Conversion Rate Ratio Predictor

Submitted in partial fulfillment of the requirements of the degree of

BACHELOR OF ENGINEERING

In

COMPUTER ENGINEERING

By

GROUP 18

1902002	**Gaurav Advani**
1902007	**Saathvik Ayyamolia**
1902057	**Shruti Jain**
1902146	**Drishti Sachwani**

Guide:

Dr. Seema Kolkur

(Associate Professor, Department of Computer Engineering, TSEC)

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Computer Engineering Department

Thadomal Shahani Engineering College

Bandra(w), Mumbai - 400 050

University of Mumbai

2022-23

CERTIFICATE

This is to certify that the Mini Project entitled **"Conversion Rate Ratio Predictor"** is a bonafide work of

Roll No.	Name
1902002	Gaurav Advani
1902007	Saathvik Ayyamolia
1902057	Shruti Jain
1902146	Drishti Sachwani

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering".**

Dr.Seema Kolkur
Guide

Dr.Tanuja Sarode
Head of Department

Dr.G.T.Thampi
Principal

Project Report Approval for B.E

This Mini Project entitled **"Conversion Rate Ratio Predictor"** by

Roll No.	Name
1902002	Gaurav Advani
1902007	Saathvik Ayyamolia
1902057	Shruti Jain
1902146	Drishti Sachwani

is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1 _____

(Internal Examiner Name & Sign)

2 _____

(External Examiner name & Sign)

Date:

Place:

Declaration

We declare that this written submission represents my ideas in my own words and where others 'ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will because for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1) _____

(Gaurav Advani- 1902002)

2) _____

(Saathvik Ayyamolia- 1902007)

3) _____

(Shruti Jain- 1902057)

4) _____

(Drishti Sachwani- 1902146)

Date:

Abstract

Conversion Rate Ratio Prediction (CRRP) is a methodology used by website owners and operators to improve the user experience and increase the likelihood of visitors taking specific actions, such as making a purchase, filling out a form, or subscribing to a newsletter. The goal of CRRP is to identify and understand the factors that contribute to the conversion of visitors into customers, and then to use that information to optimize the website and increase the conversion rate.

CRRP involves a variety of techniques and tools, including data analysis, user experience testing, and website optimization. It typically begins with gathering data about visitor behavior and interactions with the website, such as click-through rates, bounce rates, and time spent on site. This data is then analyzed to identify patterns and trends that can be used to improve the website and increase conversions.

Overall, CRRP is a powerful methodology that can help website owners and operators improve the effectiveness of their online presence, increase conversions, and ultimately grow their business. By gathering data, analyzing visitor behavior, and continually testing and optimizing the website, it is possible to create a user experience that is tailored to the needs and preferences of visitors, and that encourages them to take action and become customers.

TABLE OF CONTENTS

List of Figures

Figure No. Description Page No.

Chapter 1
Introduction

1.1 Introduction

The implementation of Conversion Rate Ratio Prediction (CRRP) as a marketing and sales methodology has become increasingly popular in recent years due to its effectiveness in helping businesses prioritize their leads and maximize their resources. CRRP involves the analysis of consumer behavior data to determine the likelihood of a lead becoming a customer. This score is calculated based on various factors such as the lead's interactions with a company's digital platforms, demographic data, and other relevant information.

To develop an effective CRRP system, businesses typically employ a combination of statistical analysis, machine learning algorithms, and domain expertise. Through the use of data analytics tools and techniques, businesses can uncover insights that can help them make more informed decisions about their marketing and sales strategies.

Once a CRRP system has been implemented, businesses can use it to segment their leads into different categories based on their likelihood of becoming a customer. The highest-scoring leads are often given the most attention and may be contacted directly by the sales team, while leads with lower scores are typically treated with follow-up marketing content to encourage them to take further action.

The effectiveness of CRRP as a marketing and sales methodology lies in its ability to help businesses make more informed decisions about how to allocate their resources. By prioritizing leads based on their likelihood of becoming a customer, businesses can focus their efforts on the leads that are most likely to convert, while minimizing their efforts on leads that are less likely to convert.

However, it is important to note that CRRP is not a one-size-fits-all solution, and its effectiveness may vary depending on the industry and the specific circumstances of the business. Therefore, it is essential for businesses to carefully evaluate their needs and develop a CRRP system that is tailored to their specific requirements[1].

In conclusion, CRRP is a valuable tool for businesses looking to optimize their marketing and sales strategies by leveraging consumer data to make more informed decisions about lead

prioritization. With the right approach and the proper tools, businesses can use CRRP to gain a competitive edge in their industry and achieve their marketing and sales goals.

1.2 Problem Statement & Objectives

Problem Statement:

One of the main challenges that businesses face in online marketing is knowing how to optimize their website and marketing campaigns to attract the right customers and drive conversions. Without this information, businesses may waste valuable resources on ineffective marketing strategies, resulting in poor conversion rates and low sales. However, with the help of CRRP, businesses can gain a better understanding of their audience and identify the most effective ways to target and convert them.

CRRP works by analyzing consumer data to identify patterns and associations in their behavior and characteristics. This information is then used to create a model that predicts the likelihood of a lead converting into a customer. This prediction can be used to identify areas for optimization in the marketing funnel, such as the website design, messaging, and targeting strategies.

Objectives:

The proposed approach for designing a Conversion Rate Predictor (CRRP) is a promising solution for businesses looking to improve their lead generation and sales performance. The approach involves the collection and processing of real customer data, which includes various demographics, customer and lead identity information. This data is then used to predict the conversion rate from website visitors to sign-ups on the X Education website.

1. The first objective in the approach is to collect data and involve data processing to make the obtained data useful to its full potential. This involves gathering data from various sources, such as website analytics, customer surveys, and social media. The data is then processed to remove any irrelevant or duplicate information and to ensure that the data is accurate and reliable.

2. The next objective is to involve the utilization of the data pertaining to the customers signing up on the X Education website. This data includes information such as the customer's age, gender, location, interests, and other relevant characteristics. By analyzing this data, businesses can gain a better understanding of their audience and develop targeted marketing campaigns to attract and convert the right customers.
3. Once the data has been collected and processed, the CRRP can then be used to predict the conversion rate from website visitors to sign-ups on the X Education website. This prediction is based on the patterns and associations identified in the data, as well as the historical performance of the website.
4. After analyzing the ratios suggested by the CRRP, different methodologies can be suggested to optimize the results throughout. For example, if the CRRP suggests that certain website pages or elements are hindering the conversion process, businesses can make adjustments to optimize those areas. Similarly, if the CRRP suggests that certain demographic or customer segments are more likely to convert, businesses can adjust their marketing messaging and targeting to focus on those groups.

