**The University of Memphis**

**Project Report**

**COMP 7118 Data Mining Fall 2022**

**Ramesh Kumar Surapuredy**

**UID: U00831213**

**Contact email:**

[**rsrprddy@memphis.edu**](mailto:rsrprddy@memphis.edu)

**Hemanth Venumbaka**

**UID: U00870132**

**Contact email:**

[**hvnmbaka@memphis.edu**](mailto:hvnmbaka@memphis.edu)

**Abstract:**

The major aim of this project is to predict which of the customers will have their loan paid or not. we are developing loan prediction system using machine learning, so the system automatically selects the eligible candidates. we are predicting the loan data by using some machine learning algorithms like Decision Tree, Logistic Regression, Random Forest, XGBoost and Support Vector Machine (SVM).

**Introduction:**

In order to predict whether to approve a loan, the lender must consider a number of background facts about the applicant. The parameters that determine whether the loan is granted include things like credit score, loan amount, lifestyle, career, and assets. It is more probable that your loan will be approved if previous borrowers with criteria similar to yours have made on-time payments.

This reliance on prior knowledge and comparisons with other applicants can be taken advantage of by machine learning algorithms, which can then be used to create a data science issue to forecast the loan status of a new applicant using a set of analogous criteria.

To determine the loan status, several collections of information from previous loan applicants use various features. This data, which may be static or time-series, can be examined by a machine learning model to estimate the likelihood that the loan would be authorized. Now let's examine a few of these datasets.

**Dataset:**

In dataset there are 614 rows and 13 columns. 12 independent variable and target variable: (614, 13).

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 data.

This dataset allows to work on supervised learning, more preciously a classification problem. This is the reason why I would like to introduce you to an analysis of this one. We have data of some predicted loans from history.I have explored dataset and found a lot interesting facts about loan prediction. The first part is going to focus on data analysis and Data visualization. The second one we are going to see the about algorithm used to tackle our problem.

**Algorithms:**

In this project we have built five different models and tested with data. The five algorithms are:

* Decision Tree
* Random Forest
* XGBoost
* Logistic Regression
* Support Vector Machine (SVM)

**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.

**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.

**Support Vector Machine (SVM):**

“Support Vector Machine” (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_1.png)

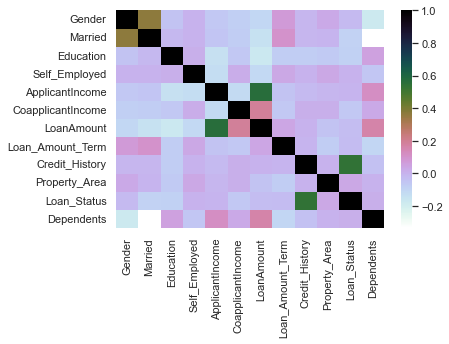
Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).

**Results:**

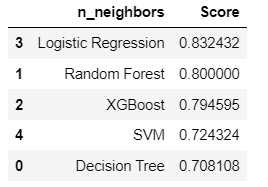
By using EDA and Feature Engineering techniques, we came to tackle the challenges. By looking at the columns description we can make many assumptions like

1. The one whose salary is more can have a greater chance of loan approval.
2. The one who is graduate has a better chance of loan approval
3. Married people would have a upper hand than unmarried people for loan approval .
4. The applicant who has less number of dependents have a high probability for loan approval.
5. The lesser the loan amount the higher the chance for getting loan.

From the below correlation matrix. We can see that Credit\_History has a positive correation. So we can conclude that the target value(Loan\_Status) is more depent on Credit\_History

****

* We constructed models taking different variables into account and found through odds ratio that credit credit history is creating the most impact on loan giving decision. Finally, we got a model with co-applicant income and credit history as independent variable with highest accuracy.
* We tested the data and got the accuracy of 83 %. Among all the models Logistic Regression gives maximum accuracy.
* Credit\_History is the variable that impacts the Loan\_Status variable (Higher Dependencies)
* The Logisitic Regression Model is the most accurate model : 83.24% Accurate



**Conclusion:**

* We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed.
* We analysed each variable to check if data is cleaned and normally distributes
* We cleaned the data and removed NA values
* We also generated hypothesis to prove an association among the independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.
* We calculated correlation between independent variables and found that applicant income and loan amount have significant relation.
* We created dummy variables for constructing the model
* We constructed models taking different variables into account and found through odds ratio that credit credit history is creating the most impact on loan giving decision
* Finally, we got a model with coapplicant income and credit history as independent variable with highest accuracy.
* We tested the data and got the accuracy of 83 %

**Reference:**

[**https://ijirt.org/master/publishedpaper/IJIRT151769\_PAPER.pdf**](https://ijirt.org/master/publishedpaper/IJIRT151769_PAPER.pdf)

[**https://www.irjmets.com/uploadedfiles/paper//issue\_6\_june\_2022/26544/final/fin\_irjmets1655891933.pdf**](https://www.irjmets.com/uploadedfiles/paper//issue_6_june_2022/26544/final/fin_irjmets1655891933.pdf)

[**https://www.slideshare.net/irjetjournal/bank-loan-prediction-using-machine-learning**](https://www.slideshare.net/irjetjournal/bank-loan-prediction-using-machine-learning)

[**https://www.kaggle.com/code/komalparakh2509/loan-prediction/data**](https://www.kaggle.com/code/komalparakh2509/loan-prediction/data)