Lending Club

Roebi de Bruijn

Introduction

- Peer-to-peer lending annuity loans
- Alternative to stock market
- Risk loan defaults: loss of interest and principal
- Claimed 4% default rate
- Grade system A-G developed by Lending Club should be
  - Grade A has lowest default rate and lowest interest rate
  - Grade G has highest default rate and highest interest rate

Goals

- How good is Lending Club's grading system of their loans?
  - Do loans with low grades indeed default less?
  - Are some grades more profitable than others?
  - Is adding more features than only grade to a classifier to predict default beneficial for its performance?
- What can we recommend to the investors of Lending Club?

Methods

- Lending Club dataset from Kaggle
  - 2007-2015
  - 887,379 loans and 74 features
  - Features about loan (32) and about borrower (42)
  - Both ongoing loans and loans that went to full term
  - Default rate 5%
- Selecting loans and features
  - Select only loans that went to full term: 252,971 (28.5%)
  - Default rate of these loans 18%
  - Select features that are known at the start of the loan
  - Remove non-predictive features like id
  - Remove features with >10% missing values
  - Remove loans with a missing value in one of the selected features (200)
  - Remaining features for prediction: 23

Methods

- Engineered features:
  - Return-of-investment: not used for prediction just exploration

$$ROI = \frac{recieved}{committed} - 1$$

  - Days-since-first-credit-line: Number of days between first credit line and issue date (instead of those two features)
  - Annual income: chance all incomes > 200,000 to 200,000
- Prediction
  - Classifiers: Logistic Regression and Random Forest
  - Features: only grade and all 23 features
  - Metrics: confusion matrix, AUC (area-under-curve)

Features (1)

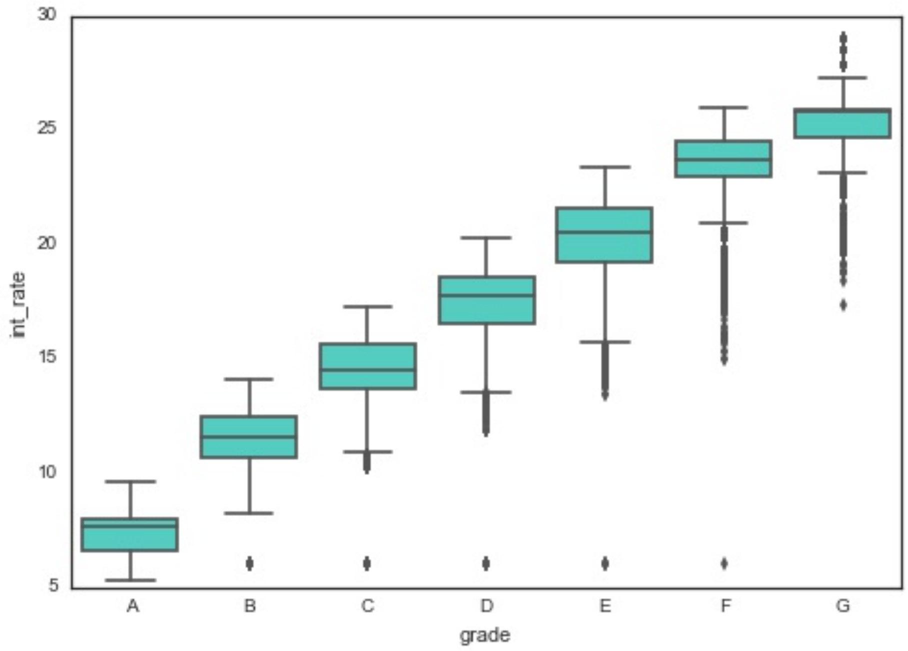
Feature	Description
term	36 or 60 months
int_rate	interest rate
installment	monthly payment
grade	A-G (low-high risk)
sub_grade	A1-G5
emp_length	employment length 0-10 (10 is 10 or more)
home_ownership	rent, own, mortgage, other
annual_inc	annual income 0-200,000
purpose	purpose provided by borrower
zip_code	format 000xx
addr_state	state borrower lives in

