Lead Scoring Case Study

Data-Driven Insights and Model-Based Strategies

Submitted By - Shimran Panigrahi, Sindhuja V, Sini Keloth

Business Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Although the company successfully generates numerous leads through various marketing channels, a significant portion of these leads do not convert into paying customers. As the sales team spends considerable time contacting leads that may not be viable, there is an urgent need to enhance the efficiency of the lead conversion process.

Specifically, the company faces the challenge of identifying high-potential leads, referred to as 'Hot Leads,' which are more likely to convert into sales. The current funnel demonstrates a large number of potential leads at the top, but only a small fraction successfully progresses to becoming customers at the bottom of the funnel.

The primary goal is to construct a logistic regression model that can score leads on a scale of 0 to 100, allowing the sales team to focus their efforts on leads with the highest conversion potential. The CEO has set a target lead conversion rate at around 80%, emphasizing the importance of targeting the right leads to meet this goal.

Business Objective

Improve Lead Conversion Rate

The foremost objective is to increase the lead conversion rate from 30% to 80% by effectively identifying and prioritizing 'Hot Leads.' This enhancement should lead to an increase in revenue generated from course sales.

Optimize Sales Efforts

By developing a scoring system for leads, the company aims to streamline the sales process. Sales personnel will invest their time and resources in leads that are statistically more likely to convert, hence reducing the time spent on less promising leads and improving overall productivity.

Data-Driven Decision Making

Establish a data-driven approach to lead management by utilizing historical data and predictive modeling. This will allow the company to make informed decisions regarding marketing strategies and customer engagement.

Scalability and Future Adaptability

Build a flexible logistic regression model that can be adjusted to accommodate evolving company requirements and market conditions. The model should be robust enough to reassess lead scoring criteria as new data and metrics become available.

Comprehensive Reporting

Deliver a well-structured report and presentation that summarizes the methodology employed, results obtained, and actionable insights derived from the analysis. This documentation should facilitate stakeholder understanding and buy-in for the new lead prioritization approach.

Analysis Approach

The analysis approach involved a comprehensive process of data collection, preprocessing, feature engineering, model selection, and evaluation. The focus was on leveraging machine learning techniques to build a predictive model that could accurately assign scores to leads based on their likelihood of conversion.

1

Data Collection & Preprocessing

Thoroughly analyzed the dataset to understand its structure and content. Clean the provided leads dataset, addressing any null values represented by 'Select' in categorical variables. This step is crucial for ensuring the data's integrity.

2

Exploratory Data Analysis (EDA)

Conduct an EDA to understand the relationships between various attributes (Lead Source, Total Time Spent, Last Activity, etc.) and the lead conversion outcome. Visualizations will help in interpreting these relationships.

3

Model Development

Identified the most relevant features using Recursive Feature Elimination (RFE).Build a logistic regression model utilizing the cleaned dataset to predict lead conversion probabilities. The model will assign a score to each lead, indicating its likelihood of converting into a paying customer.

4

Model Evaluation

Assessed model performance using accuracy, sensitivity, specificity and the ROC-AUC curve. This will help determine the model's effectiveness in distinguishing between hot and cold leads.

