International Journal for Modern Trends in Science and Technology, 8(07): 188-194, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0807027
Available online at: http://www.ijmtst.com/vol8issue07.html

Predictive Lead Scoring with Feature Selection Techniques

Ishan Dusane | Jyothi Rao

Department of Computer Engineering, K.J. Somaiya College of Engineering, Mumbai, Maharashtra, India.
Email: ishan.dusane@somaiya.edu

To Cite this Article
Ishan Dusane and Jyothi Rao. Predictive Lead Scoring with Feature Selection Techniques. International Journal for Modern Trends in Science and Technology 2022, 8(07), pp. 188-194. https://doi.org/10.46501/IJMTST0807027

Article Info
Received: 08 June 2022; Accepted: 05 July 2022; Published: 10 July 2022.

ABSTRACT

Customer Relation Management is a very crucial part of any business even if the business is a small and medium businesses. Every business needs an insight of how their business is actually behaving internally. Companies collect large amounts of data, but this data is useless unless and until proper knowledge is extracted from the data. This knowledge can be a source of competitive advantage and can help our business stand out from others. In this paper we are more interested in using such data and use it for Lead Scoring which refers to the practice of calculating and assigning a score to a particular lead (customer).This lead score helps the sales teams to improve the efficiency of their own and have a fair idea which lead will get converted to a customer and which leads won't and needs to send to the marketing team for nourishment. The purpose of this article is to show how machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, XGBoost algorithms can help us achieve our goal of predicting whether lead would be converted to customer or not. In this article we'll see how different machine learning algorithms can have different behaviours for the same data and we'll also see how data pre-processing and different types of feature Engineering were performed to get the optimal results. Results show that we can estimate the purchase probability of the lead using different machine learning algorithms and how we can optimize the results with different feature selection techniques.

KEYWORDS: ***Predictive Lead Scoring, Logistic Regression, Decision Tree, Random Forest, XGBoost, Feature Selection Techniques.***

1. INTRODUCTION

In this competitive business environment everyone needs to stand out to get to the customers. It is very important to convert leads to customers and it is not at all easy to convert a lead to a customer. In this acquisition phase companies use a lot of different techniques to convert a lead into a customer. The process begins with approaching potential customers through various channels like Email, SMS, Websites, Advertisement, taking personals information from forms. Then the potentials lead is identified and pursed by sales person. But even with this all data it is very difficult to recognize potentials leads. As sale's process is very expensive both in terms of time and money it is very important to use this resources efficiently. Here Predictive leads scoring can be very useful. Lead scoring is a procedure which is applied by an organization to predict which potential customers can be converted to a

target. Typically scores are assigned by analysing the activities performed by customers like replying to email, visiting a product website, responding to surveys. In the lead scoring process a score is assigned to a lead and a lead with the highest score is then pursued by the salesperson.The next section (Section II) consists of a review of existing methodologies considered totackle this problem statement. These methodologies include machine learning techniques. Section III discusses about the implementation details. The process of implementation right from data acquisition to final testing and comparison has been detailed. Results of the experiments are discussed in Section IV.The conclusion related to this study and the prospect of future work have been presented in Section V.

2. Related Work

Before discussing the main components of automated lead scoring, it is important to discuss the dominant approach used in practice as identified in the introduction section: manual lead scoring. According to experts, there are several problematic issues with manual lead scoring. Most importantly, manual lead scoring fails to base the recommendations on statistical support. Additionally, as typically, manual lead scoring relies on a wide set of demographic, behavioural or firmographic data, lack of some specific information for some leads with high assigned scoring weight can significantly distort the results. Finally, as the manual lead scoring process is based on a lead scoring matrix, if companies aim to keep up with the constantly changing business environment, they have to manually revaluate and update this scoring matrix continuously.

