***Introduction***

The two most pressing issues in the banking sector are: 1) How risky is the borrower? 2) Should we lend to the borrower given the risk? The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies. This model extracts and introduces the essential features of a borrower that influence the customer's loan status. Finally, it produces the planned performance (loan status). These reports make a bank manager's job simpler and quicker.

***Literature Review***

We start our literature review with more general systematic literature reviews that focus on the application of machine learning in the general field of Banking Risk Management. Since the global financial crisis, risk management in banks has to take a major role in shaping decision-making for banks. A major portion of risk management is the approval of loans to promising candidates. But the black-box nature of Machine learning algorithms makes many loan providers vary the result. Martin Leo, Suneel Sharma and k. Maddulety's extensive report has explored where Machine Learning is being used in the fields of credit risk, market risk, operational risk, and liquidity risk only to conclude that the research falls short of extensive research is required in the field.

We could not find any literature review for loan prediction for specific Machine learning algorithms to use which would be a possible starting point for our paper. Instead, since loan prediction is a classification problem, we went with popular classification algorithms used for a similar problem. Ashlesha Vaidya used logistic regression as a probabilistic and predictive approach to loan approval prediction. The author pointed out how Artificial neural networks and Logistic regression are most used for loan prediction as they are easier comparatively develop and provide the most accurate predictive analysis. One of the reasoning behind this that that other Algorithms are generally bad at predicting from non-normalized data. But the nonlinear effect and power terms are easily handled by Logistic regression as there is no need for the independent variables on which the prediction takes place to be normally distributed.

Logistic regression still has its limitations, and it requires a large sample of data for parameter estimation. Logistic regression also requires that the variables be independent of each other otherwise the model tends to overweigh the importance of the dependent variables.

A solution to this multicollinearity problem among the categorical explanatory variables is the use of a categorical principal component analysis which can be seen used by Guilder and Ozlem on a case study for housing Loan approval data. The goal of Principal component analysis is to reduce the number of m variables where many of them would be highly correlated with each other, to a smaller set of n uncorrelated variables called principal components which account for the variances between the previous m variables. Methods such as PCA are known as dimension reduction of the data. It may be suitable for scaled continuous variables but it isn’t quite an appropriate method of dimension reduction for categorical variables. Thus, the authors here used a tweaked version of PCA for categorical data called CATPCA or categorical (nonlinear) principal components analysis which is specifically developed for where the dependent variables are a mix of nominal, ordinal, or numeric data which may not have linear relationships with each other. CATPCA works by using a scaling process optimized to convert the categorical variables into numeric variables.

Similar to PCA, Zaghdoudi, Djebali & Mezni compared the use of Linear Discriminant Analysis versus Logistic Regression for Credit Scoring and Default Risk Prediction for foreseeing default risk o small and medium enterprises. Linear Discriminant Analysis (LDA) is like PCA for dimensionality reduction but instead of looking for the most variation, LDA focuses on maximizing the separability among the know categories. This subspace that well separates the classes is usually in which a linear classifier can be learned. The classification of those enterprises correctly in their original groups through both these methods was inconsequential with Logistic regression having a 0.3% better accuracy score than LDA.

Another novel approach for T.Sunitha and colleagues was to predict loan Status using Logistic Regression and a Binary Tree. Decision Tree is an algorithm for a predictive type machine learning model.

Classification and Regression Trees are referred to as CART (in short) introduced by Leo Breiman. It best suits both predictive and decision modeling problems. This Binary Tree methodology is the greedy method is used for the selection of the best splitting. Although Decision trees gave us a similar accuracy. The benefits of Decision Trees, in this case, were due to the latter giving equal importance to both accuracy and prediction. This model became successful in making a lower number of False Predictions to reduce the risk factor.

Rajiv Kumar and Vinod Jain proposed a model using machine learning algorithms to predict the loan approval of customers. They applied three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) using Python on a test data set. From the results, they concluded that the Decision Tree machine learning algorithm performs better than Logistic Regression and Random Forest machine learning approaches. It also opens other areas on which the Decision Tree algorithm is applicable.

Some machine learning models give different weights to each factor but in practice sometimes loans can be sanctioned based on a single strong factor only. To eliminate this problem J. Tejaswini and T. Mohana Kavya in their research paper have built a loan prediction system that automatically calculates the weight of each feature taking part in loan processing and on new test data the same features are processed concerning their associated weight. They have implemented six machine learning classification models using R for choosing the deserving loan applicants. The models include Decision Trees, Random Forest, Support Vector Machine, Linear Models, Neural Network and Adaboost. The authors concluded that the accuracy of the Decision Tree is highest among all models and performs better on the loan prediction system.

