REVA
UNIVERSITY
Bengaluru, India

A Project Report on

CRM-based Lead Scoring with Machine Learning

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration
In Business Analytics

Submitted By
Pradeep Thota
R19MBA63

Under the Guidance of
Phaneendra Akula
Senior Manager Data Science, Sunrise.

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

August, 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, **Pradeep Thota** hereby declare that I have completed the project work towards the second year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **CRM-based Lead Scoring with Machine Learning** under the supervision of **Phaneendra Akula, Senior Manager Data Science, Sunrise.** This report embodies the original work done by me in partial fulfilment of the requirements for the award of a degree for the academic year **2022.**

Place: Bengaluru
Date:27/08/2022

Name of the Student: Pradeep Thota
Signature of Student

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the project work entitled **CRM-based lead scoring with Machine Learning** carried out by **Pradeep Thota** with **R19MBA63,** is a bonafide student of REVA University, is submitting the second year project report in fulfilment for the award of **Master of Business Administration** in Business Analytics during the academic year 2021-2022. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said degree.

A. Phaneen

Signature of the Guide	Signature of the Director
Name of the Guide: Phaneendra Akula	Name of the Director: Dr. Shinu Abhi

External Viva

Names of the Examiners

1. Vaibhav Sahu, Strategic Cloud Engineer, Google
2. Abhishek Sinha, Data Science Manager, Capgemini

Place: Bengaluru

Date: 27/08/2022

REVA
UNIVERSITY
Bengaluru, India

Acknowledgment

I would like to thank our Hon’ble Chancellor, Dr. P Shayma Raju, Pro-Vice Chancellor, Dr. M. Dhanamjaya, Registrar, Dr. N Ramesh, and the RACE team, for supporting the RACE program specifically designed for working professionals and providing facilities and infrastructure required and conducive conditions to offer the best learning experience. I am very happy to be called a part of this program and REVA university.

Place: Bengaluru
Date:27/08/2022

Similarity Index Report

This is to certify that this project report titled **CRM-based Lead Scoring with Machine Learning** was scanned for similarity detection. Process and outcome is given below.

Software Used: Turnitin
Date of Report Generation: 26/08/2022
Similarity Index in %: 4%
Total word count: 3296
Name of the Guide: Akula Phaneendra

Place: Bengaluru
Date:27/08/2022

Name of the Student: Pradeep Thota
Signature of Student

Verified by: Dincy Dechamma

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	AI	Artificial Intelligence
2	DL	Deep Learning
3	ML	Machine Learning
4	XGB	XGBoost
5	RF	Random Forest
6	LGBM	LightGBM

List of Figures

List of Tables

Abstract

Betutelage is an educational course selling startup company with live classes targeting all levels of audiences 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 students like in which area they can improve their focus in studies etc.

Betutelage needs help in predicting the leads, these leads are the most paying customers of conversion from enquiry, now *Betutelage* needs 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. *Betutelage* leadership has also set a conversion rate target of around 70%.

In this project, going to build a classification model using CRISP-DM methodology with at least 4 classification models on the data provided by *Betutelage* using Machine Learning, Artificial Intelligence, or Deep learning to find the best model among those which have more accuracy of likely to convert as leads on both test and train data. After building the model can conclude Random Forest has good train and test accuracy with 94.3 and 92.02 respectively.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Classification Models, Leads, Random Forest.

Contents

Chapter 1: 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 to 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 Figure No.1 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* this project has built several classification models to get solutions for their problem.

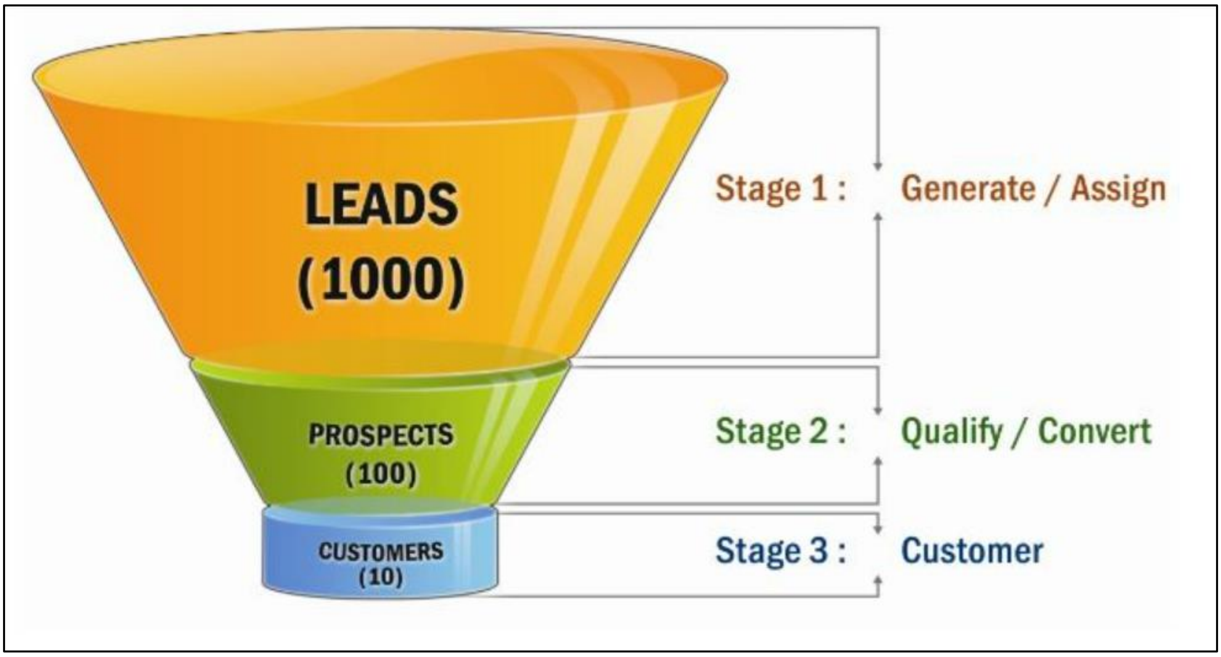


Figure No. 1: Leads conversion (Twitter).

Chapter 2: Literature Review

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 is a task for companies need to do on daily basis, even dealing with Small Data (Benhaddou & Leray, 2018). Now more interested in Lead Scoring here which refers to the practice of calculating and assigning a score to leads (business contacts or qualified prospects) of the company.

The fundamentals of the lead score are not only for the customer business but also for business-to-business matters and lead to multipliers for the market says (Sumekar & Al-Baarri, 2020). Not only running some campaigns but also calling over the telephone to a person and explaining the product is also will get the leads to the organization says (Brown & Brucker, 1987). It was clearly explained how to improve your net promotor score (Teixeira & Mendes, 2019).

So lead scoring can be increased when it is implemented with the classification models like Random forest, and logistic model (Wang et al., 2015), not only by applying this but also need to do some applications of data mining techniques in customer relationship management (Ngai et al., 2009) and dealing with imbalanced class distributions (Batista & Monard, 2002) and doing some statistical learnings from the data (Hastie et al., 2009).

After this need to deal with some missing values before building a model as they will surely impact our model (Liu et al., 2016) and need to balance the imbalanced data else the impact will be huge on the model (Luque et al., 2019) once all the necessary data preparation steps are done then can head to our model building that is Predictive analytics classification model (Shmueli & Koppius, 2011) and this model can build on any classification models and artificial intelligence classification model says (Dreiseitl & Ohno-Machado, 2002) and this can also deal with big data (McAfee & Brynjolfsson, 2012).

Once the model is built then it's very important to evaluate the model and to know the ROC (Carter et al., 2016), and metrics of the model so now will get a Lead Prioritization and Scoring model with the path to higher conversion (van der Borgh et al., 2020).

Chapter 3: Problem Statement

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, *Betutelage* needs a model were 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.

Chapter 4: Objectives of the Study

The objective of this project is to develop a Classification model for the below points:

1) Assisting the business know the leads who can convert to their paying customers so the business needs a model that can predict accuracy about that customer.

2) By the above-built model they can invest more time on non-converting leads to make them convert etc to them to get their course sold not only that they can also reduce the cost of their campaigning where there are no leads who are not turning in to their paying customers.

3) 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 the project as having a good sample of data provided by the business that has collected the dump from their server, etc., but needs to do some data preparation on top of it but the crucial part is to build a good model with great training and testing metrics of good accuracy.

Chapter 5: Project Methodology

CRISP-DM is the methodology used in this project. It involves six steps which are captured in Figure No. 5.

`

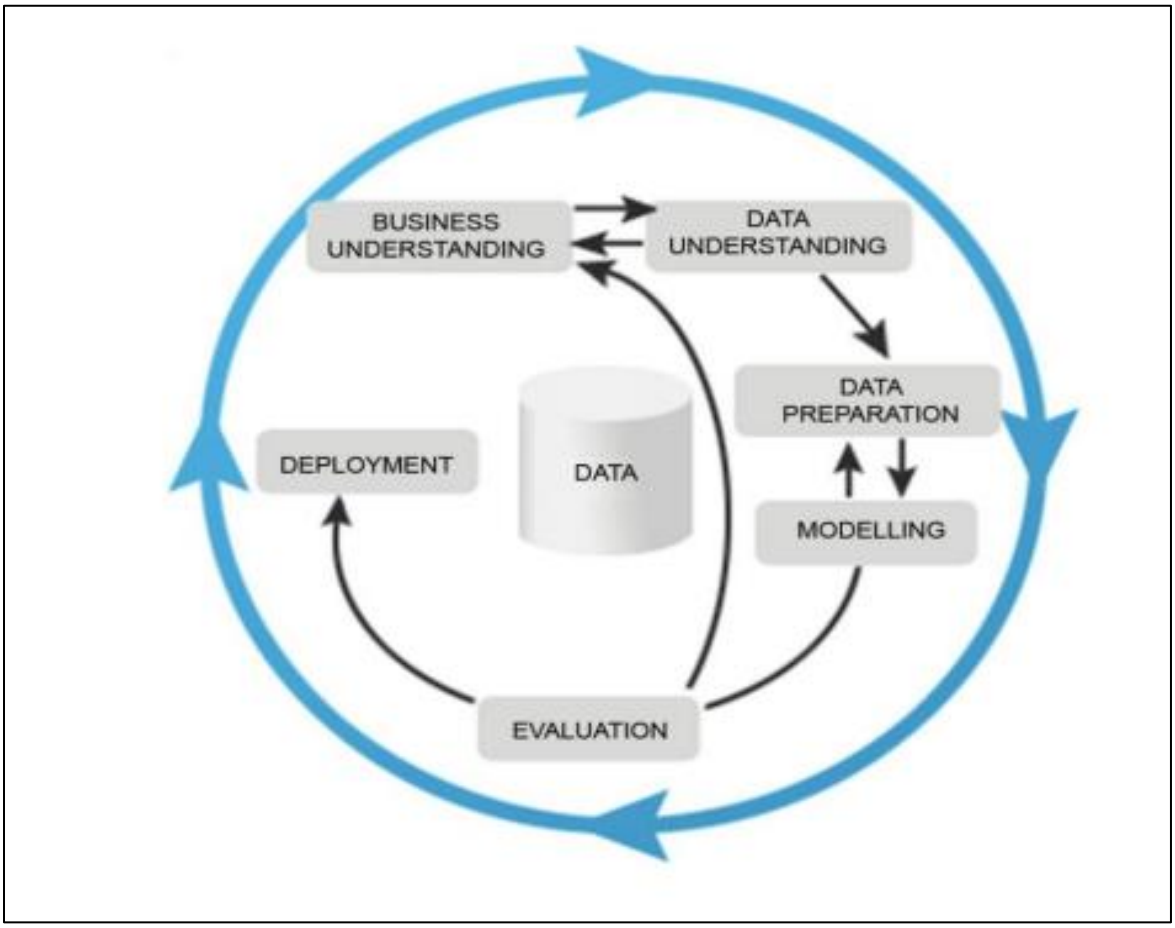


Figure No. 5: CRISP-DM High-Level Steps (Wikipedia).

