Summary Report: Lead Score Analysis

Approach to the Assignment:

The primary objective of this assignment was to analyse a dataset of leads, clean and prepare the data, build predictive models, and provide actionable insights. The process was divided into systematic steps to ensure clarity and efficiency.

1. Data Loading and Exploration:

The dataset was loaded into a pandas DataFrame, and its structure was examined using `.info()` and `.shape()`. Duplicate rows were identified and counted to ensure data integrity.

2. Missing Value Analysis:

Missing values were analysed by calculating their percentage for each column. Columns with more than 25% missing values were dropped, as they were deemed insufficiently informative for modelling. For remaining missing values, numerical columns were imputed with the median, while categorical columns were cleaned by replacing placeholder values like "Select" with `NaN`.

3. Handling High-Cardinality Columns:

Categorical columns with a high number of unique values were identified. Irrelevant columns, such as unique identifiers, were dropped to reduce noise. One-hot encoding was applied to low-cardinality categorical columns to prepare the data for machine learning models.

4. Feature Scaling:

Numerical features were standardized using `StandardScaler` to normalize their scales, ensuring better performance for models sensitive to feature magnitudes.

5. Model Selection and Hyperparameter Tuning:

Three models were chosen for evaluation: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Each model was paired with a set of hyperparameters for tuning using `GridSearchCV` . This ensured that the models were optimized for accuracy.

6. Model Training and Evaluation:

The dataset was split into training and testing sets. Each model was trained using the training set, and predictions were made on the test set. Metrics such as accuracy, precision, and recall were calculated to evaluate performance. Logistic Regression and SVM were also used to generate lead scores, providing probabilities of conversion.

7. Feature Importance Analysis:

For the Random Forest model, feature importance was analysed to identify the most influential factors in predicting lead conversion. This provided actionable insights into which features were critical for decision-making.

8. Visualization:

A lead conversion funnel was visualized using a bar plot to highlight the number of initial leads versus converted leads, providing a clear representation of the conversion process.

Lessons Learned:

1. Importance of Data Cleaning:

Handling missing values, duplicates, and irrelevant columns is crucial for building robust models. Poor data quality can significantly impact model performance.

2. Feature Engineering Matters:

Proper handling of categorical variables and scaling of numerical features can improve model accuracy and interpretability.

3. Model Selection and Tuning:

Different models have varying strengths. Logistic Regression provided interpretability, Random Forest highlighted feature importance, and SVM offered robust classification. Hyperparameter tuning was essential for optimizing each model.

4. Error Handling:

Incorporating error handling during model training (e.g., `error score’='raise'` in `GridSearchCV`) ensured that the process was resilient to unexpected issues.

5. Actionable Insights:

Beyond building models, analysing feature importance and visualizing results provided valuable insights for stakeholders, making the analysis more impactful.

This assignment reinforced the importance of a structured approach to data science projects, from data preparation to model evaluation and insight generation.