Week9:

Team member's details:

Banking Details, Vasu Sharma,

vasu.vs45@gmail.com,

United States of America,

Indiana University Bloomington,

Data Science

Problem description: ABC bank wants to sell its term deposit term. Before launching the product, the bank wants to develop a model so that they can understand which customer will buy this product.

An ML model so that they can use their resources only for the customers who will buy their product which will save them resources and time.

Some of the techniques that have been used to clean the data are as follows:

Handling Missing Values:

In the data preprocessing phase, one of the critical steps is addressing missing values. This is crucial as missing data can lead to biased or inaccurate model predictions. The approach taken for handling missing values depends on the nature of the data:

Logistic/Textual Data: Rows with missing values in columns containing categorical or textual data are removed. This is done because imputing missing values for these types of data can be challenging without introducing bias.

Linear/Mathematical Data: For columns with numerical or mathematical data, missing values are imputed using statistical measures like mean or median. This helps in maintaining the integrity of the data while filling in the gaps.

Addressing Class Imbalance with Oversampling:

In many real-world datasets, especially in binary classification problems, there can be a significant imbalance between the two classes (majority and minority). Class imbalance can lead to models that are biased towards the majority class and perform poorly on the minority class. To mitigate this issue, oversampling is employed as a powerful technique:

Understanding Class Imbalance: Before employing oversampling, it's essential to recognize the class imbalance problem. The majority class often dominates the dataset, making it challenging for the model to learn from the minority class examples.

Oversampling Technique: Oversampling is used to create more instances of the minority class, making it comparable in size to the majority class. This involves duplicating or generating synthetic examples from the existing minority class data. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling) are applied.

Balancing the Dataset: By increasing the representation of the minority class, the dataset becomes more balanced. This allows the machine learning model to learn from both classes more effectively, reducing the bias towards the majority class.

Preventing Overfitting: While oversampling can be highly beneficial, it must be used with caution to avoid overfitting. Cross-validation is often used to assess model performance and to ensure that the oversampling technique doesn't introduce noise or bias into the data.

Improving Model Performance: The ultimate goal of oversampling is to enhance model performance, especially in terms of correctly classifying the minority class. This can result in a more robust and accurate predictive model.