Project Design On Predictive Analysis Of Sales Conversion.

Using binary classification model

1st Jomol Payyapilly Jonny
MSc.Data Analytics
National College of Engineering
Dublin, Ireland
x20249250@student.ncirl.ie

2nd Juan Pablo Andres Palma Bustos
MSc.Data Analytics
National College of Engineering
Dublin, Ireland
x20225776@student.ncirl.ie

3rd Sangeetha Pillay Anil Kumar
MSc.Data Analytics
National College of Engineering
Dublin, Ireland
x20232195@student.ncirl.ie

4th Yogalakshmi Chandrasekar
MSc.Data Analytics
National College of Engineering
Dublin, Ireland
x20221665@student.ncirl.ie

***Abstract*—Today, leads are the motivation for any business. With the rise of subscription-based business models, especially among start-ups, the ability to convert leads into paying customers has never been more important. A "lead" is a potential consumer who is interested in purchasing your goods or service. Marketing and sales departments devote a large amount of time, money, and effort to lead management, which we define as the three critical phases of lead creation, qualifying, and monetizing. In this research, a project design is provided in relation to analysis of existing information on leads who want to register in a program at an academic institution, and the data is leveraged for predictive analytics.**

Keywords– Binary classification model,Sales Lead Conversion, Predictive analysis.

I. BACKGROUND AND SCOPE

The education company "X Education" offer e-learning courses to industry professionals in many fields. Anyone interested in the course can browse the website and search for a suitable one. Also, the company has used advertising strategies to offer the product on several websites and search engines like Google. The nature of the Educational business involves a significant investment of time, money and personal effort, the consideration that can make a tough decision for many people.

On the other hand, from the company's perspective, Lead generation is the process of generating client interest in a company's products or services. Leads are generated to turn an interest or enquiry into a sale. In basic terms, lead qualifying evaluates and prioritises prospects to determine the possibility of conversion. The marketing and sales departments may focus on the prioritised leads rather than the hundreds of others. Lead conversion refers to converting a qualified lead into a paying client. It encompasses all marketing tactics that intrigue a customer's interest in a product or service and encourage them to purchase.

The company is looking to identify the lead that will be converted, which is currently around 30% of range conversion. We will propose a mechanism that helps focalise the company's effort to the most interested people using predictive techniques. We may approach the Leading Scoring System from the aspect of machine learning, where we train ML models on customer features, lead origin and other information available, with the goal of Lead Converted (Yes or No). Using data mining techniques, it may be easier to train a model based on previous data to make predictions. However, the trained models must be rigorously examined before reaching a major decision.

II. GOAL OF THE PROJECT

Goal of this project is to develop a most optimum predictive model using machine learning algorithm. The model will be evaluated based on evaluation metrics not just limited to accuracy but also sensitivity, precision, specificity, and Kappa value. Additionally, based on the result how the organization could build a strategy that is impactful to both consumers, and the business will be discussed in detail. This would be achieved based on both developing the predictive model and leveraging the use of advanced visual tools such as tableau and powerbi.

III. ABOUT THE DATASET

The data was obtained from Kaggle, which is open to the public. The dataset consists of 37columns, 36 of which are features and one of which "converted" is used as a response variable. The feature columns provide information

https://www.kaggle.com/ashydv/leads-dataset?select=Leads.csv

on leads who desire to enroll in a course at an educational institution. The target variable is a categorical column that indicates whether a lead has been converted or not.The data dictionary file which is retrieved from the website itself will be utilised for further analysis. The response variable's possible categories are binary values, as illustrated below.

- 0: It indicates that the lead has not been converted, whereas
- 1: It indicates that the lead has been converted.

The table below gives further information about the variables in the dataset.

