Usage of marketing and sales data for education outcome prediction

Vitalii Chibrikov[1*]

[1*]Artificial Intelligence Engineering, İstinye University, Ayazağa Mah. Azerbaycan Cad. (Vadistanbul 4A Blok), Sarıyer/İstanbul, 34396, Türkiye.

Corresponding author(s). E-mail(s): vitalii.chibrikov@stu.istinye.edu.tr;

Abstract

According to the report published by "Facts & Factors", the global Online Education Market size was worth around USD 217 billion in 2022 and is predicted to grow to around USD 475 billion by 2030 with a compound annual growth rate (CAGR) of roughly 9.1% between 2023 and 2030. As demand for online education has grown, the market has become increasingly competitive, with providers vying for attention from a broad set of prospective students. From 2011 to 2021, the number of learners reached by massive open online courses (MOOCs) increased from 300,000 to 220 million.

In this competitive market it is very important for companies, which sell educational programs, to predict behavior and preferences of each user on base of his interaction with the educational platform, to sell the right product to the right person in the right moment. This prediction should be based on marketing data: UTM marks, type of registration, reaction on emails, time spent on product landings, etc.; and data about interaction with sales team: time from registration to phone call, duration of the phone call, description of the call results by sales-manager etc. And all data about previous interaction of the user with the educational products, if any.

The same marketing and sales data can be used for education outcome prediction after the educational contract has been concluded and education started. In this research, the importance of marketing and sales data for education outcome prediction is shown.

The purpose of this research is to predict online education platform users' behavior and educational outcome with help of ML algorithms: k-Nearest Neighbors, Random Forest, Gradient Boosting, and Artificial Neural Networks.

Keywords: Machine learning, clustering, online-education, sales, marketing, EDM

1 Introduction

We can say that the quality of education is high, then the result of education matches the expectations from this education. For state universities and commercial educational centers, it is important to achieve a high level of quality education. One of the ways to increase the quality of education is to increase the relevance of the educational process to students and improve their grades. To achieve this, a recommendation system based on the education history can be used [1]. At the same time, the importance of valid and reliable admissions criteria should not be omitted [2]. So, data that contains information about the students before they start their education is also important.

The purpose of this research is to go further and check the importance of Sales and Marketing data to the performance, drop-off rate, and final grades of the students. Data about the process of customer attraction, Universal Transverse Mercator (UTM), interaction with the sales team could also be important. In this case, the recommendations gained from the data should not only be helpful to educators and students but also to business managers, salespeople, and marketing teams. Which can improve business processes, the effectiveness of educational products, and the variety of these products.

This research consists of three parts according to customer journey stages: from interaction with marketing events and registration on the educational center's site to request of consultation (Awareness Stage), next to interaction with the sales team and educational contract formation (Consideration Stage), and finally from start of education to graduation (Use or Retention Stage) [3]. In each part, different data is used.

The aim of each part of the research is to create models for prediction of each customer journey [CJ] stage result. For Awareness stage, the result is user's phone confirmation and request of consultation. For Consideration stage it is contract sign and payment. And for Use stage, it is successful graduation.

An additional aim of the study is to create a mathematical base for a recommendation system for managers in educational companies. Machine Learning models like k-Nearest Neighbors, Random Forest, Gradient Boosting, and Artificial Neural Network, look promising based on the research [4], [5]. Recommendation system can be built with the help of classification algorithms [1].

The goal of this research is to include marketing and sales data to the classification of users of online educational platforms. And evaluate the influence of sales and marketing data on the prediction of students' performance.

2 Literature review

This section of the article consists of two parts: a review of articles about the usage of ML in Education, and a review of the application of ML in Marketing. Each part addresses the following questions: What is the current state of machine learning use in the chosen area? What are the common approaches and algorithms? What data is used? What are the common problems?

2.1 Educational Data Mining

Prediction of students' performance, at-risk students, dropout, and retention have been the subject of numerous studies in the last few years [6]. So, a new field of study with the name Educational Data Mining [EDM] was formed [7]. The main goal of EDM is to find hidden patterns in educational data and use them to predict a process or outcome of education [6].

EDM includes many approaches to data analysis, like Supervised Machine Learning (Supervised ML), Deep Learning (DL), Analysis and Statistics. Supervised ML includes up to ten types of algorithms [6]. At the same time, clustering (as Unsupervised ML) is not very often approached in EDM [6]. But articles with this type of educational data analysis can be found [4], [5].

