Predicting Loan Defaults

Kenneth Chee, Fong Yew Loong, Chen Wanlun, Tan Jiaqi

Problem Statement

- How would lenders know whether their borrowers have high default risk based on the information they submit?
- Which features are more important out of all the information submitted by borrowers?

Datasets

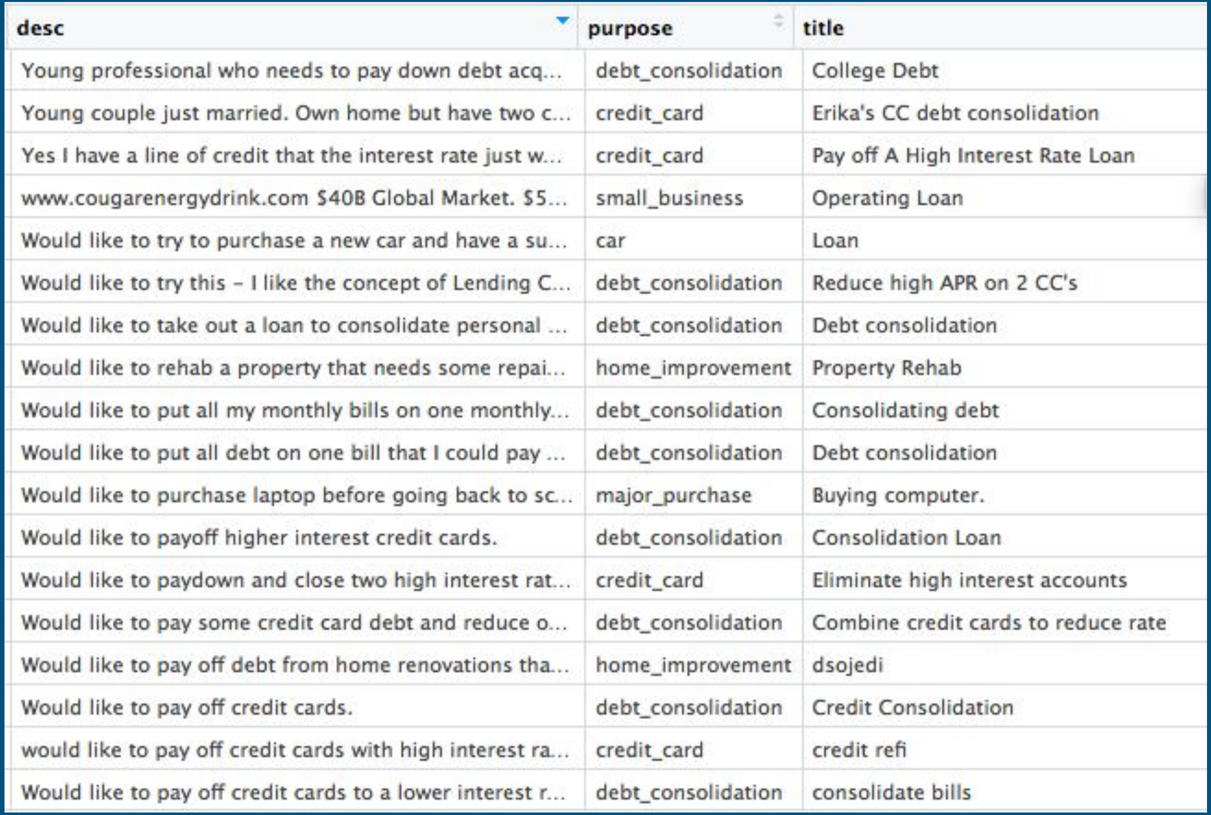
- Lending Club: peer to peer lending company which connects borrowers and investors.
- Loan data for all loans issued through the period 2008 to 2011
- Contains over 50 dimensions and 20,000 observations (n >> d yay!)

Variables

- Dependent Variable:
  - Loan status: Charged Off or Fully Paid

- Independent Variables (> 50):
  - Quantitative: Number of accounts the borrow is now delinquent, Self-reported annual income, Employment length (and more!)
  - Categorical: Purpose of Loans, Home Ownership Status
  - Qualitative: Loan Description

Dealing with Qualitative Variable



desc	purpose	title
Young professional who needs to pay down debt acq...	debt_consolidation	College Debt
Young couple just married. Own home but have two c...	credit_card	Erika's CC debt consolidation
Yes I have a line of credit that the interest rate just w...	credit_card	Pay off A High Interest Rate Loan
www.cougarenergydrink.com $40B Global Market. $5...	small_business	Operating Loan
Would like to try to purchase a new car and have a su...	car	Loan
Would like to try this - I like the concept of Lending C...	debt_consolidation	Reduce high APR on 2 CC's
Would like to take out a loan to consolidate personal ...	debt_consolidation	Debt consolidation
Would like to rehab a property that needs some repai...	home_improvement	Property Rehab
Would like to put all my monthly bills on one monthly...	debt_consolidation	Consolidating debt
Would like to put all debt on one bill that I could pay ...	debt_consolidation	Debt consolidation
Would like to purchase laptop before going back to sc...	major_purchase	Buying computer.
Would like to payoff higher interest credit cards.	debt_consolidation	Consolidation Loan
Would like to paydown and close two high interest rat...	credit_card	Eliminate high interest accounts
Would like to pay some credit card debt and reduce o...	debt_consolidation	Combine credit cards to reduce rate
Would like to pay off debt from home renovations tha...	home_improvement	dsojedi
Would like to pay off credit cards.	debt_consolidation	Credit Consolidation
would like to pay off credit cards with high interest ra...	credit_card	credit refi
Would like to pay off credit cards to a lower interest r...	debt_consolidation	consolidate bills

Dealing with Qualitative Variable

Text Analysis of loan description:

- Use term-frequency inverse document frequency to determine most frequently used words/phrases
  - Why? Expect loan descriptions with phrases such as “pay off”, “stable job” etc. to be associated with loan repayment
- Use sentiment analysis to determine positivity/negativity of description
  - Why? Expect loan descriptions with positive sentiments to be associated with loan repayment

Approaches

Linear Classification

- Split the data into training and testing sets
- Perform dimensionality reduction to select features that explain the largest proportion of variance
- Perform linear classification methods on training data
- Run K-folds validation

Random Forests

- Draw a bootstrap sample from the training data
- Train a classifier on the bootstrap sample where each split (dimension) is newly chosen for each sample
- Repeat bootstrapping and classify based on majority voting of all bootstrap samples

Boosting

- Draw a bootstrap sample, weighting misclassified observations higher each time
- Train a classifier on the bootstrap sample
- Repeat bootstrapping and classify based on majority voting of all bootstrap samples

Choosing the Best Approach

- Determine performance by Area Under ROC Curve
- Generate confusion matrices for each of the 3 approaches.
- Select the best model which gives the highest proportion of correct classifications.
- Use that in our web app!

Ideal Outcome

- Using our web app, investors will just need to upload identified key information provided by prospective borrowers.
- They will then receive a prediction of whether these borrowers are likely to default.