So, to automate the lead scoring process, the author in [1] suggested using the machine learning algorithm with aggregation techniques by considering customer activity based on lead activity. In Aggregation 1, the end date for non-customers was set as the end of the time period considered in the data, while for converted leads it was set as the same date as their first purchase. In aggregation 2, the end date for non-customers was set as the date of their last activity, while for converted leads it was set as the end of the time period considered in the dataset. In aggregation 3, the finish date for non-customers was set as their last activity date, while for converted leads it was set as the last activity date before their purchase. In aggregation 4, the finish date for non-customers was set as the date of a randomly chosen date between the first and last activity, while for converted leads it was set as the date of the last activity before the 'purchase. In aggregation 5, the end date for non- customers was set as the date of a date chosen randomly between the first and last activity, while for the converted leads the date of their purchase was set. These aggregations were applied to Logistic Regression, Decision Tree, Random Forest, Multilayer Perceptron. The best model out of all algorithms came out to be Random forest with aggregation strategy chosen as 1, with accuracy, sensitivity, specificity all 69%, and AUC of 79%.

Lead scoring Model can also be improved by input from expert knowledge as done in [2]. In this study, they present a way to build a lead scoring model with a Bayesian network. In addition to its ability to handle uncertainty, Bayesian networks are knowledge representation models that can be built from expert knowledge. In our specific context, we therefore propose to build our Lead scoring model starting from experience. We apply the usual heuristics to reduce the complexity of our model (parental divorce, NoisyOr) and three ways to estimate the parameters of our NoisyOr sub models. The only data available is used to validate our approach, with good accuracy and recall results. The data used in this study was very small. Precision got was 80%, Recall was 88%, accuracy was 75%.

In [3] quiet a few variables were used for predictions. Empirical results show that classification models can provide a probability of (partial) defection of the individual given all the individual data collected by the retailer (behavioural and demographic of customers). As a result, we are able to track down future (partial) defectors. For managers, this classification is very useful for establishing new marketing strategies towards client companies. This study has several limitations. First of all, the results are limited to the fast-moving retail consumer goods (FMCG) sector. To some extent generalizations can be made for all other companies active in a non-contractual context where defection is difficult to detect. Demographics and past purchasing behaviour were used as input into the models, based on

data from a company-internal data warehouse. However, this list of predictors can be extended with customer perceptions in order to increase model performance. Unfortunately, this type of data is typically not available in data warehouses. Logistic Regression, ARD NN, and Random Forest were applied on data.

According to [4], the most widely used machine learning models in customer relationship management include association rules mining, classification, clustering, forecasting, regression, sequence discovery and visualization. The most common machine learning algorithms used include association rule, decision tree, genetic algorithm, neural networks, K-nearest neighbour and linear as well as logistic regression. This finding was one of the main reasons for the selection of algorithms tested in the empirical study presented in the main part of this article.

3. Proposed Methodology

After performing a detailed literature review of various techniques used we have selected appropriate machine learning models suitable for our data. The overall proposed architecture of our implementation consists of many steps. These steps start with data acquisition, followed by data cleaning, data preprocessing, exploratory data analysis, designing the model, training the model, and concluding with evaluating the model by comparing it with existing architectures. The pictorial representation of said architecture can be seen in Figure 1.

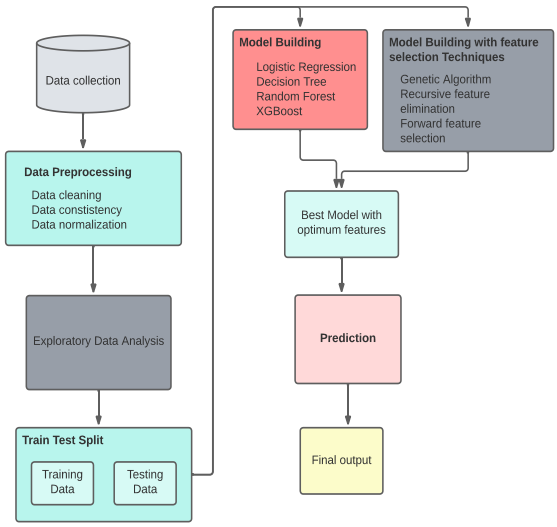

Fig 1. Proposed overall architecture

A. Dataset

The main purpose of this study is to see how predictive lead scoring can be done efficiently using different machine learning algorithms on a publicly available data on Kaggle. The data has both categorical and numerical variable. Few variable names and their type are given below.