1.3 Scope

To achieve the objective, the system used a dataset collected from an online education platform. The dataset consisted of information on search queries, impressions, clicks, conversions, and other user-related information such as demographics and geographic data. The various feature ranking and weighting schemes were then applied to identify the most important features that influence the conversion rate.

A logistic regression model was used to predict the conversion rate. The logistic regression model is a widely used classification method that is suitable for binary classification problems like this. The model was trained and tested using the collected dataset and evaluated its performance using various evaluation metrics such as accuracy, precision, recall, and F1-score.

The results showed that the model achieved an accuracy of 84% in predicting the conversion rate. The most important features that influence the conversion rate were found to be the

position of the ad, the time of the day, and the user's location. The authors then applied the feature weights to the features to find out the actual importance of each feature as well as to compare ranking and weighting methods.

The system’s contribution is its novel approach to building a predictive model of conversion rates by explicitly using feature weights. The use of feature weights is commonly overlooked by similar studies, making this thesis a valuable addition to the existing literature. The thesis also provides insights into the most important features that influence the conversion rate, which can help marketers optimize their advertising campaigns[2].

This model successfully built a predictive model of conversion rates using a logistic regression model and applied various feature ranking and weighting schemes to identify the most important features that influence the conversion rate. The explicit use of feature weights makes this model a valuable addition to the existing literature, and the insights provided into the most important features can help marketers optimize their campaigns.

Chapter 2
Review of Literature

2.1 Domain Explanation

With the increase in the popularity of e-commerce, there is a vast amount of consumer behavioral data available. Analyzing this data to understand what leads to the final consumption behavior is of great importance in improving total sales in e-commerce. Personalized search is a popular way to enhance user experience and encourage user consumption. By providing tailored recommendations and ranking the results based on conversion rate prediction, websites can increase the chances of a user making a purchase.

In this context, this paper proposes a method of combining personalized search data with a choice model to predict a user's purchase intention after clicking on an item on an e-commerce platform. The model predicts whether a user will buy an item they have clicked on, by following a sequential pattern of impression, click, and conversion.

The company in question markets its courses on various websites and search engines such as Google. When users land on the website, they may browse courses, fill up a form, or watch videos. If a user provides their email address or phone number, they are classified as a lead. The company also acquires leads through past referrals. The sales team then contacts these leads through calls or emails, and some leads get converted, while most do not. The typical lead conversion rate at X education is around 30%.

The proposed method aims to predict the post-click conversion rate of advertised search results by analyzing various features that influence a customer's online behavior. The model uses a set of related features to train predictors that can predict the post-click conversion rate. The contribution of features in the prediction task is studied by applying various feature ranking and weighting schemes. The weights are then applied to the features to determine their actual importance and compare the ranking and weighting methods. The explicit use of feature weights is often overlooked in similar studies. However, it is essential to prepare the data correctly, given the type and source of input data.

The model's goal is to predict the conversion rate of website visitors to sign-ups, i.e., future leads, using customer data, including various demographics, customer, and lead identity. Once the model predicts the conversion rate, it suggests different methodologies to optimize the results throughout.

The proposed method has significant potential to help companies optimize their websites and improve their sales. By predicting users' purchase intention, companies can provide personalized recommendations and tailor their marketing strategies accordingly. This can lead to increased customer engagement and loyalty, ultimately leading to higher sales and revenue for the company.

2.2 Review of Existing Systems

Conversion Rate Ratio Prediction is an automated method that uses machine learning algorithms to assign scores to leads based on their potential to convert into customers. Unlike the conventional lead scoring method where attribute scores are assigned manually, conversion rate ratio prediction learns potential prospects' importance scores from the available data.

The process of conversion rate ratio prediction involves constructing an automated model that can take any attribute as input. However, integrating data from various sources is a significant challenge in this method. This challenge can be addressed by manually connecting different datasets or using platforms like Google Analytics that enable integration with certain CRM platforms like HubSpot.

HubSpot is a fully-featured digital marketing platform that includes a CRM service. The CRM-generated data is the most popular data source for predictive lead scoring, consisting of features related to both personal and behavioral attributes, along with historical sales data[3].

Conversion Rate Ratio Prediction models are developed by feeding large amounts of historical data into machine learning algorithms, which identify patterns and relationships between the available data and the outcome of whether a lead will convert to a customer or not. These patterns and relationships are then used to assign scores to new leads, indicating their likelihood

of converting to a customer[13].

The benefits of conversion rate ratio prediction are numerous. For one, it automates the lead scoring process, reducing the potential for human error and ensuring consistency in lead scoring. It also enables sales and marketing teams to prioritize their efforts on the leads that are most likely to convert into customers, saving time and resources. Moreover, it helps companies identify which factors contribute most significantly to lead conversion, allowing them to adjust their strategies accordingly.

In summary, Conversion Rate Ratio Prediction is an automated method that uses machine learning algorithms to assign scores to leads based on their potential to convert into customers. The method learns from historical data and enables companies to prioritize their efforts on the leads most likely to convert, saving time and resources while improving overall conversion rates[11].

2.3 Limitations of Existing System

While conversion rate prediction methods have come a long way, there are still several limitations that need to be addressed. Here are some of the main ones:

1. Limited data: Many businesses struggle to gather enough data to accurately predict conversion rates. This is particularly true for new or small businesses that don't have a large customer base or a lot of historical data.
2. Incomplete data: Even when businesses do have a lot of data, it may not be complete or accurate. For example, customer data may be missing important information like email addresses or phone numbers.
3. Lack of context: Predictive models are only as good as the data they are trained on. If a model is trained on data from one specific period or market, it may not be effective in other contexts.
4. Over-reliance on historical data: Many predictive models are based solely on historical data, which can be limiting. If customer behavior changes or if there are external factors

that impact conversion rates (such as a pandemic), historical data may not accurately predict future outcomes.

