ADVANCED PREDICTIVE SALES SCORING USING MACHINE LEARNING.

Presented By:

Soniya Mary Sam 312317205132
Shruthi Lakshmi .S 312317205128

Under the Guidance of

Mrs. M. Poornima M.Tech IT,

OBJECTIVES

- To implement an advanced machine learning module to predict possible leads and its conversion probability for any type of organizations with their customizable parameters.

- To identify leads and classify them as hot leads, warm leads, cold leads from qualifiable leads.

- To provide ways to increase the number of customers from the initial prospects.

ABSTRACT

The marketing and sales team are the backbone of any business as they carry out the crucial role of encouraging the potential customer to purchase from them, boosting the firm's sales and revenue. To make the best use of the budget and time that goes into the process, we are implementing a predictive lead scoring model using machine learning. The model identifies the target customer using a lead scoring system that ranks leads based on their conversion probability and uses a funnel system to transform visitors into qualified leads and then as customers and promoters. The model built after the correlation of data and the test-trains splitting processes, leads to a faster lead conversion, saves a lot of time, and increases revenue proportionally.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
1.	Tejal Tandel; Sayali Wagal; Nisha Singh; Rujata Chaudhari; Vishal Badgujar **“Case Study on an Android App for Inventory Management System with Sales Prediction for Local Shopkeepers in India”;** March 2020 https://ieeexplore.ieee.org/document/9074234	Equipping the local shopkeepers with a mobile application to provide exposure to all the aforementioned benefits.	Cost effective solution that provides future sales insights.	Shopkeepers in rural areas may not have access to mobile phone and internet.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
2.	Oryza Wisesa ; Andi Adriansyah; Osamah Ibrahim Khalaf **“Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm ”** September 2020 https://ieeexplore.ieee.org/document/9249397	A brief analysis of the reliability of B2B sales using machine learning techniques.	Good accuracy in forecasting	Difficulty in dealing with big data.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
3.	Sunil K Punjabi ; Vikyhat Shetty; Shreemun Pranav; Abhishek Yadav **“Sales Prediction using Online Sentiment with Regression Model ”** May 2020 https://ieeexplore.ieee.org/document/9120936	Predicts the sales of a vehicle using sentiment analysis from various places on the internet.	Simple estimation procedure.	Sensitive to outliers.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
4.	B.Sri Sai Ramya1 , K. Vedavathi **“An Advanced Sales Forecasting Using Machine Learning Algorithm”;** May 2020 https://www.ijisrt.com/assets/upload/files/IJISRT20MAY134.pdf	Data mining techniques are used for sales forecasting such are ARIMA models and XG Boost algorithms which get better efficiency to manipulate the trending sales analysis.	XGBoost which is an expanded gradient boosting algorithm was once found to function the excellent at prediction.	Shopkeepers in rural areas may not have access to mobile phone and internet.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
5.	Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate and Prof. Dr. Nikhilkumar Shardoor **“Sales Prediction Using Machine Learning Algorithms.”** June 2020 https://www.irjet.net/archives/V7/i6/IRJET-V7I6676.pdf	Machine Learning algorithms such as Linear Regression, K Nearest Neighbors algorithm, XGBoost algorithm and Random Forest algorithm have been used to predict the sales of various outlets of the Big Mart. Various parameters such as Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracies which determine the precision of results are tabulated for each of the four algorithms. .	More accuracy. Gives an accuracy of 93.53%.	Doesn't count seasonal fluctuations

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
6.	Yulyardo and Sani M. Isa **“Predictive Business Intelligence: consumer Goods Sales Forecasting using Artificial Neural Network.”** May 2019 https://www.academia.edu/40204516/PREDICTIVE_BUSINESS_INTELLIGENCE_CONSUMER_GOODS_SALES_FORECASTING_USING_ARTIFICIAL_NEURAL_NETWORK	This method help stakeholders to know the progress, to make decision and strategy, and to evaluate the sales performance. By using Qlik Sense BI Tools and Predictive Data Mining using RapidMiner Tools.	More accuracy and better performance	Complex approach

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
7.	Sunitha Cheriyan ; Shaniba Ibrahim; Saju Mohanan ; **“Intelligent Sales Prediction Using Machine Learning Techniques ”** August 2018 https://ieeexplore.ieee.org/document/8659115	Analyzed the concept of sales data and sales forecast by using various techniques forecasting testing.	Best fit model	Can improve future accuracy.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
8.	Kumari Punam ; Rajendra Pamula; Praphula Kumar Jain **“A Two-Level Statistical Model for Big Mart Sales Prediction ”** September 2018 https://ieeexplore.ieee.org/document/8675060	Prediction of sales of a product from a particular outlet is performed via a two-level approach	Better performance	Complex approach

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
9.	Hossein Sangrody, Ning Zhou, Salih Tutun,BenyaminKhorramdel, Mahdi Motalleb,Morteza Sarailoo **“Long Term Forecasting using Machine Learning Methods.”** March 2018 http://academia.edu/longtermpaper/v5rfg	Artificial neural network(ANN), support vector regression (SVR), recurrent neural network (RNN), k-nearest neighbours (KNN), generalized regression neural network (GRNN), and Gaussian Process Regression (GPR) are used for for training and validation of the load forecasting models.	More accuracy and better performance	Complex approach

