Predictive Marketing Analytics for Better Lead Conversion in EdTech

Final report for the BDM Capstone Project

Submitted by

Name: AKHILESHWAR PANDEY

Roll number: 21F3002866

IITM Online BS Degree Program,

Indian Institute of Technology, Madras,

Chennai Tamil Nadu, India, 600036

Contents

1 Executive Summary

Blade Learners is an EdTech company that provides study materials, test series, and answer writing assistance for humanities students preparing for board exams and undergraduate entrance tests. This report is about problem of low conversion rate of leads faced by blade Learner.

This report provides a comprehensive analysis of lead conversion trends, identifying key inefficiencies in the sales funnel and suggesting data-driven strategies to enhance engagement and conversion rates. By analysing lead registration patterns, follow-up delays, subject preferences, and course conversion rates, we identify major drop-off points and propose solutions to reduce lead attrition and improve engagement strategies.

The machine learning model (Random Forest) achieved a high accuracy of 96%, demonstrating strong predictive capabilities. However, the model exhibits class imbalance, resulting in lower recall for converted leads. The feature importance analysis highlights that engagement timing, lead source, and subject selection significantly impact conversion outcomes. Additionally, the lead drop-off funnel analysis reveals that a large percentage of leads disengage after the first or second follow-up, indicating the need for a structured and automated follow-up system.

Key recommendations include reducing follow-up delays, optimizing outreach strategies, prioritizing high-converting channels, and leveraging urgency-based discounting to drive quicker decision-making. Future work will focus on refining predictive models, implementing real-time lead scoring, automating engagement workflows, and optimizing multi-channel marketing efforts. By adopting these strategies, businesses can significantly improve their lead conversion rates, optimize resource allocation, and enhance overall sales efficiency.

2 Detailed Explanation of Analysis Process/Method

Initial Approach (Preliminary preprocessing and cleaning)

This project aims to analyse lead data to identify inefficiencies in the sales funnel and improve tele-calling efforts. Many leads fail to convert due to delayed or inconsistent follow-ups, making it essential to understand engagement patterns and optimize resource allocation. To achieve this, preliminary preprocessing was conducted in MS Excel, where missing values were handled, irrelevant columns were removed. The cleaned dataset was then imported into Google Colab, where exploratory data analysis (EDA) was performed using Pandas, Seaborn, and Matplotlib to uncover trends in conversion rates. Finally, a predictive classification model was developed to rank leads based on conversion probability, allowing sales teams to focus on high-potential prospects and improve overall efficiency.

Quantitative Analysis

Quantitative analysis focuses on measurable aspects of the dataset, allowing us to identify trends and statistically significant relationships. By applying statistical techniques, we aim to uncover patterns that influence lead conversion.

Trend Analysis

- Time-series analysis was conducted to examine the registration of leads over time.
- Also, along with registration comparative time series analysis of Registration and 1st calling was done.

Statistical Summary

- Descriptive statistics were used to understand the dataset.
- Mean, median, and mode for numerical variables Diff1, Diff2, Diff3 was calculated, along with Standard deviation and variance to measure data dispersion.
- For categorical columns their frequency and their ratio with respect to total count was calculated.

Conversion Rate Analysis

- The dataset was analysed to determine the proportion of leads that successfully converted. Out of 550 leads 40 were converted, hence giving conversion rate of 7.27 percent.

Correlation Analysis

- Pearson's correlation coefficients were calculated to measure relationships between numerical variables Diff1, Diff2, Diff3.
- Heatmap was drawn to see the correlations pictorially.

Lead Drop-off and Conversion Funnel Analysis

- A funnel analysis was conducted to understand at which stage leads tend to drop off.
- The lead journey was divided into multiple stages: registration, first call, second call, third call and converted.
- Drop-off rates were calculated at each stage to identify bottlenecks where potential customers disengage.

Qualitative Analysis

Qualitative analysis involves examining non-numerical data, including text-based information, behavioural insights, and engagement patterns, to better understand lead behaviour and optimize conversion strategies.

Remarks Column which had subjective remarks was analysed and some information (like if parents intervention was sought or not-'Parents' column) was retained from it before discarding it.

Predictive Analysis

Predictive analysis was performed to estimate the probability of lead conversion using machine learning models trained on historical lead data. By identifying key factors influencing lead conversion, the model can assist sales teams in making informed decisions and increasing overall conversion rates.

Splitting The data

Data was splitted into 80:20 ratio as training and validation sets using train_test_split python library.

Handling Class Imbalance with SMOTE

The dataset exhibited class imbalance, where the number of converted leads (positive class) was significantly lower than non-converted leads (negative class). Training a model on imbalanced data could lead to bias toward the majority class, resulting in poor recall for converted leads. To address this issue, SMOTE (Synthetic Minority Over-sampling Technique) was used to generate synthetic samples for the minority class, ensuring a balanced dataset for model training.

Feature Engineering for Model Optimization

Outliers were handled using a Z-score-based method to replace extreme values with the mode of the respective feature. This method ensured that anomalous values did not negatively impact model predictions.

Diff1, Diff2, Diff3 (Time gaps between registration and follow-ups) were introduced as new features to identify delayed responses impacting conversion.

Categorical features were encoded using Label Encoder and numerical columns were scaled using standard scaler.

Since relationships between features may be non-linear, polynomial features were generated to improve model performance.

Model Selection and Justification

Several classification models were evaluated to determine the most effective approach for predicting lead conversion. The models tested include:

- **Logistic Regression**: Chosen as a baseline model due to its simplicity, interpretability, and ability to provide probability scores for conversion likelihood. While it offers explainability, it may struggle with complex decision boundaries.
- **Decision Trees**: A model that captures hierarchical decision-making by splitting data based on key attributes. Decision trees are easy to interpret but prone to overfitting when deep.

