REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

CRM-based Lead Scoring with Machine Learning

Pradeep Thota

SRN: R19MBA63

Date: 27/08/2022

MBA in Business Analytics

Capstone Project Presentation
Year: II

race.reva.edu.in

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Agenda

01 Introduction
Back Ground | Current status | Why this study

02 Literature Review
Seminal works | Summary | Research Gap

03 Problem Statement
Business Problem | Analytics Solution

04 Project Objectives
Primary & Secondary Objectives | Expected Outcome

05 Project Methodology
Conceptual Framework | Research Design

06 Business Understanding
Business Context | Monetary Impact

07 Data Understanding
Data Collection | Variables

08 Data Preparation
Pre-processing | Process | Techniques

09 Descriptive Analytics
Univariate | Bivariate | Hypothesis

10 Modeling
Machine Learning | Model Evaluation | Insights

11 Model Deployment
Applications | Demo

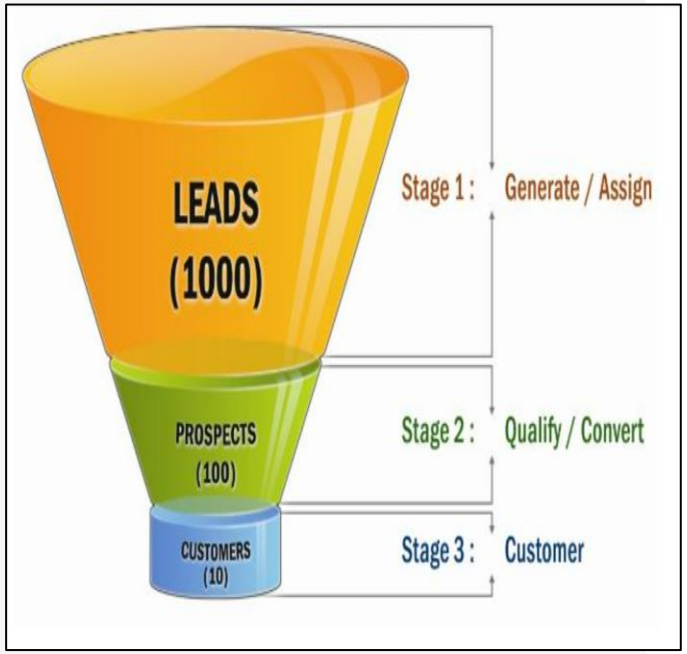
12 Suggestions and Conclusions
Insights | Next Step | Future Scope

13 Annexure
References | Publications | Plagiarism Score

Introduction

Background | Current status | Why this study

- *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.
- *Betutelage* is giving beautiful insights of students 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 in a particular subject with their courses.
- *Betutelage* along with the existing system they have entered 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.



Literature Review

Seminal works | Summary | Research Gap

- Have reviewed a minimum of 15 research papers
- 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.
- 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
- 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
- So lead scoring can be increased when it is implemented with the classification models like Random forest, logistic model, etc.
- Its always a good practice to build more classification models so can choose the better model with good metrics
- Once the model is built then it's very important to evaluate the model and to know the metrics of the model so now will get a Lead Prioritization and Scoring model with the path to higher conversion

Problem Statement

Business Problem | Analytics Solution

- 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.

Project Objectives

Primary & Secondary Objectives | Expected Outcome

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.

