LEAD SCORING ANALYSIS FOR ONLINE COURSES PLATFORM USING SCI-KIT LEARN

Akarsh Kumar Singh
Department of Computer Science and Engineering
SRM Institute of Science and Technology Ghaziabad, India
thakurak9415@gmail.com
ORCID: 0000-0003-0830-0510

Abhinav Tripathi
Department of Computer Science and Engineering
SRM Institute of Science and Technology Ghaziabad, India
at2357@srmist.edu.in

Prakhar kumar
Department of Computer Science and Engineering
SRM Institute of Science and Technology Ghaziabad, India
prakhukmr823@gmail.com

Shubh Mehrotra
Department of Computer Science and Engineering
SRM Institute of Science and Technology Ghaziabad, India
cartoonchennai@gmail.com

***Abstract*—The project aims to develop a lead scoring system using linear regression to predict the probability of a potential customer becoming a lead for a business. The system involves data collection, preprocessing, feature extraction, model training, model evaluation, deployment, and integration with other systems. The project seeks to improve the accuracy and efficiency of lead scoring by leveraging machine learning techniques. The system aims to be scalable and easily integrated into existing business workflows. The performance of the system will be evaluated through various testing techniques to ensure its accuracy and reliability in predicting lead scores**

Keywords*— *sci-kit, lead scoring, online course, linear regression.

I. INTRODUCTION

This lead score prediction provides businesses with a tool to predict the likelihood of a potential customer becoming a paying customer by analyzing various factors such as demographics, online behavior, and previous interactions with the business, the lead score prediction model can identify the potential customers who are most likely to convert to paying customers. The project uses linear regression as the primary machine learning algorithm for lead score prediction. Linear regression is a simple and interpretable algorithm that is well-suited for regression tasks, making it a suitable choice for lead score prediction. This proposal consists of several steps, such as data collection, feature extraction, pre-processing, model training, model evaluation, and deployment and integration.

The project aims to address the limitations of the existing system for lead score prediction using SVM. The SVM algorithm can be complex, sensitive to outliers, difficult to interpret, and struggle with non-linear relationships and large or imbalanced datasets. By using linear regression as the primary machine learning algorithm, the proposed system can provide a simpler and more interpretable alternative to the existing system.

The lead score prediction project has several potential benefits for businesses, including increased efficiency and effectiveness in identifying potential customers who are more likely to convert to paying customers. By identifying these potential customers, businesses can target them with personalized marketing campaigns and increase their chances of conversion. Additionally, by integrating the lead score prediction model with other systems such as CRM systems or marketing automation platforms, businesses can enhance its impact and maximize its potential benefits

II. OBJECTIVE

The lead score prediction project aims to provide businesses with an efficient and effective tool to identify potential customers who are most likely to convert to paying customers. The project aims to leverage ML Algorithms such as linear regression which is used to analyze enormous amounts of data and identify patterns and characteristics of potential customers that are associated with higher conversion rates. The ultimate goal of the project is to help businesses increase their efficiency and effectiveness in identifying potential customers and targeting them with personalized marketing campaigns. By accurately predicting lead scores, businesses can prioritize their efforts and resources towards potential customers who are most likely to convert, improving their chances of conversion and maximizing their return on investment.

Additionally, Lead score prediction is a vital aspect of any successful marketing campaign. However, the existing system for lead score prediction using SVM has certain limitations that need to be addressed. One of the biggest challenges is the complexity of the SVM algorithm, which makes it difficult to interpret the results.

Overall, the project's goal is to leverage the power of linear regression to create a more accessible and user-friendly lead score prediction system that delivers actionable insights and drives business growth.

III. LITERATURE REVIEW

The retail industry largely adopted various inventory management programs, and some retailers even use prediction software to forecast future projected sales. However, much of India's daily buying is done in small stores. The owners of such small businesses may not have the cash to invest in proprietary applications to set up an inventory management system. Needless to add, the same is true for any sales forecasting program. As a result, many retailers wind up stockpiling a large number of useless and non-profitable products, resulting in financial losses. A mobile application that delivers all of the functionality of a

point-of-sale system as well as future sales analytics is a very cost-effective and accessible solution to this problem. It will allow store owners to handle their current product purchases and billing. The predicted sales analysis will assist them in modifying their product and supply investments, ensuring maximum revenues. If a store has a relevant product that meets the wants of its customers, its customer base will grow. The Economic Times reported in May 2019 that the number of smartphone users in India is predicted to increase by 84% to 859 million by 2022, up from 468 million in 2017. It is realistic to expect that many store owners will have cell phones in the coming years. As a result, providing local shops with a mobile application will be beneficial because it will expose them to all of the above-mentioned benefits

