**Capstone Project 1: In-Dept Analysis (Machine Learning)**

**Problem Statement**

How can automating the loan eligibility process identify highly qualified customers? How can automation help with marketing by targeting potential customers? How can eligibility automation reduce risk and increase profitability?

**Data Description**

The original data contains 614 instances/observations and 13 features/variables (LoanID, Gender, Marital Status, Self Employed, Education, Number of Dependents, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount Term, Property Area, Credit History, and Loan Status)

**Hypothesis**

Factors that may affect the approval of a loan are the factors worth considering when determining the hypothesis. Here are a few,

Income: The total income (combined applicant and coapplicant) is a determining factor. The higher the total income the higher the chances of being approved.

Credit History: Having a history of previous loans help to improve approval possibility because it indicates loan worthiness.

Monthly payments: The lower to monthly payment to income the greater the chances of approval.

Term: The shorter the term, the chances of approval increases however this is also dependent previously mentioned factors.

Loan amount: The lower the loan amount the greater the chance of approval.

**Solution**

This project is a classification problem in which the two prediction classes are yes or no, binary classification. The data calls for a supervised machine learning technique; Random Forest, Naives Bayes, Logistic Regression, and Support Vector Machine (SVM). The models will be evaluated using precision, recall, and accuracy scores.

Steps taken after the pre-processing of the data.

1. Identify the independent variables and the target variable (LoanStatus)
2. Import test\_train\_split module, to split data into training and test set
3. Import evaluation matrix, to determine the performance of the model
4. Import classifiers
5. Instantiate classifier
6. Fit the training data to the classifier.
7. Predict the test data using the trained classifier
8. Determine the model performance

**Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.791  Recall : 1.0  Accuracy : 0.834 | Precision: 0.833  Recall : 0.979  Accuracy : 0.855 | Precision: 0.784  Recall : 1.0  Accuracy : 0.828 | Precision: 0.784  Recall : 1.0  Accuracy : 0.828 |

After changing how missing data are treated.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.845  Recall : 0.973  Accuracy : 0.851 | Precision: 0.862  Recall : 0.946  Accuracy : 0.851 | Precision: 0.852  Recall : 0.973  Accuracy : 0.857 | Precision: 0.84  Recall : 0.982  Accuracy : 0.851 |

Feature Engineering: Taking the log of the TotalIncome and MonthlyPaymentNoInterest

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.845  Recall : 0.973  Accuracy : 0.851 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.852  Recall : 0.973  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 |

C optimization and feature selection (based on coef) of Linear Regression model and GridSearchCV for RandomForest model.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | Random Forest | Naïve Bayes | SVM |
| Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 | Precision: 0.846  Recall : 0.982  Accuracy : 0.857 |

Based on the manipulations done, Naïve Bayes seems to be the model that outperform the others, allowing for more generalization.