Credit Risk and Loan Default Prediction

Executive Summary:
The goal of this project is to predict the probability of loan default using historical data from Lending Club. High-risk borrowers were identified using borrower characteristics, loan attributes, and credit profile features available at approval time. Careful feature engineering and missing value imputation was done to avoid data leakage and capture key characteristics of borrowers. A deep learning, feedforward neural network achieved key metrics that greatly outperformed baseline heuristics, correctly identifying over two-thirds of borrowers who actually defaulted (ROC-AUC: 0.717, recall: 0.680). The model showed little overfitting and high generalizability of unseen data. The model was interpreted using SHAP analysis, which showed that recent credit inquiries, loan sub-grade, and annual income were the strongest predictors of default risk.

Introduction:
Loan default prediction is a core problem in consumer credit risk management. Financial institutions must balance approval rates with default exposure, making accurate and interpretable risk models critical for decision-making. Traditional models such as logistic regression remain popular due to their interpretability, but modern machine learning models can capture nonlinear relationships and interaction effects that are difficult to model explicitly. This project frames loan default prediction as a binary classification task and explores the use of a feedforward neural network to model complex credit risk patterns.

The main objectives of this project are:

1. To correctly identify high-risk borrowers (i.e., maximize ROC-AUC and recall metrics).
2. To determine the strongest predictors of default risk.

Data Description:
The "Lending Club Loan Data" dataset was obtained from Kaggle. It contains data on over 2.2 million loans issued between 2007 and 2015, including the current loan status and latest payment information.. The dataset contains key attributes about the borrowers, such as credit scores, number of finance inquiries, address including zip codes and state, and collections, among other features.

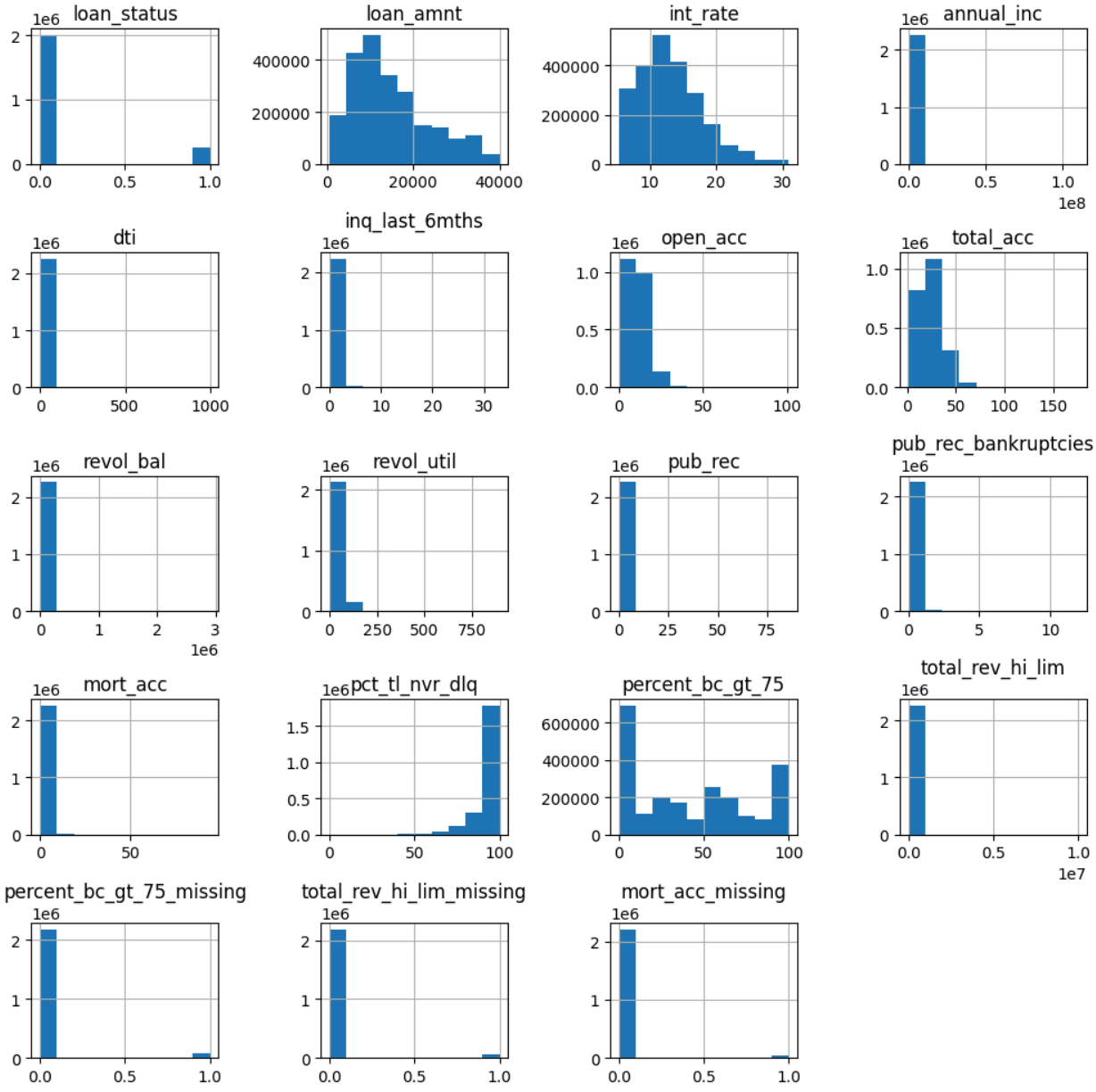
Only features available at loan origination were retained. Post-origination variables such as remaining principal, payment history, hardship indicators, and collection-related fields were excluded to prevent data leakage.