Column name	Type
id	Numeric
Prospect ID	Numeric
Lead Number	Numeric
Lead Origin	String(Categorical)
Lead Source	String(Categorical)
Do Not Email	Factor(Binary)
Do Not Call	Factor(Binary)
Converted	Numeric (response variable)
TotalVisits	Numeric (Categorical)
Total Time Spent on Website	Numeric (Continuous)
Page Views Per Visit	Numeric (Continuous)
Last Activity	String(Categorical)
Country	String
Specialization	String
How did you hear	String
current occupation	String
What matters most	String
Search	String
Magazine	Factor(Binary)
Newspaper Article	Factor(Binary)
X Education Forums	Factor(Binary)
Newspaper	Factor(Binary)
Digital Advertisement	Factor(Binary)
Through Recommendations	Factor(Binary)
Receive More Updates	Factor(Binary)
Tags	String
Lead Quality	String
Supply Chain Content	Factor(Binary)
Get updates on DM Content	Factor(Binary)
Lead Profile	String
City	String
Asymmetrique Activity Index	Factor(Categorical)
Asymmetrique Profile Index	Factor(Categorical)
Asymmetrique Activity Score	Numeric (Continuous)
Asymmetrique Profile Score	Numeric (Continuous)
I agree to pay the amount through cheque	Factor(Binary)

Table.1 Data Description

IV. ETHICAL CONCERNS

Ethical concerns seem to be an unexplored matter in the data analytics field. The rocketing uses of data and the new ability to collect, store and process new data define the digital revolution, but it does not say anything about the externalities. Hence, the faster development of the method and techniques but barely understood their consequences. Therefore, when we talk about ethical concerns, we have to consider the individual, organizational and societal dimensions [1].

On the other hand, social science has widely discussed this matter, especially in the survey, interview and experimental research fields, such as sociology and psychology [2]. Interaction between accumulated knowledge and ethical concerns and the development of data analytics methods is the pending task to fill the gap between ethical and unethical practices.

A. DATA PRIVACY

There are no significant issues with personal data privacy in this project. First of all, the data set was already published for public use and free access online.

Secondly, guaranteeing the confidentiality of the leads is a priority. When people fill up the form providing email addresses or phone numbers, they are classified as leads. However, every ID for each lead was encoded, and sensible information such as email and phone numbers is not accessible in the data set.

Furthermore, when they fill up the form, they are asked whether they want to be emailed or called about the course or neither avoid any undesired spam.

Finally, the target is to predict which potential lead will be converted. But, again, that is a voluntary inscription throw the form. And The goal will help to focalize the resource and communication to people interested, giving a free choice to surf the web not to be involved with excess adverts.

B. POSITIVE IMPACTS

The benefits are for every person who wants to improve their skills in a specific domain. However, an exploratory analysis showed that it has been more interesting for unemployed people who wish to return to the labour market with upgraded aptitudes.

Improving the capacity to predict which potential lead will be finally converted can be especially beneficial for the company to aim all the products and communication efforts to seriously and genuinely interested people. But also, the customer that has been categorized as "hot lead" can have a personalized experience.

Thirdly, societal benefits could be met here because upskilling people who want to return to the labour market can be a valuable opportunity for many companies who wish to specialize their teams and face new challenges in the industry.

C. BIAS BASED ON ETHNICITY

The bias in the data is one of the more significant issues when we are analyzing data, especially when the data set have private information of people or companies or their

characteristics. The data set used in this project had already managed the risk and correctly encoded the sensible fields.

However, it is possible to see that most leads in the data set are from India. Still, it is due to products offered having been thought for the Indian labour market, even when it is open to people from other countries. Therefore, it is essential to understand that not every bias or localization is bad itself, but that is why proper business understanding is a crucial matter.

