

LOAN APPLICATION STATUS PREDICTION MODEL



Submitted by:
UJJWAL PRATIK

Contents

ACKNOWLEDGMENT	3
INTRODUCTION	4
ABOUT LOAN.....	4
PROBLEM STATEMENT.....	4
DATA ANALYSIS	5
COLUMNS IN THE DATAFRAME.....	5
sample.....	5
OBJECTIVE	5
DATA DESCRIPTION.....	5
METHODOLOGY	6
METRIC USAGE	6
SYSTEM REQUIREMENTS	7
APPROACH	8
IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES.....	8
TESTING OF IDENTIFIED APPROACH(Algorithms)	9
KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION	10
CONCLUSION	11
LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE	11
REFERENCES	12

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT LOAN

The term loan refers to a type of credit vehicle in which a sum of money is lent to another party in exchange for future repayment of the value or principal amount. In many cases, the lender also adds interest and/or finance charges to the principal value which the borrower must repay in addition to the principal balance. Loans may be for a specific, one-time amount, or they may be available as an open-ended line of credit up to a specified limit.

PROBLEM STATEMENT

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc. Load Application Status Prediction is a task that can be done based on historical information of the customer and bank. By checking the dataset already existed regarding the status of the Load Application and creating a model will help us to Predict the further Loan Application Status.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- Loan ID
- Gender
- Married
- Dependents
- Education
- Self-employed
- Applicant Income
- Co applicant Income
- Loan Amount
- Loan Amount Term
- Credit History
- Property Area
- Loan Status (Target Variable)

sample

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0	Rural
Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0	Rural
Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0	Urban
Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0	Urban
Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0	Semiurban

OBJECTIVE

- Our main objective is 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.
- All the parameters will be analysed through Machine Learning algorithms like Logistic Regression, AdaBoost Classifier, Random Forest Classifier, Decision Tree Classifier, Gaussian NB, Support Vector Classifier etc which will help to predict the Loan Status.

DATA DESCRIPTION

- The source of data is taken from GitHub.
- Data Link is (https://github.com/dsrscientist/DSDData/blob/master/loan_prediction.csv)

METHODOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learning techniques to determine the loan status because it gives the best outcome.
- The independent variable is Gender, Married, Dependents, Education, Self Employed, Applicant Income, Co Applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, and dependent Variable is Loan Status.

METRIC USAGE

- a. Logistic Regression.
- b. AdaBoost Classifier.
- c. Random Forest Classifier.
- d. Gaussian NB
- e. Decision Tree Classifier.
- f. Support Vector Classifier

SYSTEM REQUIREMENTS

HARDWARE and SOFTWARE REQUIREMENTS and TOOLS USED

- a) Hardware Requirement:
 - i. Intel core i5
 - ii. 8 GB Ram
- b) Software Requirement:
 - i. Python 3.x with packages:
 - 1. Pandas: Data analysis and manipulation tool
 - 2. NumPy: Provide support for mathematical functions, random number etc.
 - 3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
 - 4. Seaborn: is a library mostly used for statistical plotting in python.
 - 5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- **Logistic Regression:** Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- **AdaBoost Classifier:** s a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
- **Random Forest Classifier:** The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.
- **Gaussian NB:** Gaussian NB is based on the Naive Bayes theorem with the assumption of conditional independence between every pair of features given the label of the target class.
- **Decision Tree Classifier:** A decision tree is a flowchart-like tree structure in which the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. A Decision Tree consists of Nodes: Test for the value of a certain attribute.
- **Support Vector Classifier:** Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. In addition to this, an SVM can also perform non-linear classification.
- **Cross-Validation-Score:** a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- **Grid Search CV:** This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- **ZScore:** Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- **Label Encoder:** Label Encoding refers to converting the labels into numeric form.
- **IMB Learn Anaconda:** imbalanced-learn is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. It is compatible with scikit-learn and is part of scikit-learn-contra projects.
- **AUC-ROC CURVE:** The AUC-ROC curve tells us visualize how well our machine learning classifier is carrying out. It is one of the most significant evaluation metrics for examining any classification model's performance. It is also called as AUROC (Area Under the Receiver Operating Characteristics).

TESTING OF IDENTIFIED APPROACH(Algorithms)

- a. Train Test Split
- b. Label Encoding
- c. IMB Learn
- d. Logistic Regression
- e. AdaBoost Classifier
- f. Gaussian NB
- g. Decision Tree classifier
- h. Support Vector Classifier
- i. Grid Search CV
- j. Cross Validation
- k. AUC ROC CURVE

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Compared between two columns.
- Checked and removed outliers through zscore method.
- Removed skewness present in the dataset.
- Done Oversampling.
- Cross validate the accuracy score from overfitting.
- Done AUC ROC score for better understanding.
- Hyper Parameter tuning.

CONCLUSION

As our conclusion we proclaim that, after checking accuracy score, cross validation, Ensemble Techniques, and checking AUC score we declare Random Forest Classifier is predicting approx. 80% accuracy is best suited model for our purpose of predicting Loan Approval status.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.