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
10.	Ping-feng Pai And Chia-hsin Liu **“Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values.”** October 2018 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8481411	Comparing the performance of two time series models namely Winters model and linear exponential smoothening on the simulated datasets. Deseasonalizing procedures were employed to deal with different types of data.	Best fit approach	Doesn't count seasonal fluctuations

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
11.	Hendrik Setyo Utomo; Rabini Sayyidati; Oky Rahmanto **“Implementation of mobile-based monitoring sales system in Semi Tani Shop”** November 2017 https://ieeexplore.ieee.org/abstract/document/8304137	Aims to implement the sales information system of agricultural store sales based on Mobile and test the usability of the application.	Fast response time	Accuracy can be improved

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
12.	Du Hong; Cui Bo **“Sale forecasting method in dynamic environment based on ARMA(1,1)”** April 2011 https://www.researchgate.net/publication/252003459_Sale_forecasting_method_in_dynamic_environment_based_on_ARMA11	Uses the ARMA (1,1) model to forecast the circle and irregular factor separated from multiplication model.	Improved the forecasting accuracy	Time consuming approach .

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
13.	Zhang Xiaoyan; Wu Qiong; Yu Xiaoran **"Recognitions on Mixed Sales Confirmation"** November 2011 https://www.researchgate.net/publication/254016956_Recognitions_on_Mixed_Sales_Confirmation	Discusses the problems with segmenting non-taxable labor services.	Pays great attention on mixed sale transaction.	May not be personalised to business.

LITERATURE SURVEY

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
14.	Bai Yuewei; Wei Shuangyu; Lou Binchao **“Research and Development on Lean Collaborative Software System for Sales Activity Management”;** May 2009 https://dl.acm.org/doi/10.1109/IFITA.2009.562	Lean platform modeling technology for sales activities collaborative management according to the requirement of small-to-medium size (STMS) companies in China on distributed sales activities collaborative management and sales data consolidated management. In the application, STMS companies normally sell their products via some direct sales channels built by them, e.g., sales offices, or sub-companies which spread around different cities.	Low system implementation cost, customized functionalities, and short software development time.	Integration is difficult.

LITERATURE SURVEY: BASE PAPER

SI.NO	AUTHOR,TITLE ,YEAR	DESCRIPTION	ADVANTAGE	DISADVANTAGE
15.	Carlos Aguilar-palacios , Sergio Muñoz-romero , And José Luis Rojo-álvarez **“Cold-Start Promotional Sales Forecasting Through Gradient Boosted-Based Contrastive Explanations”;** July 2020 https://ieeexplore.ieee.org/document/9149573	This method presents the cold-start forecasts in relation to the observed promotional sales of other products, which we call neighbors. They are selected based on a measure of closeness to the predicted promotion, which is derived from the variable importance calculated during the training of the regressors. . The results on real-market data also show that the proposed method performs at a similar level to widespread methods such as conventional CatBoost, NGBoost or AutoGluon, and has the advantage of providing interpretability	Improved the forecasting accuracy	Sensitive to outliers.

EXISTING SYSTEM AND DISADVANTAGES

- Human interventions make the process more expensive and time consuming as there are always room to mistakes and default leads.
- The existing model is more expensive and less accurate. The lead might be possibly using other products or not from the same segment as the target market is.
- The generated leads cannot be qualified as hot or warm leads without testing as the sales team would have to do them.

DISADVANTAGES:

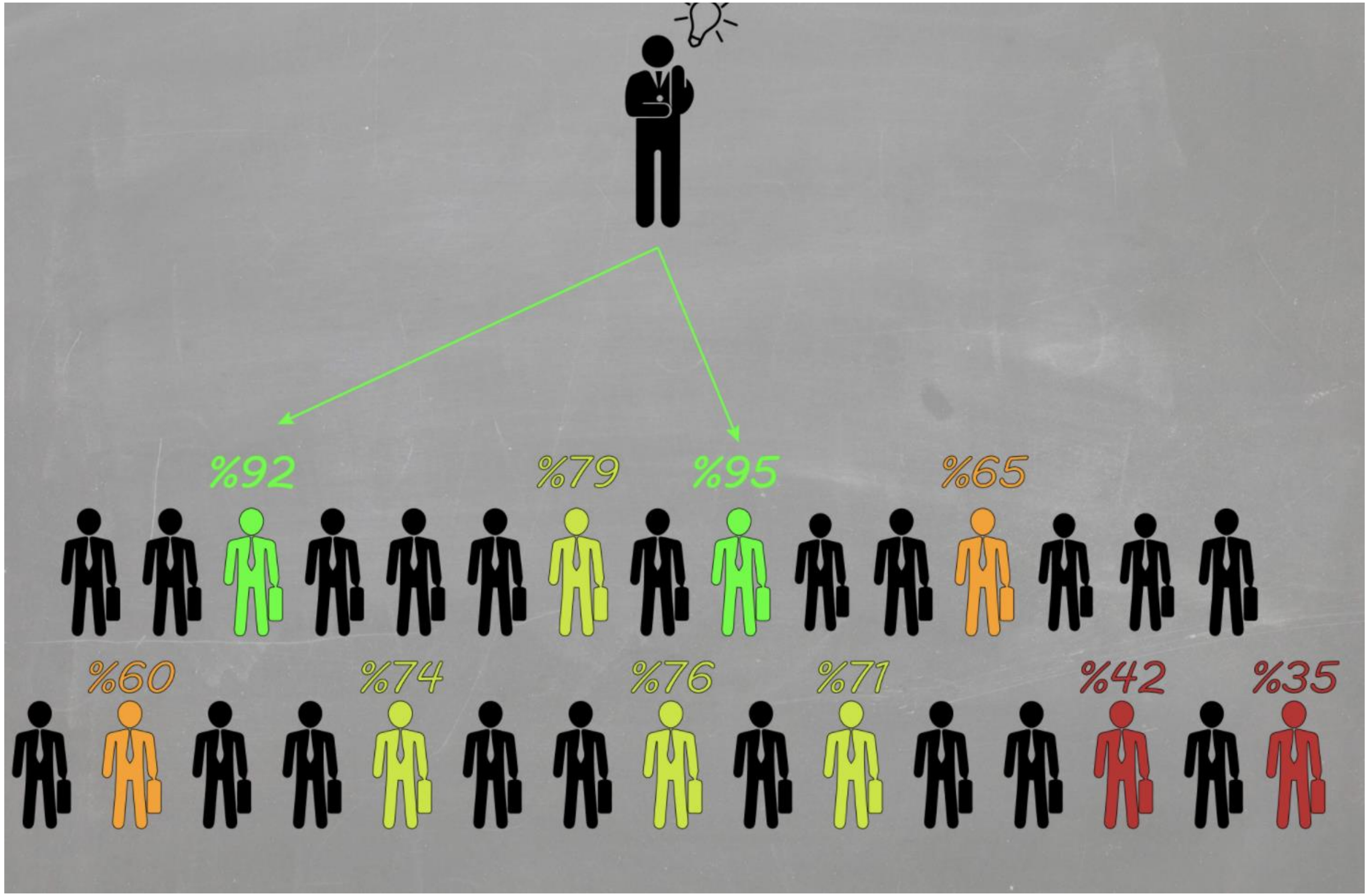
- Expensive to implement
- Non-qualified leads
- More time is spent, and less revenue generated

