, the following preprocessing steps were applied:

- Handled Missing Data: Removed columns with too many missing values and filled others with suitable replacements. For categorical data it filled with “Unknown” or “Not Specified” and for continuous data it filled with mean or median.
- Dropped Unnecessary Columns: Eliminated duplicate or irrelevant columns that didn’t contribute to the model.
- Outlier Detection: Use IQR (Inter Quantile Range) method for detecting the outliers and cap extreme values or remove them to get cleaner numeric data with minimized impact of outliers.
- Feature Engineering: To modify or edit the features in the dataset. To combine one or more features to make it single one to reduce the complexity of the model.

4.3 Modeling Approaches

Multiple machine learning algorithms were evaluated for lead score conversion. As our usecase is a classification model so that we used classification algorithms.

- Logistic Regression: A simple classification method that estimates the probability of a lead converting by fitting data to a logistic curve. As our project is binary classification so logistic regression is used, it performs well on binary classification.
- Decision Tree: Builds a tree-like model of decisions by splitting data based on feature values, making it easy to interpret outcomes. It suited for regression & classification problems, but it overfits the model.
- Random Forest: An ensemble method follows parallel approach that combines multiple decision trees to improve accuracy and reduce overfitting by averaging their results. Used for regression as well as classification problem.

4.4 Hyperparameter Tuning

To enhance model performance, Grid Search was employed for hyperparameter tuning. Parameters such as the number of estimators, maximum depth (for tree-based models), and learning rate (for boosting methods) were optimized using cross-validation to prevent overfitting and ensure generalization.

4.5 Evaluation Metrics

Models were evaluated using the following metrics:

- Accuracy: Represents the percentage of total correct predictions but can be unreliable when the dataset has imbalanced classes like more non-converted leads.
- Precision: Measures how many leads predicted as "converted" were actually correct, helping reduce false positives and improve targeting accuracy in marketing campaigns.
- Recall: Indicates how many actual converted leads were correctly identified by the model, ensuring fewer missed opportunities in lead follow-ups.
- ROC-AUC: Shows the model’s ability to differentiate between converted and non-converted leads across various thresholds, with higher scores reflecting better performance. To compare the performance of different algorithms ROC-AUC curve will be used. It’s a plot between FPR and TPR.

5. Results & Discussion

Based on the final dataset described in the previous section, three different machine learning algorithms were selected to be tested motivated by the findings in our literature review on the most widely used algorithms in customer relationship management:

- Logistic Regression (LR) [14]: A well-established generalized linear model frequently applied to binary classification tasks, estimating the likelihood of class membership based on input features. As our problem is binary classification so logistic regression is used, it performs well on binary classification.
- Decision Trees (DT) [20]: These models build hierarchical decision rules based on dataset features. They are widely appreciated for their interpretability and ability to explain predictions through logical if-then conditions. It suited for regression & classification problems, but it overfits the model.
- Random Forests (RF) [21]: A tree-based ensemble method that mitigates the overfitting tendency of individual decision trees by generating multiple de-correlated trees and averaging their outputs to make predictions. To compare the performance of different algorithms ROC-AUC curve will be used. It's a plot between FPR (False Positive Rate) and TPR (True Positive Rate).

To evaluate model effectiveness, several metrics derived from the confusion matrix were used [14]. These include counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), where a "positive" denotes a successfully converted lead. In addition to basic accuracy, metrics like precision, recall, sensitivity, and specificity were calculated to better understand the model's behavior under different types of errors.

A key evaluation metric applied was the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), which measures the model's ability to differentiate between classes by plotting true positive rates against false positive rates at various threshold levels.

Given the class imbalance in the dataset, SMOTE (Synthetic Minority Oversampling Technique) was used to create a more balanced training set. Additionally, 10-fold cross-validation was implemented to ensure robust and unbiased model performance estimates.

An initial exploration of data aggregation strategies was conducted, and after evaluating potential biases, a single aggregation method was selected for final model refinement and detailed analysis.