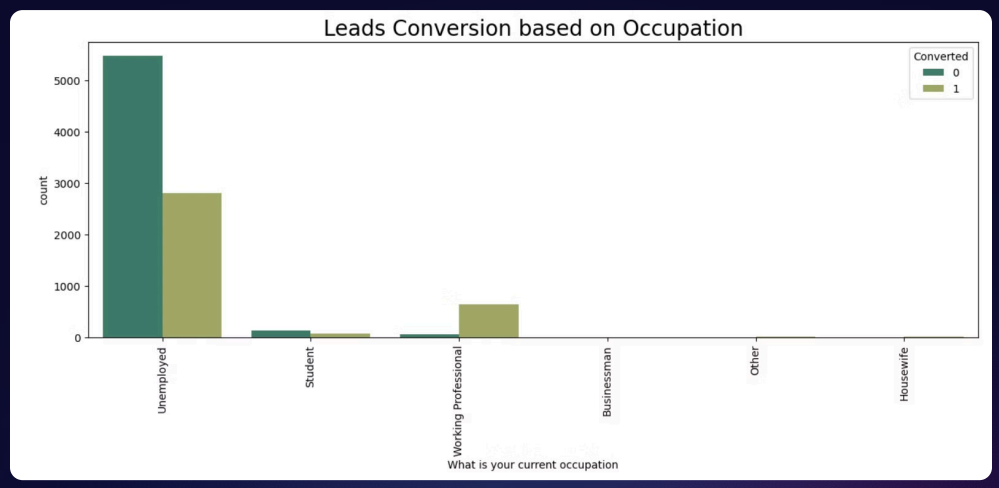
5

Recommendations and Reporting

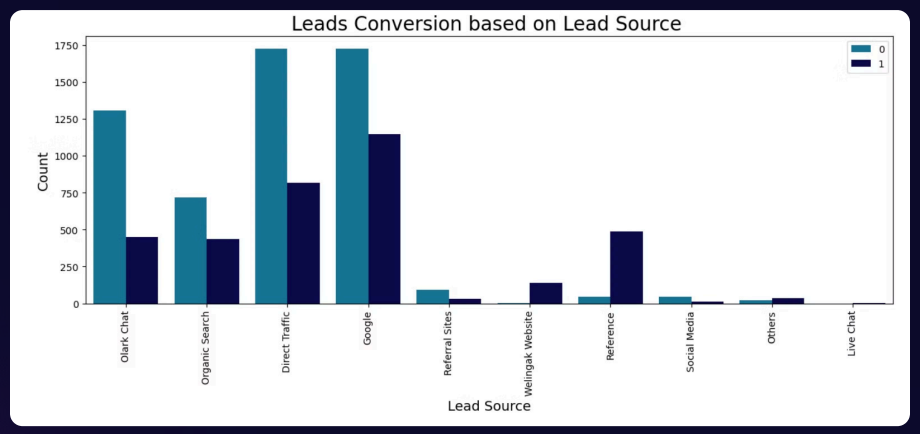
Prepare a detailed report, summarizing the analysis process, results obtained, and recommendations for the sales team. Include visualizations to aid in understanding key insights and findings.

Key Insights from EDA

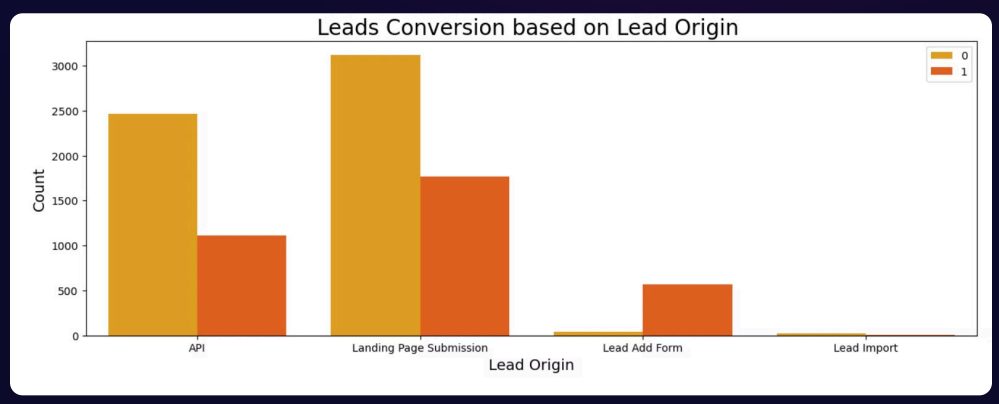
Exploratory data analysis (EDA) revealed valuable insights into the characteristics of high-converting leads. Analyzing data patterns and trends helped identify key features that contributed to lead conversion.



