Chapter 3: Methodology

Following the data science project life cycle, this chapter describes the research technique used to create and validate BERI's automated recruitment system. It contains a thorough explanation of feature engineering, data pre-processing, data sources, and data gathering techniques.

**3.1 Data Science Project Life Cycle**

The data science project life cycle involves several key stages:

* **Problem Definition:** articulating the project's goals and problem in clear terms.
* **Data Collection:** collecting pertinent information from several sources.
* **Pre-processing of the data:** data need to be cleaned and cleared
* **Exploratory Data Analysis (EDA):** It is a way of finding the required data through information and trends.
* **Modeling:** Modeling is a way of developing machine learning models to solve specific problems.
* **Evaluation:** utilizing the proper measures to evaluate the models' performance.
* **Deployment:** The implementation of a model in an actual environment is called deployment.
* **Monitoring and Maintenance:** It means keeping an eye on the model's performance and implementing any required modifications.

**3.2 Data Sources and Data Collection Methods**

**3.2.1 Data Sources**

Social networking sites and BERI's CRM system will be the main sources of data for this project. The information will consist of:

* **Social Media Data:** Details gathered from user profiles, posts, and interactions on social media sites like Facebook, LinkedIn, and Twitter.
* **CRM Data:** Past recruiting information, such as contact information, interaction logs, and student profiles.

**3.2.2 Data Collection Methods**

* **Social Media APIs:** Gathering pertinent data by using APIs offered by social media networks. In order to collect user profiles and activity data, tools like the Facebook Graph API and LinkedIn API will be used.
* **Web scraping:** Using web scraping methods to obtain data from social networking sites in situations when access to APIs is restricted or nonexistent.
* **CRM Integration:** Using database queries and API calls, BERI's CRM system is used to extract pertinent data.

**3.3 Data Pre-processing**

**3.3.1 Data Cleaning**

* **Taking Care of Missing Values:** Locating, adding, or eliminating missing values from the dataset.
* **Removing Duplicates:** To maintain data integrity, remove duplicate entries.
* **Data Consistency:** Maintaining uniformity in data formats, including date formats and category variables, is known as data consistency.

**3.3.2 Data Transformation**

* **Normalization and Scaling:** Standardizing numerical properties through the application of normalization or scaling procedures.
* **Coding Categorical Variables:** Utilizing methods such as label encoding or one-hot encoding to translate categorical variables into numerical representation.

**3.3.3 Feature Engineering**

* **Developing New Features:** To enhance model performance, new features can be created using the data that already exists. Developing engagement measurements, for instance, from social media exchanges.
* **Feature Selection:** To lower dimensionality and boost efficiency, find and pick the most pertinent characteristics for the model.

**3.4 Analysis with Exploratory Data (EDA)**

**3.4.1 Descriptive Statistics**

To comprehend the distribution and central patterns of the data, summary statistics for numerical and categorical variables are calculated.

**3.4.2 Data Visualization**

Data visualization methods like scatter plots, chart formats, and histograms can be used to find relationships and trends in the data.

**3.4.3 Correlation Analysis**

Examining feature correlations in order to recognize any multicollinearity problems and comprehend feature interactions.

**3.5 Modeling**

**3.5.1 Model Selection**

Choosing suitable machine learning algorithms according to the properties of the data and the description of the problem. Neural networks, logistic regression, random forests, and decision trees are examples of potential models.

**3.5.2 Model Training**

Using the pre-processed data, train the chosen models. To ensure robustness, this entails dividing the data into training and testing sets and applying cross-validation procedures.

**3.5.3 Model Evaluation**

Using suitable measures for model evaluation, such as F1-score, recall, accuracy, and precision. This aids in the deployment process of choosing the model that performs the best.

**3.6 Deployment**

**3.6.1 Integration with CRM System**

Putting the chosen strategy into practice within BERI's CRM system to allow for automatic data gathering and individualized outreach to potential students.

**3.6.2 User Interface Development**

Creating an intuitive user interface so that the BERI hiring team can communicate with the model and obtain insights.

**3.7 Monitoring and Maintenance**

**3.7.1 Performance Monitoring**

Keeping an ongoing eye on the model's performance to make sure it's successful in locating and attracting potential students.

**3.7.2 Regular Updates**

Upgrading the model and data sources on a regular basis to adjust to evolving trends and enhance accuracy over time.