Lead Scoring Case Study: Improving Conversion Prediction – X Education

This presentation outlines our comprehensive analysis of lead conversion factors for an educational company. We examined a dataset of 9,240 leads with 37 variables to build a predictive model that identifies which leads are most likely to convert into paying customers.

Our analysis follows a systematic approach: data cleaning, exploratory data analysis, feature engineering, model building with logistic regression, and evaluation. The findings provide actionable insights to optimize the lead nurturing process and improve conversion rates.

by Duy Pham and Vy Pham

Project Overview and Objectives

1 Business Context

The education company X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines. When these individuals land on the website, they might browse courses or fill out forms for the course or watch videos.

2 Problem Statement

The conversion rate of leads to paying customers is currently low at around 30%. The company needs to identify the most promising leads (hot leads) to focus their sales efforts efficiently.

3 Project Goal

Build a lead scoring model that assigns a score to each lead, indicating the likelihood of conversion. This will help the sales team prioritize leads with higher conversion probability.

Dataset Description

Dataset Size

The initial dataset contained 9,240 leads with 37 variables, including both categorical and numerical features. After cleaning, we retained approximately 69% of the original data (6,373 records).

The target variable is "Converted," which indicates whether a lead eventually became a paying customer (1) or not (0).

Key Variables

The dataset includes lead source information, engagement metrics (time spent on website, total visits), lead details (occupation, specialization), and interaction history (last activity, notable actions).

We also derived new metrics like average time per visit and average time per page view to better understand user engagement patterns.

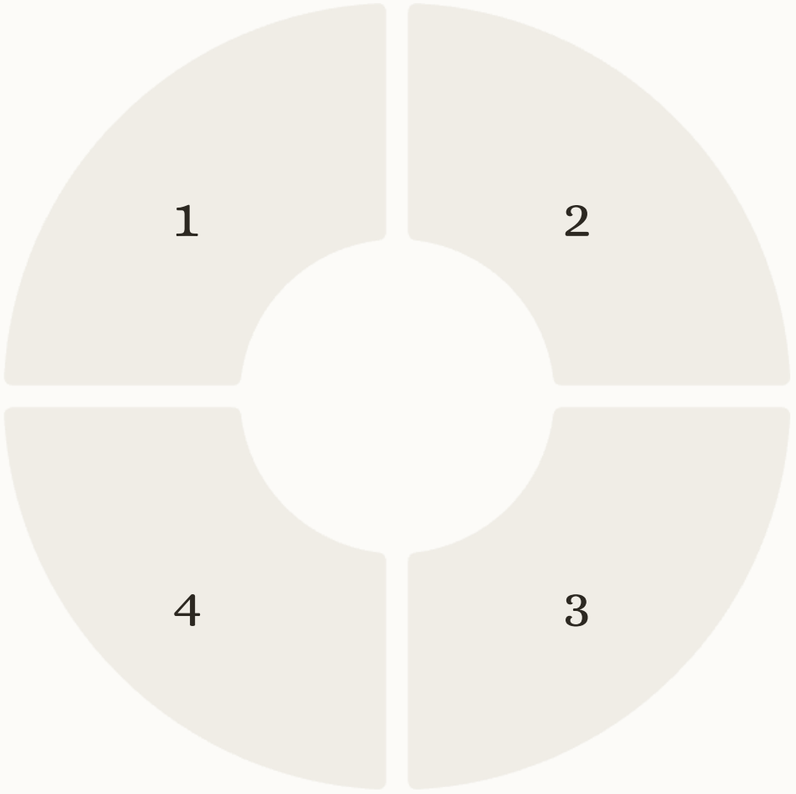
Data Cleaning: Handling Missing Values

Step 1: Identify Missing Data

Several columns had significant missing values. For example, "Lead Quality" had 51.6% missing values, while "Asymmetrique Activity Index" had 45.6% missing values.

Step 2: Remove Low-Value Columns

Columns with over 3,000 missing values (>32% of the dataset) were dropped as they provided insufficient information for meaningful analysis.



Step 3: Address Remaining Nulls

For key variables with fewer missing values, we removed the specific rows with nulls rather than losing the entire feature. This balanced data completeness with information retention.

Step 4: Handle 'Select' Values

Many categorical variables contained 'Select' as a value, indicating no selection was made. These were treated as missing values during the analysis.

