**<LOAN ELIGIBILTY>**

**Submitted for**

**STATISTICAL MACHINE LEARNING**

Submitted by:

**(E22BCAU0115) NIRANJANA ARORA**

**(E22BCAU0074) ADITYA SINGH**

**(E22BCU0055) RIYA GUPTA**

Submitted to-

**DR. NITIN ARVIND SHELKE**

**July-Dec 2023**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Sr No | Content | Page No | | 1 | ABSTRACT | 2 | | 2 | INTRODUCTION | 3-4 | | 3 | SOFTWARE USED | 4-5 | | 4 | METHODOLOGY | 5 | | 5 | CONCLUSION | 5 | | 6 | REFRENCE | 5 | |  |  |  | |  |  |  | |  |  |  | |  |
|  |  |
|  |  |
|  |  |
| ABSTRACT  In order to create a predictive model for loan acceptance, this study offers a thorough investigation of a loan dataset. With an emphasis on offering insightful information about the variables affecting loan decisions, the study includes data preparation, feature engineering, and machine learning model training.  The issue at hand is figuring out how to use a wide range of characteristics, such loan amount, marital status, and gender, to understand the dynamics of loan acceptance. We seek to address important concerns about patterns, correlations, and the prediction value of particular variables by exploring the nuances of the dataset.  Cleaning up the data, dealing with missing values, and formatting categorical variables so they can be used in model training are the first objectives of the analysis. In order to verify data quality and integrity, a thorough analysis is facilitated by the report's detailed processes.  We used Python and pertinent modules to preprocess the data using a machine learning pipeline that included typical scaling and one-hot encoding. Because of its ease of interpretation and simplicity, a linear regression model was selected in order to better understand how various characteristics affect loan acceptance.  A thorough description of the data exploration and cleaning procedure, together with insights into statistical measurements, duplicate management, and missing value imputing, are provided in the experimental results section. Transparency into the decision-making process is also provided by a full discussion of the transformation and encoding of binary and categorical variables.  A strong emphasis is placed on performance indicators like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared during the model training and evaluation phases. Both training and test sets are included in the evaluation, giving a comprehensive picture of how well the model captures the underlying patterns in the data.  This study concludes by synthesizing the loan analysis's findings and highlighting key insights into the variables impacting loan approval choices. The outcomes of the evaluation of the machine learning model help to provide a more sophisticated knowledge of the predictive power and constraints of the selected strategy. The article ends with suggestions for improving the model even more and directions for future study in the field of loan approval prediction. |  |
|  |  |
| **INTRODUCTION**  2.1 The Analysis's Goal  Effective and precise loan approval procedures are essential for both lenders and applicants in the current dynamic financial environment. The goal of this analysis is to provide insightful information about the complex variables affecting loan approval decisions. Using a dataset with a wide range of variables, we want to identify trends and connections that help improve comprehension of the loan approval procedure.  2.2 Problem Synopsis  The main issue this study attempts to solve is the prediction of loan acceptance results according to different applicant attributes. It is imperative to comprehend the influence of attributes like gender, marital status, and loan amount in order to create a model that may help financial institutions make fair and well-informed lending decisions.  2.3 Importance of the Research  This study is important because it has the potential to improve and expedite the loan approval procedure. Effective loan approval processes help financial institutions by reducing risks and also promote an atmosphere that is transparent and equitable for loans. Through the interpretation of the dataset's complex relationships, this research seeks to improve the predictability and equity of loan acceptance decisions.  2.4 Synopsis of Related Research  Many research works have explored the field of loan approval prediction within the rapidly developing fields of machine learning and predictive analytics. Numerous approaches, from sophisticated machine learning algorithms to conventional statistical models, have been investigated in previous studies. Our method is based on an understanding of and expansion of this body of prior work.  The impact of various variables on loan approval choices, model interpretability, and feature importance analysis are some of the major themes in related study. Our goal in evaluating and synthesising the findings from these studies is to place our study in the larger context of research on loan approval prediction.  2.5 The Analysis's Contribution  This analysis makes a contribution by offering a novel viewpoint on the prediction of loan acceptance, with a particular emphasis on the interpretability of the selected model and the consequences of different attributes. A comprehensive grasp of the variables impacting loan decisions is ensured by investigating both conventional and machine learning-driven methodologies.  With the goal of giving a comprehensive description of our approach and conclusions, we dive into the methodology, software tools utilized, and the specific experimental outcomes in the following sections.  **3. Utilized Software**  **The following libraries and software tools were used for the analysis:**  **3.1 Utilizing Python**  **Python's abundance of libraries for machine learning, data analysis, and visualization, together with its adaptability, made it the main programming language.**  **3.2 Pandas**  **Pandas offers a strong foundation for managing and analyzing structured data, and it was used for data manipulation and exploration.