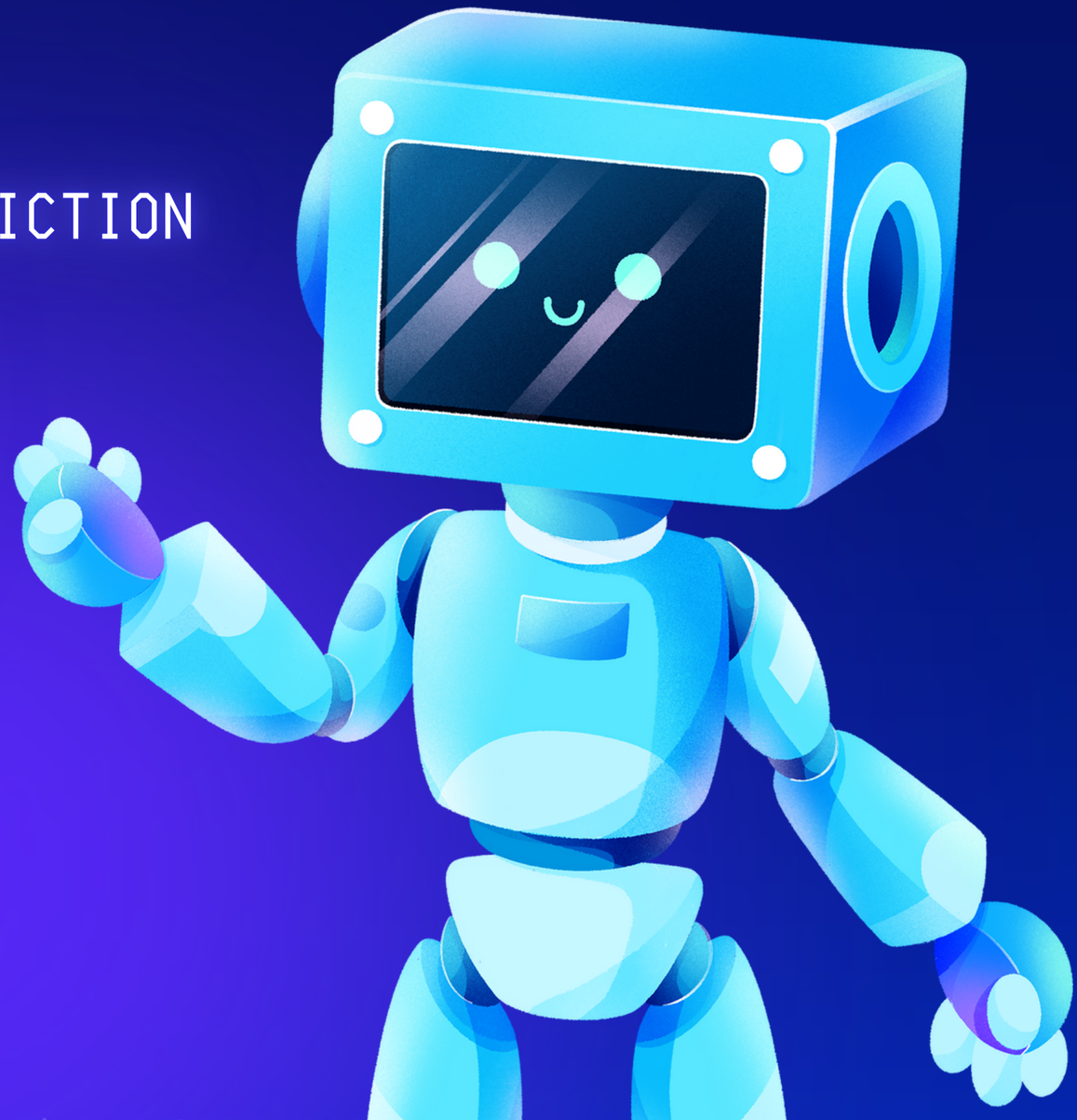
AUTOMATED LOAN ELIGIBILITY PREDICTION

PROJECT

By Oscar Mulei



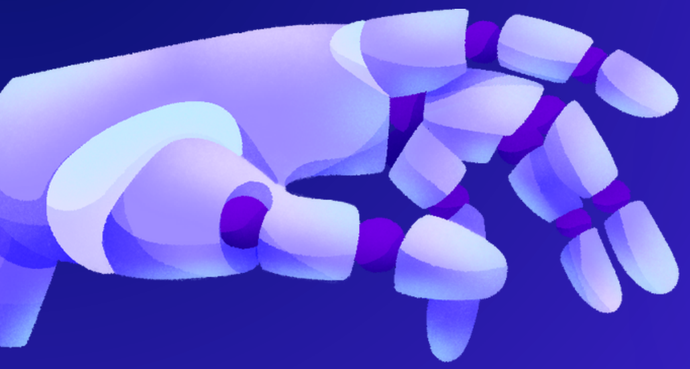
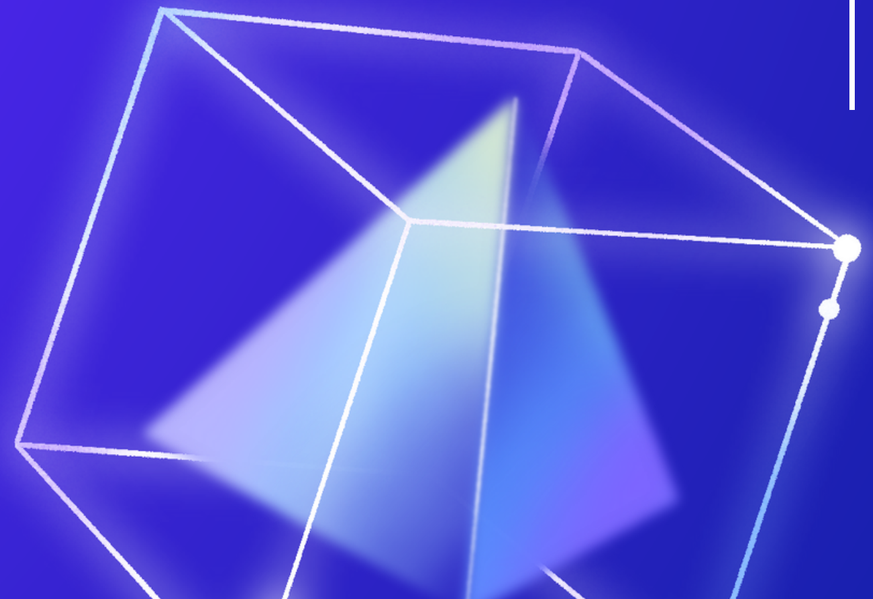


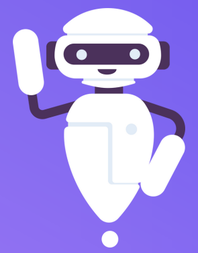
TABLE OF CONTENTS



INTRODUCTION

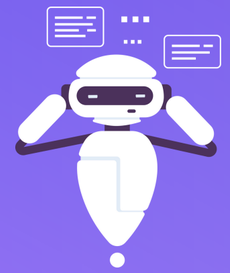
Nairobi Housing Finance Company offers all types of housing loans. They are present in all urban, semi-urban, and rural locations. Customers apply for a house loan once the company verifies their loan eligibility. The organization wants to automate the loan eligibility process (in real time) based on the information provided by the consumer when filling out the online application form. These characteristics include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. To automate this process, they created a problem to identify client segments that are eligible for loan amounts so that they may directly target these customers.

PROJECT OBJECTIVES



IMPROVE EFFICIENCY AND CUSTOMER EXPERIENCE

ENHANCE RISK MANAGEMENT



OPTIMIZE OPERATIONAL COSTS

MACHINE LEARNING

Automation

Machine learning automation can significantly benefit businesses and provide strong reasons for securing funding by addressing various critical aspects of operations and decision-making. Here are five strong business points for funding related to machine learning automation:

1. Consistency and Accuracy
2. Scalability
3. Speed and Responsiveness
4. Risk Management

These factors underscore why automating the loan eligibility prediction process is a critical investment, as it ensures consistent, scalable, and responsive operations while effectively managing risk.