- **Random Forest**: Ensemble models that leverage multiple decision trees to improve predictive accuracy. Random Forest reduces overfitting by averaging multiple trees.

After testing and comparing model performance, **Random Forest** was selected as the primary model due to its ability to handle both categorical and numerical data, robustness to missing values, and improved predictive power.

Hyperparameter Tuning for Optimization

To enhance model accuracy and generalization, **hyperparameter tuning** was conducted using **GridSearchCV**, an exhaustive search method that evaluates different parameter combinations through cross-validation. The parameters optimized for the **Random Forest** model included:

- **n_estimators (Number of trees)**: Higher values help improve accuracy but increase computational cost.
- **max_depth (Tree depth)**: Controls the complexity of individual trees to prevent overfitting.
- **min_samples_split (Minimum samples required to split a node)**: Determines when a node should be split to ensure meaningful partitions.
- **min_samples_leaf (Minimum samples per leaf)**: Prevents model over-complexity by ensuring each leaf contains enough samples.

The tuning process identified the best parameter combination, significantly improving the model's ability to generalize well on unseen data.

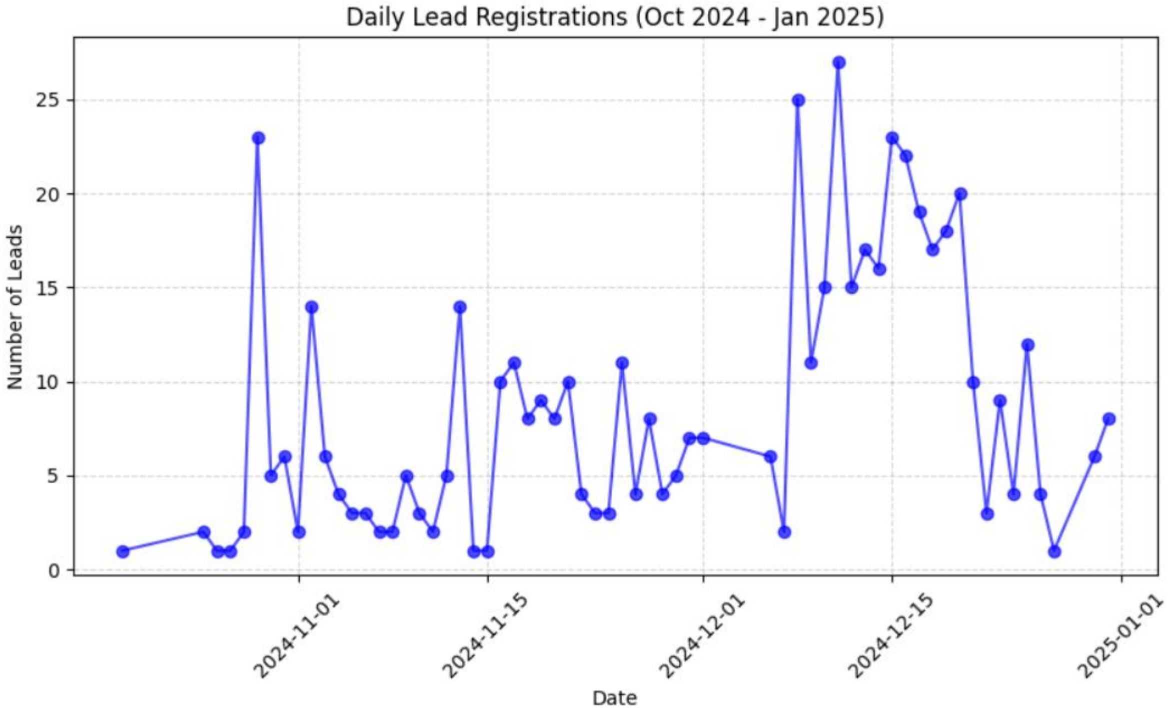
Model Evaluation metrics

- Accuracy: Measures overall correctness.
- Precision, Recall, and F1-score: Assess how well the model distinguishes between converted and non-converted leads.
- ROC-AUC Score: Used to assess the probability estimates provided by models.
- Confusion Matrix: Examined false positives and false negatives to ensure a balanced classification approach.

