Predictive Analysis Of Sales Lead Conversion in E-Learning courses using Support Vector Machine.

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

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

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

4th Yogalakshmi Chandrasekar
MSc.Data Analytics
National College of Ireland
Dublin, Ireland
X20221665@student.ncirl.ie

***Abstract*—Nowadays, data can make the difference between whether enterprises can adapt to the new challenges or not. Therefore, lead conversion is crucial, especially in e-commerce and essential fields such as education. Using predictive analytics techniques, we present a solution for increasing the range of predictability of lead conversion to an 80% of accuracy using Support Vector Machine in the E-learning courses offered by X Education company..**

***Index Terms*—Sales Lead Conversion, SVM Model, Binary classification, Predictive analysis.**

I. Introduction

Leads are becoming the motivating force behind several enterprises. Lead conversion refers to converting a qualified lead into a paying client of a company's products or services. It has never been more vital than now, especially with the rise of subscription-based business practices, especially among start-ups. As a result, the marketing and sales departments may prioritise leads rather than hundreds of other potential clients.

For example, the education company "X Education" offers e-learning courses to industry professionals and is looking for a mechanism that helps to focalises the company's effort using predictive analysis.

They used advertising strategies to provide the product on several websites and search engines like Google. As a result, the lead conversion rate is around 30%. However, 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.

This paper presents a Leading Scoring System as a predictive solution technique that helps identify the potential lead using a Support Vector Machine, significantly improving the resources' efficiency and increasing the sales performance of Education X company. Moreover, in the next session, we present a state-of-the-art Support Vector Machine technique applied in related fields, which explains why we use this technique over others available. After that, we give the methodology based on CRISP-DM and their following steps, and finally, our result of the model's performance and the conclusion of this research.

II. Literature Review

Many research have been carried on lead conversion in sales and services. In this section, a literature study was conducted in order to finalize the prediction model.

In recent years many companies have made significant investments in predictive analytics techniques to improve their performance and the knowledge of customers. Support Vector Machine (SVM) seems to be a good standard and one of the predilect for this purpose, even in a field like product-service systems (PSS) [1] using this technique to identify customer needs and translate them into specific PSSs. Furthermore, in areas like Education and e-commerce in particular, SVM shows to be a very consistent technique for this purpose, even when the author concludes that the better accuracy is obtained using a collective approach than single classification technique [2].

The identification of lead users is a must for the success of any company. In Su et al [3], the researchers go deep into the three-factor that helps identify lead users. These factors are general user attributes (personal information), user activity attributes (behaviour) and user knowledge level attributes (demand-based knowledge factors and innovative skills factors) instead of just gender and age for customer collaborative product innovation (CCPI). And how the SVM was established to achieve more convenient, effective, and reliable results in leading user identification than other traditional methods for this task. Also, SVM was able to effectively solve the different costs of unbalanced data and sample misidentification and improved the identification performance from the perspective of misidentification cost and identification accuracy [3].

The author of [4] research use an SVM model to predict a customer's future purchasing chances based on past behav-

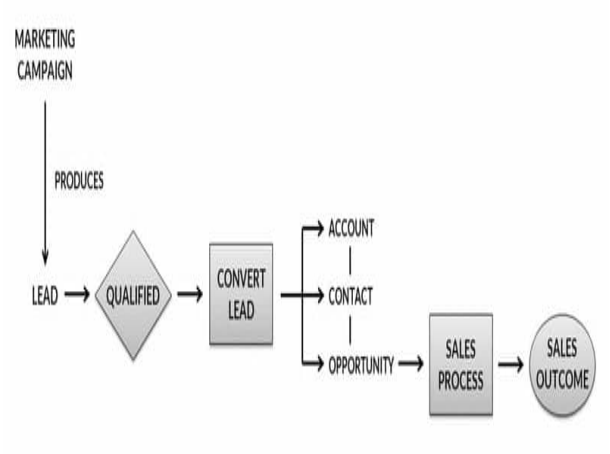

Fig. 1. Workflow diagram

ior. Linear separable and linearly inseparable situations were employed by the author (using RBF kernel). The accuracy of the model with the RBF kernel function is 98.7166%, and the F1 score on the model is 0.046823. The effects of applying a different kernel function and a varying percentage of positive and negative samples on model prediction outcomes are minor.

A comparison of models for predicting the sales conversion of car insurance promotional calls was included in [5].Logistic Regression, kNN, Decision Tree, and SVM tied for best model.