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.

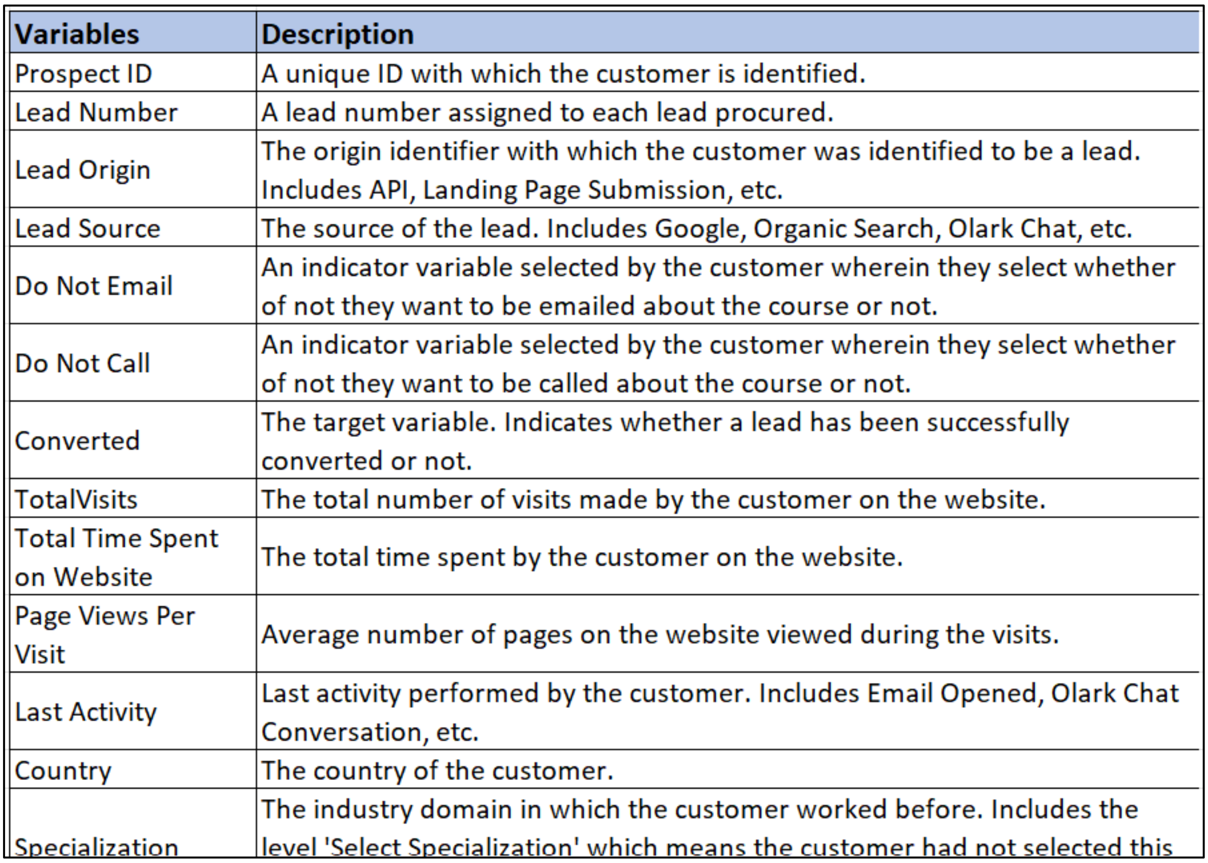
Deployment - The goal of this stage is to determine the model deployment strategy based on evaluation results and a plan for monitoring and maintenance of models in the business environment.

Chapter 6: Business Understanding

As the part of business understanding, this project has a very 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, so now business can concentrate on this low lead score customers to make them as their paying customers applying appropriate strategies.

Chapter 7: Data Understanding

Data comprises structured data which is eligible for the Classification model and it is in the CSV file format. Data is collected from the company-maintained CSV file format and maintained manually. Table No.7.1 shows the legend of the data for more understanding.



Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this

Table No. 7.1: Data Variables and their description.

List of Visualization data understanding 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.

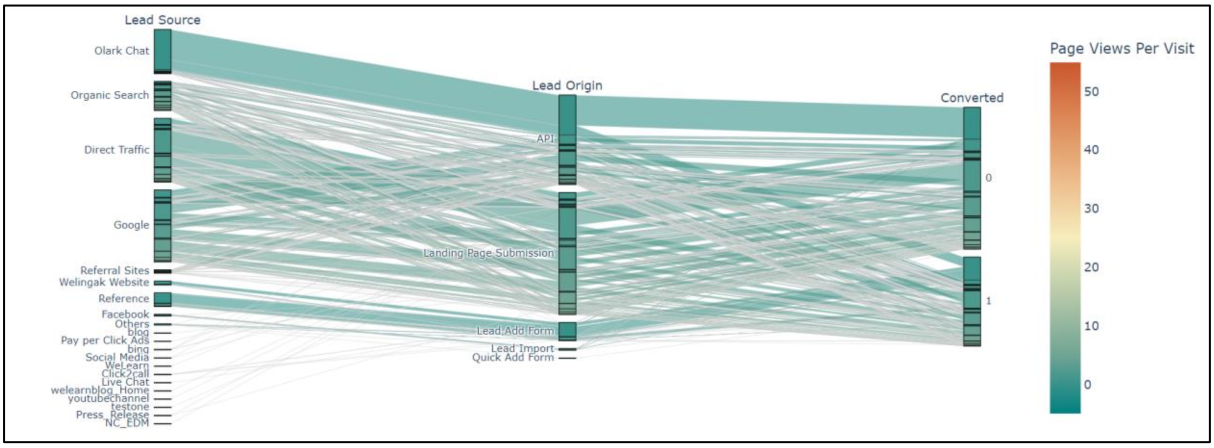


Figure No. 7.2: Leads for converted and non-converted with data provided as it is.

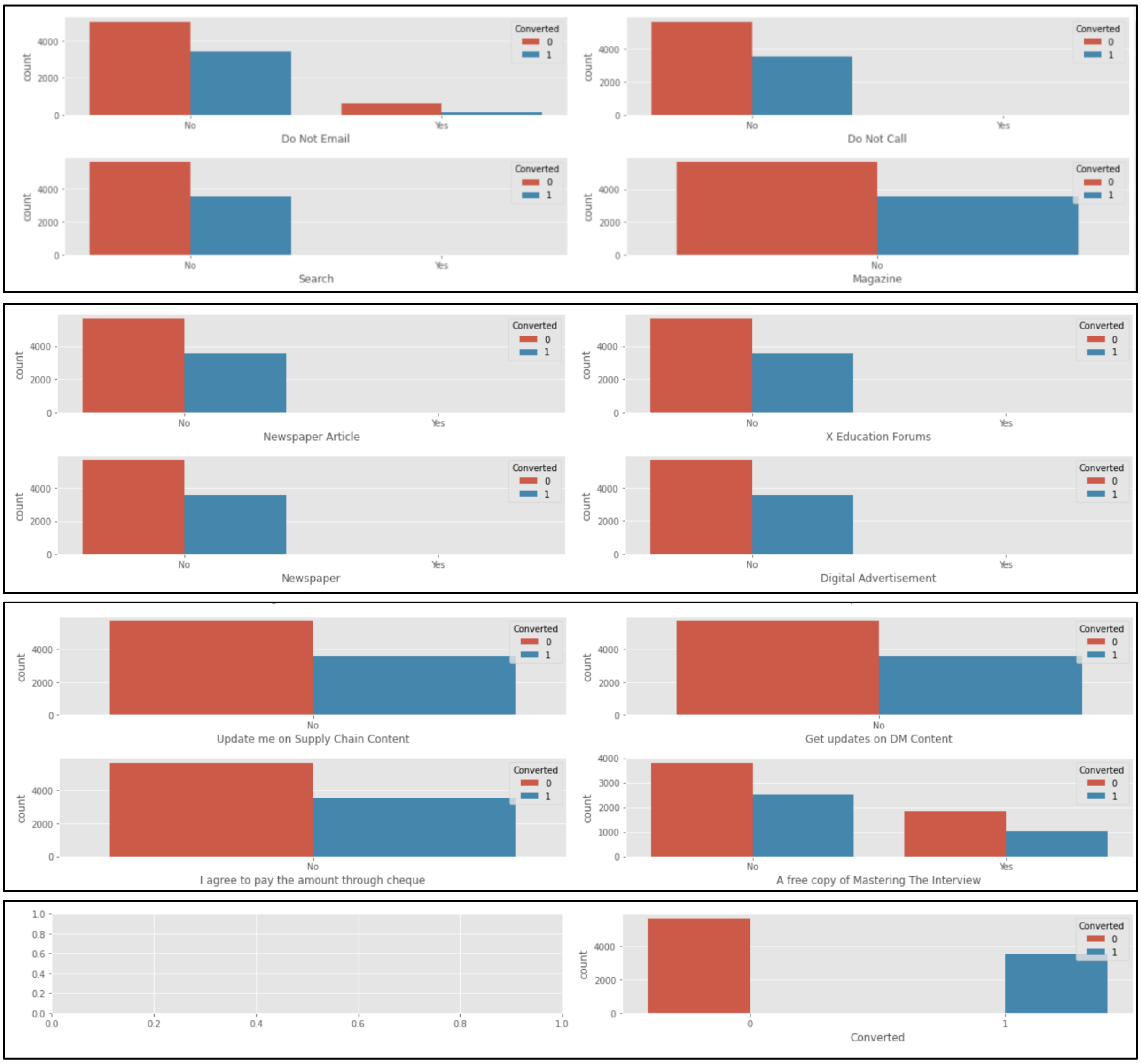


Figure No. 7.3: Lead conversion for individual variables

Based on Figures No. 7.2 and Figure No. 7.3, the target variable is having a 61.5:38.5 ratio, in the classification model, with 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 because who does not like freebies? The reason may be has a large proportion of the audience is "Unemployed" and the only thing they are interested about upskilling themselves and not giving priority to the interview preparation in the early stage. Also, there are certain columns from which are not going to infer much information as most of the values is "No" so will be going to drop the same in the later stage.

