Loan Prediction Using Machine Learning Models

**Summary:**

We’ve created a simple python script which runs multiple machine learning models including SVM and Logistic Regression, among others, which achieves accuracy of around 80% on average

**Preprocessing:**

The script starts by removing the ID field from the attached .CSV file, as it won’t be used to identify data and it’s a unique value for each row, so it wouldn’t affect the model’s accuracy at all.

Next, all null attributes are dropped from the Dataset and all values are factorized into numeric values using pandas.factorize() which is performed on every value in the dataset.

**Models applied:**

To acquire a correct estimation of the accuracy of the models, the fitting and predicting processes are repeated a number of times, which is set by a variable inside the code, our reported accuracies were recorded after 15 predictions on randomized test samples, using a different random\_state in each run, but keeping the random states the same between all models in each run.

This means that a large number of training and testing datasets are created and stored.

Text

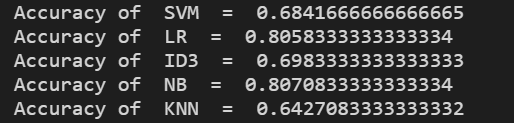
Description automatically generatedHere is the code for this process:

We’ve used a polynomial SVM with 12 degrees as it produced the most accuracy before the processing time was too large for my machine to handle, it’s also that the dataset after removal of the ID column has 12 columns which helped settle on this degree.

We’ve set a large maximum of iterations on the Logistic Regression model to avoid an error which results after. it’s run on too many rows, however we found more accuracy keeping the iterations to run on the entire training sub-dataset.

The dataset is split into 80-20 portions for training and testing, those datasets are stored in the datas[] list.

**Results:**



We’ve found that Logistic Regression produces the best results at around 80.5% on the testing datasets

**Post Scripta (Bonuses Section):**

**[feature selection-data scaling- extra models]**

We added two new models, Gaussian Naïve Bayes and K-Nearest-Neighbors, it’s found that GNB produces the best results on average with a slight increase over the accuracy of the Logistic Regression model.

Next we plotted feature selection that was performed using Recursive Feature Elimination (RFE), as the results of this method varies by models used, we applied each model on three different datasets each had it’s feature selection done by a different models.

Models used for Feature Selection were: Linear SVM, Logistic Regression and decision tree identifier.

Unfortunately, It’s rather hard to run this volume of models on this volume of datasets, so it’s done only once, this means that, unlike the above results, these are not averages, and may be skewed.

Text

Description automatically generated

These are the accuracies for Linear SVM, Logistic Regression and Decision Tree Identifier respectively.

All using only 9 features from the original dataset.

there is a slight increase in accuracy but without running the tests more times it’ll be hard to assess whether the increase is probable or not.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Note that the importance is in descending order, so the most important features are given a value of ‘1’

In these particular plots, the most important ‘4’ features are to be left by the end of the RFE process.