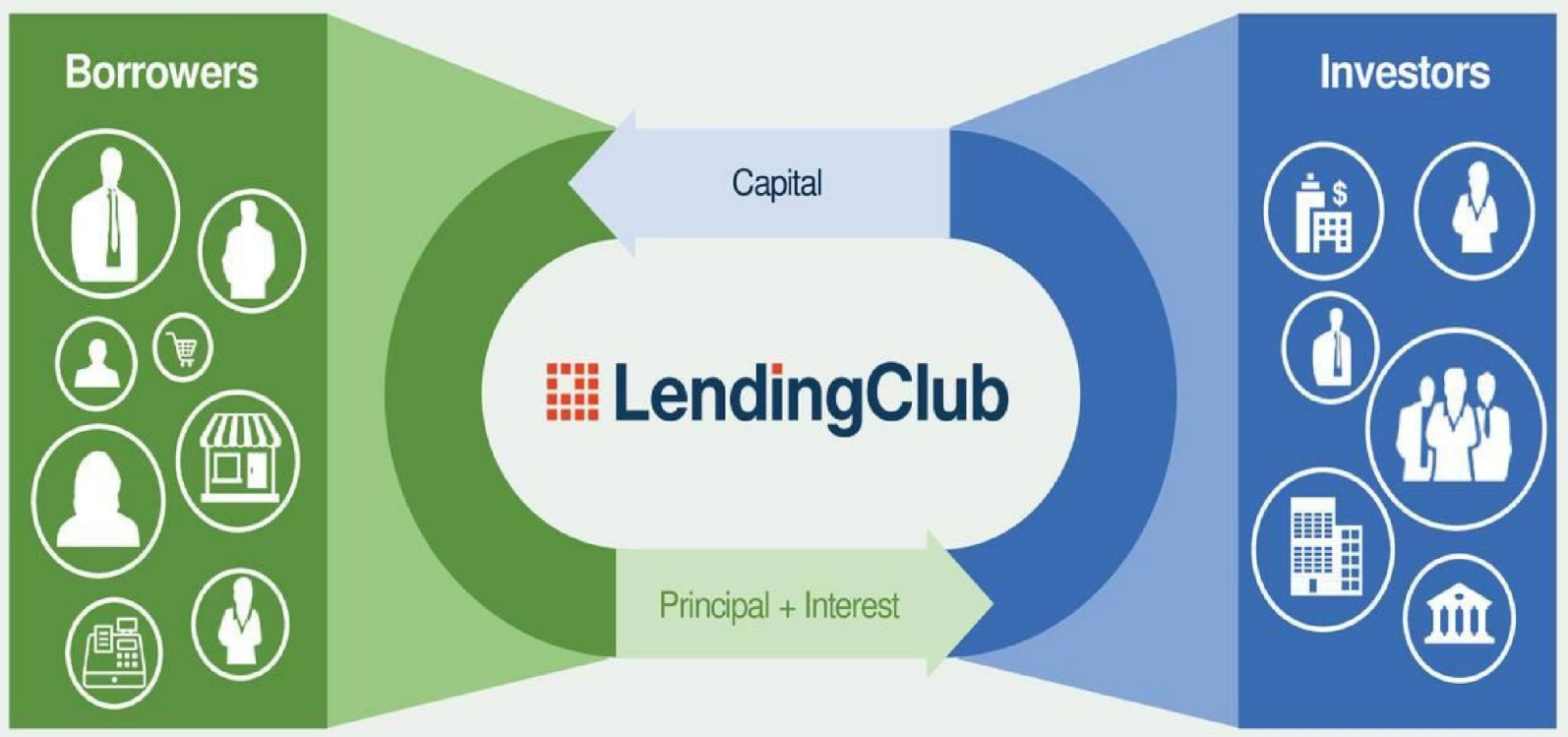
Lending Club

Loan Approval Optimization

Michael Son

Lending Club Business Operation



Problem

- **Charged Off Loans**
- **Default Loans**
- **Delayed Payments**

Solution

- **Identify “Bad Loans”**

Stakeholders

Investors

LendingClub

Data Information

- Source: Lending Club Loan Data 2007-11 - dataset by jaypeedevlin
- Data pertaining to **2007-2011**
- **Loan Status** and relevant financial information
- Number of entries: **42,538**
- Number of features: **115**