PROPOSED SYSTEM AND ADVANTAGES

- The lead scores are predicted based on machine learning(supervised binary classification algorithm) techniques that identify the right leads and helps us to save ample time to converse with next leads.
- These predictions can influence business strategy and can be done mutually with human resource or can be programmed independently based on the business requirement.
- The system makes the client prospect to qualify lead conversion easier as the lead is targeted and warm when the proposal is sent.
- This is an enterprise grade data science problem at hand which is resolved using machine learning to understand deep insights on customers and make good business decisions.

ADVANTAGES:

- Faster lead conversion(Direct)
- Saves time and increase revenue generation
- Saves the spend and expense with easy implementation



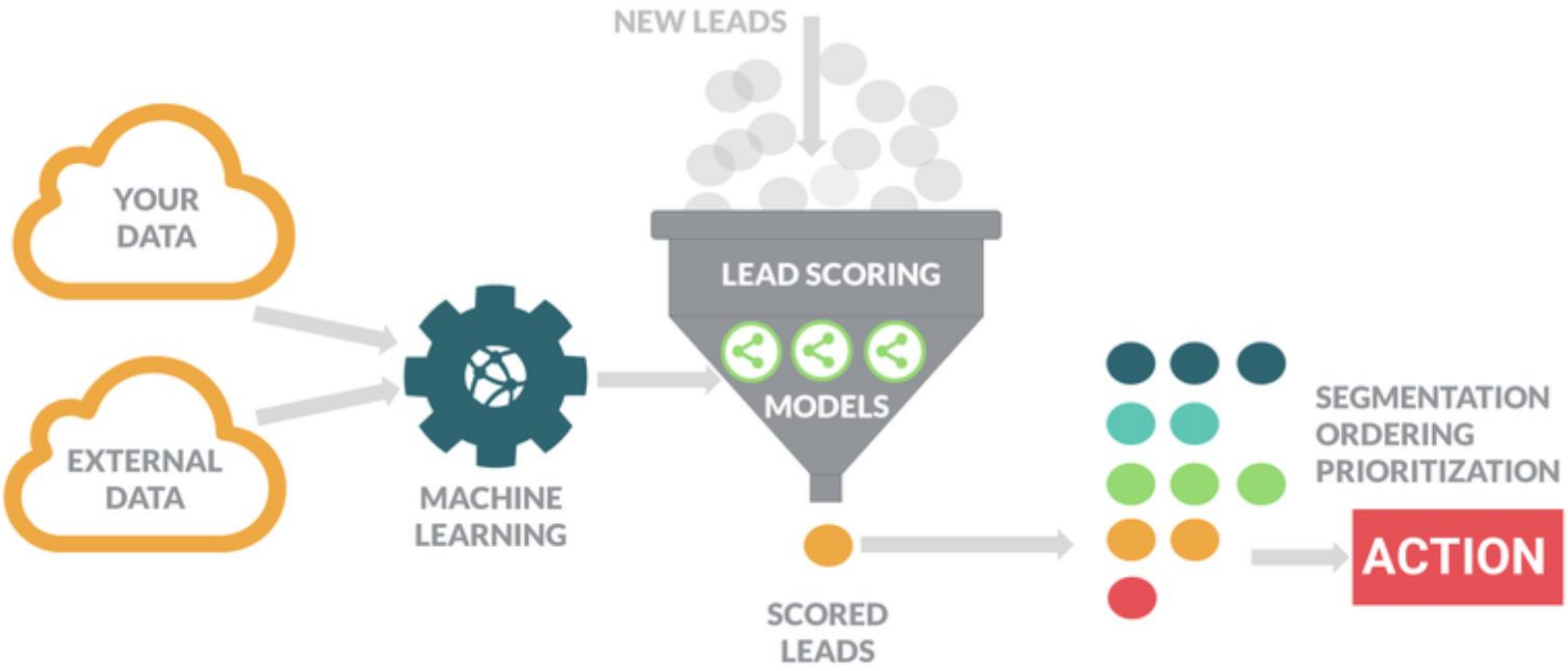
PREDICTIVE SALES SCORING

APPLICATION

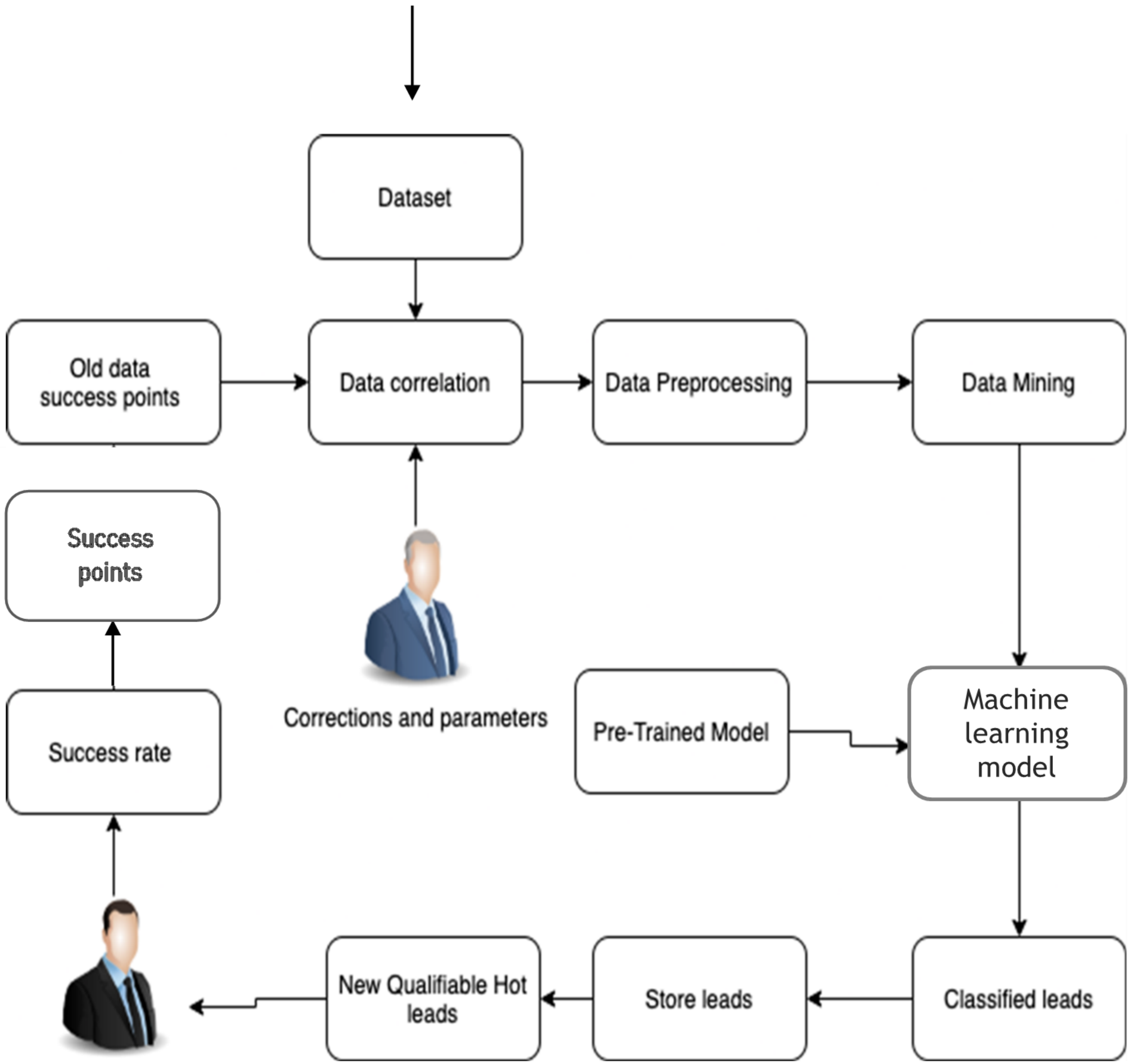
Where this can be applied?

- Organizations
- Sales team
- B2B Companies
- Lead generator models
- Any product/service consumer facing companies.

