

LoanTap-Credit Risk Analysis & Predictive Modeling Report

Prepared By:

Manish Nandi, Data Analyst

Date: 15/11/2025

Project Title:

Credit Risk Modelling Using Logistic Regression & LightGBM

1. Executive Summary

This report presents a data-driven credit-risk evaluation framework developed using Logistic Regression and LightGBM models. The primary objective is to predict high-risk loan applicants and provide actionable insights to reduce default rates while optimizing loan approvals.

The LightGBM model demonstrated strong recall performance (**85% at a 0.63 threshold**), enabling early identification of risky borrowers and supporting informed lending decisions.

2. Business Problem

LoanTap aims to **minimize loan defaults** while maintaining healthy loan disbursement volumes.

Key challenges include:

- Identifying customers likely to default before approval.
- Balancing risk control with revenue generation.
- Reducing operational costs from manual underwriting.

To address these challenges, a predictive model was built to **classify applicants as high-risk or low-risk** based on historical loan performance.

3. Data Overview

- Total observations: **~400k rows**
 - Features included: demographic details, financial metrics, loan history, credit behavior.
 - Target label:
 - **1 = Good borrower**
 - **0 = Defaulter**
-

4. Data Pre-processing & Feature Engineering

Cleaning

- Removed missing or irrelevant values
- Encoded categorical variables
- Corrected inconsistent data formats

Feature Engineering

Combine 3-4 features to catch variances of each feature in one column

- $\text{loan_to_income_ratio} = \text{loan_amnt} / \text{annual_inc} * \text{installment} / \text{int_rate}$
- $\text{dti/revol_bal} = \text{dti} / (\text{revol_bal} * \text{grade})$
- Normalized and transformed skewed financial fields
- Handled multicollinearity (checking VIF)
- Handle imbalance data using SMOTE in Logistic Regression

These engineered features helped capture borrower repayment stress more effectively.

5. Modeling Approach

Two models were implemented:

4.1 Logistic Regression

- Baseline interpretable model
- Useful for understanding directional impact of variables
- Achieved ROC-AUC: **69%**

4.2 LightGBM

- Gradient boosting decision tree model
 - Superior performance on non-linear & high-dimensional data
 - High recall with decent precision after threshold tuning
 - Achieved ROC-AUC: **70%**
 - Selected for final deployment
-

5. Threshold Tuning and Evaluation

At a tuned threshold of **0.63**, LightGBM delivered:

- **ROC-AUC:** 0.70
- **Accuracy :** 70%

- **Recall** : 85%
- **Precision** : 25%

Why this threshold?

A threshold of ~0.63 provides a **strong recall**, which is important because:

- Missing defaulters is **more costly** than incorrectly flagging safe customers.
 - High recall reduces *expected loss*.
 - Low precision is acceptable in lending because flagged customers can undergo extra verification rather than automatic rejection.
-

6. Key Drivers of Default (Feature Importance)

Typical drivers include:

- High loan amount
- Low income level
- High existing debt
- Longer tenure
- Delayed past payments

These drivers help shape credit policy rules.

7. Actionable Business Insights

1. High-Risk Borrowers Can Be Flagged Early

The model identifies **85% of defaulters**, helping avoid risky disbursements.

2. Loan Approval Process Can Be Tiered

Using prediction scores:

Risk Score	Action	Business Impact
High (≥ 0.63)	Manual verification / stricter rules	Reduced defaults
Medium	Check employment + income docs	Balanced risk & approval
Low	Fast-track approval,	Increased throughput

3. Pricing Strategy Optimization

High-risk borrowers can:

- Be charged higher interest
- Offered smaller loan amounts
- Given shorter repayment tenure

This improves risk-adjusted profitability.

4. Portfolio Risk Monitoring

The model can score existing customers weekly/monthly to detect potential defaults early.

5. Data Quality Improvements

Features like debt to income ratio, revolving balance, and mortgage account influence prediction heavily → LoanTap should ensure accurate collection of these values.

8. Business Impact

Using this model, LoanTap can expect:

- **Reduction in approval of high-risk loans**
- **Lower NPA / default percentage**
- **Higher profit through optimized pricing**
- **Improved compliance and underwriting standards**
- **Better customer targeting**

Even a **5–10% reduction in defaults** can translate into significant financial savings.

8. Limitations

- Model performance depends on data quality.
- Precision is low; business must handle false positives carefully.
- Needs periodic retraining for stability.

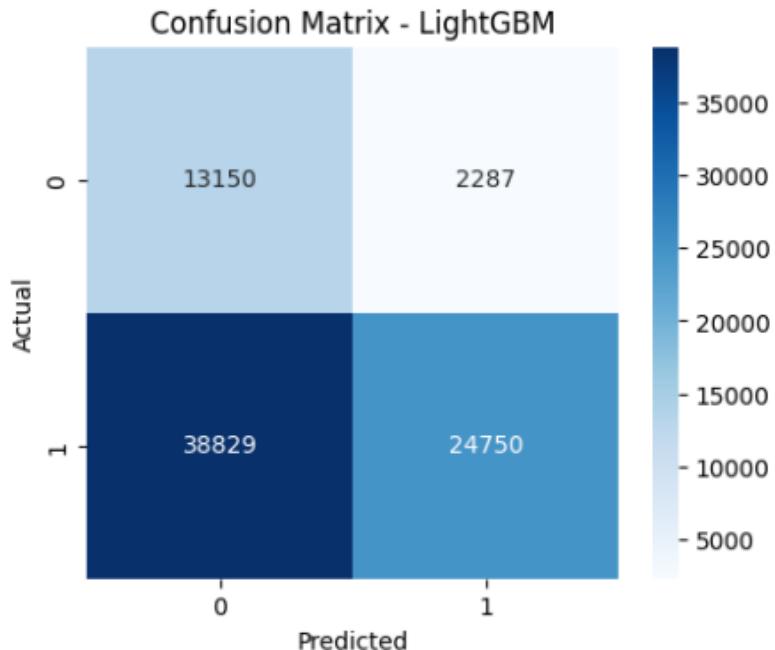
9. Conclusion

This project successfully built a predictive credit risk model using Logistic Regression and LightGBM. With a recall of ~85% at a business-optimized threshold and an ROC-AUC of 0.70, the model provides **reliable and actionable insights** to minimize financial risk.

