CRM-based Lead Scoring with Machine Learning

Pradeep Thota
REVA Academy for Corporate Excellence (RACE), REVA University
REVA University
Bengaluru, India
pradeeot.ba06@reva.edu.in

Rashmi Agarwal
REVA Academy for Corporate Excellence (RACE), REVA University
REVA University
Bengaluru, India
rashmi.agarwal@reva.edu.in

Akula Phaneendra
REVA Academy for Corporate Excellence (RACE), REVA University
REVA University
Bengaluru, India
akula.res@reva.edu.in

***Abstract*— *Betutelage* is a startup organization that offers educational courses and provides live classes to audiences of different skill levels. They are a million-rupee revenue generator backed by investors who were impressed with their concept of providing valuable insights to students in areas where they can improve their focus on studies etc.**

***Betutelage* needs assistance in predicting their leads. These leads represent their highest-paying enquiry conversion customers. *Betutelage* needs a model that can assign a score to each lead so that their customers will have a good conversion rate when the lead score is high and vice versa.**

By using this model, *Betutelage* aims to invest more time in non-converting leads and convert them into paying customers. They will also be able to reduce the cost of their campaigning in areas where there are no leads.

This paper has opted to study 4 classification machine learning models, which are Random Forest, Gradient Boosting, LightBGM, and Catboost, using CRISP-DM methodology with the data provided by *Betutelage*. The aim is to find the best model among these models that have the highest accuracy to convert leads based on both test and train data. The outcome of this investigation demonstrates that Random Forest, with train and test accuracy of 94.3% and 92.02%, respectively, has the highest accuracy.

Keywords*— *Artificial Intelligence, Machine Learning, Deep Learning, Classification Models, Leads, Random Forest, Gradient Boosting, LightBGM, Catboost.

I. Introduction

Betutelage is an educational course selling startup company with live classes targeting all levels of audience and they are a million rupees revenue generators which are funded by some of the investors by seeing their vision where they are giving beautiful insights of student in which area they can improve their focus in studies, to know where their area of interest lies and how to make them get interested on a particular subject with their courses.

Now *Betutelage* along with the existing system they have entered into online courses for professional, academic, etc, to know the leads for their existing system and the new system they are looking for help to build a classification model to know the leads for their business, that who are likely to convert into the paying customers, for this, business have provided some data which they have collected from several sources to build a model.

Hence, based on the results they need to know how these leads are coming and how can they reduce their expenditure on unnecessary campaigns so that they can invest more on the path where they will get more monetary gain. So, to help *Betutelage* have to build several classification models to get solutions for their problem.

II. State of Art

For all organizations leads are very important, leads are a person or a company who are interested in the products, services, or offerings of the organization. Customer Relationship Management (CRM) is a task for companies which needs to be done on the daily basis, even when dealing with small data [1].

The fundamentals of the lead score are not only for the customer business but also matters for business-to-business which leads to multipliers for the market [2]. Not only running some campaigns but also calling over the telephone to a person and explaining the product will get the leads to the organization says [3]. It is clearly explained how to improve the net promotor score [4].

Lead scoring can be increased when it is implemented with the classification models like Random Forest, and logistic model [5], not only by applying these techniques but also need to do some applications of data mining techniques in CRM [6] and dealing with imbalanced class distributions [7] and doing some statistical learnings from the data [8].

Dealing with some missing values before building a model is also important as they will surely impact the model [9] and is also required to deal with the imbalanced dataset to avoid its impact of it on the huge dataset [10]. Once all the necessary data preparation steps are completed, one can build the model for predictive analytics [11]. This model can build on any classification method and artificial intelligence classification model [12]. This can also deal with big data [13].

Once the model is built, then it is important to evaluate the model, to know the Receiver Operating Characteristic (ROC) curve [14], and the metrics of the model. This information will be useful to get a Lead Prioritization and Scoring model with the path to higher conversion [15].

III. Problem Statement and Objective of the Study

Betutelage is an Indian-based startup company of educational selling courses with live classes targeting all levels of the audience and the company is based out of Bengaluru. *Betutelage* needs help in predicting the leads, these leads are the most paying customers of conversion from enquiry, The company requires a model assigning the score to each of the leads so that their customers have a good conversion rate when the lead score is high and vice versa.