Chapter 8: Data Preparation

After finishing 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 seem not to have any relevance

- 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.

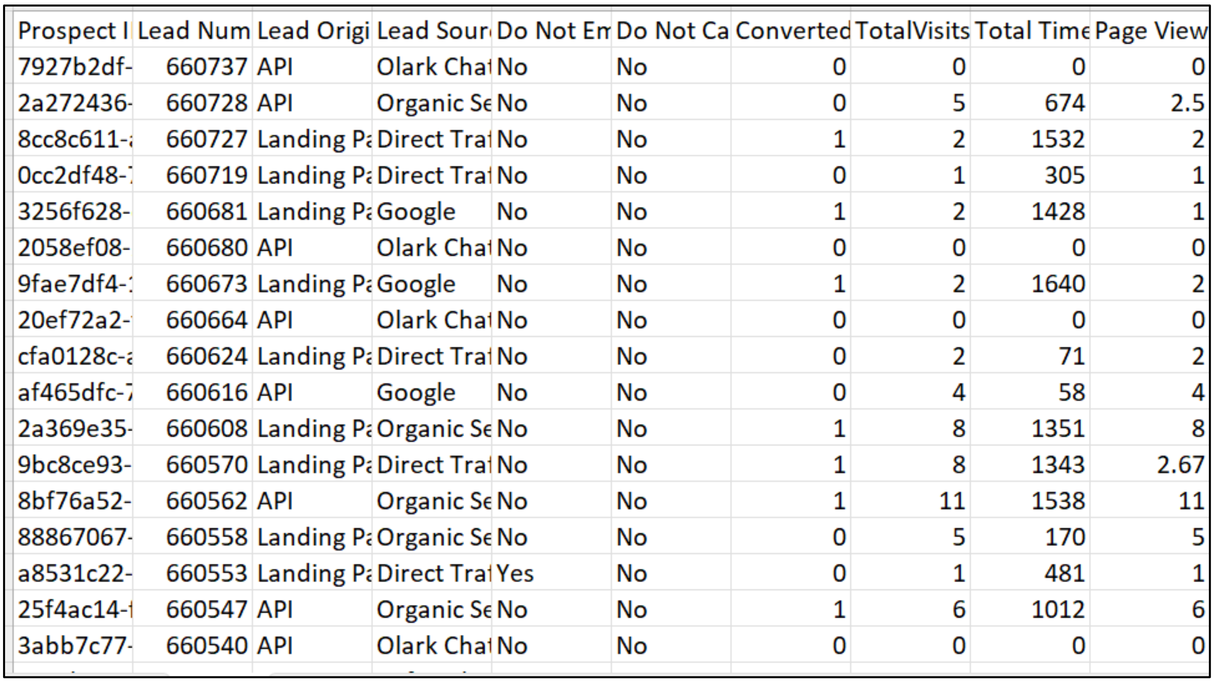
- Based on the different variables from the data which tell about the preferences and background of the people being approached as potential leads for business, try to first analyze the variables that seem to cause high conversion rates and also identify any correlations or patterns between the variables 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.

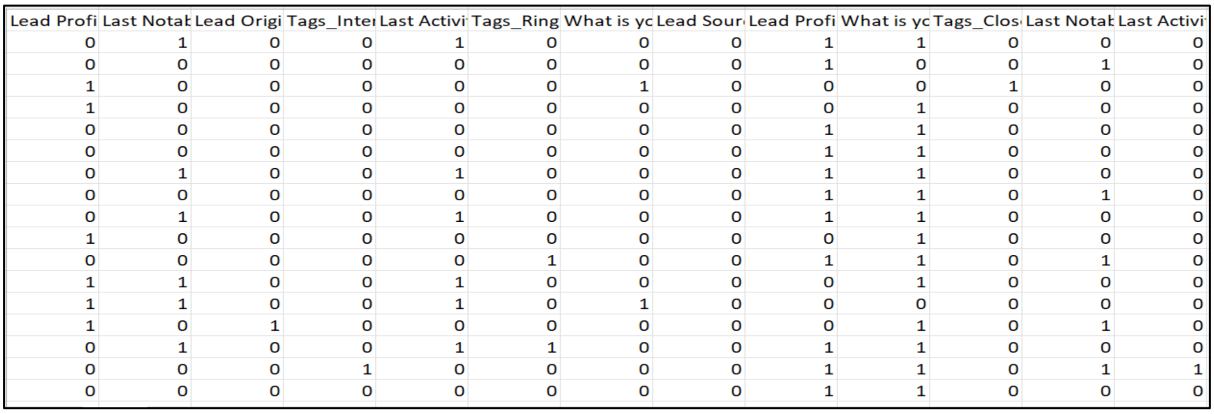
- Lastly, find out the top variables that impact the lead conversion and summarize them so that it enables the Client Sales Team to identify the potential customers.

As the client has given a good sample of data Table No. 8.1 requires minimal preparation, after necessary modifications fed the corpus as it is and once the data is prepared it looks like Table No. 8.2.



Prospect I	Lead Num	Lead Origi	Lead Sour	Do Not Em	Do Not Ca	Converted	TotalVisits	Total Time	Page View
7927b2df-	660737	API	Olark Chat	No	No	0	0	0	0
2a272436-	660728	API	Organic Se	No	No	0	5	674	2.5
8cc8c611-	660727	Landing Pa	Direct Tra	No	No	1	2	1532	2
0cc2df48-	660719	Landing Pa	Direct Tra	No	No	0	1	305	1
3256f628-	660681	Landing Pa	Google	No	No	1	2	1428	1
2058ef08-	660680	API	Olark Chat	No	No	0	0	0	0
9fae7df4-	660673	Landing Pa	Google	No	No	1	2	1640	2
20ef72a2-	660664	API	Olark Chat	No	No	0	0	0	0
cfa0128c-	660624	Landing Pa	Direct Tra	No	No	0	2	71	2
af465dfc-	660616	API	Google	No	No	0	4	58	4
2a369e35-	660608	Landing Pa	Organic Se	No	No	1	8	1351	8
9bc8ce93-	660570	Landing Pa	Direct Tra	No	No	1	8	1343	2.67
8bf76a52-	660562	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 Tra	Yes	No	0	1	481	1
25f4ac14-	660547	API	Organic Se	No	No	1	6	1012	6
3abb7c77-	660540	API	Olark Chat	No	No	0	0	0	0

Table No. 8.1: Raw Data Corpus



Lead Profi	Last Notab	Lead Origi	Tags_Inte	Last Activi	Tags_Ring	What is yo	Lead Sour	Lead Profi	What is yo	Tags_Clos	Last Notab	Last Activi
0	1	0	0	1	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0	0
0	1	0	0	1	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	1	0
0	1	0	0	1	0	0	0	1	1	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	1	1	0	1	0
1	1	0	0	1	0	0	0	0	1	0	0	0
1	1	0	0	1	0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	1	0	1	0
0	1	0	0	1	1	0	0	1	1	0	0	0
0	0	0	1	0	0	0	0	1	1	0	1	1
0	0	0	0	0	0	0	0	1	1	0	0	0

Table No. 8.2: Data after necessary cleanup activities

Chapter 9: Modeling

Data modeling and flow for this project are as follows in Figure No. 9.

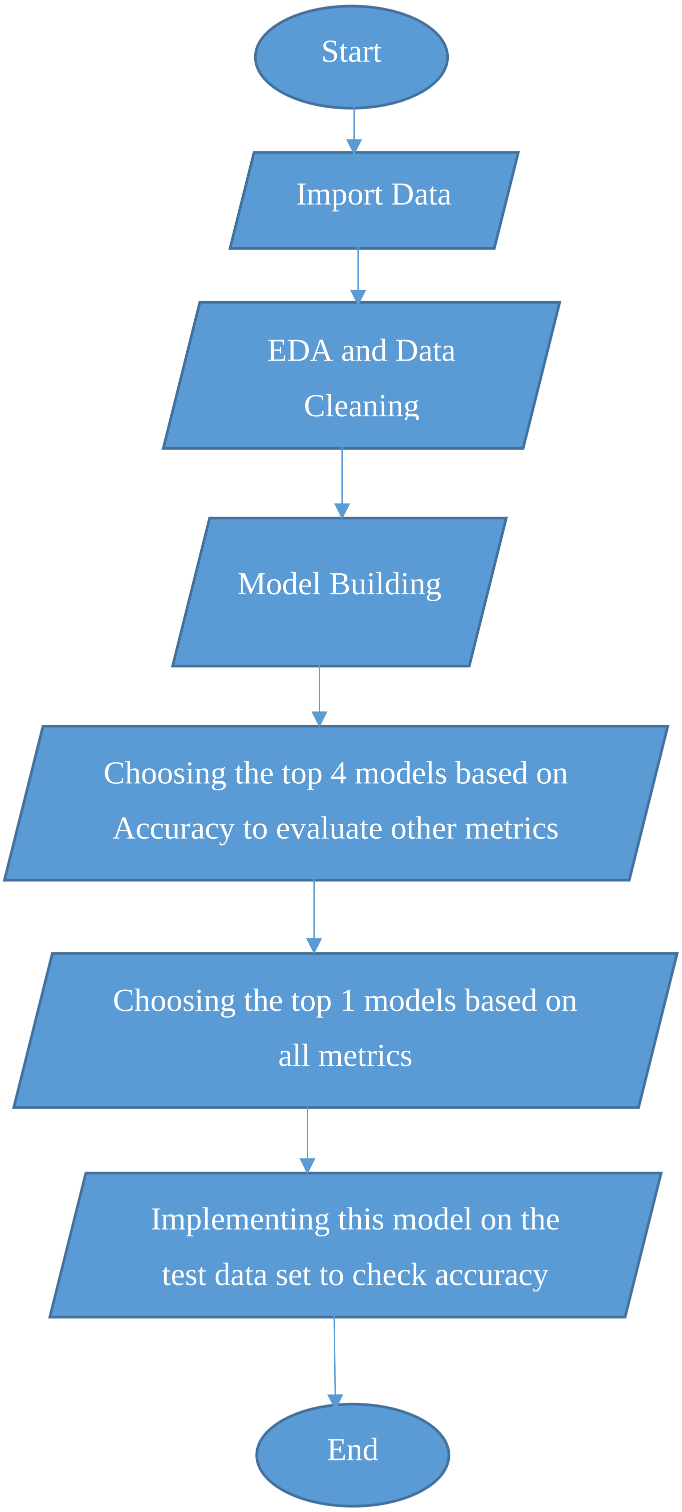

Start
Import Data
EDA and Data Cleaning
Model Building
Choosing the top 4 models based on Accuracy to evaluate other metrics
Choosing the top 1 models based on all metrics
Implementing this model on the test data set to check accuracy
End

Figure No. 9: 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 step is done then built the model based on 12 different classification methods **RandomForest, Adaboost, ExtraTree, BaggingClassifier, GradientBoosting, DecisionTree, KNN, Logistic, SGD Classifier, MLPClassifier, NaiveBayes, LightGBM, Catboost** and will finalize based on their metrics for final testing on test data.