5. Difficulty in choosing the right features: Choosing the right features to include in a predictive model can be challenging. Businesses need to decide which data points are most relevant and how to weight them appropriately[4].
6. Limited visibility into the customer journey: Many conversion rate prediction methods only consider data from a single touchpoint, such as a website visit or email open. However, customers typically engage with businesses across multiple channels and touchpoints, which can make it difficult to accurately predict conversion rates.
7. Lack of personalization: Many websites still use a one-size-fits-all approach to conversion rate optimization, where all users are presented with the same page. However, users have different preferences and behaviors, and personalizing the experience can lead to higher conversion rates.
8. Limited scope: Conversion rate optimization usually focuses on optimizing a single webpage or user flow. However, users interact with websites and businesses in many ways, and optimizing one page may not fully address their needs and desires.
9. Short-term focus: Many conversion rate optimization techniques focus on short-term gains, such as increasing clicks or sign-ups, rather than long-term success, such as improving customer loyalty and retention.
10. Limited human expertise: While machine learning and other automated techniques can help optimize conversion rates, they lack the human expertise to fully understand user behavior and preferences. This limitation can lead to missed opportunities for optimization.
11. Complexity: Conversion rate optimization can be complex, involving many variables and data sources. This complexity can make it difficult to fully understand the impact of changes and optimize accordingly.

Chapter 3
Proposed System

3.1 Analysis

Conversion rate ratio prediction is a critical area of focus for businesses looking to optimize their sales and marketing efforts. By predicting the likelihood that a customer will take a desired action, such as making a purchase, businesses can target their efforts more effectively and improve their overall conversion rates. In this analysis, we'll take a closer look at the various approaches and tools used for conversion rate ratio prediction, including machine learning algorithms, data preprocessing, feature selection, and model selection.

Machine Learning Algorithms for Conversion Rate Ratio Prediction:

Machine learning algorithms are a critical tool for conversion rate ratio prediction. These algorithms analyze large amounts of data and identify patterns and trends that can help predict future conversion rates. Some common machine learning algorithms used for conversion rate ratio prediction include logistic regression, decision trees, and neural networks.

Logistic regression is a statistical model that is used to analyze relationships between a dependent variable (in this case, conversion rate) and one or more independent variables (such as website traffic, customer behavior, and demographic data). Logistic regression is a binary classification model, which means it predicts a binary output (1 or 0) based on the input features.

Decision trees are another popular machine learning algorithm for conversion rate ratio prediction. A decision tree is a hierarchical model that uses a tree-like structure to represent decisions and their possible consequences. The tree consists of nodes that represent decisions and branches that represent possible outcomes. Decision trees are particularly useful when the relationship between the input features and the target variable is complex and nonlinear[8].

Neural networks are a more complex machine learning algorithm that is designed to model complex relationships between input features and the target variable. Neural networks consist of layers of interconnected nodes (neurons) that are designed to simulate the way the human brain processes information. They are particularly useful when the relationship between the input features and the target variable is highly complex and nonlinear.

Data Preprocessing for Conversion Rate Ratio Prediction:

Data preprocessing is a critical step in the conversion rate ratio prediction process. Preprocessing involves cleaning and transforming data, removing outliers, and handling missing data. The quality of the data used for analysis is critical to the success of conversion rate ratio prediction. If the data is incomplete, inaccurate, or biased, it can lead to inaccurate predictions.

Cleaning and transforming data involve removing irrelevant data points, transforming data into a format that is suitable for analysis, and converting categorical data into numerical data. Removing outliers involves identifying data points that are significantly different from the rest of the data and either removing them or treating them separately. Handling missing data involves identifying missing data points and either imputing them (filling in the missing values) or removing them.

Feature Selection for Conversion Rate Ratio Prediction:

Feature selection is the process of selecting the most relevant input features for conversion rate ratio prediction. The choice of features used for analysis can have a significant impact on the accuracy of conversion rate ratio prediction. It's important to choose features that are relevant to the target variable (conversion rate) and that have a strong correlation with it.

There are several approaches to feature selection, including filter methods, wrapper methods, and embedded methods. Filter methods involve selecting features based on their statistical significance or correlation with the target variable. Wrapper methods involve selecting features based on their impact on the performance of a machine learning model. Embedded methods involve selecting features as part of the training process for a machine learning model.

Model Selection for Conversion Rate Ratio Prediction:

Model selection is the process of choosing the most appropriate machine learning model for conversion rate ratio prediction. Different machine learning models have different strengths and weaknesses, and it's important to choose the one that is best suited for the specific problem at hand[10].

In conclusion, conversion rate ratio prediction is a critical area of focus for businesses looking to optimize their sales and marketing efforts. Machine learning algorithms are a critical tool for predicting conversion rates, and there are several approaches to data preprocessing, feature selection, and model selection that can improve the accuracy of predictions. When choosing a machine learning model, it's important to consider the strengths and weaknesses of each model and select the one that is best suited for the specific problem at hand. With the right approach, businesses can optimize their sales and marketing efforts and improve their overall conversion rates.

3.2 Design Details

Conversion rate optimization is an essential process for any business that has an online presence. The goal is to increase the percentage of website visitors who complete a desired action, such as making a purchase or filling out a contact form. To achieve this, businesses must predict and optimize the conversion rate of their website continually. In this article, we will discuss the design details with flowchart for Conversion Rate prediction and optimization.

1. Collect Data: The first step in predicting and optimizing conversion rate is collecting relevant data. This data can include website traffic, user demographics, and behavior. Gathering this data can be achieved by using various tools such as Google Analytics, heat maps, and surveys. The data collected should be comprehensive and accurate as it forms the basis for the subsequent analysis.

2. Data Preprocessing: Once data is collected, it must be cleaned and preprocessed. This involves removing any errors or inconsistencies, dealing with missing values, and ensuring data consistency across different sources. Preprocessing is a critical step as it ensures that the data is ready for analysis.

3. Feature Extraction: After preprocessing, the next step is to extract important features that will help predict the conversion rate. The features extracted can include click-through rates, time spent on the website, and page views. Feature extraction is important as it ensures that the right data is fed into the predictive model.

4. Build Model: The fourth step is to build a predictive model that can predict the conversion rate. The model can be built using various machine learning algorithms, such as linear regression, logistic regression, or decision trees. The algorithm chosen depends on the type of data and the specific problem being addressed. The model's goal is to provide insight into the factors that influence the conversion rate, such as website layout, copy, and imagery.

5. Model Validation: After building the model, it's necessary to validate it by splitting the data into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate the model's performance. The model's accuracy and precision can be measured using metrics such as mean squared error or root mean squared error. Model validation is important as it ensures that the model is reliable and can be used to make accurate predictions.

6. Optimize Conversion Rate: Based on the model's predictions, businesses can optimize their website or landing page to improve the conversion rate. This optimization can include testing various design elements, such as call-to-action buttons, color schemes, and layout. Businesses can use A/B testing to compare different versions of the website to determine which design elements work best.