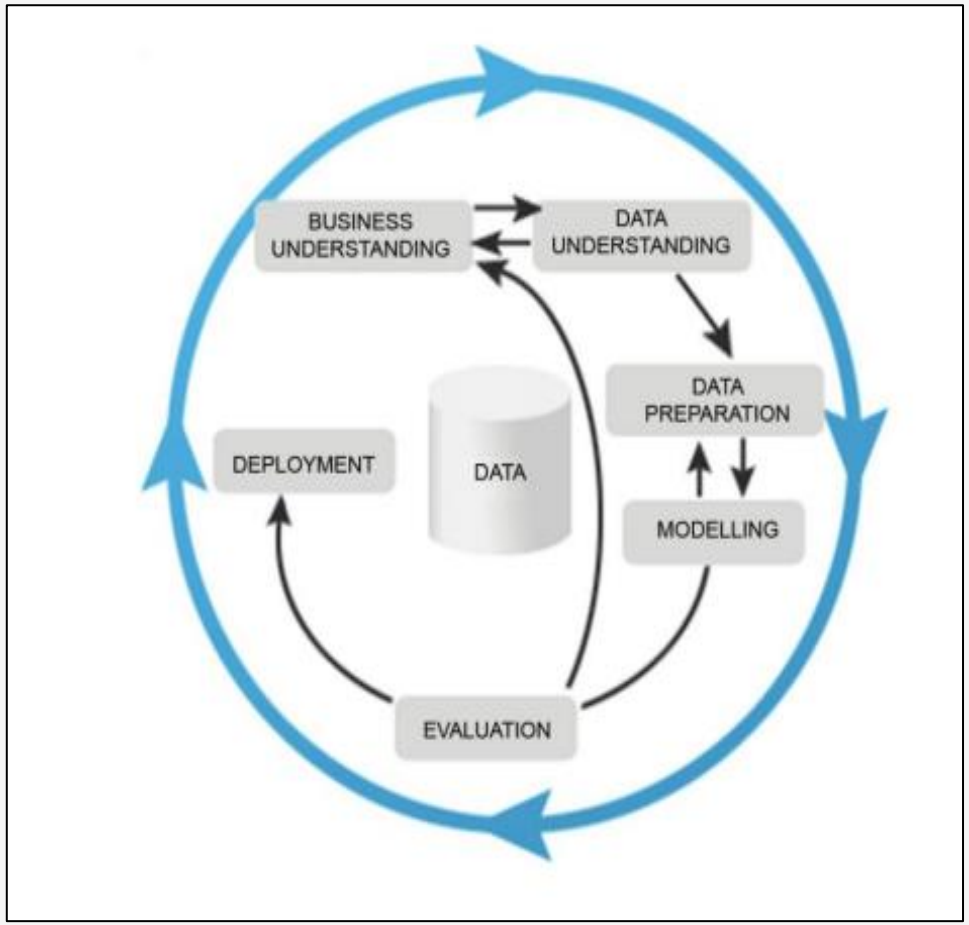
2) Data collection is not a crucial part of the project as having a good sample of data provided by the business that has collected the dump from their server, etc., but needs to do some data preparation on top of it.

3) Need to help the business by building classification models with appropriate techniques using Machine learning, Deep learning, Artificial Intelligence, etc. with accuracy with both train and test data.

Project Methodology

Conceptual Framework | Research Design

- **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.
- **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.



Business Understanding

Business Impact | Challenges | Monetary Impact

- As 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.
- Now the business can concentrate on the low lead score customers to make them as their paying customers by applying appropriate strategies.

Data Understanding

Data Collection | Variables

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 its maintained manually. The table 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

Data Preparation

Pre-processing | Techniques

- 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.

