Yuleica Ledezma
UCB Data Analytics Bootcamp April-Oct 2024

Project Title:

Predictive Lead Conversion for Optimizing Online Training Sales and Enhancing Conversion Rates

Executive Summary:

This project's main purpose is to develop a logistic regression model for predicting lead conversion to improve current lead conversion of 30% to 75%. By assigning a lead score to each lead, the sales team can focus their efforts on high-potential prospects. The goal is to improve lead conversion rates by better identifying which leads are more likely to convert into paying customers.This solution aims to optimize the sales pipeline, improving both efficiency and revenue.

Problem Statement:

Online training (X Education) markets courses to potential customers through various online channels. Despite capturing leads via forms and website interactions, only 30% of these leads convert into paying customers. The goal is to develop a model that identifies high-conversion leads by assigning a predictive score to each lead, optimizing the sales team's efforts.

Objectives:

- Develop a predictive lead scoring model.
- Increase the lead conversion rate from 30% to 75%.
- Provide actionable insights for improving sales targeting and outreach efficiency.
- Ensure the model is easy to integrate with existing sales workflows.

Data Overview:

The dataset used includes lead data sourced from CSV file (9,241 rows x 37 columns) from Kaggel, capturing features like lead origin, engagement metrics, and historical conversion outcomes. Data types are mostly categorical and numerical, including variables such as time spent on the website and response to marketing campaigns. The dataset will be cleaned and preprocessed to ensure accuracy before modeling.

Methodology:

The project involves:

- **Data Cleaning and Preprocessing**: Handling missing values, outliers, and categorical data encoding.
- **Exploratory Data Analysis (EDA)**: distributions, and correlations between features.
- **Feature Engineering**: Creating new variables based on user engagement levels.

- **Model Building**: Logistic Regression will be employed to predict lead scores. Hyperparameter tuning will optimize the model's performance.
- **Model Evaluation**: Performance metrics such as precision, recall, and F1 score will be used.

Tools and Technologies:

- **Languages**: Python
- **Libraries**: Pandas, Scikit-learn, Matplotlib
- **Tools**: Jupyter Notebook, GitHub, Kaggle (dataset)
- **Machine Learning**: Logistic regression model development and tuning

Project Timeline:

- **09/22 - 09/28 Week 1**: Data collection, cleaning, and EDA
- **09/29 - 10/05 Week 2**: Model building and testing
- **10/06 - 10/12 Week 3**: Model evaluation and optimization
- **10/13 - 10/17 Week 4**: Final report and presentation

Key Deliverables:

- Cleaned and preprocessed dataset
- Logistic regression model with performance metrics
- A comprehensive report summarizing findings and recommendations
- Visualizations of key insights and model performance
- A presentation of the project for stakeholders

Success Criteria:

- Model accuracy above 75%
- Reduction in time spent on low-potential leads
- Clear increase in lead conversion rates after implementation

Risk Management:

- **Data Quality Issues**: Data validation and cleaning steps.
- **Overfitting**: Cross-validation and hyperparameter tuning.

Budget and Resources:

- **Software**: Free Python libraries and Kaggle.
- **Human Resources**: Data analyst and collaboration with the sales team for feedback and integration (in theory)

Stakeholders:

- **Data Analyst**: Responsible for model development and testing

- **Sales and Marketing Team**: End users of the lead scoring system
- **Upper Management**: Overseeing the project's alignment with business goals

Ethics and Compliance:

- Ensuring data privacy and compliance with relevant regulations (e.g., GDPR).

Conclusion:

This project will significantly impact the efficiency of X Education's sales efforts by improving lead targeting. The predictive lead scoring model is expected to enhance conversion rates, leading to increased revenue and a more streamlined sales process.

Project Schema (subject to change)

1. **Data Preprocessing**
   - 1.1. **Data Loading -** Load the dataset into an environment for analysis .
   - 1.2. **Data Inspection -** Handle missing values, column types, and summary statistics.
   - 1.3. **Exploratory Data Analysis (EDA) -**
     - 1.3.1. **Data Cleaning:** Remove duplicates, handle outliers, fix structural errors, and deal with missing data.
2. **Feature Engineering and Transformation**
   - 2.1. **Discretization:** Convert continuous variables into discrete ones if necessary
   - 2.2. **2.2 Categorical Encoding:** Use one-hot encoding for categorical features.
   - 2.3. **2.3 Variable Transformation:** Apply log or square-root transformations for skewed data.
   - 2.4. **Scaling:** Normalize or standardize numerical features.
   - 2.5. **Dimensionality Reduction:** Use Principal Component Analysis (PCA) to reduce feature dimensions.
3. **Model Implementation**
   - 3.1. **Supervised Learning:** Logistic Regression, Decision Trees, Random Forest, Neural Networks**.**
   - 3.2. **Model Evaluation:** Calculate baseline metric (accuracy, precision, recall) and performance on unseen data.
4. **Model Optimization**
   - 4.1. **Hyperparameter Tuning:**
   - 4.2. **Regularization:**
   - 4.3. **Visualizations:** Learning curves, feature importance
5. **Presentation**
   - 5.1. **Clear Objectives:** Communicate the problem statement and solution.

5.2. **Model Performance:** Use visualizations (learning curves, comparison of models).
5.3. **Interactivity (optional):** If meaningful, incorporate interactive visualizations.