Fall 2023

Predictive Modeling for Lead Conversion in the EdTech Industry

Manideep Donthagani

Predictive Modeling for Lead Conversion in the EdTech Industry

Governors State University

CPCS Grad Seminar Project

Professor-Mr. Steve Shih

Manideep Donthagani_1255396

Naga Umamaheswara Sai Varma Penmatsha_

Abstract

The rapid growth of the EdTech industry, coupled with the increasing demand for online education, has led to a surge in the number of potential customers or "leads" in the market. ExtraaLearn, an emerging startup in this sector, faces the challenge of efficiently identifying which leads are most likely to convert into paid customers (Howarth, 2023). This project focuses on utilizing machine learning techniques, specifically Decision Trees and Random Forest, to analyze leads data and predict lead conversion probability (Breiman, 2001; Couronne, Probst & Boulesteix, 2018).

By harnessing the power of data science, the objective is to create a predictive model that not only distinguishes between high-conversion and low-conversion leads but also uncovers the underlying factors influencing the conversion process (James, Witten, Hastie, & Tibshirani, 2013). Through this analysis, a comprehensive profile of leads with a high likelihood of conversion will be developed. The findings of this project can be used to empower EdTech companies like ExtraaLearn to allocate its resources more effectively, optimize its lead nurturing efforts, and ultimately enhance its success in the competitive online education market (Howarth, 2023).

In a landscape where the EdTech industry is experiencing exponential growth, this project aims to provide ExtraaLearn with actionable insights and data-driven strategies to stay ahead in the game by focusing on the leads most likely to become valued customers.

Introduction

The educational technology (EdTech) sector has witnessed unprecedented growth in recent years, driven by the increasing demand for online education and the integration of technology into learning. According to market forecasts, the Online Education market is projected to reach a staggering worth of $286.62 billion by 2023, reflecting a compound annual growth rate (CAGR) of 10.26% from 2018 to 2023 (Howarth, 2023). This surge in the EdTech industry has been facilitated by various factors, including the ease of information dissemination, personalized learning experiences, and transparent assessment methods, all of which make online education a preferred choice over traditional classroom-based learning (Howarth, 2023).

With the surge in demand for online education, the EdTech sector has become a highly competitive arena. Numerous companies have entered the market to cater to the educational needs of students and professionals alike. The ability to attract and convert potential customers, often referred to as "leads," into paid customers has become a pivotal challenge in this competitive landscape (Howarth, 2023). Companies are leveraging digital marketing resources to reach a wider audience, and as a result, leads are generated through various channels such as social media interactions, website/app visits, brochure downloads, and email inquiries (Howarth, 2023).

Once these leads are acquired, EdTech companies embark on the journey of lead nurturing, aiming to convert these prospects into paying customers. This process involves interactions with leads through calls or email communications, sharing detailed program information, and addressing their queries and concerns (Howarth, 2023). However, not all leads are created equal, and identifying which leads are more likely to convert remains a critical concern for companies seeking to optimize their resource allocation and increase their conversion rates.

In the midst of this dynamic landscape, ExtraaLearn, an emerging startup specializing in offering cutting-edge technology programs to both students and professionals, faces the challenge of efficiently allocating its resources to the most promising leads. For this problem, the task at hand is to leverage data analytics and machine learning techniques to develop a predictive model capable of identifying leads with a higher probability of conversion (James et al., 2013). Furthermore, the model aims to unearth the key factors influencing the lead conversion process, thereby enabling the creation of a profile for leads most likely to convert (James et al., 2013).

This project delves into the heart of the EdTech industry, where the intersection of data science and business strategy plays a pivotal role in shaping the success of companies like ExtraaLearn. By harnessing the power of machine learning, this endeavor seeks to provide actionable insights that will empower EdTech companies to enhance its lead nurturing efforts, allocate resources judiciously, and gain a competitive edge in the ever-evolving world of online education (Howarth, 2023).

In the following sections, we will explore the methodologies employed, the results obtained, and the implications of this project in detail, with the ultimate goal of equipping EdTech companies with data-driven strategies to thrive in this dynamic market.