Methodology:

1. Data Preprocessing
    a. Missing values were handled using feature-specific strategies based on domain knowledge. Structural missingness (e.g., mortgage accounts) was imputed with zeros, while skewed financial variables were imputed using median or

distribution-aware approaches. Missing flags were added as features where deemed appropriate. Features with negligible missingness were handled via row removal.

b. Skewed continuous variables were transformed using the Yeo-Johnson power transformation to stabilize variance and improve neural network training. Data distributions before transformations are shown below.



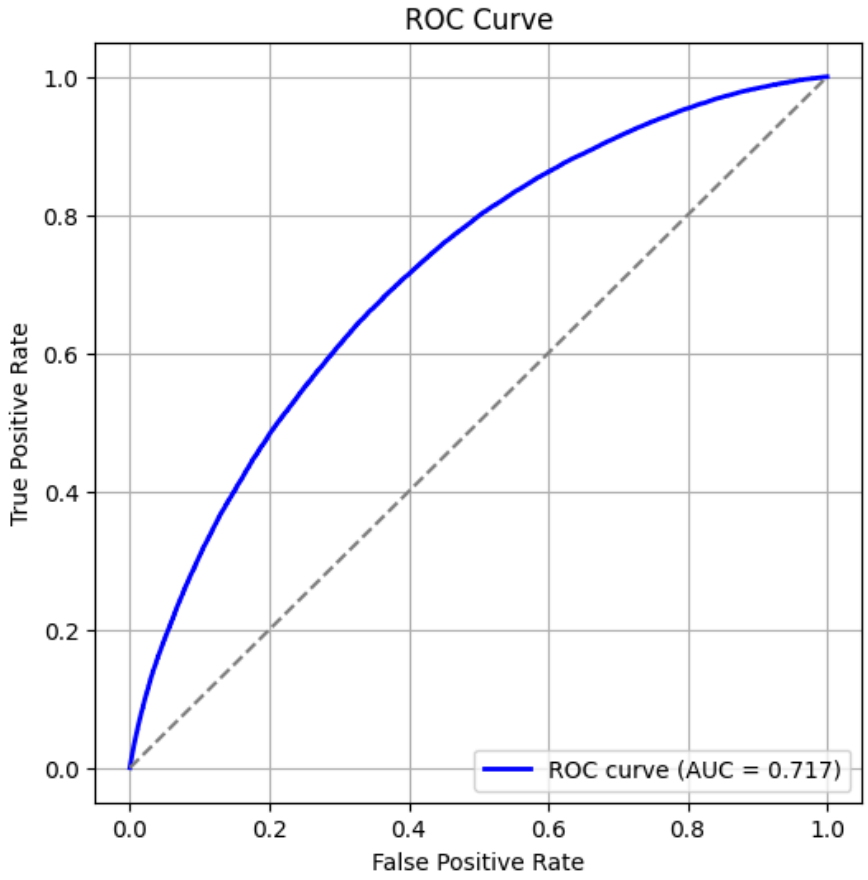
c. Continuous features were standardized and categorical variables were one-hot encoded.