Features (2)

Feature	Description
delinq_2yrs	number of times >30 days late in 2 years
inq_last_6mths	number of credit inquiries in 6 months
open_acc	number of open credit lines
pub_rec	number of derogatory public records
revol_bal	total credit revolving balance
revol_util	ratio credit used versus all credit of borrower
total_acc	number of credit lines
acc_now_delinq	number of accounts borrower is now delinquent
loan_amnt	loan amount
dti	debt-to-income
loan_status	fully paid or charged off; feature to predict
days_since_first_credit_line	days between issue date and first credit line

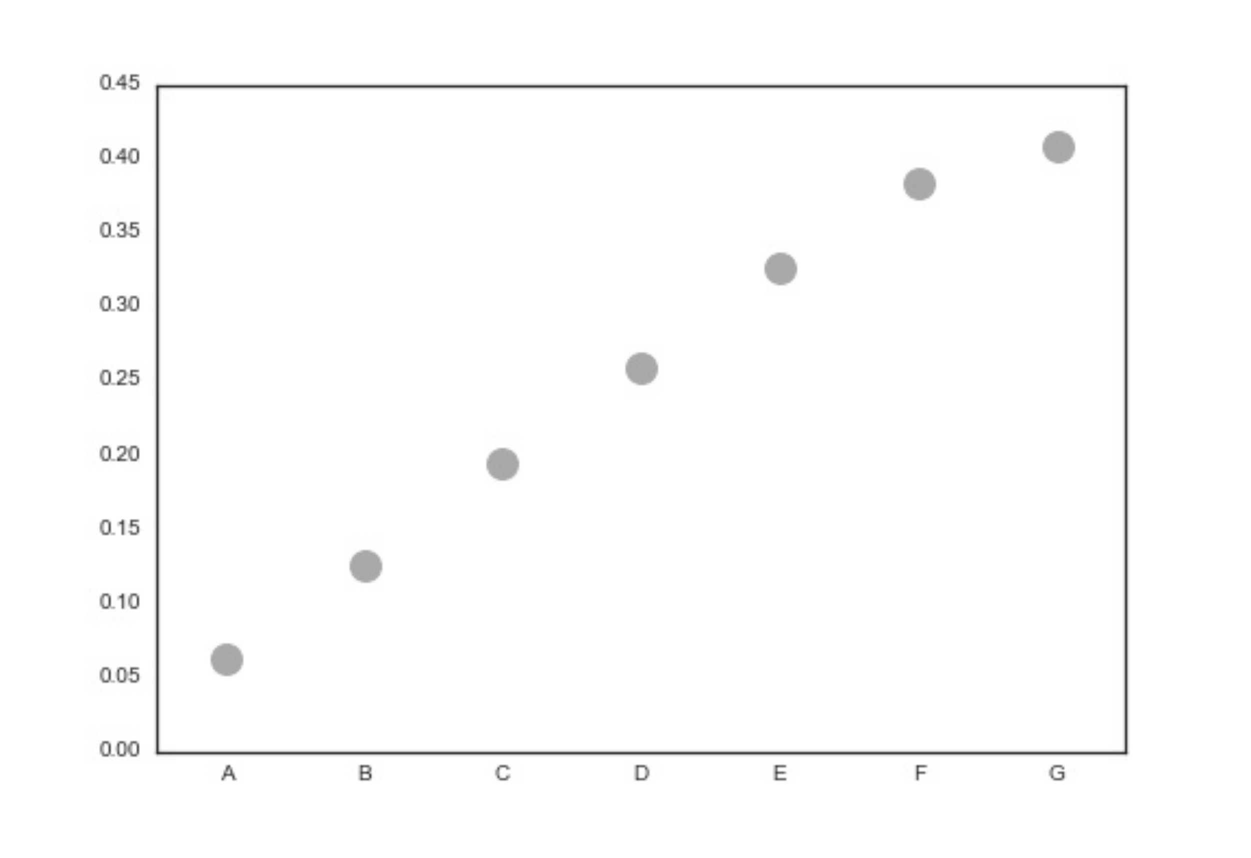
Results - exploration

- Interest rate mostly higher with riskier grade to make riskier loans still attractive