Handling Categorical Variables

Identification

1

We identified all object-type columns in the dataset that needed to be converted to numerical format for modeling. This included variables like Lead Origin, Lead Source, and Specialization.

Value Assessment

2

For each categorical variable, we examined value counts to understand distribution. Variables with highly imbalanced classes (e.g., 99% "No" responses) were dropped as they provided little discriminatory power.

Dummy Variable Creation

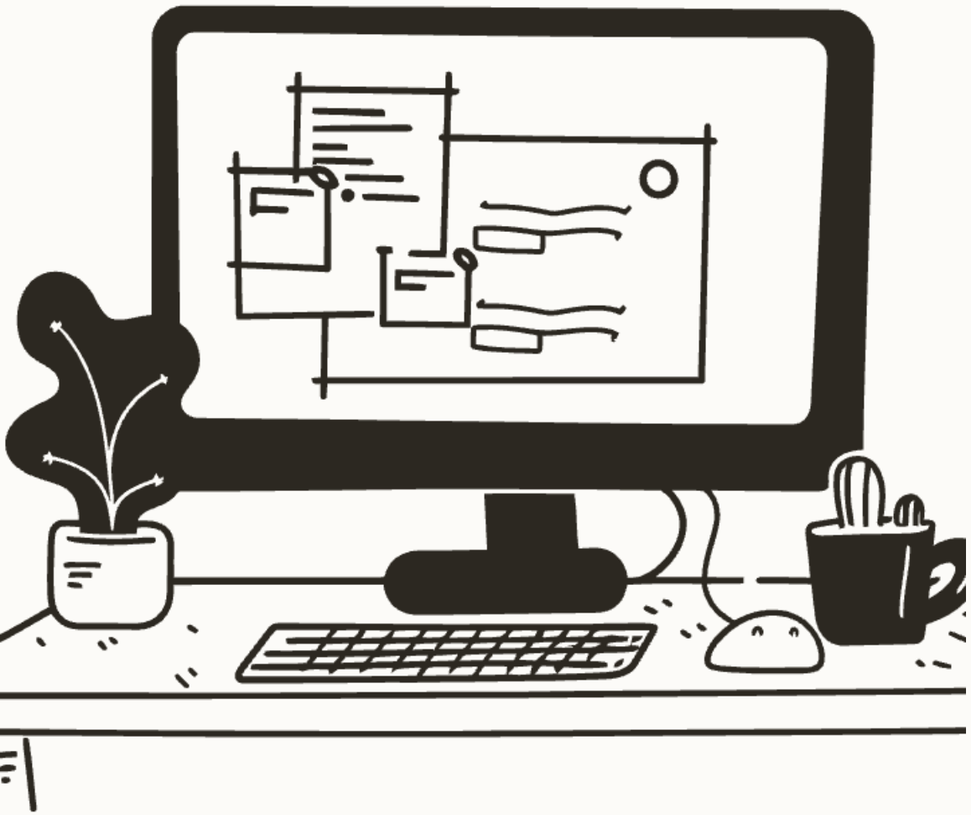
3

We used one-hot encoding to convert categorical variables into binary indicators. For multilevel variables like Specialization, we created multiple binary columns, dropping the first level to avoid multicollinearity.

Final Preparation

4

After encoding, we had a much wider dataset with 77 predictor columns ready for modeling, with all categorical information properly represented numerically.



Feature Engineering

Derived Engagement Metrics

We created new features to better capture user engagement: "Avg_Time_Per_Visit" (Total Time Spent / Total Visits) and "Avg_Time_Per_View" (Total Time Spent / Total Page Views). These metrics provide deeper insights into how intensively leads interact with the website.

Feature Scaling

Numerical variables with different scales were normalized using MinMaxScaler. This ensures that variables like "Total Time Spent on Website" (ranging from 0-2272) don't overshadow variables with smaller ranges during model training.

Feature Selection

We used Recursive Feature Elimination (RFE) to identify the most relevant predictors, starting with all variables and iteratively removing the least important ones until reaching 15 optimal predictors.

Correlation Analysis

1

Strong Positive Correlations

Total Time Spent on Website and Conversion showed a positive correlation of 0.32, indicating that leads who spend more time exploring the site are more likely to convert. Similarly, TotalVisits and Page Views Per Visit showed a moderate correlation of 0.49.

2

Channel Analysis

Lead sources showed varying correlation with conversion. The "Welingak Website" channel had one of the strongest positive associations with conversion, while certain other channels showed negative correlations.