Features

- Loan Amount
- Term
- Installment
- Grade
- Employment Length
- Home Ownership
- Annual Income
- Verification Status
- Loan Status
- Purpose
- Loan Title
- Address State
- Debt/Income Ratio
- Delinquency
- Earliest Credit Line
- Inquiry
- # of Open Credit Lines
- Public Records
- Total Credit Revolving Balance
- ...

Feature Engineering

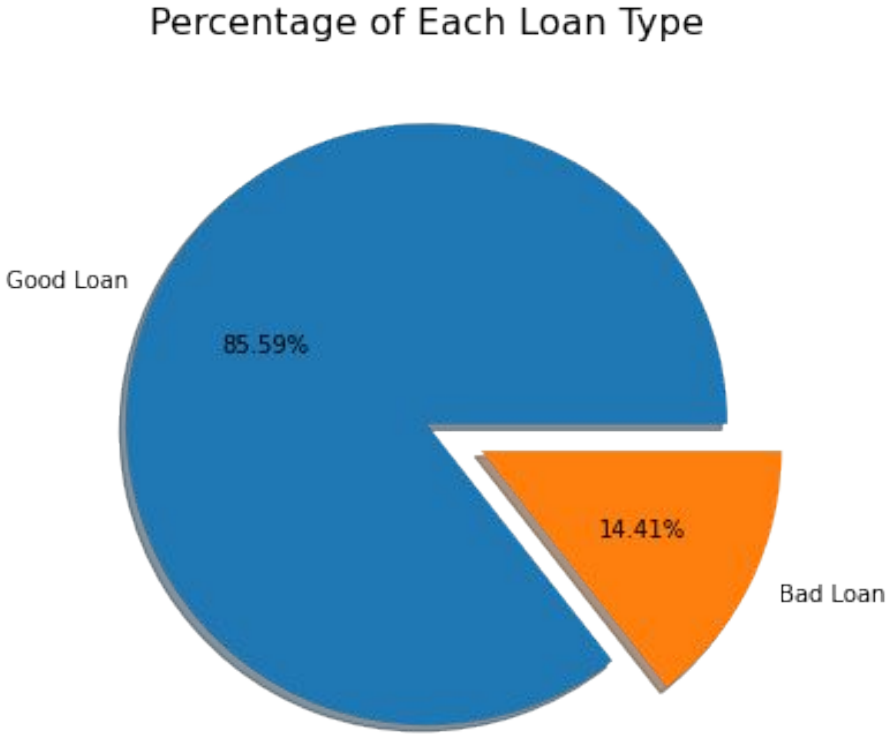
1. Data Cleaning
   - Removed empty columns/null values

