

Loan Risk Prediction Report Summary

Objective: The goal of this project was to build a machine learning model to predict whether a client is high risk or low risk for loan approval based on financial data. The target variable, Risk_Flag, is 1 if the client is high risk and 0 if the client is low risk.

Data Exploration and Visualization:

- The income distribution showed a wide range of values with a right skew.
- The target variable Risk_Flag showed a significant imbalance, with more low-risk clients than high-risk clients.
- Various categorical features such as marital status, house ownership, car ownership, profession, city, and state were analyzed to understand their distributions.

Feature Engineering:

- Missing values were handled, and categorical features were encoded using label encoding.
- Numerical features were scaled using standard scaling to bring them to a comparable range.

Model Building and Evaluation:

Five models were trained and evaluated:

1. Logistic Regression
2. XGBoost Classifier
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. AdaBoost Classifier

The Random Forest Classifier showed the best performance:

- Accuracy: 0.95
- Precision: 0.60
- Recall: 0.55
- F1-Score: 0.57

Confusion Matrix Breakdown:

- **True Positives (TP):** High-risk clients correctly identified as high-risk.
- **False Positives (FP):** Low-risk clients incorrectly identified as high-risk.
- **False Negatives (FN):** High-risk clients incorrectly identified as low-risk.
- **True Negatives (TN):** Low-risk clients correctly identified as low-risk.

Important Factors:

- The feature importance analysis from the Random Forest model indicated that certain features had a higher impact on the prediction of loan risk.

Conclusion:-

The Random Forest Classifier was selected as the final model due to its superior performance. The report includes detailed visualizations, model performance metrics, and feature importance insights.

The Solution link is :-