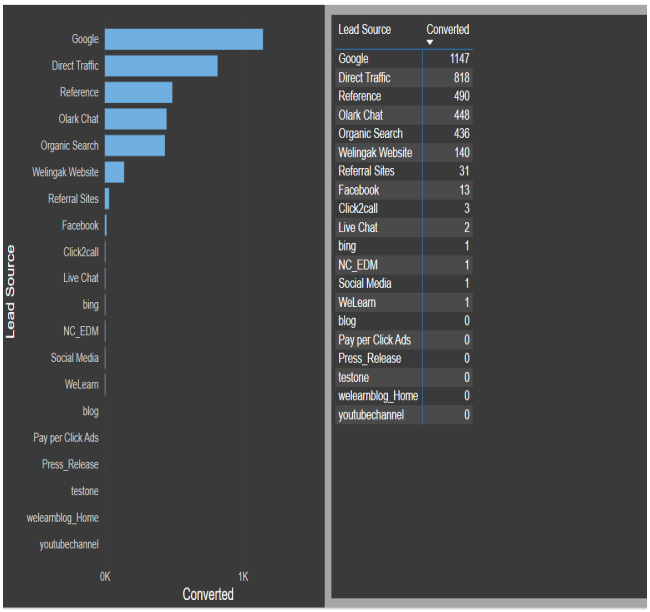
V. BUSINESS VALUES

Accurately predicting the leads conversion is extremely crucial for any kind of businesses. Especially in the educational sector, it is highly essential to quantify the prospective customers in order to start a batch, hire faculties and arrange facilities for the course. This analysis would also bring positive value to the business by precisely predicting the quality leads and redirect the sales force towards them. By predicting the leads based on the historical data available with the institution it is possible to predict the possible revenue that could be generated in the upcoming months or quarters. This predictive analysis add values to the business by showing clear picture of the future revenue this would subsequently help them ascertain the sustainability of the business in the market.

VI. PRELIMINARY VISUALIZATION

We utilized Power BI and Tableau for preliminary visualization. We generated dashboards for conversions by lead source, conversions by visits, conversions by location, and conversions by employment across locations. We were able to deduce a few things from our preliminary visualization, which are detailed below.

A. Lead Conversion by Lead Source



Lead Source	Converted
Google	1147
Direct Traffic	818
Reference	490
Olark Chat	448
Organic Search	436
Welingak Website	140
Referral Sites	31
Facebook	13
Click2call	3
Live Chat	2
bing	1
NC_EDM	1
Social Media	1
WeLearn	1
blog	0
Pay per Click Ads	0
Press_Release	0
testone	0
welearnblog_Home	0
youtubechannel	0

Fig 1: Lead conversion by lead source

This dashboard was created with Power BI. From this, we may deduce that Google was the dominant source of leads, as seen in Fig.2. Even if the conversion rate is low, referral sites and Facebook help to increase the conversion rate. The lead conversion is not aided by the blog, pay-per-click ads, press release, testone, welearnblog Home, or YouTube channel.