7. Monitor and Refine: Once the website is optimized, it's essential to continually monitor it and refine the optimization strategies to improve the conversion rate further. Monitoring can involve using tools such as heat maps and analytics to track user behavior and website performance. Refining the optimization strategy can involve testing different design elements, such as headlines, images, and product descriptions.

In conclusion, Conversion Rate prediction and optimization is a continuous process that requires businesses to collect and preprocess data, extract features, build a predictive model, optimize the website, monitor and refine the optimization strategies, and continually evaluate the model's performance. By following these steps and using the right tools and techniques, businesses can improve their website's conversion rate, increase revenue, and improve customer satisfaction.

3.3 Methodology

The methodology for building a conversion rate predictor can be broken down into several steps. The first step is to find a suitable dataset that contains the necessary information required to predict the conversion rate accurately. This dataset can be obtained from a variety of sources, including Customer Relationship Management (CRM) software, web analytics tools, and third-party data providers. Once the dataset is obtained, it needs to be processed using different methods to clean and homogenize the data. This involves removing any missing or erroneous data, standardizing the data format, and applying a naming convention to make the data consistent and easier to work with.

Once the dataset is prepared, it is split into two parts: a training dataset and a testing dataset. The training dataset is used to build the predictive model using logistic regression. Logistic regression is a common method used for binary classification problems like predicting conversion rates, where the output variable has two possible outcomes (in this case, converted or not converted).

After the model is trained, it is important to evaluate its performance on new, unseen data. This is where the testing dataset comes in. The model is applied to the testing dataset to see how well it can predict conversion rates for this new data. One way to evaluate the model's performance is to create a Receiver Operating Characteristic (ROC) curve. The ROC curve helps to visualize the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity). A higher Area Under the Curve (AUC) indicates better performance of the model.

Another way to evaluate the model's performance is by creating a Confusion Matrix. The Confusion Matrix provides a detailed breakdown of the model's predictions. It shows the number of true positives (correctly identified converted customers), true negatives (correctly identified non-converted customers), false positives (incorrectly identified as converted), and false negatives (incorrectly identified as non-converted).

After evaluating the model's performance, it is important to use the results to improve the conversion rate. One approach to this is by analyzing the reliability of the model's predictions. This involves understanding the factors that are contributing to the conversion rate and how those factors can be improved. By identifying the factors that are driving conversion rates and optimizing those factors, the conversion rate can be improved.

Finally, the optimization techniques and suggestions can be presented to the user to help them make informed decisions and improve their conversion rates. This could include suggestions such as improving the website's design, targeting specific customer demographics, or optimizing the sales process. By providing actionable insights, the conversion rate predictor can help businesses make data-driven decisions and ultimately improve their bottom line. The methodology proposed can be seen in figure 3.1[14].

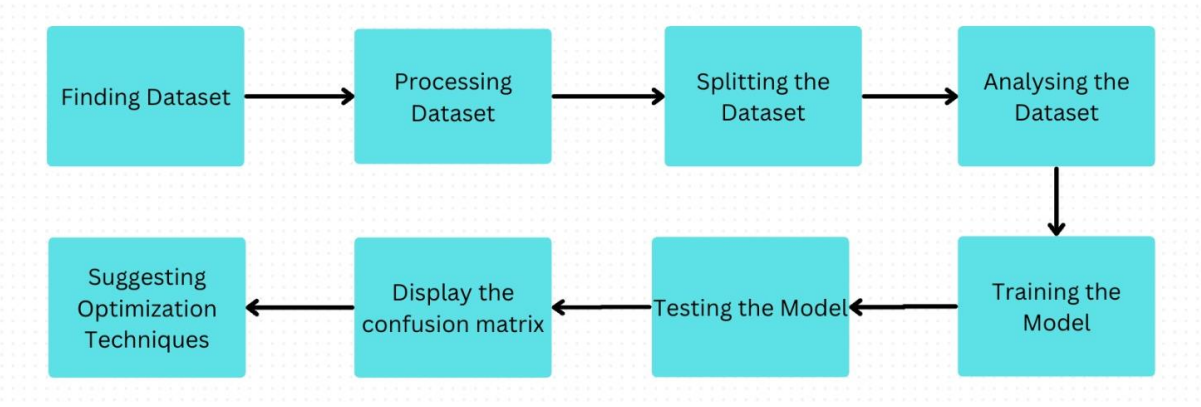

Fig 3.1 Methodology Proposed

Chapter 4

Implementation Details

4.1 Experimental Setup

Dataset Description

The features mentioned above can be used to create a comprehensive customer profile that takes into account their demographics, interests, and behavior patterns. By understanding the customer's preferences and motivations, businesses can tailor their marketing strategies to increase the likelihood of a conversion.

User profiles can be particularly useful in identifying patterns in customer behaviour. For instance, if a particular city has a higher conversion rate than others, this may indicate that there is a specific demographic in that area that is more likely to make a purchase. Similarly, by analyzing the last activity of a user, businesses can understand how frequently a customer interacts with their brand and adjust their marketing efforts accordingly.

The context page and context timestamp features can also provide valuable insights into customer behavior. For example, if a customer clicks on a product while browsing a specific category page, this may indicate a higher level of interest in that product than if they clicked on it while on the homepage. Understanding these context-based patterns can help businesses tailor their marketing efforts to the specific needs and interests of each customer.

In addition to these key features, there are other factors that can impact conversion rates, such as pricing, product quality, and shipping times. By analyzing these factors alongside user and context-based features, businesses can gain a more comprehensive understanding of their customers' behavior and optimize their marketing strategies accordingly.

Overall, the success of a conversion rate prediction algorithm depends on the careful selection and analysis of a range of features that can influence customer behavior. By understanding the customer's preferences and motivations, businesses can create targeted marketing strategies that are more likely to result in a conversion.

There are 9240 tuples in the dataset, out of which, 7392 have been used for training the model

and the rest 1848 have been used for testing the model. There are 37 columns, that is, different parameters for evaluation, as it can be seen in figure 4.1.