Another research [6] looked at whether it was feasible to forecast the conversion rate (CVR) by examining the readability of the landing page language. The information pertains to landing page content and conversion rates. ML methods were employed to forecast conversions. Five machine learning models were created. The accuracy of the machine learning model described using the SVM method appears promising for application. a research [7] also shows SVM on intrinsic risk reduction was used to customer churn prediction to increase the prediction abilities of machine learning approaches. Artificial neural networks, decision trees, and logistic regression were used to compare the approach. The best accuracy rate, hit rate, coverage rate, and lift coefficient are shown in this model.

III. Hypothesis

Any predictive analysis model is build majorly on techniques of statistics. The preliminary step in build machine learning algorithm is to put in place the hypothesis. As described in the project design proposal earlier. This project is built based on the hypothesis that sales conversion has no dependency on count of website visit. The alternate hypothesis is that website view does have impact on the sales lead conversion. As obtained from the visualisation, the machine learning model also confirmed that page view count does not have correlation on the conversion. So the null hypothesis is valid. However, the interesting information is length of time spent on website had positive correlation with the sales lead conversion.

IV. CRISP-DM

The life cycle of this model is described by the six phases of the CRoss Industry Standard Process for Data Mining (CRISP-DM). The stages of the cycle are depicted in the diagram as Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment.

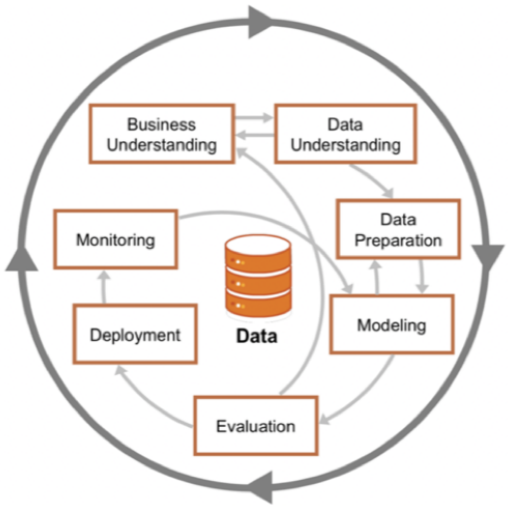

Fig. 2. CRISP-DM Diagram

1) Business Understanding: Lead conversion is a collaborative sales and marketing process that entails turning leads into customers via nurturing strategies such as behavior automation, retargeting, and email nurturing. Accurately predicting the leads conversion is extremely crucial for any kind of businesses. It is critical, especially in the educational industry, to quantify prospective clients in order to start a batch, hire faculty, and provide course facilities. This predictive analysis would also add value to the firm by accurately forecasting excellent leads and directing the sales force toward them. It is critical, especially in the educational industry, to quantify prospective clients in order to start a batch, hire faculty, and provide course facilities. This analysis would also add value to the firm by accurately forecasting excellent leads and directing the sales force toward them. In every predictive analytics project, validations and for model creation and efficient business rules evaluation are required.

V. Data Understanding

A. Data Preparation

In any predictive analytical model data exploration and preparation is the most crucial step. Almost 60 percentage of any data analytical project spent on data cleaning and preparation. If the data preparation step is handled correctly with proper conviction. The model which is developed using the dataset will provide valuable business insights. In this project, the machine learning algorithm employed was Support Vector Machine (SVM). In order for the SVM machine learning model to work and also to make predictions data need to be formatted in a specific ways. In reality the data that are

retrieved are not perfect for using them straightly in a machine learning model.

Challenges such as outliers, imbalance in the data, missing value, and text data are to be handled carefully. For instance, if there is a missing value the model could not use it. If there is invalid values it will lead to inaccurate predictions or even sometimes misleading outcomes. Even if the data is clean it would still lack useful information in the context of business. So enrichment of the feature must be taken place. When considered all these points a clean and well organized dataset can be prepared. This would lead to more accurate and practical outcome of the model. The leads dataset that is used in this study had multiple challenges on different level. So extensive and careful steps were taken in this study to retain the records as much as possible and also to bring a quality data for the model building.The coding language used for this project is R and IDE employed was RStudio.