Predicting loan defaulters is an important process of the banking system as it directly affects profitability. However, loan default data sets available are highly imbalanced which results in poor performance of the algorithms. Lifeng Zhou and Hong Wang in their call for paper made loan default prediction on imbalanced data sets using an improved random forests approach. In this approach, the authors have employed weights in decision tree aggregation. The weights are calculated and assigned to each tree in the forest during the forest construction process using Out-of-bag (OOB) errors. The experimental results conclude that the improved algorithm performs better and has better accuracy than the original random forest and other popular classification algorithms such as SVM, KNN, and C4.5. The research opens improvements in terms of efficiency of the algorithm if parallel random forests can be used for further work.

Anchal Goyal and Ranpreet Kaur discuss various ensemble algorithms. Ensemble algorithm is a supervised machine learning algorithm that is a combination of two or more algorithms to get better predictive performance. They carried out a systematic literature review to compare ensemble models with various stand-alone models such as neural network, SVM, regression, etc. The authors after reviewing different literature reviews concluded that the Ensemble Model performs better than the stand-alone models. Finally, they concluded that the concept of combined algorithms also improves the accuracy of the model.

Data Mining is also becoming popular in the field banking sector as it extracts information from a tremendous amount of accumulated data sets. Aboobyda Jafar Hamid and Tarig Mohammed Ahmed focused on implementing data mining techniques using three models j48, bayesNet, and naiveBayesdel for classifying loan risk in the banking sector. The author implemented and tested models using the Weka application. In their work, they made a comparison between these algorithms in terms of accuracy in classifying the data correctly. The operation of sprinting happened in a manner that 80% represented the training dataset and 20% represented the testing dataset. After analyzing the results the author came up with the results that the best algorithm among the three is the J48w algorithm in terms of high accuracy and low mean absolute error.

***Methodology***

Machine Learing and Concepts

Four machine learning models have been used for the prediction of loan approvals. Below are the description of the models used:

Logistic Regression

This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome. The logistic function used is a logit function which is a log of odds in the favor of the event. Logit function develops a s-shaped curve with the probability estimate similar to a step function.

Decision Trees

This is a supervised machine learning algorithm mostly used for classification problems. All features should be discretized in this model, so that the population can be split into two or more homogeneous sets or subsets. This model uses a different algorithm to split a node into two or more sub-nodes. With the creation of more sub-nodes, homogeneity and purity of the nodes increases with respect to the dependent variable.

Random Forest

This is a tree based ensemble model which helps in improving the accuracy of the model . It combines a large number of Decision trees to build a powerful predicting model. It takes a random sample of rows and features of each individual tree to prepare a decision tree model. Final prediction class is either the mode of all the predictors or the mean of all the predictors.

XGBoost

This algorithm only works with the quantitative variable. It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed.

Libraries for Data Analysis

The models are implemented using Python 3.7 with listed libraries:

Pandas

Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, 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

Seaborn is a python library for building graphs to visualise data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data.

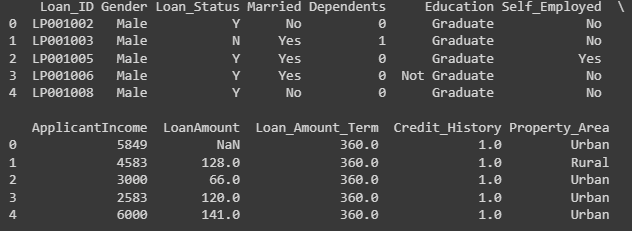
Sklearn

This python library is helpful for building machine learning and statistical models such as clustering, classification, regression etc. Though it can be used for reading, manipulating and summarizing the data as well, better libraries are there to perform these functions.

Understanding the Dataset

The machine learning model is trained using the training data set. Every new applicant details filled at the time of application form acts as a test data set. On the basis of the training data sets, the model will predict whether a loan would be approved or not. We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan\_Status in train dataset and 12 independent variables in test dataset. The Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status are all categorical

The Dataset features are :



The Datatypes of features are as :

Loan\_ID object

Gender object

Married object

Dependents object

Education object

Self\_Employed object

ApplicantIncome int64

CoapplicantIncome float64

LoanAmount float64

Loan\_Amount\_Term float64

Credit\_History float64

Property\_Area object

Loan\_Status object

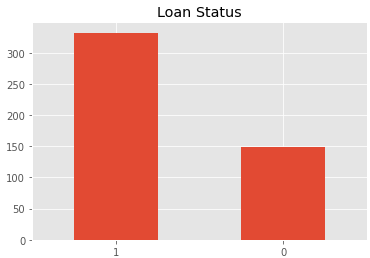
dtype: object

***Exploratory Data Analysis***

***Visual Analysis***

Target Variable - Loan Status

We will start first with an independent variable which is our target variable as well. We will analyse this categorical variable using a bar chart as shown below. The bar chart shows that loan of 322 (69.16 % ) people out of 480 was approved



***Predictor Variables***

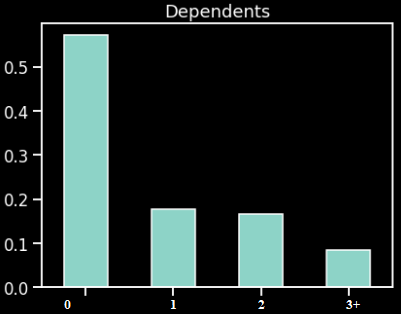
There are 3 types of Independent Variables: Categorical, Ordinal & Numerical.