Finding and preparing high-quality data are the most challenging tasks in predictive learning analytics [6]. Three-quarters of all publications have used available data sources and records found in public datasets or through university/school repositories. According to [6], more than 90% of studies use educational records, questionnaires or both. Others do not state the origin of the data. No studies on the usage of marketing data were mentioned in the articles with EDM studies [6], [7]. Most studies tend to focus on the common educational characteristics of students, like overall grades, ignoring their individual characteristics [8].

In addition to difficulties with high-quality data preparation, there are problems and gaps in EDM. First, students' data used for prediction is often presented in imbalanced datasets [9]. So, special approaches are needed to work with these datasets. Next, there are difficulties in accessing personal data and privacy issues [10].

The dataset which is used in this research is unbalanced, but privacy issues are not the problem.

2.2 ML in Sales and Marketing

The current trend in marketing is to become more data-driven. To sell the right product to the right customers at the right time, information must be managed systematically, and meaningful information must be extracted from generated data [11]. Applications of big data analytics and ML in marketing are mainly connected to social media analysis, product and purchasing decision-making, and advertising [11]. The main goal of data analysis is to get insights from the raw data and to create a link between marketing and information systems to use marketing intelligence for user behavior prediction [12].

Applications of ML in Marketing are related to the prediction of market prices, forecasting of demand or purchase patterns, classification or prices of products, differences in prices, probabilities of abandonment, retention, or cancellation, and the customer lifetime value. There are studies predicting satisfaction and brand recommendation, as well as purchase intention [13].

The main sources of Marketing data can be divided into two categories: the first, open sources: social networks (Twitter, Instagram, and Facebook), online reviews and videos, or comments on YouTube. Review opinions are presented by film classification data, hotel websites, and Yelp. The second are internal proprietary databases such as

telephone numbers, matching products in the shopping cart, daily sales, and online purchases [13].

In terms of types of algorithms, ML in Marketing shows the same algorithms application as ML in Education. Supervised and unsupervised learning are common, but reinforcement learning, and hybrid methods are also used. In the present articles, the most used techniques are the artificial neural network (ANN), followed by the convolutional neural network (CNN) [13].

The main problem with ML in Marketing is the quality of the data. It is high-volume, fast-moving and not matched with traditional database structures. Also, privacy management, security, and ethical issues of gathering private data from customers should be mentioned [11].

3 Methodology

As it has been mentioned before, each modeling stage of the customer journey has its own data, but the same goal of prediction – on base of user's characteristics predict user's chances of successfully finishing the stage.

This part of the article has three sections. One section for one step of CJ. Each section has five subsections: Business Understanding, Data understanding, Data preparation, Modeling, and Evaluation. Corresponding to the first five phases of CRISP-ML methodology [14]. The Deployment stage of CRISP-ML is discussed in "Results and discussion" section of the article.

3.1 Awareness stage of CJ

Business understanding

The main goal of this part of CJ is to attract a potential customer, make him briefly familiar with the products and to interest him in registering in the educational platform. This part is successful if the customer registered in the platform and confirmed his phone number. Having the confirmed phone number make it possible for the Sales team to interact with the customer and lead help him to select the product and buy it. So, prediction of Awareness stag of CJ is prediction of phone confirmation on base of limited data which known about the customer after registration.

Data understanding

As it was mentioned before, finding high-quality data in EDM is one of the most challenging tasks. 75% of articles use data from public datasets or university/school repositories [6]. And as one of the sources of marketing data, researchers use proprietary databases [13]. So, in both EDM and ML in marketing, data is the challenging part of the work, and both use open sources and proprietary bases as a source of data. The source of data for this research is a proprietary database of online education schools, "Otus education" [15] and "Ketus education" [16].

As far as the data is private and belongs to commercial educational companies, special attention should be paid to the questions of permissions to access the data, data distribution, and protection of personal information. All required permissions were obtained before the beginning of the work.

Personal information is not included in the datasets obtained from the companies. All names, phones, emails and other personal information were removed from the dataset before the data was received. Sources of data from both companies have similar database structures, which serve as part of their back-end of learning management systems (LMS).

The data for prediction of phone confirmation consists only of user characteristics obtained in the process of registration and information the customer left about himself in the personal cabinet on the educational platform.

The data we worked in this stage contains records of online users' registrations in 2022 year. It consists of 20 columns, and 95508 rows. Fields of this data with statistical information shown in Table 1.