The objective of this paper is to develop a classification model for the following points:

1) Assisting the business to know the leads who can convert to their paying customers, so the business needs a model that can predict accuracy about the customer.
2) By the above-built model they can invest more time on non-converting leads to make them convert, so that their course gets sold, not only that, they can also reduce the cost of their campaigning cost where there are no leads and those who are not turning into their paying customers.

Hence, need to help the business by building classification models with appropriate techniques using Machine learning, Deep learning, Artificial Intelligence, etc.

Data collection is not a crucial part of this development, because there is a good sample of data provided by the business that was collected from their server. The crucial part is data preparation and building a good model with great training and testing metrics of good accuracy.

IV. METHODOLOGY

Cross-Industry Standard Process for Data Mining (CRISP-DM) is the methodology used in this paper. It involves six steps which are captured in Fig. 1.

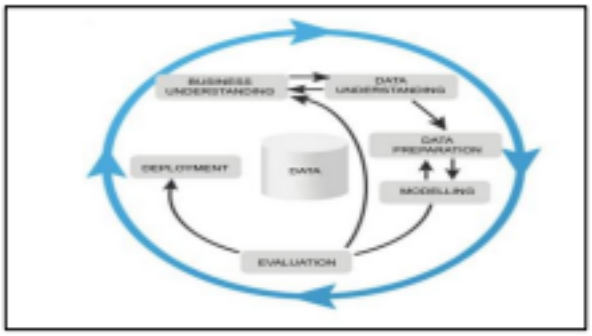


Fig. 1: CRISP-DM [16].

Business Understanding — The goal of this stage is to understand the business goal and then convert it into a measurable and specific project goal and then formalize it as a problem statement.

Data Understanding — The goal of this stage is to gather data and then explore and comprehend the data.

Data Preparation – The goal of this stage is to select the final data which will be relevant to the data mining objectives, and clean and transform the data.

Data Modelling - The goal of this stage is, to apply the modeling techniques and record them.

Model Evaluation – The goal of this stage is, to assess the degree to which the model meets the business requirements and to test the model in real applications.

Deployment - The goal of this stage is to determine the model deployment strategy based on evaluation results and a plan for monitoring and maintenance of models in the business environment.

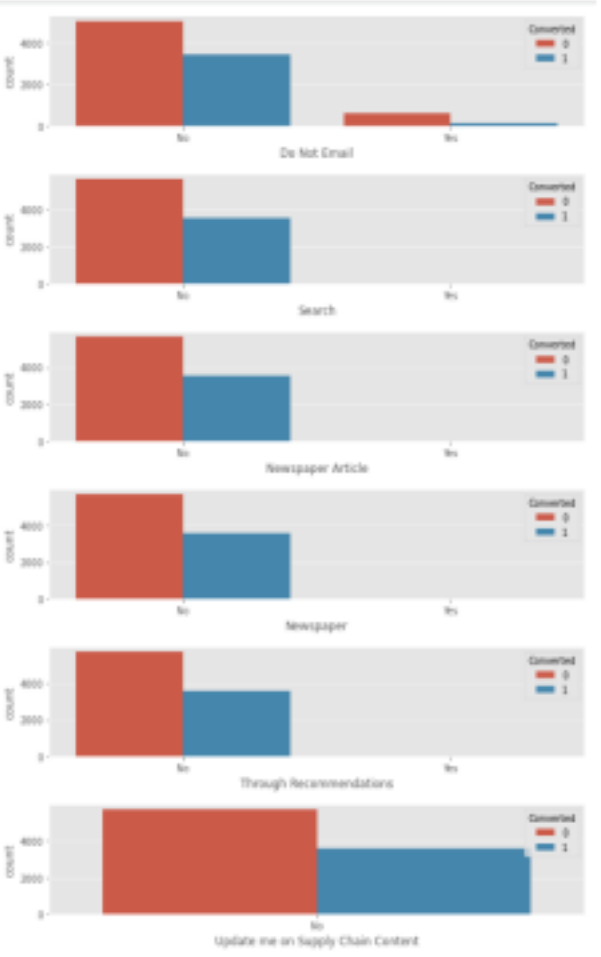
V. BUSINESS AND DATA UNDERSTANDING

As part of business understanding, this paper has a clear problem statement that the client needs to know the promising leads who can become their customers by taking up the course.

