National College of Ireland

Identifying Factors Contributing to Lead Conversion Using Machine Learning to Gain Business Insights

MSc Research Project
MSCDADJAN22

Mansi Sharma
Student ID: x21143315

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. Mansi Sharma……

Student ID: ……x21143315@student.ncirl.ie……..……

Programme: …… Data Analytics…… **Year:** ……2022……..

Module: …… MSc. Research Project…….……

Supervisor: …… Qurrat Ul Ain …….……

Submission Due Date: ……01-02-2023…….……

Project Title: …… Identifying Factors Contributing to Lead Conversion Using Machine Learning to Gain Business Insights…….……

Word Count: ……7691…… **Page Count**……20…….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Mansi Sharma……

Date: ……29-01-2023……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identifying Factors Contributing to Lead Conversion Using Machine Learning to Gain Business Insights

Mansi Sharma
x21143315

Abstract

Digital marketing has become a major factor contributing to a company's revenue. Since, the people invest most of their time online, it is important do promotions online in order to run a business. Companies promote their products online in a lot of ways such as showing advertisements, sending newsletters and so on. This makes people get interested in their products and are called leads. Since not every lead will turn into customer, it is beneficial to identify those leads who have the potential and nurture them in becoming customers. There are a number of factors which contribute in converting a lead to a customer. The goal of this research is to predict the lead conversion with the help of machine learning and determine the most contributing factors in this process. The study uses 5 different machine learning algorithms, namely, Logistic Regression, Decision Tree, Random Forest, CatBoost and XGBoost in predicting the lead conversion. It also aims to visualise the results so that it is easily understandable to the non-technical stakeholders and thus, to help sales and marketing teams in planning to target leads. Among the best performing models, importance of all the variables are calculated from Logistic Regression, Random Forest and CatBoost and the most contributing ones are selected. These variables are then used to create visualizations in Tableau that will help businesses in making marketing and sales strategies.

Keywords: Lead Conversion, Binary Classification, Feature Importance, Tableau

1 Introduction

1.1 Background

Marketing is about developing relationships with the target audience at a suitable time and location. Today, that means you need to connect with them online, where they are already spending time. This is where Digital Marketing comes into the picture which refers to internet marketing activities in general. To communicate with their present and potential customers, businesses use digital platforms such as search engines, social media, email, and other websites. Digital marketing is the act of communicating with customers online, where they spend the majority of their time, using a number of digital platforms and approaches. Depending on the objectives of their marketing strategy, marketers might utilize both free and paid channels to promote a larger campaign. Any individual who expresses interest in a company's goods or services in any way, shape, or form is considered a lead. Leads usually hear back from a company after initiating contact by providing personal data for an offer, trial, or subscription. Once a potential consumer sees a business's website and provides their contact information, they go from the generating phase to the conversion phase. The process where a lead converts into a customer by purchasing a product or service from a company is known as lead conversion. A process by which a potential consumer learns about your

product, exhibits interest in it, and then progresses to become a paying customer is known as a lead funnel as shown in Fig. 1. *(Joshi, 2018)*

Fig. 1 Leads Funnel

A sales funnel (shown in Fig. 2) is a part of the lead funnel which refers to the route that potential consumers take on their way to making a purchase. It helps the businesses in understanding the activities of the leads which they can use in making appropriate marketing strategies to covert more leads into customers. *(Lead Funnel Definition, Stages, and Strategy, 2022)*

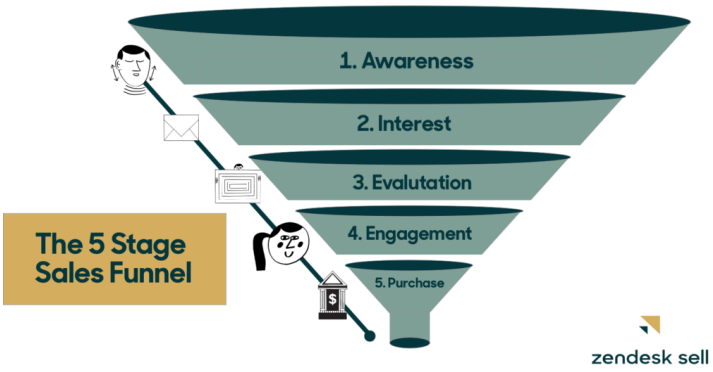

Fig. 2 Sales Funnel

Lead conversion is more than just ensuring leads visit a company's website; it is also about turning them into opportunities, nurturing them, and converting them into customers. Lead conversion depends on several factors like the origin of the lead, their occupation, number of times they visit the company's website, whether they sign up for a newsletter, and so on.

1.2 Motivation

A precise lead prediction strategy can assist marketing and sales teams prioritize targeted leads and reply to them in a timely manner, increasing their chances of becoming clients. Most of the research has been done in the lead scoring area which is the process of assigning values to leads based on their behaviour in relation to their interest in products. It can help in targeting potential customers to some extent but lead conversion prediction actually forecasts which leads will become customers on the basis of various attributes. Having implemented lead conversion, the sales team can determine which leads require more attention and which do not. This research aims to predict whether a lead converts into a customer depending upon the previously stated factors. Additionally, valuable business insights will be fetched from the data visualization, understandable to both technical and non-technical stakeholders.

1.3 Research Question

How can machine learning be used to predict lead conversion and which factors majorly influence the lead conversion?

1.4 Research Objectives

The research aims to achieve the below objectives:

- To apply 5 different machine learning algorithms, namely, Logistic Regression, Decision Tree, Random Forest, CatBoost and XGBoost, and determine the best performing algorithm with the help of evaluation methods.
- To find that factors which contribute to the conversion of leads into customers.
- To build interactive dashboards containing data visualizations to help sales and marketing teams in targeting leads.

1.5 Report Structure

This report has been separated into many segments. The background of the issue, the research question, and the objectives are covered in the introduction. The second portion is the literature review, which helps familiarize the reader with the previous work and provides support for the research. The following section of the document describes the data pre-processing and recommended methodology and procedures for the research. The next section discusses the implementation of the experiments/methods utilized in the research. The utilized approaches are evaluated in a separate section along with the analysis on Tableau dashboards. On the basis of dashboard analysis, business recommendations are stated in the next section. Lastly, the conclusion provides a succinct summary of the full research project.

2 Related Work

2.1 Predicting Lead Conversion with Machine Learning

An efficient lead scoring approach may assist marketing and sales teams in prioritizing leads, responding quickly to those leads, and eventually increasing the likelihood that those leads will convert into customers. *(Jadli et al., 2022)* evaluated and studied the prediction capabilities of many machine learning algorithms used to predict lead conversion. They employed five distinct categorization methods and assessed them based on their respective precision, recall, specificity, and F1 scores. Although, the accuracy of the models was quite high, but the variables responsible for the prediction were not talked about in the study. This research doesn't use gradient boosting algorithms and neither performs EDA in order to fetch business insights which could have been beneficial to the non-technical stakeholders.

The study *(Nygård & Mezei, n.d.)* utilises machine learning algorithms in order to automate lead scoring systems. It was concluded that the random forest method had the greatest overall performance among all the models. In addition to the automated procedure, it provides some insight into the manual lead scoring process. On the other hand, no comparison could be drawn between the two. To compare the models, the authors employed accuracy, Area Under the Curve (AUC), sensitivity, and specificity. The study has achieved good accuracy with all the models. However, the study doesn't use any gradient based algorithms for the prediction.

