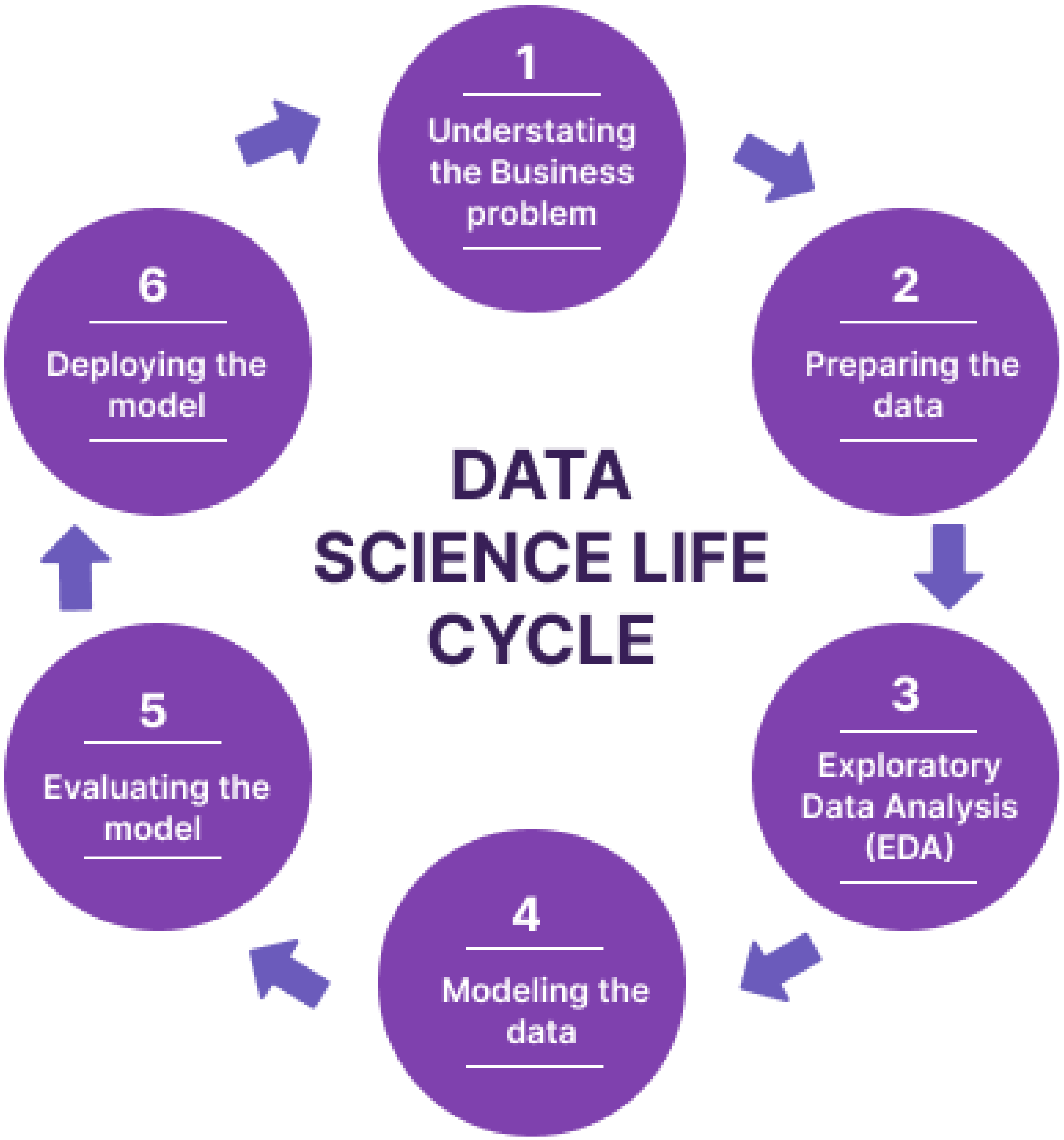
Enhancing Home Loan Approvals with Machine Learning

Leveraging Data Science for Smarter Decision Making

Data Science Life Cycle



- **Definition**: The data science life cycle is a systematic approach to solving complex problems using data-driven methodologies.
- **Application in the project:**
  - **Data Acquisition**: Gathering historical data on home loan applications from internal databases.
  - **Data Preparation**: Cleaning, preprocessing, and formatting the data to ensure consistency and quality.
  - **Exploratory Data Analysis (EDA)**: Gaining insights into the dataset, identifying patterns, and understanding relationships between variables.
  - **Model Building:** Experimenting with various machine learning algorithms to build predictive models for home loan approvals.
  - **Model Evaluation:** Assessing model performance using metrics such as accuracy, precision, recall, and ROC-AUC.
  - **Deployment**: Deploying successful models into the loan approval system to automate decision-making and improve efficiency.

Project Overview

Business Problem:

- **Description**: Challenges in the home loan approval process leading to delays and suboptimal decisions.
- **Impact**: Delays can result in customer dissatisfaction and lost opportunities for the company.

Business Objective:

- **Objective**: Enhance the home loan approval process to improve customer satisfaction, reduce default rates, and increase business efficiency.
- **Hypothesis**: "With machine learning, we can optimize our home loan approval system to ensure timely and accurate decisions, leading to improved customer outcomes and business performance."

Scope:

- **Focus:** Leveraging machine learning techniques to automate and optimize the loan approval process
- **Limitations:**Human judgment and domain expertise remain critical in certain aspects of the process.

Data Overview

Description:

- Dataset: Historical data on home loan applications.
- Size: 981records and 13 features.
- Data Types: Mix of categorical (e.g., gender, marital status) and numerical (e.g., applicant income, loan amount) variables.
- Target/Loan Status – Y (422) vs N (192)

Missing values in categorical variables (Gender, Married, Dependents, Self_Employed) were filled with the mode, while 'Unknown' was introduced for additional missing values. Numerical variables (LoanAmount, Loan_Amount_Term, Credit_History) were imputed with their respective mean values. Due to the target variable's significance, rows with missing values in the Loan_Status column were dropped to maintain data integrity.

Modeling

1. A bespoke machine learning model underwent preprocessing steps before training, ensuring the data was properly formatted and cleaned to enhance model performance.
2. In contrast, AutoML was employed directly without the need for preprocessing, leveraging its automated capabilities to handle data preprocessing tasks seamlessly.
3. Despite the distinct preprocessing requirements, the outcomes yielded by both approaches exhibited striking similarity, highlighting the effectiveness of both bespoke and AutoML methodologies in addressing the home loan approval task.

Recommendation

1. While the accuracy of the Logistic Regression model is 75.61%, the TPOT AutoML model achieved a slightly higher accuracy of 77.24%.
2. The bespoke ML model, although slightly outperformed by AutoML, offers several advantages:
   - Transparency: We have a clear understanding of the preprocessing steps and algorithms used, providing greater interpretability.
   - Time Efficiency: The bespoke model requires less time for training, which could be advantageous for real-time prediction scenarios.
3. AutoML serves well as a baseline model:
   - It provides a benchmark against which bespoke models can be compared.
   - Its automated approach streamlines the model selection process, saving time and effort in model development.