1) Missing Data: The very challenging part of pre-processing the data for machine learning algorithm is the missing values. The missing value could be because of data corruption, missed to record data and there is no information to record for a particular variable. However, machine learning model could not be executed with missing data. In this dataset apart plain missing values there were other forms of data recorded which found to as less use as missing values. They were identified during each step of data cleaning and handled properly. Following are the examples: ”NA”, ”NULL”,””,” ”,”Select”. What is interesting is most of the website user did not make any selection for a few questions and the default choice ”Select” was populated as entries in several occurrences. Using ”.isna()” function number of missing row was identified and dropped and saved as new data frame programmatically.

2) Handling Imbalance and Outliers: This data did not suffer from outliers; However, the data had a lot of imbalance in multiple variables. Due to page constraint discussion will be on just one such example. In ”Do.Not.Email” column 4204 had ”no” as response and only 334 responses were ”yes”. However it was insightful to understand that majority of customers prefers to receive emails this sort of data imbalance will cause inaccuracy in predictions if we include them in the model. Hence the variable is removed from the dataframe.

3) Feature Reduction: This dataset had totally 9240 records and 37 columns in the beginning. At the end of the data cleaning process it came down to 4535 records and 12 columns. At the beginning of data processing records that old alphanumeric unique values such as *Tags, Lead.Quality, City, Country, Asymmetrique.Activity.Index, Asymmetrique.Profile.Index, Asymmetrique.Activity.Score, and Asymmetrique.Profile.Score*. The columns *city* and *country* were redundant as it is an online course.

4) Standardization of categorical variables: Another crucial step before applying the model is to standardize the categorical variable. A machine learning model could not understand text or category. We need to normalize it. The

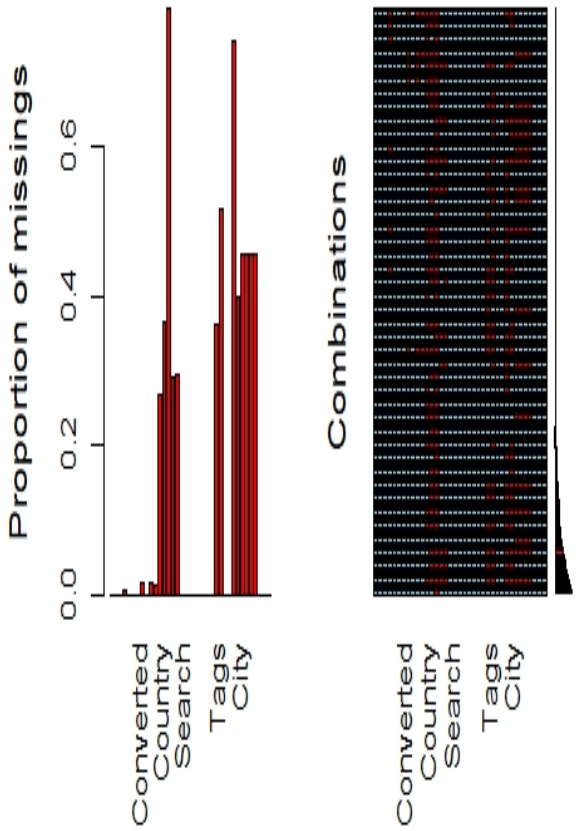

Fig. 3. Missing values

```
> table(df_5$Do.Not.Email)

   0    1
4204  331
```

Fig. 4. Imbalance in Do.Not.Email

predictor variables that are dichotomous in nature are assigned 1 for positive entry and 0 for negative entry. The predictor variables that has entries more then two category were one hot encoded.

B. Modeling

As discussed extensively in the literature review, all the classification models that were suitable for predicting binary classification were carefully analysed based on the results that was obtained from the previous studies, support vector machine (SVM) algorithm was considered to be employed in this study. SVM is a highly sophisticated supervised machine learning algorithm which can be applied for classification and also regression. The problem statement here in this study is to predict whether the sales lead will get converted to paid curriculum or not. The predicted variable is ”converted” it has

```
y_pred   0   1
     0 419  85
     1 136 493

               Accuracy : 0.8049
                 95% CI : (0.7806, 0.8276)
    No Information Rate : 0.5102
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.609

 Mcnemar's Test P-Value : 0.00077

            Sensitivity : 0.8529
            Specificity : 0.7550
         Pos Pred Value : 0.7838
         Neg Pred Value : 0.8313
             Prevalence : 0.5102
         Detection Rate : 0.4351
   Detection Prevalence : 0.5552
      Balanced Accuracy : 0.8039

       'Positive' Class : 1
```

Fig. 5. Confusion Matrix - Support Vector Machine

two types of responses "Yes" and "No". Yes was converted as 1 and No to 0. In SVM, each data will be plotted as a point in n dimensional space. The classification will be performed by identifying the hyper-plane that distinguishes the classes of interest efficiently. The data was split into training and test data in 70-30 proportion. The model was trained using the training dataset and later validated using the test data set.