The purpose of this study (B. Gouveia & Costa, 2022) was to predict lead conversion and identify the gaps that are leading the leads to not becoming customers. The authors used Logistic Regression, kNN and Random Forest for the prediction among which the LR performed the best and identified the variables contributing the most to lead conversion.
The authors *(Lee et al., 2021)* in this research attempted to predict online customer behaviour by using Gradient Boosting models and used several evaluation metrics to assess the models. However, a wider range of parameter candidates together could have been covered with machine learning models other than those used in this study. Also, it did not analyse the marginal effects which could have been helpful in providing economic explanation.
The research *(Chaudhuri et al., 2021)* attempted to increase the knowledge of online customer purchasing behaviour for an e-commerce platform by forecasting it using deep learning techniques on a massive sample of multidimensional data. Prediction approaches utilizing machine learning, such as Decision Tree, Random Forest, Support Vector Machines, and Artificial Neural Network (ANN) where ANN performed the best. However, the data was very specific to European market which might not be generalizable for other locations.
With the advancement of online payment mechanisms and e-commerce platforms, an increasing number of clients are opting for online purchasing. The Catboost model is used in this study *(Dou, 2020)* to examine and forecast if customers would buy a specific product. The results could have been validated more by using other machine learning techniques. The accuracy, particularly the recognition of a few categories, could have been enhanced.

2.2 Binary Classification through Machine learning

The authors *(Mostafa & Hasan, 2021)* presented and They compared Artificial Neural Network (ANN), Random Forest, and Support Vector Machine, which were used to tell the difference between blood donors and non-donors with hepatitis, fibrosis, and cirrhosis diseases. In all models, SVM and Random Forest did better than ANN.
One of the most common problems that the telecommunications industry and other industries and organizations that offer subscription-based services face is customer churn. The authors *(Iranmanesh et al., 2019)* provide a solution to the problem of predicting customer churn by leveraging the telecommunications sector data set and ANN to evaluate the variables driving customer churn and optimize the approaches by experimenting with different activation functions. The study offers light on the precision and performance of various techniques, such as altering epoch, batch size, neuron layer count, activation function, and optimizer.

2.3 Predicting Potential Customers

The purpose of this study *(Deng et al., 2018)* is This study aims to determine the Click-Through Rate (CTR) for sponsored advertisements on commercial online search engines in order to improve the user experience and generate revenue. Utilizing Recurrent Neural Networks (RNN) with Long Short-Term Memory, the prediction is made (LSTM). The research is confined to adverts displayed in online browsers that are not based on the information of the users, but rather on their browsing history.
The research *(Yeo et al., 2020)* assesses the conversion rate and probability that a customer will purchase a particular product. The forecast is based on both conventional and marketplace search. A combined modelling of both patterns is recommended on the basis of

the purchasing decision process. In a dynamic market environment, conversion projections and predictability are more accurate than existing baselines, according to the study's findings.

2.4 Data Analysis by Tableau for Decision Making

This study *(Hoelscher & Mortimer, 2018)* demonstrates the relevance of data analysis by using a data visualization tool Tableau to demonstrate how it might be turned into information that could aid in decision-making. The research uses different case studies and analyses their respective data in order to make business strategies on them. After this, the study discovers relevant trends in the data in order to compile and publish this information in order to improve the corporate decision-making process.

The major objective of business intelligence and analytics is to help in decision-making and boost the profit of the corporation. Tableau is one of the many business-related applications used to analyse and visualize regularly generated data of various forms. The primary objective of this paper *(Jena, 2019)* to demonstrate how quickly forecasting and analysis can be conducted using Tableau by utilizing a superstore's information to anticipate sales and profit for the next four quarters of the following year. In addition, it evaluates all business intelligence and analytics frameworks based on factors like complexity, speed, etc.

2.5 Summary

In conclusion, the problem is not solved because the non-technical decision makers need to be shown something which is easy to understand and is reliable to make strategies if a company wants to enhance its business. For this, Tableau is used in this research so that trends can be analysed and improvements can be made in marketing and sales practices to convert more leads into customers. Also, tree based and gradient based models will be used to get more accurate models and the set of variables that contribute to the lead conversion.

3 Research Methodology

This research employs the CRoss Industry Standard Process for Data Mining (CRISP-DM) CRISP-DM methodology as it is a suitable fit for this type of research which demands business understanding. The flowchart of CRISP-DM is given in Fig. 3. *(Hotz, 2018)*

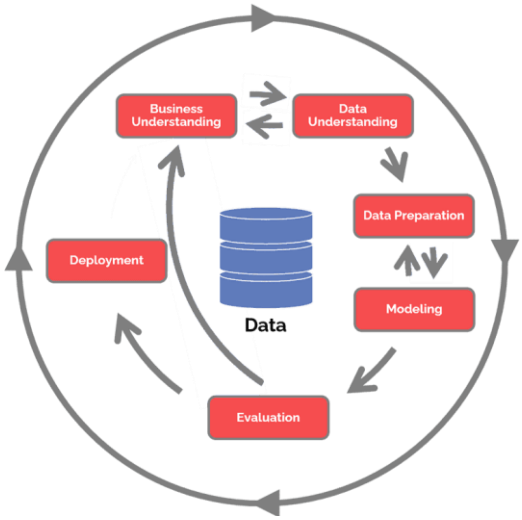

Fig. 3 CRISP-DM Methodology

3.1 Business Understanding

This kind of project that require data analysis, it is necessary to begin with an understanding of the domain's background knowledge, as well as a clear knowledge of the business objectives at the time, as this will allow for a perfect transformation of data into information using data mining algorithms. Lead conversion entails using a combination of variables to convert leads into customers. At each phase, the company must provide opportunities for the lead to take action in order to proceed to the next step.

3.2 Data Understanding

For this project, the dataset is sourced from Kaggle (*Lead Scoring Dataset*) and it includes 9240 rows and 37 columns. Among 37 columns, 30 attributes are categorical variables while 7 are numerical. The dataset contains information on every lead (who has signed up on the website) of the company and their different activities. The data was obtained from Kaggle and was utilized in an UpGrad Case Study. It is owned by a corporation that provides online courses to students, working people, and others. People who want to buy their products go to their website and search for courses. Once they visit the website, they may investigate the courses and submit a form containing their personal details, they are classified as leads. After acquiring the leads, sales team members begin making calls, sending emails, etc. The attributes of the dataset are nothing but the general information about those leads, their preferences and efforts that sales and marketing team make in order to convert them successfully. Fig. 4 depicts the dataframe before it was pre-processed.

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0000000000	0	0.0000000000	...	No	Select	Select
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0000000000	674	2.5000000000	...	No	Select	Select
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0000000000	1532	2.0000000000	...	No	Potential Lead	Mumbai

Fig. 4 Dataframe

3.3 Data Preparation

3.3.1 Data Cleaning

Firstly, missing/null values were checked in the data and it was observed that 17 columns had missing values while rest 20 didn't. In order to treat this, columns with more than 3000 missing values were dropped from the dataframe as the features didn't contain any information which might have affected the prediction. For the rest of the columns, imputation technique was carried out in order to replace the null values. For the columns 'Lead Source', 'Country', 'How did you hear about X Education', 'What is your current occupation', 'What matters most to you in choosing a course', 'Lead Profile' and 'Last Activity', missing values were replaced by mode value. For the columns 'TotalVisits' and 'Page Views Per Visit', missing values were replaced by median value. The dataframe was also checked for duplicate

values but none were found. After all the cleaning, the shape of the dataset came out to be (9240, 16).

3.3.2 Feature Selection

Next, the percentage of space occupied by values in categorical columns was calculated. Categorical columns with more than 90% of the same values in them were dropped from the dataframe which included columns like ‘Receive More Updates About Our Courses’ ‘Update me on Supply Chain Content’, ‘Get updates on DM Content’, and so on.