3 Results and Findings

Time series Analysis

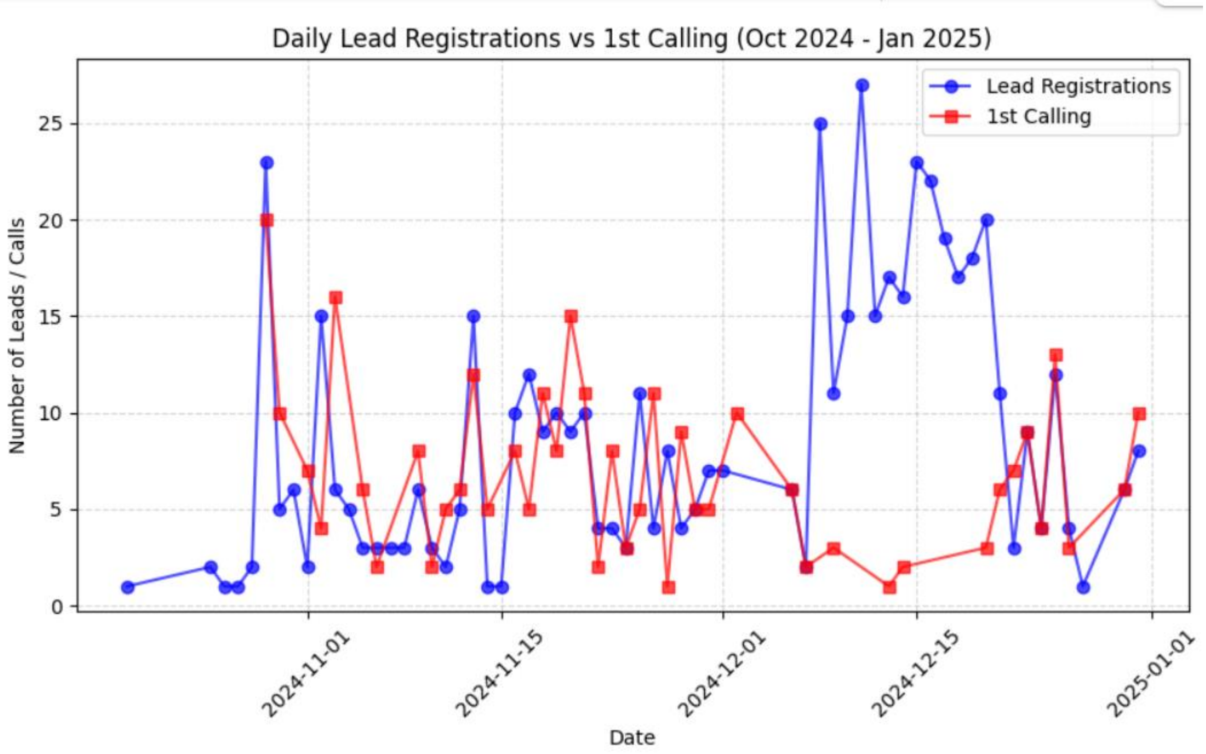
- Following graph was plotted to observe trends in new lead sign-ups over time, helping to identify peak periods and any seasonal fluctuations. This visualization allows to assess whether marketing campaigns or external factors influence lead generation rates. Understanding these patterns can aid in optimizing outreach strategies and resource allocation for follow-ups.



Graph 1: Daily Lead Registration

- Graph shows that the number of daily lead registrations varies significantly throughout the period, with noticeable peaks and troughs, but there is a clear surge in daily lead registrations during December 2024, specifically in mid-December.

This graph was plotted to analyse the response time between lead registration and the first follow-up call. Delays in initial engagement can impact conversion rates, so this visualization helps assess how efficiently leads are being contacted. By identifying trends or inconsistencies, we can improve follow-up strategies to enhance conversion chances.



Graph 2: Daily Lead Registrations vs 1st Calling Date

The graph illustrates daily lead registrations (blue line) and the number of first calls made (red line) between October 2024 and January 2025. Following observations are endorsed by the graph:

1. **Lead Registrations vs. First Calls:** There are multiple instances where lead registrations spike significantly, but the corresponding first calls do not increase proportionally, especially during the mid of December. This anomaly suggests potential delays in reaching out to leads.

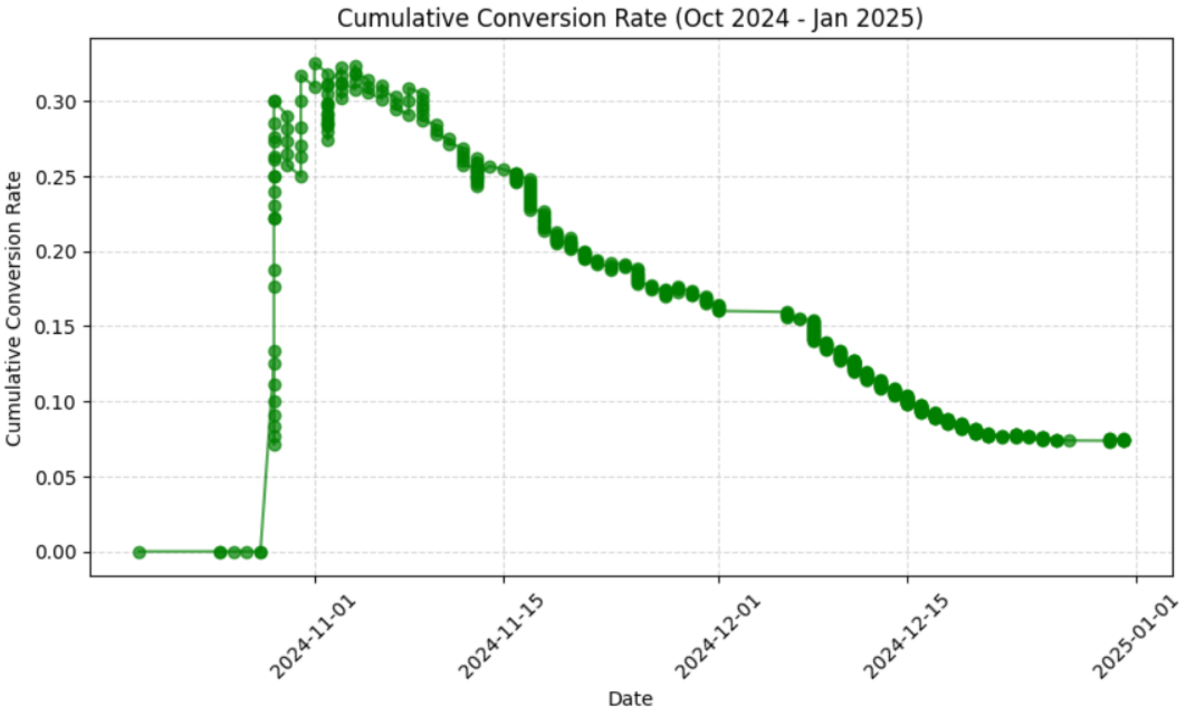
2. **Periodic Spikes:** Peaks in lead registrations are evident in early November and mid-December, possibly due to marketing campaigns or seasonal trends. However, first calls do not consistently follow the same trend, indicating inefficiencies in follow-up processes.