2. Feature Selection
   - Removed redundancy/data leakage
   - Retained relevant/useful features

3. Conversion to Numerical Dtype
   - Numerical: **revol_util**
   - Ordinal: **grade/emp_length**

4. New Feature Added
   - **fico_range_avg**

5. Target Feature Engineering
   - Excluded loans in-progress
   - New classification: **loan_type**
   - Binarization

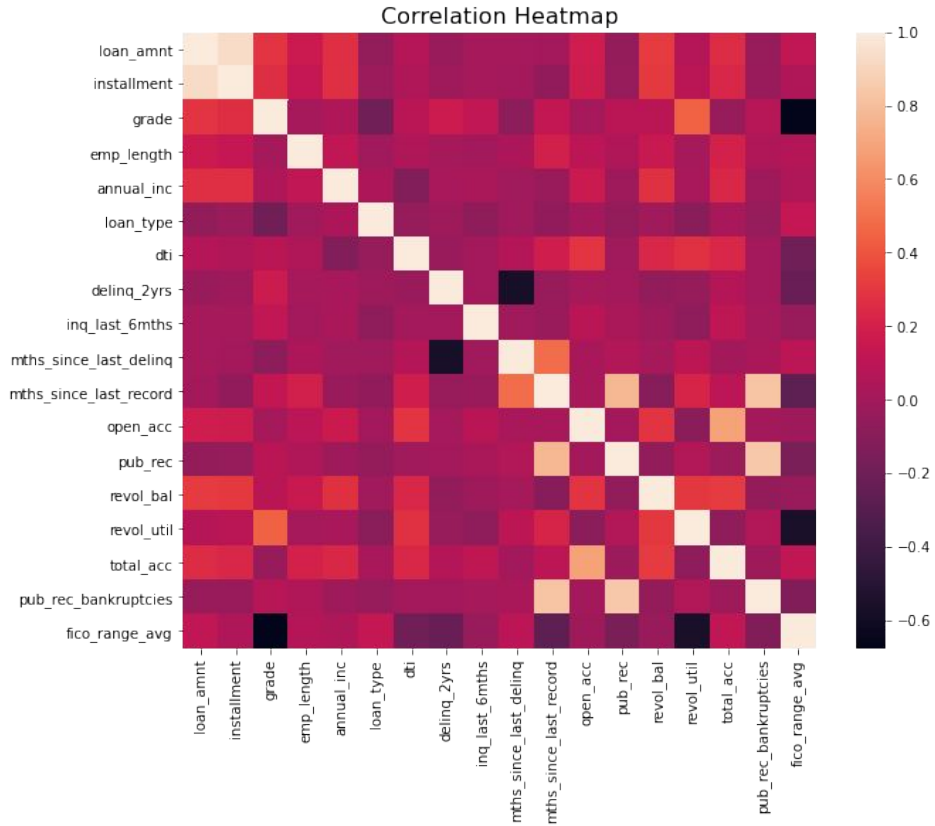


Good Loan

- Fully Paid

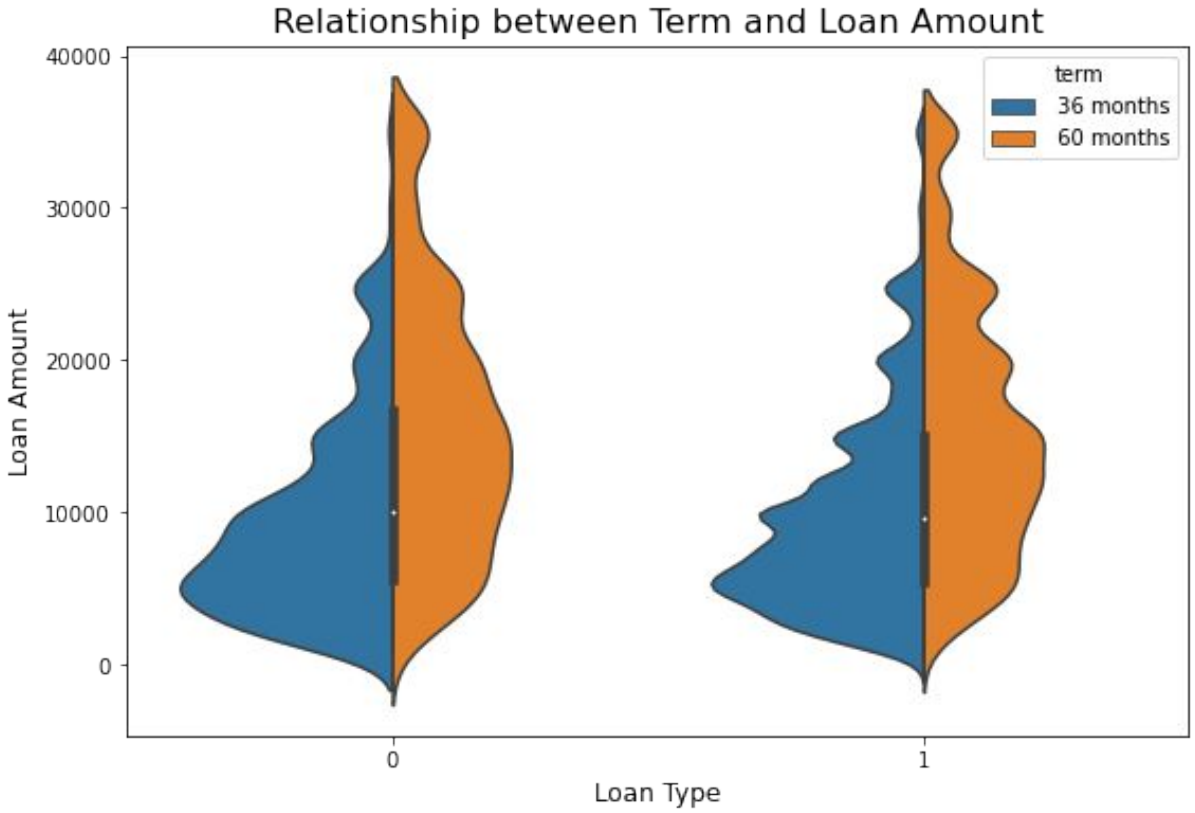
Bad Loan

- Charged Off
- Does Not Meet the Credit Policy

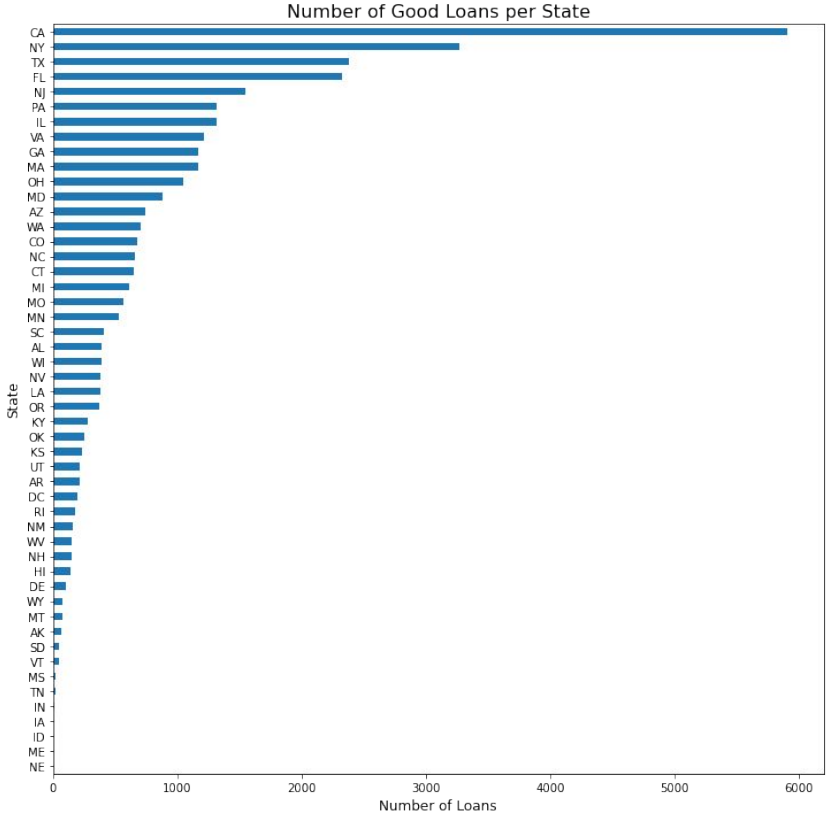


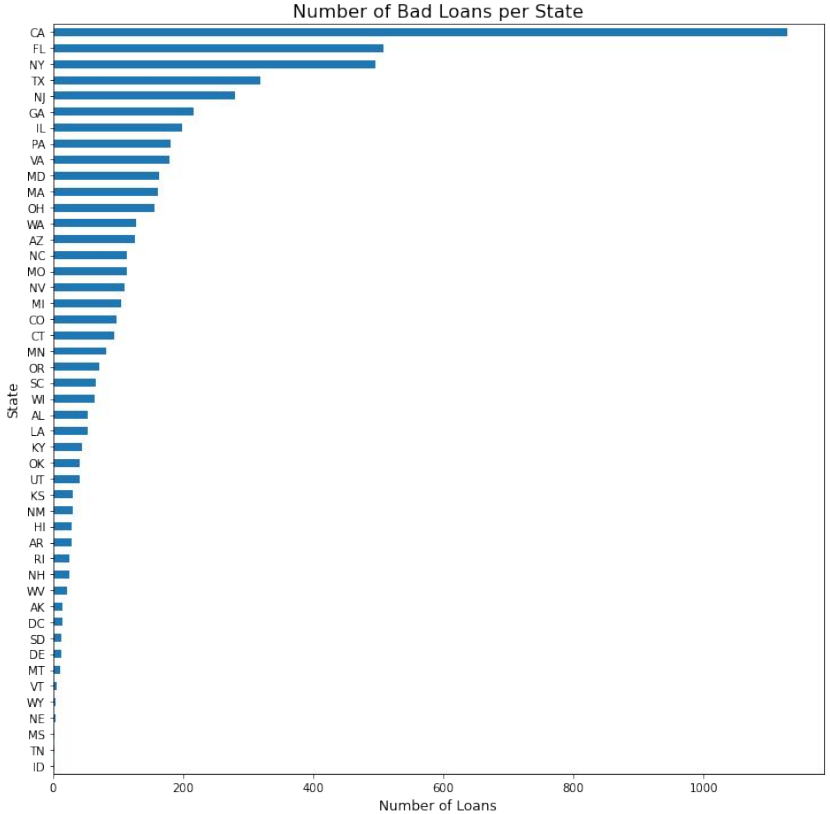
- Notable **loan_type** correlations
  - **grade**
  - **fico_range_avg**
  - **revol_util**
  - **Inq_last_6mths**

- Intrinsically linked features
  - **Installment/loan_amnt**
  - **pub_rec/pub_rec_bankruptcies**
  - **pub_rec/mths_since_last_record**
  - **open_acc/total_acc**
  - **...**