Table 1 Variables of Awareness stage

Name	Type	Missing	Count of TRUE	Unique values
ID	Number	0	-	95508
UTM_SOURCE	String	6795	-	57
UTM_MEDIUM	String	5542	-	30
UTM_CAMPAING	String	11813	-	241
UTM_TERM	String	18827	-	3782
TYPE	String	23026	-	3
GENDER	String	90363	-	2
CITY_SET	Boolean	0	8627	2
COMPANY_SET	Boolean	0	2934	2
IS_SUBSCRIBE	Boolean	0	21020	2
PARTNER_CAMPAIGN_SET	Boolean	0	5525	2
AVATAR_SET	Boolean	0	1172	2
WORK_SET	Boolean	0	2720	2
BLOG_NAME_SET	Boolean	0	94510	2
REAL_EMAIL_SET	Boolean	0	3	2
EMAIL_CONFIRMED	Boolean	0	1356	2
IS_MAIL_DISABLED	Boolean	0	240	2
IS_MAIL_UNSUBSCRIBE	Boolean	0	624	2
BIRTH_DAY_SET	Boolean	0	4409	2
PHONE_CONFIRMED	Boolean	0	59271	2

Data preparation

The data preparation step of this work includes two parts: data cleaning, and transformation of the data to a form suitable for ML models.

Data cleaning includes: removing of unusable variables, imputing missing variables, removing of variables which contains mostly the same value, and reduce cardinality.

After data cleaning, columns ID, UTM_CAMPAIGN, UTM_TERM, and REAL_EMAIL_SET were removed. All missing variables were substituted by "Missing" value. Cardinality of UTM_SOURCE and UTM_MEDIUM were reduced to values which are presented in more than 1000 rows. All other variables were substituted with value "rare_value".

The second part of data processing is preparation for ML. As one can see from the Table 1 our data contains fields with categorical values. Before further processing, categorical values were transformed to numerical with One-Hot Encoding.

Next step of data preparation is removing of outliers. Outliers are values within a dataset that vary greatly from the others. According to NumPy documentation [17] outliers can be found by comparing the local density of a sample to the local densities of its neighbors, one can identify samples that have a substantially lower density than their neighbors. These are considered outliers.

Finally, the data was split to train and test subsets. The train subset was undersampled to has equal rows with TRUE and FALSE values of dependent variable (PHONE_CONFIRMED). Thus, train data set is balanced.

All values in the final matrix are 0 or 1, so no normalization needed.

Modeling

After data preparation, we are ready to analyze the data with ML algorithms. As it was mentioned before, the aim of this stage of customer journey is prediction of PHONE_CONFIRMED variable value on base of information the used left at the moment of registration and filling of the fields of a personal cabinet.

Prediction of binomial value is a well-known binary classification problem. Based on the previous scientific researches which were mentioned in literature review section of the article, we chose the following ML model for this prediction: k-Nearest Neighbour [KNN], Artificial Neural Network [ANN], Random Forest [RF], and Gradient Boosted Trees [GBT].

All models trained with the same train dataset and tested with the same test dataset for evaluation of the model performance.

k-NN is the simplest model used with the following parameters: k=5, metric="cosine", algorithm="brute". Experiments with k=3 and k=7 and other metrics were also organized, but gave worse results.

ANN is used with the following parameters: two hidden layers with 16 nodes in each, loss function=BinaryCrossentropy, optimizer=Adam with learning_rate=0.0001, activation="softmax".

RF is used with the parameters: max_depth=100, random_state=0, criterion="gini". Experiments with other criterion also organized, but it doesn't influence the result significantly.

GBT is used with parameters: n_estimators=100, learning_rate=1.0, max_depth=5, random_state=0.

Evaluation

Comparison of main metrics of prediction is shown in Table 2.

As we can see, all models show approximately the same result. Models based on trees slightly better.

Recall is higher than Precision, and it is good result for business purposes. If the prediction of PHONE_CONFIRMED is YES, then marketing team should try to reach the person by marketing activities and support his possible interest of the educational products.

Table 2 Result metrics of PHONE_CONFIRMED prediction

Model	Accuracy	Recall	Precision	F1
kNN	0.670	0.771	0.715	0.742
ANN	0.693	0.828	0.716	0.768
RF	0.696	0.845	0.713	0.773
GBT	0.696	0.847	0.712	0.773

The results shown in Table 2 indicate, that having only data available at the moment of registration and personal cabinet fulfillment of a new user is enough to predict result of the first CJ stage with nearly 70% accuracy. At also indicate, that this CJ stage data can be useful to predict results of other stages of CJ, including success of education.

Next section of this article describe modeling of the other stages of customer journey with using data typical for the stage, and all data about the customer including marketing data.

3.2 Consideration stage of CJ

This section is not ready yet. Here will be description of prediction of the CJ second stage results, related to his interaction with the educational platform and Sales team. Right now, I am in the process of data collection for this stage.

3.3 Use stage of CJ

This section is not ready yet. Here will be description of educational outcome prediction.