3

Activity Correlations

Last activities like "Had a Phone Conversation" and "SMS Sent" correlated positively with conversion, suggesting that direct communication with leads significantly improves conversion chances.

Model Building Approach

Data Splitting

We divided the dataset into training (70%) and testing (30%) sets to ensure unbiased evaluation. This gave us 4,461 observations for training and 1,912 for testing the model.

Feature Selection

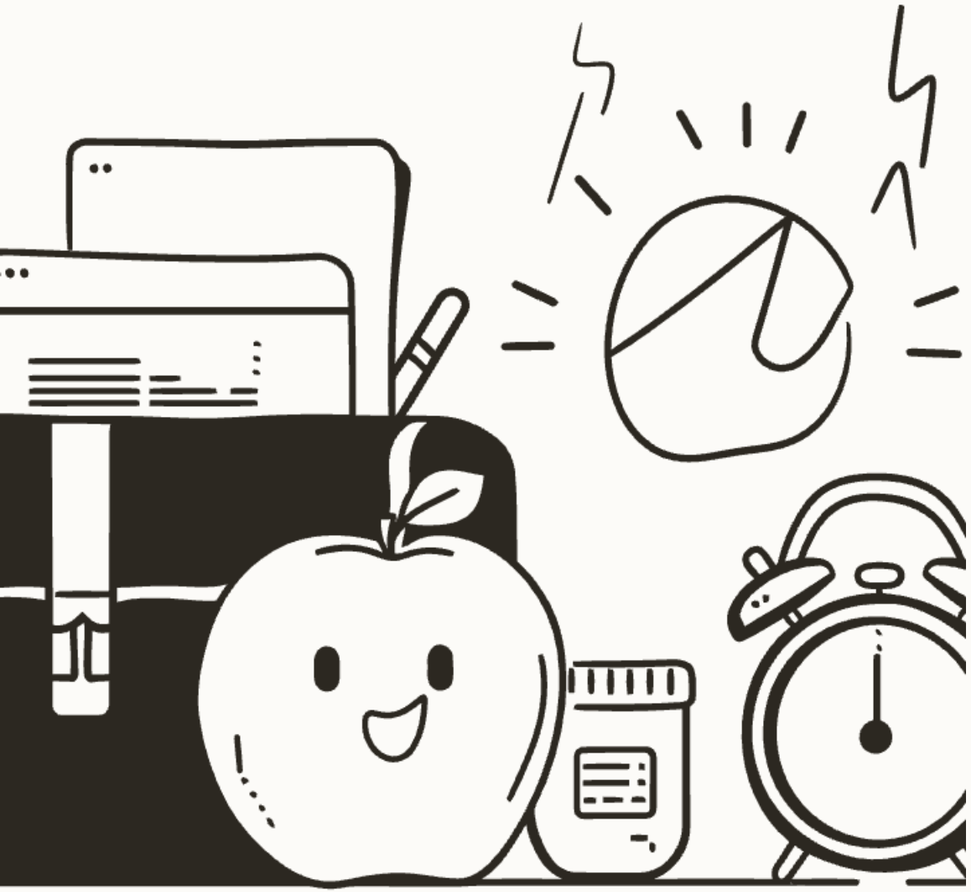
Using Recursive Feature Elimination (RFE), we identified the 15 most important predictors from the large feature set. This helps create a simpler, more interpretable model while maintaining predictive power.

Logistic Regression

We chose logistic regression as our modeling technique since it provides both predictions and interpretable coefficients. This allows us to understand which factors most heavily influence lead conversion.

Model Refinement

We iteratively improved the model by removing variables with high p-values (>0.05) or high Variance Inflation Factors (VIF), which indicate multicollinearity. This resulted in our final model with 11 predictors.



Key Predictors of Lead Conversion

Time Spent on Website

This emerged as the strongest predictor with a coefficient of 5.76. For every standardized unit increase in time spent on the website, the log-odds of conversion increase by 5.76, translating to significantly higher conversion probability.

Lead Add Form

Leads who came through the Lead Add Form had much higher conversion rates (coefficient: 4.02). This suggests that leads who take the time to fill out detailed forms demonstrate higher intent and commitment.

Phone Conversation

Leads who had a phone conversation with the sales team were significantly more likely to convert (coefficient: 2.84). This highlights the importance of direct, personal communication in the conversion process.