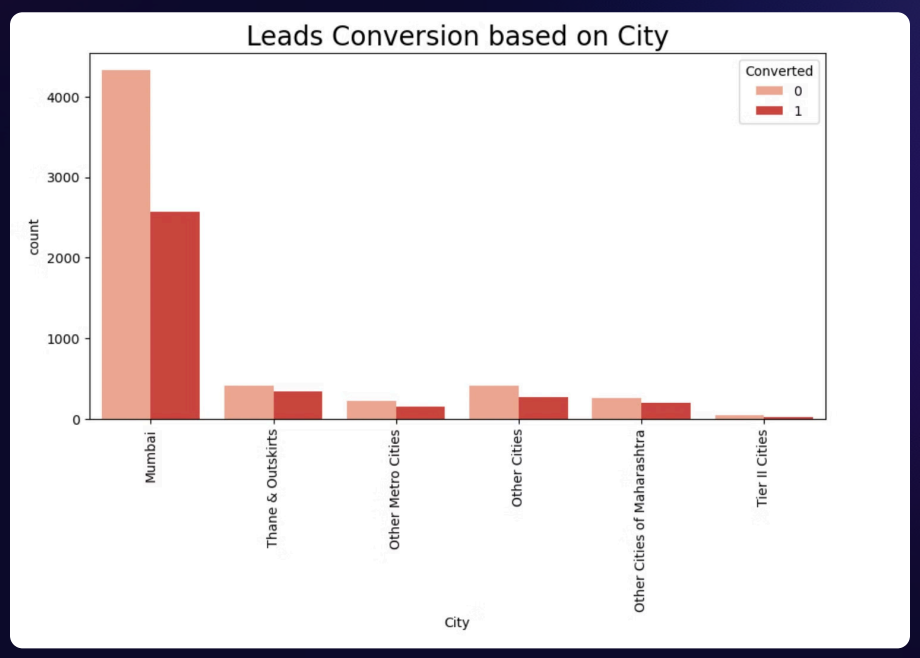
- **Unemployed Individuals:** The majority of leads come from unemployed individuals. This suggests that X Education's marketing efforts might be particularly effective in reaching individuals seeking career transitions or upskilling.
- **Occupation:** The majority of leads were unemployed. This provides a further indication of X Education's potential to reach individuals seeking career changes and upskilling opportunities.



- **Dominant Lead Sources:** Leads generated through Google and direct traffic are the most common. This indicates that X Education's online presence and search engine optimization are performing well.
- **Welingak Website:** The Welingak website has the highest conversion rate. This suggests that X Education should explore potential partnerships or strategies to increase leads from this specific source.



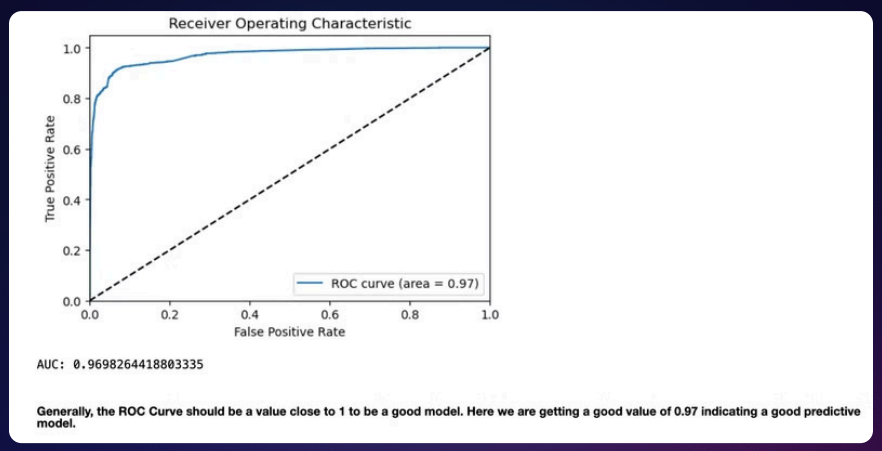
- **Lead Add Form:** "Lead Add Form" shows a high conversion rate, indicating effectiveness. However, the low volume of leads generated through this source suggests exploring methods to drive more leads through this channel.