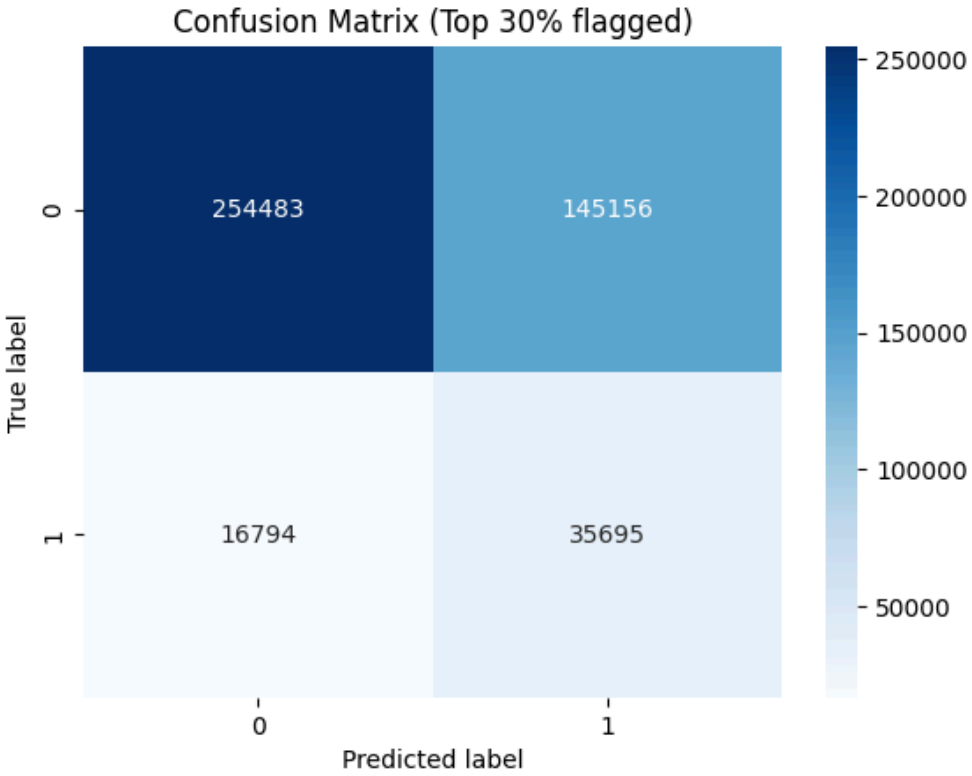
2. Model Building

a. A feedforward neural network was implemented using Keras with the following characteristics:
    i. Fully connected dense layers with ReLU activations
    ii. Binary cross-entropy loss function
    iii. Adam optimizer
    iv. Regularization via dropout
    v. Callbacks, including early stopping and a learning rate schedule
    vi. Evaluation using ROC-AUC and classification metrics
b. Neural networks were selected to capture nonlinear relationships and interaction effects common in credit risk data.