FLOW DIAGRAM



ARCHITECTURE DIAGRAM



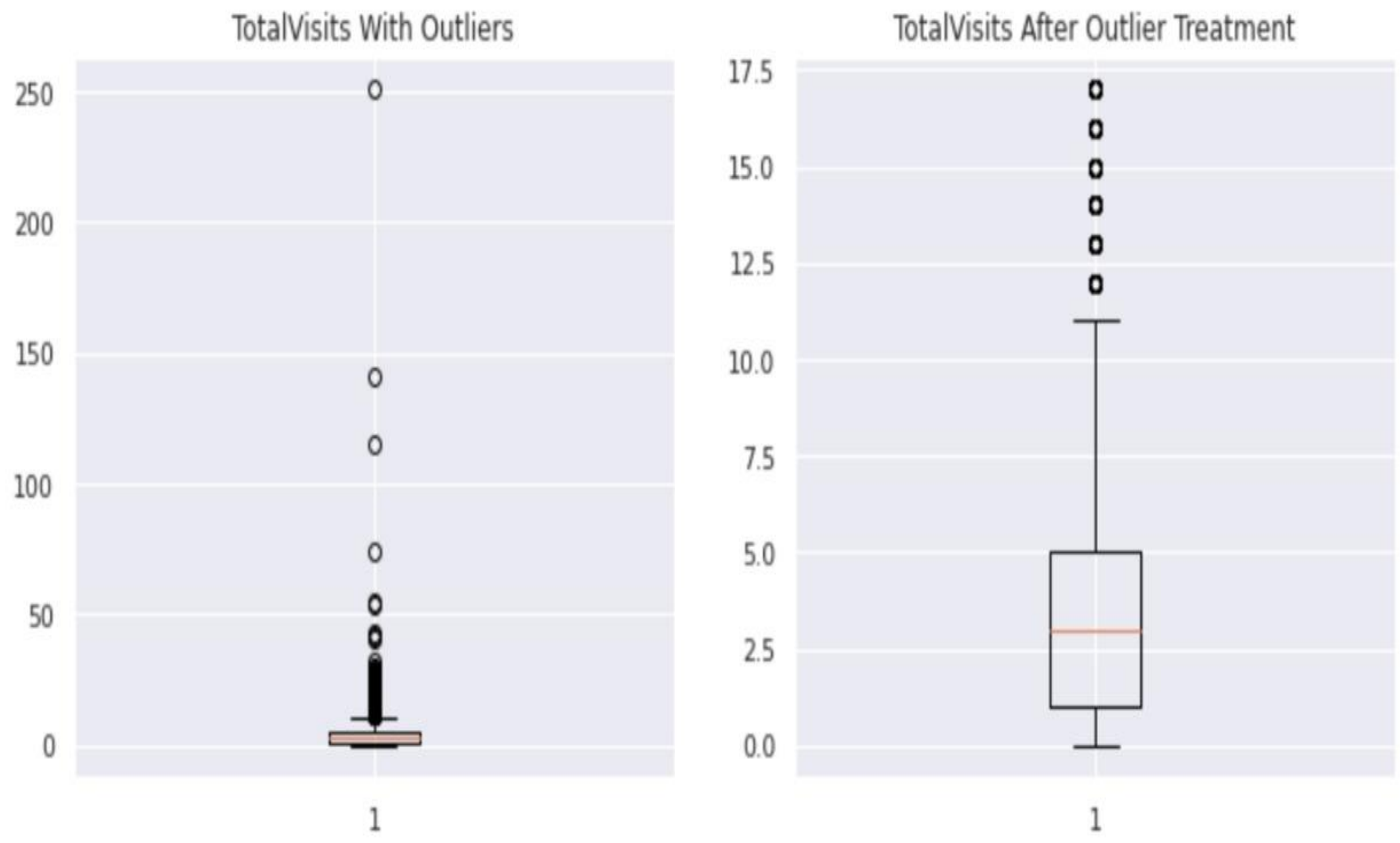
LIST OF MODULES

- DATA PREPROCESSING
- MACHINE LEARNING PROCESS
- LEAD PROCESSING
- DATA MINING

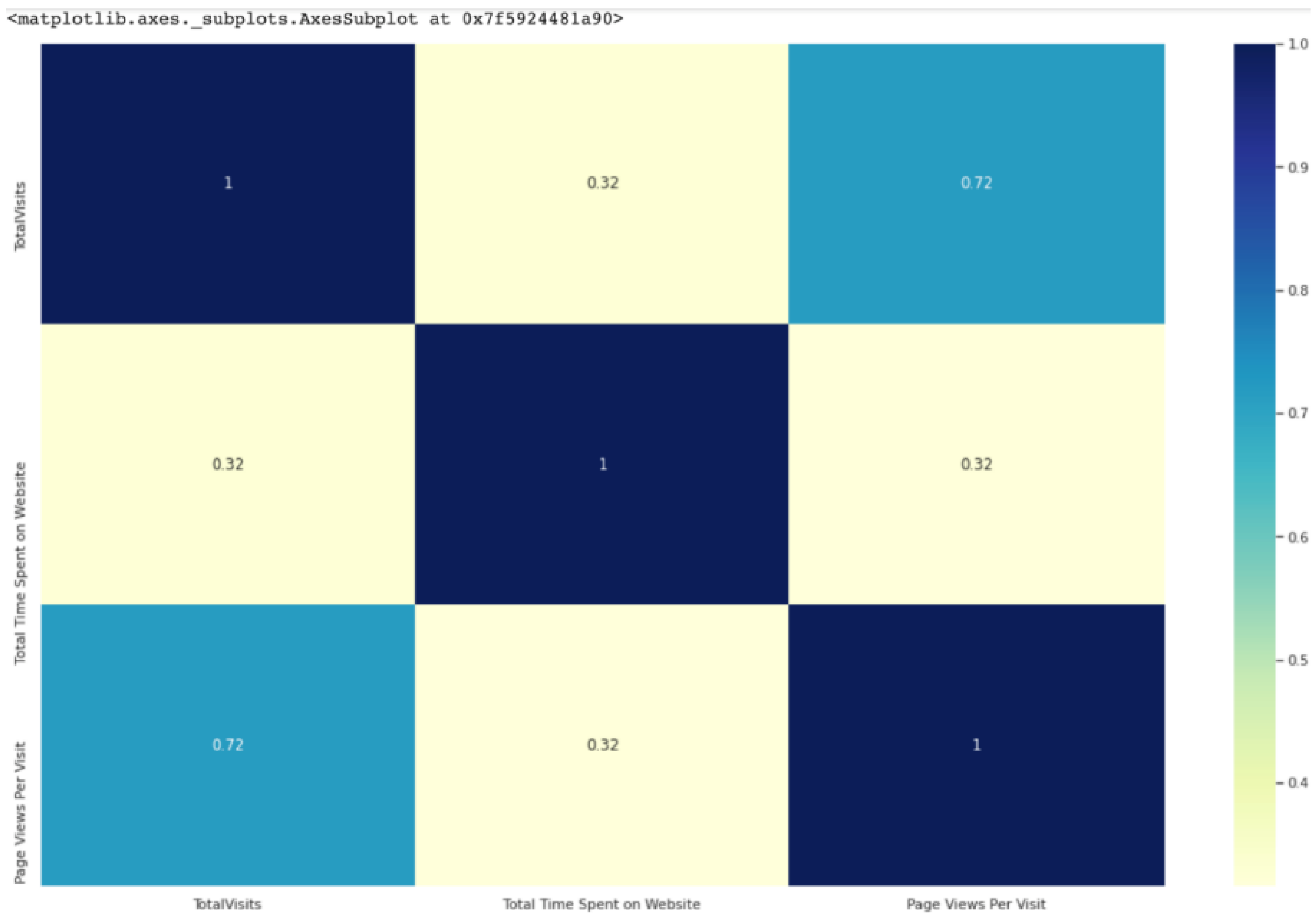
DATA PREPROCESSING

- The input raw data often comes in unusable forms and requires to be cleaned before using it to build a model.

  - By handling missing and noisy data(outliers).
- Exploratory data analysis is performed by visualizing categorical and numerical variables through Univariate and Bivariate analysis.
- Since the data has varying scales, it is transformed by performing standardization using StandardScaler().
- Data correlation is performed by plotting against a Heat Map to find highly correlated features.
- Here, we doesn’t process the type of lead directly but convert the raw data to classifiable format.



OUTLIERS TREATMENT USING IQR
(INTERQUARTILE RANGE OUTLIER)