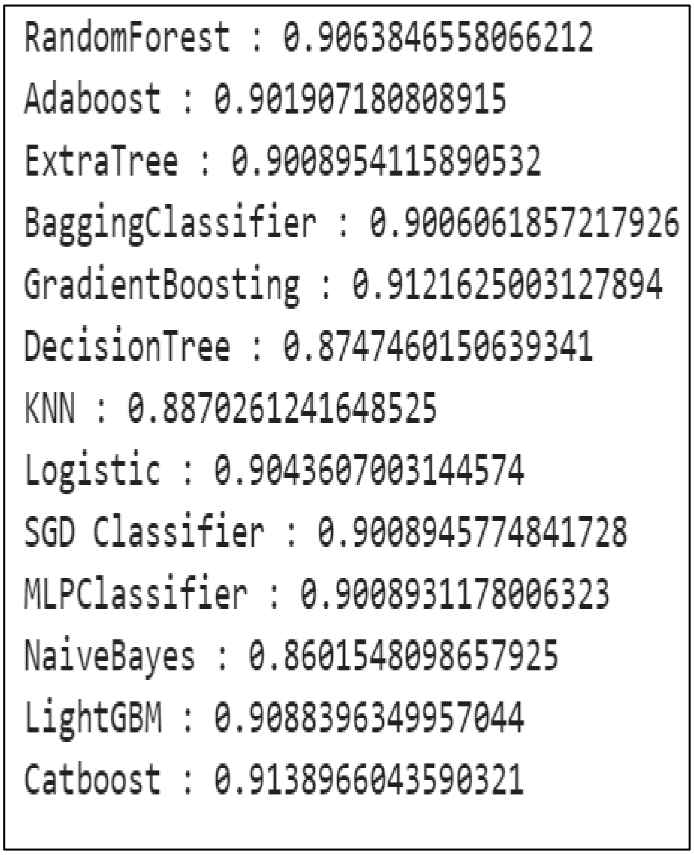
After building models on several classifiers only **RandomForest Classifier, GradientBoosting, LightGBM & Catboost** classifiers have been chosen for the next level based on top accuracy for checking other metrics like precision, recall, f1 score, and others.

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

Chapter 10: Model Evaluation

Model Evaluation for this project is as follows:

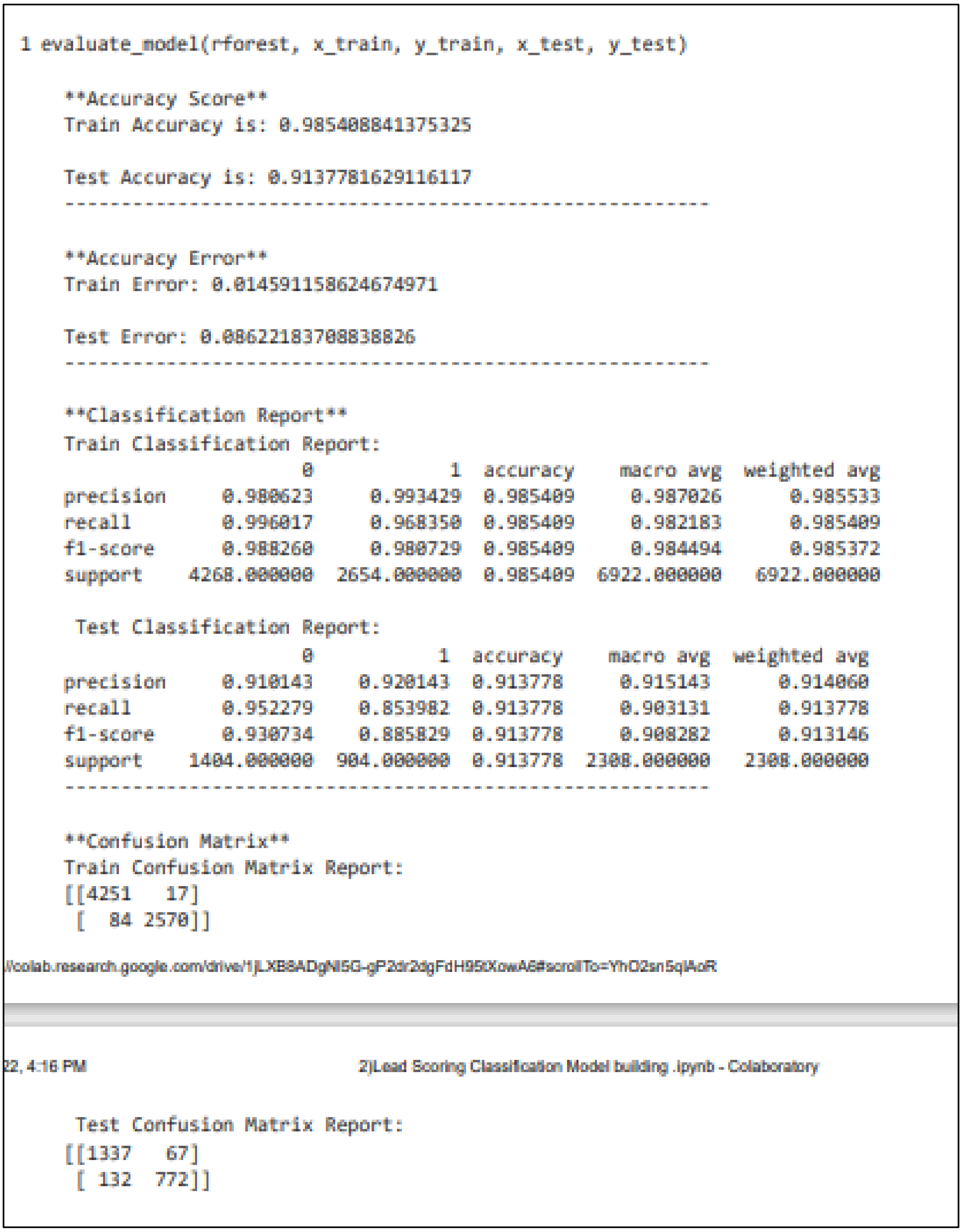
Information Based on the Classification Models:



```
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
```

Figure No. 10.1: Accuracy of different models

Based on the results shown in Figure No 10.1, one can choose the RandomForest Classifier, GradientBoosting, LightGBM & Catboost and shown in Table No. 10 which 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.



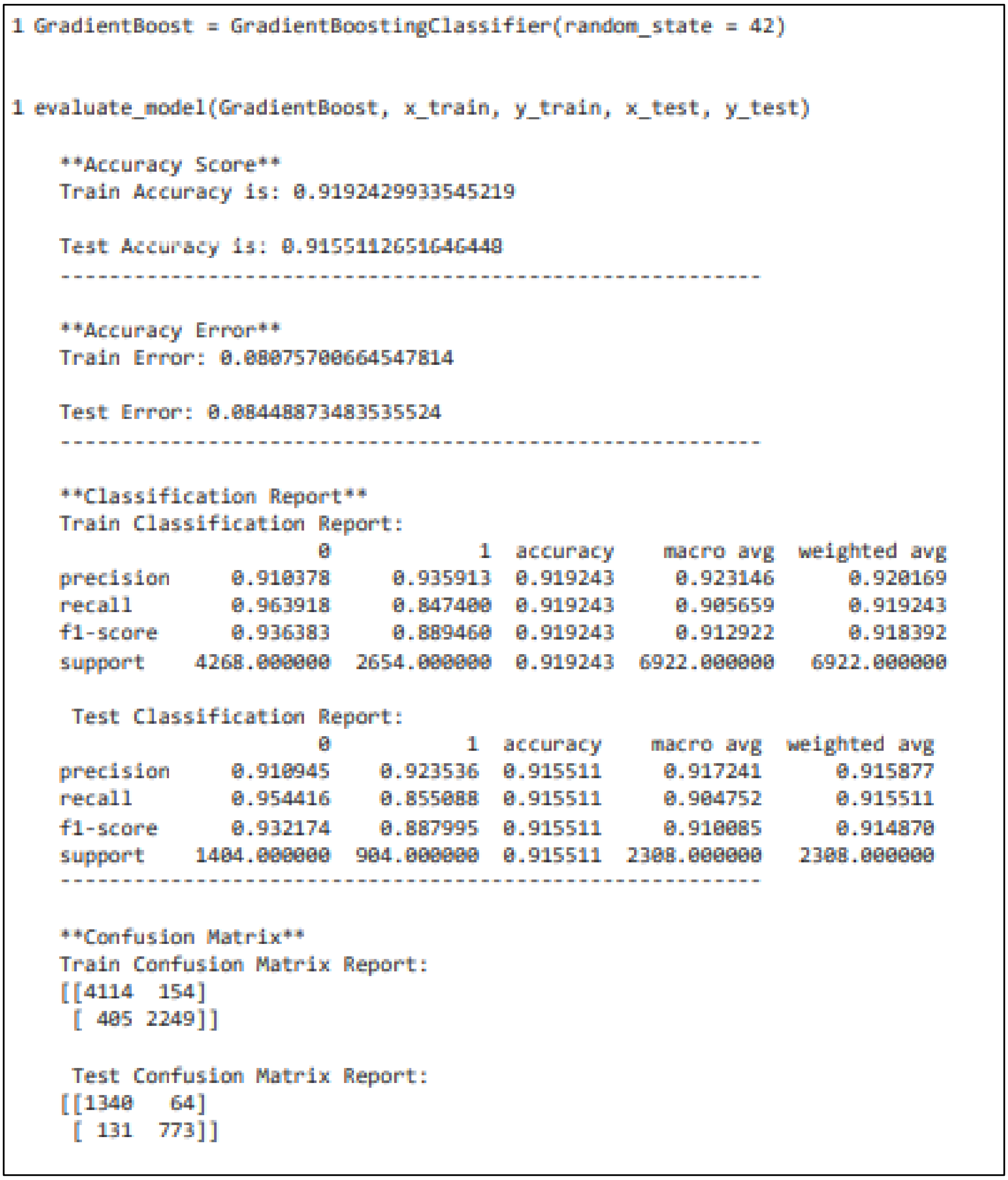
```
evaluate_model(rforest, x_train, y_train, x_test, y_test)

    **Accuracy Score**
    Train Accuracy is: 0.985408841375325

    Test Accuracy is: 0.9137781629116117
    ----------------------------------------------------------

    **Accuracy Error**
    Train Error: 0.014591158624674971

    Test Error: 0.08622183708838826
    ----------------------------------------------------------

    **Classification Report**
    Train Classification Report:
                          0            1  accuracy    macro avg  weighted avg
    precision     0.980623     0.993429  0.985409     0.987026      0.985533
    recall        0.996017     0.968350  0.985409     0.982183      0.985409
    f1-score      0.988260     0.980729  0.985409     0.984494      0.985372
    support    4268.000000  2654.000000  0.985409  6922.000000   6922.000000

     Test Classification Report:
                          0           1  accuracy    macro avg  weighted avg
    precision     0.910143    0.920143  0.913778     0.915143      0.914060
    recall        0.952279    0.853982  0.913778     0.903131      0.913778
    f1-score      0.930734    0.885829  0.913778     0.908282      0.913146
    support    1404.000000  904.000000  0.913778  2308.000000   2308.000000
    ----------------------------------------------------------

    **Confusion Matrix**
    Train Confusion Matrix Report:
    [[4251   17]
     [  84 2570]]
```