Data exploration reveals diverse lead characteristics. The distribution of lead sources (Figure 2) shows that prospects come from multiple channels (e.g. Google, Organic Search, etc.) without a single dominant source, underscoring the need to treat *"*Unknown*"* as a category during imputation. The distribution of total website visits (Figure 3) is right-skewed: most leads made only a few visits, but a minority visited frequently. Similarly, the page views per visit distribution (Figure 4) indicates most leads browsed a small number of pages per visit, with few high-engagement outliers. These behavioral features reflect varying levels of engagement.

Lead Source Analysis (Figure 2): Leads came from various platforms (Google, ads, chat); missing sources were treated as 'unknown' to retain all data.

Website Visit Behavior (Figure 3): Most leads visited the site only 1–3 times, showing low engagement; data was right-skewed and scaled for modeling.

Page Views per Visit (Figure 4): Similar skew observed—most users viewed few pages, while some viewed many, indicating high intent; this was used as a behavioral signal.

Last Activity Engagement (Figure 5): Actions like email opened or SMS sent reflect engagement level; rare but important activities were retained or grouped for better prediction.

Country Information and Imputation (Figure 6): 26.6% of entries lacked country info; missing values were filled using city data where possible, mostly as 'India' or 'unknown'.

City-wise Distribution (Figure 7): A few cities (e.g., Mumbai, Thane) generated most leads; similar cities were combined, and city was kept as a key categorical feature**.**

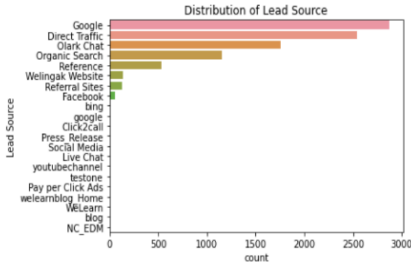

Figure 2

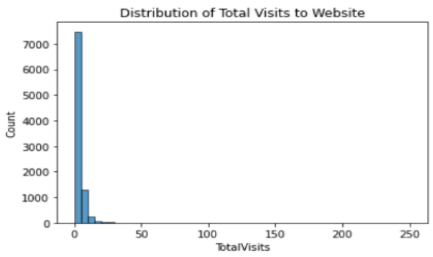

Figure 3

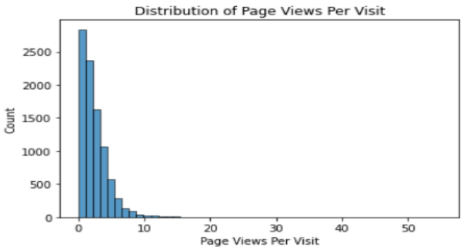

Figure 4

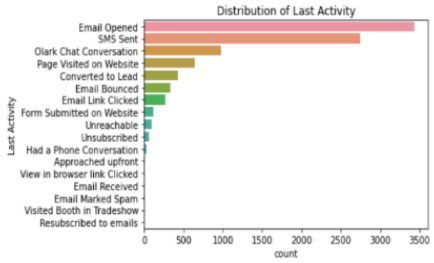

Figure 5

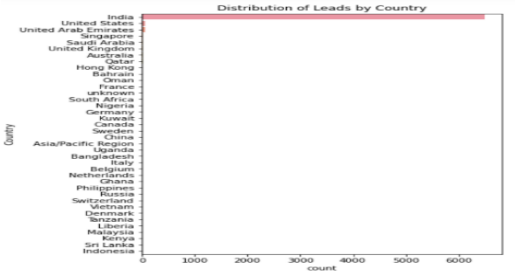

Figure 6

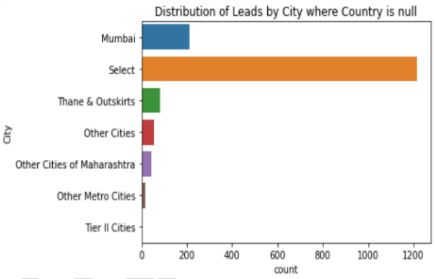

Figure 7

In summary, the exploratory analysis reveals that leads differ significantly in their origin, level of engagement, and geographical location. We move forward with the cleaned and encoded dataset into the process of training the model.

