# Loan Default Prediction
# Model Documentation

## 1. Statement of Purpose:

This model predicts the likelihood of a Lending Club-approved loan being charged off (i.e., defaulting) versus being fully paid. The primary purpose is to provide third-party lenders with a more accurate and fair risk assessment tool than traditional credit scoring methods. The model aims to achieve high predictive accuracy while mitigating bias related to protected characteristics.

This model addresses the critical challenge of accurately assessing loan default risk for Lending Club-approved loans. Unlike traditional credit scoring methods, which may perpetuate bias and overlook key predictive signals, this model leverages advanced machine learning techniques to provide third-party lenders with a more precise and equitable risk assessment tool. The primary goal is to significantly enhance loan portfolio risk management by improving loan approval decisions, optimizing pricing strategies, and ensuring compliance with fair lending regulations. This enhanced risk assessment will enable lenders to reduce defaults, improve profitability, and gain a competitive advantage in the market. The model prioritizes high predictive accuracy while actively mitigating biases related to protected characteristics, ensuring fair and responsible lending practices. This solution directly addresses the problem of inaccurate creditworthiness assessment, which leads to increased loan defaults, higher operational costs, and reduced profitability for third-party lenders. The model offers a substantial value proposition by reducing default rates, improving profitability, enhancing compliance, and providing a competitive differentiation in the market, ultimately benefiting both lenders and borrowers.

**Business Use Case:** Enhancing Loan Portfolio Risk Management for Third-Party Lenders

This project provides a solution to improve loan portfolio risk management and reduce defaults for third-party lenders. A machine learning model accurately predicts the likelihood of loan default for Lending Club-approved loans, offering a superior risk assessment tool compared to traditional methods.

**Problem:** Inaccurate creditworthiness assessment leads to increased loan defaults, operational costs, and reduced profitability for third-party lenders. Traditional methods may perpetuate bias and overlook crucial information.

**Solution**: Our model uses Lending Club data and advanced machine learning techniques to predict loan defaults with high accuracy while mitigating demographic biases. This enables lenders to:

- Improve Loan Approval Decisions: Reduce defaults by approving only low-risk loans.
- Optimize Pricing Strategies: Adjust interest rates based on precise risk levels.
- Enhance Regulatory Compliance: Adhere to fair lending regulations.
- Gain a Competitive Advantage: Offer more responsible and accurate lending.
- Value Proposition: The model offers significant value through:
- Reduced Default Rates: Substantial cost savings.
- Improved Profitability: Optimized pricing and reduced losses.
- Enhanced Compliance: Mitigation of legal and reputational risks.
- Competitive Differentiation: Attraction of borrowers and investors.

**Target Audience:** Third-party lenders of all sizes seeking to improve lending processes, expand into new markets, and enhance regulatory compliance. The model provides a fairer and more accurate risk assessment than traditional methods.

## 2. Foundational Data

The foundation of this model rests on the "accepted_2007_to_2018Q3.csv.gz" dataset sourced from a public GitHub repository maintained by Fiddler Labs. This dataset contains a comprehensive history of accepted loans, encompassing a wide range of borrower characteristics and loan terms. The data's temporal range (2007-2018Q3) provides a substantial historical context. The size of the dataset, comprising millions of loan records, enables the training of robust and reliable predictive models. However, the dataset's scale necessitates a multi-stage cleaning and preprocessing pipeline to ensure data integrity and model efficiency.

The initial data cleaning involved focusing exclusively on loans with a "Fully Paid" or "Charged Off" status. This binary classification approach simplifies the predictive task and improves model accuracy by excluding uncertain or ambiguous cases (loans still in progress). A threshold of 50% missing data was established to remove features with extensive missingness, preventing the introduction of bias. Furthermore, features considered "cheat data" – those unavailable at the time of the loan application but only revealed afterward – were eliminated. This strategic data filtering process ensured that the model's predictions are solely based on information accessible at the loan application stage.

The preprocessed data underwent a 70/30 stratified split into training and testing subsets. Stratification based on the 'Charged_Off' outcome guarantees that the class distribution remains consistent across the train and test datasets, preventing potential bias and ensuring a more reliable model evaluation.

## 3. Features & Feature Engineering

The model employs a diverse set of features to predict loan defaults. This feature set includes both quantitative and qualitative data, reflecting various aspects of the borrower's financial profile and the loan's characteristics. Numerical features quantify aspects such as loan amount (loan_amnt)*(Fig 4)*, interest rate (int_rate), debt-to-income ratio (dti) *(Fig 3)*, credit score (fico_score), revolving balance (revol_bal), and the total number of credit accounts (total_acc). The inclusion of these metrics allows the model to capture the borrower's creditworthiness and financial standing.