3. **Gaps Between Registration and Calls:** Some days show high lead registrations but low calling activity, which might lead to lost opportunities.

Cumulative Conversion Rate

The Cumulative Conversion Rate graph was plotted to track the overall efficiency of the lead conversion process over time. By visualizing how conversions accumulate, we can assess whether our strategies are consistently improving or if there are periods of stagnation.

Following graph shows the cumulative conversion rate from October 2024 to January 2025, tracking how effectively leads are being converted over time.



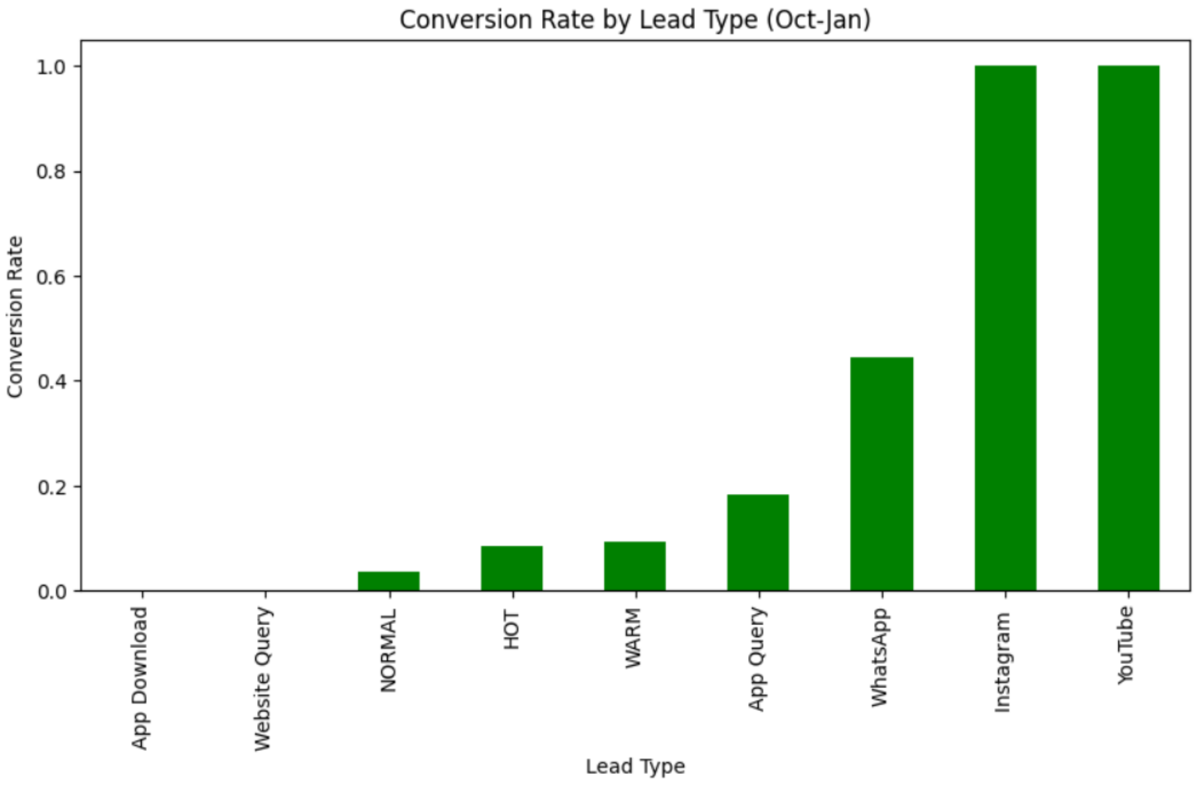
Graph 3: Cumulative Conversion Rate

The conversion rate followed a sharp increase to around 30%. During start of November, this suggests a successful campaign or an operational shift that significantly improved lead conversion. Sales head Faran told that this was because of starting of a new batch during that time.

Further this shows a brief period of stabilisation and then continuous downtrend during December which is expected for anomaly mentioned above during mid of December, i.e. less calls were made despite surge in daily registrations.

Conversion Rate by Lead Type

The motivation for plotting this graph was to evaluate the conversion rates across different lead sources and identify the most effective channels for acquiring customers. By analysing this data, we can optimize marketing strategies and focus efforts on platforms that yield higher conversions. This helps in improving lead nurturing processes and reallocating resources toward high-performing channels while identifying gaps in underperforming ones.