5.1 Model Performance: Table 2 provides an overview of the key performance metrics of our models, including accuracy, precision, recall, and F1 score, on both the training and test datasets. Logistic regression attained an AUC of 0.9217 on the training set and 0.9094 on the test set. The decision tree achieved an accuracy of 0.9105 (train) and 0.8956 (test). The random forest achieved an accuracy of 0.9106 in both the training and test sets. All models have high AUCs (>0.89), but logistic regression is slightly better than the other methods on the validation set. This implies that the generalization is most effective, possibly because regularization helps prevent overfitting, while the tree and forest models show slightly lower test AUC scores. Model accuracy of the training set accuracy of testing set.

Model	AUC of Training Set	AUC of Testing Set
Logistic Regression	0.9217	0.9094
Decision Tree	0.9105	0.8956
Random Forest	0.9106	0.9000

Table 2: AUC Scores of Different Models

Figure 8 (below) displays the receiver operating characteristic (ROC) curves for each model on the test set. We observe that the logistic model's receiver operating characteristic (ROC) curve is slightly above the random forest's, resulting in the highest area under the curve (AUC). The variation is slight, suggesting that all models are fairly effective in distinguishing between the two groups. Considering the logistic model's simplicity and interpretability, we choose it as the primary scoring model for our subsequent analysis.

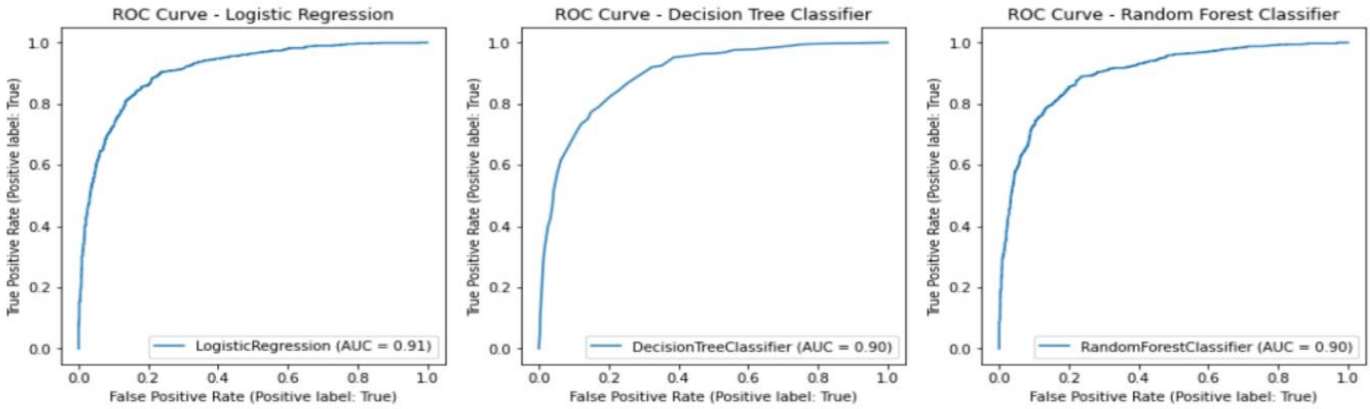

Figure 8: ROC Curves of Different Models

The chosen logistics model assigns a probability score for conversion to each lead. To translate this into a sales strategy, we calculate a lift curve. Based on the data, the average conversion rate was around 38%. To achieve an 80% conversion rate (more than double the baseline), it would be necessary to prioritize the leads with the highest scores. Our lift analysis indicates that reaching out to approximately the top 30–35% of leads based on their score results in an estimated conversion rate of around 80%. In essence, if sales reps focus on the top third of scored leads, they can achieve the CEO's goal of achieving an 80% conversion rate. By automating lead prioritization, the team can prevent wasting effort on low-probability leads and instead focus on allocating resources to those that have the highest potential for enrollment.