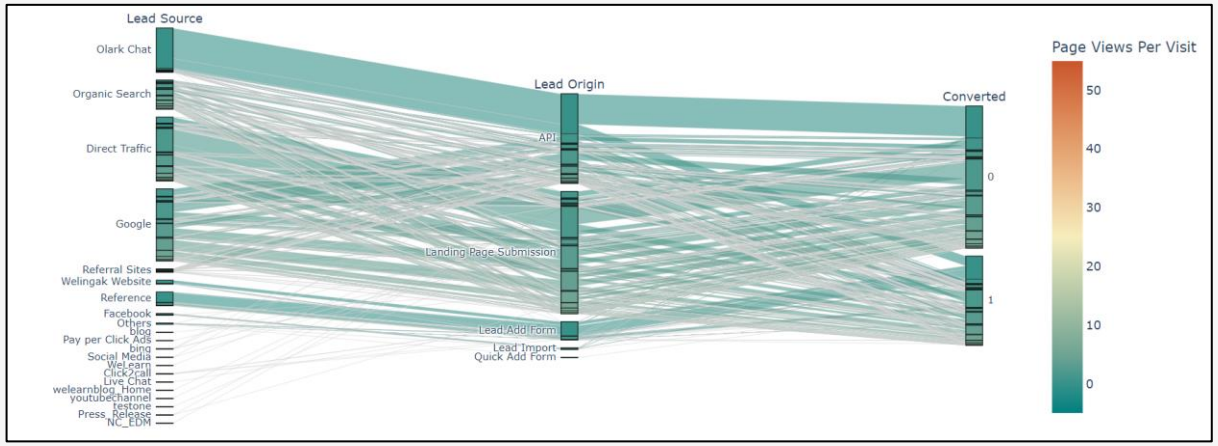
Data Preparation

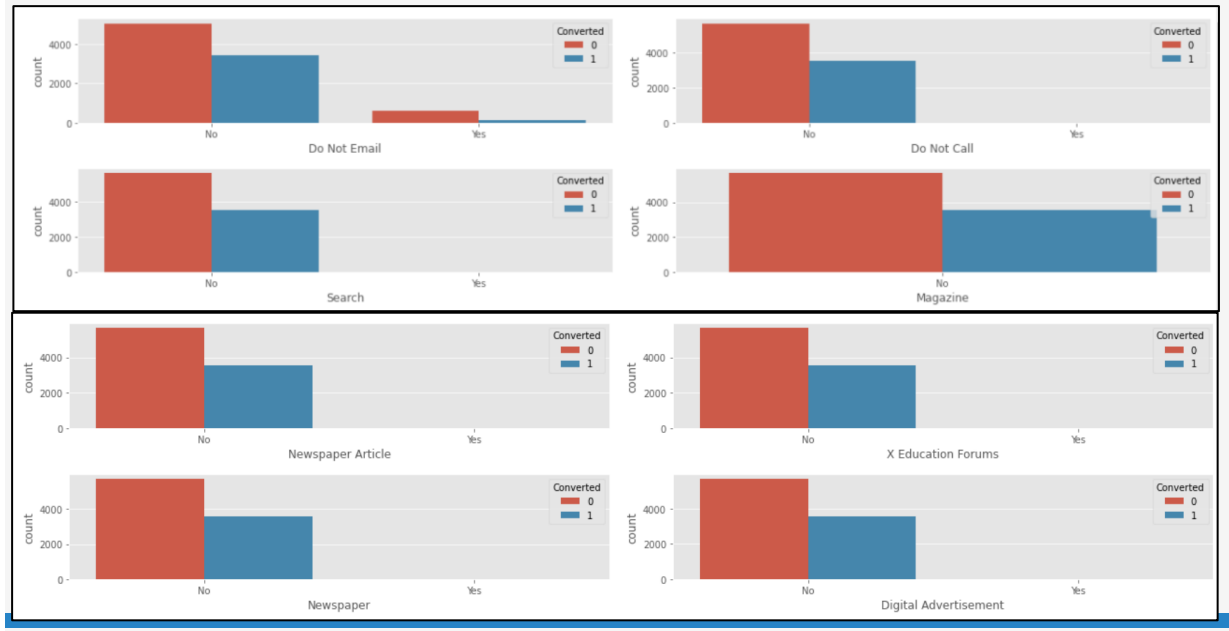
Pre-processing | Techniques

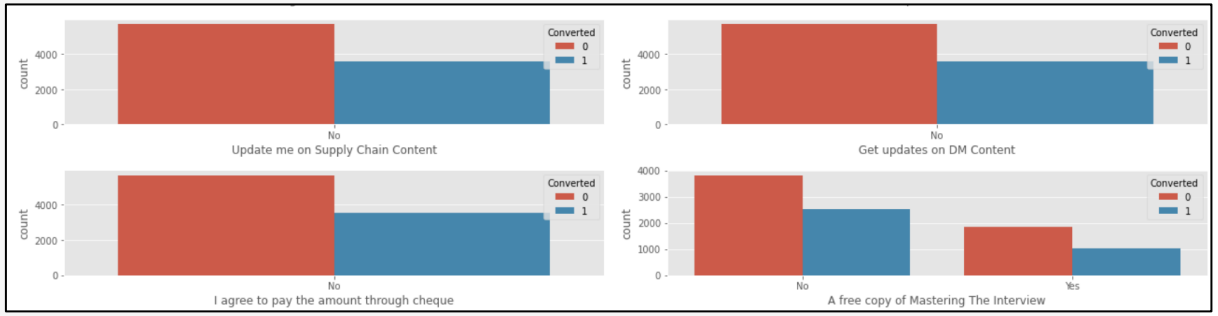
Prospect I	Lead Num	Lead Origi	Lead Sour	Do Not Em	Do Not Ca	Converted	TotalVisits	Total Time	Page View
7927b2df-	660737	API	Olark Chat	No	No	0	0	0	0
2a272436-	660728	API	Organic Se	No	No	0	5	674	2.5
8cc8c611-a	660727	Landing Pa	Direct Traf	No	No	1	2	1532	2
0cc2df48-7	660719	Landing Pa	Direct Traf	No	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
9fae7df4-1	660673	Landing Pa	Google	No	No	1	2	1640	2
20ef72a2-	660664	API	Olark Chat	No	No	0	0	0	0
cfa0128c-a	660624	Landing Pa	Direct Traf	No	No	0	2	71	2
af465dfc-7	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 Traf	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 Traf	Yes	No	0	1	481	1
25f4ac14-f	660547	API	Organic Se	No	No	1	6	1012	6
3abb7c77-	660540	API	Olark Chat	No	No	0	0	0	0