Categorical Features

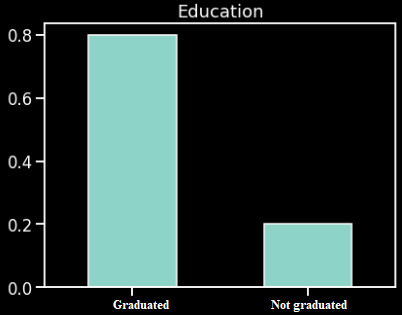
1. Gender
2. Marital Status
3. Employment Type
4. Credit History

Ordinal Features

1. Number of Dependents



1. Education Level



1. Property or Area Background



Numerical Features

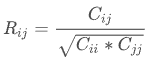
1. The Applicant's Income
2. The Co-Applicant's Income

***Visualizing Correlations via HeadMap***

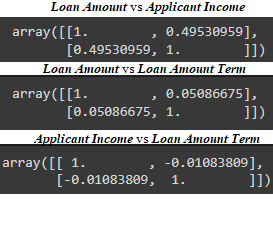
******

***Visualising Correlations between Quantitative variables***

corrcoef returns the correlation coefficient matrix of the variables..



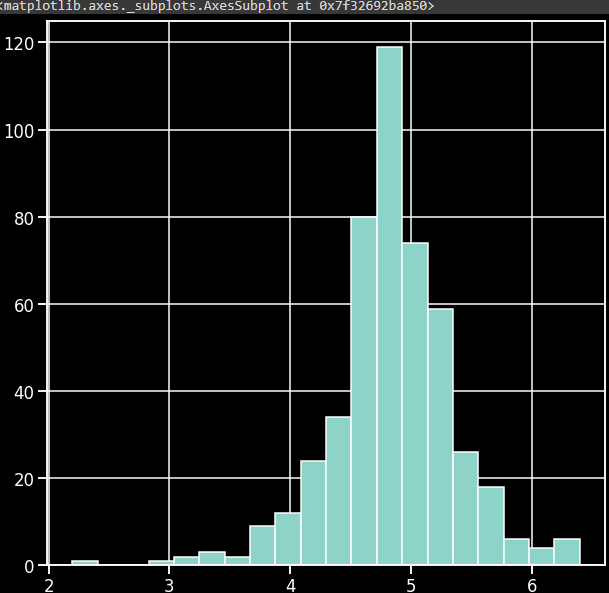
R lies between -1 to 1.



* Correlation between loan amount and applicant income is 49.5%
* Correlation between loan amount and applicant income is 50.86%
* Correlation between loan amount and applicant income is -1.08% (which denotes they both change in opposite direction).

***Outlier***

Due to outliers in the Loan Amount. the data in the loan amount is skewed towards the right, which means bulk of the data is towards the left. We remove this skewness by doing a log transformation. A log transformation doesn't effect affect the smaller values much but reduces the larger values. So the distribution becomes normal.

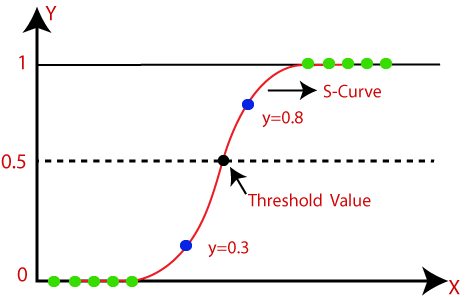


***Logistic Regression***

Introduction-

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
* 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 is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
* In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
* The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

Logistic Regression in Machine Learning-



Logistic Function (Sigmoid Function):

* The sigmoid function is a mathematical function used to map the predicted values to probabilities.
* It maps any real value into another value within a range of 0 and 1.
* The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
* In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.
* Assumptions for Logistic Regression:
* The dependent variable must be categorical in nature.
* The independent variable should not have multi-collinearity.

Logistic Regression Equation:

* The Logistic regression equation can be obtained from the Linear Regression equation.
* In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):



Type of Logistic Regression: On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial**: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial**: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
3. **Ordinal**: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

***Program***

train\_x = pd.get\_dummies(train\_x)

test\_x  = pd.get\_dummies(test\_x)

print('shape of training data : ',train\_x.shape)

print('shape of testing data : ',test\_x.shape)



model = LogisticRegression()

model.fit(train\_x,train\_y)

predict = model.predict(test\_x)

print('Predicted Values on Test Data',predict)

print('\n\nAccuracy Score on test data :',accuracy\_score(test\_y,predict))