3.3.3 Data Transformation

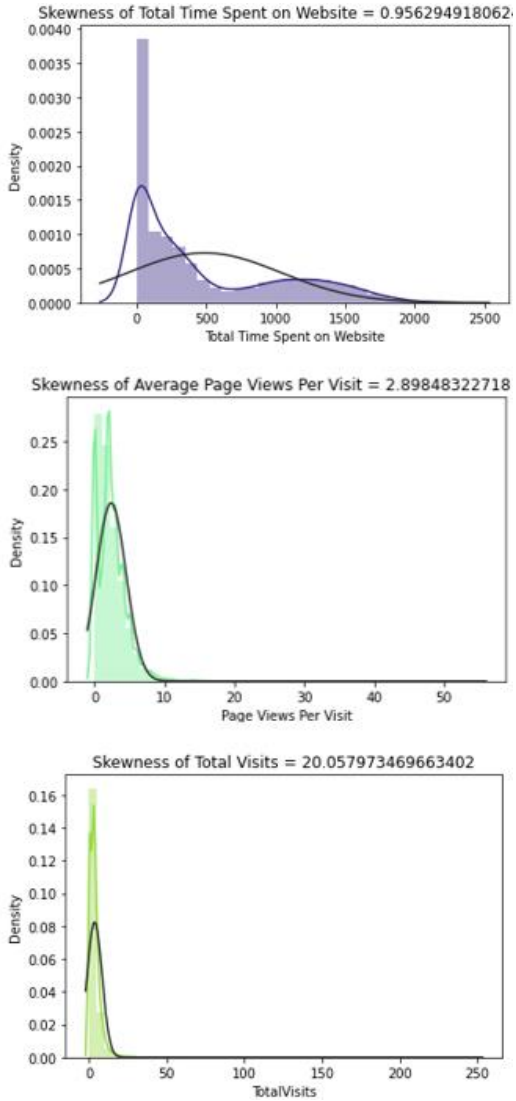

Fig. 5 Skewness After Taking Square Root

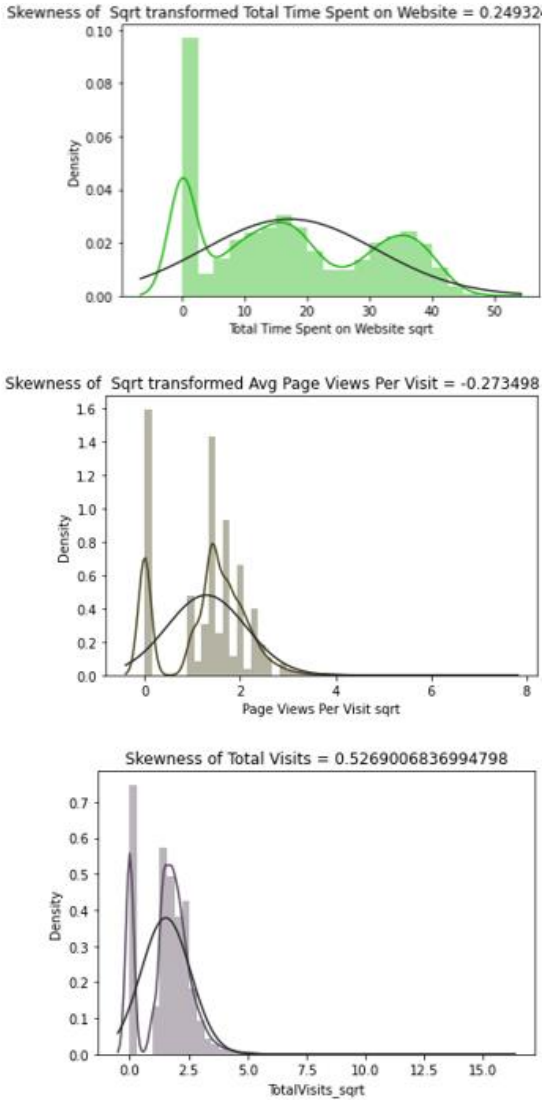

Fig. 6 Skewness After Taking Square Root

After cleaning the data, skewness of the numerical variables was checked. Skewness is the measure of how asymmetrical the data is. For this, ‘scipy.stats’ package was used which contains the statistical functions required to plot graphs. In order to reduce the skewness of the columns 'Total Time Spent on Website’, 'Page Views Per Visit' and 'TotalVisits', their square roots were taken. Fig. 5 and 6 represent numerical variables before and after their skewness was reduced. In Fig. 5, it can be seen that all three numerical variables are highly

skewed to the left. On taking their square roots, the skewness has highly reduced and the plots looked more like a normal distribution as shown in Fig. 6.

Furthermore, outlier analysis was carried out to identify any outliers present in the dataset. It is an important process which removes inaccurate data and aids in better accuracy of the model. The process was carried out for columns 'TotalVisits_sqrt', 'Total Time Spent on Website sqrt' and 'Page Views Per Visit sqrt' wherein outliers were detected and later on removed from the data using ‘quantile’ function from ‘NumPy’. Fig. 7 and 8 represent 2 numerical variables before and after outlier treatment was carried out.

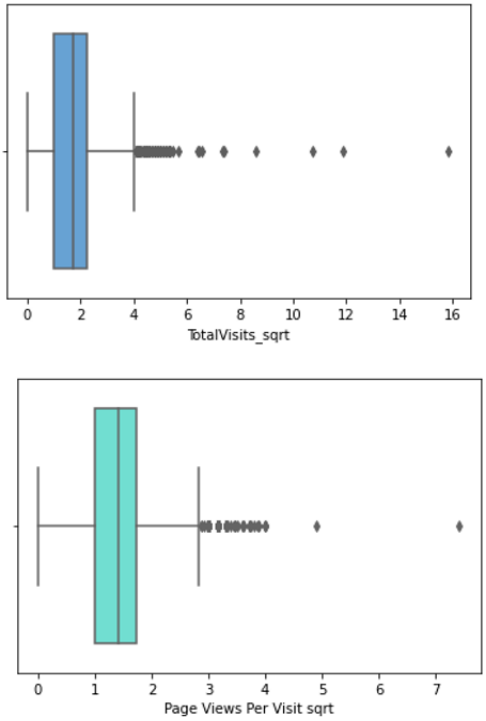

Fig. 7 Before Treating Outliers

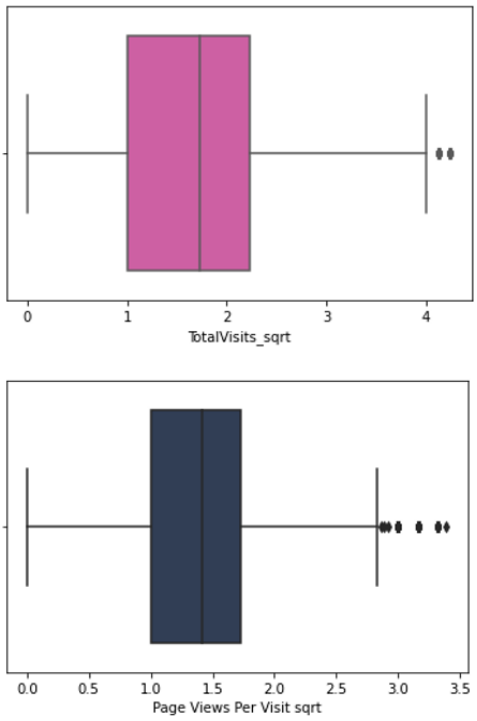

Fig. 8 After Treating Outliers

Most machine learning models do not directly employ attribute, which are often represented through text, hence it is important to get numerical encodings for categorical attributes. For this purpose, One Hot Encoding has been used which was imported from sklearn library. One Hot Encoding is applied because the categorical variables are nominal in this case.