Negative Predictors of Conversion

Average Time Per Visit

Interestingly, while total time spent on the website is positive, the average time per visit shows a negative coefficient (-3.79). This suggests that multiple shorter visits are more valuable than fewer longer visits. It may indicate that leads who need to think about the decision over multiple sessions are more serious.

Email Preferences

Leads who selected "Do Not Email" were significantly less likely to convert (coefficient: -1.42). This suggests that willingness to receive communications is an important indicator of interest and openness to the offering.

Student Occupation

Leads who identified as students had lower conversion rates (coefficient: -2.37), as did unemployed individuals (-2.53). This likely reflects financial constraints or different decision-making priorities among these groups.

Model Performance: Accuracy and Precision

79.1%

Accuracy (Default)

Using a standard 0.5 probability threshold

79.5%

Accuracy (Optimized on Train set)

With threshold at 0.42 (optimal cutoff)

79.2%

Precision

Among predicted conversions, actual rate

77.8%

Recall

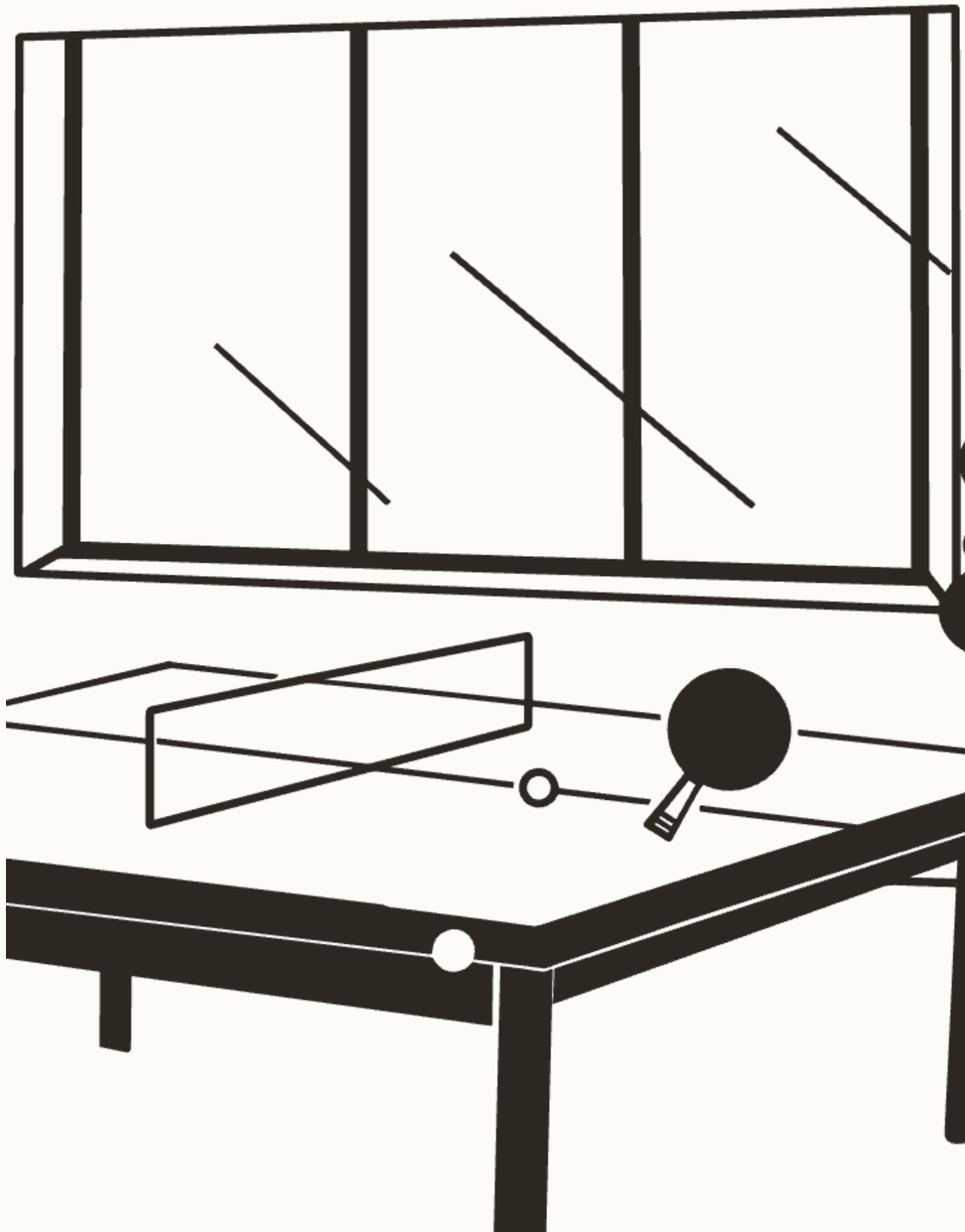
Ability to find all actual conversions

Our model demonstrates strong performance across multiple metrics. The accuracy increased from 79.1% to 79.5% when we optimized the threshold from 0.5 to 0.42. This means that for approximately 8 out of 10 leads, our model correctly predicts whether they will convert or not. The precision of 79.2% indicates that when our model predicts a lead will convert, it's right about 78% of the time.

ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate at various threshold settings. Our model achieved an Area Under the Curve (AUC) of 0.86, which indicates excellent discriminative ability. This means the model is very good at distinguishing between leads that will convert and those that won't.

The ROC curve helped us identify the optimal probability threshold of 0.42, which balances sensitivity (finding actual conversions) and specificity (avoiding false positives). This threshold optimization is crucial for business applications where different types of errors may have different costs.