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5924481a90>
```

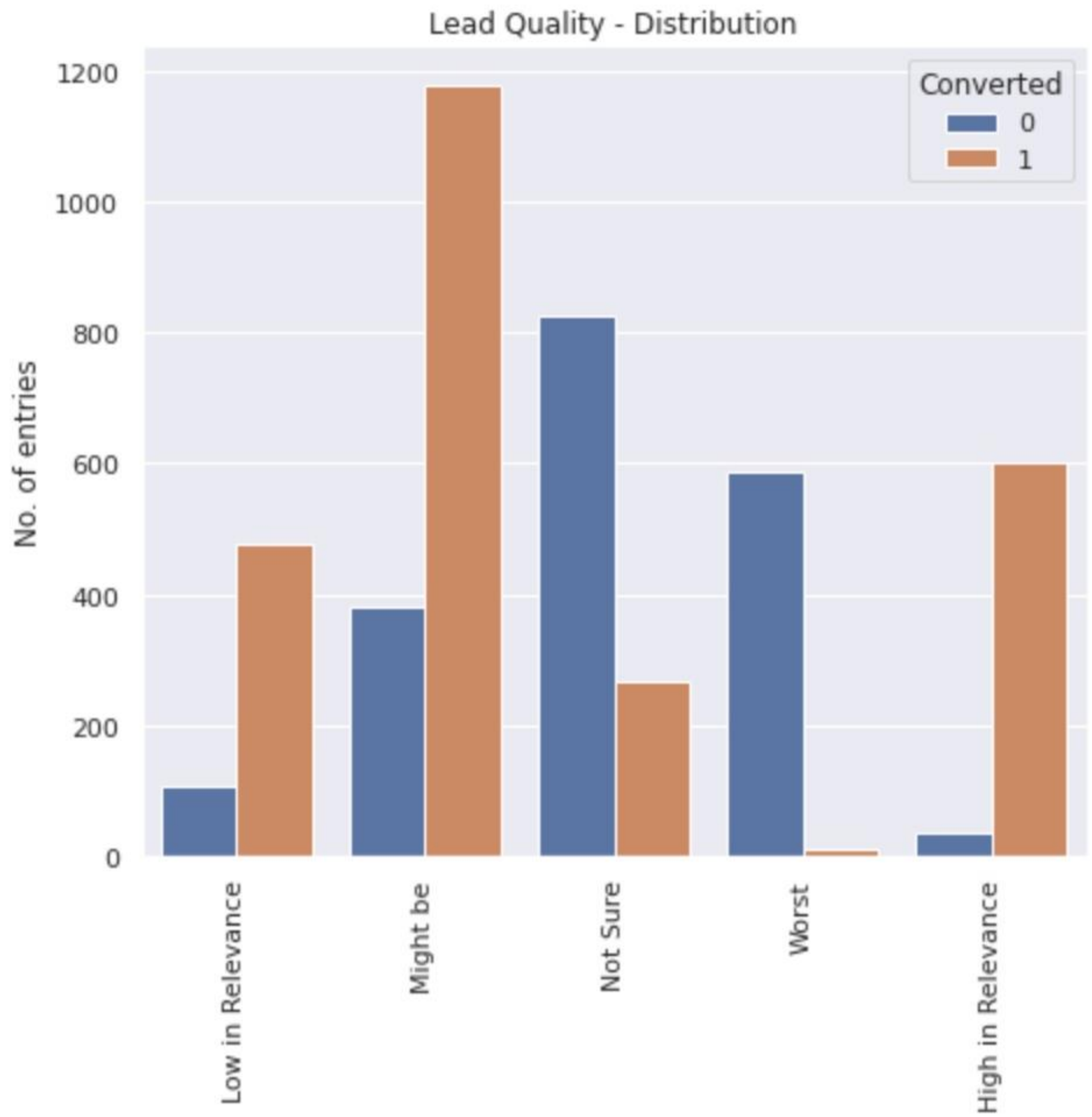
CORRELATION HEAT MAP

MACHINE LEARNING PROCESS

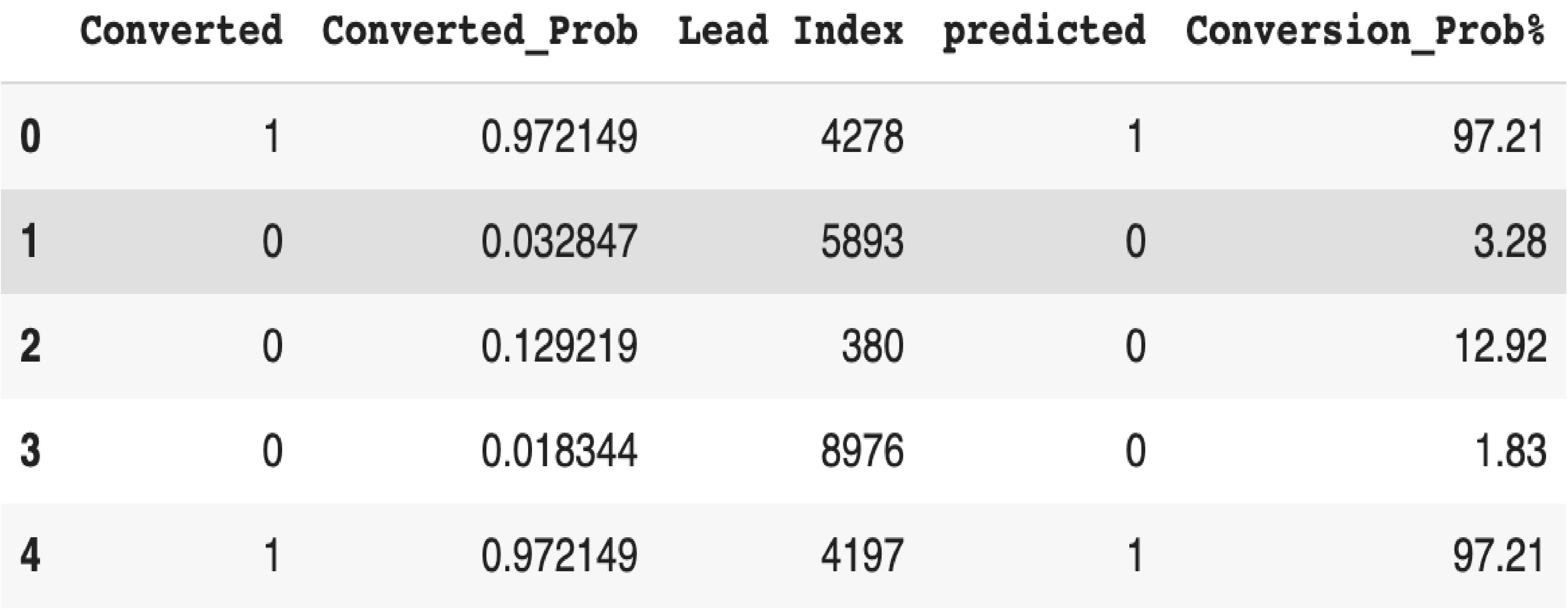
- Initially the categorical variables are encoded, as the numerical data processing is more efficient in ML.
- The preprocessed is then split into test data and train data. The target variable to be predicted is identified and moved to a separate dataframe.
- A logistic regression model was built for the binary classification problem.
- Feature selection is performed using Recursive Feature Elimination(RFE) method to reduce data dimensionality by ranking the significant features.
- Multicollinearity between various attributes were checked using Variance Inflation factor(VIF)
- The AUC(Area Under the Curve) in the plotted ROC(Receiver Operating Characteristic) curve was 0.95.This implied that the model had good ability to distinguish between classes.
- The built model predicted the Test data with an accuracy of 89.23%.

LEAD PROCESSING

- Through this process the possible leads are generated for the sales team to convert the prospects into leads and close the sale.
- In the given dataset, leads are scored on scale of 1-20. The leads are classified into 3 categories namely Hot, Warm and Cold based on their conversion probability.
- The prediction model enables business to push the Hot leads as a priority into the sales funnel, to convert them into Customers before they turn warm or cold. This helps the business to make unique offers to that customer to encourage sales.