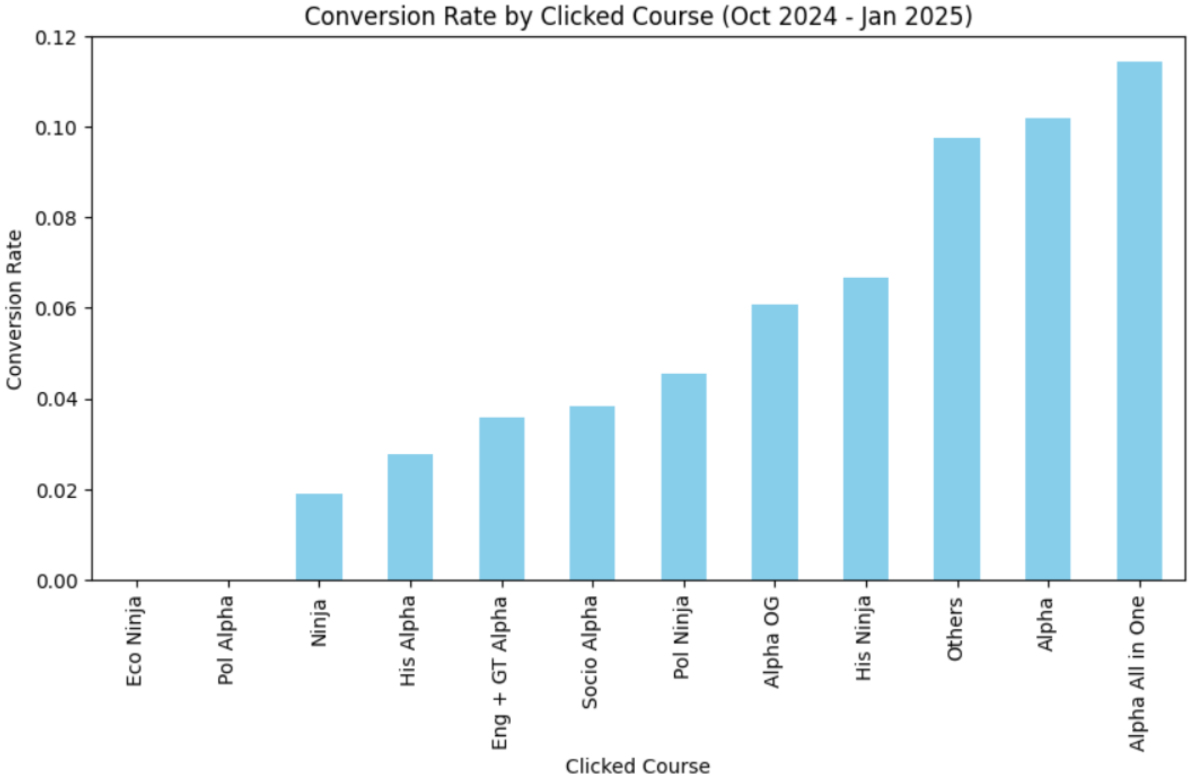


Graph 4: Conversion Rate by Lead Type

Although leads from WhatsApp, Instagram and YouTube are less in volume but their conversion rate is much higher as compared to other leads. Also, 'NORMAL' which happens to be the most frequent lead type is having very less conversion rate.

Conversion rate by Lead Type

- The motivation for plotting this graph was to analyse the conversion rates of different courses based on user clicks. By understanding which courses attract the most interest and lead to actual enrolments, we can refine marketing strategies and course offerings. This insight helps in prioritizing high-performing courses while identifying those that may need better promotion or curriculum enhancements.
- Management Team had intuition that ‘Alpha OG’ was most attractive course hence they suggested it to students who couldn’t decide themselves which course to take. But this graph suggests otherwise.

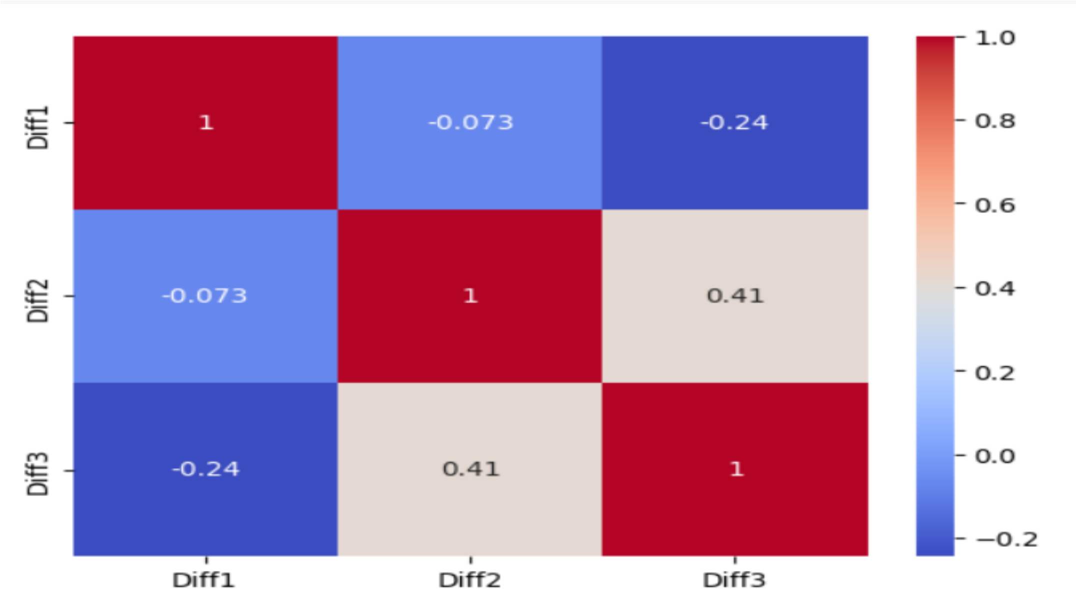


Graph 5: conversion Rate by Clicked course

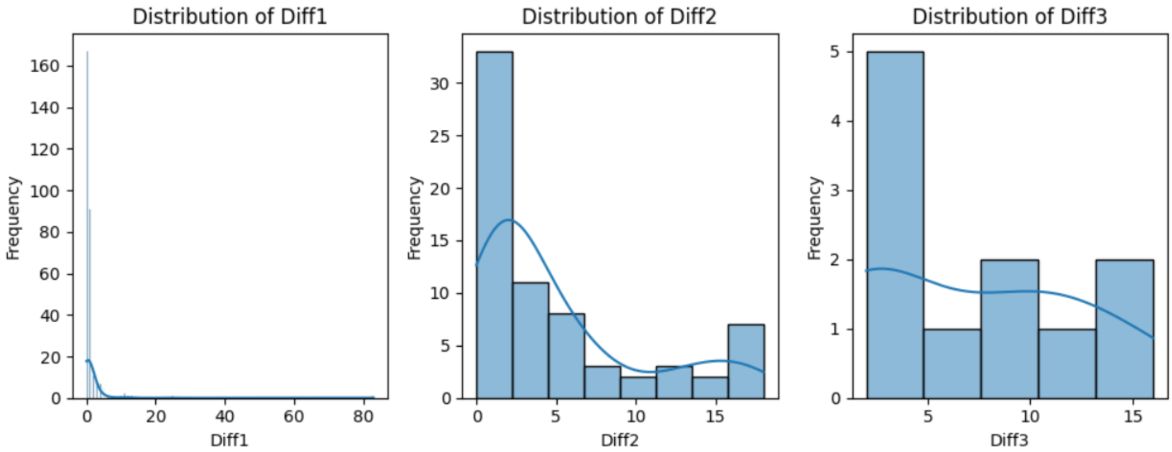
- The graph above shows that ‘Alpha All in One’ has highest Conversion rate. We know that Alpha OG has highest frequency but its conversion rate is lower than three courses.