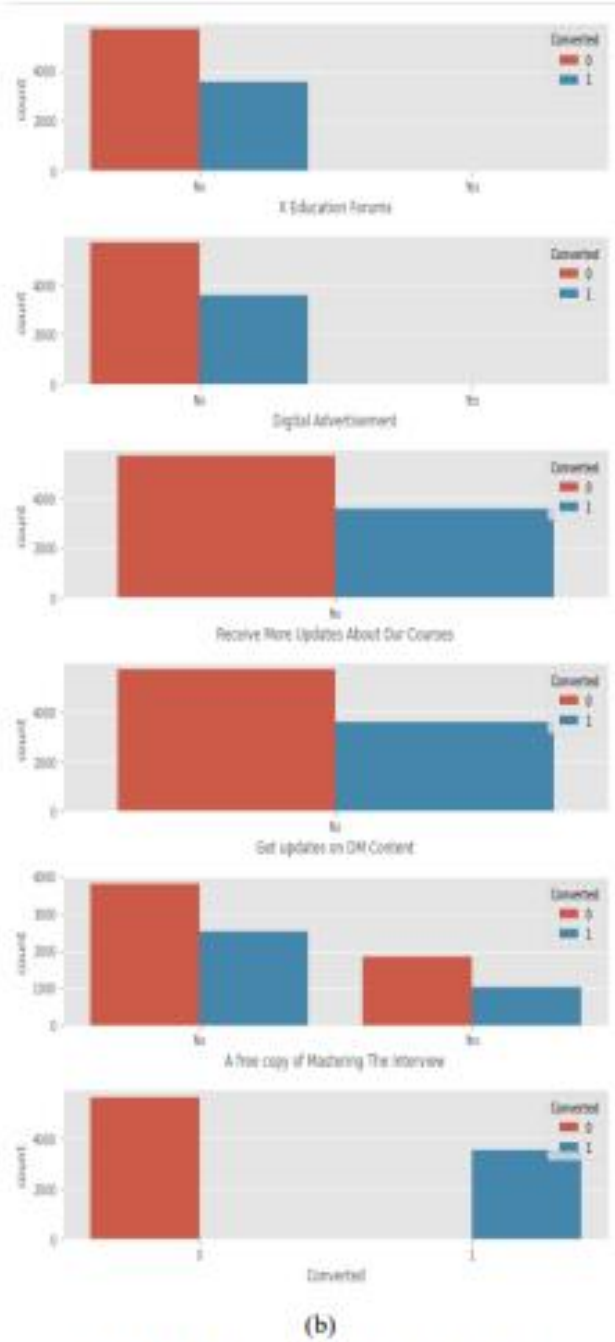
So, the business can conclude that customer who has the highest lead score will be having high conversion chances, and the customer who has the lowest lead score will be having low conversion chances.

Now businesses can concentrate on these low lead score customers to make them as their paying customers by applying appropriate strategies.

The list of visualization data understanding is as follows:
Users who come from the "Olark Chat" source usually have a Lead Origin "API" and most of them are not able to convert. When it comes to "Reference" businesses have a lead origin of "Lead Add Form" and mostly got converted.



(a)



(b)

Fig. 2: Lead conversion for individual variables

Based on Fig. 2 (a) and Fig. 2 (b), the target variable is having a 61.5:38.5 ratio, in the classification model. This ratio can be considered a balanced dataset. The proportion of users who do not convert is high as compared to the users who converted. Also, the users are not much interested in "Free Copy of Mastering the Interview" which is weird. The reason may be their disliking of freebies. Another reason may be the large proportion of "Unemployed" audience. The only thing they are interested in upskilling themselves and not giving priority to the interview preparation in the early stage.

Also, there are certain columns, which are not going to infer much information as most of the values are "No", hence those will be dropped in the later stage.