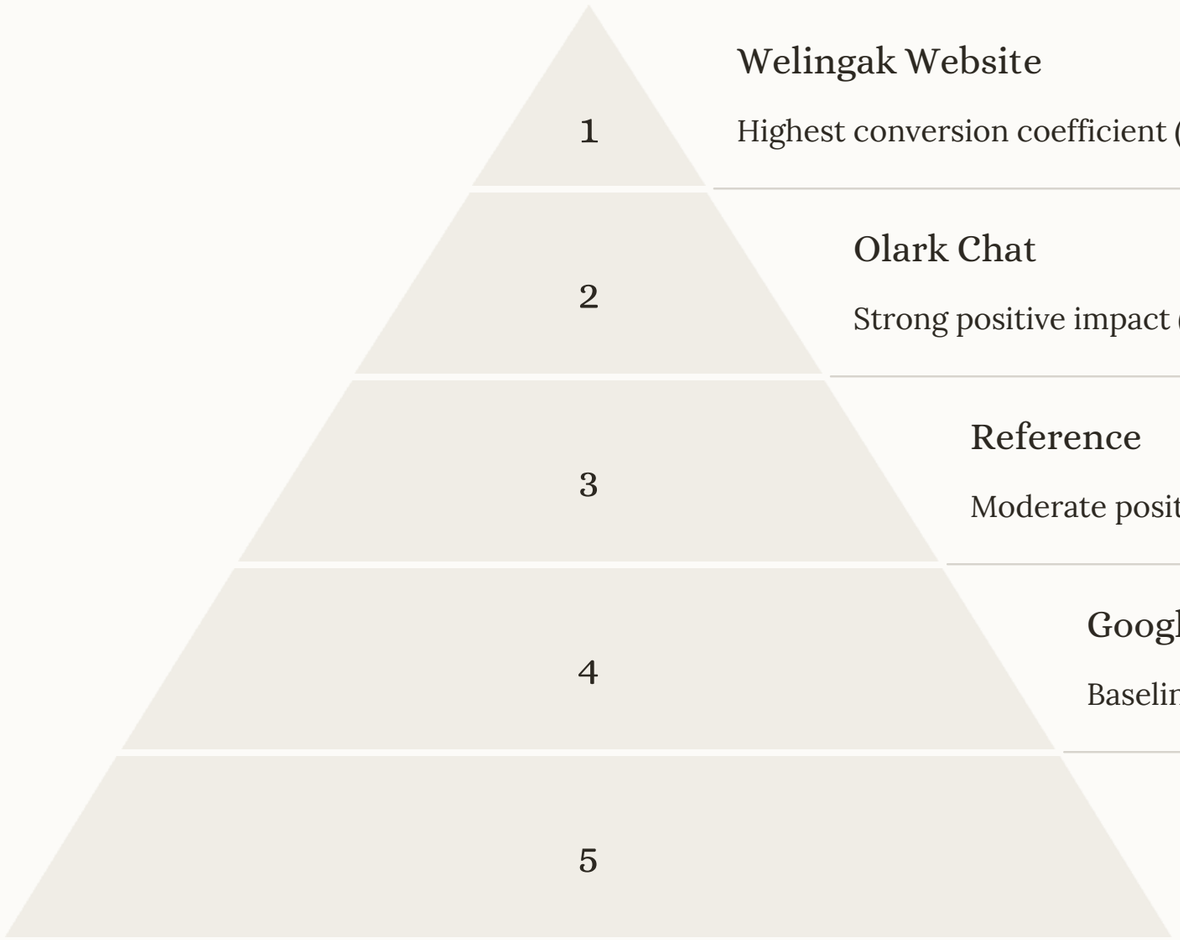


Threshold Optimization: Precision-Recall Tradeoff

When determining the optimal probability threshold for classifying leads, we faced a critical tradeoff between precision and recall. Lower thresholds increase recall (finding more potential conversions) but decrease precision (more false positives). Higher thresholds do the opposite.

After analyzing this tradeoff, we selected 0.44 as our final threshold for the test set, which achieved the best balance of precision (78.2%) and recall (77.2%). This optimization ensures that sales resources are allocated efficiently while still capturing a high percentage of potential conversions.

Lead Source Impact on Conversion



Welingak Website

Highest conversion coefficient (2.15)

Olark Chat

Strong positive impact (1.28)

Reference

Moderate positive impact

Google

Baseline conversion rate

Direct Traffic

Lower conversion potential

Different lead sources showed varying effectiveness in generating conversions. The Welingak Website emerged as the most valuable channel with a coefficient of 2.15, indicating leads from this source are significantly more likely to convert. Olark Chat was also highly effective (coefficient: 1.28), suggesting that interactive engagement channels produce higher-quality leads.