//colab.research.google.com/drive/1jLXB8ADgNI5G-gP2dr2dgFdH95tXowA6#scrollTo=YhO2sn5qlAoR

22, 4:16 PM 2)Lead Scoring Classification Model building .ipynb - Colaboratory

```
     Test Confusion Matrix Report:
    [[1337   67]
     [ 132  772]]
```

Figure No. 10.2: Metrics of Random Forest

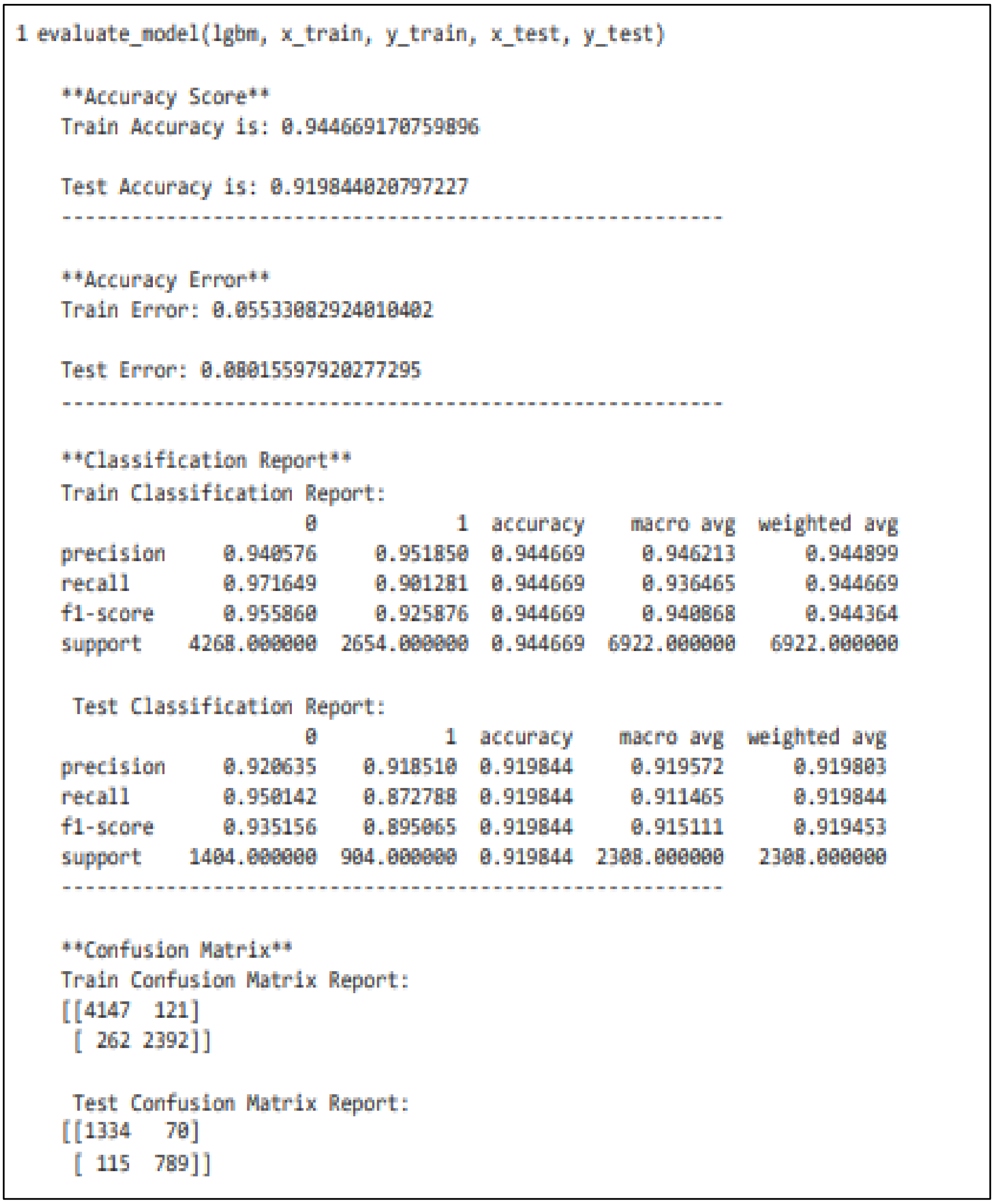


```
GradientBoost = GradientBoostingClassifier(random_state = 42)
```

```
evaluate_model(GradientBoost, x_train, y_train, x_test, y_test)
```

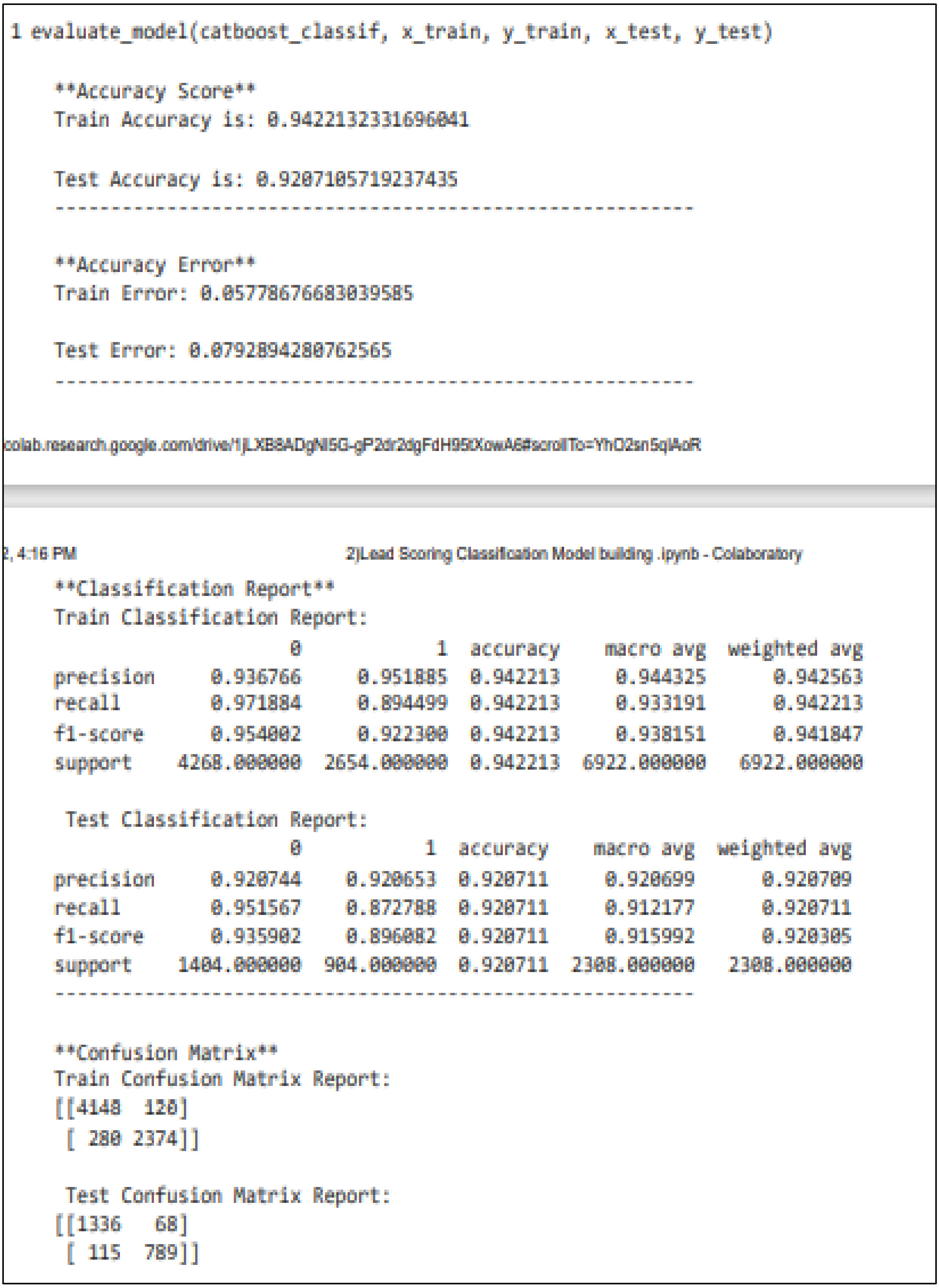
```
**Accuracy Score**
Train Accuracy is: 0.9192429933545219

Test Accuracy is: 0.9155112651646448
---------------------------------------------------------

**Accuracy Error**
Train Error: 0.08075700664547814

Test Error: 0.08448873483535524
---------------------------------------------------------

**Classification Report**
Train Classification Report:
                     0            1  accuracy    macro avg  weighted avg
precision     0.910378     0.935913  0.919243     0.923146      0.920169
recall        0.963918     0.847400  0.919243     0.905659      0.919243
f1-score      0.936383     0.889460  0.919243     0.912922      0.918392
support    4268.000000  2654.000000  0.919243  6922.000000   6922.000000

 Test Classification Report:
                     0           1  accuracy    macro avg  weighted avg
precision     0.910945    0.923536  0.915511     0.917241      0.915877
recall        0.954416    0.855088  0.915511     0.904752      0.915511
f1-score      0.932174    0.887995  0.915511     0.910085      0.914870
support    1404.000000  904.000000  0.915511  2308.000000   2308.000000
---------------------------------------------------------

**Confusion Matrix**
Train Confusion Matrix Report:
[[4114  154]
 [ 405 2249]]

 Test Confusion Matrix Report:
[[1340   64]
 [ 131  773]]
```

Figure No. 10.3: Metrics of Gradient Boost



```
evaluate_model(lgbm, x_train, y_train, x_test, y_test)

    **Accuracy Score**
    Train Accuracy is: 0.944669170759896

    Test Accuracy is: 0.919844020797227
    ---------------------------------------------------------

    **Accuracy Error**
    Train Error: 0.05533082924010402

    Test Error: 0.08015597920277295
    ---------------------------------------------------------

    **Classification Report**
    Train Classification Report:
                         0            1  accuracy    macro avg  weighted avg
    precision     0.940576     0.951850  0.944669     0.946213      0.944899
    recall        0.971649     0.901281  0.944669     0.936465      0.944669
    f1-score      0.955860     0.925876  0.944669     0.940868      0.944364
    support    4268.000000  2654.000000  0.944669  6922.000000   6922.000000

     Test Classification Report:
                         0           1  accuracy    macro avg  weighted avg
    precision     0.920635    0.918510  0.919844     0.919572      0.919803
    recall        0.950142    0.872788  0.919844     0.911465      0.919844
    f1-score      0.935156    0.895065  0.919844     0.915111      0.919453
    support    1404.000000  904.000000  0.919844  2308.000000   2308.000000
    ---------------------------------------------------------

    **Confusion Matrix**
    Train Confusion Matrix Report:
    [[4147  121]
     [ 262 2392]]

     Test Confusion Matrix Report:
    [[1334   70]
     [ 115  789]]
```