B. Lead Conversion by Visits

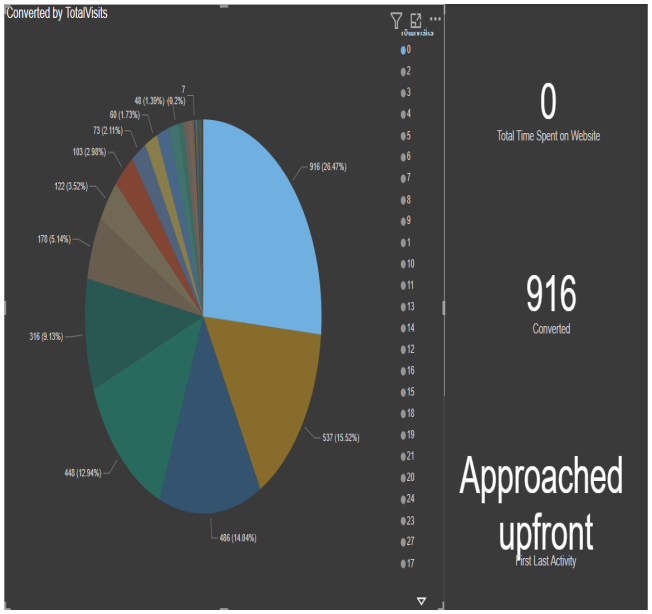


Fig.2 Lead conversion by zero visits

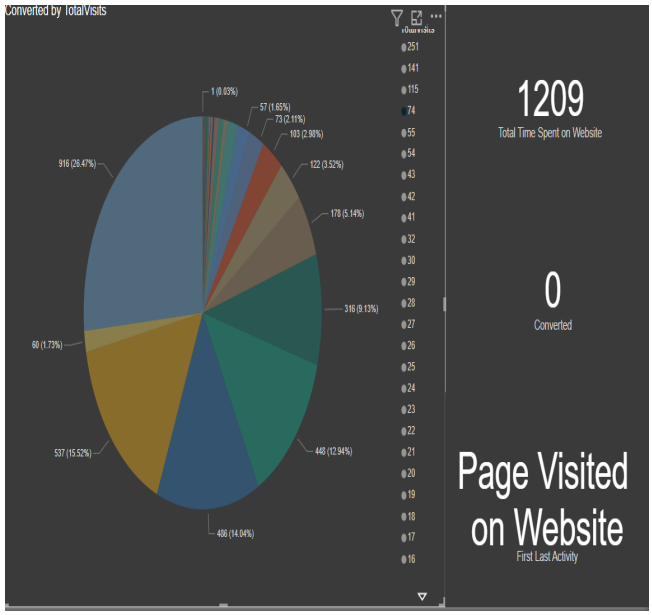


Fig.3 Lead conversion did not happen even with 74 visits

We were curious to see the conversion rate based on the number of times a person visited this page because it is a website. Surprisingly, the majority of lead conversions (916)

as shown Fig. 2 occurred without the client viewing the website, and the customer's last activity was to approach them directly. Customers prefer to speak with salespeople over visiting websites, as seen by this. Fig.3 shows that a consumer viewed the page 74 times and spent more than 1209 seconds on it, but he/she did not convert the lead.

C. Lead Conversion with respect to Locations

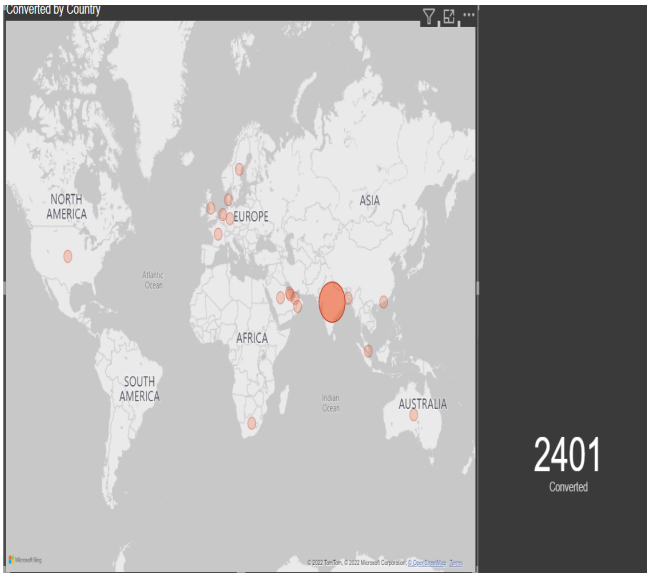


Fig. 4 Lead conversion by Location (Count shown is for India)

This was a fascinating dashboard to check which countries had the greatest lead conversions. And it was evident from the first glance at Fig.4 that it was in India (2401) rather than other countries.

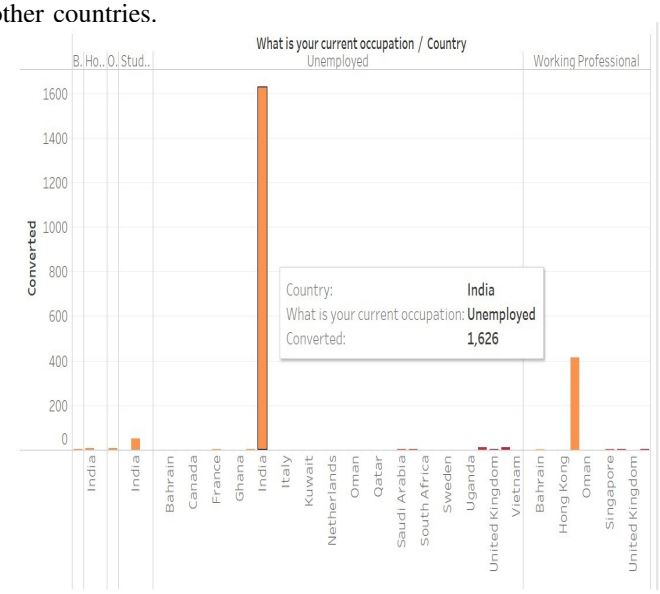


Fig. 5 Lead Conversion by Job Title in Relation to Locations

VII. HYPOTHESIS

Predictive analysis is based substantially on statistical approach. Before we begin building any predictive model it is crucial to establish the hypothesis. The hypothesis is basically divided in to null hypothesis and alternate hypothesis. The null hypothesis are actually based on the facts considering the preliminary analysis. Alternate hypothesis argues null hypothesis does not hold true. Considering the insights from the visualization null and alternate hypothesis for the study is stated below. Null hypothesis: The number of visits to the website does not impact the sales conversion.