VI. Data Preparation

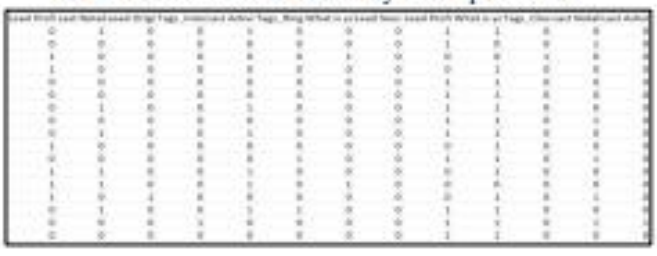
After finishing the data understanding, the data preparation steps are as follows:

- The data available with us qualifies for the classification model and can apply the same to see if a lead converts into a customer or not.
- Firstly, clean the data to improve its quality by eliminating variables that are not relevant.
- Combine low-frequency categories into a new category to compress the number of categories for improving the analysis.
- Identify and treat the missing values and the outliers in the data to stabilize the data set.
- Based on the different variables from the data which tell about the preferences and background of the people being approached as potential leads for business, try to first analyze the variables that seem to cause high conversion rates and also identify any correlations or patterns between the variables during EDA (Exploratory Data Analysis) phase.
- Then train and create a classification model which would predict the lead conversion with good sensitivity and accuracy scores.
- Evaluate the above model on the test data to predict the lead conversion and check the model sensitivity and accuracy scores.
- Lastly, find out the top variables that impact the lead conversion and summarize them so that it enables the client sales team to identify the potential customers.

Table No. 1: Raw Data Corpus

Prospect I	Lead Num	Lead Origi	Lead Sour	Do Not Em	Do Not Ca	Converted	TotalVisits	Total Time	Page View
7927b2df-	660737	API	Olark Cha	No	No	0	0	0	0
2a272436-	660728	API	Organic Se	No	No	0	5	674	2.5
8cc8c611-	660727	Landing Pa	Direct Traf	No	No	1	2	1532	2
0cc2df48-	660719	Landing Pa	Direct Traf	No	No	0	1	305	1
3256f628-	660681	Landing Pa	Google	No	No	1	2	1428	1
2258ef08-	660680	API	Olark Cha	No	No	0	0	0	0
9fae7df4-	660673	Landing Pa	Google	No	No	1	2	1640	2
20ef72a2-	660664	API	Olark Cha	No	No	0	0	0	0
cfa0128c-c	660624	Landing Pa	Direct Traf	No	No	0	2	71	2
a4483e8c-3	660618	API	Google	No	No	0	4	58	4
2a369e35	660608	Landing Pa	Organic Se	No	No	1	8	1351	8
9ba8ce93-	660570	Landing Pa	Direct Traf	No	No	1	8	1343	2.67
8bf76a52-	660563	API	Organic Se	No	No	1	11	1538	11
88867067-	660558	Landing Pa	Organic Se	No	No	0	5	170	5
a8531c22-	660553	Landing Pa	Direct Traf	Yes	No	0	1	481	1
25f4ac24-7	660547	API	Organic Se	No	No	1	6	1012	6
5abdb7c77-	660540	API	Olark Cha	No	No	0	0	0	0

Table No. 2: Data after necessary cleanup activities



As the client has given a good sample of data Table No. 1 requires minimal preparation, after necessary modifications fed the corpus as it is and once the data is prepared it looks like Table No. 2.

VII. MODELING AND MODEL EVALUATION

Data modeling and flow for this development are as follows in Fig. 3.