Basic raw features in the project dataset are:

1. Prospect ID
2. Lead Number
3. Lead Origin
4. Lead Source
5. Do Not Email
6. Do Not Call
7. Converted
8. Total Visits
9. Total Time Spent on Website
10. Page Views Per Visit
11. Last Activity
12. Country
13. Specialization
14. How did you hear about X Education?
15. What is your current occupation?
16. What matters most to you in choosing a course?
17. Search
18. Magazine
19. Newspaper Article
20. X Education Forums
21. Newspaper
22. Digital Advertisements
23. Through Recommendations
24. Receive More Updates About Our Courses
25. Tags
26. Lead Quality
27. Update me on Supply Chain Content
28. Get updates on DM Content
29. Lead Profile

30. City
31. Asymmetrique Activity Index
32. Asymmetrique Profile Index
33. Asymmetrique Activity Score
34. Asymmetrique Profile Score
35. I agree to pay the amount through cheque
36. A free copy of Mastering The Interview
37. Last Notable Activity

Prospect ID	7927b2df-8bba-4d29-b9a2-b6e0beafe620	2a272436-5132-4136-86fa-dcc88c88f482	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	3256f628-e534-4826-9d63-4a8b88782852
Lead Number	660737	660728	660727	660719	660681
Lead Origin	API	API	Landing Page Submission	Landing Page Submission	Landing Page Submission
Lead Source	Olark Chat	Organic Search	Direct Traffic	Direct Traffic	Google
Do Not Email	No	No	No	No	No
Do Not Call	No	No	No	No	No
Converted	0	0	1	0	1
TotalVisits	0.0	5.0	2.0	1.0	2.0
Total Time Spent on Website	0	674	1532	305	1428
Page Views Per Visit	0.0	2.5	2.0	1.0	1.0
Last Activity	Page Visited on Website	Email Opened	Email Opened	Unreachable	Converted to Lead
Country	NaN	India	India	India	India
Specialization	Select	Select	Business Administration	Media and Advertising	Select
How did you hear about X Education	Select	Select	Select	Word Of Mouth	Other
What is your current occupation	Unemployed	Unemployed	Student	Unemployed	Unemployed
What matters most to you in choosing a course	Better Career Prospects	Better Career Prospects	Better Career Prospects	Better Career Prospects	Better Career Prospects
Search	No	No	No	No	No
Magazine	No	No	No	No	No
Newspaper Article	No	No	No	No	No

Fig 4.1 Dataset

4.2 Software and Hardware Setup

Hardware Setup:

1. Operating system: The recommended operating systems for this project are Linux (specifically Ubuntu 16.04 to 17.10) or Windows (7 to 10). These operating systems are commonly used in the development community and have good support for Python and its related packages. Other operating systems like macOS can also be used, but may require additional setup or configuration

2. RAM: RAM stands for Random Access Memory and is a type of computer memory used for storing and accessing data quickly. The recommended RAM for this project is
3. 2GB, but 4GB is preferable as it allows for smoother and faster performance. Having enough RAM is important for running large datasets and complex algorithms, which are common in data science projects.
4. Python 3.6: Python is a popular programming language used in data science and machine learning. The recommended version for this project is Python 3.6, as it has good support for the required packages and features. Python can be installed through various methods depending on the operating system being used, including through package managers like apt-get (for Ubuntu) or through an installer from the official website (for Windows).
5. Related packages: Python has a rich ecosystem of packages and libraries for data science and machine learning. For this project, specific packages and libraries will be required, such as pandas (for data manipulation), scikit-learn (for machine learning algorithms), and matplotlib (for data visualization). These packages can be installed using Python's package manager pip, which is included with most Python installations.

Software Setup:

Matplotlib:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

Seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see some of the things that you can do with seaborn, and then check out the tutorials or API reference to find out how.

To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stack overflow, which has a dedicated channel for seaborn.

NumPy:

It provides:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities and much more

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Logistic Regression Algorithm:

Logistic regression aims to solve classification problems. It does this by predicting categorical outcomes, unlike linear regression that predicts a continuous outcome.

In the simplest case there are two outcomes, which is called binomial, an example of which is predicting if a tumor is malignant or benign. Other cases have more than two outcomes to classify, in this case it is called multinomial. A common example for multinomial logistic regression would be predicting the class of an iris flower between 3 different species.

Pandas:

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open-source data analysis/manipulation tool available in any language. It is already well on its way toward this goal.

Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data need not be labeled at all to be placed into a pandas data structure

Chapter 5
Results and Discussion

5.1 Performance Evaluation Parameters

Confusion Matrix:

A confusion matrix is a table that is commonly used to evaluate the performance of a classification model. It is a useful tool that helps us visualize how well our model is doing by showing the number of correct and incorrect predictions made by the model. It is also known as an error matrix, contingency table, or classification matrix.

A confusion matrix consists of four different metrics: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These metrics are derived from the outcomes of the classification model.

True Positive (TP): The number of positive instances that are correctly classified as positive.

False Positive (FP): The number of negative instances that are incorrectly classified as positive.

True Negative (TN): The number of negative instances that are correctly classified as negative.

False Negative (FN): The number of positive instances that are incorrectly classified as negative.

ROC Curve:

The Receiver Operating Characteristic curve is a graphical representation of the performance of a binary classifier that varies its decision threshold. It is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold settings.

The TPR is also known as sensitivity or recall, which is the proportion of actual positive cases that are correctly identified by the classifier. The FPR is the proportion of actual negative cases that are incorrectly identified as positive by the classifier.

In a ROC curve, the diagonal line represents the classifier with no predictive power, where the TPR and FPR are equal to each other (50/50 chance). A good classifier should have a curve that is as close to the top-left corner of the graph as possible, indicating high TPR and low FPR.

The ROC curve is useful for visualizing and comparing the performance of different classifiers, as well as for selecting the optimal decision threshold for a particular classification problem. It can also be used to analyze the trade-off between TPR and FPR, depending on the specific needs and goals of the classification task.

Area Under Curve (AUC):

AUC stands for "Area Under the Curve" and is a metric used to evaluate the performance of binary classification models. It represents the area under the Receiver Operating Characteristic (ROC) curve, which is a plot of the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds.

The AUC is a useful metric because it provides a single number that summarizes the performance of the model across all possible classification thresholds. The value of the AUC ranges from 0 to 1, with a higher value indicating better performance. A model with an AUC of 1 is perfect, meaning that it has a TPR of 1 and an FPR of 0 across all classification thresholds.

A model with an AUC of 0.5 is no better than random guessing, as it has the same probability of correctly identifying positive and negative examples as randomly guessing. A model with an AUC between 0.5 and 1 is better than random guessing, with higher values indicating better performance.

The AUC is particularly useful when dealing with imbalanced datasets, where the number of positive examples is much smaller than the number of negative examples. In such cases, a model that always predicts the majority class can achieve a high accuracy, but may have poor performance in terms of identifying the minority class. The AUC provides a more balanced evaluation of the model's performance, as it considers both the TPR and FPR across all classification thresholds.