Categorical features represent qualitative aspects of the loan and borrower, offering nuanced insights into default risk. These features encompass the loan grade (grade/sub_grade), home ownership status (home_ownership), verification status of income (verification_status), loan purpose (purpose), borrower's state (addr_state), initial listing status of the loan (initial_list_status), application type (application_type), and employment length (emp_length). One-hot encoding transformed these categorical features into a numerical format suitable for model training.

Several transformations were applied to enhance model performance.

- Numerical features (annual_inc, revol_bal) were log-transformed (using np.log10(x + 1)) to reduce skewness and improve model fit *(Fig 1 and Fig 2)*.
- Temporal features such as issue_d (loan issue date) and earliest_cr_line (earliest credit line) were converted to numerical representations of the years and months since these events occurred, providing valuable temporal context to the loan application process.

## 4. Models

Two distinct machine learning models were trained and evaluated for their ability to predict loan defaults:

- **XGBoost** (Extreme Gradient Boosting): This model is a gradient boosting algorithm known for its high predictive accuracy and efficiency. Its tree-based nature allows for handling both numerical and categorical features effectively. The model was configured with 100 trees (n_estimators=100) and no maximum depth limitation (max_depth=None) during the initial stage to maximize feature importance evaluation.

- **Random Forest**: A popular ensemble learning method, the Random Forest model combines multiple decision trees to make predictions. This approach reduces overfitting and improves model robustness. Similar to the XGBoost model, it used 100 trees (n_estimators=100).

**Model Performance Comparison:**

Both models achieved acceptable accuracy. However, the Random Forest model exhibited significant underfitting, demonstrated by a near-perfect training accuracy contrasting sharply with a much lower testing accuracy. The XGBoost model performed more robustly, with testing accuracy only slightly lower than training accuracy, indicating better generalization to unseen data. The XGBoost model displayed more balanced performance across the classes, indicating a superior ability to discern between 'Fully Paid' and 'Charged Off' loans."

# 5. Winning Model Selection & Reasoning

Given the substantial overfitting observed in the Random Forest model and the relatively balanced performance of the XGBoost model across training and testing sets, the ***XGBoost model is selected as the preferred model***. The superior generalization capability of XGBoost, evidenced by its more consistent performance on unseen data, makes it a more reliable and practical choice for real-world loan default prediction. The slight decrease in accuracy compared to Random Forest is far outweighed by the significantly improved robustness to overfitting. This choice also allows for exploring feature importance using XGBoost's built-in features, which can enhance model interpretability.

# 6. Interpretability/Explainability

Understanding the factors contributing to a model's predictions is crucial for building trust and facilitating responsible lending practices. For XGBoost, feature importance scores, calculated internally by the algorithm, quantify the relative contribution of each feature in the model's predictions. A bar plot (detailed in the code) visually ranks features based on their importance, allowing for immediate identification of the most influential factors. This analysis reveals that grade, term, issue_d_year, mort_accc and int_rate  are consistently ranked among the most influential features, providing actionable insights for lenders in assessing default risk.

While the current code doesn't generate Partial Dependence Plots (PDPs), these would provide even deeper insight. PDPs visually depict the marginal effect of each feature on the predicted probability of default, independent of the other features. For instance, a PDP for int_rate would precisely illustrate how the predicted default probability changes as the interest rate varies, holding other variables constant. PDPs would provide a more nuanced interpretation of individual feature effects. Analyzing these plots side-by-side could further clarify the predictive behavior of the model and its specific vulnerabilities.

# 7. Risk Management

The XGBoost model, despite being selected for its robustness, is still susceptible to several inherent risks:

- Model Drift : Our data does not take into account the recent data that includes the pandemic. The data is only till 2018. Pandemic has caused a lot of uncertainties in the market. This has the potential for making our model less reliable.

- XGBoost can be sensitive to noise or outliers in the dataset, as the boosting process may attempt to fit these aberrations.

To control these risks, multiple mitigations must be implemented:

- The data during and after the pandemic can tell us about these uncertain situations faced. Training the model with updated data from 2019 can potentially make the model take into account real-time data which will make the model perform better.
- Perform thorough data preprocessing, including outlier detection and removal, and apply feature scaling when necessary.
- XGBoost assumes that the training data distribution mirrors the future data it will encounter. Changes in the distribution of loan applications (e.g., shifts in borrower demographics or economic conditions) can lead to model degradation. Regular model retraining and monitoring are essential for maintaining accuracy over time.
- Regularization: Incorporate L1 or L2 regularization into the XGBoost training process, directly penalizing complex models and preventing overfitting.
- Robust Feature Engineering: Use domain expertise to carefully select, engineer, and transform features, reducing the influence of noise and bias.
- Continuous Model Monitoring: Regularly monitor the model's performance on a live data stream, tracking key metrics and retraining or updating the model as needed to account for model drift. Employ techniques like A/B testing to evaluate model performance before deploying changes to production systems.
- Explainable AI (XAI) Techniques: Integrate additional explainability techniques beyond feature importance, such as SHAP values, to comprehensively understand the model's decisions, promoting transparency and detecting potential bias.