Alternate hypothesis: The number of visits to the website has impact on sales conversion to some extent.

Whether to accept or fail to reject the null hypothesis will be handled in the further stages of this project.

VIII. APPLICABLE TECHNIQUES

The target variable in this study is binary-categorical. So supervised classification algorithm would be a best fit to apply on this dataset. There were multiple techniques employed by scholars in their research papers. The subsequent part of this section would be comprised of the techniques used by researchers on similar dataset or the similar problem statement.

- Zhenchang Xia [3] had applied ForeXGBoost model on their dataset to perform the sales prediction and they achieved high classification accuracy.
- Rajan Gupta [4] applied binary logistic regression algorithm in their study to predict the customer behaviour on dynamic pricing of products in an E-commerce dataset.
- Pedro José Pereira [5] discussed elaborately on applying decision tree and random forest techniques on their dataset to perform predictive analysis on mobile marketing sales conversion.
- Verena Eitle [6] conducted study on predicting the lead conversion using numerous classification algorithms such as catboost, random forest, Support Vector Machine, XGBoost, and Decision Tree. This study also suggest that predicting the lead conversion during the very early phase is difficult and gives inaccurate results.
- Lu´ıs Miguel Matos [7] in their research has applied XGboost and Logistic Regression to perform the classification to identify the user conversion rate (CVR).
- Jungwon Lee [8] had applied eXtreme Gradient Boosting model to predict the purchase conversion of online customers in their study.
- Marcelo Tallis [9] discussed about Historic Conversion Rate Feature Model (HCRFM), Time Decay Weighting Model (TDWM), and Mixture of Long-Term and Short-Term Model (MLTSTM) to perform the predictive analysis on sales conversion.

IX. CONCLUSION AND FUTURE WORK

This paper provides an outline of project design which comprises the goal, quick discussion on the dataset chosen, data type, distribution, and the background scope of it. It also discusses ethical concerns, business insights, values and preliminary visualization. This study also discusses in length about the appropriate techniques that can be applied on the data based on the list of literatures reviewed. Visualisation indicates that the time spent on website plays very slight role in conversion which is interesting and surprising. However, it is too early to decide. Further study will make it clearer.

For the predictive analyses of binary classification numerous techniques are available, however, the one that is most appropriate and best fit will be applied for this dataset.

REFERENCES

[1] Ethical implications of big data analytics,Ida Asadi Someh, Christoph Breidbach,Graeme Shanks, Michael Davern,2016.

[2] Ethical dilemmas and reflexivity in qualitative research, Anne-Marie Reid, Jeremy M. Brown, Julie M. Smith, Alexandra C. Cope, Susan Jamieson, 2018.

[3] ForeXGBoost: passenger car sales prediction based on XGBoost, Zhenchang Xia, Shan Xue, Libing Wu, Jiaxin Sun, Yanjiao Chen, and Rui Zhang, 2020.

[4] A Machine Learning Framework for Predicting Purchase by online customers based on Dynamic Pricing, Rajan Gupta and Chaitanya Pathak, ScienceDirect,2014.

[5] Multi-objective Grammatical Evolution of Decision Trees for Mobile Marketing user conversion prediction,Pedro José Pereira, Paulo Cortez , Rui Mendes, ScienceDirect,2020.

[6] Business Analytics for Sales Pipeline Management in the Software Industry: A Machine Learning Perspective, Verena Eitle and Peter Buxmann, 2019.

[7] A Comparison of Data-Driven Approaches for Mobile Marketing User Conversion Prediction, Lu´ıs Miguel Matos, Paulo Cortez, Rui Mendes, and Antoine Moreau, IEEE,2018.

[8] A Comparison and Interpretation of Machine Learning Algorithm for the Prediction of Online Purchase Conversion, Jungwon Lee, Okkyung Jung, Yunhye Lee, Ohsung Kim and Cheol Park, MDPI, 2021.

[9] Reacting to Variations in Product Demand: An Application for Conversion Rate (CR) Prediction in Sponsored Search, Marcelo Tallis and Pranjul Yadav, IEEE, 2018.