3. Model Interpretability
    a. SHAP was applied using a model-agnostic KernelExplainer
        i. Feature importance and directionality were determined
        ii. Dependence plots demonstrated feature effects and interactions

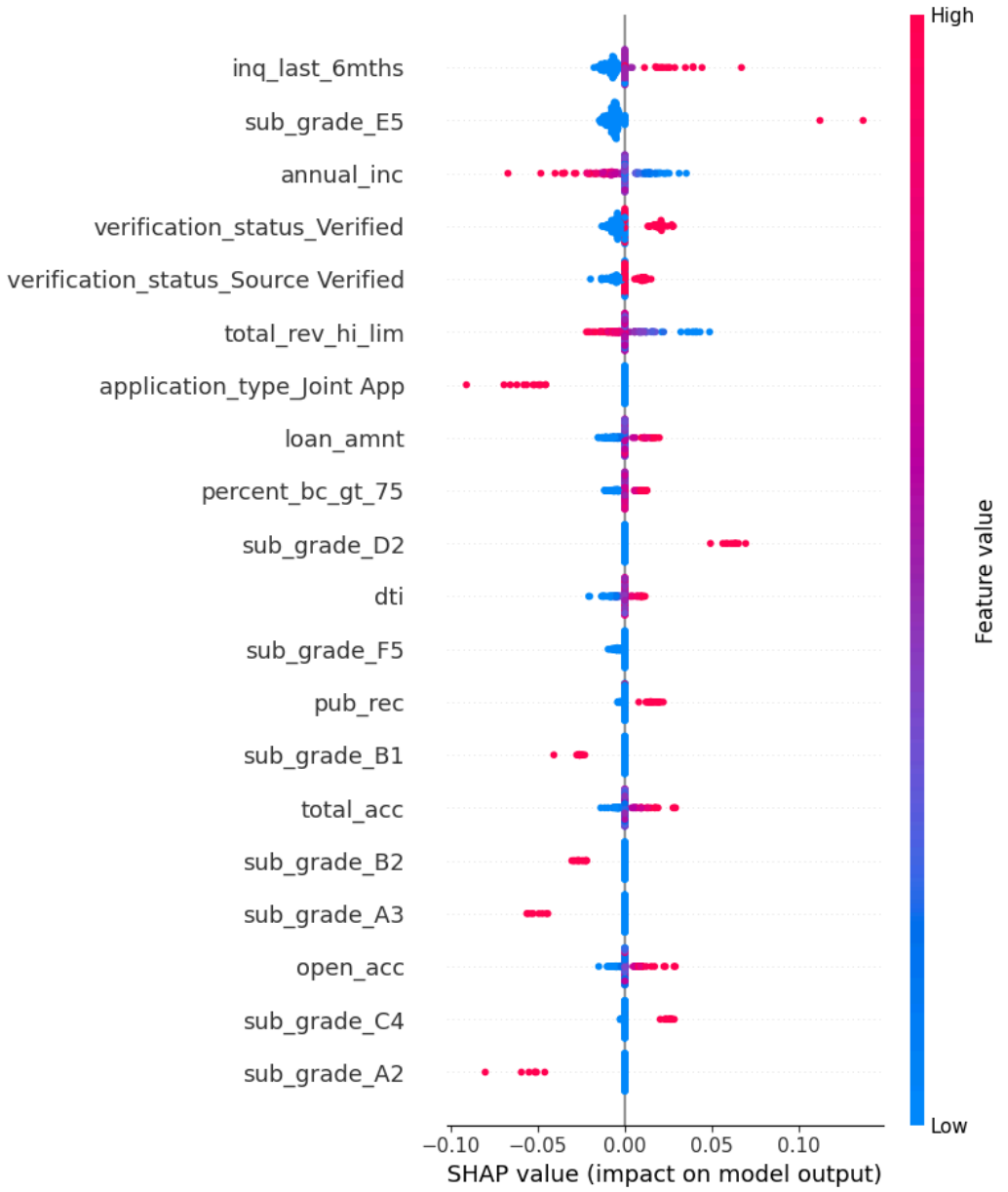
Results:
The model achieved a ROC-AUC score of 0.717, a recall of 0.680, and a precision of 0.197. The optimal balance between recall and precision was determined by flagging various top percentages of predicted probabilities as a loan default and then choosing the best threshold based on desired recall and precision metrics. The confusion matrix summarizes how predictions compare to observed values.