# Appendix:

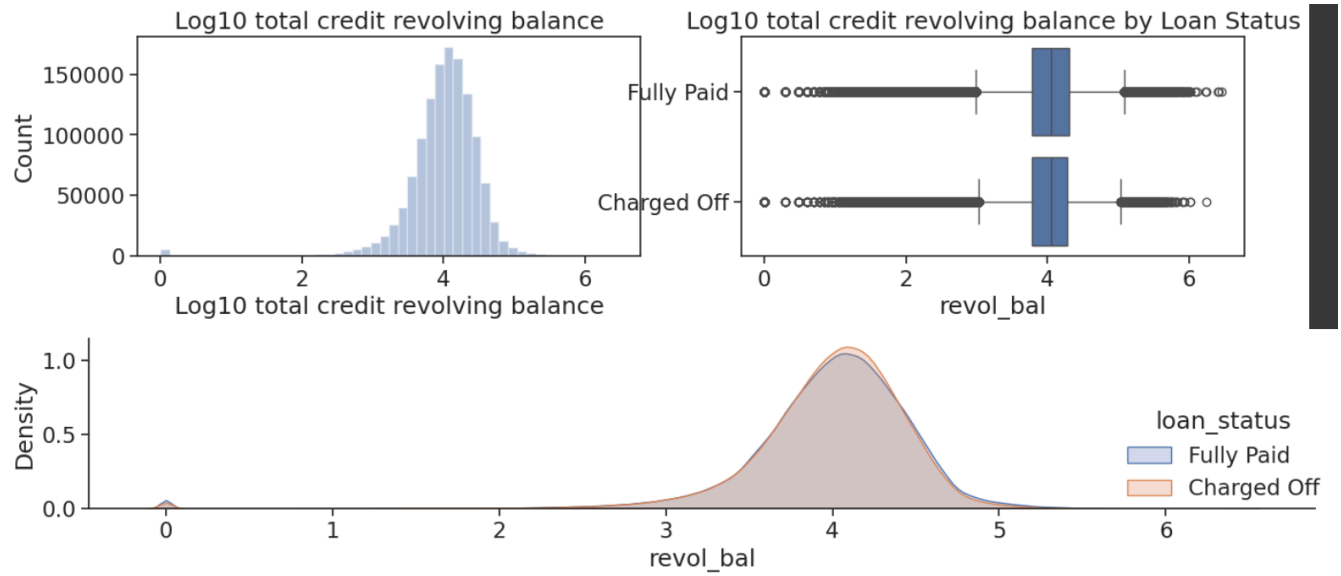


**Figure 1: Loan Status**

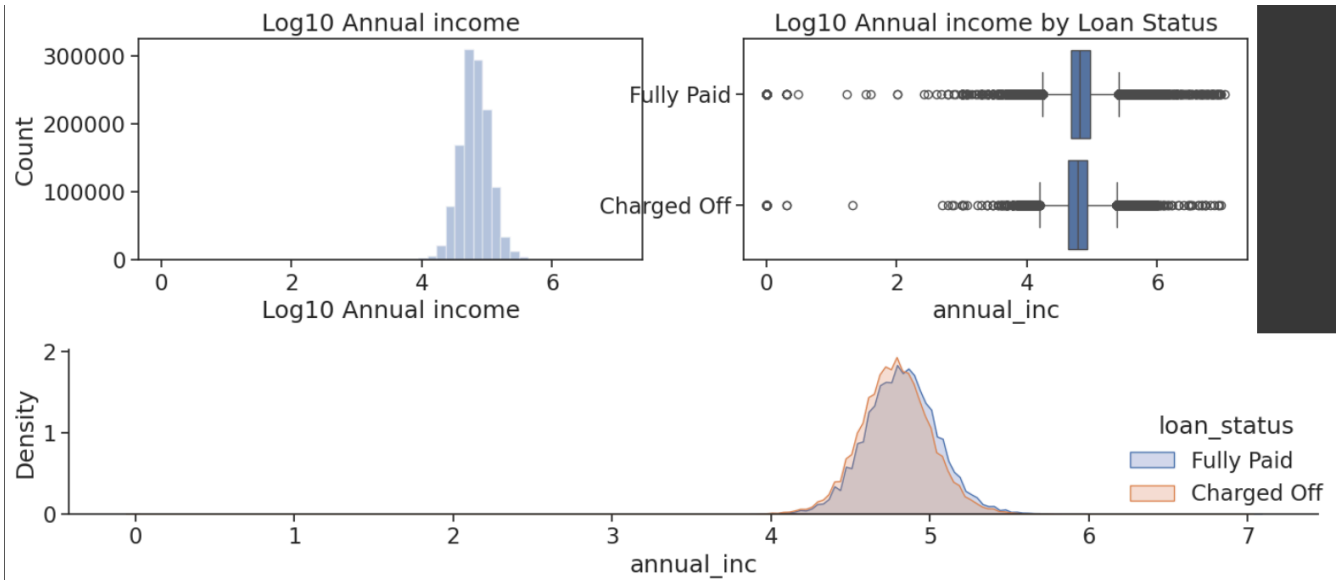


**Figure 2: Annual Income by Loan Status**