Figure No. 10.4: Metrics of LightGBM



```
1 evaluate_model(catboost_classif, x_train, y_train, x_test, y_test)

    **Accuracy Score**
    Train Accuracy is: 0.9422132331696041

    Test Accuracy is: 0.9207105719237435
    ---------------------------------------------------------

    **Accuracy Error**
    Train Error: 0.05778676683039585

    Test Error: 0.0792894280762565
    ---------------------------------------------------------
```

colab.research.google.com/drive/1jLXB8ADgNI5G-gP2dr2dgFdH95tXowA6#scrollTo=YhO2sn5qlAoR

2, 4:16 PM 2)Lead Scoring Classification Model building .ipynb - Colaboratory

```
    **Classification Report**
    Train Classification Report:
                          0            1  accuracy    macro avg  weighted avg
    precision     0.936766     0.951885  0.942213     0.944325      0.942563
    recall        0.971884     0.894499  0.942213     0.933191      0.942213
    f1-score      0.954002     0.922300  0.942213     0.938151      0.941847
    support    4268.000000  2654.000000  0.942213  6922.000000   6922.000000

     Test Classification Report:
                          0           1  accuracy    macro avg  weighted avg
    precision     0.920744    0.920653  0.920711     0.920699      0.920709
    recall        0.951567    0.872788  0.920711     0.912177      0.920711
    f1-score      0.935902    0.896082  0.920711     0.915992      0.920305
    support    1404.000000  904.000000  0.920711  2308.000000   2308.000000
    ---------------------------------------------------------

    **Confusion Matrix**
    Train Confusion Matrix Report:
    [[4148  120]
     [ 280 2374]]

     Test Confusion Matrix Report:
    [[1336   68]
     [ 115  789]]
```

Figure No. 10.5: Metrics of Catboost

Model	Train Precision	F1-Score	Recall	Train Accuracy	Test Accuracy
Random Forest	98.06	98.8	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%

Table No. 10: Model Metrics

A) Model Accuracy:

1) Random Forest:

When it involves training accuracy Figure No. 10.2, Random Forest has the 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% Figure No. 10.3 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% Figure No. 10.4, 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% Figure No. 10.5. In the Catboost set of rules, have the best look at accuracy compared to Random Forest, Gradient Boosting, and LightGBM.

B) 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%.

C) 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:

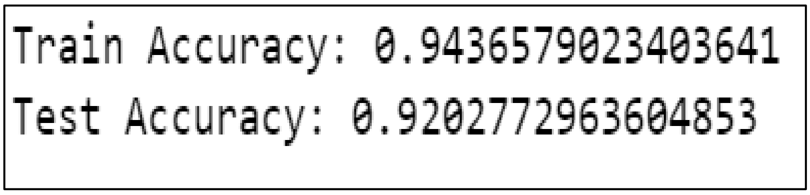
On the training dataset for the class label "0," our F1 rating is popping out to be 93.63% while for the class label "1" it's miles coming as 88.94%. For checking out the test dataset, the F1-rating for the class label "0" has been decreased to 93.21% while for the class label "1" it's miles 88.79%.

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%.



```
Train Accuracy: 0.9436579023403641
Test Accuracy: 0.9202772963604853
```

Figure No. 10.6: Random Forest Metrics on a test data set

Finally, when implementing the learnings to the test model and calculating the conversion probability based on the Sensitivity metric & cutting off and found the train accuracy value to be 94.36%, the test accuracy was 92.02% as per Figure No. 10.6.

Some of the key drivers for lead conversion of the project 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

Chapter 11: 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.

Chapter 12: 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.

Chapter 13: Conclusions and Recommendations for future work

After this project, 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.

Bibliography

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*.

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

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

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

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

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

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

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

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

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

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

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

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

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

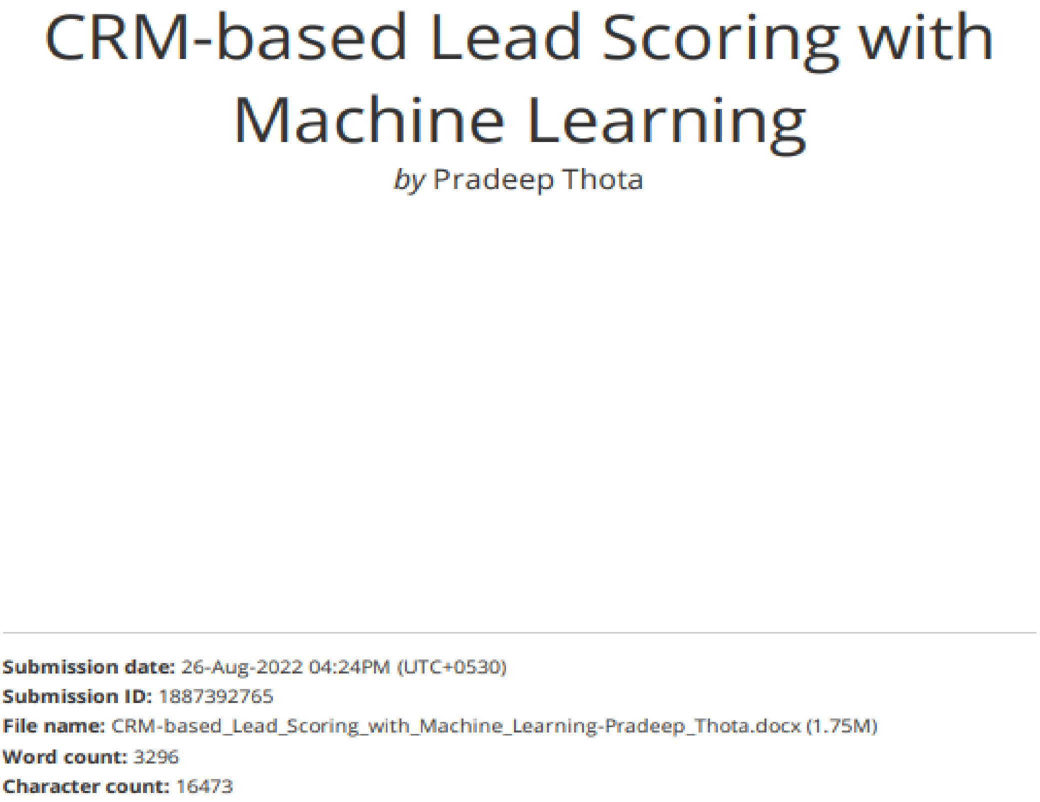
https://twitter.com/due/status/869257062701834240?lang=bg

https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

https://betutelage.com/

Appendix

Plagiarism Report[1]

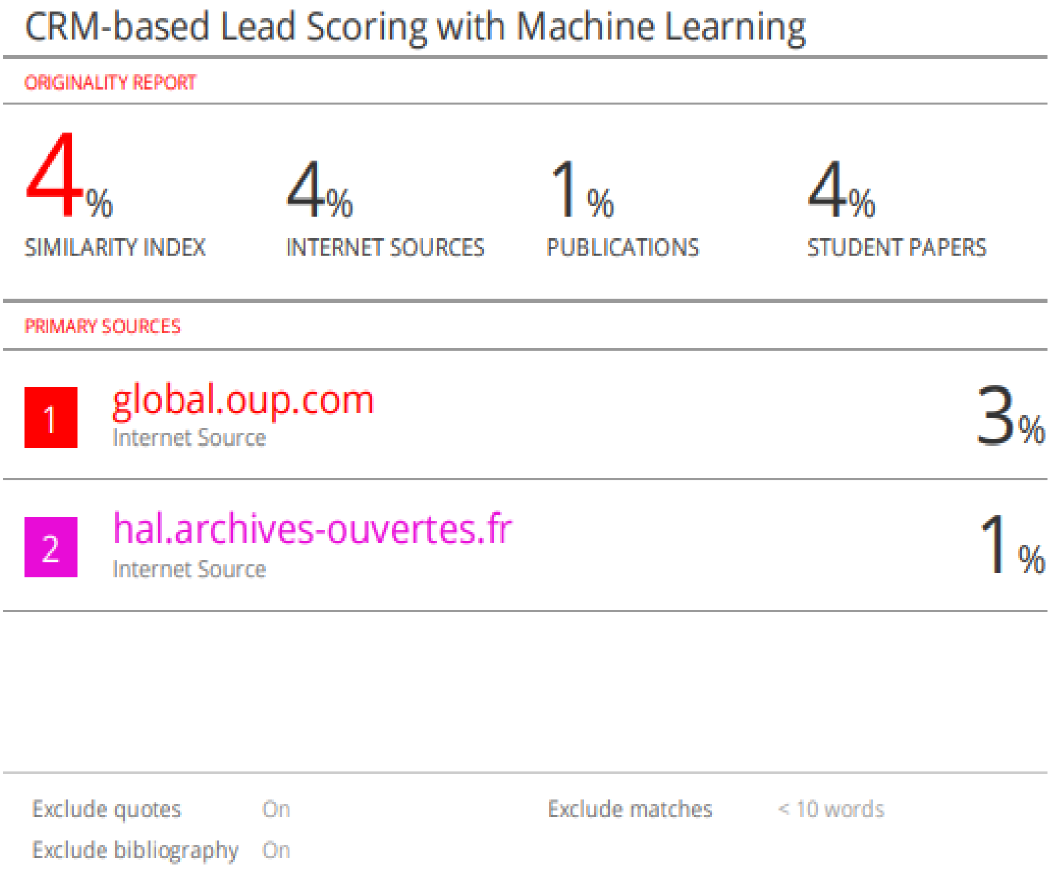


CRM-based Lead Scoring with Machine Learning

by Pradeep Thota

Submission date: 26-Aug-2022 04:24PM (UTC+0530)
Submission ID: 1887392765
File name: CRM-based_Lead_Scoring_with_Machine_Learning-Pradeep_Thota.docx (1.75M)
Word count: 3296
Character count: 16473

[1] Turnitn report to be attached from the University.