5.2 Implementation Results and Discussion

The results of our project are:

1. Logistic Regression-

Logistic regression is a commonly used statistical method for analyzing a dataset in order to predict a binary outcome, such as whether or not a lead will convert. In this case, the model was trained on a dataset of features related to advertising items, user profiles, and online shop profiles, and used to predict the conversion rate of leads in a testing dataset[5].

The model was able to correctly predict the conversion status of 656 leads in the testing dataset, giving an overall accuracy of 84%. This indicates that the logistic regression model was effective at identifying the features that are most strongly correlated with conversion, and using them to make accurate predictions about which leads are most likely to convert.

However, it's worth noting that accuracy alone is not always the best metric for evaluating the performance of a logistic regression model. For example, if the dataset is imbalanced (meaning that there are far more instances of one class than the other), the model may be biased towards predicting the more common outcome, leading to a high accuracy score but poor overall performance[6].

Therefore, it's important to evaluate the performance of the logistic regression model using a range of metrics, such as the ROC curve and AUC score, as discussed earlier. This can help to provide a more nuanced understanding of how well the model is performing, and identify areas where improvements can be made.

Threshold	Accuracy
0.0	0.371
0.1	0.671
0.2	0.784
0.3	0.822
0.4	0.834
0.5	0.824
0.6	0.811
0.7	0.794
0.8	0.771
0.9	0.728
1.0	0.629

Fig 5.1 threshold vs accuracy table of logistic regression

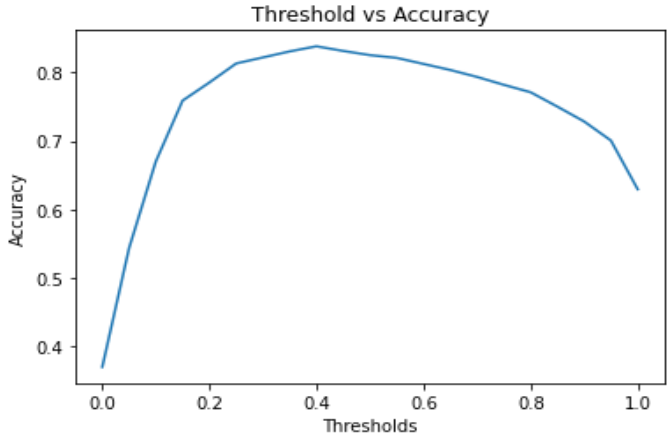

Fig 5.2 threshold vs accuracy curve of logistic regression

Figure 5.2 is the curve of threshold vs accuracy based on the table in figure 5.1. Figure 5.1 is the table for the relevancy of lead conversion. The model has been used to predict the conversion of leads in the testing dataset. The model predicted that out of the 1848 leads, 656 leads were converted. We attain a maximum accuracy of 83.7% at 0.4 threshold value.

Further, a relevancy threshold of 0.4 has been set to categorize leads as converted. This means that if a lead has a relevancy score of 0.4 or above, it will be considered as converted by the model. It provides information about the performance of the conversion rate prediction model and the criteria used to categorize leads as converted.

2. Optimization-

In order to improve conversion rates, it is important to not only identify the relevant factors that influence conversion, but also to understand the reliability of these factors. Reliability refers to the consistency of the effect of a particular factor on the outcome, in this case, the conversion rate[7].

To determine the reliability of the relevant factors, we can analyze the conversion rates of different subgroups within each factor category. For example, within the "occupation" factor

category, we can analyze the conversion rates of housewives, doctors, teachers, and other occupations to determine which group has the highest reliability in terms of conversion.

Once we have determined the group with the highest reliability for each relevant factor, we can use this information to optimize our conversion strategy. For example, if we find that housewives have the highest reliability in terms of conversion for the "occupation" factor, we can target our marketing efforts towards this group in order to increase the overall conversion rate.

Furthermore, we can also use this information as a suggestive technique to improve conversion. For instance, we can provide recommendations to the user on how to improve the conversion rates based on the reliability of different factors. This can include suggestions such as targeting specific occupations or age groups, or optimizing the placement of advertisements on certain pages of the website.

By using reliability analysis as a tool for optimization, we can improve the overall conversion rates and ultimately achieve better business outcomes[9].

Reliability = global rate / group mean. If reliability is lower than 1, the group has lower reliability. And if group reliability is around 1, it's not different from global rate. And if it's above 1, it means that it has significantly more reliability that leads will convert their contracts.

```
[ ]  grouped_train = train_X.groupby('what_is_your_current_occupation').converted.agg(['mean'])
     grouped_train['diff'] = global_mean - grouped_train['mean']
     grouped_train['reliability'] = grouped_train['mean']/global_mean
     grouped_train
```

what_is_your_current_occupation	mean	diff	reliability
Businessman	0.400000	-0.014000	1.036269
Housewife	1.000000	-0.614000	2.590674
Other	0.147387	0.238613	0.381832
Student	0.377778	0.008222	0.978699
Unemployed	0.431573	-0.045573	1.118065
Working Professional	0.924107	-0.538107	2.394060

Housewife and Working Professional categories are most likely to be converted. There are sharp differences between values.

Fig 5.3 reliability of what_is_your_current_occupation

last_activity	mean	reliability
Approached upfront	1.000000	2.570694
Converted to Lead	0.135021	0.347098
Email Bounced	0.101064	0.259804
Email Link Clicked	0.232258	0.597064
Email Marked Spam	1.000000	2.570694
Email Opened	0.355397	0.913617
Form Submitted on Website	0.202703	0.521087
Had a Phone Conversation	0.785714	2.019831
Olark Chat Conversation	0.083789	0.215395
Other	0.822581	2.114603
Page Visited on Website	0.265714	0.683070
Resubscribed to emails	1.000000	2.570694
SMS Sent	0.644089	1.655757
Unreachable	0.288136	0.740708
Unsubscribed	0.333333	0.856898
View in browser link Clicked	0.250000	0.642674

Fig 5.4 reliability of last_activity

lead_profile	mean	reliability
Dual Specialization Student	1.000000	2.570694
Lateral Student	1.000000	2.570694
Other Leads	0.371528	0.955084
Potential Lead	0.901961	2.318665
Select	0.302527	0.777705
Student of SomeSchool	0.019868	0.051073