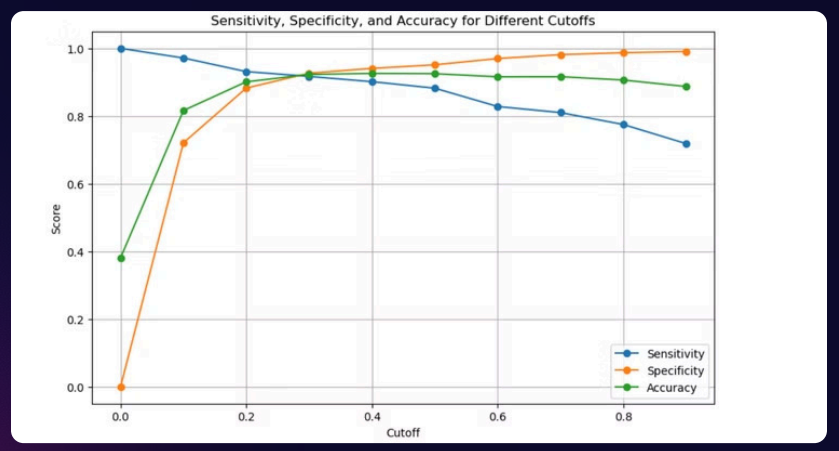


- **City:** The analysis revealed that Mumbai is the most frequent city, so consider focusing on targeted marketing efforts within that region.

Model Performance

The lead scoring model exhibited strong performance in predicting lead conversion. The ROC curve has a values of 0.97 indicating a good predictive model. The model achieved an accuracy score of 92.29%, indicating that it correctly classified . The Sensitivity has a score of 91.70% and Specificity has a score of 92.65% suggesting that the model is effective in identifying both positive and negative cases accurately, making it a well-balanced model suitable for applications where both types of errors need to be minimized.





1

Accuracy

Training Data: 92.29%, **Test Data:** 92.78%

The model is highly accurate in predicting lead conversion. It correctly predicts around 92% of the cases in both the training and test data. This is a positive sign that the model is generally reliable in its predictions.

2

Sensitivity

Training Data: 91.70%, **Test Data:** 91.98%

The model has high sensitivity. It correctly identifies a significant percentage of actual conversions. This is crucial for X Education, as it means the model is good at identifying potential paying customers and minimizing the risk of missing out on valuable leads.

3

Specificity

Training Data: 92.65%, **Test Data:** 93.26%

The model also exhibits high specificity. It effectively identifies leads who are unlikely to convert. This is important because it helps X Education avoid wasting resources on leads with a low probability of conversion.

The model demonstrates a strong overall performance, with high accuracy, sensitivity, and specificity. This suggests that X Education can confidently rely on the model's predictions to prioritize leads and focus their sales efforts on those most likely to convert.

Conclusion

The analysis successfully developed a robust Logistic Regression model for predicting lead conversion for X Education. The model demonstrates excellent performance on both training and test data, achieving high accuracy, sensitivity, and specificity, highlighting its ability to accurately predict lead conversion. Key findings include:

- The majority of leads come from unemployed individuals, suggesting a focus on career transition and upskilling marketing.
- Google and direct traffic are the most prominent lead sources.
- The Welingak website showcases a high conversion rate, indicating a potential partnership opportunity.
- The "Lead Add Form" demonstrates a high conversion rate but needs more lead generation.
- The model's high accuracy (92.29% on training, 92.78% on testing), sensitivity (91.70% on training, 91.98% on testing), and specificity (92.65% on training, 93.26% on testing) confirm its ability to identify potential customers and avoid wasting resources on unqualified leads.

This model allows X Education to focus their sales and marketing efforts on the most promising leads, leading to higher conversion rates and improved business outcomes. Further exploration of advanced algorithms and hyper-parameter tuning could potentially enhance the model's performance even more.