Distribution and correlation between Numerical Features

To know about correlation between different calling gaps this heatmap was plotted.



Graph 6: Heatmap showing Correlation between Diff1, Diff2, Diff3



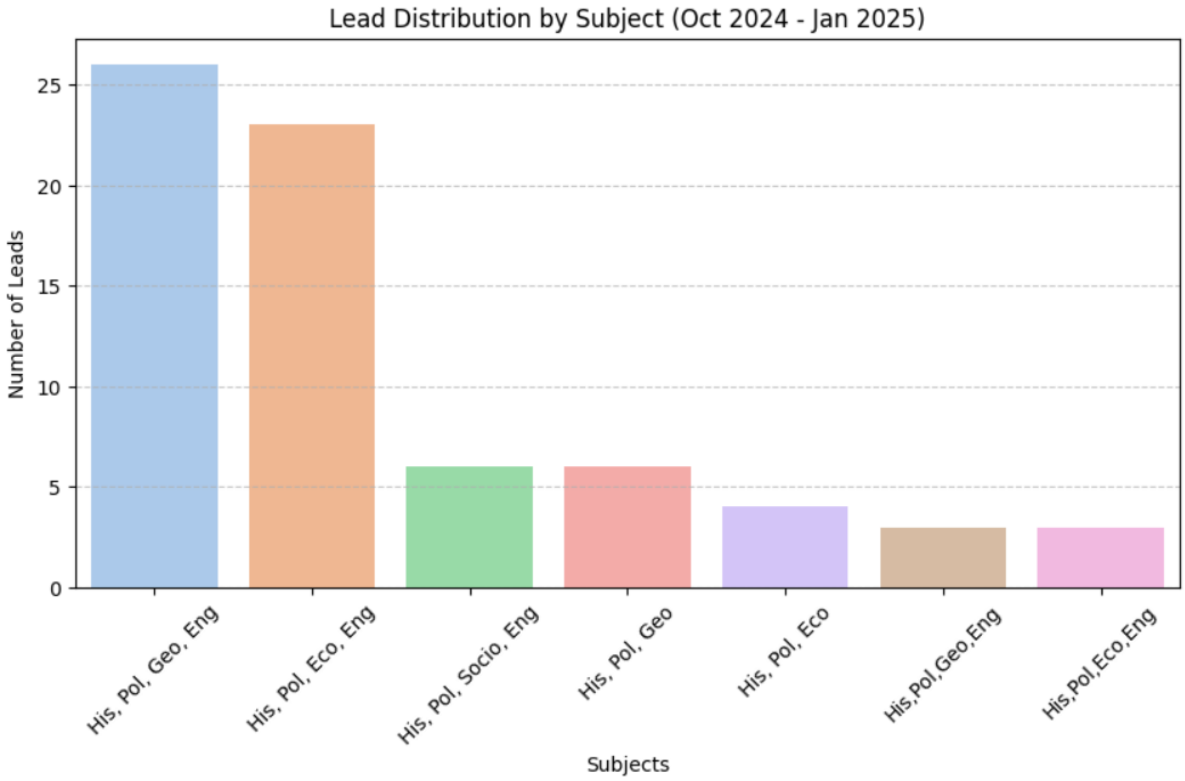
Graph 7: Distribution of Diff1, Diff2, Diff3

- Diff1 is highly right-skewed, leading to weak correlations with Diff2 (-0.073) and Diff3 (-0.24). Its extreme skewness may obscure linear relationships.

- Diff2 and Diff3 have a moderate positive correlation (0.41), indicating some common trend, supported by their relatively wider spread distributions.
- Skewness affects correlation strength, suggesting that transformations (e.g., log-scaling) might reveal stronger relationships between variables.

Lead Distribution by Subject

- The motivation for plotting this graph was to analyse the distribution of leads across different subject combinations. Understanding which subject combinations generate the most interest can help in optimizing course offerings and marketing efforts.

