Lead Score Analysis

Naren Shriram, Virang Singh Tomar, Prof. Sworna Kokila

Abstract: Marketing and sales have been major pillars in any business processing services either product or service-based organizations, they are the ones which enable the companies to survive and are able to increase their capability to make money and deliver the right products to the right clients. But it would always be hard for the sales team to process this without hurdles as the leads they generate and get are almost out of segment or the ones who already are using an alternative for the product, this also increases the company spend and reduces the rate of closing the deal. To overcome this, we are implementing a solution on predictive lead scoring by using machine learning to provide the sales and marketing people with advance knowledge on customers with ways to target and close most deals which save a lot of time and increases revenue proportionately.

I. INTRODUCTION

Some of the most important business decisions in the current, cutthroat business environment have to do with customer acquisition. Companies use a variety of techniques to try to convert leads into customers during the acquisition phase of the customer life cycle. Several lead scoring procedures have been created and put into use in order to make this process as time- and money-efficient as it is possible from the perspective of the companies.

Organizations generally use lead scoring to determine which customer lead to prioritise. In most cases, the evaluation is based on the actions taken by the potential client when interacting with the business via various channels. Emails and website visits are examples of this. The leads are scored depending on the relevance of each action, and the leads with the highest overall scores are subsequently pursued by salespeople, according to a simple model. Manual lead scoring is the name given to this method.

The goal of this article is to show how the automation of the manual process of calculating lead can be done using machine learning. Machine learning is also used to improve the process of lead score analysis. To do this, real-world data is used to create and assess several machine learning models that serve as the foundation for automated lead scoring, as well as to show common problems that arise during the overall data preparation process. Also, we attempt to demonstrate how the lead score results might assist in revealing other business insights, such as the significance of various customer touch points, using a variety of data visualisation tools. Understanding “how machine learning and data analytics can be used to automate lead scoring and create business insights for the decision makers” is the article's research goal.

This article is divided into different sections including brief literature review on the topic, machine learning and automation of lead score analysis process followed by the description of the data used to study, discussing the result achieved and finally conclusion.

II. BACKGROUND

These days, businesses develop and gather data, and they follow the same patterns. A major part of marketing, currently referred to as relationship marketing, is "the continual process of engaging in collaborative actions in programmes with immediate and end-user consumers to develop or enhance mutual economic, social, and psychological value, profitably" . Relationship marketing is heavily dependent on the availability of digital data, which is becoming more and more important for firms as they strive to maintain their competitiveness. Collecting this digital data allows organizations to collect data on how possible future customers and interested people, i.e. leads, have interacted with various online communication channels available.

Customer relationship management can be greatly improved by tracking these behaviours and using various advanced business analytics tools or machine learning on the data gathered. These helpful data are gathered using a variety of online channels, including e-commerce websites, software, and email. In general, the key findings of numerous studies support the assertion that organisations should pursue data-driven decisions when implementing a (automated) lead scoring solution to replace or at least supplement manual lead scoring in the presence of this opportunity to use data in marketing and customer relationship management rather than relying on gut instinct or business intuition.

To rephrase these observations for our particular context, we can say that automated marketing is the process of utilising data from following potential leads’ online activities to discover behavioural patterns of these potential customers that can help in identifying the ones who are more likely to become actual customers. Although there are many tools that can assist these processes in an automated manner, there aren’t many studies looking at how businesses may use these tools to help potential customers as they engage in various stages of the B2C sales process [8]. We provide a quick literature evaluation on lead scoring and machine learning applications in automated customer relationship management based on this brief talk.