Fig 5.5 reliability of lead_profile

lead_source	mean	reliability
Click2call	0.500000	1.285347
Direct Traffic	0.322326	0.828601
Facebook	0.205882	0.529261
Google	0.403499	1.037271
Olark Chat	0.247967	0.637449
Organic Search	0.389752	1.001932
Press_Release	0.000000	0.000000
Reference	0.918429	2.361000
Referral Sites	0.246575	0.633870
Social Media	0.500000	1.285347
Welingak Website	0.987654	2.538957
bing	0.250000	0.642674
blog	0.000000	0.000000
welearnblog_Home	0.000000	0.000000
youtubechannel	0.000000	0.000000

Fig 5.6 reliability of lead_source

what_matters_most_to_you_in_choosing_a_course	mean	reliability
Better Career Prospects	0.496218	1.275625
Flexibility & Convenience	1.000000	2.570694
Other	0.134271	0.345170

Fig 5.7 reliability of what_matters_most_to_you_in_choosing_a_course

lead_origin	mean	reliability
API	0.310584	0.798415
Landing Page Submission	0.363866	0.935388
Lead Add Form	0.922197	2.370686
Lead Import	0.205882	0.529261
Quick Add Form	1.000000	2.570694

Fig 5.8 reliability of lead_origin

specialization	mean	reliability
Banking, Investment And Insurance	0.480769	1.235911
Business Administration	0.440909	1.133442
E-Business	0.437500	1.124679
E-COMMERCE	0.238095	0.612070
Finance Management	0.456364	1.173171
Healthcare Management	0.546512	1.404914
Hospitality Management	0.508197	1.306418
Human Resource Management	0.453815	1.166620
IT Projects Management	0.365854	0.940498
International Business	0.391753	1.007076
Marketing Management	0.483193	1.242142
Media and Advertising	0.456000	1.172237
Operations Management	0.473868	1.218169

Fig 5.9 reliability of specialization

In figure 5.3, the reliability of the current occupation of leads is relevant when they are housewives and working professionals. They are most likely to get converted since they have a reliability which is significantly more than one. Similarly, in figure 5.4, the last activity of the leads is relevant if leads are approached upfront, have had a phone conversation or resubscribed to the emails, they are most likely to get converted. In figure 5.5, the lead profile is relevant. Dual specialization students and lateral students are most likely to get converted. In figure 5.6, the lead source is an important parameter. The leads who have come via

references or Welingak website are most likely to get converted. In figure 5.7, the course on the website matters the most. Flexibility and convenience are important for the leads who are most likely to get converted. In figure 5.8, lead origin is reliable. The leads who submit the Lead Add Form and Quick Add Form are most likely to get converted. In figure 5.9, it is noticed that the specialization of the leads is not reliable in the prediction of the conversion of the leads[12].

3. Confusion Matrix-

```
df_scores = pd.DataFrame(scores)
df_scores.columns = ['thresholds','tp','fp','fn','tn']
df_scores[::10]
```

	thresholds	tp	fp	fn	tn
0	0.0	548	931	0	0
10	0.1	532	472	16	459
20	0.2	506	276	42	655
30	0.3	466	182	82	749
40	0.4	434	126	114	805
50	0.5	388	99	160	832
60	0.6	345	75	203	856
70	0.7	295	53	253	878
80	0.8	240	31	308	900
90	0.9	160	14	388	917
100	1.0	0	0	548	931

Fig 5.10 threshold and their tp fp fn tn values

	thresholds	tp	fp	fn	tn	tpr	fpr
0	0.0	548	931	0	0	1.000000	1.000000
10	0.1	532	472	16	459	0.970803	0.506982
20	0.2	506	276	42	655	0.923358	0.296455
30	0.3	466	182	82	749	0.850365	0.195489
40	0.4	434	126	114	805	0.791971	0.135338
50	0.5	388	99	160	832	0.708029	0.106337
60	0.6	345	75	203	856	0.629562	0.080559
70	0.7	295	53	253	878	0.538321	0.056928
80	0.8	240	31	308	900	0.437956	0.033298
90	0.9	160	14	388	917	0.291971	0.015038
100	1.0	0	0	548	931	0.000000	0.000000

Fig 5.11 tpr and fpr values

In figure 5.9, we found out the different true positive, true negative, false positive, false negative values at different thresholds. In figure 5.10, we were able to determine the tru positive rate and false positive rate at different threshold values.

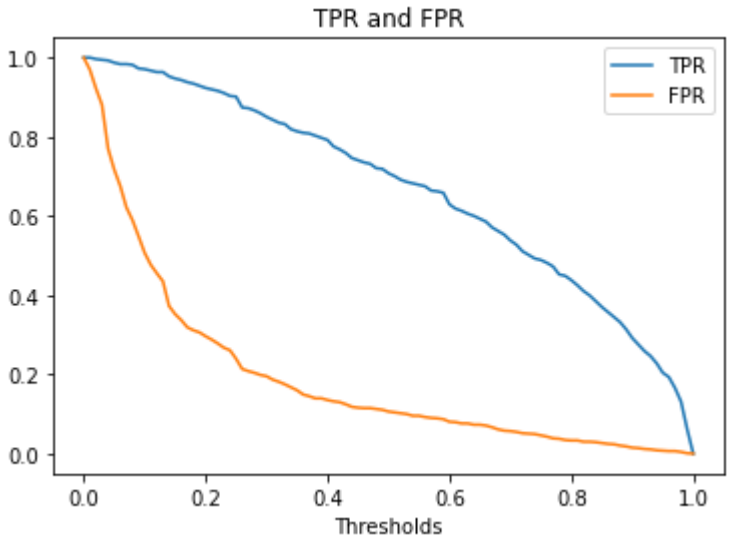

Fig 5.12 tpr and fpr vs threshold

We have plotted true positive rate and false positive rate vs the threshold as it can be observed in the figure 5.11.

4. ROC Curve and AUC-

   In our project, an ROC curve was implemented as a method for evaluating the performance of the logistic regression model in predicting the conversion rates of leads.