- By predicting the right leads, businesses will be able to target their marketing to customers that are most likely to purchase. This saves a lot of time and increases the revenue proportionately.



LEAD QUALITY

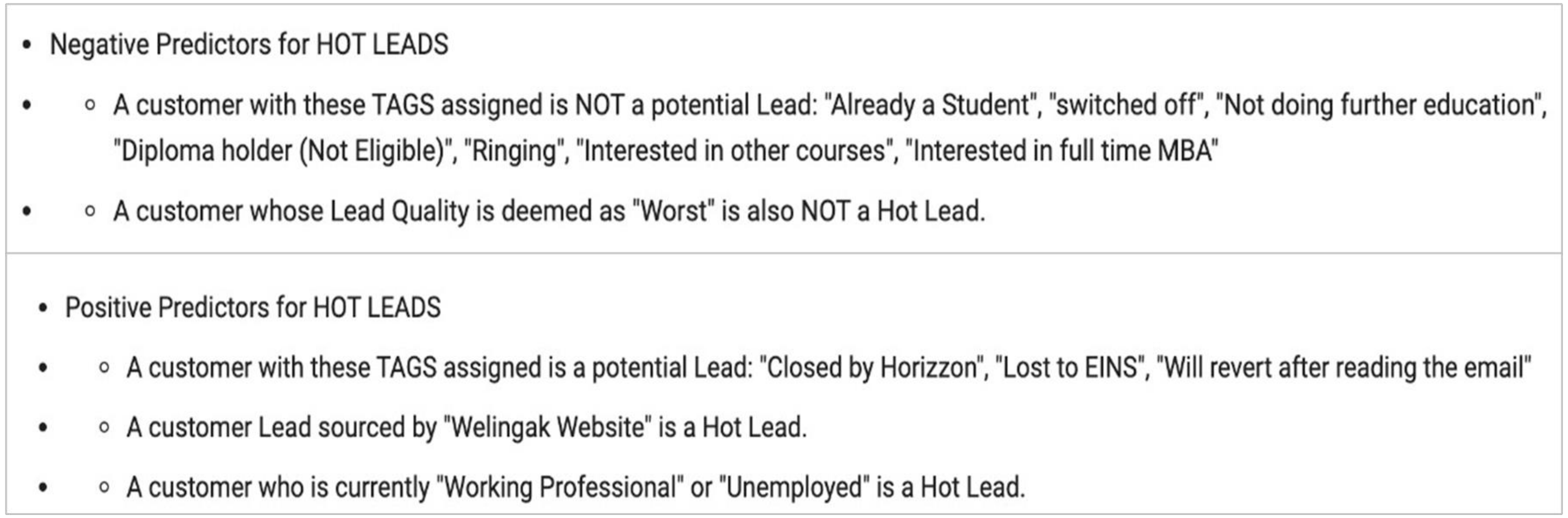


	Converted	Converted_Prob	Lead Index	predicted	Conversion_Prob%
0	1	0.972149	4278	1	97.21
1	0	0.032847	5893	0	3.28
2	0	0.129219	380	0	12.92
3	0	0.018344	8976	0	1.83
4	1	0.972149	4197	1	97.21

LEAD PROCESSING

DATA MINING

- Data mining is undertaken to discover the patterns in the large data set.
- Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.
- Predictive data mining is used to forecast trends that help business leaders to make better decisions and marketing strategies.



- Negative Predictors for HOT LEADS
  - A customer with these TAGS assigned is NOT a potential Lead: "Already a Student", "switched off", "Not doing further education", "Diploma holder (Not Eligible)", "Ringing", "Interested in other courses", "Interested in full time MBA"
  - A customer whose Lead Quality is deemed as "Worst" is also NOT a Hot Lead.

- Positive Predictors for HOT LEADS
  - A customer with these TAGS assigned is a potential Lead: "Closed by Horizzon", "Lost to EINS", "Will revert after reading the email"
  - A customer Lead sourced by "Welingak Website" is a Hot Lead.
  - A customer who is currently "Working Professional" or "Unemployed" is a Hot Lead.

CUSTOMER BEHAVIORAL ANALYTICS AFTER DATA MINING

CONCLUSION

- For building a predictive sales scoring model, past customer data was used. Initially, the raw data was cleaned by removing irrelevant observations and handling missing data.