Normalization of the features was done which makes the variables more consistent with one another, allowing a model to forecast accurate output. For this Min-Max scaler is used where all values between minimum and maximum are scaled to 0-1 based on the original value.

3.3.4 Exploratory Data Analysis

EDA is used to look into the data and summarize the most important findings. With the help of statistical summaries and graphical representations, it is used to find trends or patterns. For this, Tableau will be used to gain better insights into their data to offer the best customer experience.

3.4 Data Modelling

After the dataset has been pre-processed and the model has been built, the implementation will begin. Following algorithms will be used for the prediction:

3.4.1 Logistic Regression

Logistic regression is a supervised machine learning approach that predicts the likely outcome of binary classification tasks. Because the target variable in the dataset is dichotomous, logistic regression will be a suitable approach. It is extremely beneficial for predicting a dichotomous outcome and can handle a high number of variables which is one of the key advantages of logistic regression. *(What Is Logistic Regression?)*

3.4.2 Decision Tree Classifier

The internal nodes represent the features of a dataset, the branches represent the decision rules, and the leaf nodes represent the outcome. In a decision tree, the process of deciding the class of a given dataset begins at the tree's root node. Based on the comparison, this algorithm moves to the next node along the branch. The way a binary tree is built, with each logical statement being checked one at a time until the final "yes" or "no" target is reached, makes it a good tool for making "Yes" or "No" these predictions. *(Machine Learning Decision Tree Classification Algorithm - Javatpoint)*

3.4.3 Random Forest Classifier

A random forest is composed of numerous decision trees, each of which forecasts a value for the probability of target variables. Then the final outcome is calculated by averaging the probability. By picking data points with replacement, the first dataset samples are generated. Next, a subset of the available input variables is employed to generate decision trees rather than all of them. *(Random Forest | Introduction to Random Forest Algorithm)*

3.4.4 CatBoost Classifier

CatBoost is a high-performance open-sourced decision tree gradient boosting library. Typically, categorical characteristics in the dataset must be encoded before using them in algorithms. However, the CatBoost algorithm uses one-hot encoding to automatically encode those categorical attributes into numeric values. It is a ready-made classifier according to scikit-learn's that automatically handles categorical features. (T. Gouveia, 2021)

3.4.5 XGBoost Classifier

The XGBoost library employs a gradient boosting decision tree technique. Boosting is an ensemble method that use fresh models to fix flaws created by prior models. Models are added in succession until no more advancement is feasible. Utilizing sequentially built shallow decision trees and a highly scalable training technique that reduces overfitting, it generates accurate results. *(How to Create a Classification Model Using XGBoost in Python)*

3.5 Evaluation

In this step, it will be evaluated how well the models meet the objectives. Following evaluation matrices will be used to evaluate the performance of the models. *(Czakon, 2022)*

3.5.1 Accuracy

It tells how many times a model correctly classified an item in the dataset in comparison to the total. Accuracy is a good metric to use if the dataset is well-balanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.5.2 Precision

Precision is just the accuracy that is calculated only for classes that are positive. It tells how often it's correct when a class is classified as positive.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

3.5.3 Recall

It is the ratio between the number of predicted examples belonging to a class and the number of actual examples belonging to that class.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

3.5.4 F1 Score

This is the harmonic mean of accuracy and recall and is likely the most prevalent metric for assessing binary classification algorithms. If the F1 score increases, the model has improved accuracy, recall, or both.

$$F1 = 2.\frac{Precision \times Recall}{Precision + Recall}$$

3.5.5 ROC (Receiver Operating Characteristic) Curve

It is a probability curve depicting the True Positive Rate (TPR) vs the False Positive Rate (FPR) at various threshold levels. Essentially, it distinguishes between the "signal" and the "noise." A description of the ROC curve, the Area Under the Curve (AUC) indicates how effectively a classifier can distinguish between groups.

3.6 Feature Importance – Selecting Most Contributing Variables

Once the prediction is completed, the importance of variables will be calculated from Logistic Regression, Random Forest and CatBoost. In Logistic Regression, it is known as coefficient while in Random Forest and CatBoost, it is known as feature importance. All three will be compared and common variables will be selected for the visualization.

3.6.1 Logistic Regression Coefficient

Coefficients are usually calculated to study the contribution of each independent feature in the model. In logistic regression, the coefficients (in Fig. 9) represent the change in logit for each unit change in the predictor variable. It is accomplished by computing the impact of the predictor on the exponential function of the regression coefficient – the odds ratio. The coefficients for individual variables will be calculated from 'coef_' from the scikit-learn library. *(Sklearn.Linear_model.LogisticRegression)*

$$\operatorname{logit} p = \sigma^{-1}(p) = \ln \frac{p}{1-p} \quad \text{for} \quad p \in (0,1)$$

Fig. 9 Logistic Regression Logit Formula

3.6.2 Random Forest Feature Importance

The importance of a characteristic is calculated as the ratio between the decrease in node impurity and the likelihood of reaching that node. The probability of a node may be calculated by dividing the total number of samples by the number of samples that reach the node. The higher the value, the greater the importance of the feature. It is given by the formula (in Fig. 10) and can be calculated using the fitted attribute 'feature_importances_' from the scikit-learn library.*(Ronaghan, 2019)*

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

Fig. 10 RF Feature Importance Formula

Here, RFfi sub(i)= the importance of feature i calculated from all trees in Random Forest, normfi sub(ij)= the normalized feature importance for i in tree j and T = total number of trees.

3.6.3 CatBoost Feature Importance

The feature importance score shows how effective each feature was in building the boosted decision trees inside the model. The more important an attribute is, the more often decision trees use it to make decisions. The sum of feature importance score of all the variables is equal to 100. It is calculated using 'feature_importances_' attribute. (*Feature Importances*)

4 Design Specification

The data needed for the research is obtained via Kaggle. The data is then cleaned to prepare it for further processing. Data transformation is performed in order to properly organize the data. Finally, machine learning models are applied, their performances are evaluated and the best performing models are selected. By calculating feature importance from best models, the major contributing attributes are determined. Tableau is used to visualize these attributes in order to get business insights. Fig. 11 represents the architecture for this study.

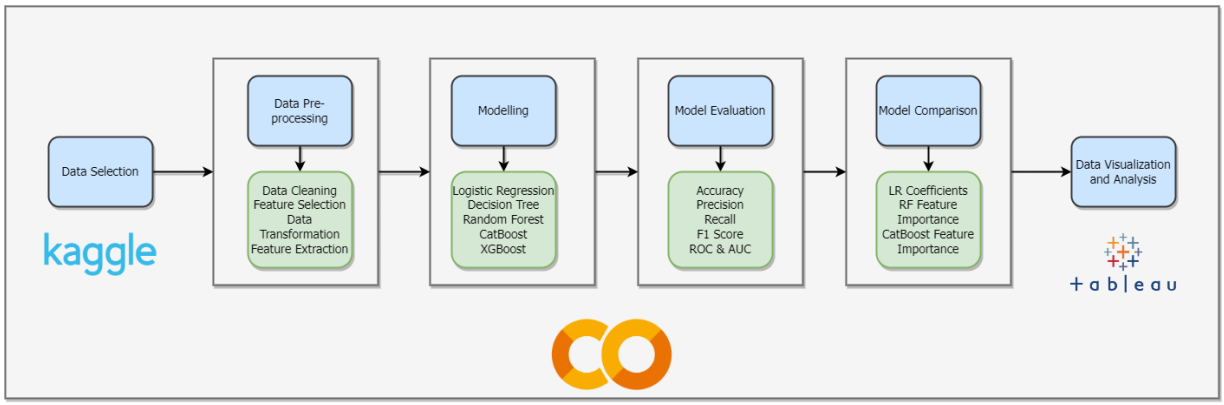

Fig. 11 Research Architecture

5 Implementation

5.1 Environmental Setup

This research study's implementation was conducted on a 64-bit computer with Windows Operating System and 8 GB RAM. In order to develop the models, python coding language was used. The code was executed using Google Colaboratory which runs on Google Cloud. The latest version (3.8.8) of Python was used for the coding. Tableau desktop has been used to create visualization and interpret data. Tableau version 2022.2 has been used for this.