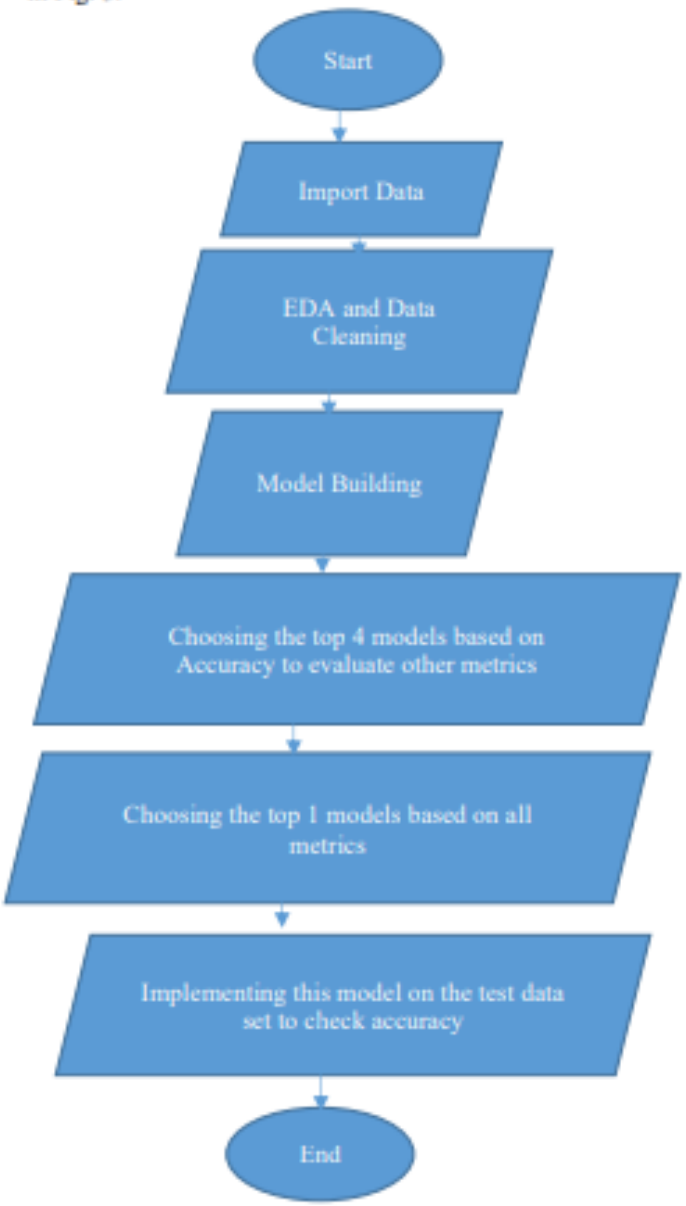


Fig. 3: Flow Chart

Once importing the necessary packages and corpus files then all the data is taken to check Exploratory Data Analysis (EDA) to get insights into the data.

As the data preparation is done then building the models based on 12 different classification methods RandomForest, Adaboost, Extra Tree, Bagging Classifier, Gradient Boosting, Decision Tree, K Nearest Neighborhood (KNN), Logistic, Stochastic Gradient Decent (SGD) Classifier, Multi-layer perceptron (MLP) Classifier, Naïve Bayes, Light Gradient Boosting Machine (GBM), Catboost and will finalize based on their metrics for final testing on validation data.

After building models on mentioned classifiers only RandomForest Classifier, Gradient Boosting, LightGBM & Catboost classifiers have been chosen for the next level based on top accuracy for checking other metrics like precision, recall, f1 score, and other metrics.

By checking all the metrics, can consider the **RandomForest Classifier** for the next step to predict the leads with the validation data and check the accuracy of the test data.

The model Evaluation for this paper is as follows:

Information Based on the Classification Models:

```
RandomForest : 0.9063846558066212
Adaboost : 0.901907180808915
ExtraTree : 0.9008954115890532
BaggingClassifier : 0.9006061857217926
GradientBoosting : 0.9121625003127894
DecisionTree : 0.8747460150639341
KNN : 0.8870261241648525
Logistic : 0.9043607003144574
SGD Classifier : 0.9008945774841728
MLPClassifier : 0.9008931178006323
NaiveBayes : 0.8601548098657925
LightGBM : 0.9088396349957044
Catboost : 0.9138966043590321
```

Fig. 4: Accuracy of the models

Based on the results shown in Fig. 4, one can choose the RandomForest Classifier, Gradient Boosting, Light GBM, Catboost and shown in Table No. 3 will test the other metrics to see the in-depth performance of these four models based on several different metrics to choose the best model for our analysis.

