**A PROJECT REPORT**

**ON**

**LOAN APPROVAL PREDICTION USING**

**VARIOUS MACHINE LEARNING ALGORITHMS**

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

Data Analytics, Machine Learning and AI using Python



*Submitted By*

**Rajan Jangir (**Engineering College Bikaner, Bikaner**)**

*Under the guidance of*

**Mr. Bipul Shahi**

**ACKNOWLEDGEMENT**

My sincere gratitude and thanks towards my project paper guide Bipul Shahi, Corporate Trainer, Developer, Traveller!!! IOT, Artificial Intelligence, Robotics, Cloud Computing, Android Apps !!!

It was only with his backing and support that I could complete the report. He provided me all sorts of help and corrected me if ever seemed to make mistakes. I have no such words to express my gratitude. I acknowledge my sincere gratitude to MR. SHUBHAM JANGIR (my elder brother) for their valuable guidance and helping attitude even in their very busy schedule. Without his help and support I couldn’t even start the work. So I am grateful to him. And at last but not the least, I acknowledge my dearest parents for being such a nice source of encouragement and moral support that helped me tremendously in this aspect. I also declare to the best of my knowledge and belief that the Project Work has not been submitted anywhere else.

**ABSTRACT**

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. In recent years many researchers worked on loan approval prediction systems. Machine Learning (ML)techniques are very useful in predicting outcomes for large amount of data. In this paper three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Logistic Regression machine learning algorithm is better as compared to Decision Tree and Random Forest machine learning approaches.

**INTRODUCTION**

Now a day’s people rely on bank loans to fulfill their needs. The rate of loan applications increases with a very fast speed in recent years. Risk is always involved in approval of loans. The banking officials are very conscious about the payment of the loan amount by its customers. Event after taking lot of precautions and analyzing the loan applicant data, the loan approval decisions are not always correct. There is need of automation of this process so that loan approval is less risky and incur less loss for banks

Artificial Intelligence AI is an emerging technology now a day. The application of AI solves many problems of the real world. Machine Learning is an AI technique which is very useful in prediction systems. Figure 1 is showing a basic model of machine learning. It creates a model from a training data. While making the prediction the model which is developed by training algorithm (which is machine learning) is used. The machine learning algorithm trained the system using a fraction of the data available and test the remaining data.

The machine Learning techniques can be applied on a sample test data first and then can be used in making prediction related decisions. This paper applied the machine learning approaches in solving loan approval problem of banking sector. Next section discusses the literature survey.

Then proposed work, results and analysis are discussed. Finally, conclusion and future scope is discussed which is followed by the references used in this paper.

**1.Problem statement**

A Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a data set.

**2.Data Collection**

The training set will be used for training the model, i.e. our model will learn from this data. It contains all the independent variables and the target variable. The test set contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. There are 13 columns of features and 614 rows of records in the training set and 12 columns of features and 367 rows of records in the test set. The dataset variables are summarized as below:

| **Variable** | **Description** |
| --- | --- |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Under Graduate) |
| Self\_Employed | Self employed (Y/N) |
| ApplicantIncome | Applicant income |
| CoapplicantIncome | Coapplicant income |
| LoanAmount | Loan amount in thousands |
| Loan\_Amount\_Term | Term of loan in months |
| Credit\_History | credit history meets guidelines |
| Property\_Area | Urban/ Semi Urban/ Rural |
| Loan\_Status | Loan approved (Y/N) |

**3.Technology and Concepts**

**Machine Learning**

Learning algorithms are widely used in computer vision applications. Before considering loan approval related tasks, we are going to have a brief look at basics of machine learning.

Machine learning has emerged as a useful tool for modelling problems that are otherwise difficult to formulate exactly. Classical computer programs are explicitly programmed by hand to perform a task. With machine learning, some portion of the human contribution is replaced by a learning algorithm. As availability of computational capacity and data has increased, machine learning has become more and more practical over the years, to the point of being almost ubiquitous.