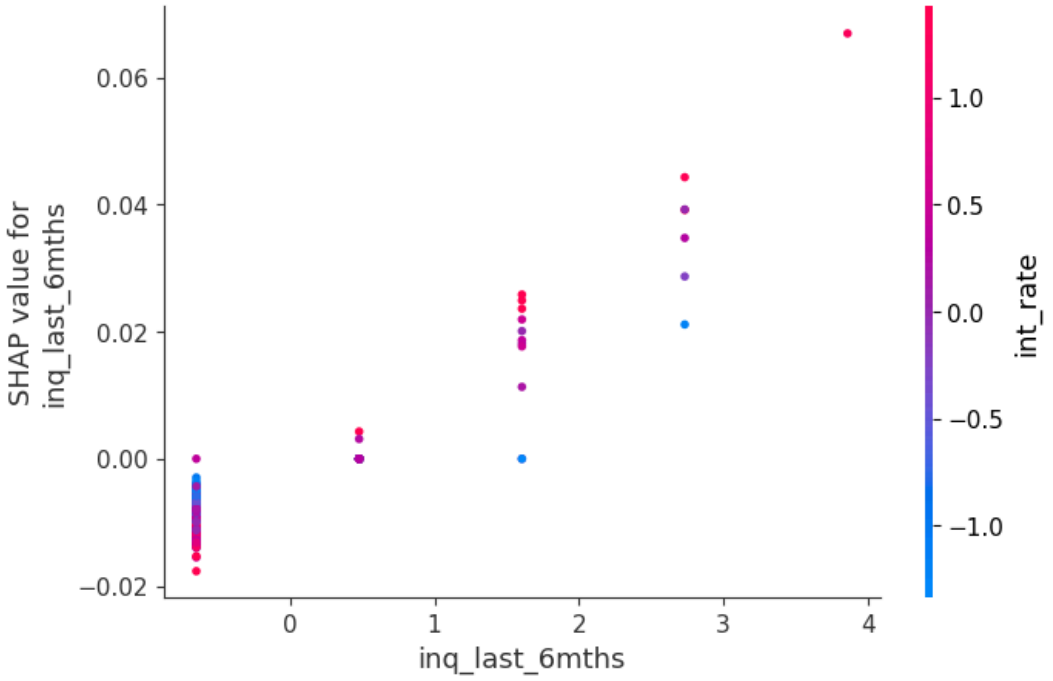

ROC Curve
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.717)
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix (Top 30% flagged)
True label
Predicted label
254483
145156
16794
35695
0
1
250000
200000
150000
100000
50000

Key findings from SHAP analysis included the following:

- Recent credit inquiries (inq_last_6mths) were the strongest predictor of default, with risk increasing monotonically as inquiry count rose.
- Credit sub-grade was highly influential, with worse sub-grades significantly increasing default probability.
- Interest rate amplified risk, particularly in combination with high inquiry counts.
- Income and revolving credit limits exhibited protective effects, reducing predicted default risk.
- Joint applications and verified income statuses were associated with lower risk.





Discussion:
The model successfully captures nonlinear risk patterns and interaction effects that are difficult to represent using traditional linear models. The SHAP dependence analysis highlights compounding risk factors, such as high inquiry volume combined with elevated interest rates, demonstrating the added value of neural networks in credit modeling.

Given that the cost of a false negative is greater than the cost of a false positive, the model achieved good performance based on ROC-AUC and recall metrics. Additionally, the model achieved nearly identical metrics for the training and validation sets, indicating excellent generalizability to unseen data.

Based on feature importance determined through SHAP, borrows with many recent credit inquiries, low loan sub-grades, and low annual incomes are higher risk for defaulting and should be more carefully screened.

Conclusion and Next Steps:
This project successfully predicted borrowers with a high risk of defaulting loans. The model identifies intuitive and economically meaningful drivers of loan default while maintaining strong predictive performance.

Possible next steps are:

1. To compare performance against other supervised models such as logistic regression and gradient boosting.
2. To explore fairness and bias analysis across demographic proxies.