Reference dataset’s link:

<https://www.kaggle.com/datasets/laotse/credit-risk-dataset>

Variables:

Numerical Variables:

|  |  |
| --- | --- |
| Id | A primary key is the column or columns that contain values that uniquely identify each row in a table. |
| Age | The borrower’s age |
| Annual Income | The amount of money you receive during the year into your bank account before any deductions. |
| Employment Length | Year of Employment means a period of service of 12 months |
| Loan Amount | An amount of money loaned at interest by a bank to a borrower |
| Interest Rate | The amount a lender charges a borrower |
| Percent Income | Loan Amount / Annual Income |
| Credit History Length | How long have you held a credit facility (credit card, or a loan) |

Categorical Variables:

|  |  |
| --- | --- |
| Home Ownership | Owning a house |
| Loan Intent | The documentation addressed to the Applicant/Developer of an interest or intent to provide funding, setting forth the writer’s intention to negotiate the financing and stating the amount, interest rate, security, repayment terms and including the minimum debt service coverage ratio and loan-to-value ratio used by the lender to size the financing, as applicable |
| Loan Grade | Loan grading is a classification system that involves assigning a quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest |
| Loan Status (0 is non default 1 is default) | Indicates where your loan is in the process. |
| Historical Default | A default happens when a borrower fails to make required payments on a debt, whether of interest or principal. Historical Default means the previous stage of default. |

Main Objective:

How do we know whether we can get a bank loan or not? In the banking system, there will be a scorecard point to analyse whether this person is available and qualified to receive a loan. The probability of default will be considered in this scorecard point. The lower the probability of default, the better the chances of getting the loan. As an example, those with a probability of less than 0.05% will be classified as A grade and will be the most welcome in the banking system to obtain a loan. So, how can we calculate the probability of default? We will create a model and count the accuracy based on the binning and weightage that we used. In the Python file, I cleaned the data and ran the full model, Weighted Logistic Regression, RFE model, and RFECV model to discover which model provided the most accuracy and performance.

Data Cleaning, EDA:

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Reference Link: <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

Model Used:

Logistic Regression.

Logistic Regression is the best regression model to use when the target variable is binary. It is used to describe and explain the relationship between one dependent binary variable and one or more independent variables. Since our primary goal is to see the output: Loan Status, where 0 indicates that the borrower is not in default, while 1 indicates that the borrower is in default. Logistic Regression consider as the best model to used.

Class Imbalance Model Used:

Weighted Logistic Regression.

Reference Link: <https://towardsdatascience.com/weighted-logistic-regression-for-imbalanced-dataset-9a5cd88e68b>

RFE & RFECV:

Selecting optimal features is important part of data preparation in machine learning. It helps us to eliminate less important part of the data and reduce a training time in large datasets.

RFE:

Recursive Feature Elimination (RFE) is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

Reference Link: <https://machinelearningmastery.com/rfe-feature-selection-in-python/>

RFECV:

Recursive Feature Elimination Cross Validation (RFECV) performs recursive feature elimination with cross-validation loop to extract the optimal features. Scikit-learn provides RFECV class to implement RFECV method to find the most important features in a given dataset.

Performance of Model:

|  |  |
| --- | --- |
| Accuracy | Quite essential classification metric.  Easily suited for binary and multiclass classification problem. |
| Precision | When we want to be very sure of our prediction.  Ie. If we are building a system to predict if we should decrease the credit limit on a particular account, we want to be very sure about our prediction or it may result in customer dissatisfaction. |
| Recall | Used when we want to capture as many positives as possible.  Ie. If we are building a system to predict if a person has cancer or not, we want to capture the disease even if we are not very sure. |
| F1 Score | Used when we want to have a model with both good precision and recall.  Ie. If you are a police inspector and you want to catch criminals, you want to be sure that the person you catch is a criminal (Precision) and you also want to capture as many criminals (Recall) as possible. |
| AUC | AUC is the area under the ROC curve.  It measure how well predictions are ranked.  Ie. If you are a marketer want to find a list of users who will respond to a marketing campaign. AUC is a good metric to use since the predictions ranked by probability is the order in which you will create a list of users to send the marketing campaign.  It measure the quality of the model’s predictions irrespective of what classification threshold is chosen, unlike F1 score or accuracy which depend on the choice of threshold. |

Conclusion: From our case, I will decide to use F1 Score to observe the performance.

Diagnostics / Goodness of Fit (Weighted Logistic Regression):

We must decide which model to use. As a result, we should indeed run the test to see how well a model fits a given set of data or predicts a future set of observations.

Log Loss Ratio Test:

Log loss, aka logistic loss or cross-entropy loss.

This is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of a logistic model that returns y\_pred probabilities for its training data y\_true. The log loss is only defined for two or more labels.

The log-likelihood was represent the higher the log-likelihood, the better a model fits a dataset.

However, since the Log loss functions defined as the negative log-likelihood, means the lower the log-loss, the better the performance of the model.

Psuedo R^2:

In ordinary least square (OLS) regression, the R Square statistics measures the amount of variance explained by the regression model. The value of R Square ranges in 0 and 1, with a larger value indicating more variance is explained by the model (higher value is better).