CRM-based Lead Scoring with Machine Learning

ORIGINALITY REPORT

4%	4%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

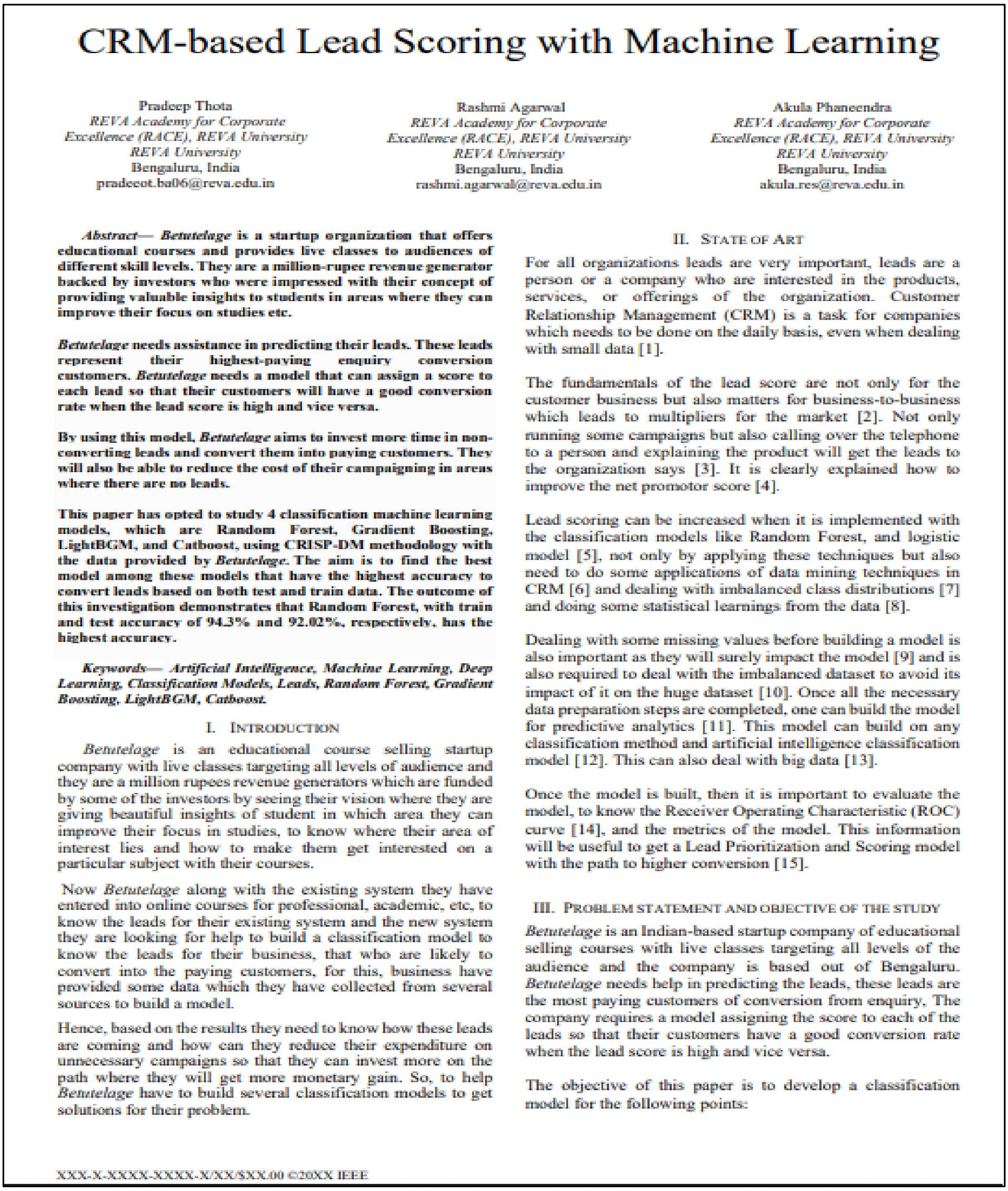
1	global.oup.com Internet Source	3%
2	hal.archives-ouvertes.fr Internet Source	1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Publications in a Journal/Conference Presented/White Paper[2]

Paper Submitted:

Pradeep Thota, Rashmi Agarwal, Akula Phaneendra, “CRM-based Lead Scoring with Machine Learning” 9th International Conference on Business Analytics and Intelligence, IIMB. Submission Date: 21st October 2022. paper ID: 2234



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:

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

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.

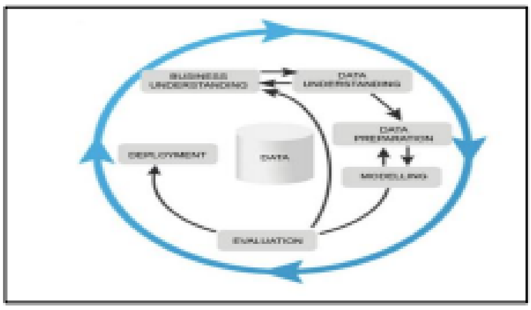


Fig. 1: CRISP-DM [16].

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.

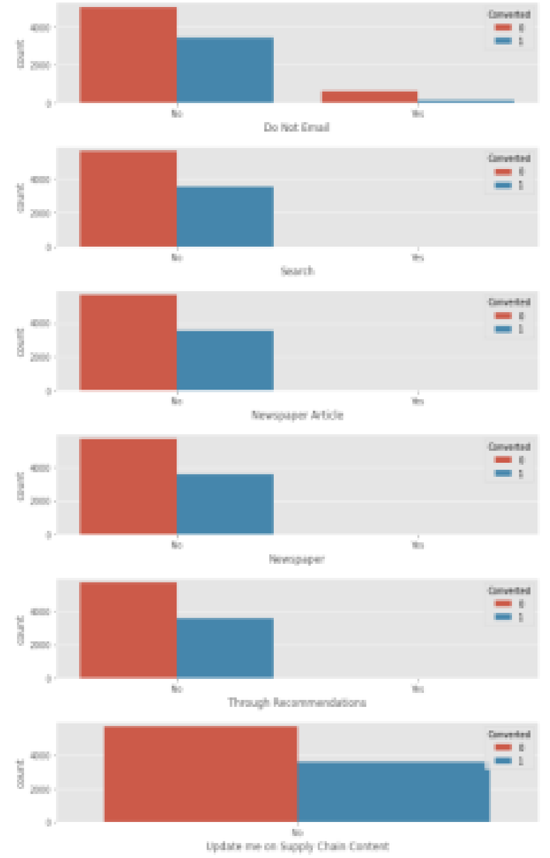
V. Business and Data Understanding

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.

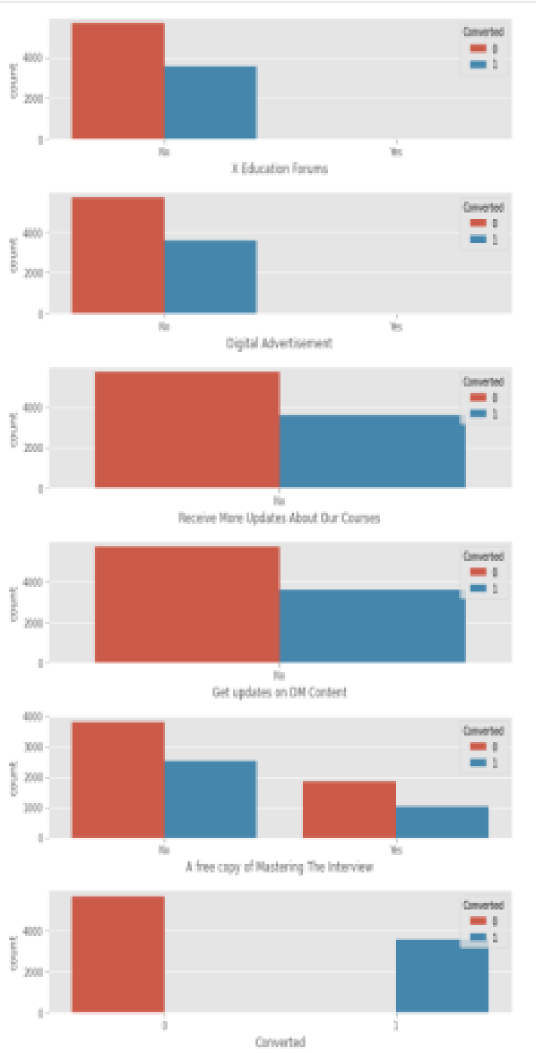
So, the business can conclude that customer who has the highest lead score will be having high conversion chances, and the customer who has the lowest lead score will be having low conversion chances.

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.

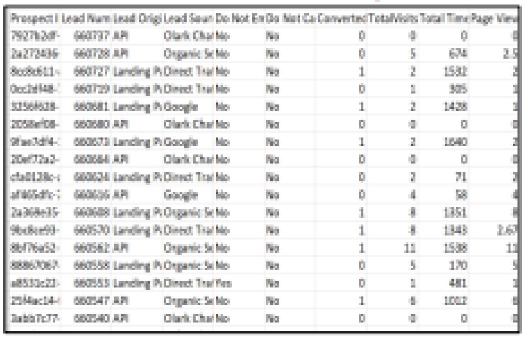
Also, there are certain columns, which are not going to infer much information as most of the values are "No", hence those will be dropped in the later stage.

VI. Data Preparation

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

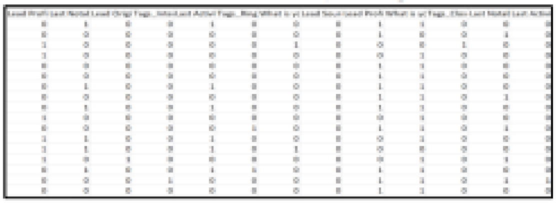
- The data available with us qualifies for the classification model and can apply the same to see if a lead converts into a customer or not.
- Firstly, clean the data to improve its quality by eliminating variables that are not relevant.
- Combine low-frequency categories into a new category to compress 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 Chat	No	No	0	0	0	0
2a272436	660728	API	Organic Se	No	No	0	5	674	2.5
8cc8c611-	660727	Landing P	Direct Traf	No	No	1	2	1532	2
0cc2df48-	660719	Landing P	Direct Traf	No	No	0	1	305	1
3256f628-	660681	Landing P	Google	No	No	1	2	1428	1
2058ef08-	660680	API	Olark Chat	No	No	0	0	0	0
9fae7df4-	660673	Landing P	Google	No	No	1	2	1640	2
20ef72a2-	660664	API	Olark Chat	No	No	0	0	0	0
cfa0128c-	660624	Landing P	Direct Traf	No	No	0	2	71	2
af465dfc-	660616	API	Google	No	No	0	4	58	4
2a369e35-	660608	Landing P	Organic Se	No	No	1	8	1351	8
9bc8ce93-	660570	Landing P	Direct Traf	No	No	1	8	1343	2.67
8bf76a52-	660562	API	Organic Se	No	No	1	11	1538	11
88867067-	660558	Landing P	Organic Se	No	No	0	5	170	5
a8533c22-	660553	Landing P	Direct Traf	Yes	No	0	1	481	1
2594ac14-	660547	API	Organic Se	No	No	1	6	1012	6
3a8b7c77-	660540	API	Olark Chat	No	No	0	0	0	0

Table No. 2: Data after necessary cleanup activities



Lead Profi	Last Notab	Lead Origi	Tags_Inter	Last Activi	Tags_Ring	What is yo	Lead Sour	Lead Profi	What is yo	Tags_Clos	Last Notab	Last Activ
0	1	0	0	1	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0	0
0	1	0	0	1	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	1	0
0	1	0	0	1	0	0	0	1	1	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	1	1	0	1	0
1	1	0	0	1	0	0	0	0	1	0	0	0
1	1	0	0	1	0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	1	0	1	0
0	1	0	0	1	1	0	0	1	1	0	0	0
0	0	0	1	0	0	0	0	1	1	0	1	1
0	0	0	0	0	0	0	0	1	1	0	0	0

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.