Smart data mining approaches with accurate prediction models and great dependability are required for sales prediction analysis. Most market segments, in essence, rely on the know-how base and demand trend forecast for analysis of Business To Business (B2B) sales data. Sales data are offered on how a telecommunications company should manage its sales team, goods, and budgetary processes. Telecommunication Companies can survive the industry fight and improve it's market growth thanks to precise estimates. In this study, intelligible predictive models are studied and analyzed using machine learning approaches to improve future sales projections. Traditional forecasting systems struggle to deal with massive data and sales forecasting accuracy. This study provides a brief examination of the dependability of B2B sales using machine learning approaches. The final section of this study explains a variety of sales forecast methodologies and treatments. Based on the performance evaluation, the best-suited predictive model for B2B sales trend forecasting is provided. The results of projection, estimation, and analysis are described in terms of the dependability and consistency of effective prediction and forecasting methodologies. The findings of this analysis are intended to produce dependable, accurate, and effective forecasting data, which will be a significant resource for sales forecasting. According to research, the Gradient Boost Algorithm offers a high level of accuracy in predicting future B2B sales prediction, with MSE = 24,743,000,000.00 and MAPE = 0.18.

The main goal of the sales prediction regression model is to anticipate automobile sales using sentiment analysis from various websites. The online presence of a car, as well as its brand, are important factors in automobile sales. Many more criteria, however, are essential and will be covered in this work. In today's industry, sales forecasting benefits not only the manufacturer but also several other companies that make vehicle parts or accessories. It is also beneficial to retailers, showroom owners, and service mechanics. We used linear regression for sentiment analysis and polynomial regression for sales prediction in this application Sales forecasting is a method of predicting future sales. Accurate income estimates enable the retail market to make informed business decisions and predict short- and long-term performance. Sales forecasting enables a shop to estimate its projected future revenues for income earned over a specific time period. As a result, time is an important factor in sales forecasting. A time sequence is a sequence of information components that are indexed in time order. The time sequence analysis evaluates observations as a set of statistics at specific intervals across time, with the goal of determining trends, cycles, and seasonal variances that can be used to forecast future events. With the advancement of data technology, large retailers have begun to utilize statistical approaches such as index numbers, time collection, and a couple of regression evaluations for the purpose of income forecasting. For prediction, the XG Boost method was used in this research. The program makes use of sales data from Rossman, which operates over 8000 drug stores across seven European countries. Machine Learning is transforming every walk of life and has become a major contributor in real-world scenarios. The revolutionary applications of Machine Learning can be seen in every field including education, healthcare, engineering, sales, entertainment, transport, and several more; the list is never-ending. The traditional approach of sales and marketing goals no longer helps the companies to cope with the pace of the competitive market, as they are carried out with no insights into customers ' purchasing patterns. Major transformations can be seen in the domain of sales and marketing as a result of Machine Learning advancements. Owing to such advancements, various critical aspects such as consumers 'purchase patterns, target audience, and predicting sales for the recent years to come can be easily determined, thus helping the sales team in formulating plans for a boost in their business. The aim of this paper is to propose a dimension for predicting the future sales of Big Mart Companies keeping in view the sales of previous years. A comprehensive study of sales prediction is done using Machine Learning models such as Linear Regression, K-Neighbors Regressor, XGBoost Regressor, and Random Forest Regressor. The prediction includes data parameters such as item weight, item fat content, item visibility, item type, item MRP, outlet establishment year, outlet size, and outlet location type

A. Abbreviations and Acronyms

SVM- Support Vector Machine

CRM- Customer Relationship Management

ML- Machine Learning

MAPE- Mean Absolute Percentage Error

LSSVR- Lead Squares Support Vector Regression

PCA- Principal Component Analysis

IV. PROPOSED SYSTEM

The proposed system using linear regression for lead score prediction is a simpler and more interpretable alternative to the existing system that uses SVM. As a supervised learning approach that can be used for regression tasks, linear regression is a good option for predicting lead scores. The suggested system entails a number of phases, such as data collection and pre-processing, feature extraction, model training, model evaluation, deployment, and integration.

