LOAN APPROVAL PREDICTION MODEL USING PYTHON REGRESSION FOR A MID-SIZED LENDING FIRM

1. Introduction

A mid-sized consumer lending company faced inconsistencies and delays in evaluating loan applications. Their manual process relied heavily on underwriter judgment, leading to inefficiencies and applicant dissatisfaction. The company approached us to build a Python-powered regression model that could predict the probability of loan approval using applicant data.

2. Objective

- To quantify how key financial and demographic indicators influence loan approvals
- To develop a multiple linear regression model in Python with dummy and interaction variables
- To deliver a backend-ready model and interpretable visual report for internal use

3. Data Provided by Client

The client shared a dataset of 1,000 historical loan applications with the following fields:

- Loan_Status (approved or rejected)
- Applicant_Income
- Loan_Amount
- Credit_Score
- Education_Level (Graduate / Not Graduate)
- Employment_Type (Salaried / Self-Employed)

The goal was to model Loan_Status as a probability using the other fields as predictors. No personally identifiable information (PII) was shared.

4. Methodology

4.1 Data Preparation

- Cleaned missing values
- Created dummy variables for education and employment type
- Built interaction terms (e.g., Income × Education) to reflect non-linear effects
- Performed outlier detection using Z-scores and removed 2% high-leverage records

4.2 Model Development

- Used Statsmodels to run a multiple linear regression
- Checked for multicollinearity using VIF
- Validated assumptions of linearity, normality, and homoscedasticity

4.3 Output Delivered

- Python .ipynb file with all steps documented
- Visual report in PDF showing coefficients, confidence intervals, and key charts
- Excel file with model predictions for the full dataset
- Recommendations on how to integrate this into their loan workflow

5. Key Results

- Model R-squared: **0.62**, Adjusted R-squared: **0.60**
- **Credit Score** and **Income** had the strongest influence on approval probability
- The interaction term showed income mattered more for non-graduates than graduates
- The model helped rank new applicants by likelihood of approval

6. Business Impact

- Manual review time reduced by **40%**
- Underwriting team used model outputs to pre-screen applications
- Helped justify approval/rejection decisions with numeric evidence
- Improved customer transparency during application follow-ups

7. Future Scope Suggested

- Transition to a logistic regression model for binary classification

- API-based integration into their CRM system for real-time predictions
- Expand model to include new variables such as marital status and number of dependents
- Ongoing retraining every 6 months to maintain model accuracy with new applicant trends