SUMMARY REPORT: BUILDING A LEAD SCORING MODEL FOR X EDUCATION

Introduction:

This report summarizes the process and findings of building a lead scoring model for X Education using logistic regression to identify "hot leads" for their online courses. The goal was to improve lead conversion rates from the current 30% to a target of 80%.

Data and Methods:

The project analyzed a dataset of past leads with various attributes like lead source, website activity, and conversion status. Data preprocessing involved handling missing values, transforming categorical variables, and addressing outlier issues. Data preprocessing involved:

- **Cleaning:** Renaming columns, dropping unnecessary features, handling missing values, and addressing issues with specific categorical columns.
- **Transformations:** Converting binary variables to 0/1, creating dummy variables for categorical features, and handling outliers.
- **Preparation:** Splitting data into training and testing sets, scaling features, and analyzing correlations.

Model Building and Evaluation:

Logistic regression models were built, using all available features. Conversion rate is 38.5%

Key Findings and Results:

- **Feature Importance:** The model identified several key features contributing to lead conversion, including:
  - **Positive impact:** Lead source ("Reference", "Other Social Sites"), Not opting out of email ("do_not_email")
  - **Potential impact:** Total time spent on website, Page views per visit, Specialization, Source of information about X Education
- **Subjective Questions:**
  - Top 3 variables for conversion probability:
    - lead_source_Reference
    - lead_source_Other Social Sites
    - do_not_email
  - Top 3 categorical variables for conversion focus:
    - lead_source_Reference (positive contribution)
    - lead_source_Other Social Sites (positive contribution)
    - lead_source_Olark Chat (positive contribution)
  - Strategy for internship period: Prioritize leads by predicted conversion probability and segment by "lead_source". Utilize interns for initial

outreach, personalize communication, track performance, and leverage multi-channel outreach with limited-time incentives. Respect "do_not-email" preferences.

Learnings and Reflections:

- **Challenges:** Challenges encountered included handling data quality issues like missing values and categorical variable levels, and selecting the most relevant features for the model.
- **Limitations:** Potential limitations include model assumptions, data representativeness, and generalizability to future lead cohorts.
- **Future Improvements:** Further exploration could involve optimizing feature engineering, trying different model architectures, and incorporating real-time lead behavior data.

Recommendations:

- **Feature engineering:** Refining feature selection and creating domain-specific features could improve model performance.
- **Actionable strategies:** Utilizing insights from the model (top features, lead segmentation) to prioritize outreach efforts and personalize communication can enhance lead nurturing and conversion rates.

Conclusion:

This project successfully built a lead scoring model that identifies promising leads with potential to improve conversion rates. Continued refinement and integration with real-time data can further enhance X Education's lead nurturing and conversion strategies.