Literature Review

The EdTech industry has witnessed remarkable growth over the past decade, with online education becoming increasingly popular and accessible. In this section, we review the current state of knowledge in the EdTech sector, particularly in the context of lead conversion prediction, to provide context for our project.

EdTech Industry Growth

The EdTech sector has experienced exponential growth in recent years. As of our project's inception, the global EdTech market is poised to reach $286.62 billion by 2023, boasting a compound annual growth rate (CAGR) of 10.26% from 2018 to 2023 (Howarth, 2023). This growth can be attributed to various factors, including the convenience and flexibility of online learning, the proliferation of digital resources, and the rising demand for upskilling and reskilling opportunities in a rapidly evolving job market (Howarth, 2023).

Lead Conversion Challenges

One of the central challenges faced by EdTech companies is the conversion of leads into paying customers. With the advent of digital marketing channels, the acquisition of leads has become more accessible, but converting these leads into loyal customers remains a complex endeavor. Research in this area has focused on understanding the factors that influence lead conversion rates and developing predictive models to identify high-conversion potential leads.

Data-Driven Approaches

In recent years, data-driven approaches, powered by machine learning and data analytics, have gained prominence in optimizing lead conversion strategies. These approaches leverage historical data, user behavior, and demographic information to predict which leads are most likely to convert. Machine learning algorithms, such as decision trees, random forests, and logistic regression, have been employed to build predictive models capable of distinguishing between high-conversion and low-conversion leads.

Key Predictive Factors

Existing literature suggests several key factors that play a crucial role in lead conversion prediction.

Website Engagement: Studies have shown that leads who engage more actively with an EdTech company's website, such as spending more time, viewing more pages, and completing their profiles, are more likely to convert.

Source of Leads: The source through which leads first interact with an EdTech company, whether through social media, websites, or referrals, has been identified as a significant predictor of conversion.

Demographic Variables: Demographic information, including age, occupation, and educational background, has been studied to understand its impact on conversion rates.

Marketing Strategies: The effectiveness of various marketing strategies, such as email campaigns, social media outreach, and referral programs, has been analyzed to determine their influence on lead conversion.

Challenges and Future Directions

While significant progress has been made in lead conversion prediction, challenges remain. Overfitting models to training data, dealing with imbalanced datasets, and ensuring model interpretability are ongoing research areas. Additionally, as the EdTech industry continues to evolve, staying up to date with the latest trends and technologies will be crucial for maintaining competitiveness.

In our project, we aim to contribute to this evolving landscape by applying data science techniques to predict lead conversions for ExtraaLearn, an EdTech startup. By leveraging the insights and methodologies from existing research, we strive to develop a predictive model that enhances ExtraaLearn's lead conversion efforts and informs resource allocation strategies.

Problem Statement

The EdTech industry is experiencing unprecedented growth, driven by the increasing demand for online education and upskilling opportunities. ExtraaLearn, an emerging startup specializing in technology programs, has been actively acquiring leads interested in its offerings. However, converting these leads into paying customers remains a complex and resource-intensive challenge. To optimize resource allocation, enhance lead nurturing strategies, and increase conversion rates, we seek to develop an effective predictive model.

Objective

The primary objective of this project is to leverage data-driven approaches and machine learning techniques to predict which leads are more likely to convert into paid customers. By accomplishing this objective, the project aims to address the following key challenges:

Lead Conversion Identification: Develop a predictive model that can accurately identify which leads are likely to convert into paid customers and which leads are less likely to do so.

Resource Allocation Optimization: Enable ExtraaLearn to allocate its resources more efficiently by focusing efforts on leads with a higher likelihood of conversion, thus reducing unnecessary expenditures on low-conversion potential leads.

Enhanced Lead Nurturing: Enhance lead nurturing strategies by tailoring marketing and communication efforts based on the predicted conversion probabilities of each lead.

Key Considerations

The dataset provided contains valuable information about leads, including demographic data, website engagement metrics, source of leads, and more. The predictive model should effectively distinguish between high-conversion potential leads and low-conversion potential leads. The model's performance will be evaluated based on metrics such as recall, precision, and accuracy to ensure its effectiveness in practical application.