Data collection and pre-processing involve collecting data from various sources and pre-processing it to ensure its accuracy, completeness, and relevance for the lead score prediction model. Feature extraction involves identifying the most important features that can be used to predict the lead score. Techniques such as correlation analysis, PCA, or RFE can be used for this step. Model training involves using preprocessed and feature-extracted data to train the linear regression model. The model is

trained by finding the best fit line through the data points that reduce the squared error between the target variable's actual values and its anticipated values. Evaluation of the model entails assessing how well the trained linear regression model using techniques such as MSE or R-squared.

The proposed system using linear regression has several advantages, including simplicity, interpretability, scalability, and flexibility. Linear regression is a straightforward and simple method, which makes it simpler for developers to comprehend and analyze the model. It is also highly interpretable, making it easier to understand how the model is making its predictions and take appropriate actions based on the predictions. Additionally, linear regression can scale well to large datasets and handle a wide range of data types. However, the proposed system also has some limitations. Linear regression makes several assumptions about the data, including linearity, independence, homoscedasticity, and normality. Violations of these assumptions can lead to poor model performance. Additionally, linear regression is particularly prone to overfitting when the number of features is high in comparison to the number of observations. The simplicity and interpretability of linear regression can make it easier for developers to understand and interpret the model, while its scalability and flexibility can make it suitable for lead score prediction tasks that involve large amounts of data and different data types. However, developers should also consider the assumptions and limitations of linear regression when designing and implementing the system.

V. METHODOLOGIES

A. Data Collection and preprocessing

Data collection and pre-processing is a critical step in lead score prediction using linear regression. The accuracy of the lead score prediction model can be significantly impacted by the quality and accuracy of the data that was collected. Data pre-processing ensures that the collected data is accurate, complete, and relevant for the lead score prediction model.

Data collection may involve collecting data from various sources, such as CRM systems, marketing automation platforms, and social media platforms. The collected data may include information such as customer demographics, customer behavior, and customer interactions with the business.

To make sure the data is precise and pertinent for the lead score prediction model, it must be preprocessed when it is obtained. Several steps may be involved in data preparation, including

1) Data cleaning: The data cleaning stage entails getting rid of any irrelevant or duplicate data from the collected data. This step may also involve filling in missing data values using techniques such as imputation.

2) Data transformation: The data transformation step involves converting the data gathered into a format that is suitable for the lead score prediction model. This step may involve converting categorical data into numerical data using techniques such as one-hot encoding.

3) Data Normalization: The data normalization step involves scaling the data to ensure that it is on a similar scale. This step may involve techniques such as min-max normalization or standardization.

4) Feature Selection: The feature selection step involves selecting the most important features from the preprocessed data for the lead score prediction model. This step may involve using techniques such as correlation analysis or principal component analysis (PCA).

B. Future Extraction

Feature extraction is a crucial step in lead score prediction using linear regression. Finding the most crucial features from the pre-processed data that can be used to forecast the lead score is known as feature extraction. The aim of feature extraction is to reduce the dimensionality of the data while maintaining as much pertinent information as you can.

C. Figure

a) The given ROC curve is used in the lead scoring analysis for online course platforms using sci-kit-learn. It plots the True Positive Rate (sensitivity) on the x-axis and the False Positive Rate (1 - specificity) on the y-axis. It helps evaluate the model's ability to distinguish between converted and non-converted leads. The curve shows the trade-off between sensitivity and specificity at different classification thresholds, allowing for the selection of an optimal threshold. The area under the curve (AUC-ROC) summarizes the model's performance, with a higher value indicating better discrimination. This analysis aids in comparing models, optimizing thresholds, and making informed decisions about lead prioritization.