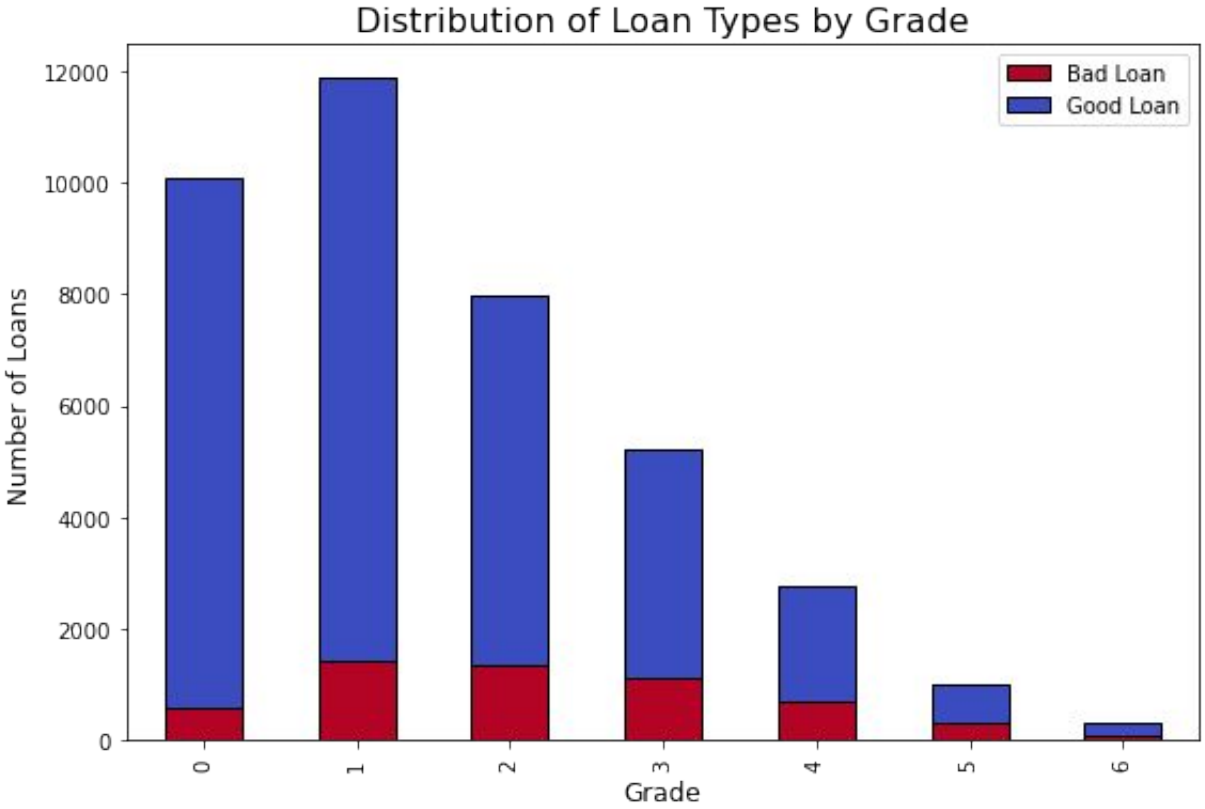


Loan amount tends to increase with longer term regardless of the loan type





States with highest # of good loans ≈ States with highest # of bad loans



Grade Description: 0 as worst, 6 as best

Modeling Overview

- Supervised Learning
- Binary Classification
  - 1: Good Loan
  - 0: Bad Loan
- Highly Imbalanced Data
- Machine Learning Tools: Scikit-Learn, Imbalanced-Learn, XGBoost

Modeling Procedure

I. Data Preprocessing

1. One-Hot Encoding
2. Datetime Objects to Ordinal Numeric
3. Training Test Split (80%: 20%)
4. Feature Standardization
5. Minority Class Oversampling

II. Randomized Search

- Stratified K-Fold with cv = 5
- n_iter = 30
- scoring = “roc_auc”