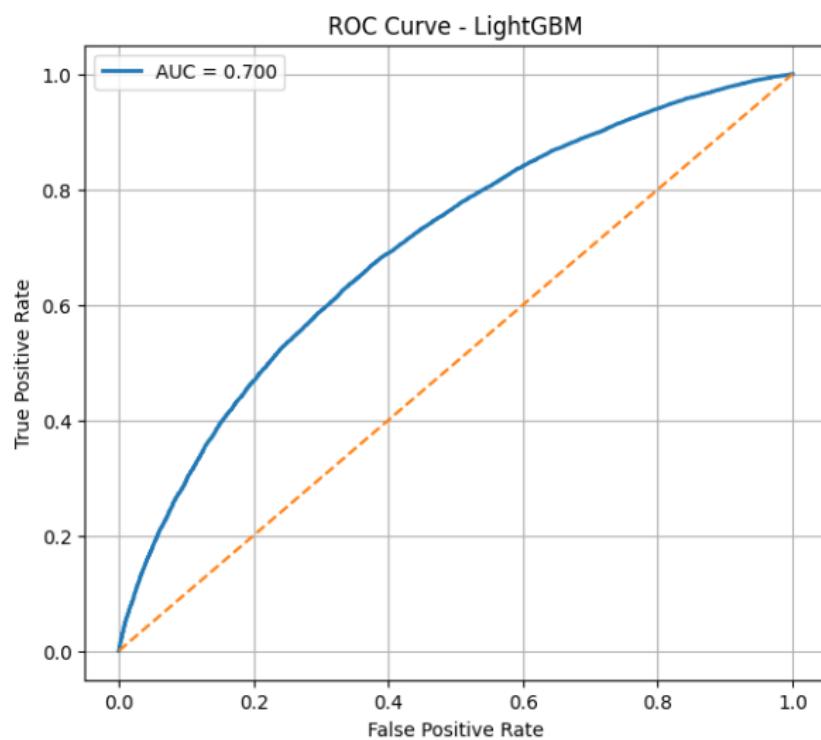
It is ready for integration into the underwriting workflow, scorecards, or risk-based pricing systems.

Appendix

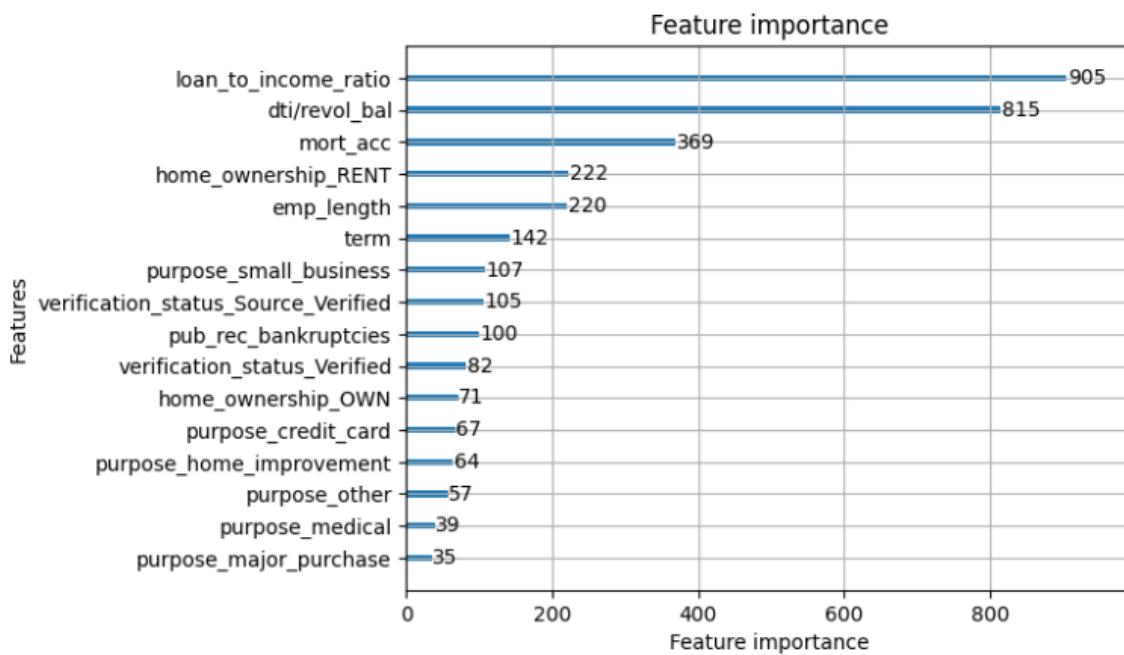
A. Confusion Matrix



B. ROC Curve



C. Feature Importance Plot



D. Classification report

```
Classification Report:  
precision    recall    f1-score   support  
  
      0       0.253     0.852     0.390     15437  
      1       0.915     0.389     0.546     63579  
  
accuracy                           0.480     79016  
macro avg       0.584     0.621     0.468     79016  
weighted avg     0.786     0.480     0.516     79016
```

ROC-AUC Score: 0.7002562046219567