C. Evaluation

The performance of the model was evaluated with the help of confusion matrix. Addition to accuracy various other statistical functions of the SVM model were summarised in R. They are Kappa value, sensitivity, specificity, pos pred value, Neg pred value, Accuracy, Balanced accuracy. The detailed value of each statistical function and the summary of the model can be found in the figure below.

Additionally, AUC curve is also checked.f Finally, The value of specificity 0.7550, sensitivity 0.8529, Kappa value 0.609, and AUC was 0.8039.

VI. Conclusion and Future Work

This paper's primary purpose was to give a solution to improve the earlier lead detection in E-learning courses using predictive analysis. Our solution using the Support Vector Machine achieves an accuracy of 80%, which is considered excellent in the literature and means a significatively improvement compared with the 30% of the regular investment in publicity in searching engines.

However, our performance has some limitations. First of all, the quality of the data. Predictive analysis can improve considerably the early detection of the lead, which is essential for any business. But often, the quality of the data is not the best, and improvements in this aspect can help to use better the data collected.

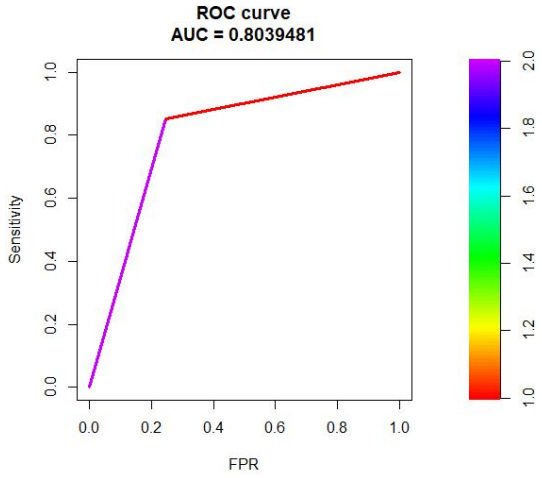

Fig. 6. ROC Curve

Secondly, our solution was concentrated only on Support Vector Machine, but as was mentioned by Manohar et al [2] a collective approach can be very substantial for classification problems. Therefore, we consider it is a valuable contribution to implement in future work.

In conclusion, our project was successful in its own goals. Still, for future appliances or extensions for a solution able to be performed in a real-world context, the considerations mentioned might be considered to improve the quality of data and the range of techniques for solving classification problems and their appliances in predicting leads for E-learning courses.

References

[1] H. Long, L. Wang, J. Shen, M. Wu, and Z. Jiang, "Product service system configuration based on support vector machine considering customer perception," *International Journal of Production Research*, vol. 51, pp. 5450–5468, 2013.

[2] E. Manohar, P. Jenifer, M. S. Nisha, and B. Benita, "A collective data mining approach to predict customer behaviour," *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021.

[3] J. Su, X. Chen, F. Zhang, N. Zhang, and F. Li, "An intelligent method for lead user identification in customer collaborative product innovation," *J. Theor. Appl. Electron. Commer*, vol. 16, pp. 1571–1583, 2021.

[4] X. Liu and J. Li, "Using support vector machine for online purchase predication," *International Conference on Logistics, Informatics and Service Sciences (LISS)*, pp. 1–6, 2016.

[5] D. Gopagoni, P. Lakshmi, and P. Siripurapu, "Predicting the sales conversion rate of car insurance promotional calls," *athore, V.S., Dey, N., Piuri, V., Babo, R., Polkowski, Z., Tavares, J.M.R.S. (eds) Rising Threats in Expert Applications and Solutions. Advances in Intelligent Systems and Computing*, vol. 1187, 2021.

[6] Korniichuk, Ruslan, and M. Boryczka, "Conversion rate prediction based on text readability analysis of landing pages," *Entropy (Basel, Switzerland)*, vol. 23, p. 11, 2021.

[7] G.-e. XIA and W.-d. JIN, "Model of customer churn prediction on support vector machine," *ystems Engineering - Theory and Practice*, vol. 28, pp. 71–77, 2008.