- Response variable- Converted.
- Numerical variable- Total time spend on website, TotalVisits, Pages Views Per Visits.
- Categorical- Lead Origin, Lead, Source.

Data set contain 9240 rows and 37 features including numerical, categorical and response variable

B. Data cleaning

After the dataset collection, data cleaning was performed. The dataset cleaning includes following steps.

1.Handling missing values

The features who have more than 50% of missing value were dropped. For categorical variable who didn't had missing value more that 50% in that case the missing value were replace by most occurring instances.

2.Making Data consistent

In some places the the data was inconsistent for example in the "Lead Source" column there were two values for same Lead Source, like "Google" and "google".This both values can get treated as different values. So in such places data was made consistent

C. Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and examine data sets and summarize their main characteristics, often using data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to find patterns, find discrepancies, test hypotheses, or test assumptions.

EDA is primarily used to see what data can be revealed outside of formal modeling or hypothesis testing work and provides a better understanding of the data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. There are mainly four types of EDA: Univariate non-graphical,Univariate analysis, Multivariate Non-Graphical, Multivariate Graphical.

We used Univarite analysis for our study. As as result of EDA many insight full information about data were discovered. Many features had only one value and no information could be drawn from such feature so they were not passed to model.

D.Implementation Details

1.Logistic Regression

Logistic Regression is one of the simplest, easy to train, easy to interpret and robust models one can use for binary classification.It is very appealing because: 1. A closed-form solution for the posterior probabilities is available (as opposed to probit); 2. The basic assumption of logit (the logarithm of the ratio of group-conditional densities is linear in the parameters) is satisfied by many families of distributions [6]; Due to the fact that the accuracy is not suitable for this situation, we will have to use another measurement to decide which cut-off value to choose, the ROC curve. when we say that when we choose our cut-off value, we are striking a balance between the false positive rate (FPR) and the false negative rate (FNR), you can think of this as the objective function for our model, where we are looking for to minimize the number of mistakes we are making or so called to cost. Well, the purpose of the ROC curve is used to visualize and quantify the trade-off we are making between the two measures. This curve is created by plotting the true positive rate (TPR) on the y axis versus the false positive rate (FPR) on the x axis at various cut-off settings (between 0 and 1). We will calculate accuracy, sensitivity and specificity from 0 to 1 and plot them [7]. The optimal cut-off we got is 0.3. All results with probabilities greater than 0.3 will be marked as "1" and below those values will be marked as "0". This increased the the overall performance of model by 2%

2. Decision Tree

Decision Tree is a supervised learning technique that can be used for both classification and regression problems, but is primarily preferred for solving classification problems. It is a tree-structured classifier, where the internal nodes represent the characteristics of a data set, the branches represent the decision rules, and each leaf node represents the result.Decision trees are very “natural” constructs, in particular when the explanatory variables are categorical[12]. They are best model expandability. But some time the tend to overfit. With default configuration of decision tree there was a difference of 12% in recall and 10% in accuracy of train and test data with train being on higher side. So clearly our model was overfitting with default configurations. We can use minimal cost complexity pruning for avoid overfitting. In Decision treewe can control overfitting by ccp_alpha tuning. In Fig. 2we can see train and test behaviour for various values of alpha. We can see as the alpha values increases train and test recall converges. For higher values it converges more but the value of recall falls.

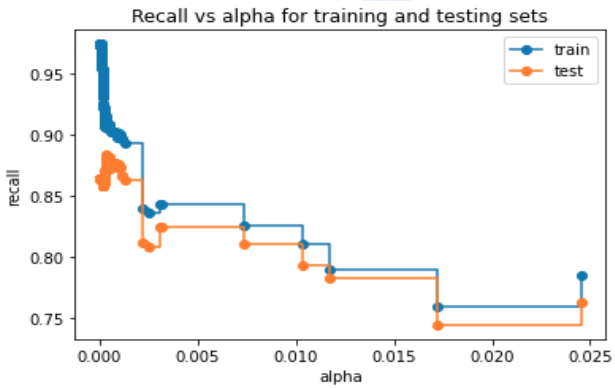

Fig 2. Recall vs alpha for training and test set

We can also avoid overfitting in Decision Tree by fine tuning “max_depth” and “min_sample_leaf”. In Fig. 3. we can see how recall of train and test values differ for different values of max_depth. We can see as the max_depth model tends to overfit.