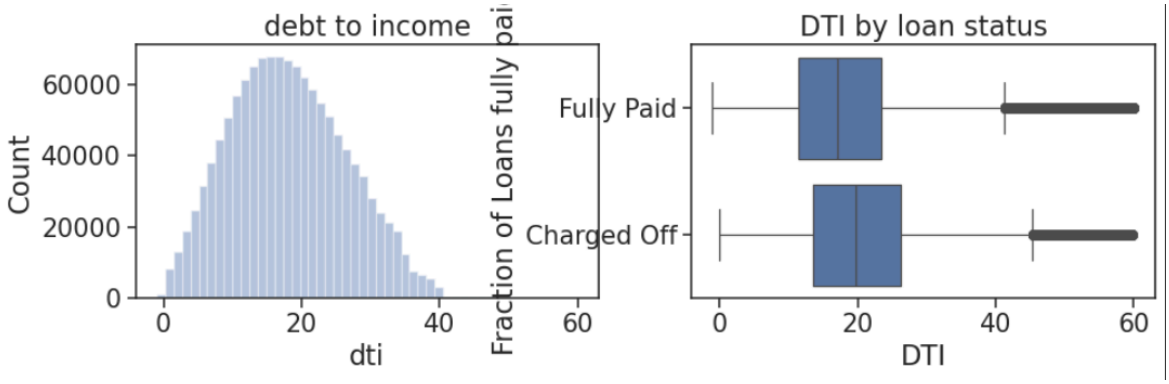


**Figure 3: Debt to Income & DTI by Loan Status**

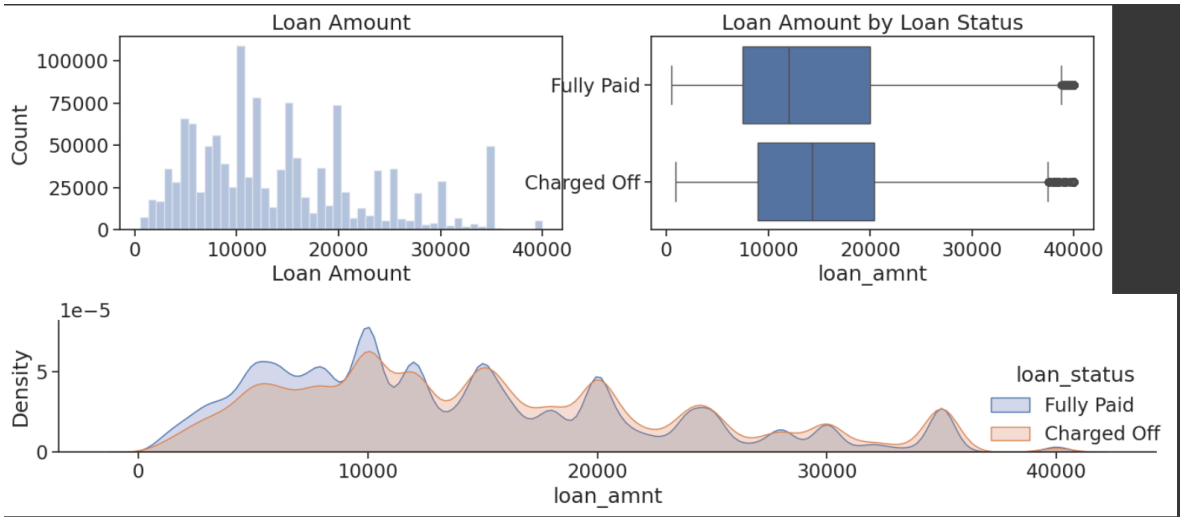


**Figure 4: Loan Amount by Loan Status**