Communication Channel Effectiveness

Phone Conversations

Having a phone conversation with a lead was one of the strongest predictors of conversion (coefficient: 2.84). This direct, personalized communication method creates a connection that significantly improves conversion chances.

SMS Communication

Leads who received SMS communications showed higher conversion rates (coefficient: 1.16). This suggests that text messaging is an effective channel for engaging with potential customers and driving conversions.

Email Engagement

Leads who opened emails showed moderate conversion increases. Conversely, those who opted out of emails were much less likely to convert (-1.42). This highlights the importance of email as a nurturing channel.

Demographic Insights: Occupation Impact

0.68

Working Professional

Highest conversion propensity

-2.53

Unemployed

Lowest conversion likelihood

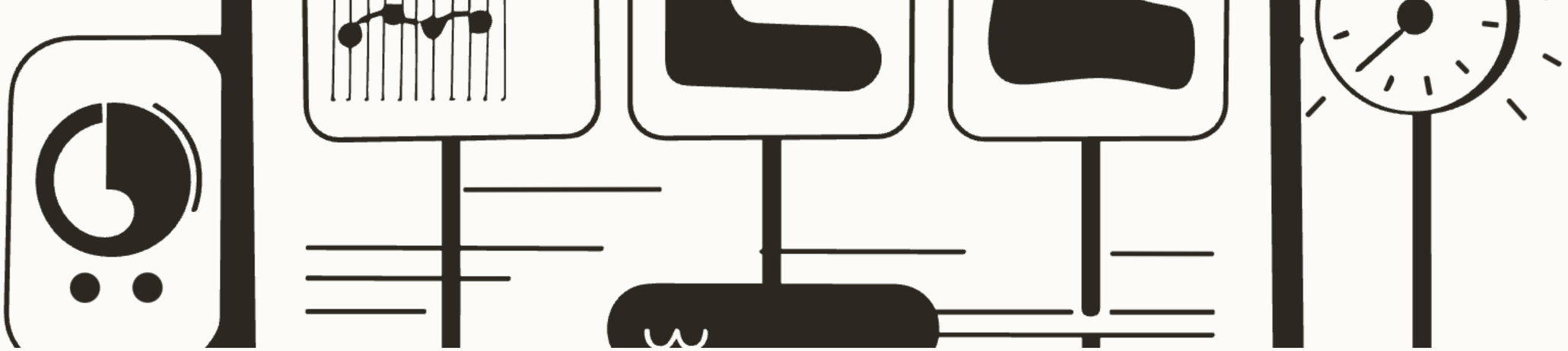
Positive Conversion Indicators

Working professionals showed the highest propensity to convert (coefficient: 0.68), suggesting they have both the decision-making authority and financial means to purchase courses. Businessmen showed a slight positive impact (coefficient: 0.12) on conversion likelihood.