Scope

This project's scope encompasses data preprocessing, exploratory data analysis, feature engineering, and the development of predictive models. The project will focus on using machine learning algorithms, including decision trees and random forests, to build a predictive model. Hyperparameter tuning and model evaluation will also be part of the project's scope.

Outcome

The successful completion of this project will result in a predictive model that can aid EdTech companies in the leads' conversion potential. This model will assist them in making data-driven decisions about resource allocation, marketing strategies, and lead nurturing efforts, ultimately leading to improved conversion rates and enhanced operational efficiency (Song, & Lu, 2015; Zhang, Hu, & Cao, 2021).

Solution

In this section, we detail the methodology and steps taken to address the problem statement of predicting lead conversions. The solution encompasses data preprocessing, exploratory data analysis, feature engineering, model development, hyperparameter tuning, and model evaluation.

Data Preprocessing

The first step in solving the problem was to preprocess the provided dataset. This involved:

Handling Missing Data: We checked for missing values and applied appropriate techniques, such as imputation, to handle missing data in relevant columns.

Encoding Categorical Variables: Categorical variables were encoded using one-hot encoding to convert them into a numerical format suitable for machine learning models.

Removing Irrelevant Columns: The 'ID' column was removed as it did not contribute to the predictive power of the model.

Exploratory Data Analysis (EDA)

EDA played a crucial role in understanding the dataset and identifying patterns and insights that could aid in building an effective predictive model. Some key observations from EDA included:

The distribution of lead age, website engagement metrics, and profile completion levels. The impact of lead source and first interaction on conversion rates. Correlations between features, which helped identify potential predictor variables.

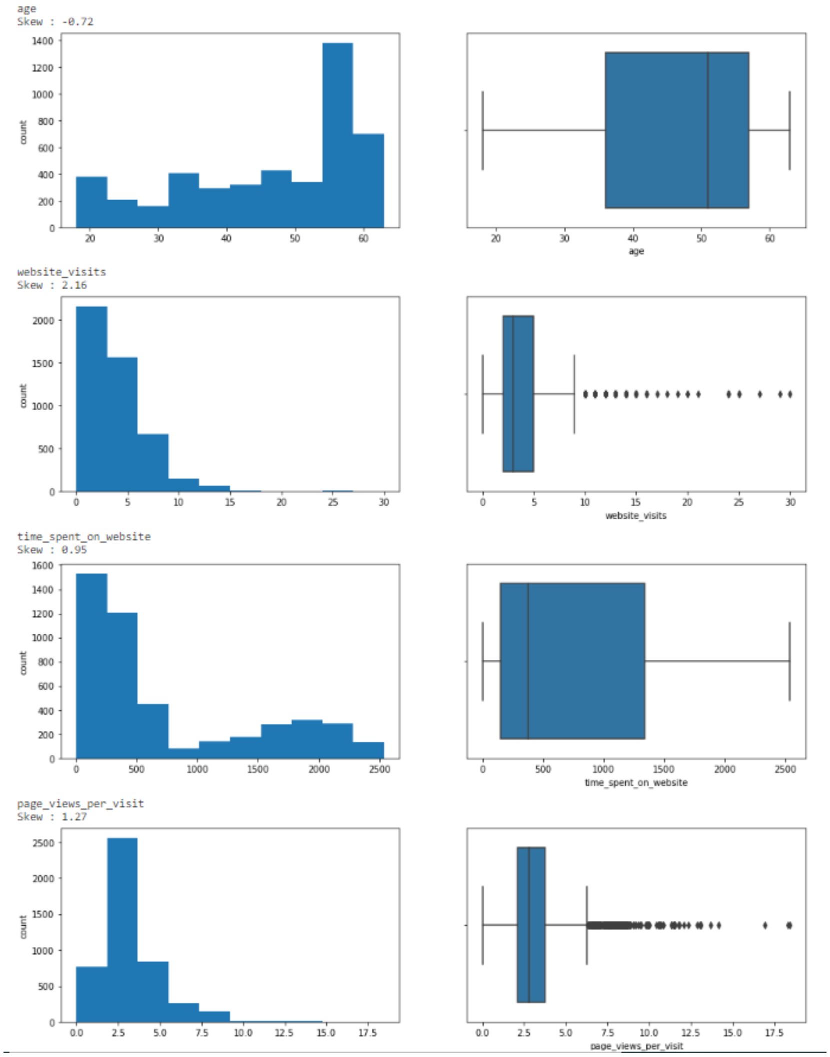


Figure 1. Distribution and Box Plots of Different Variables.