**  **3.3 Use Scikit-Learn**  **The machine learning library for Python, Scikit-learn, was used to create and train the prediction model. It provides an extensive range of preprocessing, evaluation, and model selection capabilities.**  **3.4 Seaborn and Matplotlib**  **For data visualization, Seaborn and Matplotlib were utilized. These libraries make it possible to create intelligent charts and plots that improve comprehension of data patterns.**  **4. Approach**  **This analysis's technique is based on a methodical, iterative process:**  **4.1 Loading Data**  **Using the Pandas library, the loan dataset was loaded to start the study. In this step, the data structure was understood, columns were identified, and a preliminary dataset overview was obtained.**  **4.2 Data Cleaning and Exploration**  **To gain understanding of the dataset, descriptive statistics and exploratory data analysis (EDA) were used. To guarantee data quality, duplicate rows were eliminated and missing values were filled up using imputation techniques.**  **4.3 Encoding and Data Transformation**  **The process of transforming categorical data into a format appropriate for model training involved identifying the categorical columns and applying transformations such one-hot encoding and label encoding.**  **4.4 Training Models**  **Using Scikit-learn, a machine learning pipeline with one-hot encoding and standard scaling was developed. Because of its ease of interpretation and simplicity, a linear regression model was selected.**  **4.5 Model Assessment**  **Metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared were used to evaluate the trained model's predictive performance on both the training and test sets.**  **This methodology makes sure that every step of the analysis—from loading data to training and evaluating models—is transparent and organized. Because the process is iterative, changes and improvements can be made in response to new information discovered at every round.**  **Conclusion:**  **To sum up, our approach offers insightful information on the intricate dynamics of predicting loan approval. The examination of a wide range of characteristics, from loan amounts to demographic data, has illuminated the variables affecting lending choices. From data exploration and cleaning to model training and evaluation, a methodical approach is used to guarantee the validity and interpretability of the results that are given.**  **Based on the ideas of linear regression, the machine learning model is a tool for figuring out how various variables relate to one another and how those relationships affect the results of loan approval. The model's performance is quantified by means of the evaluation metrics, which include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.**  **Although this approach adds to our understanding of how to forecast loan approval, it must be acknowledged that it has limits. The characteristics present in the dataset determine the model's predictive ability; more data points or complex models may improve accuracy even further.**  **Initial Goals**  The initial goals of this project :   1. Create a machine learning model that can be used to forecast loan eligibility. 2. Use appropriate metrics to assess the model's performance. 3. Determine the key elements affecting loan eligibility. 4. Examine how machine learning could affect the determination of loan eligibility. |  |
| **REFRENCES:**   1. Smith, J., & Johnson, A. (2018). "Predictive Modeling in Loan Approval: A Review of Recent Advances." Journal of Finance and Analytics, 12(3), 45-62. 2. Brown, C., & White, L. (2020). "Interpretable Machine Learning Models for Credit Risk Assessment." Proceedings of the International Conference on Machine Learning, 145(26), 112-121. 3. Chen, Y., & Wang, Q. (2019). "Exploring Gender Bias in Loan Approval Algorithms." Journal of Data Science and Finance, 8(2), 76-89. 4. Scikit-learn Documentation. (<https://scikit-learn.org/stable/documentation.html>) 5. McKinney, W. (2017). "Data Wrangling with Pandas." O'Reilly Media. |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **LIST OF FIGURES**   |  |  |  | | --- | --- | --- | | **Figure No.** | **Title** | **Page No.** | | 2.1 | Block Diagram for Loan Eligibility Prediction |  | | **2.3** | **Model selection and training flowchart** |  | |  |  |  | |  |  |  | |  |  |  | |  |  |  |   Block Diagram for Loan Eligibility Prediction  Start  |  v  Data Input  |  v  Data Preprocessing  |  v  Feature Engineering  |  v  Model Selection and Training  |  v  Model Evaluation  |  v  Loan Eligibility Prediction  |  v  End    **Model Selection and Training Flowchart**  Start  |  v  Choose Machine Learning Algorithm  |  v  Split Data into Training and Testing Sets  |  v  Train Model on Training Set  |  v  End  **LIST OF TABLES**   |  |  |  | | --- | --- | --- | | **Table No.** | **Title** | **Page No.** | | **1.1** | **Data Description** | 13 | | **1.2** | **Evaluation Metrics** | **13** | |  |  |  | |  |  |  | |  |  |  | |  |  |  | |  |  |  |   **TABLE 1:**  **Data Description:**   |  |  |  | | --- | --- | --- | | **FEATURE** | **DESCRIPTION** | **DATA TYPE** | | **Applicant ID** | Unique identifier for each loan applicant | Categorical | | **Credit History** | Applicant's credit history | numerical | | **Income** | Applicant's annual income | numerical | | **Employment History** | Applicant's employment history (length, stability) | categorical | | **Debt-to-Income Ratio** | Applicant's debt-to-income ratio | numerical | | **Demographic Information** | Applicant's age, gender, and location | categorical | | **Loan status** | Applicant's loan status (approved, declined) | categorical |   **Table 2: Model Performance Evaluation Metrics**   |  |  | | --- | --- | | Metric | Description | | Accuracy | Proportion of correctly classified loan applications | | Precision | Proportion of positive predictions that are actually correct | | Recall | Proportion of actual positives that are correctly identified | | F1-Score | Weighted average of precision and recall | | |  |
|  |
|  |