A. Manually calculating lead scoring

Manual lead scoring has a number of drawbacks. Most notably, manual lead scoring doesn’t use statistical evidence to back the recommendations. A lack of precise information for some leads with a high assigned scoring weight can also greatly skew the results, as manual lead scoring often relies on a wide range of demographic, behavioural, or firmographic data. Furthermore, because the human lead scoring method is based on a lead scoring matrix, businesses must manually analyse and update this scoring matrix regularly if they want to stay up with the rapidly changing business environment.
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 | +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

Figure 1: Example manual lead scoring matrix

Figure 1 shows an illustration of a scoring matrix. An experiment with 800 leads were manually rated. They discovered no statistically significant difference between randomly selecting leads that had not been scored at all and being able to convert scored leads that were assessed to be “ready for sales”. It is impossible for someone without statistical knowledge to accurately score or weigh these activities. Also, it takes a lot of effort to continually alter the ratings, and that time could be used more productively in other ways.

B. Components of lead scoring

Lead scoring is task of customer relation management. Lead scoring is process of assigning a numeric value to potential lead customer. A higher lead score suggests that the contact, or lead, is more likely to engage with the organisation. According to, low priority leads should be involved in lead nurturing initiatives while high priority leads should be forwarded to sales. The selection of variables used in the lead scoring models is the most important process that significantly affects the output quality of the lead scoring system. Explicit data (obtained directly from the customer's input) and implicit data (obtained indirectly via collecting data on the behaviours of potential leads) are the two basic categories into which the acquired data can be divided . The lead score models of the best-performing businesses typically comprised three or more implicit variable features.

Predictive marketing, which is defined as “a customer-centric marketing approach that attempts to enrich the customer's experience throughout the customer life cycle,” is when predictive analytics is used to score leads. The availability of technology that captures data that was previously inaccessible to the average marketer has made this experience possible. The sharp drop in processing costs is another element that helps predictive marketing succeed.

Statistical models or machine learning algorithms are two examples of the insights that can be produced from data using a collection of approaches called predictive analytics . The three primary categories of machine learning algorithms are supervised, unsupervised, and reinforcement learning. Getting a numerical value that forecasts the possibility of a client lead becoming a sale is the primary objective of lead scoring. This is an example of a problem that falls under the category of supervised learning. Using historical information about previous leads, such as their characteristics and the observed outcome of the lead (whether it became a customer or not), we try to develop a model that can forecast the outcome for new leads.

A lead scoring model is constructed utilizing Bayesian networks. This approach allows combining expert knowledge and historical data in a straightforward manner requiring a small amount of data. In the impact of utilizing modern information technologies such as machine learning to improve the efficiency of managing the customer journey, including how to effectively shorten the customer journey and related sales cycle in business-to-business firms using new technologies.

C. Machine learning examples from customer relationship management

we show some relevant application of machine learning in customer relationship management to gain insight. According to their literature review, the most widely used machine learning models in customer relationship management include classification, association rules mining, clustering, regression, forecasting, sequence discovery and visualization. The most common machine learning algorithms used include association rule, decision tree, genetic algorithm, neural networks, K-nearest neighbor and linear as well as logistic regression

A decision support tool is constructed that aids in predicting customer loyalty in a non-contractual setting using random forest, logistic regression and neural networks. Logistic regression was included as a comparison point for the more advanced models. The random forest algorithm is used in lieu of a decision tree algorithm due to their robustness and superior performance. The model is evaluated using accuracy and AUC. The model was successful in detecting future partial defection and there were no noticeable differences in the models created by the three algorithms.
III. METHODOLOGY

In the study, the general recommended process for predictive analytics in information systems research is applied. With the focus of the research being on the construction and evaluation of possible predictive machine learning models for automated lead scoring.

A. Data description and data preprocessing

- In computing, data is information that has been translated into a form that is efficient for movement or processing.
- Data preprocessing is a component of data preparation. Data preparation and filtering can take a lot of time. It is done to check the quality of data in terms of its accuracy, completeness, consistency, timeliness, validity and interpretability.
- We begin by attempting to transform data from a predetermined form to one that is significantly more desirable and usable.
- Methods for data preprocessing used are:
  - Univariate analysis: Univariate analysis explores each variable in the existing data set, separately. It is the simplest technique to analyse data. Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summarise it, and then find some pattern in the data.
  - Removing NULL values: removing the null values or drop the null values from the columns depends on how much value is present or how many are there.
  - Checking the dependencies of variable: examining the interdependencies between various factors so that we can determine which features will work best in our situation and examine the correlation.
  - Removing the outliers: Outliers are those data that are significantly different from the rest of the data set. They are anomalies that skew the data distribution and are generally caused due to inconsistent data entry. The outliers can be detected/removed using Z-score, Scatter plot or Interquartile range(IQR).