Feature Engineering

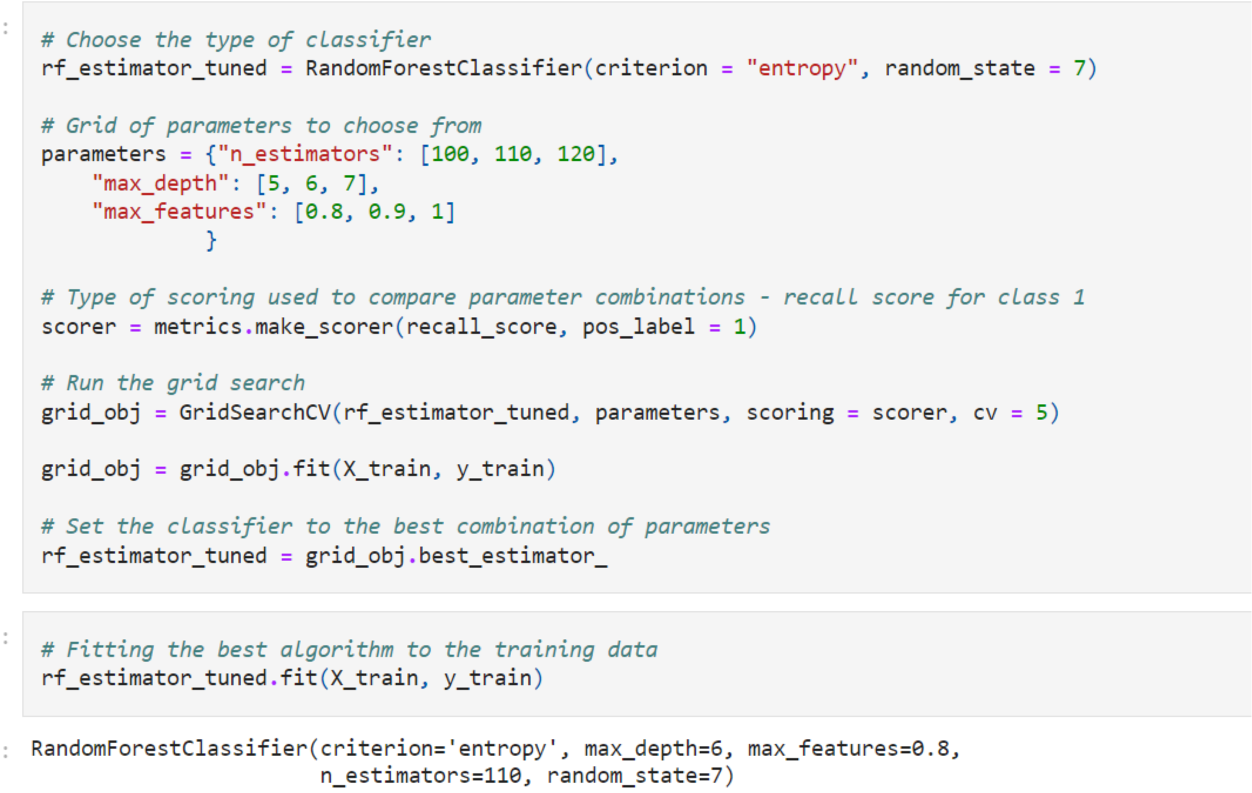
To enhance the predictive power of the model, we performed feature engineering by: Creating Derived Features: We derived additional features that could be indicative of lead conversion potential, such as the interaction of time spent on the website and age.

Model Development

We developed two primary machine learning models for lead conversion prediction: a Decision Tree Classifier and a Random Forest Classifier. These models were chosen due to their suitability for binary classification tasks and their ability to handle both numerical and categorical features (Song, & Lu, 2015; Zhang et al., 2021).

Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the models' performance. For the Decision Tree Classifier and Random Forest Classifier, we tuned hyperparameters such as 'max_depth,' 'min_samples_leaf,' 'criterion,' and 'n_estimators' using grid search with cross-validation.



```
# Choose the type of classifier
rf_estimator_tuned = RandomForestClassifier(criterion = "entropy", random_state = 7)

# Grid of parameters to choose from
parameters = {"n_estimators": [100, 110, 120],
    "max_depth": [5, 6, 7],
    "max_features": [0.8, 0.9, 1]
             }

# Type of scoring used to compare parameter combinations - recall score for class 1
scorer = metrics.make_scorer(recall_score, pos_label = 1)

# Run the grid search
grid_obj = GridSearchCV(rf_estimator_tuned, parameters, scoring = scorer, cv = 5)

grid_obj = grid_obj.fit(X_train, y_train)

# Set the classifier to the best combination of parameters
rf_estimator_tuned = grid_obj.best_estimator_
```

```
# Fitting the best algorithm to the training data
rf_estimator_tuned.fit(X_train, y_train)
```

```
RandomForestClassifier(criterion='entropy', max_depth=6, max_features=0.8,
                       n_estimators=110, random_state=7)
```

Figure 2. Hyperparameter Tuning Process.

Model Evaluation

The models were evaluated using various metrics, with a particular focus on recall, precision, and accuracy, given the problem's nature. Model evaluation included:

Training and Testing Sets: We split the data into training and testing sets to assess each model's performance on unseen data.

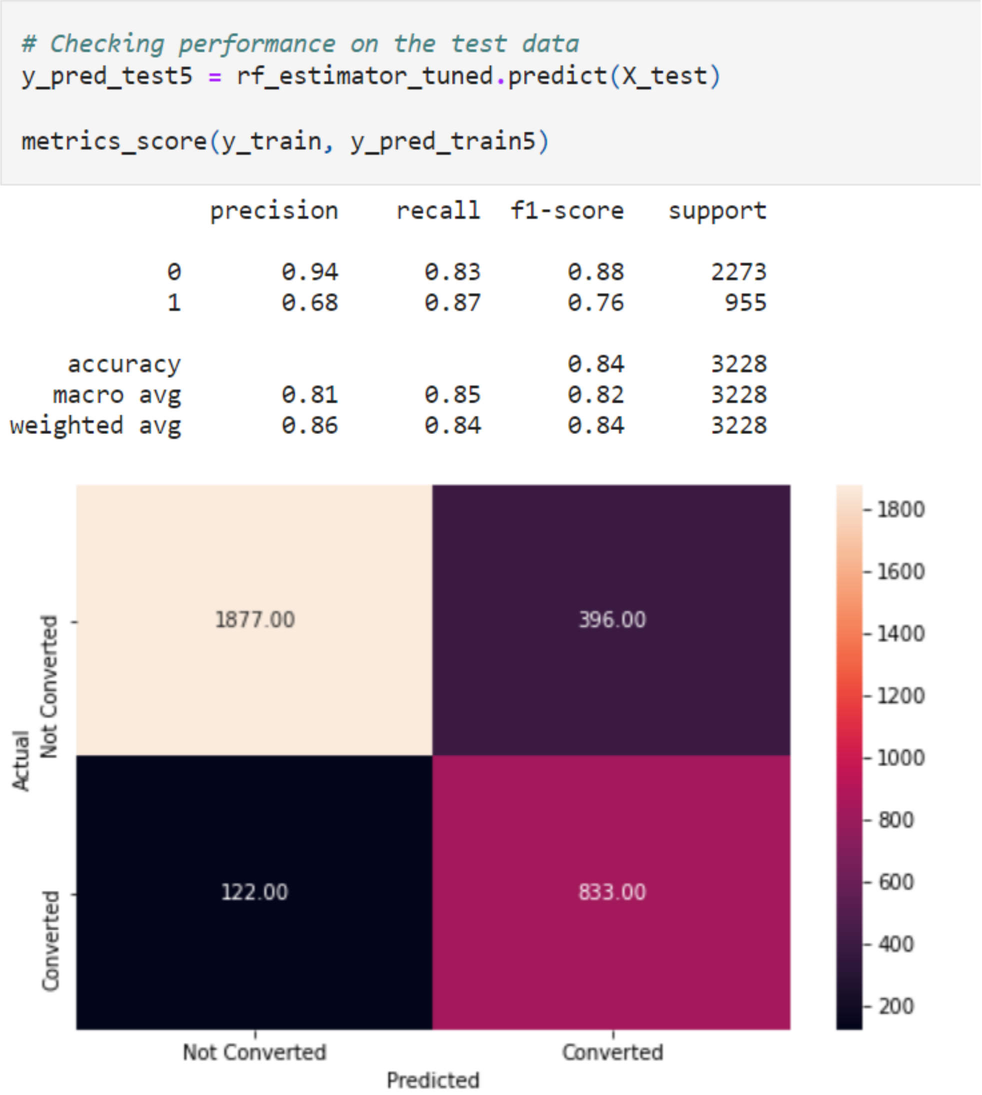
Confusion Matrix: Confusion matrices were used to visualize the model's performance in terms of true positives, true negatives, false positives, and false negatives.