- Exploratory data analysis helped to visualize the distribution of customers against various attributes.

- The data was transformed and scaled before reducing the dimensionality using Recursive feature elimination.

- Data correlation was performed and multicollinearity was handled followed by test-train splitting for model building.

- A logistic regression model was built that predicted conversion probability with an accuracy score of 89.23%

FUTURE ENHANCEMENT

- Bias of the model can be improved to increase the rigidity and flexibility of the model and prevent from being an under-fitting model.

- Variance can be channelized to prevent the model from picking up noisy data. Regularization can be used for better feature selection.

- Bagging and Boosting methods can be added to increase complexity and improve success rate

REFERENCES

1. Aberdeen Group. Collaborative Product Commerce: Delivering Product Innovation at Internet Speed. Aberdeen Group, 1999, 12(9): 1~9.
2. Aberdeen Group. Beating the Competition with Collaborative Product Commerce: Leveraging the Internet for New Product Innovation. Aberdeen Group Inc, 2000 Preface.
3. L. Gao, X.Y. Shao et al. Collaborative Product Commercials. Journal of China Mechanical Engineering. 2001, 12(2): 168~173.
4. K. Johns. Beyond supply: Collaboration and CPC. Manufacturing Systems. 2000, 18(6): 32.
5. Y.B. Luo et al. Analysis and comments for the tools based on CPC. Journal of Computer Engineering and Application. 2003, 29: 70~74.

THANK YOU