5.2 Modelling Parameters

For this research, 5 different classification models namely, Logistic Regression, Decision Tree, Random Forest Classifier, CatBoost Classifier, and XGBoost Classifier, have been used with 'X' as all independent variables and 'Y' as the dependent variable 'Converted'. The data was split into train and test set at a split ratio of 7:3, leaving 6468 records for training and 2772 records for testing.

After split the train and test data, **Logistic Regression** model was applied to the dataset for which 'LogisticRegression' was imported from sklearn library and the model was evaluated. Next, coefficients were calculated and only those features with coefficient greater than zero were considered for the next fitting. Further, VIF (Variance Inflation Factor) was calculated for each variable using the 'variance_inflation_factor' from 'statsmodels.stats.outliers_influence' library. Variables with VIF > 5 were removed and the model was again applied and evaluated.

For **Decision Tree**, the 'DecisionTreeClassifier' package was imported from sklearn library. The model was applied on the original train and test sets and it was evaluated for the same.

In **Random Forest**, prediction was first done using 'RandomForestClassifier' from 'sklearn.ensemble' library. To increase the performance of the model, Hyperparameter Tuning was carried out using the method 'RandomizedSearchCV' from Scikit-Learn's library. It was observed that the best parameters were n estimators = 100, min samples split = 10, min samples leaf = 4, max features = auto, max_depth= 50 and bootstrap = True.

For **CatBoost Classifier**, the 'catboost' library was used. The model was applied on the original train and test sets and it was evaluated for the same.

Lastly, for **XGBoost Classifier**, the 'xgboost' library was used. The model was applied on the original train and test sets and it was evaluated for the same.

5.2.1 Comparison of Feature Sets from Different Models

In order to enhance the performance, a group of features was selected from the test and train sets. For Logistic Regression, coefficients were calculated, and features (total 55) with coefficients larger than zero were included for the subsequent fitting. To further improve the performance, VIF was calculated and features (total 49) with VIF <= 5 were selected for next fitting. For Random Forest, hyperparameter tuning was done to increase model performance. Best parameters were determined first and then were used to calculate the evaluation metrics which significantly increased the performance of the model. In case of other models, original test and train sets with 109 features were used to fit the model and evaluate it.

6 Results & Discussion

6.1 Exploratory Data Analysis Results

The skewness of the numerical variables was evaluated after cleaning the data. The square roots of the columns 'Total Time Spent on Website,' 'Page Views Per Visit,' and 'TotalVisits' were calculated to decrease the skewness. After calculating their square roots, the skewness was greatly decreased, and the plots resembled a normal distribution. Outlier analysis was also performed to find any outliers in the sample. Outliers were found and deleted from the data using this technique, which was repeated for all three numerical variables. The characteristics were normalized, which makes the variables more consistent with one another and allows the model to anticipate output more correctly. The variables were visualized in Tableau and dashboards were created which have been discussed in section 6.4 in the report.

6.2 Evaluation

This section will provide a complete overview of the study's results and key findings. It will throw light on all the evaluation methods considered for the classification problem. The evaluation is performed on the test set of the models.

The accuracy of these models shows that out of all the lead conversion predictions made by the model, what percentage was correctly predicted. In this case, both classes of the target variable are equally important so that marketing team can make focus on the future customers properly. After evaluating all the 5 models, the accuracy of Logistic Regression, Decision Tree, Random Forest Classifier, CatBoost Classifier, and XGBoost Classifier is 83.26%, 79.11%, 85.28%, 85.02% and 85.17% respectively. Higher accuracy means that both converted and non-converted leads will be predicted correctly which will help businesses in prioritising those leads which are predicted to be converted.

Precision is a measure that is important to evaluate as it shows the percentage of correctly converted leads out of all the converted ones. It is used when the cost of raising false alerts is high. Precision value should be high otherwise if leads do not actually convert as per prediction, company will be in loss after it has invested on them. The highest precision score is that of Decision Tree which is 88.54% and that of Random Forest is 84.95%.

Another evaluation metric called Recall of the model has been calculated. The recall value of CatBoost classifier is 77.62% and is the highest among all whereas Random Forest has a recall value of 75.48%. This shows that how correctly the models have predicted that the leads will turn into actual customers. The recall value shows the number of leads predicted to be converted among the leads that were actually converted.

F1 score has also been calculated enables the combination of precision and recall into a single metric that captures both properties. The highest F1 score is coming out to be of CatBoost classifier which is 80.11% whereas for Random Forest classifier is 79.94%.

The ROC curves for the best performing models are shown the Fig. 12. The steeper the top-left curve, the better the ROC AUC score, which indicates a successful model. It tells the probability of randomly selected converted lead is ranked higher than randomly selected non-converted lead. LR, RF classifier and CatBoost models have ROC curve pointing to the top left and AUC equal to 0.92, which indicates good model performance.

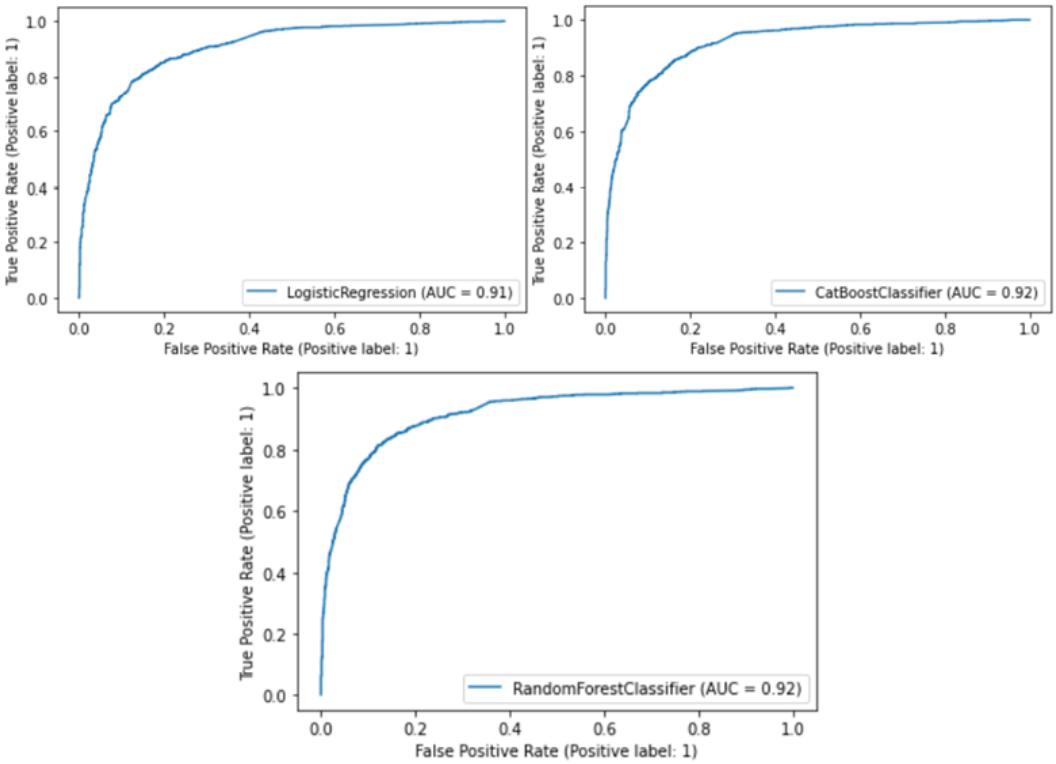

Fig. 12 Models with ROC and Highest AUC

	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	CatBoost Classifier	XGBoost Classifier
Accuracy	83.26%	79.11%	85.28%	85.02%	85.17%
Precision	82.29%	88.54%	84.95%	82.77%	83.77%
Recall	72.51%	53.11%	75.48%	77.62%	76.69%
F1 Score	77.09%	66.39%	79.94%	80.11%	80.07%
AUC	0.91	0.86	0.92	0.92	0.92

Table 1 Evaluation Metrics Summary

Table 1 shows the summarised version of performance of all the models combined. It can be said that the Logistic Regression, Random Forest Classifier, CatBoost and XGBoost models performed well but the former three models will be considered to extract the most important features since their evaluation metrics are equally good.