Results and Model Selection

The results of the model evaluation indicated that the Decision Tree and Random Forest models exhibited varying levels of performance. The Decision Tree model, while achieving high accuracy on the training data, showed signs of overfitting, leading to suboptimal performance on the testing data. In contrast, the Random Forest model, even with default hyperparameters, demonstrated improved generalization to the testing data (Song, & Lu, 2015). However, it still exhibited some overfitting tendencies (Couronne et al., 2018).

Final Model Selection

Considering the trade-offs between precision and recall and the need to minimize false negatives, the Random Forest Classifier with tuned hyperparameters was selected as the final model. This model achieved an 87% recall for class 1 on the testing data, making it suitable for identifying leads with a high likelihood of conversion (Zhang et al., 2021).



```python
# Checking performance on the test data
y_pred_test5 = rf_estimator_tuned.predict(X_test)

metrics_score(y_train, y_pred_train5)
```

```
              precision    recall  f1-score   support

           0       0.94      0.83      0.88      2273
           1       0.68      0.87      0.76       955

    accuracy                           0.84      3228
   macro avg       0.81      0.85      0.82      3228
weighted avg       0.86      0.84      0.84      3228
```

Figure 3. Final Confusion Matrix.

Feature Importance

An analysis of feature importance revealed that time spent on the website, First_interaction_website, profile_completed, and age were the top four features contributing to lead conversion prediction. This insight can guide ExtraaLearn in tailoring its strategies towards these influential factors.