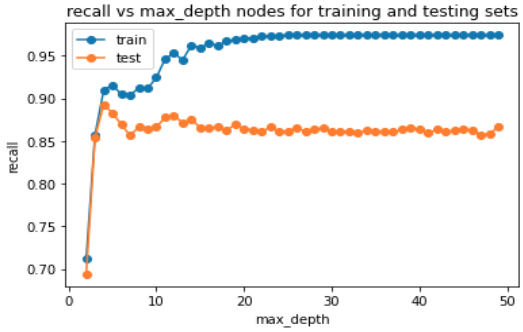

Fig 3. Recall vs max_depth for training and test set

Min_sample_leaf is the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_sample_leaf training samples in each.In Fig. 3. after fine tuning max_depth and min_sample_leaf we can see convergence between train and test set

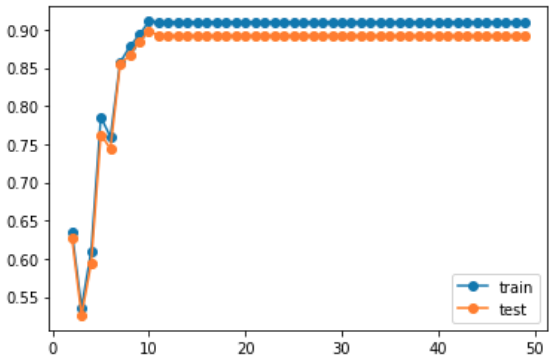

Fig 4. Train and test recall of fine tuned model

3. Random Forest

Decision trees have become very popular for solving classification tasks because they can deal with predictors measured at different measurement levels (including nominal variables) and because of their ease of use and interpretability [8]. However, they also have their disadvantages such as lack of robustness and suboptimal performance [9]. Random Forest are more robust and overcomes disadvantages by ensemble of trees. But sometimes it tends to overfit. In our case the accuracy and recall of train is 10% more than test data. But we can avoid overfitting by parameter tuning.To avoid overfitting in we have tuned max_depth, min_sample_split, max_leaf_nodes and n_estimators. GridSearchCV was used with 10-fold cross validation to find the best combination of max_depth, min_sample_split, max_leaf_nodes and n_estimators. After fine tuning the difference between train and test values came down from 10% to 1% i.e. fine tuned model was more generalized compared to model with default configuration.

4. Extreme Gradient Boosting (XGBoost)

Before we move to XGBoost first we need to understand Bagging and Boosting. Bagging classifier is a estimator that fits the base classifier each on random subset of the original data set and then combines their individual predictions (by voting or averaging) to form final prediction. Such estimator can be used to reduce variance of black-box estimators like decision tree by including randomization into its construction procedure and making it a set. In gradient enhancement each predictor corrects the error of it predecessor. XGBoost is implementation of Gradient Boosting trees. In this algorithm the trees are created in sequential form. Weights play a very important role in XGBoost. The weights are assigned to independent variable which are then inserted into tree which predicts the results. The weights which are not correctly predicted variables from the previously tree are increased are send to second tree. The individual classifier then come together to give a strong and accurate model. XGBoost has aninbuilt L1 and L2 Regularization feature. As we saw in case of Decision tree and Random Forest for our data, both models were getting overfitted with default configuration. But that was not the case with XGBoost. XGBoost didn't get overfiitted with default configuration. But we can still fine tune the model to increase the performance of the model. We fine tune the model in terms of gamma, max_depth, min_child_weight, subsamples, colsample_bytree, learning_rate. GridSearchCV was used with 10-fold cross validation to find best combinations of hyper parameter.