6. Conclusion & Future Scope

This case study highlights the use of AI-driven lead scoring in EdTech using logistic regression, achieving strong predictive performance (AUC ~0.91), this model is used for classification model & it performs well on binary classification. Prioritizing the top 35% of leads led to an estimated 80% enrollment rate, showcasing the model's effectiveness. Integrating this system into a CRM can help sales teams focus on high-potential leads in real time, improving conversion rates.

In the future, the integration of this system into a full-scale CRM (Customer Relationship Management) platform can be further enhanced with real-time analytics, AI-driven lead nurturing, and personalized communication workflows. Advanced machine learning models can continuously learn from new data to improve lead scoring accuracy. Additionally, integrating with marketing tools can allow sales teams to engage high-potential leads across multiple platforms , increasing overall conversion rates. Scalability to support larger datasets and cross-functional team collaboration will also open up opportunities for broader adoption across various industries beyond EdTech.

The model can be enhanced with real-time data feeds, integrated with marketing automation tools, and adapted using deep learning techniques for more complex behavior patterns. Expanding to multi-channel engagement and personalizing outreach based on lead behavior could further boost enrollment efficiency.

References

[1] Syam, N. and Sharma, A., (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Industrial Marketing Management, 69, pp.135-146.

[2] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

[3] Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision-making. American Economic Review, 106(5), 133-39.

[4] Sheth J. N., Parvatiyar A., Sinha M., (2015). The conceptual foundations of relationship marketing: Review and synthesis. Journal of economic sociology, 16(2), 119-149.

[5] Leeflang, P. S., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. European management journal, 32(1), 1-12.

[6] Chorianopoulos, A. (2016). Effective CRM using predictive analytics. John Wiley & Sons.

[7] Duncan, B. A., & Elkan, C. P. (2015, August). Probabilistic modeling of a sales funnel to prioritize leads. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1751-1758). ACM.

[8] Järvinen, J., & Taiminen, H. (2016). Harnessing marketing automation for B2B content marketing. Industrial Marketing Management, 54, 164-175.

[9] Marion, G. (2016). Lead Scoring is Broken. Here's What to Do Instead. URL: https://medium.com/marketing-on autopilot/lead-scoring-is-broken-here-s-what-to-do-instead 194a0696b8a3 (Retrieved 24.09.2018)

[10] Bohlin, E. (2017). Sorting Through the Scoring Mess. URL: https://www.siriusdecisions.com/blog/sorting-through the-scoring-mess (Retrieved 24.09.2018)

[11] Benhaddou, Y., & Leray, P. (2017, October). Customer Relationship Management and Small Data—Application of Bayesian Network Elicitation Techniques for Building a Lead Scoring Model. In Computer Systems and Applications (AICCSA), 2017 IEEE/ACS 14th International Conference on (pp. 251-255). IEEE.

[12] Michiels, I. (2008). Lead Prioritization and Scoring: The Path to Higher Conversion. Aberdeen Group.

[13] Artun, O., & Levin, D. (2015). Predictive marketing: Easy ways every marketer can use customer analytics and big data. John Wiley & Sons.

[14] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

[15] Adam, M.B. (2018). Improving complex sale cycles and performance by using machine learning and predictive analytics to understand the customer journey (Doctoral dissertation, Massachusetts Institute of Technology).

[16] Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

[17] Wouter, B., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviorally-loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, 164(1), 252-268.

[18] Kim, Y., & Street, W. N. (2004). An intelligent system for customer targeting: a data mining approach. Decision Support Systems, 37(2), 215-228.

[19] Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. MIS quarterly, 553 572.

[20] Karim, M., & Rahman, R. M. (2013). Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. Journal of Software Engineering and Applications, 6(04), 196.

[21] Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472-484.