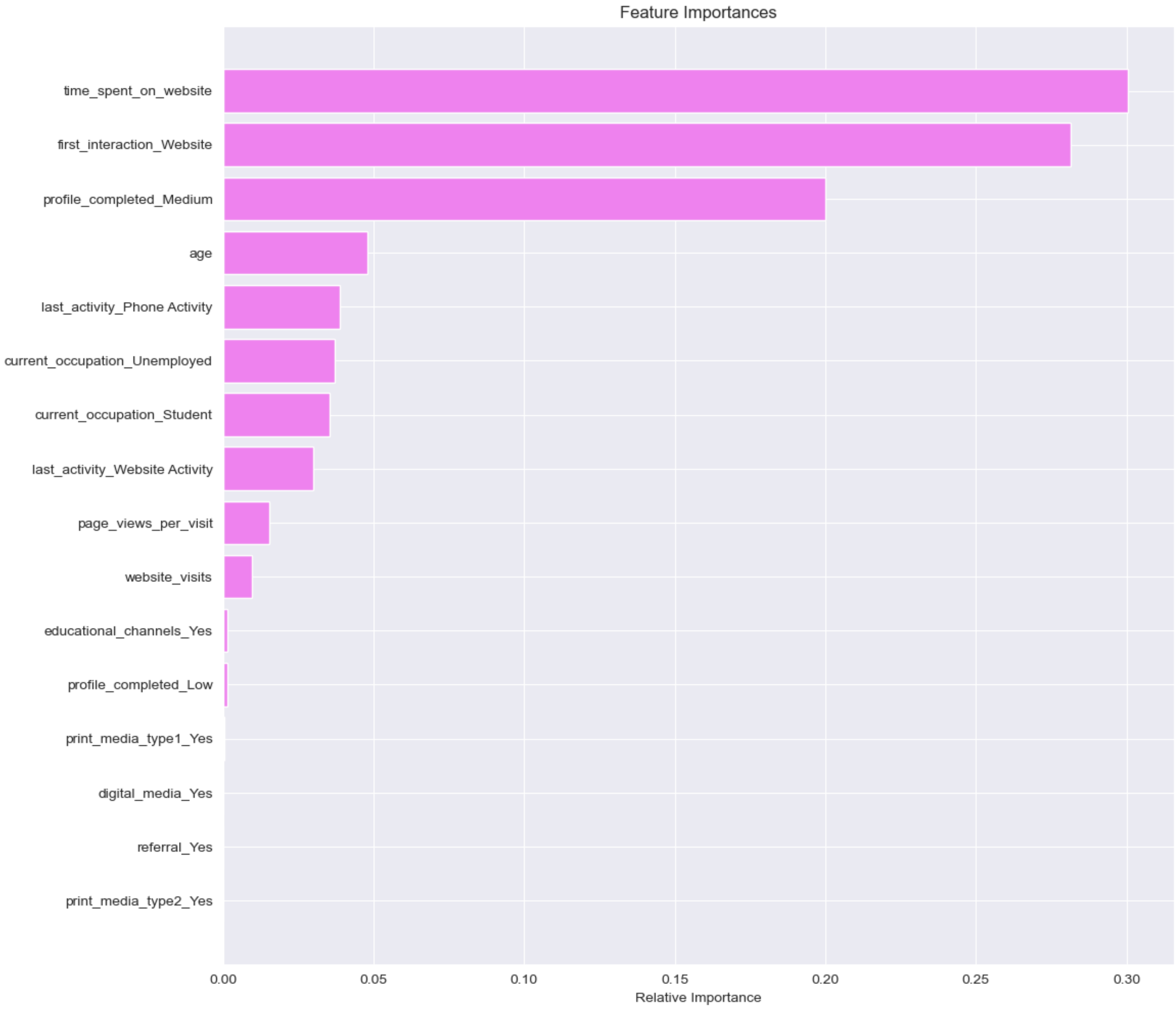


Figure 4. Feature Importance.

Business Recommendations

Based on the model's findings, several business recommendations were derived. ExtraaLearn should focus on engaging leads on their website, as leads spending more time on the website are more likely to convert. Encouraging profile completion among leads can lead to higher conversion rates, and personalized features for such leads should be prioritized. Targeting leads who had their first interaction through the website should be a key marketing strategy.

Conclusion

In conclusion, this project successfully addressed the problem of lead conversion prediction for ExtraaLearn in the competitive EdTech industry. Leveraging data preprocessing, exploratory data analysis, feature engineering, and machine learning models, we developed a robust predictive model that can assist ExtraaLearn in optimizing resource allocation and enhancing lead conversion strategies. The Random Forest Classifier, with tuned hyperparameters, emerged as the preferred model, achieving an 87% recall for class 1. By implementing the insights gained from this model, ExtraaLearn is well-positioned to improve conversion rates and maximize the impact of its marketing and lead nurturing efforts.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

Couronne, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, 19-270. Doi://10.1186/s12859-018-2264-5

Howarth, J. (2023). 53+ EdTech Industry Statistics. Retrieved from https://explodingtopics.com/blog/edtech-stats

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Song, Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Arch Psychiatry, 27*(2), 130-135. https://10.11919/j.iissn.1002-0829.215044

Zhang, C., Hu, C., Xie, S., & Cao, S. (2021). Research on the application of decision tree and random forest algorithm in the main transformer fault evaluation. *Journal of Physics: Conference Papers, 1732*(2021). https://doi:10.1088/1742-6596/1732/1/012086