Later move to Latex

Question3 - Bank (**Speed up bank loaning process**)

This is our target variable

(i) Exploratory Data Analysis and Data Pre-Processing

1. Get Rid of Duplicate Data-Points:

Identical data-points can repeat many times over if the training data is huge in size. Therefore, to prevent bias during modeling, it is important to remove duplicate data-points.

2. Handling Highly Correlated Features:

Clustering and correlation plots can help find out if two features are strongly correlated or offer the same information.

As a general rule, if the correlation between the two features is higher than 99%, you can safely remove one of them.

The threshold (for correlation) percentage can be decided on the basis of the problem at hand.

3. Handling Low-Variance Features:

You can remove a feature if its variance is too low.

Such a feature remains constant in a dataset and cannot explain or influence the variation in the target variable.

4. Handling Imbalanced Data:

In the case of imbalanced data sets, you can

Oversample the class with lesser data points (you can use SMOTE or create duplicate data points)

Undersample the class with more data points (you can remove a few similar data points)

5. Handling Missing Values:

There are different ways to handle missing values in a data set after you are done importing the libraries and the data set.

High Percentage of Missing Values: You can drop a feature having more than 40–50% missing values.

Low Percentage of Missing Values: If the missing values for a feature are very low, you can drop the rows that contain missing values.

Imputation: The data is rarely complete; data can be missing due to numerous reasons: not captured, captured but not available, etc. In this scenario, you can continue with analysis after estimating the missing value. The process is called imputation. You can impute the missing values with the mean or median for a numerical feature and mode for a categorical feature.

6. Encoding Categorical Features:

At times, some data is in qualitative (text) form. In this case, you will find categories in text form in the data.

Such categorical features need to be converted into numerical data as most data models are based on mathematical equations and calculations and take numerical data as input.

You can use one-hot (variable binary representation) or label encoding if there aren’t too many categorical features. Otherwise, you may need to use supervised ration.

7. Feature Scaling:

Scaling is a method deployed to standardize the range of features or independent variables.

Various features in a data set will vary in their scale.

Since some features may dominate the rest, it is recommended to have all of them on the same scale.

8. Dimensionality Reduction:

This preprocessing step is important when you’re dealing with big data sets having hundreds or thousands of features.

You can use the Principal Component Analysis (PCA) technique here.

In this technique, the linear combination of a set of original features is transformed into a new set of features by reducing the size of feature space while retaining maximum information possible.

9. Train and Test Sets:

Check if the distribution of train and test sets is the same. Otherwise, the analysis will make no sense.

As a general rule, 20% of the data set is allocated to the test set and the remaining 80% is allocated to the training set. You will train a machine learning set on the training set and test it on the training set to check how well it can predict.

Shuffle the data set so that your model learns about the various data points in a single iteration.

(ii) Feature Selection

(iii) Appropriate Machine Learning Techniques Used

(iv) Results and Discussion

(v) Deployment (Visualization)