5. Feature Selection Techniques

Wrapper methods: In wrapper methods the feature selection is based on specific machine learning algorithm that we are trying to fit to a certain data. We try to achieve best feature combination against the evaluation criterion. The evaluation criterion can be simple measure of performance that depends on type of problems. For regression it can be p value, R squared, R squared adjusted, in the same way for classification it can be accuracy, recall, f1 score. Finally select the combinations of features that provides optimum results for specified machine learning algorithm.

5.1. Recursive Feature Elimination (RFE)

Recursive Feature Elimination is a feature selection method that fits the model and removes the weakest features until the specified number of features are achieved. Features are selected based on coef_ or feature_importance_ attributes of the model and recursively delecting small features per cycle. RFE attempts to eliminate dependencies and collinearity that may exsist in the model.

5.2. Forward Feature Selection (FFS)

In Forward Feature Selection the process starts with the

empty set and one by one the features are added until we get specified number of features. First the best feature is selected. Then the pair functions are formed using one of remaining features. Then considering these two best features triplets are formed. Process continues until we get specified number of features.

5.3. GeneticSelectionCV (GSCV)

GeneticSelectionCV is a genetic feature module from sklearn-genetic provided by scikit-learn. Genetic algorithms mimic the process of natural selection to search the optimum values of function. More details about the module can be found on [10]

E. Model Evaluation

Different indicator can be used to evaluate the machine learning model. Using testing dataset and model prediction for same dataset we can construct confusion metric as shown in Fig. 5.

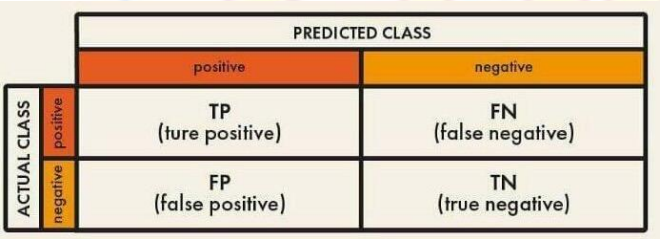

Fig 5. Confusion Metrics

To increase performance of model at any instance we should try to increase TP and TN scores and minimize FN and FP.

1. Precision: The proportion of number of correct positive prediction among entire positive prediction.

Precision = TP/(TP+FP)

2. Recall: The proportion of correctly number of correct positive prediction among the entire positives.

Recall = TP/(TP+FN)

3. f1 score: f1 score is popular classification metrics used over accuracy when the data is imbalanced (i.e. the quantity of a particular class is greater than another class)

f1 = 2*(precision*recall/precision+recall)

4. Area Under Curve (AUC): F1 score is applicable for any particular point on ROC curve with different thresholds. AUC (area under the ROC curve) indicates how well probabilities from positive classes are separated from negative classes.

4. RESULTS

After data collection, preprocessing and Exploratory data analysis the data was passed to the machine learning model which were found best during literature survey. The train and test data was split into 70:30 ratios. The models were fined tuned to increase the performance. All the code was implemented in Python language. Data preprocessing, Exploratory data analysis was done with the help of libraries like pandas, matplotlib. Machine Learning models was implemented by Sklearn. This experimental study uses a relevant data to predict the likelihood of a potential customers(leads) to get converted into customers based on historical data of the customers. The results of the experiments can be seen in Table 1.

Table I. The Experiment results

Models	Recall	Precision	F1 score	AUC
LR	0.8843	0.9196	0.9016	0.9034
DT	0.8956	0.9002	0.8978	0.9101
RF	0.9201	0.8968	0.9083	0.9125
XGB	0.9130	0.8724	0.8922	0.9184

Comparing the results of constructed models, it is clear that Random Forest is outperforming all other models, achieving highest recall of all 92.01%, followed by XGBoost with recall of 91.30%

Feature Selection Techniqueslike Recursive Feature Elimination(RFE), Forward Feature Selection (FFS), GenecticSearchCV(GSCV)was applied to all the machine learning model to optimize model byselecting optimal features. Table II, Table III, Table IV shows performance metrics for all machine learning models with feature selection. **Note**: The numbers in the parenthesis indicates the number of feature selected.