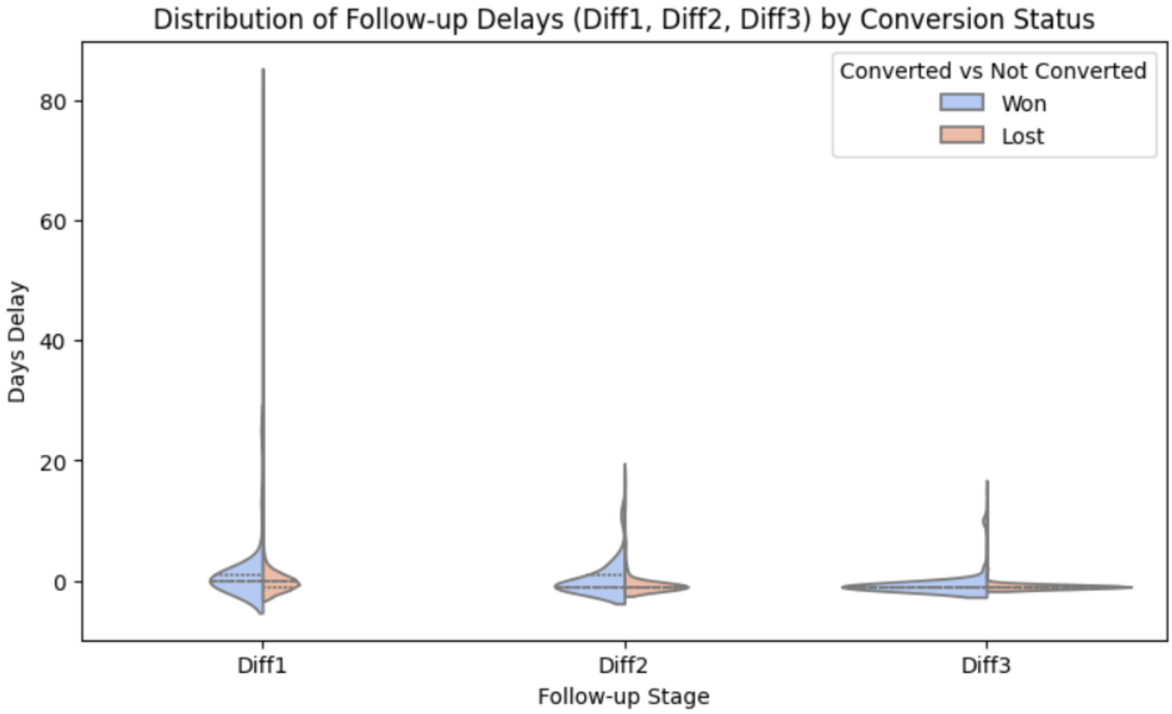


Graph 8: Lead Distribution by Subject

- Most popular subject combinations: "His, Pol, Geo, Eng" and "His, Pol, Eco, Eng" have the highest number of leads, indicating strong interest in these subject combinations.
- History Political English are the subject sought in every combinations.

Understanding Follow Up Delays

- To analyse the time gaps between follow-ups and their impact on conversion rates, this violin plot was plotted. This visualization helps in understanding whether longer delays correlate with lower conversion chances and if successful conversions tend to follow a specific follow-up timing pattern. By comparing distributions between converted and lost leads, we can determine optimal follow-up intervals and refine engagement strategies to maximize conversions.

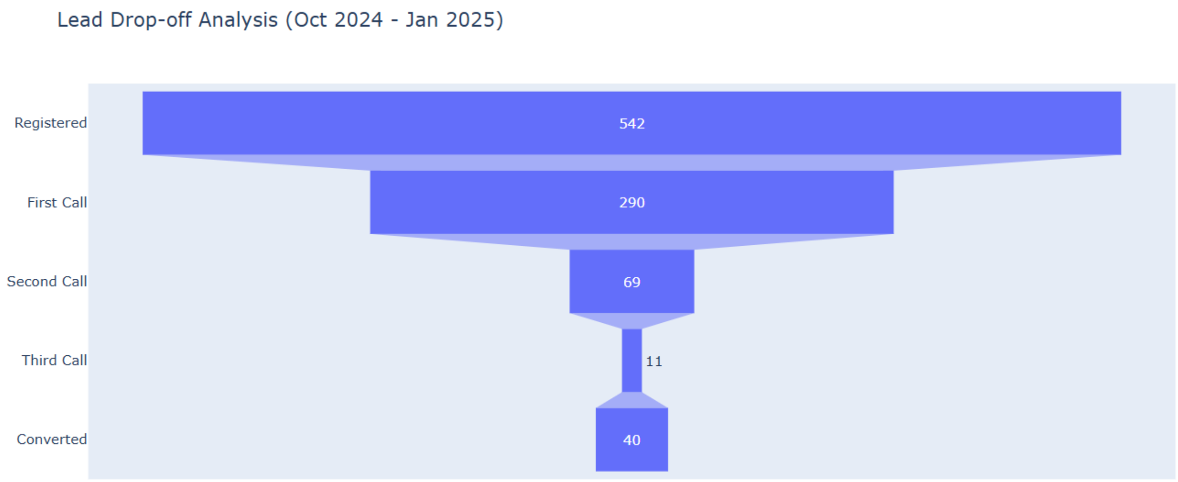


Graph 9: Distribution of Follow-up delays (Violin Plot)

- Longer follow-up delays (Diff1) are highly skewed with extreme outliers, suggesting that some leads experience significant delays, which may negatively impact conversion.
- Converted leads ("Won") tend to have slightly lower delays across all stages (Diff1, Diff2, Diff3) compared to lost leads, indicating that timely follow-ups might improve conversion rates.
- Variance decreases in later follow-up stages (Diff2, Diff3), meaning that delays are more consistent, but initial follow-up timing (Diff1) has a larger spread, making it a critical factor in conversion outcomes.

Lead Drop-off Analysis

- In order to visualize the stage-wise attrition of leads throughout the conversion process, Lead Drop-off Analysis Funnel was plotted. This helps in identifying where the majority of drop-offs occur, allowing for targeted improvements in follow-up strategies. A sharp decline at any stage indicates potential inefficiencies, such as ineffective communication, lack of engagement, or process delays. By analysing this funnel, we can refine outreach methods, enhance lead nurturing, and ultimately improve conversion rates.

