Lead Scoring - Case Study

Project Summary

The project focuses on building a predictive model to optimize lead conversion for X Education, an online course provider. The primary goal was to identify 'Hot Leads' with a higher likelihood of conversion, helping the sales team prioritize outreach and improve efficiency.

1. Data Exploration and Cleaning: The initial dataset was explored to understand its structure. The dataset included various features such as Lead Origin, Lead Source, Total Visits, Time Spent on the Website, and more. During this stage, missing values were identified and handled appropriately. Categorical variables were transformed using one-hot encoding, and numerical variables were standardized. Outlier detection and removal ensured the dataset was clean and ready for modeling.

2. Feature Engineering: Key derived metrics and new features were created to better capture patterns in the data. This included calculating engagement scores and grouping categories with low frequencies. Feature selection methods, such as correlation matrices and recursive feature elimination (RFE), were used to identify the most predictive variables.

3. Model Building: The predictive model was developed using a combination of logistic regression and gradient-boosted decision trees (XGBoost). The logistic regression model provided interpretable coefficients, while XGBoost helped capture non-linear relationships. Hyper-parameter tuning was performed using grid search cross-validation to find the optimal combination of parameters.

4. Model Evaluation: The model was evaluated on both the training and test datasets. Key performance metrics included accuracy, sensitivity, and specificity. The model achieved a little over 80% accuracy on both sets, indicating strong generalization. The ROC curve and AUC score were used to assess the model's discriminatory power, while precision-recall curves provided insight into handling class imbalance.

5. Probability Cutoff Optimization: The model's probability outputs were analyzed to identify the optimal cutoff point. By plotting accuracy, sensitivity, and specificity against different probability thresholds, the intersection point was identified as the most balanced cutoff. This allows the sales team to maximize conversions while minimizing false positives and negatives.

6. Insights and Learnings: One major takeaway was the importance of feature selection in preventing over-fitting. The model demonstrated that focusing on a subset of highly relevant variables significantly improved performance. Key insights revealed that leads originating from "Welingak Websites," "References," and "Olark Chat," as well as working professionals who spent more time on the website, had a higher likelihood of conversion. In contrast, leads

whose last activity was "Olark Chat Conversation," those from "Landing Page Submissions," and those who chose "Do Not Email" as "Yes" were less likely to convert. This data-driven filtering process enabled the sales team to channel their efforts into the most promising prospects.

7. Conclusion: The project successfully built a predictive model that helps X Education improve its lead conversion rate by focusing on high-potential leads, leading to higher conversions and maximum ROI. By identifying actionable insights, such as prioritizing outreach to engaged leads and avoiding low-potential ones, the model empowers the sales team to work more efficiently, saving time and resources while increasing revenue. The learnings from this assignment highlight the value of cleaning data thoroughly, engineering meaningful features, and optimizing model performance through iterative evaluation.