Table No. 3: Model Metrics

Model	Train Precision	F1-Score	Recall	Train Accuracy	Test Accuracy
Random Forest	88.06	88.8	99.60	98.5%	91.3%
Gradient Boost	91.03	93.63	96.39	91.9%	91.5%
LightBGM	94.05	95.5	97.1	94.4%	91.9%
CatBoost	93.67	95.40	97.1	94.2%	92.07%

Model Accuracy:

1) Random Forest:

When it involves training accuracy, Random Forest has an accuracy of 98.5% while test accuracy has declined to 91.3% which is good sized drop.

2) Gradient Boosting:

For provided dataset, have a training accuracy rate of 91.9% while looking at the test dataset, has an accuracy rating of 91.5% which is quite top as there may be no tons accuracy drop compared to Random Forest

3) LightGBM:
The LightGBM set of rules offers us a training accuracy of 94.4%, looking at a test accuracy of 91.9%.

4) CatBoost:
Under Catboost, have a training accuracy of 94.2% while looking at test accuracy of 92.07%. In the Catboost set of rules, have the best look at accuracy compared to Random Forest, Gradient Boosting, and Light GBM.

Model Precision:
1) Random Forest:
When it involves training precision for our elegance labels, have a precision rating of 98.06% for the class label "0" and 99.3% for the class label "1" while on taking a look at the test dataset this has decreased. On checking out the dataset, the precision rating for the class label "0" is popping out to be 91.01% whilst for the class label "1" its miles popping out to be 92.01%.

This indicates that our version calls for parameters wishes to be alternated because the rating has come down drastically at the checking out dataset.

2) Gradient Boosting:
On our train data facts for the class label, "0" have a precision rating of 91.03% while for the class label "1" has a precision rating of 93.59%. On checking out the test dataset for our elegance label "0" this has been expanded from 91.03% to 91.09% whilst for sophistication label "1" that is barely down i.e; 92.35% however nonetheless it's miles quite top compared to Random Forest.

3) Light GBM:
When it involves LightGBM, our training precision rate for the class label "0" is popping out to be 94.05% while for the class label "1" its miles coming to 95.18%.

For some distance, because of the checking out test dataset concern, the precision rating of the class label "0" is popping out to be 92.06% whilst for the class label "1" its miles popping out to be 91.85%.

4) CatBoost:
Under CatBoost, for the class label "0" below the training dataset our precision score is popping out to be 93.67% while for the class label "1" its miles popping out to be 95.18%.

For checking out the test dataset, the precision rating elegance label "0" it's miles barely down from 93.67% to 92.07% whilst for sophistication label "1" it's miles popping out to be 92.06%.

F1-Score:
1) Random Forest:

When testing the F1-Score for Random Forest Classifier on the training dataset, its miles pop out to be 98.8% for the class label "0" while 98.09% for the class label "1".

On checking out the test dataset, our F1 rating has come down from 98.8% to 93.0% for the class label "0" while for the class label "1" it miles popping out to be 88.5% which is a once more massive drop.

2) Gradient Boosting:
On the training dataset for the class label "0," our F1 rating is popping out to be 93.63% while for the class label "1" it's miles coming as 88.94%. For checking out the test dataset, the F1-rating for the class label "0" has been decreased to 93.21% while for the class label "1" it's miles 88.79%.

3) LightGBM:
On the training dataset for the class label "0," our F1-rating is popping out to be 95.5% while for the class label "1" it's miles coming as 92.5%. For checking out the test dataset, the F1-rating for the class label "0" has been decreased to 93.5% whilst for the class label "1" it's miles 89.5%.

4) CatBoost:
On the training dataset for the class label "0," our F1-rating is popping out to be 95.40% whilst for sophistication label "1" its miles come as 92.2%. For checking out the test dataset, the F1-score for the class label "0" has been right down to 93.59 % whilst for label "1" it's miles 89.6%.

Table No. 4: Random Forest Metrics on validation data set.

Train Accuracy	0.9436
Test Accuracy	0.9202

Finally, when implementing the learnings to the test model and calculating the conversion probability based on the Sensitivity metric and cutting off and found the train accuracy value to be 94.36%, the test accuracy was 92.02% as per Table No. 4.