Variable for the analysis were filtered out based on the following criteria:
- correlation with the output label
- number of unique values (in categorical variables)
- number of missing values

From the main dataset, the following datatypes were selected:
- Identifier: links the same contacts in different n data sets
- Location: specifies the region of the lead
- Marketing unit: specifies the location of marketing unit
- Date created and modified: timestamp for events related to the lead
- Email address domain: specifies the email domain
- Contact status and time:

B. Model creation and training the model

The Model creation
- We will develop a model and specify its parameters using statsmodels' GLM (generalised linear model).

Training Phase
- Now we train the data with our GLM model, this model does not need any GPU powered computation.
- For this, we will use pandas, numpy, matplotlib, and seaborn.

Splitting our Data Set into Training Data and Test Data
- scikit-learn makes it very easy to divide our data set into training data and test data. To do this, we’ll need to import the function train_test_split from the model_selection module of scikit-learn.

Model Performance Evaluation (Algorithms used):
- random forest
- Gradient Boosting
- LightGBM
- CatBoost

Hyper parameter Tuning: When creating a machine learning model, one has to make choices as how to define architecture of your model. Parameter which shapes the model architecture is known as hyperparameters, thus the
process of creating your ideal model architecture is referred to as hyperparameter tuning. Example: How many trees should I include in my random forest? Or what should be the maximum depth allowed for my decision tree?

Random Forest algorithm: Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.

- It can be used for both Classification and Regression problems in ML.
- It is based on the concept of ensemble learning.
- Ensemble learning is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Gradient Boosting Algorithm

- Machine learning is one of the most popular technologies to build predictive models for various complex regression and classification tasks.
- Boosting is one of the popular learning ensemble modeling techniques used to build strong classifiers from various weak classifiers. It starts with building a primary model from available training data sets then it identifies the errors present in the base model. After identifying the error, a secondary model is built, and further, a third model is introduced in this process. In this way, this process of introducing more models is continued until we get a complete training data set by which model predicts correctly.

LightGBM

- LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.
- It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks.

CatBoost

- CatBoost or Categorical Boosting is an open-source boosting library developed by Yandex. In addition to regression and classification, CatBoost can be used in ranking, recommendation systems, forecasting and even personal assistants.

IV. RESULTS

Based on the final dataset described in the previous section, four 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:

- Random Forest
- Gradient boosting
- LightGBM
- CatBoost

In order to evaluate the performance of the constructed machine learning models, as it is common in practice, different evaluation metrics based on the confusion matrix are used [14]. By differentiating between correct and incorrect classifications on the two possible output classes, we can count true positive (TP), true negative (TN), false positive (FP) and false negative predictions (FN). In this paper, a positive case refers to a converted lead and negative case refers to leads with no actual purchase. Additionally to the basic accuracy measure, in order to account for the different types of errors, we can calculate metrics such as precision, recall, sensitivity and specificity. The final evaluation measure utilized in this paper is the Area under the Curve (AUC) that can be obtained by calculating the area under the Receiver Operating Characteristic (ROC) curve. ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) across different probability thresholds.
The vast majority of leads indicate "No" under Do Not Call, suggesting that the sales team is free to contact them by phone. The conversion rate among leads that indicate "No" under Do Not Call is 38.52%. While very few leads have a "Yes" under Do Not Call, the conversion rate among these appears to be 100%.