4 Results and discussion

Final results and discussion will be ready after I finish 3.2 and 3.3 sections of the article.

Right now, the prediction of Awareness stage gives promising result. 70% accuracy in prediction of the less data-reach stage is good result for the current work.

All models used for the prediction showed approximately the same results. And as far as the models are very different from each over internally, it indicates that 70% accuracy is not a feature of a model, but hidden information in the data.

On base of the result of Awareness stage prediction, recommendation system for marketing team can be built. It can indicate users who potentially tend to confirm phone number, finishing the registration process. Additional marketing activities can be organized for these users as far as they are more loyal and ready for further interaction.

Further development of the marketing data can include research of the following questions: can we understand which rows of data influence the prediction the most?

can we identify which parameters (values) of marketing activities lead to phone confirmation? which data we need to add to improve the prediction? These questions may become topics of next studies.

References

[1] Fernández-García, A.J., Rodríguez-Echeverría, R., Preciado, J.C., Manzano, J.M.C., Sánchez-Figueroa, F.: Creating a recommender system to support higher education students in the subject enrollment decision. IEEE Access **8**, 189069–189088 (2020) https://doi.org/10.1109/ACCESS.2020.3031572

[2] Mengash, H.A.: Using data mining techniques to predict student performance to support decision making in university admission systems. IEEE Access **8**, 55462–55470 (2020) https://doi.org/10.1109/ACCESS.2020.2981905

[3] Mondragon, G., Méndez, K., Mauricio, D., Díaz, E.: Evaluation model of the digital experience in the retail sector using customer journey, pp. 1–4 (2019). https://doi.org/10.1109/INTERCON.2019.8853635

[4] Feng, G., Fan, M., Chen, Y.: Analysis and prediction of students' academic performance based on educational data mining. IEEE Access **10**, 19558–19571 (2022) https://doi.org/10.1109/ACCESS.2022.3151652

[5] Feng, G., Fan, M., Ao, C.: Exploration and visualization of learning behavior patterns from the perspective of educational process mining. IEEE Access **10**, 65271–65283 (2022) https://doi.org/10.1109/ACCESS.2022.3184111

[6] Shafiq, D.A., Marjani, M., Habeeb, R.A.A., Asirvatham, D.: Student retention using educational data mining and predictive analytics: A systematic literature review. IEEE Access **10**, 72480–72503 (2022) https://doi.org/10.1109/ACCESS.2022.3188767

[7] Ozyurt, O., Ozyurt, H., Mishra, D.: Uncovering the educational data mining landscape and future perspective: A comprehensive analysis. IEEE Access **11**, 120192–120208 (2023) https://doi.org/10.1109/ACCESS.2023.3327624

[8] Chen, Z., Cen, G., Wei, Y., Li, Z.: Student performance prediction approach based on educational data mining. IEEE Access **11**, 131260–131272 (2023) https://doi.org/10.1109/ACCESS.2023.3335985

[9] Bujang, S.D.A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., Ghani, N.A.M.: Multiclass prediction model for student grade prediction using machine learning. IEEE Access **9**, 95608–95621 (2021) https://doi.org/10.1109/ACCESS.2021.3093563

[10] Fernández-García, A.J., Preciado, J.C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J.M., Sánchez-Figueroa, F.: A real-life machine learning experience

for predicting university dropout at different stages using academic data. IEEE Access **9**, 133076–133090 (2021) https://doi.org/10.1109/ACCESS.2021.3115851

[11] Miklosik, A., Evans, N.: Impact of big data and machine learning on digital transformation in marketing: A literature review. IEEE Access **8**, 101284–101292 (2020) https://doi.org/10.1109/ACCESS.2020.2998754

[12] Chaudhari, V., Damle, M.: Strategic decisions using machine learning with interpretative structural modelling (ism) on digital platform data for marketing intelligence, pp. 58–64 (2023). https://doi.org/10.1109/ICICT57646.2023.10134285

[13] Duarte, V., Zuniga-Jara, S., Contreras, S.: Machine learning and marketing: A systematic literature review. IEEE Access **10**, 93273–93288 (2022) https://doi.org/10.1109/ACCESS.2022.3202896

[14] Schröer, C., Kruse, F., Gómez, J.M.: A systematic literature review on applying crisp-dm process model. Procedia Computer Science **181**, 526–534 (2021) https://doi.org/10.1016/j.procs.2021.01.199 . CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020

[15] Otus Education (2024). https://otus.ru Accessed 2024-05-01

[16] Ketus Online Education (2024). https://ketus.io Accessed 2024-05-01

[17] NumPy documentation (2022). https://numpy.org/doc/stable/ Accessed 2023-10-01