Negative Conversion Indicators

Students and unemployed individuals were significantly less likely to convert (coefficients: -2.37 and -2.53 respectively). Housewives also showed a slightly negative tendency (coefficient: -0.25).

This data suggests that targeting efforts should prioritize working professionals. For student segments, different pricing or financing options might be needed to improve conversion rates.



Engagement Metrics: Time and Visits

Total Time Spent (Positive)

The total time a lead spends on the website was the strongest positive predictor of conversion (coefficient: 5.76). This indicates that engaging content that keeps visitors on the site longer directly contributes to higher conversion rates.

Average Time Per Visit (Negative)

Interestingly, the average time per visit showed a negative relationship with conversion (coefficient: -3.79). This suggests that multiple shorter visits may be more valuable than fewer lengthy visits.

Visit Pattern Implications

The combination of these findings suggests an ideal engagement pattern: leads who return to the site multiple times, even for shorter sessions, show higher intent and conversion probability than those who browse extensively in a single visit.

Key Recommendations

1

Optimize Lead Sources

Increase investment in high-converting channels like Welingak Website and Olark Chat. Consider reducing spend on lower-performing channels or revamping their approach to improve quality.

2

Enhance Communication Strategy

Prioritize phone conversations for high-potential leads. Implement a robust SMS communication strategy as a complement to email marketing. Ensure email content is engaging to improve open rates.

3

Improve Website Engagement

Design the website to encourage repeat visits rather than one-time lengthy browsing. Create content that builds interest over multiple sessions. Use remarketing to bring visitors back to the site.

4

Segment Marketing Approach

Tailor marketing strategies based on occupation segments. Focus primary efforts on working professionals while developing special offers or financing options for student and unemployed segments.

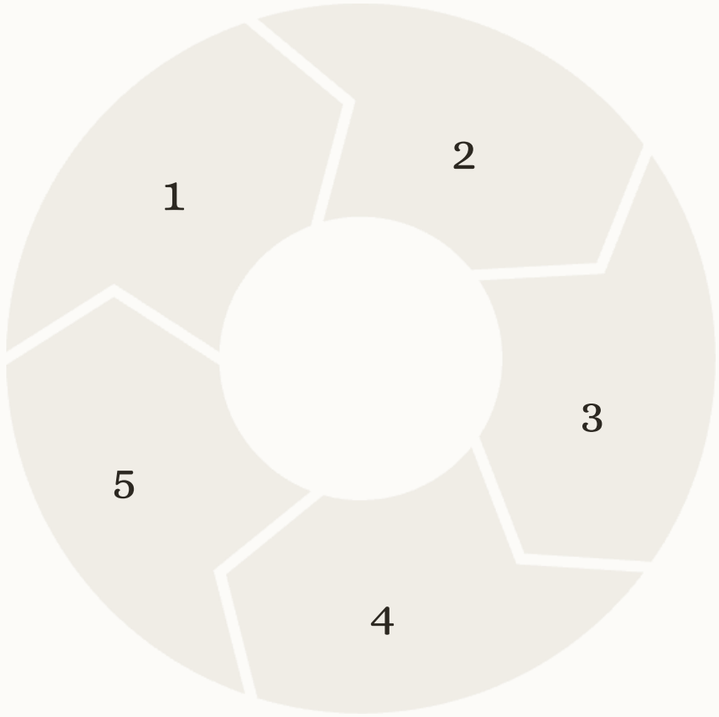
Implementation Strategy & Next Steps

Score Implementation

Deploy the lead scoring model in production to automatically generate conversion probability scores for all new leads

Model Refinement

Continuously update the model with new data and refine based on performance feedback



Resource Optimization

Integrate scores into CRM systems and redesign the lead assignment process to prioritize high-potential leads. Adjust the threshold depends on the manpower and strategies

Marketing Channel Optimization

Reallocate marketing budget based on channel performance insights from the model

Performance Monitoring

Track key metrics including model accuracy, lead-to-customer conversion rates, and sales efficiency

To maximize the impact of our lead scoring model, we recommend a phased implementation approach. Begin by using the model to score and prioritize incoming leads, followed by integrating these scores into your existing CRM and sales processes. Continuously monitor the model's performance and refine it as new data becomes available.
Moreover, base on the strategies of each period, the threshold can be adjusted to adapt with the resource at the same time (base on subject questions)