6.3 Feature Importance – Selecting Most Contributing Variables

After the Logistic Regression model was applied, the coefficients of independent variables were calculated. Next, feature importance was calculated for each independent variable after Random Forest model was applied. Lastly, feature importance of the CatBoost model was calculated after fitting the model. Then, the coefficient values of LR and feature importance of the variables from RF and CatBoost were compared and the variables with significant

values in all three categories were selected for visualization. Fig. 13 represents some of the significant variables with their coefficients from Logistic Regression, feature importance from Random Forest and the feature importance from Catboost classifier. Around 40 variables were selected for the visualization from this table.

Column Name	LR	RF	CB
Total Time Spent on Website sqrt	0.092355	0.229802964	21.05626586
Lead Profile_Potential Lead	1.118291	0.095859044	6.466877586
TotalVisits_sqrt	0.256254	0.031019249	5.76960577
Page Views Per Visit sqrt	-0.86958	0.026839799	5.065679469
Last Notable Activity_SMS Sent	0.49845	0.077509107	7.308473815
Last Activity_SMS Sent	0.773689	0.071143289	3.331298326
Lead Profile_Unknown	-0.645692	0.056954146	1.98377511
Lead Origin_Lead Add Form	2.095012	0.047344107	5.251991291
What is your current occupation_Working Professional	1.126254	0.049317306	5.667249766
What is your current occupation_Unemployed	-1.049653	0.033302391	2.248741053
Specialization_Unknown	-0.304139	0.017514749	1.84006364
City_Mumbai	0.074768	0.006152061	0.899161229
Last Activity_Email Opened	0.400982	0.012664677	1.270919075
Specialization_Finance Management	0.144341	0.003022652	0.433527428
Lead Source_Google	-0.324936	0.005596709	0.358910975
Specialization_Human Resource Management	-0.062161	0.003294562	0.765965754
Lead Source_Direct Traffic	-0.629205	0.008253404	0.908792646
Lead Origin_Landing Page Submission	-1.007614	0.010620747	1.377162212

Fig. 13 Feature Importances of Random Forest, CatBoost and Logistic Regression

6.4 Tableau Dashboards & Analysis

The variables from comparing feature importance were selected for the visualization. Using them, 3 Tableau dashboards have been created - Lead Background, Post-Prediction and Comparison between before and after prediction.

6.4.1 Leads' Background

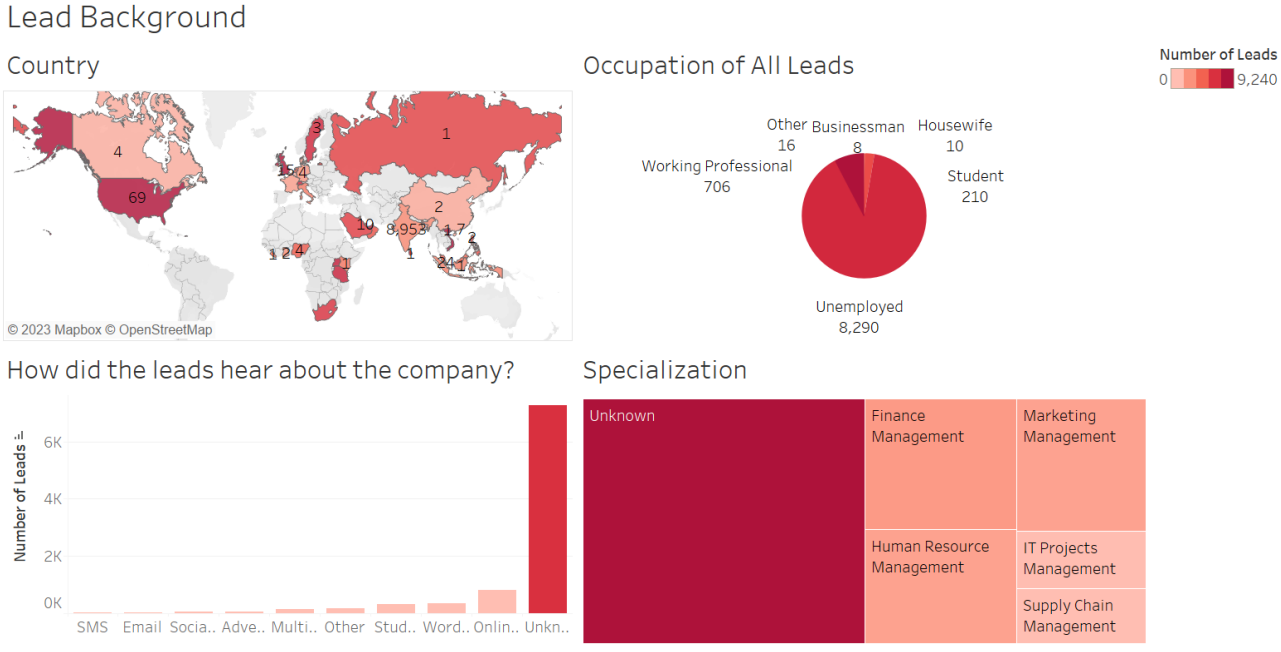

Fig. 14 Leads' Background Dashboard

Fig. 14 shows the background of the leads. It mainly depicts the countries where the leads are coming from, their occupation, how they heard about the company and the industry in which the leads have worked before i.e., their specialization. In the first section, which is a map, it can be clearly said that almost all the leads are from India which shows that people from India are the most interested in buying the company's products followed by the USA. The top-right chart depicts the different professional backgrounds of the leads. It says that most of the interested people are unemployed. Other than that, people who are either working professionals or students are interested in the company. The bottom left chart represents the places where the users heard about the company. Most of the known sources were online search and word of mouth. Finally, the bottom right chart shows the specialization of most of the leads which include finance management, HR management, marketing management, and so on. This dashboard basically shows the leads' background so that the businesses can analyse them and make strategies and the weak areas can become profitable.