A. Model accuracy

Model Accuracy:
- Random Forest: When it comes to train accuracy, Random Forest have the accuracy of 98.4677% while test accuracy has been declined to 91.693% which is significant drop.
- Gradient Boosting: For train dataset, we have a accuracy score of 91.7350% while for test dataset, we have a accuracy score of 91.657% which is pretty good as there is no much accuracy drop as compared to Random Forest.
- LightGBM: The LightGBM algorithm gives us a train accuracy of 94.582% while test accuracy of 91.549%.
- CatBoost: Under Catboost, we have a train accuracy of 94.05% while test accuracy of 92.018%. In Catboost algorithm, we have the highest test accuracy as compared to Random Forest, Gradient Boosting, LightGBM. (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

B. Model precision

- Random Forest: When it comes to train precision for our class labels, we have a precision score of 97.95% for class label "0" and 99.30% for class label "1" while on test dataset this has been reduced. On testing dataset, precision score for class label "0" is coming out to be 91.84% while for class label "1" it is coming out to be 91.42%.
  - This indicating that our model requires parameters needs to be change as the score has come down significantly on the testing dataset.
- Gradient Boosting: On our training data for class label "0" we have a precision score of 90.71% while for class label "1" we have a precision score of 93.63%.
  - On testing dataset for our class label "0" this has been increased from 90.71% to 91.37% while for class label "1" this is slightly down i.e; 92.18% but still it is pretty good as compared to Random Forest.
- Light GBM: When it comes to Light GBM, our training precision score for class label "0" is coming out to be 94.12% while for class label "1" it is coming out to be 95.37%.
  - As far as the testing dataset concern, the precision score of class label "0" is coming out to be 92.11% while for class label "1" it is coming out to be 90.56%.
- CatBoost: Under CatBoost, for class label "0" under training dataset our precision score is coming out to be 93.43% while for class label "1" it is coming out to be 95.15%.

C. Probability Calibration

Two diagnostic tools that help in the interpretation of probabilistic forecast for binary (two-class) classification predictive modeling problems are ROC Curves and Precision-Recall curves. A common way to compare models that predict probabilities for two-class problems is to use a ROC curve. We can plot a ROC curve for a model in Python using the roc_curve() scikit-learn function.

The function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. The function returns the false positive rates for each threshold, true positive rates for each threshold and thresholds.
Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. Precision is a ratio of the number of true positives divided by the sum of the true positives and false positives. It describes how good a model is at predicting the positive class. Precision is referred to as the positive predictive value.

In order to map predicted values to probabilities, we use the sigmoid function. The function maps any real values into another value between 0 and 1. In machine learning, we use sigmoid to map prediction to probabilities.

V. RESULTS

In conclusion, lead score analysis is a valuable tool for businesses to evaluate and prioritize potential leads based on their likelihood to convert into customers. By analyzing various factors such as demographics, behavior, and engagement, businesses can assign a score to each lead and focus their resources on those with the highest scores. One of the primary benefits of lead score analysis is its ability to improve the efficiency of the sales process. Instead of spending time and resources on leads that are unlikely to convert, sales teams can prioritize their efforts on leads with the highest scores, increasing their chances of success. Additionally, lead score analysis can help businesses identify areas for improvement in their lead generation and nurturing strategies, allowing them to make data-driven decisions to optimize their processes. Another advantage of lead score analysis is its ability to provide valuable insights into customer behavior and preferences. By analyzing the factors that contribute to lead scores, businesses can gain a deeper understanding of what drives customer engagement and loyalty, and adjust their strategies accordingly. However, there are also some limitations to consider when using lead score analysis. In conclusion, lead score analysis is a powerful tool that can help businesses streamline their sales processes, improve customer engagement, and make data-driven decisions to optimize their strategies. However, it is important to approach lead score analysis with caution, ensuring that the input data is accurate and complete, and understanding that lead scores are not always a guarantee of success. With proper implementation and ongoing analysis, lead score analysis can be a valuable asset for businesses looking to grow and succeed in today's competitive marketplace.

REFERENCES

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), 2022