It can be used in two ways:

* *Supervised Learning*
* *Unsupervised Learning*

**Machine Learning Algorithms**

Machine learning algorithm which are used in this work to make a model are as follows:

1. Logistic Regression

2. Decision Tree

3. Random forest.

**1.Logistic Regression**

Logistic Regression (LR) is a machine learning technique. The LR is very commonly used to solve binary classification problem. There are following basic postulation:

1. Binary logistic regression has binary dependent variables.

2. In binary regression dependent variables have level 1.

3. The included variables should have meaning. All included independent variables should be self-reliant.

4. The independent variables are related to the log odds linearly.

5. The sample size should be large for LR.

**2.Decision Tree**

Decision TREE is a supervised ML technique which is non parametric in nature. It has predefined target variable which is generally used in problem classification. It is useful for classification and regression both. It works categorical & continuous both for input and output variables.

**3.Random Forest**

Random Forest (RF) is a very useful machine learning algorithm. It is mostly used in areas such as classification, regression analysis etc. At the training time RF algorithm creates many decision trees. RF is a supervised learning approach which need a test data for the model for training. It creates random forests for the problem set and then find the solution using these random forests.

**4.Results and Analysis**

Three machine learning approaches are applied on the test data to predict the loan approvals of loan requests. Python programming language is used to implement machine learning algorithms. For training 70 percent data is used and 30 percent data is used for testing. The prediction accuracy of the different ML approaches is calculated and compared. The training data set is shown in figure 2. On the basis of this train data set (shown in figure 1), system analyze rest of 30 percent data and predict the results in term of loan status either accepted or rejected.

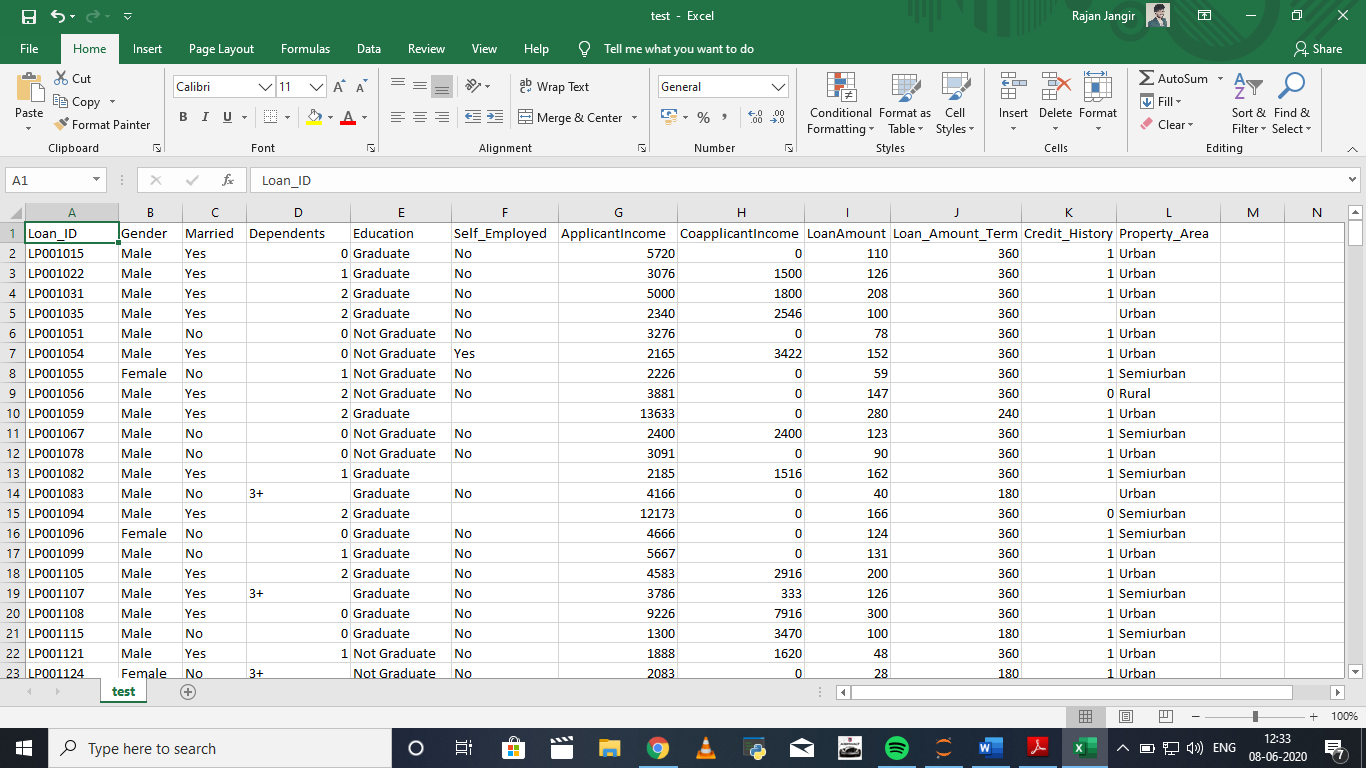


Figure-1. Testing data set

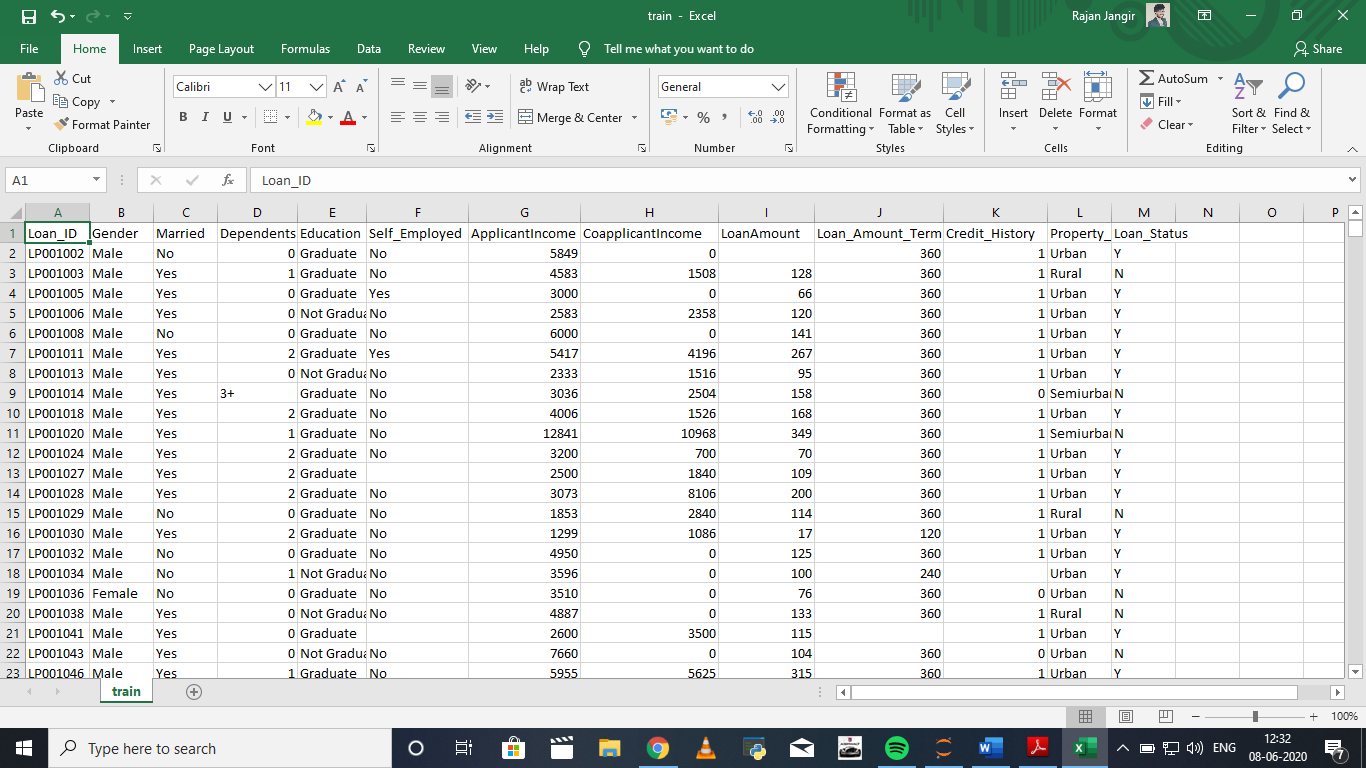


Figure-2. Training data set

Comparison analysis of training and testing accuracy for three machine learning algorithms is shown in table -1.

Table-1: Comparison of training and testing accuracy of machine learning algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr.No.** | **Machine learning Algorithm** | **Accuracy Percentage** | |
| **Training** | **Testing** |
| 1. | Linear Regression | 81% | 79% |
| 2. | Decision Tree | 100% | 70% |
| 3. | Random Forest | 100% | 77% |

Results with loan status by applying the Logistic Regression (shown in figure-3(a)), Decision Dree (shown in figure-3(b)) and Random Forest (shown in figure-3(c))

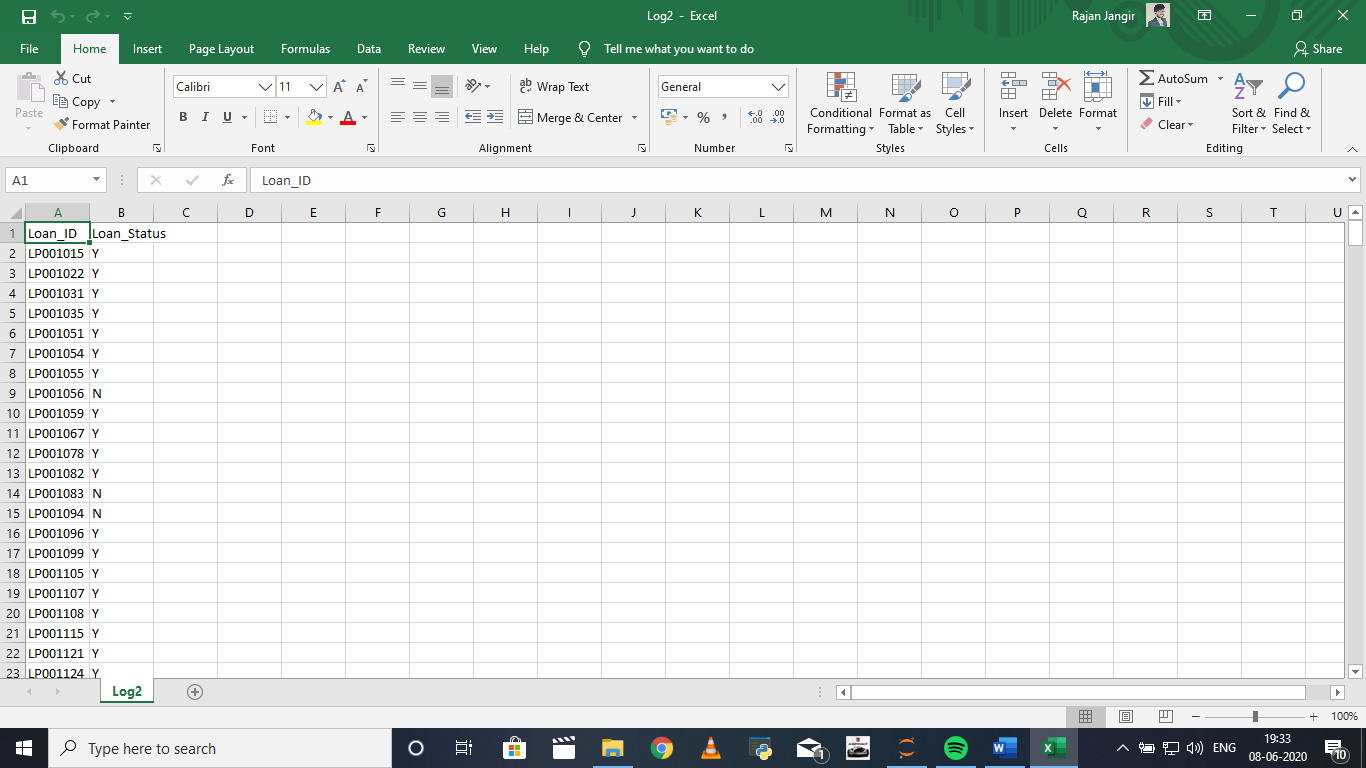


Figure-3(a). Logistics Regression Result with Loan Status.

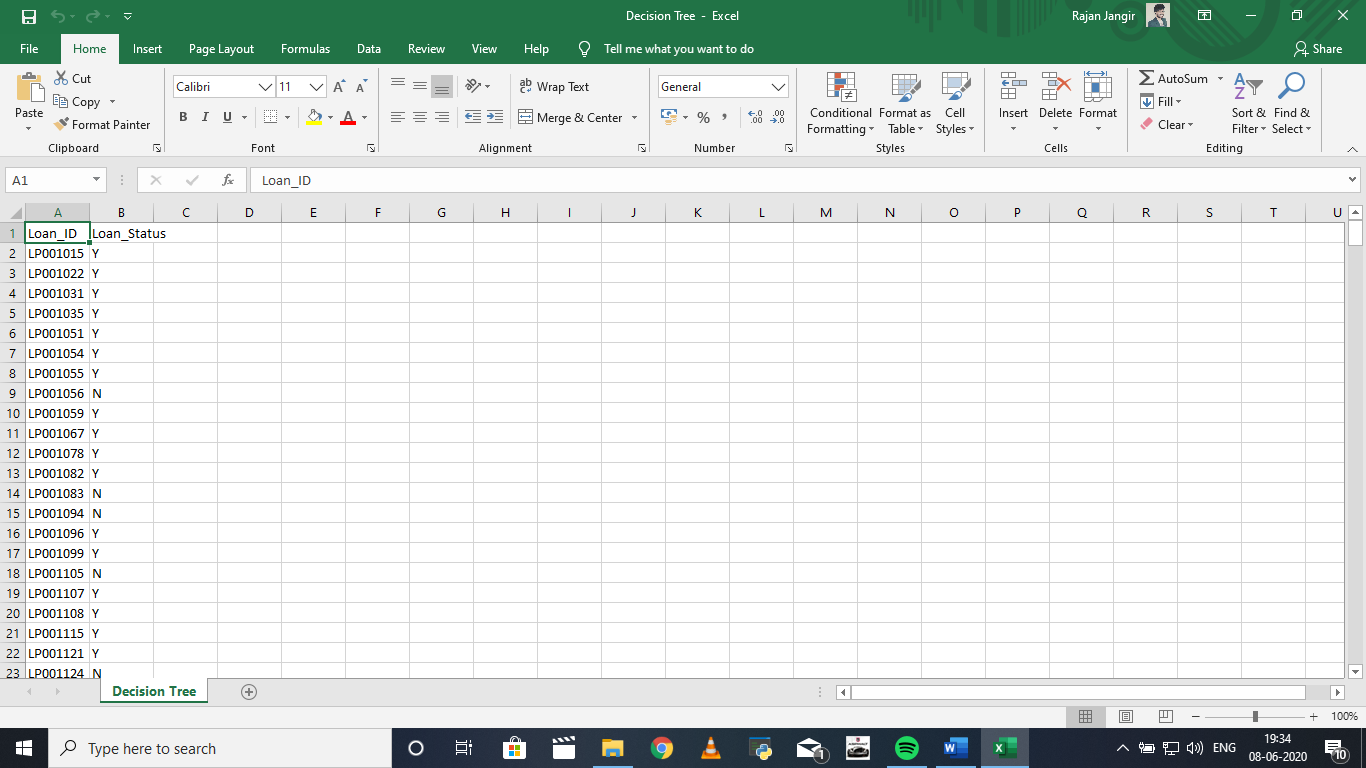


Figure-3(b). Decision Tree Result with Loan Status.

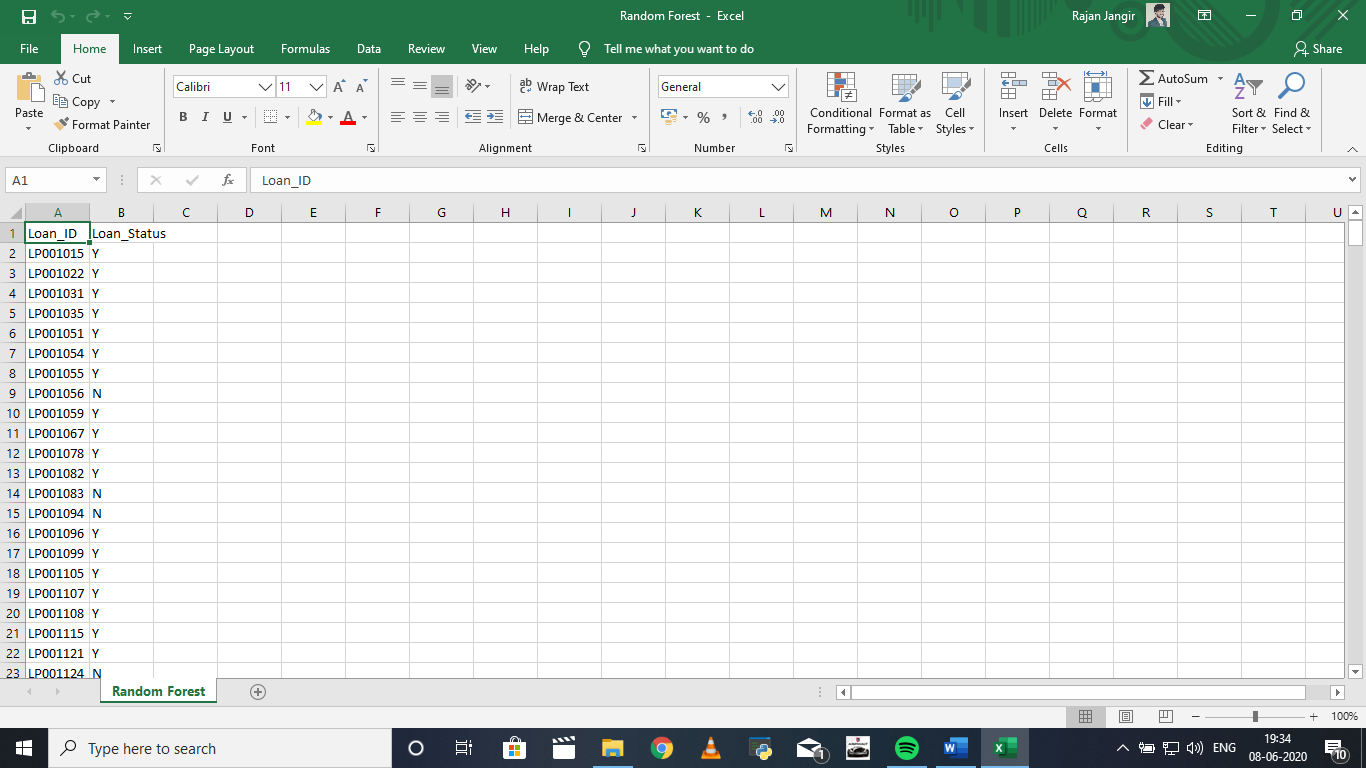


Figure-3(c). Random forest Result with Loan Status

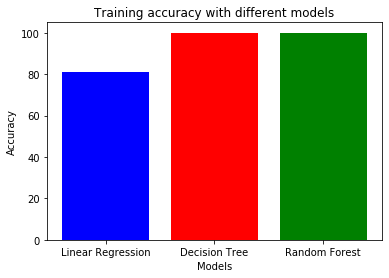


Figure-4. Training accuracy with different machine learning algorithms.

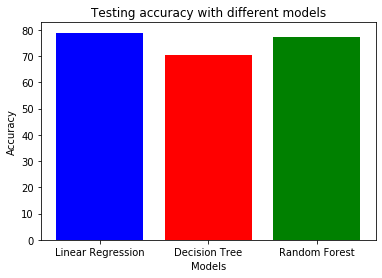


Figure-5. Testing accuracy with different machine learning algorithms.

**Conclusion**

After trying and testing 3 different algorithms, the best accuracy on the public leaderboard is achieved by Logistic Regression (79%), followed by Random Forest (77%) and Decision Tree performed the worst (70%). While new features created via feature engineering helped in predicting the target variable, it did not improve the overall model accuracy much. On the whole, a logistic regression classifier provides the best result in terms of accuracy for the given dataset, without any feature engineering needed. Because of its simplicity and the fact that it can be implemented relatively easy and quick, Logistic Regression is often a good baseline that data scientists can use to measure the performance of other more complex algorithms. In this case, however, a basic Logistic Regression has already outperformed other more complex algorithms like Random Forest and Decision Tree, for the given dataset.

A rigorous analysis of other machine learning algorithms other than these three can also be done in future to investigate the power of machine learning algorithms for loan approval prediction.

*In summary, this project is helpful to me in understanding how to build a machine learning model and what are the steps one should go through to build a robust model. If you have any feedback for this project, feel free to contact me via my*[*LinkedIn*](https://www.linkedin.com/in/rajan-jangir-374a97173/)

Bibliography

* <https://www.kaggle.com/>
* <https://medium.com/>
* <https://towardsdatascience.com/>