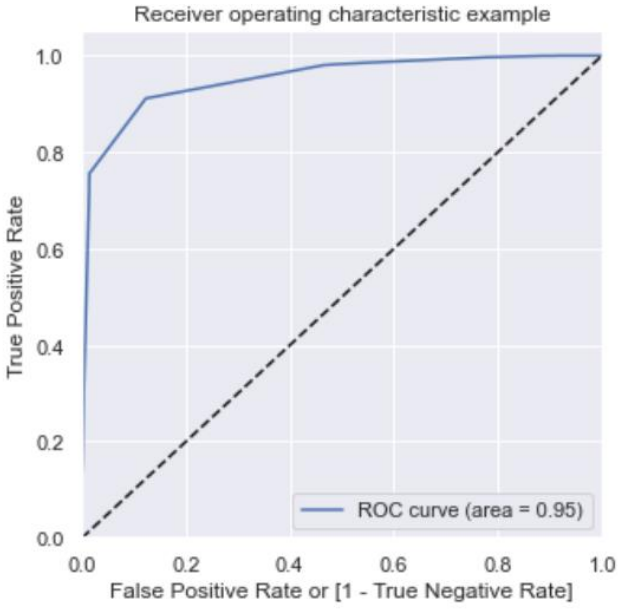

Fig. 1 ROC Curve

CONCLUSION

In conclusion, the proposed system using linear regression for lead score prediction is a simpler and more interpretable alternative to the existing system that uses SVM. The suggested system goes through a number of processes, including data collection and preprocessing, feature extraction, model training, model evaluation, deployment, and integration. By using appropriate techniques for each of these steps, developers can ensure that the linear regression model accurately predicts lead scores and is integrated with other systems to enhance its impact. Linear regression has several advantages, including simplicity, interpretability, scalability, and flexibility. Because of its straightforward and understandable algorithm, the model is simpler for organizations to comprehend and interpret. It also supports a variety of data types and scales well to huge datasets. By providing accurate and interpretable predictions of lead scores, the proposed system using linear regression has the potential to help businesses identify potential customers that are more likely to become paying customers.

However, developers should also consider the limitations and assumptions of linear regression when designing and implementing the system. Violations of these assumptions can lead to poor model performance, and linear regression can be prone to overfitting. By considering these limitations and using appropriate techniques to address them, developers can ensure that the proposed system using linear regression provides accurate and interpretable predictions of lead scores for businesses

REFERENCES

[1] H. Shi, "Evaluation of CET-4 and CET-6 Courses on Online LearningPlatform," 2021 IEEE 3rd International Conference on Computer Science and Educational Informatization (CSEI), Xinxiang, China, 2021, pp. 61-64, doi: 10.1109/ CSEI51395.2021.9477641.

[2] H. Cao, "Research on the Construction of Network Course Platform Based on Big Data," 2021 International Conference on Computer Network, Electronic, and Automation (ICCNEA), Xi'an, China, 2021, pp. 207-210, doi: 10.1109/ ICCNEA53019.2021.00053

[3] Erni-cassola G, Zadjelovic V, Gibson MI, Christie-oleza JA. Distribution of plastic polymer types in the marine environment; A meta-analysis. *Journal of Hazardous Materials* 2019;**369**:691-8

[4] J. -C. Liang, J. -Y. Wang and S. -H. Chou, "A designed platform for programming courses," 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), Kitakyushu, Japan, 2020, pp. 290-293, doi: 10.1109/ IIAIAAI50415.2020.00064

[5] J. Xie, Y. He, and Q. Qin, "Research on the Construction of the Quality Evaluation Indicator System of "Online Golden Course" Based on Experimental Analysis," 2021 2nd International Conference on Information Science and Education (ICISE-IE), Chongqing, China, 2021, pp. 855-859, doi: 10.1109/ ICISEIE53922.2021.00195.

[6] Y. -J. Chang and W. -L. Hsu, "Online Teaching Practice in Chinese Culture Course," 2020 IEEE Eurasia Conference on IOT, Communication, and Engineering.

[7] X. Liu and G. Guo, "Online Immersive Platform Image Processing Algorithm for Art Courses based on Virtual Equipment Digital Media Technology," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 1181-1184, doi: 10.1109/ICICT54344.2022.9850680

[8] H. Yao, "Application of Audio Processing Based on Image Semantic Segmentation in Applied Mathematics Online Course," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 1527-1530, doi: 10.1109/ICSSIT53264.2022.9716388.

[9] S. Wang, Y. Liu, F. Song, X. Xie, and D. Yu, "Research on Evaluation System of User Experience With Online Live Course Platform," in IEEE Access, vol.9, pp.23863-23875, 2021, doi: 10.1109/ACCESS.2021.3054047.

[10] G. X. Chen, X. Q. Qu, L. P. Huang, L. Huang, C. Zhou, and M. Y. Qiao, "Modeling-Eliciting Activities in an Online Engineering Course for improving Conceptual Learning, Professional Skill, Interaction," in IEEE Access, vol. 10, pp.87767-87777, 2022, doi: 10.1109/ACCESS.2022.3199690.