**Loan Default Prediction**

**Dataset Description and Preprocessing Steps:**

The dataset used is derived from a LendingClub loan dataset, where a subset of features was selected to focus the model on the most impactful predictors of loan default. The target variable is binary whether a loan is "Fully Paid" or "Charged Off."

**Selected Features Include:**

* Numeric: annual\_inc, loan\_amnt, int\_rate, installment, dti, open\_acc, revol\_util
* Categorical: term, emp\_title, emp\_length, purpose, home\_ownership, verification\_status, application\_type, grade, sub\_grade

**Preprocessing Steps:**

* Filtered data to include only relevant loan statuses (Fully Paid, Charged Off).
* Handled categorical features using Label Encoding.
* Scaled numerical features using StanderedScaler.
* Handled class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) to balance the target classes.
* Split data into training and test sets using an 80/20 ratio.

**Models Implemented and Rationale:**

1. **LightGBM Classifier**

LightGBM was chosen for this loan default prediction task due to its speed, accuracy, and suitability for structured financial data. It outperforms many boosting algorithms by using a histogram-based, leaf-wise growth strategy, which leads to faster training and higher accuracy with less overfitting.

The model natively handles categorical features, reducing preprocessing effort and preserving data quality. It is also memory-efficient, making it ideal for large datasets and limited-resource environments.

Given the class imbalance in loan defaults, LightGBM’s ability to adjust class weights and integrate with techniques like SMOTE makes it particularly effective. Additionally, its built-in feature importance tools provide interpretability, helping to identify key risk factors driving loan default predictions.

**Key Insights and Visualizations:**

* **Loan Status Distribution**: Majority of loans are "Fully Paid", confirming class imbalance which was addressed using SMOTE.

A blue rectangle with white text

AI-generated content may be incorrect.

* **Feature Correlation**: High correlation observed between loan\_amnt, installment, and int\_rate, which are critical predictors of default.

A screenshot of a computer screen

AI-generated content may be incorrect.

* **Application Type Analysis**: Individuals (vs. joint applications) had a different default risk pattern.

A screenshot of a graph

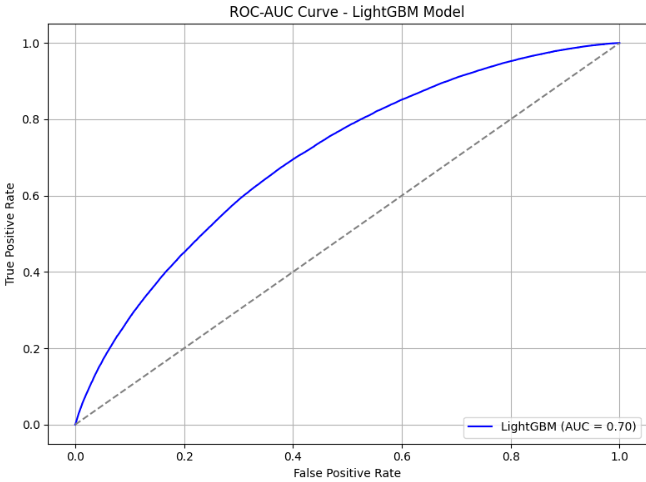
AI-generated content may be incorrect.

* **Important Features (via Gain Importance in LightGBM)**:
  + installment, int\_rate, and open\_acc contributed most to the prediction.

A graph with numbers and text

AI-generated content may be incorrect.

* ROC-AUC curves and Confusion Matrices for Light Bgm model



A diagram of a blue and white chart

AI-generated content may be incorrect.

**Challenges Faced and Solutions:**

* **Imbalanced Classes**: Original data was skewed toward "Fully Paid" loans.  
  Solution: Applied SMOTE to synthesize more "Charged Off" instances.
* **High Cardinality in Categorical Features**: Fields like emp\_title had too many unique values.  
  Solution: Retained only impactful features and used Label Encoding strategically.
* **Overfitting Risk:** LightGBM’s leaf-wise tree growth can lead to overfitting, especially on noisy or small datasets.  
  Solution: Used train-test split and ROC-AUC curves to evaluate model generalization. Applied LightGBM’s regularization parameters to control model complexity and reduce overfitting.
* **Feature Redundancy**: Features like loan\_amnt, installment, and int\_rate were closely correlated.  
  Solution: Used correlation heatmaps and feature importance scores to guide selection.