Predicting Defaults of Personal Loans On Lending Club

Boya Yan, Mengmei Chen, Yuze Liu

Motivation

- Lending Club: the world's largest peer-to-peer lending platform
- Business model:
  - Lower interest rates
  - Origination fees from borrowers and service fees from investors
- Problems:
  - Information asymmetry
  - Uncollateralized loans
- Solution: a better screening mechanism
- **Our goal: use Machine Learning algorithms to predict defaults of personal loans for Lending Clubs**

Data

- https://www.lendingclub.com/info/download-data.action
- Detailed personal information provided by the borrowers when they applied, e.g.
  - Borrower's ID
  - Last payment date
  - The amount of the loan
  - Employment length
  - Credit score range
  - Debt-to-income ratio
- Information on whether the loans were fully paid or charged off
- Imbalanced data: 87.3% fully paid vs 12.7% charged off

Preprocessing - Nontext Features

- Selected 22 useful numerical/categorical features including:
  - Employment length
  - Home ownership status: Rend, Own, Mortgage, Other
  - State: NY, NJ...
  - Annual income
  - Debt-to-income ratio
  - The number of delinquencies in the past two years
  - The number of open credit lines
- Removed irrelevant features such as:
  - Borrower's ID
  - Next payment date
  - Issue date of the loan
  - Loan grade (LC's score)
- Transformed categorical features into Boolean features → 86
- Normalized all feature values to [0,1]

Preprocessing - Text Features

- Loan description: the description the borrowers provided on why they needed the loan
  - “Debt-Paying, Wife-Helping”
  - “I want to get out from the burden of credit card debt”
- Processing method:
  - ❖ Stop words
  - ❖ Stemming
  - ❖ Length of description
  - ❖ Logistic Regression to obtain probability
  - ❖ Append to the feature vector

Preprocessing: Data splitting

- Total: 15000
- Training: 9600
- Validation 2400
- Test: 3000

Evaluation Metrics:

True values

Predicted values

	True	False
True	TP	FP
False	FN	TN

- Accuracy: 1 - error rate
- Sensitivity: TP / (TP + FN)
  - Classify out people who can pay back
- **Specificity: TN / (TN + FP)**
  - **Classify out people who cannot pay back**
- Positive Predictive Value: TP / (TP + FP)

Baseline Accuracy

- Random Guessing: 50%
- Lending Club Accuracy: 61%
  - It assigns each of its loan a grade ranging from A to F and bases the interest rate on that grade
  - When we set the threshold at B, we got the highest accuracy 61%

Model: Logistic Regression (62%~63%)

Accuracy: 62.7%

Sensitivity: 61.94%

Specificity: 63.4%

PPV: 61.7%

Model: SVM (62%~63%)

	L1	L2	Poly	RBF	Sigmond
Accuracy	62.7%	63.5%	62.3%	62.4%	62%
Sensitivity	61.9%	60.7%	60.8%	60.8%	61.3%
Specificity	63.4%	66.2%	63.5%	64%	62.6%
PPV	61.8%	63.1%	61.5%	61.7%	61%

Model: Decision Tree (62%~63%)

Accuracy: 62.13%

Sensitivity: 59.52%

Specificity: 64.61%

PPV: 61.57%

Model: Random Forest (Around 67%)

Accuracy: 67.12%

Sensitivity: 63.8%

Specificity: 70.3%

PPV: 67.2%

- Highest accuracy
- Bagging method and random choice of features to prevent overfitting. Low variance.
- Faster to train than SVM