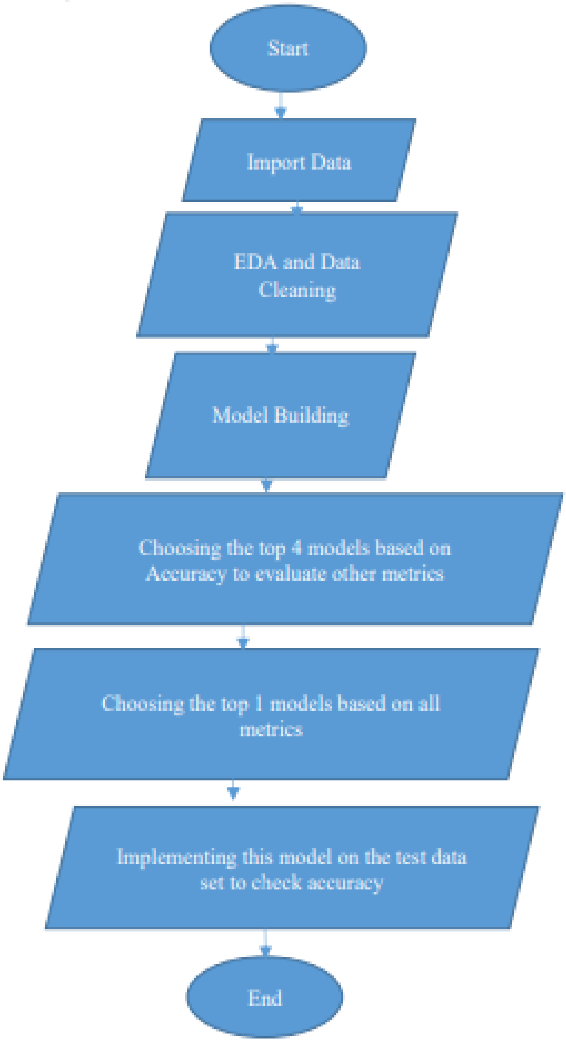

Start
Import Data
EDA and Data Cleaning
Model Building
Choosing the top 4 models based on Accuracy to evaluate other metrics
Choosing the top 1 models based on all metrics
Implementing this model on the test data set to check accuracy
End

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.

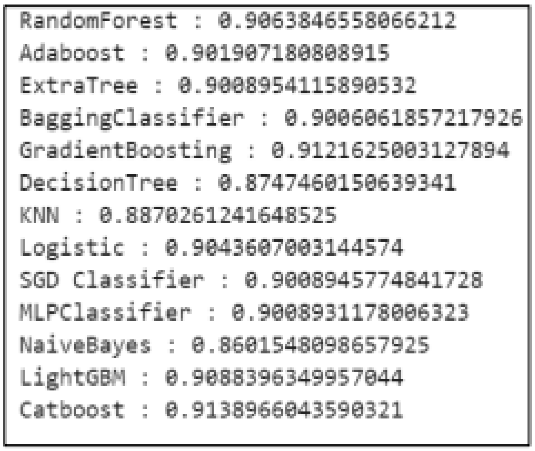
As the data preparation is done then building the models based on 12 different classification methods RandomForest, Adaboost, Extra Tree, Bagging Classifier, Gradient Boosting, Decision Tree, K Nearest Neighborhood (KNN), Logistic, Stochastic Gradient Decent (SGD) Classifier, Multi-layer perceptron (MLP) 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	98.06	98.8	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%

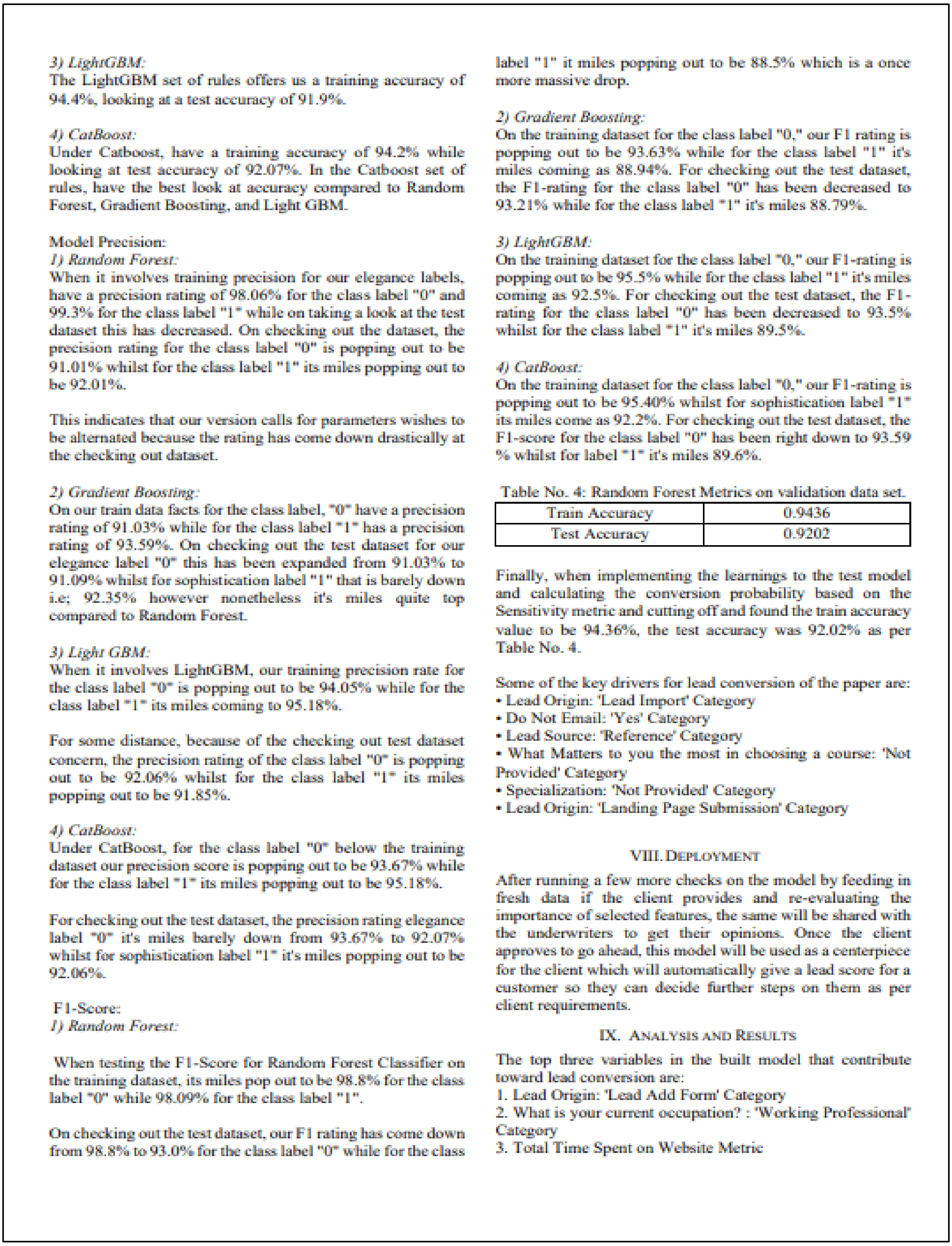
Model Accuracy:

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

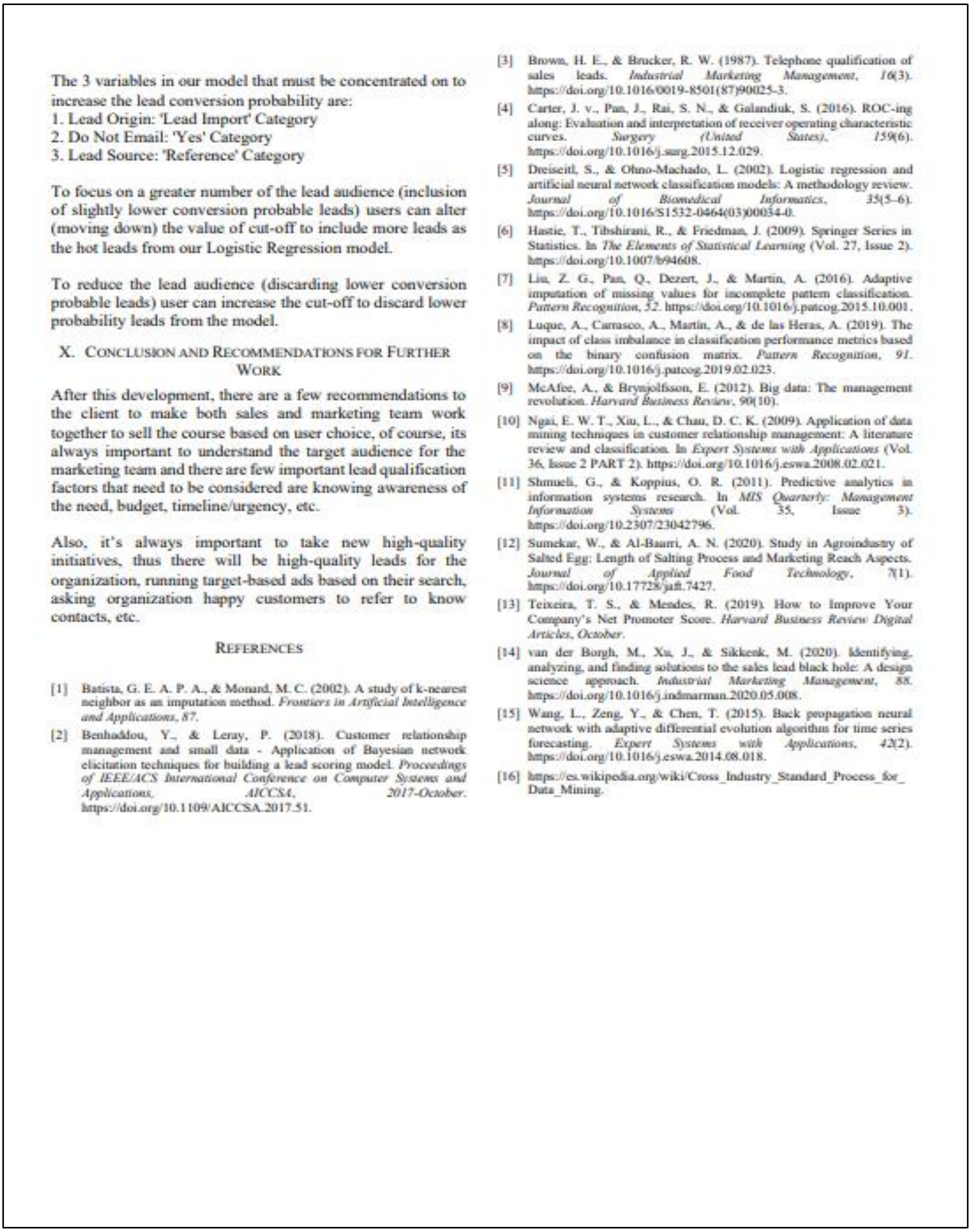
VIII. Deployment

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.