AUTOMATION PROCESS

01

Data Collection and Preprocessing

02

Model Building and Training.

03

Automation

04

Deployment and Monitoring

DATA COLLECTION AND PREPROCESSING

SUMMARY FOR PRESENTATION: EXPLORATORY DATA ANALYSIS (EDA)

System Setup and Data Loading:

- Utilized Python and libraries like Pandas, Numpy, Seaborn, and Matplotlib.
- Loaded data from two CSV files: "train" and "test."
- Safeguarded the original datasets by creating copies.

Understanding the Data:

- Training dataset: 12 independent variables and 1 target variable (Loan_Status).
- Test dataset: Comparable features, excluding Loan_Status.
- Data types included object (categorical), int64 (integer), and float64 (numerical).
- Categorical variables: Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status.
- Numerical variables: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, and Credit_History.
- Training dataset: 614 rows and 13 columns; Test dataset: 367 rows and 12 columns.

SUMMARY FOR PRESENTATION: EXPLORATORY DATA ANALYSIS (EDA)

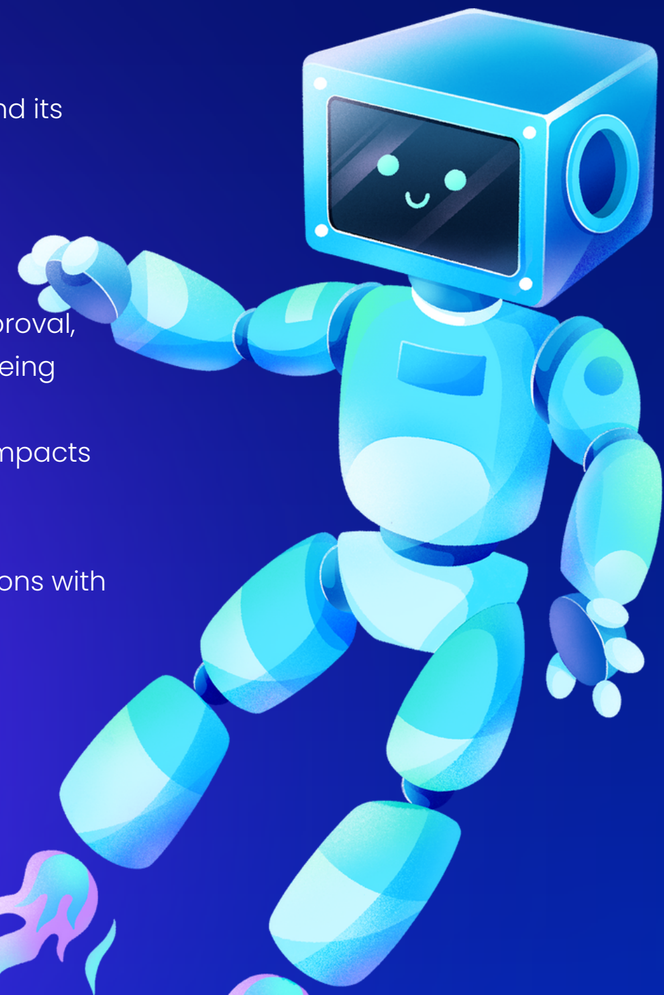
Univariate Analysis:

- Individual examination of variables.
- Analysis of the target variable, Loan_Status, using frequency tables, percentage distribution, and bar plots.
- Visualization of categorical, ordinal, and numerical features.
- Notable Observations:
  - Approximately 69% of applicants were approved for loans.
  - Categorical features (e.g., Gender, Married, Self_Employed, Credit_History) exhibited varying proportions.
  - Dependents were mostly '0,' 'Graduate' was the predominant Education level, and 'Semiurban' dominated the Property_Area.