Evaluate on Test Data with Random Forest

Accuracy: 67.5%

Sensitivity: 63.5%

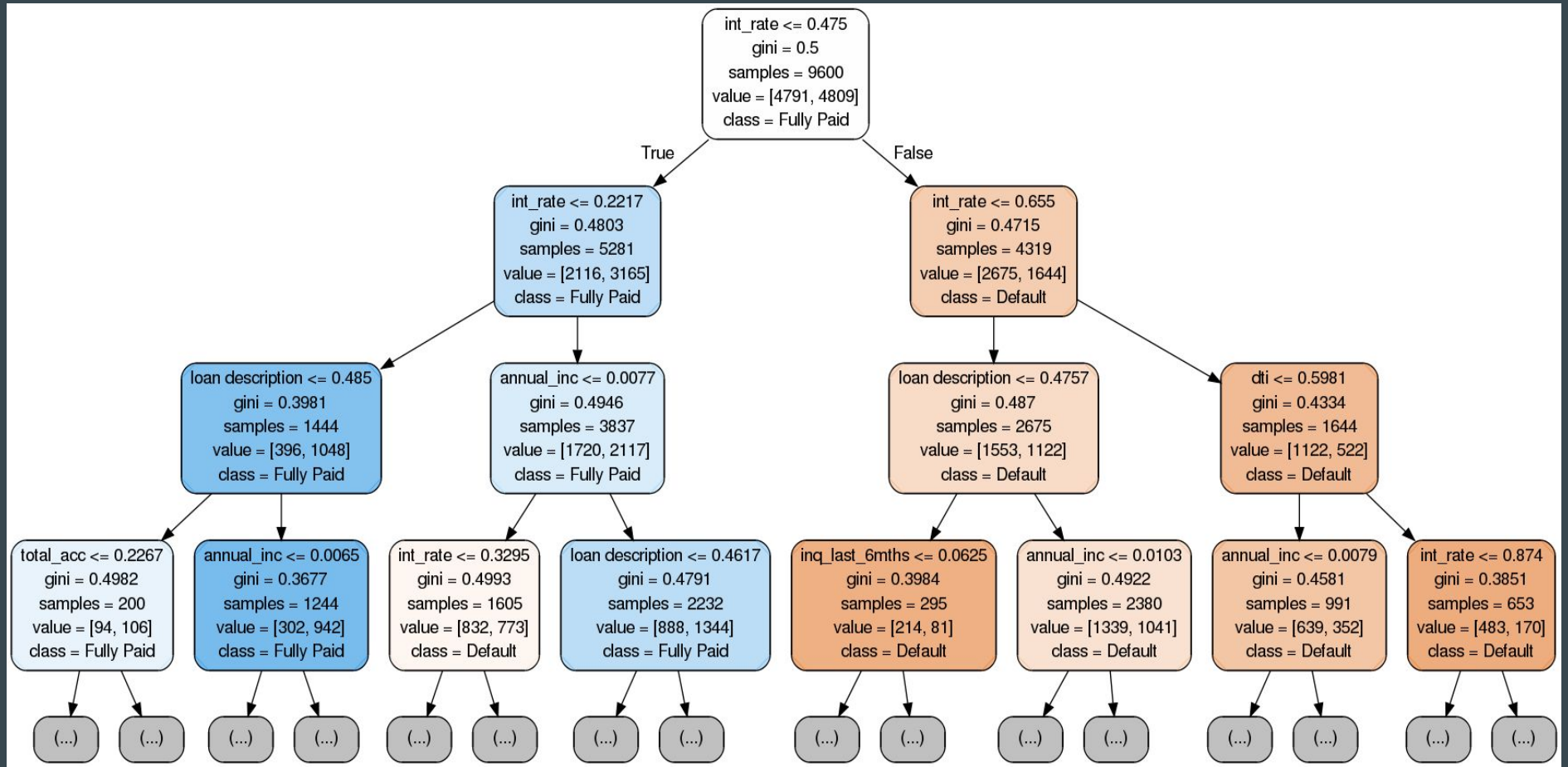
Specificity: 71.7%

PPV: 69.7%

Qualitative Analysis: Linear SVM

Features with Highest Weights	Features with Lowest (Negative) Weights
Annual Income States: ID, DC, MT, WY, KS, WV Total credit revolving balance (?) Number of derogatory public records (???)	Loan Amount The number of inquiries in past 6 months (excluding auto and mortgage inquiries) Debt to Income Ratio installment(The monthly payment owed by the borrower if the loan originates.) States: NJ, RI, FL, OK, NV, OR

Qualitative Analysis: Decision Tree



Qualitative Analysis: Random Forest

Feature Importance from high to low:

- Debt-to-Income ratio
- Annual income
- Revolving balance
- Installment
- Total number of accounts
- Earliest credit line
- Loan amount
- FICO score
- Employment length

Conclusion & Future Plan

- Our Strategy is effective
  - Accuracy: 67% > 61%
  - Specificity is high
- PCA or Topic Model on text data
- Clustering on data to see patterns in features

THANK YOU!