**Machine Learning**

The ultimate goal of machine learning is to design algorithms that automatically help a system gather data and use that data to learn more. Systems are expected to look for patterns in the data collected and use them to make vital decisions for themselves.

**Problem Definition**

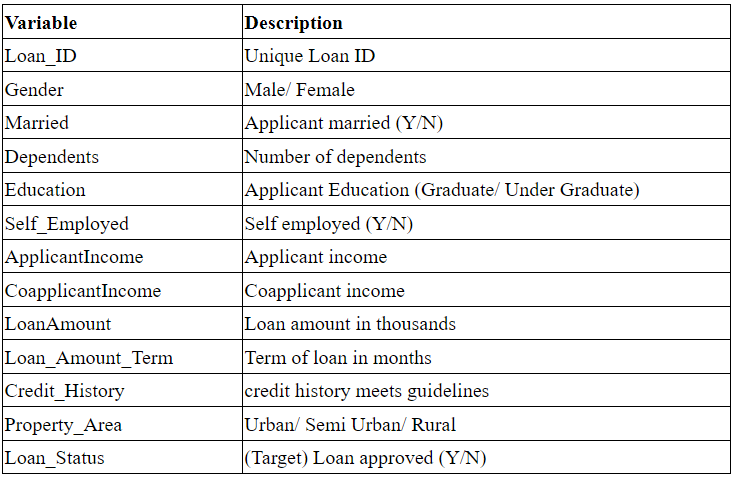
You have to build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset. The details mentioned are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. Dataset is provided to identify the customer segments that are eligible for loan amounts so that they can specifically target these customers. This is a Binary Classification problem in which we need to predict our Target label which is “Loan Status”.

Loan status can have two values: Yes or NO

Yes: if the loan is approved

NO: if the loan is not approved

So, using the training dataset we will train our model and try to predict our target column that is “Loan Status” on the test dataset.



This is a supervised classification problem to be trained with algorithms like:

Logistic Regression

Decision Tree

Random Forest

**Data Analysis**

Firstly, we need to import the required dataset as well as libraries.

Pandas is a Python package to work with structured and time series data. The data from various file formats such as csv, excel, sql etc can be imported using Pandas. It is a powerful open-source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling.

Seaborn is a python library for building graphs to visualise data. It provides integration with pandas.

Sklearn is helpful for building machine learning and statistical models such as clustering, classification, regression etc.

Data exploration & Analysis means reviewing such things as the type and distribution of data contained within each variable, the relationships between variables and how they vary relative to the outcome you're predicting or interested in achieving.

Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values. This also involves removing features with high cardinality (many unique values) as they lead to misleading results. For e.g., date.

Replacing With Mean is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first. Mode is the most frequently occurring value. It is used in the case of categorical features. Median is the middlemost value. It’s better to use the median value for imputation in the case of outliers.

**EDA Concluding Remark**.

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. It helps data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA provides a better understanding of data set variables and the relationships between them.

count plot() method is used to Show the counts of observations in each categorical bin using bars.

A histogram is the most commonly used graph to show frequency distributions. It looks very much like a bar chart, but there are important differences between them.

Predictor Variables

There are 3 types of Independent Variables: Categorical, Ordinal & Numerical.

Categorical Features

• Gender

• Marrital Status

• Employment Type

• Credit Hystory

Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories are not known

Ordinal data are discrete integers that can be ranked or sorted.For example, the distance between first and second may not be the same as the distance between second and third.

**Correlation**

Correlation explains how one or more variables are related to each other. It’s a bi-variate analysis measure which describes the association between different variables

It can be inferred from the below bar plots that in our observed data:

1. 80% of loan applicants are male in the training dataset.
2. Nearly 70% are married
3. About 75% of loan applicants are graduates
4. Mostly, People with income within 10000 have applied for loans. The histogram is rightly skewed that means lower values.
5. Nearly 85–90% loan applicants are not self-employed. This means, they are working with some institute or org. and not have business of their own.
6. The loan has been approved for more than 65% of applicants.
7. Normal loan amount term is 360 days.

**Pre-Processing Pipeline**.

This includes checking skewness, removing outliers and labelling the data.

Skewness is simply a reflection of a data set in which activity is heavily condensed in one range and less condensed in another.

Data skewed to the right is usually [a result of a lower boundary in a data set](http://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm) (whereas data skewed to the left is a result of a higher boundary).

[Outliers](https://statisticsbyjim.com/glossary/outliers/) are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions.

Z score is also called standard score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.

Z score = (x -mean) / std. deviation

If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

**Label Encoding** refers to converting the labels into a numeric form so as to convert them into the machine-readable form. It is an important pre-processing step for the structured dataset in supervised learning.

Object-type data has been transformed to integer type.

Loan ID has been dropped due to cardinality and dependents data-type has been changed separately from object to integer after replacing 3+ by 3.

**Building Machine Learning Models.**

Train/Test is a method to measure the accuracy of your model.

It is called Train/Test because you split the the data set into two sets: a training set and a testing set.

80% for training, and 20% for testing.

You train the model using the training set.

You test the model using the testing set.

Train the model means create the model.

Test the model means test the accuracy of the model.

**Decision Tree** is a Supervised learning technique that can be used for both classification and Regression problems.It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

 "**Random Forest** is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

**Logistic regression** predicts the output of a categorical dependent variable. Therefore , the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic regression uses the concept of predictive modelling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

**Accuracy Score**

Accuracy is a metric used in classification problems used to tell the percentage of accurate predictions. We calculate it by dividing the number of correct predictions by the total number of predictions.

(True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives).

**Hyperparameter Tuning**

Hyperparameters control the over-fitting and under-fitting of the model. Optimal hyperparameters often differ for different datasets. To get the best hyperparameters the following steps are followed:

1. For each proposed hyperparameter setting the model is evaluated

2. The hyperparameters that give the best model are selected.

Choosing the correct hyperparameters for machine learning or deep learning models is one of the best ways to extract the last juice out of your models.

[Hyperparameter tuning](https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624) (or hyperparameter optimization) is the process of determining the right combination of hyperparameters that maximizes the model performance. It works by running multiple trials in a single training process.

GridSearchCV searches for best set of hyperparameters from a grid of hyperparameters values.

Model hyperparameters in different models:

Learning rate in gradient descent

Number of iterations in gradient descent

Number of layers in a Neural Network

Number of neurons per layer in a Neural Network

Number of clusters(k) in k means clustering

**After hyper tuning the parameters, the accuracy score turned out to be 82%.**

**ROC Curve-** An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

True Positive Rate

False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

TPR=TPTP+FN

False Positive Rate (FPR) is defined as follows:

FPR=FPFP+TN

An ROC curve plots TPR vs. FPR at different classification thresholds

Concluding Remarks.

**Saving the model**

In machine learning, while working with [scikit learn](https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/)library, we need to save the trained models in a file and restore them in order to reuse them to compare the model with other models, and to test the model on new data. The saving of data is called Serialization, while restoring the data is called **Deserialization**.

The pickle module implements a fundamental, but powerful algorithm for serializing and de-serializing a Python object structure.

Pickle model provides the following functions –

pickle.dump to serialize an object hierarchy, you simply use dump().

pickle.load to deserialize a data stream, you call the loads() function.

**Loaded model score is 79.31%**