SUMMARY FOR PRESENTATION: EXPLORATORY DATA ANALYSIS (EDA)

Bivariate Analysis:

- Testing hypotheses related to loan approval and its relationship with categorical variables.
- Visualizing the proportions of approved and unapproved loans using stacked bar plots.
- Key Findings:
  - Marital status appeared to influence loan approval, with a higher proportion of married applicants being approved.
  - Other categorical variables showed varying impacts on loan approval.
- Numerical variables, like ApplicantIncome and LoanAmount, did not display significant correlations with loan approval.

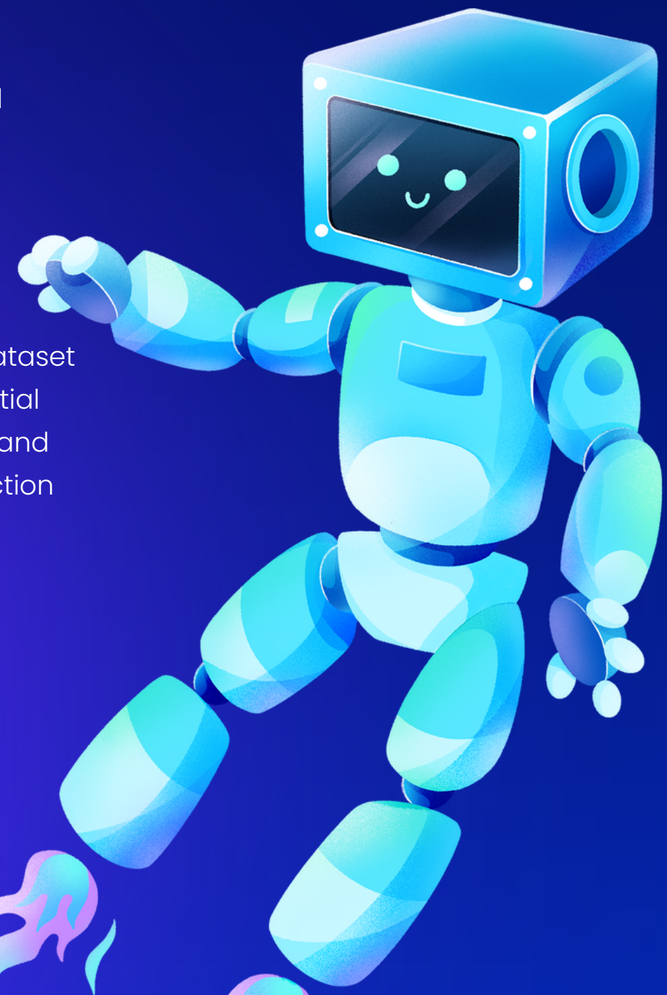


SUMMARY FOR PRESENTATION: EXPLORATORY DATA ANALYSIS (EDA)

Correlation Analysis:

- Examination of correlations between numerical variables using a heatmap.
- Strong positive correlations observed between ApplicantIncome and LoanAmount, as well as Credit_History and Loan_Status.

This EDA process provided critical insights into dataset characteristics, variable relationships, and potential factors affecting loan approval. Further analysis and preprocessing will be necessary for feature selection and model development in the Automated Loan Eligibility Prediction project.



MODEL BUILDING AND TRAINING.

MODELS

1. **Logistic Regression with Stratified k-folds Cross Validation:** A simple model for loan eligibility with improved reliability through data partitioning.

2. **Decision Tree:** Provides a transparent decision-making process based on input features, especially suitable for categorical data.

3. **Random Forest:** An ensemble of decision trees for enhanced accuracy and robustness in predictions.

4. **XGBoost**: A powerful algorithm for complex offering high accuracy, feature importance, and regularization.

RESULTS AND ACHIEVEMENTS

01

Logistic Regression:

- Mean Validation Accuracy: 0.8013
- Performance: Highest accuracy among the models

02

Decision Tree:

- Mean Validation Accuracy: 0.7149
- Performance: Lower accuracy compared to other models

03

Random Forest:

- Mean Validation Accuracy: 0.7947
- Performance: Good accuracy

04

XGBoost:

- Mean Validation Accuracy: Approximately 0.7752
- Performance: Moderate accuracy

THANK YOU!