6.4.2 Leads After Conversion

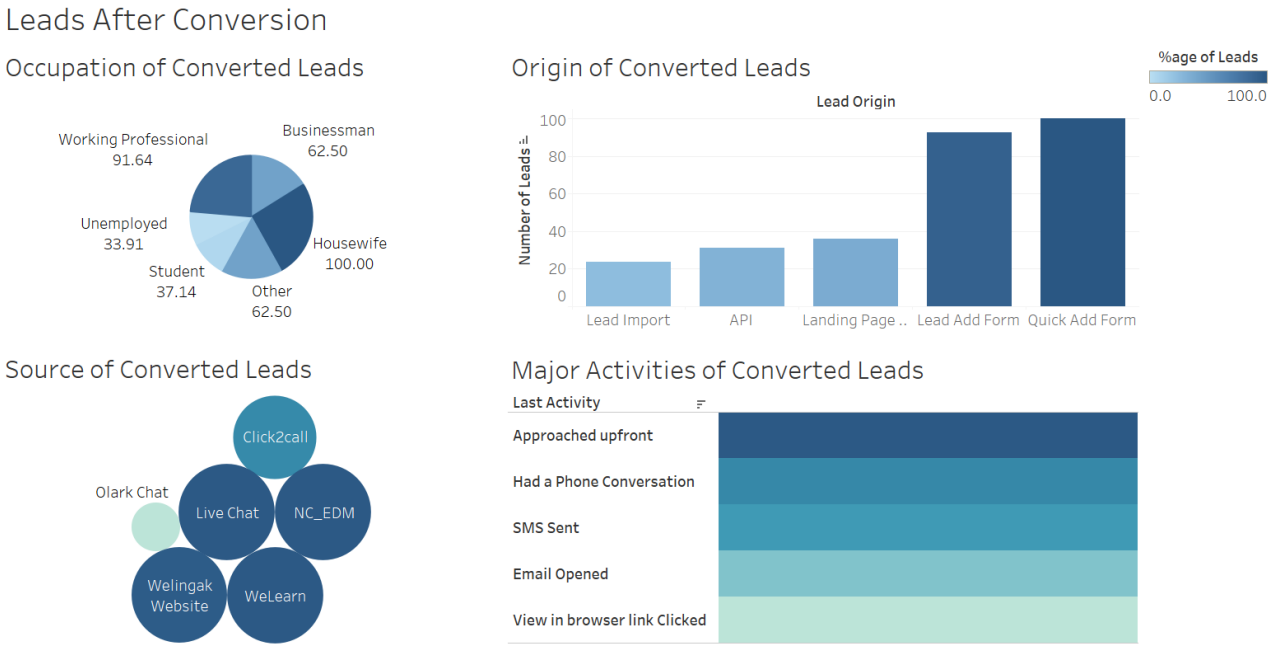

Fig. 15 After Conversion Dashboard

Fig. 15 shows the charts after the leads have been converted into customers. The top-left pie-chart shows the percentage of leads from different occupations that turned into customers. It can be said that all the housewives that were interested in the company's products bought them and became its customers. Additionally, the conversion rate of working professionals is also very high. Contrary to this, unemployed leads had the least conversion rate among all. The top-right chart shows the 5 different lead origins among which most converted leads come from the 'Quick Add Form', followed by 'Lead Add Form'. The bottom left chart represents the sources that the leads have come from and converted into customers. Here, WeLearn, Welingak Website and Live Chat have been the most contributing sources to a successful lead conversion. The bottom right chart shows the major activities that happened between the potential leads and the company's customer representatives. Most converted

leads had a direct conversation with the company's representatives. This is followed by phone conversation, SMS, Email and so on.

6.4.3 Comparison - Leads before and after Conversion

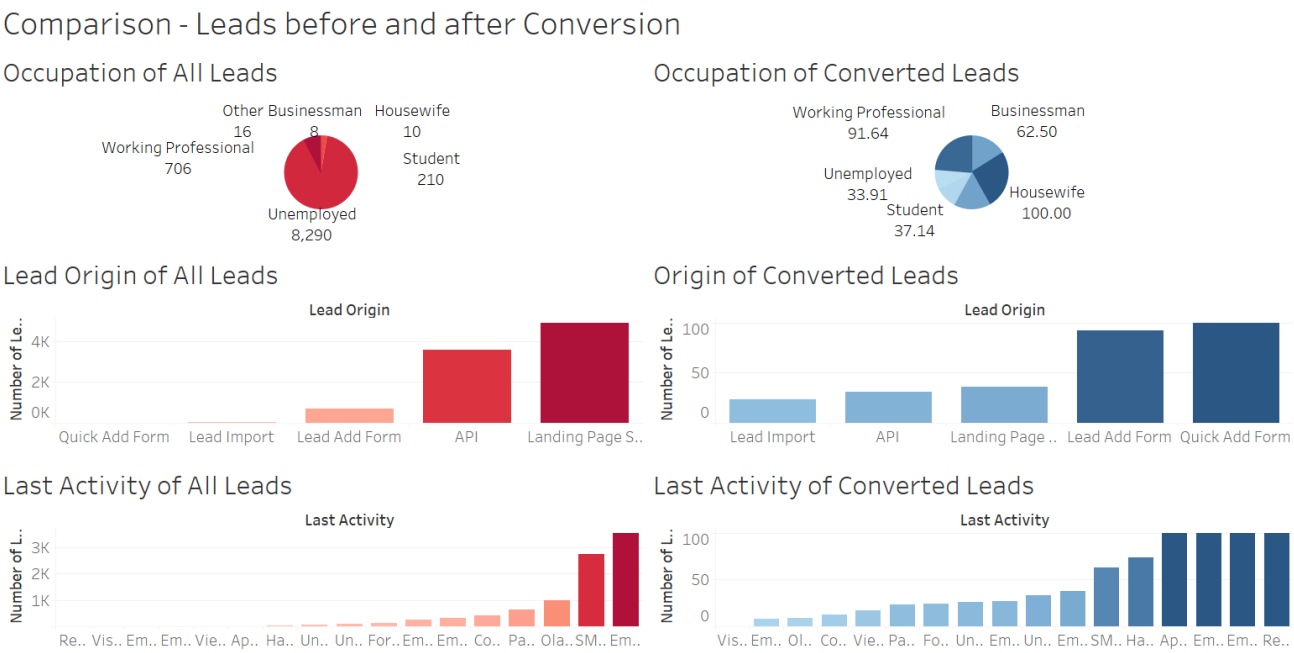

Fig. 16 Comparison Dashboard

This dashboard in Fig. 16 compares different features before and after the leads were converted into customers. It analyses which business have worked and which haven't so weak areas may be strengthened. In the first row, occupations of the leads and converted leads have been compared. The leads that were unemployed were 8290 in number as it can be seen on the left pie-chart whereas only 33.91% of the could actually convert into customers as shown by the right pie-chart. Similarly, only 37.14% of 210 students became customers. On the other hand, the number of leads who were either working professionals or housewives was less in comparison but their conversion rate was very high. In the second row, the origins of the leads and converted leads have been compared. The lease number of people were qualified as leads from 'Quick Add Form' as seen on the left bar graph but the greatest number of conversions (100% conversion rate) happened for the leads from the same origin. Similarly, the conversion rate of the leads from 'Lead Add Form' is 92.48% despite of the low number (718) of leads before conversion. Most number of people that were qualified as leads was through 'Landing Page Submission' which has a very low conversion rate i.e., about 36%. Similarly, 'API' has very less conversion rate which is about 31% out of 3580 qualified leads. The company can do an analysis on the lead origins and check the reason behind the origins that did not perform well. On the contrary, it can plan to get a greater number of leads through 'Quick Add Form' and 'Lead Add Form' as their conversion rate is very high compared to the others. In the third row, the last activity of the leads and converted leads have been compared. Most number of people that were qualified as leads had their last activity as 'Email Opened' as shown on the left bar graph while only 37.7% of them converted to customers. On the contrary, the activity of least number of leads was 'Resubscribed to Emails' but it had the highest conversion rate. People that were

‘Approached Upfront’ were converted into customers at a 100% rate. Also, people who ‘Had a Phone Conversation’ were converted at a rate of 73% and people who with ‘SMS Sent’ were converted at a rate of 62%. Also, all the leads were converted with activity as ‘Email Received’. These data demonstrate that leads directly contacted by customer representatives via phone, SMS, or Email were converted at a high rate.