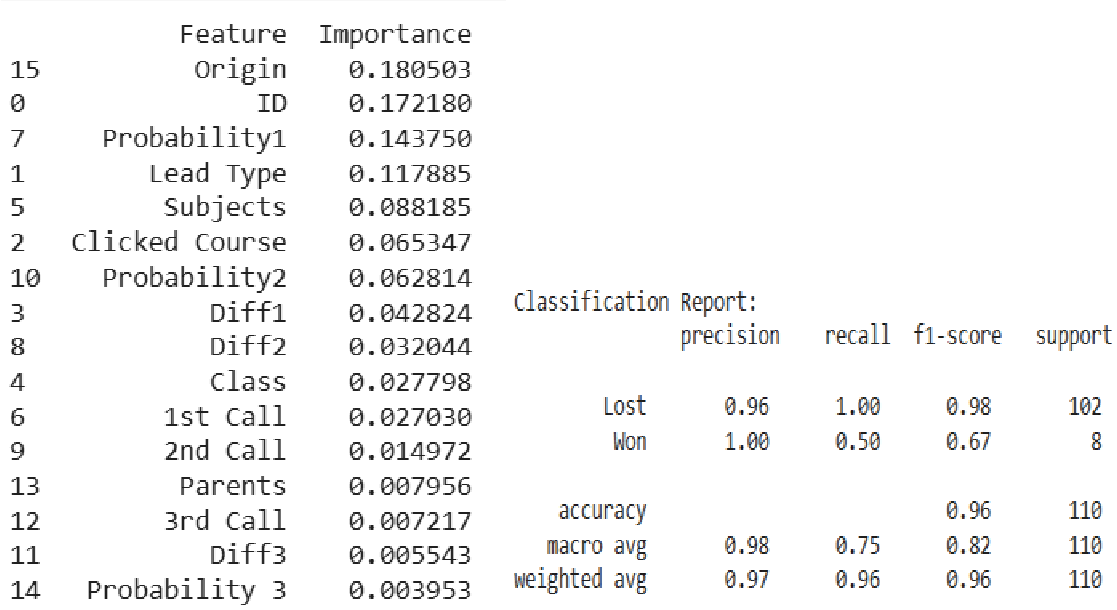


Graph 10: Lead Drop-Analysis Funnel

- Significant drop-off after registration: Only about 53 percent of registered leads proceed to the first call, highlighting the need for better engagement strategies post-registration.
- Drastic decline after the second call: Only 11 out of 69 leads make it to the third call, indicating a major drop-off point that requires improved follow-up or re-engagement tactics.
- Final conversion rate is low (7.27%): Only 40 leads out of 542 eventually convert, suggesting that refining the sales process and reducing early-stage drop-offs could significantly improve overall conversion.

Findings from ML Model

After training the ML model Random Forest Model was found to be best suited for the purpose, but due to high scores on train set suggests model overfitted on the data. Further when model was tested on validation set it showed similar results. Hence model was retained and would be tried to improved further as more data is collected. Following are Evaluation metrics.



```
             Feature  Importance
15            Origin    0.180503
0                 ID    0.172180
7       Probability1    0.143750
1          Lead Type    0.117885
5           Subjects    0.088185
2     Clicked Course    0.065347
10      Probability2    0.062814
3              Diff1    0.042824
8              Diff2    0.032044
4              Class    0.027798
6           1st Call    0.027030
9           2nd Call    0.014972
13           Parents    0.007956
12          3rd Call    0.007217
11             Diff3    0.005543
14     Probability 3    0.003953
```

```
Classification Report:
              precision    recall  f1-score   support

        Lost       0.96      1.00      0.98       102
         Won       1.00      0.50      0.67         8

    accuracy                           0.96       110
   macro avg       0.98      0.75      0.82       110
weighted avg       0.97      0.96      0.96       110
```

Feature Importance and Evaluation metrics

4 Interpretation of Results and Recommendations

Interpretation of Results

The analysis reveals key insights into the lead conversion process, highlighting both strengths and areas for improvement. The classification report shows that while the model achieves high accuracy (96%), it struggles with recall for converted leads, indicating a need for better detection of high-potential leads. The confusion matrix further confirms this by showing that some actual conversions were misclassified as lost, reinforcing the need for improving model sensitivity to the minority class. The AUC-ROC score (0.99) suggests that the model is highly capable of distinguishing between converted and lost leads, but the imbalance in class distribution might affect real-world performance. The lead drop-off analysis identifies major disengagement points, particularly after the first and second follow-up calls, signalling the importance of structured follow-up efforts. Additionally, lead source and subject-wise distribution graphs indicate that certain marketing channels and subject combinations perform better in attracting engaged leads.

Recommendations

- The lead drop-off funnel shows that only about 53 percent of leads proceed to the first call. Hence, I suggest that more sales persons should be hired and nearly all significant leads should be reached for first call.
- As the violin plot of follow-up delays suggests timely follow-ups increase conversion rates, hence structured follow-up system ensuring calls are made within 24-48 hours of registration should be implemented.
- The lead type conversion analysis shows that Instagram, WhatsApp, and YouTube leads have higher conversion rates, hence ad expenditure and engagement efforts on these high-performing channels should be increased
- Leads from Website Queries and App Downloads have lower conversion rates. Revamp landing pages, improve chatbot interactions, and offer incentives (discounts/free trials) for these leads.

- The lead distribution by subject shows that History, Political Science, and English attract the most leads. Create personalized messaging and course recommendations to convert more leads from these subjects.
- Ensure all important features (like contact info, demographics, past interactions, and engagement history) are fully captured and accurately recorded to improve predictive accuracy. Missing or incomplete data can reduce the effectiveness of the model, leading to misclassification of high-potential leads.
- Include columns such as Board, State etc and include economical information of students.
- Introducing time-sensitive discounts can create a sense of urgency and drive faster conversions. This could be tried on leads marked as ‘high chance’.