Bureau Assignment

Avinash Yadav

September 2024

1 Introduction

The goal of this assignment is to predict the loan status, whether the application will be approved or not.

2 Preliminary - Analysis

The test data has following columns :-

Column	Non-Null Count	Dtype
DEALER ID	10000	int64
APPLICATION LOGIN DATE	10000	object
HDB BRANCH NAME	9999	object
HDB BRANCH STATE	9146	object
FIRST NAME	10000	object
MIDDLE NAME	2855	object
LAST NAME	9319	object
mobile	10000	int64
AADHAR VERIFIED	10000	object
Cibil Score	5703	object
MOBILE VERIFICATION	10000	bool
DEALER NAME	9996	object
TOTAL ASSET COST	4892	float64
ASSET CTG	4892	object
ASSET MODEL NO	10000	int64
APPLIED AMOUNT	10000	int64
PRIMARY ASSET MAKE	10000	object
Primary Asset Model No	10000	object
Personal Email Address	10000	object
MARITAL STATUS	5106	object
GENDER	10000	object
DOB	10000	int64

Continued on next page

Table 1 – *Continued from previous page*

Column	Non-Null Count	Dtype
AGE	10000	int64
ADDRESS TYPE	6688	object
EMPLOY CONSTITUTION	5002	object
EMPLOYER NAME	4990	object
EMPLOYER TYPE	5002	object
Pan Name	8947	object
name	10000	object
vpa	7213	object
upi$_{n}ame$	7211	object
Phone Social Premium.a23games	1	float64
Phone Social Premium.amazon	8084	float64
Phone Social Premium.byjus	8052	float64
Phone Social Premium.flipkart	8168	float64
Phone Social Premium.housing	8224	float64
Phone Social Premium.indiamart	8225	float64
Phone Social Premium.instagram	3370	float64
Phone Social Premium.isWABusiness	1573	float64
Phone Social Premium.jeevansaathi	8171	float64
Phone Social Premium.jiomart	410	float64
Phone Social Premium.microsoft	8128	float64
Phone Social Premium.my11	2	float64
Phone Social Premium.paytm	8243	float64
Phone Social Premium.rummycircle	1	float64
Phone Social Premium.shaadi	8221	float64
Phone Social Premium.skype	8215	float64
Phone Social Premium.toi	8057	float64
Phone Social Premium.whatsapp	1573	float64
Phone Social Premium.yatra	9	float64
Phone Social Premium.zoho	8218	float64
phone$_{d}igitalage$	9996	float64
phone$_{n}ameMatchScore$	9996	float64
phone$_{p}honeFootprintStrengthOverall$	9994	object
Application Status	10000	object

Mutual Information and Feature Selection

Mutual Information (MI) is a concept from information theory that measures the amount of information one random variable contains about another. In the context of machine learning and feature selection, it quantifies how much knowing the value of one variable (e.g., a feature) reduces uncertainty about another variable (e.g., the target). Unlike correlation, mutual information can capture both linear and non-linear relationships between variables.

How Mutual Information Works

Mathematically, mutual information between two variables X (a feature) and Y (the target) is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Where:

- $p(x,y)$ is the joint probability distribution of X and Y,
- $p(x)$ and $p(y)$ are the marginal probabilities of X and Y, respectively.

If X and Y are independent, the mutual information is zero, as knowing X does not provide any information about Y. On the other hand, higher values of mutual information indicate a stronger relationship between X and Y.

Mutual Information for Feature Selection

In feature selection, the goal is to select features that are most relevant to predicting the target variable. Mutual information helps by identifying which features contribute the most information about the target. Here's how it assists in this process:

1. **Handling Non-Linear Relationships:** Traditional metrics like correlation only measure linear relationships between features and the target. Mutual information can detect both linear and non-linear dependencies, making it more versatile for feature selection.
2. **Measuring Redundancy:** Mutual information can help detect redundant features. If two features provide similar information about the target, mutual information can be used to measure the redundancy and select one over the other.
3. **Capturing Dependencies:** Mutual information is effective when dealing with categorical or ordinal variables, as it can capture dependencies beyond simple numeric relationships.
4. **Filter Method for Feature Selection:** Mutual information can be used as a filter method, where features are ranked based on their mutual information with the target variable. Features with low mutual information scores contribute little to predicting the target and can be removed from the dataset.

Advantages of Mutual Information in Feature Selection

- **Non-parametric:** It does not assume any underlying distribution of the data.

- **Flexible:** Can be applied to both categorical and continuous variables.
- **Effective with Complex Data:** Captures non-linear relationships, which traditional techniques may miss.

3 Approach

Our goal is to predict application status. A simple mutual information analysis shows that some of the columns are much more relevant than other columns. Unfortunately these are the columns if the most null values. I followed the following steps.

1) Do a basic cleaning, to drop unecessary columns, clear whitespaces and bring uniformity by converting everything to lower case.

2) The columns which more than 70 percent null values are dropped. The columns in which null values less than 10 percent, null values are replaced by the mode (Only categorical)

3) Preliminary investigation suggests that we need cibil score and employer type. There are many null values in both the columns. So I trained gradient boosting (XGB) predictors for both.
Now similar investigation shows employer type to be dependent on address type and marital status. Now there are null values in these columns also, I trained XGB predictors for them also.

I reduced addresses to 3 types only. personal, rented and parents/spouse owned.
For numerical columns I observed that mean and median are nearby, so I replaced all null values with median.
In the end I predicted all null values of cibil score, address type, marital status, and employer type using different classifiers I trained.

For final prediction I used a logistic regression model.