6.5 Discussion

The study aimed to predict the lead conversion prediction and based upon that, it intended to draw business insights via visualizations. The raw data had some columns which had a large percentage of missing values and some had more than 90% of the spaces containing the same values. Due to this, the columns had to be dropped from the dataframe as they would have affected the modelling. In addition, the numerical variables were highly skewed due to which square root transform was applied. To make the data more efficient, techniques like normalization and outlier analysis were performed. The authors of the paper *(Jadli et al., 2022*) have applied different models and got decent accuracy but in order to utilize this model in a business organization, it is necessary to present it in a way that it is understandable to both technical and non-technical audience. This is why Tableau dashboards have been used to make them understand what led to the leads’ conversion and what did not and it will help in building marketing and sales strategies.

7 Business Recommendations

Based on the dashboards, several business strategies can be made for the sales and marketing teams to plan accordingly. Some of them are given below:

- In countries other than India, the company can enhance its products and build promotional strategies to gain more leads and customers where the lead count is less.
- Among all the leads, those who are students and unemployed have a very less conversion rate. This can either mean that they are interested in the products but the sales team is not nurturing the leads properly which is making the company lose its potential customers. Another reason could be that the leads do not find the products appropriate for themselves. In this case, the company needs to modify its products according to the leads’ necessities.
- Lead count of businessmen, housewives and students is lesser in comparison. The company can introduce products that are useful for so that more people from these backgrounds get interested in purchasing the products.
- Very less leads heard about the company through advertisements. The reason can either be that the adverts are not being shown on relevant places or they might need enhancements. Hence, the company needs to make a better marketing strategy for ads.
- Major activities include direct contact between the leads and customer representatives which shows that approaching the people upfront has the most impact on them which eventually makes them customers of the company.
- Leads coming by filling forms are getting converted into customers at a higher rate than those from API. Therefore, the company should plan in a way that more leads come thorough the forms.

8 Conclusion and Future Work

The purpose of this research was to determine the best performing model to predict lead conversion and also to visualize and analyse data with the help of Tableau dashboards. The idea of the research was to compute the conversion of the leads and find out the most contributing factors to it which can be visualized for the companies to improve their sales and marketing strategies. Out of all the models, features were selected from Logistic Regression, Random Forest and CatBoost which performed the equally best in terms of both precision and accuracy. Precision is vital in this case since organizations must invest time and money in leads that will become customers. Both groups of target variables are significant in corporate strategy planning, hence accuracy is also crucial. Once the prediction was done, significant variables from the data were identified in order to create visualizations in Tableau. Most influential factors were found to be origin, profile, occupation, time spent on website, number of visits and source of the leads. Analysis of Tableau dashboards can assist businesses in determining marketing performance. By comparing several factors, they may improve weak areas, make strategic strategies, and polish ideas.

For future work, a company can predict whether its own leads are converting into customers or not by providing a dataset with the same variables. This can be done by using their data as a test set and making the prediction with the help of the train set which is already available from this study. Also, other models like SVM, Naïve Bayes, kNN, etc., can be applied to the dataset and the best performing model can be determined. For visualization, tools such as PowerBI, framework like Streamlit or library like Gradio can be also used. Lastly, more lead nurturing-related attributes can be added to the dataset. This will help companies improve their sales and marketing strategies and, in turn, their revenue.

References

Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems*, *149*, 113622. https://doi.org/10.1016/j.dss.2021.113622

Czakon, J. (2022, July 21). *24 Evaluation Metrics for Binary Classification (And When to Use Them)*. Neptune.Ai. https://neptune.ai/blog/evaluation-metrics-binary-classification

Deng, W., Ling, X., Qi, Y., Tan, T., Manavoglu, E., & Zhang, Q. (2018). Ad Click Prediction in Sequence with Long Short-Term Memory Networks: An Externality-aware Model. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1065–1068. https://doi.org/10.1145/3209978.3210071

Dou, X. (2020). Online Purchase Behavior Prediction and Analysis Using Ensemble Learning. *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 532–536. https://doi.org/10.1109/ICCCBDA49378.2020.9095554

Feature importances. (n.d.). Retrieved December 13, 2022, from https://catboost.ai/docs/features/feature-importances-calculation

Gouveia, B., & Costa, O. (2022). Industry 4.0: Predicting lead conversion opportunities with machine learning in small and medium sized enterprises. *Procedia Computer Science*, *204*, 54–64. https://doi.org/10.1016/j.procs.2022.08.007

Gouveia, T. (2021, March 16). CatBoost Quickstart—ML Classification. *Nerd For Tech*. https://medium.com/nerd-for-tech/catboost-quickstart-ml-classification-f1d7fb70fea8

Hoelscher, J., & Mortimer, A. (2018). Using Tableau to visualize data and drive decision-making. *Journal of Accounting Education*, *44*, 49–59. https://doi.org/10.1016/j.jaccedu.2018.05.002

Hotz, N. (2018, September 10). What is CRISP DM? *Data Science Process Alliance*. https://www.datascience-pm.com/crisp-dm-2/

How to create a classification model using XGBoost in Python. (, 00:00). https://practicaldatascience.co.uk/machine-learning/how-to-create-a-classification-model-using-xgboost

Iranmanesh, S. H., Hamid, M., Bastan, M., & Nasiri, M. M. (2019). *Customer Churn Prediction Using Artificial Neural Network: An Analytical CRM Application*. 13.

Jadli, A., Hamim, M., Hain, M., & Hasbaoui, A. (2022). TOWARD A SMART LEAD SCORING SYSTEM USING MACHINE LEARNING. *Indian Journal of Computer Science and Engineering*, *13*(2), 433–443. https://doi.org/10.21817/indjcse/2022/v13i2/221302098

Jena, B. (2019). An Approach for Forecast Prediction in Data Analytics Field by Tableau Software. *International Journal of Information Engineering and Electronic Business*, *11*(1), 19–26. https://doi.org/10.5815/ijieeb.2019.01.03

Joshi, M. (2018, December 21). *What is Lead Funnel and How to Build One For Your Business*. LeadSquared. https://www.leadsquared.com/what-is-lead-funnel

Lead funnel definition, stages, and strategy. (2022, January 28). Zendesk. https://www.zendesk.com/blog/lead-funnel/

Lead Scoring Dataset. (n.d.). Retrieved December 7, 2022, from https://www.kaggle.com/datasets/662012015c226b1fa0444a4d7b38730f18ad6fe44fd3ca8057acec9dd17a211f

Lee, J., Jung, O., Lee, Y., Kim, O., & Park, C. (2021). A Comparison and Interpretation of Machine Learning Algorithm for the Prediction of Online Purchase Conversion. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(5), Article 5. https://doi.org/10.3390/jtaer16050083

Machine Learning Decision Tree Classification Algorithm—Javatpoint. (n.d.). Retrieved December 7, 2022, from https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Mostafa, F. B., & Hasan, E. (2021). *Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data* [Preprint]. Health Systems and Quality Improvement. https://doi.org/10.1101/2021.04.26.21256121

Nygård, R., & Mezei, J. (n.d.). *Automating Lead Scoring with Machine Learning: An Experimental Study*. 10.

Random Forest | Introduction to Random Forest Algorithm. (n.d.). Retrieved December 7, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Ronaghan, S. (2019, November 1). *The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark*. Medium. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Sklearn.linear_model.LogisticRegression. (n.d.). Scikit-Learn. Retrieved December 8, 2022, from https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

What is Logistic Regression? - Definition from SearchBusinessAnalytics. (n.d.). Business Analytics. Retrieved December 7, 2022, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

Yeo, J., Hwang, S., kim, sungchul, Koh, E., & Lipka, N. (2020). Conversion Prediction from Clickstream: Modeling Market Prediction and Customer Predictability. *IEEE Transactions on Knowledge and Data Engineering*, *32*(2), 246–259. https://doi.org/10.1109/TKDE.2018.2884467