Some of the key drivers for lead conversion of the paper are:
- Lead Origin: 'Lead Import' Category
- Do Not Email: 'Yes' Category
- Lead Source: 'Reference' Category
- What Matters to you the most in choosing a course: 'Not Provided' Category
- Specialization: 'Not Provided' Category
- Lead Origin: 'Landing Page Submission' Category

VIII. Deployment

After running a few more checks on the model by feeding in fresh data if the client provides and re-evaluating the importance of selected features, the same will be shared with the underwriters to get their opinions. Once the client approves to go ahead, this model will be used as a centerpiece for the client which will automatically give a lead score for a customer so they can decide further steps on them as per client requirements.

IX. Analysis and Results

The top three variables in the built model that contribute toward lead conversion are:
1. Lead Origin: 'Lead Add Form' Category
2. What is your current occupation? : 'Working Professional' Category
3. Total Time Spent on Website Metric

The 3 variables in our model that must be concentrated on to increase the lead conversion probability are:

1. Lead Origin: 'Lead Import' Category
2. Do Not Email: 'Yes' Category
3. Lead Source: 'Reference' Category

To focus on a greater number of the lead audience (inclusion of slightly lower conversion probable leads) users can alter (moving down) the value of cut-off to include more leads as the hot leads from our Logistic Regression model.

To reduce the lead audience (discarding lower conversion probable leads) user can increase the cut-off to discard lower probability leads from the model.

X. Conclusion and Recommendations for Further Work

After this development, there are a few recommendations to the client to make both sales and marketing team work together to sell the course based on user choice, of course, its always important to understand the target audience for the marketing team and there are few important lead qualification factors that need to be considered are knowing awareness of the need, budget, timeline/urgency, etc.

Also, it's always important to take new high-quality initiatives, thus there will be high-quality leads for the organization, running target-based ads based on their search, asking organization happy customers to refer to know contacts, etc.

References

[1] Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbor as an imputation method. *Frontiers in Artificial Intelligence and Applications*, *87*.

[2] Benhaddou, Y., & Leray, P. (2018). Customer relationship management and small data - Application of Bayesian network elicitation techniques for building a lead scoring model. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2017-October*. https://doi.org/10.1109/AICCSA.2017.51.

[3] Brown, H. E., & Brucker, R. W. (1987). Telephone qualification of sales leads. *Industrial Marketing Management*, *16*(3). https://doi.org/10.1016/0019-8501(87)90025-3.

[4] Carter, J. v., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery (United States)*, *159*(6). https://doi.org/10.1016/j.surg.2015.12.029.

[5] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6). https://doi.org/10.1016/S1532-0464(03)00034-0.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Springer Series in Statistics. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). https://doi.org/10.1007/b94608.

[7] Liu, Z. G., Pan, Q., Dezert, J., & Martin, A. (2016). Adaptive imputation of missing values for incomplete pattern classification. *Pattern Recognition*, *52*. https://doi.org/10.1016/j.patcog.2015.10.001.

[8] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*. https://doi.org/10.1016/j.patcog.2019.02.023.

[9] McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10).

[10] Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. In *Expert Systems with Applications* (Vol. 36, Issue 2 PART 2). https://doi.org/10.1016/j.eswa.2008.02.021.

[11] Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. In *MIS Quarterly: Management Information Systems* (Vol. 35, Issue 3). https://doi.org/10.2307/23042796.

[12] Sumekar, W., & Al-Baarri, A. N. (2020). Study in Agroindustry of Salted Egg: Length of Salting Process and Marketing Reach Aspects. *Journal of Applied Food Technology*, *7*(1). https://doi.org/10.17728/jaft.7427.

[13] Teixeira, T. S., & Mendes, R. (2019). How to Improve Your Company's Net Promoter Score. *Harvard Business Review Digital Articles*, *October*.

[14] van der Borgh, M., Xu, J., & Sikkenk, M. (2020). Identifying, analyzing, and finding solutions to the sales lead black hole: A design science approach. *Industrial Marketing Management*, *88*. https://doi.org/10.1016/j.indmarman.2020.05.008.

[15] Wang, L., Zeng, Y., & Chen, T. (2015). Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. *Expert Systems with Applications*, *42*(2). https://doi.org/10.1016/j.eswa.2014.08.018.

[16] https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining.