**Dissertation Report: Loan Prediction Approval**

**Aims and Objectives:**

**Aim:** The aim of this dissertation report is to make a machine learning algorithm to predict the loan prediction approval rate for the customer. Loan is one of the most critical parts of any banking business as it ensures the capital return to the bank with lucrative interest over it. Machine learning algorithms are being implemented to solve the problem of prediction of loan approval to the customer based on the profile associated with it, the use of machine learning algorithms have been used for this purpose for quite a long period, however the selection of best machine learning approach for building the model to get inference for generating the approval is a discussion of its own. Upon observation at very basic lower-level machine learning algorithms is nothing but the parametric mathematical equation that involves the branches of linear algebra, statistics, and calculus, this makes the understanding and implementation of this mathematics dependent algorithm quite difficult and takes a crystal clarity in the concept to select the best algorithm for this downstream implementation. This report aims to provide the solution for this very purpose by solving the defined problem using two separate algorithms that is:

**Random Forest** and **Support Vector Machine.** The main differentiating factor this report aims to describe is the use of **Grid Search CV** a type of hyperparameter tunning that will find the best parameters for both the machine learning algorithms, that will otherwise require domain knowledge of math and test-trials to obtain best hyperparameters.

**Objectives:** The objectives proposed in this report are summarized below:

* Use of two machine learning models that is **Random Forest** and **Support Vector Machine.**
* Use of Data visualization techniques like countplot, stripplot and violinplot.
* Use of Feature Engineering and extensive visualization to study relation between independent and dependent variables
* Use of model specific feature scaling using minmax scaler
* Transformation of dataset with accordance to avoid the data leakage.
* Use of cross validation for regularization, ensuring removal of overfit prone models.
* Use of **Grid Search CV** for Hyperparameter Tuning.
* Use of three different evaluation metrics that is **Classification Report, Confusion Matrix, ROC-AUC Score.**

**Literature Review:** The notion of building a standalone software that leverages the power of machine learning algorithm to predict the loan approval to the customer has always been a class problem set in the field of machine learning, the nature of the problem itself makes it popular as it provides an amalgamation of machine learning usage in real life business problem solving. Loan generates interest which contribute to a significant chunk of profit share for any bank, this factor alone defines the importance of loan approval prediction, any bank key focus is to operate the business effectively by providing the loan waivers to the customers including individuals, company etc. The interest that is generated along the installments paid back to the bank periodically provides the bank with profit and the extra liquid capital to redirect for investment.

Credit Scoring mechanisms were introduced as the direct consequence of the importance of loan as a source of profit for the Banks , credit scoring mechanism like CIBIL made to assess the creditworthiness of the individual user in India for loan approval is treated as critical metric to determine the decision of loan lending and at what terms to the loan applier , it ranges from 300 to 900 , where the higher number implies more likeliness to get the loan at better terms , the financial literacy of this factor is essential to determine the loan approval and can promote better financial understanding among the applier of loan (Laxmanan, 2021)

There is various type of machine learning algorithms used to solve this popular problem, the category of supervised machine learning algorithms are used. Supervised machine learning algorithms works on labelled dataset, where the outcome is to get the prediction of the new data based on the model made by training on the historical data, the subcategory of supervised machine learning algorithm used for loan approval prediction is known as classification in which the data is discrete type well labelled in nature (Singh et al., 2016). Some of the earliest machine learning algorithms that were used for the purpose are Logistic Regression [Sheikh et al. (2020)] that was able to predict the likelihood of customer to default on loans and become Non-Performing Assets. As the time progressed various type of machine learning algorithms were also explored to obtain better result, which lead to the use of Support Vector Machine and Random Forest Classifiers for loan approval prediction.

Support Vector machine is a potent supervise machine learning algorithm used in classification and regression based problems , the main essence of it lies on finding the best possible hyperplane that separates multiple classes in the feature space, it does by separating the support vector which is essentially data points of different classes by maximizing the margin between them , the further distance between this support vector the better separation of class takes place which results in better classification. (Suthaharan & Suthaharan, 2016).

Random Forest is an ensemble learning technique that construct makes multiple base machine learning model during training and outputs the mode for classification tasks. The training of the individual baseline model is done through bootstrapping the sample of the original derived dataset and by each node a random combination of feature is used for training the base model and the final prediction of the model is obtained through aggregation of all the baseline models. This unique ensembling model ensures the model wont overfit (Biau & Scornet, 2016).

Choosing appropriate hyperparameter is also critical in ensuring the model performance, Grid search and Cross validation when paired with machine learning algorithms proves to provide better accuracy in multiclass scenarios (Budiman, 2019) , while Grid search uses pre-defined grid of parameter to assess the best set of hyperparameter the cross validation in the Grid Search CV act as a regularizer ensuring the model made through grid is cross validated at each fold to avoid overfit. The implementation of hyperparameter search techniques like this proves to be beneficial for solving problems like loan prediction approval.

**Methodology:**

The methodology proposed in this report is divided into the various sections:

* Data Selection: For the rapid model deployment purposes, the data selected for this study is a static data derived from Kaggle:

[https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset]

* Libraries used: This study involves the use of python packages such as: Seaborn, Matplotlib.pyplot, Pandas, Numpy, Sklearn and pickle.

.

* Data Visualization: For data visualization the seaborn library was used extensively, the main motive was to find the different relationship among the dependent and independent variables and decide if there is a scope of remove of any feature that does not make sense to the model fit , in initial plots it was evident that the categorical features were not making much of a sense , however when the stripplot were plotted on same feature some hidden data points were discovered leading to the conclusion of hidden relationships , this discovery assisted a lot in the data preprocessing step.
* Data Preprocessing: The data used in the methodology was preprocess to prepare for the models effectively , starting with removal of the features like loan-id and features that didn’t make any significance to the model and target feature , then the data was feature engineered into two separate dataset for Random Forest and SVM , the categorical features like education and self-employed and the target feature itself where mapped into binary form to make them usable for model fit. Along with this conversion of categorical features MinMax scaling was also implemented to normalize the non-categorical variables ensuring better interpretability and fast convergence for both models. At last, the dataset was divided into a split of 8:2 for train and test subsets and after the creation of this subsets the MinMax scaling was used to normalize the non-categorical features, the scaling was done in such a way that the data leakage was avoided completely by carefully using fit transform for trainable data and transform for test data.
* Hyperparameter Tuning: To get the best model fit to the dataset the use of Grid Search CV was carried out for both models, the instances of both models were initialized, and a default grid of parameter was uses as a standard procedure to initialize the Grid Search CV, the Cross Validation a form of regularization technique was also defined to avoid overfit. The returned hyperparameter were used in both models for best model fit possible.
* Model Selection: For the loan prediction to black box category models were used that is **Random Forest** and **Support Vector Machine,** the reason behind selection of this machine learning algorithm was primarily the nature of the models, these models are known for performing well on binary classification task, they introduce non linearity between features that lead to discovery of hidden relationship between target and features , both models can handle the categorical and non-categorical feature well and the effective ensembling in Random Forest and Ability of SVM to work on high dimensional space make this models ideal for our purpose of binary classification. Neural networks were other candidates that suit our purpose but since the data was very less and factors like overfit, convergence and less interpretability of the neural network proved to be the cause of not using it.
* Model Evaluation: The model was evaluated by using the subset of train and test dataset that were obtained from the original dataset in the ration of 8:2. To assess their accuracy and performance three different metrics were used, they were Classification report, Confusion Matrix and ROC-AUC score, the following table provide the overall performance of both models.

The higher Recall and F1-Score of Random forest provides the insight about the rate at which the model has successfully classified the data points into the classes that they belong to, the second metrics used in the model that is ROC-AUC and Accuracy favored the Random Forest again as the model True positive rate versus False positive rate at various points proved to be better , the last metric used to determine the evaluation was confusion matrix and by observing the confusion matrix of Random Forest it labelled more data points to True positive and True Negatives correctly which justifies the high score and better classification report for Random Forest when compared to the Support Vector Machine Classifier. From the business point of view this result implies that the Random Forest model-based loan approval system will have less tendency to approve the loan to user who are prone to be defaulters and high tendency to approve the loan to the user that will return the loan amount periodically.

**Classification and ROC-AUC Score:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1-Score** | **Support** |
| Random Forest |  |  |  |  |
| Class 0 | 0.98 | 0.96 | 0.97 | 318 |
| Class 1 | 0.98 | 0.99 | 0.98 | 536 |
|  |  |  |  |  |
| SVM |  |  |  |  |
| Class 0 | 0.90 | 0.92 | 0.91 | 318 |
| Class 1 | 0.95 | 0.94 | 0.95 | 536 |
|  |  |  |  |  |

Table 1: Classification Report for Random Forest and SVM.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Random Forest** | **Support Vector Machine** |
| **Accuracy** | 0.98 | 0.93 |
| **ROC-AUC Score** | 0.97 | 0.93 |

Table 2: Accuracy for Random Forest and SVM.

**Confusion Matrix:**

**Random Forest:**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** |
| **Actual 0** | **306** | **12** |
| **Actual 1** | **7** | **529** |

**Support Vector Machine:**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** |
| **Actual 0** | **292** | **26** |
| **Actual 1** | **32** | **504** |

Table 3: Confusion Matrox for Random Forest and SVM.

**Conclusion:**

* **Key Findings:** The Loan Approval model made by using SVM and Random Forest algorithms resulted in the optimal accuracy of 94% and 98% respectively. The methodology used to make the model demonstrated the optimal results and cleared all the objectives thoroughly. The highlight of the work demonstrated in this report implies the use of hyperparameter tuning methods like Grid Search CV can result in fast and better model making to provide the solution to domain specific problem like loan approval, the use of Grid Search CV promises the optimal hyperparameter for various machine learning algorithms without requiring the extensive knowledge of mathematics behind the model arguments and methodology.
* **Limitations:** This work shed the light towards the use of hyperparameter search techniques to obtain optimal results, however there are various limitations in the approach:
  + The model selected are of black box nature which are difficult to interpret leading to low confidence in deploying this model in real world.
  + Data used in this model is of static nature that makes the model prone to overfit.
  + Data used to run the model was small which leads the model to start remembering instead of generalizing.
  + Grid Search CV Parameters were chosen without any domain knowledge indicating that there might exist a combination of parameters for it that will obtain ever better hyperparameter set.
  + Real Life Banking data contain more features than the provided data which makes the model too simple for real world.
  + The Model was prepared without the use of domain knowledge which may become the cause of not leveraging the data correctly.
  + The Evaluation metrics used does not make business sense and hard to evaluate the real-life effectiveness.
  + The approach of using deep learning models were not explored, there exist a wide variety of algorithms and deep learning architectures that can perform even better then proposed methodology.
* **Future Scope:** Using the key finding from this work, the problem can be further explored through:
  + Using XAI to obtain local and global explanation to enhance interpretability this will create increase the chance of accepting the proposed model in real life prediction.
  + Including a non-static data through API available which will ensure the model gets more data to trained leading to better results.
  + Using a large Dataset for model fit in case of unavailability of the API since baking system will not provide the sensitive info can be viewed as the secondary option to train the model effectively.
  + Exploring neural networks architectures that specialize in binary classification could lead to discovery of even better prediction model that can be used to solve the problem better.
  + Including the metrics that make business sense can enhance the real-life value of the proposed model, leading to adoption of the model into banking system.
  + Using other hyperparameter tuning methods can lead to yield of better hyperparameters combinations.
  + Including Ablation study on the proposed methodology can reveal the hidden nuanced relationship between the features.

**References:**

Sheikh, M.A., Goel, A.K. and Kumar, T., 2020, July. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 490-494). IEEE.

Suthaharan, S. and Suthaharan, S., 2016. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp.207-235.

Biau, G. and Scornet, E., 2016. A random forest guided tour. *Test*, *25*, pp.197-227.

Singh, A., Thakur, N. and Sharma, A., 2016, March. A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom)* (pp. 1310-1315). Ieee.

Budiman, F., 2019. SVM-RBF parameters testing optimization using cross validation and grid search to improve multiclass classification. *Научная визуализация*, *11*(1), pp.80-90.

Laxmanan, S., Awareness of Credit Score Mechanism in India: A Study with reference to Credit Information Bureau India Limited (CIBIL).

**Submitted by: Udit Rawat**

**Email: uditrawat1905@gmail.com**