III. Training with Tuned Parameters

IV. Performance Evaluation

- Evaluation Metric: F1, ROC AUC

Techniques/Algorithms Used

Resampling Technique

- SMOTE from Imbalanced-Learn

Weighting Technique

- class_weight = “balanced” in Scikit-Learn classifiers

Hyperparameter Tuning Technique

- Randomized Search from Scikit-Learn

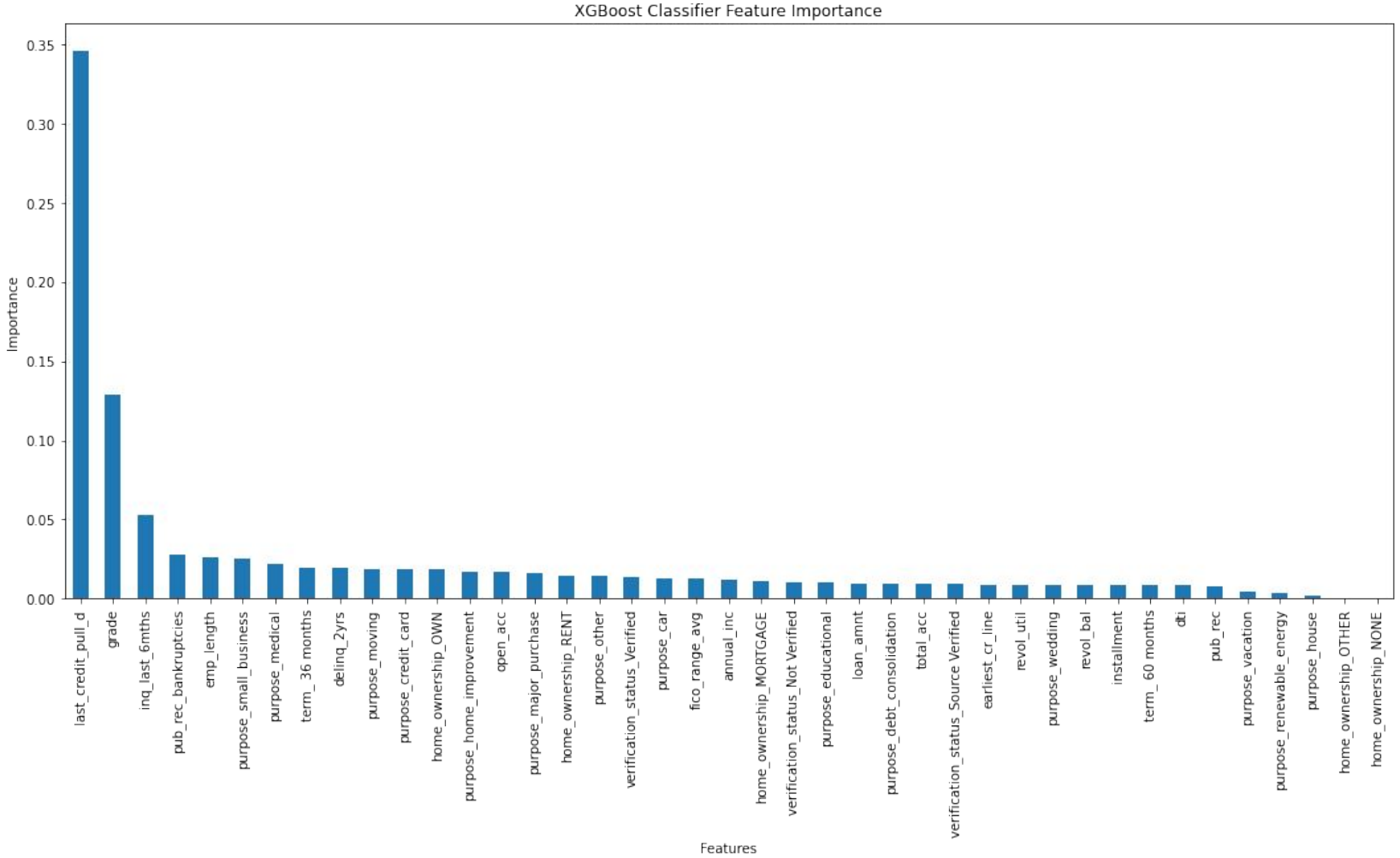
Classification Algorithms

1) Logistic Regression
2) Random Forest
3) Support Vector Machine
4) XGBoost

Model Comparison

Model	ROC AUC	Minority F1	Majority F1
Logistic Regression	0.78	0.47	0.80
Random Forest	0.74	0.53	0.91
Support Vector Machine	0.67	0.41	0.87
XGBoost	0.80	0.56	0.89

Best: XGBoost Worst: Logistic Regression


XGBoost Classifier Feature Importance
Importance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
last_credit_pull_d
grade
inq_last_6mths
pub_rec_bankruptcies
emp_length
purpose_small_business
purpose_medical
term_ 36 months
delinq_2yrs
purpose_moving
purpose_credit_card
home_ownership_OWN
purpose_home_improvement
open_acc
purpose_major_purchase
home_ownership_RENT
purpose_other
verification_status_Verified
purpose_car
fico_range_avg
annual_inc
home_ownership_MORTGAGE
verification_status_Not Verified
purpose_educational
loan_amnt
purpose_debt_consolidation
total_acc
verification_status_Source Verified
earliest_cr_line
revol_util
purpose_wedding
revol_bal
installment
term_ 60 months
dti
pub_rec
purpose_vacation
purpose_renewable_energy
purpose_house
home_ownership_OTHER
home_ownership_NONE
Features

Assumptions/Limitations

- Primary Assumption
  - Validity of the data

- Limitations
  - Outdated (year 2007-2011)
  - Large volume of missing features
  - Limited entries (only 39239 entries)

Conclusion

- Only 24 out of 115 features used
- Best model: XGBoost
- Primary features of importance: **last_credit_pull_d**, **grade**, **inq_last_6mths**
- Prospective improvement
  - Up-to-date dataset
  - Hyperparameter tuning with Grid Search
  - Alternative classifier algorithms: Neural Network, Deep Learning, etc.