Summary Report

Introduction

X Education, an online course provider for industry professionals, faces challenges in converting leads into paying customers. The company's current conversion rate is approximately 30%, and they aim to improve it to around 80% by identifying and prioritizing high-potential leads. This project focuses on building a logistic regression model to assign lead scores, allowing the sales team to concentrate on the most promising prospects.

Process Followed

The analysis followed a structured approach to ensure accuracy and efficiency in predicting lead conversions:

1. **Data Exploration and Preprocessing:**
    - Analysed ~9000 leads with categorical and numerical attributes.
    - Identified and handled missing values, ensuring data consistency.
    - Treated 'Select' values as missing data and removed redundant variables.
    - Standardized numerical features and one-hot encoded categorical variables for compatibility with the model.
2. **Feature Selection:**
    - Conducted correlation analysis to identify relationships between variables.
    - Applied Recursive Feature Elimination (RFE) to retain the most significant predictors.
    - Ensured interpretability while balancing model performance.
3. **Model Development:**
    - Built a logistic regression model to predict lead conversion likelihood.
    - Split data into training (70%) and testing (30%) sets for validation.
    - Evaluated model performance using accuracy, precision, recall, F1-score, and AUC-ROC.
    - Tuned model parameters to optimize predictive power.
4. **Model Interpretation & Business Insights:**
    - Identified key factors affecting lead conversion, such as time spent on the website, lead source, and last activity.
    - Assigned lead scores between 0 and 100, prioritizing high-potential leads.
    - Recommended tailored engagement strategies based on lead scores.
5. **Implementation Recommendations:**
    - Integrated lead scoring into the sales team's workflow to optimize resource allocation.

- Recommended a targeted follow-up approach, focusing on high-score leads.
- Suggested periodic model updates to incorporate new data trends.

Learnings and Insights

The project provided several key insights:

- **Data Quality Matters:** Handling missing values and removing redundant features significantly improved model accuracy.
- **Engagement Drives Conversion:** Leads who spent more time on the website and interacted frequently with marketing content had a higher likelihood of conversion.
- **Lead Source is Critical:** Organic search, referrals, and direct traffic showed a stronger correlation with conversion rates.
- **Efficient Resource Allocation:** The lead scoring system enables the sales team to prioritize leads with higher conversion potential, improving efficiency and reducing wasted effort.
- **Continuous Model Improvement:** Regular retraining of the model with new data will enhance prediction accuracy and adapt to changing customer behaviours.

Conclusion

By implementing a data-driven approach, X Education can significantly improve its lead conversion strategy. The logistic regression model effectively differentiates hot leads from cold ones, enabling the sales team to focus on high-potential customers. Future work should explore more advanced machine learning models, such as decision trees or ensemble methods, and incorporate real-time scoring for enhanced decision-making. This approach ensures sustainable improvements in conversion rates and overall business performance.