Table II. Performance Metrics comparing all models combined with Recursive Feature Selection

Recursive Feature Selection			
Models	**Recall**	**Precision**	**AUC**
LR	0.8604(45)	**0.892(45)**	0.9066
DT	0.8756(16)	0.9002(16)	0.9101
RF	0.9169(30)	0.8677(30)	0.9142
XGB	0.9250(44)	0.8695(44)	0.9160

Table III. Performance Metrics comparing all models combined with Genetic Search CV

Genetic Search CV			
Models	**Recall**	**Precision**	**AUC**
LR	0.8624(30)	**0.8924(30)**	0.9034
DT	0.8705(14)	0.9071(14)	0.9099
RF	0.9150(49)	0.8545(49)	0.9131
XGB	0.9034(48)	0.8654(48)	0.9102

Table IV. Performance Metrics comparing all models combined with Forward Feature Selection

Forward Feature Selection			
Models	**Recall**	**Precision**	**AUC**
LR	0.8341(42)	0.9217(42)	0.9053
DT	0.8756(24)	0.9002(24)	0.9566
RF	0.9160(42)	0.8531(42)	0.9122
XGB	0.9079(44)	0.8735(44)	0.9165

Comparing Table II, III, IV we can see that Random Forest with Recursive Feature Elimination with 91.69 % of recall with just 30 features out of 85. XGBoost with Recursive Feature Elimination and Random forest with Genetic Search CV are competitive models with 92.50% and 91.50% recall respectively but with 14 to 19 features more that Random Forest with Recursive Feature Elimination

5. Conclusion And Future Scope

Predictive Lead Scoring is one of the most popular business problems which helps businesses to come up with new opportunities in business. In this study we saw how different machine learning algorithms can be used to achieve predictive lead scoring. Different machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost to classify that will the potential leads get converted into customers or not. We also saw how feature selection methods like Recursive Feature Selection, Forward Feature Selection, GeneticSearchCV can be used to optimized model by selecting optimal features.The best model (without any feature selection technique) was Random Forest with highest recall of 92.01%. We decreased the complexity of model by feature selection and best model(with feature selection) was Random Forest with Recursive Feature Elimination with recall of 91.96 % with just 30 features.There is still room for improvement in feature selection part where we can use meta-heuristic algorithms like Spider Monkey Optimization for feature selection.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

References

[1] Robert Nygård, József Mezei "Automating Lead Scoring with Machine Learning: An Experimental Study", 2020 53rd Hawaii International Conference on System Sciences

[2] Youssef Benhaddou, Philippe Leray, "Customer Relationship Management and Small Data - Application of Bayesian Network Elicitation techniques for building a Lead scoring model", IEEE/ACS 14th International Conference on Computer Systems and Applications

[3] E.W.T. Ngai , Li Xiu , D.C.K. Chau "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications 36 (2009) 2592–2602

[4] Wouter Buckinx, Dirk Van den Poel, "Customer base analysis: partial defection of behaviorally loyal clients in a non-contractual FMCG retail setting", European Journal of Operational Research.

[5] Jadli, Aissam, et al. "Towards a Smart Lead Scoring System Using MachineLearning.", Indian Journal of Computer Science and Engineering (IJCSE)

[6] Anderson, J.A., 1982. Logistic discrimination. In: Krishnaiah, P.R., Kanal, L.N. (Eds.), Handbook of Statistics, vol. 2, pp. 169–191

[7] https://ethen8181.github.io/machine-learning/unbalanced/unbalanced.html#Choosing the Suitable Cutoff Value

[8] Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classifica- tion. Wiley, NY.

[9] Dudoit,S.,Fridlyand,J.,Speed,T.P.,2002.Comparisonof discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97 (457), 77–87.

[10] GeneticSearchCv. Avalable: https://sklearn-genetic.readthedocs.io/en/latest/