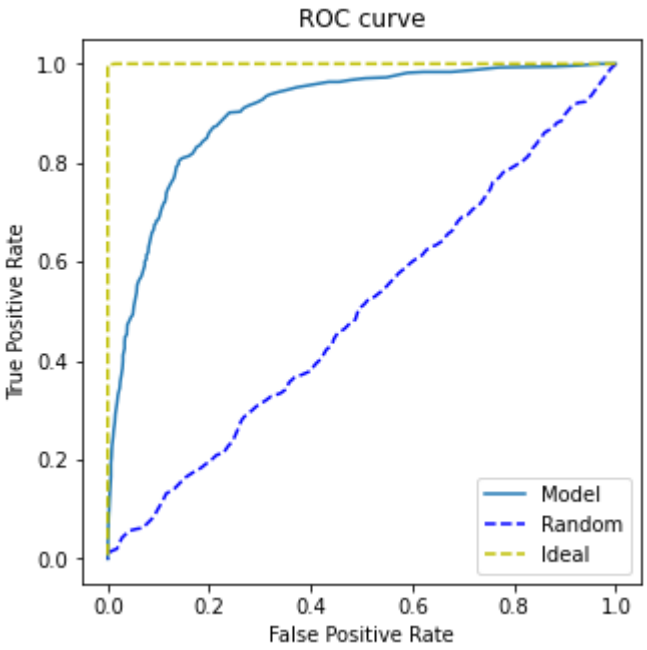

Fig 5.13 tpr vs fpr of roc curve

```
[ ]  auc(fpr,tpr)

     0.9019243886567304
```

Fig 5.14 auc value

We have plotted the TPR vs FPR for the ROC curve. Hereby, we have plotted the ROC curve for our model as it can be seen in figure 5.12. The baseline makes it easier to see how far the ROC curve of our model is from that of a random model. The top-left corner (0,1) is the "ideal spot": the closer our models get to it, the better.

We then found out the area under the ROC curve which is 0.9 as it can be observed in figure 5.13.

0.9 AUC indicates that the model is reasonably good. The closer the AUC is to one, the better the model.

Chapter 6

Conclusion and Future Work

Our model is important for using a combination of features to build a predictive model for conversion rate prediction. The results of the experiments showed that the model with reliability and feature engineering performed better than the Logistic Regression model. This suggests that the features used to build the model are probably not linearly related to the target label.

The model included raw features, cross features, and statistical features, but there are many more types of features that need to be taken into consideration in the future, such as ranking type features, trend type features, cross features of query and users, competition type of features, and more. It is suggested that different combinations of features could lead to a surprisingly good predictor.

The model only used Logistic Regression to build predictive models, but further work could be done with other algorithms and stacking of multiple algorithms. This approach could potentially improve the accuracy of the predictive model.

In the model, only click samples were used to build the model, but unclicked samples could potentially be used to supplement training. For example, if a query searches out many items and users probably only click one of them, information about other items that customers have seen could be extracted from other users who used the same query but clicked a different item. By organizing the whole dataset, some of the unclicked samples could be extracted, and this information could potentially improve the predictive model of conversion rate.

In conclusion, the model highlights the importance of using an accurate predictive model for conversion rate prediction. Further research could be done to explore the potential of using unclicked samples to supplement training and improve the predictive model.

References

[1] Dietmar, J., & Matle L.(2017). Investigating Personalized Search in E-Commerce. Association for the Advancement of Artificial Intelligence, www.aaai.org

[2] Pappas, I. O., Kourouthanassis, P. E., Giannakos, M.N., & Chrissikopoulous, V. (2017). The interplay of online shopping motivations and experiential factors on personalized e-commerce: A complexity theory approach. Telematics and Informatics, 34, 730-742

[3] Fan, Z., & Sagar, S. K. (2006). Predicting Online Customer Shopping Behavior.
Emerging Trends and Challenges in Information Technology Management, ½ Gefen, D., Darahanna, E., & Straub, D.W. (2003). Trust and TAM in online shopping:
An integrated model. MIS Quarterly, 27(1), 51-90

[4] QuanLu, Shengjun Pan, LiangWang, Junwei Pan, FengdanWan and Hongxia Yang. 2017. A Practical Framework of Conversion Rate Prediction for Online Display Advertising. In Proceedings of .e 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining', Halifax, Nova Scotia - Canada, July 2017 (KDD), 9 pages.

[5] T. Hastie, R. Tibshirani and J. Friedman, "Fitting Logistic Regression Models," i The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer, 2016, pp. 120-121.

[6] S. Lemeshow and D. W. Hosmer, "Model building strategies and methods for logistic regression," i Applied logistic regression, New York, John Wiley & Sons, INC, 2000, p. 93

[7] Bertelsen, S. M. (2012). Conversion Rate Optimisation: Barriers to Adoption. Master Thesis, IT University of Copenhagen, Digital Design and Communication, Copenhagen. Retrieved March 5, 2019, from https://crothesis.files.wordpress.com/2013/01/conversion-rate-optimisation-barriersto-adoption_sisse-bertelsen.pdf

[8] Croxen-John, D., & Tonder, J. v. (2017). E-Commerce Website Optimization: Why 95% of Your Website Visitors Don't Buy, and What You Can Do About it. Kogan Page Ltd.

[9] Shukairy, A. (2017, February 24). What is Conversion Rate Optimization (CRO) and Why Is It Important? Retrieved from www.invespcro.com: https://www.invespcro.com/blog/what-is-conversion-rate-optimization/

[10] Andrus, A. 2020. What is Conversion Rate? How to Calculate and Improve Your Conversion Rate. Disruptive Advertising. Retrieved on 11 August 2020. Available at: https://www.disruptiveadvertising.com/conversion-rate-optimization/conversion-rate/

[11] Saleem, H & Uddin, M. K. S. & Habib-ur-Rehman, S & Saleem, S & Aslam, A. M. 2019. Strategic Data Driven Approach to Improve Conversion Rates and Sales Performance of e-Commerce Websites. International Journal of Scientific & Engineering Research. Retrieved on 10 August 2020. Available at: https://www.citefactor.org/journal/pdf/Strategic-Data-Driven-Approach-to-ImproveConversion-Rates-and-Sales-Performance-of-E-Commerce-Websites.pdf

[12] Berezhnaya, Anastasia 2016 “Conversion Rate Optimization: Visual Neuro Programming Principles (Bachelor Thesis Helsinki Metropolia University of Applied Science)

[13] H Bryan 2013 Conversion Marketing: Convert Website Visitors to Buyers (AudioInk Publishing)

[14] P Soonsawad 2013 Developing New Model for Conversion Rate Optimization: A Case Study (International Journal of Business and Management) vol 8 no 10 pp 41-51.

Acknowledgement

We would like to express our gratitude and thanks to Dr. Seema Kolkur for her valuable guidance and help. We are indebted for her guidance and constant supervision as well as for providing necessary information regarding the project. We would like to express our greatest appreciation to principal Dr. G.T. Thampi and the head of the department Dr. Tanuja Sarode for their encouragement and tremendous support. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of the project.

Gaurav Advani

Saathvik Ayyamolia

Shruti Jain

Drishti Sachwani