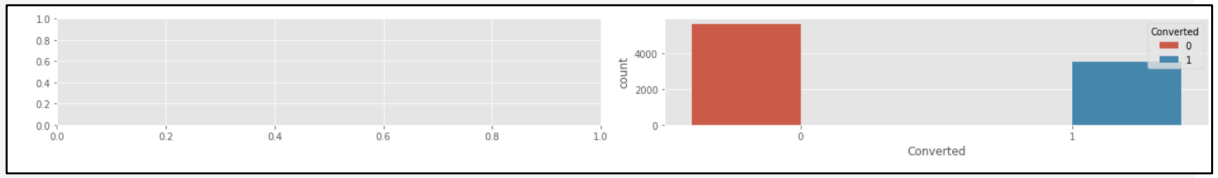
Lead Profi	Last Notat	Lead Origi	Tags_Inter	Last Activi	Tags_Ring	What is yc	Lead Sour	Lead Profi	What is yc	Tags_Clos	Last Notat	Last Activi
0	1	0	0	1	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	0	0
0	1	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

EDA


Lead Source
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
Others
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Origin
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Converted
0
1
Page Views Per Visit
50
40
30
20
10
0


Converted
0
1
count
No
Yes
Do Not Email
Do Not Call
Search
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement


Converted
0
1
count
No
Update me on Supply Chain Content
Get updates on DM Content
I agree to pay the amount through cheque
A free copy of Mastering The Interview


Converted
0
1
count
Converted

Descriptive Analytics

Multivariate Analysis | Hypothesis

- Based on the Figures, 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 because who does not like freebies? The reason may be has a large proportion of the audience is "Unemployed“.
- 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 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.

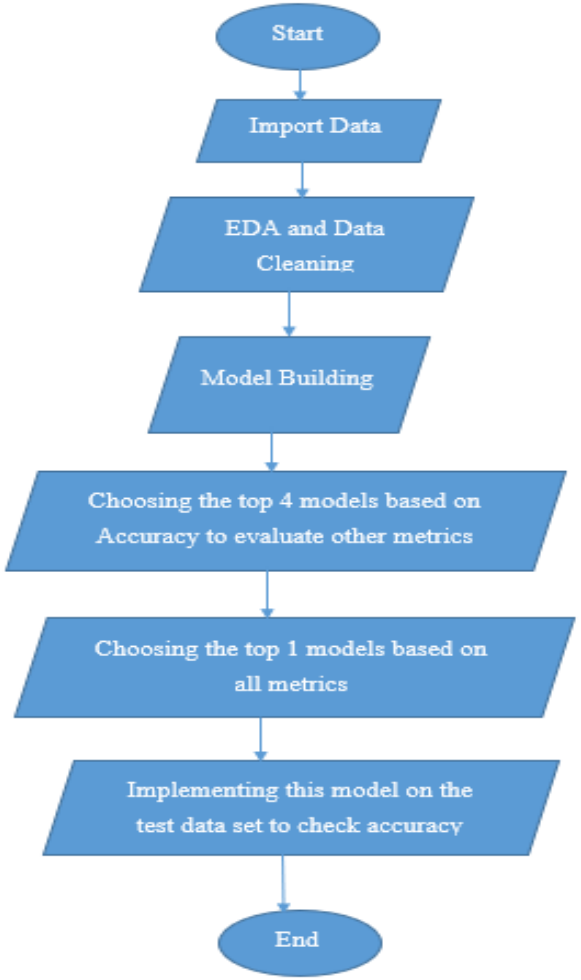
Modeling

Modeling Techniques | Modeling Process | Model Building

Have built 12 different classification methods i.e., **RandomForest, Adaboost, ExtraTree, BaggingClassifier, GradientBoosting, DecisionTree, KNN, Logistic, SGD Classifier, MLPClassifier, NaiveBayes, LightGBM, Catboost**.

After building models on several classifiers considered **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.



Model Evaluation

Results | Interpretation | Insights

- Initially, built 12 Classification models
- In these 12 models, based on accuracy here considering only the top 4 models for checking all other metrics in-depth.

Formula for metrics is as follows:

1) Accuracy = TP + TN / (TP + FP + FN + FP)
2) Precision = TP / TP + FP
3) Recall = TP / TP + FN
4) F1 = 2* Precision X Recall / Precision + Recall

```
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
```

Model Evaluation

Results | Interpretation | Insights

- The Top 4 models are RandomForest, Gradient Boosting, LightGBM, and Catboost.

```
1 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]]
```



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

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


1 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]]
```

Model Evaluation

Results | Interpretation | Insights

```
1 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]]
```

```
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]]
```

Model Evaluation

Results | Interpretation | Insights

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%

The above table shows the metrics of different Classification models from this we are choosing Random Forest as our final model for further evaluation

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

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.

Model Deployment

Demonstration

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.

Results and Insights

Key Findings | Suggestions

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

Conclusion and Future Work

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.

References

Bibliography | Webliography

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

References

Bibliography | Webliography

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

References

Bibliography | Webliography

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

REVA
UNIVERSITY

Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Thank You!