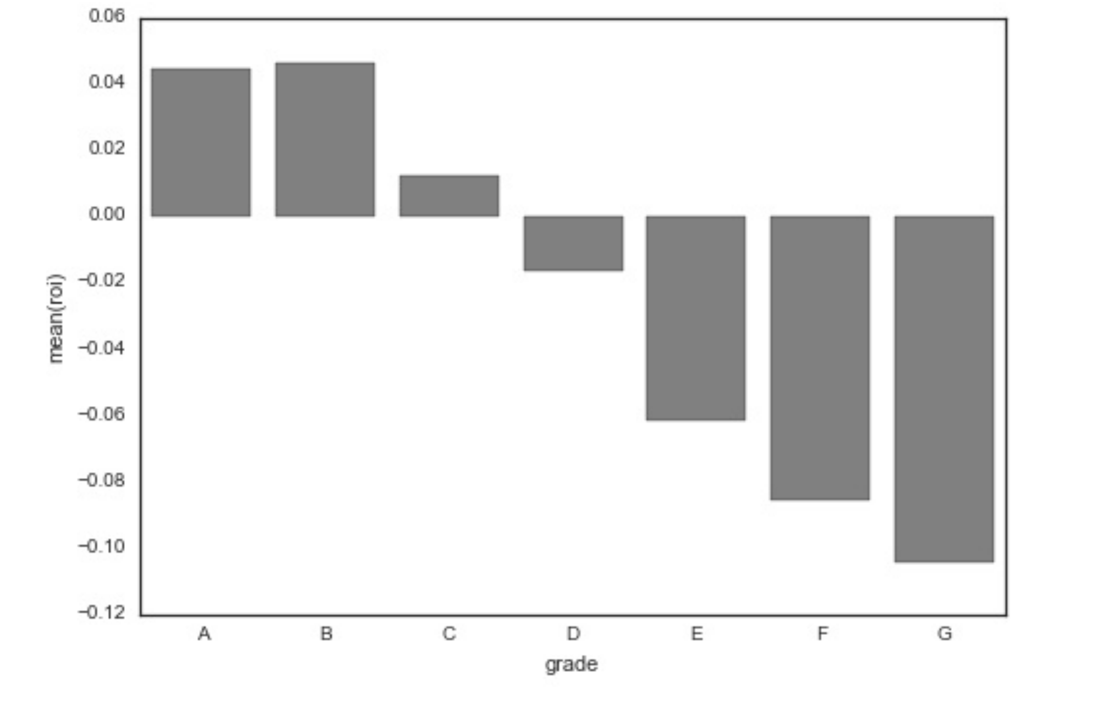
Results - exploration

- Grade indeed predictive of default rate



Results exploration

- Only grades A (4.5%), B (4.6%) and C (1.2%) on average profitable



Results - prediction

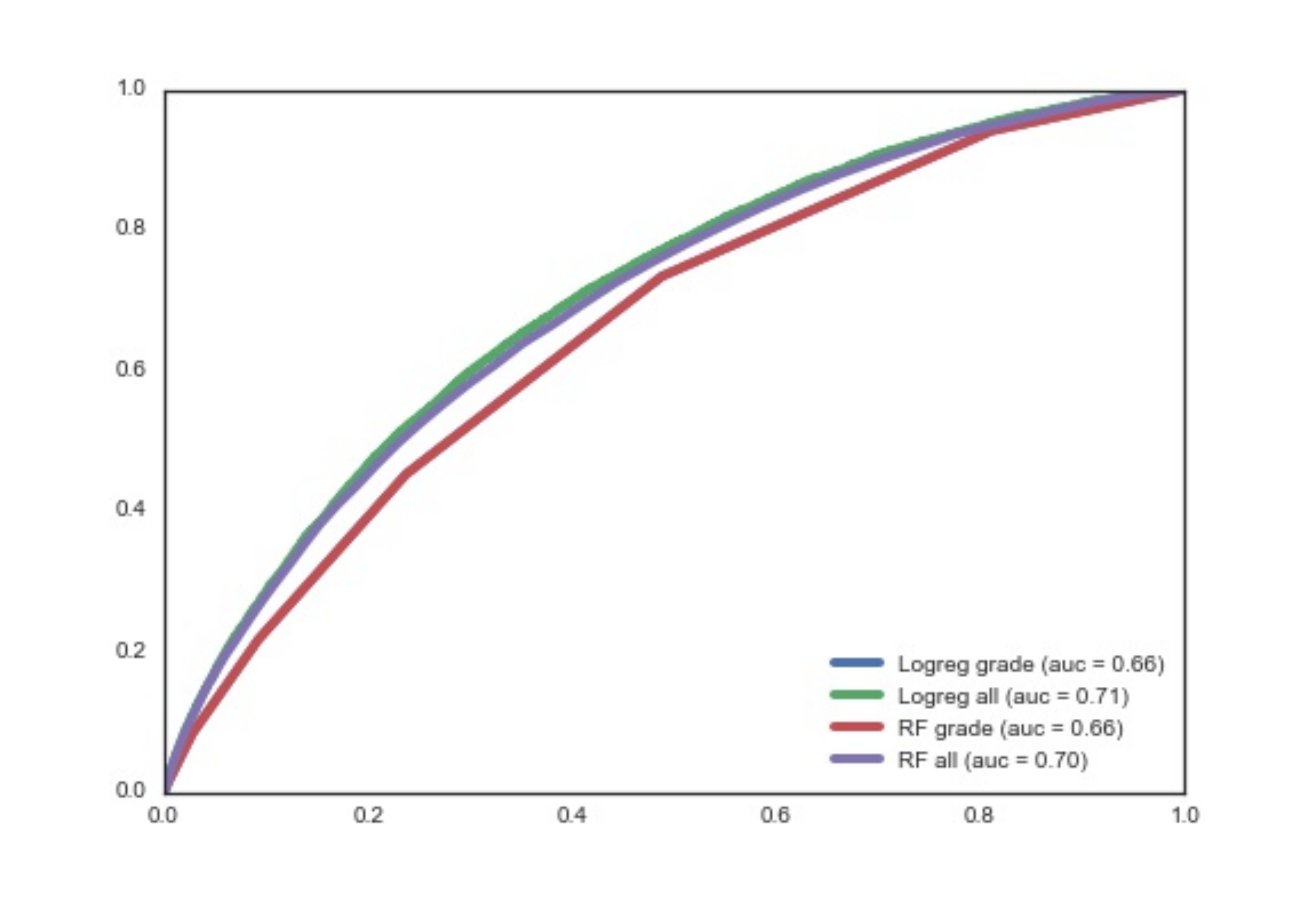
LogReg Grade	Predict Default	Predict Paid
Actual Default	236	13,302
Actual Paid	330	61,964

LogReg All	Predict Default	Predict Paid
Actual Default	547	12,991
Actual Paid	548	61,746

RF Grade	Predict Default	Predict Paid
Actual Default	0	13,538
Actual Paid	0	62,294

RF All	Predict Default	Predict Paid
Actual Default	582	12,956
Actual Paid	572	61,722

Results prediction



Results - prediction

- In general the algorithms just predicted mostly ‘Paid’ since that was the dominant class (82%) this leads to unrealistically high accuracies
- AUC was chosen as evaluation metric since it is uninfluenced by unequal classes (random always 0.5)
- Random Forest and Logistic Regression were not very different
- Logistic Regression all features performed best (AUC 0.71), but differences were marginal between only grade (AUC 0.66)and all features.
- Features most important and significant for increase in performance: interest rate, annual income, term and dti

Conclusions

- How good is Lending Club's grading system of their loans?
  - Loans do indeed default less in less risky grades, grade is a very good indicator of default risk
  - Grades A and B are the most profitable in average with everything riskier than C not being profitable on average
  - Adding more features than grade is only marginally beneficial in predicting default rate
  - The features found to help performance most next to grade were: interest rate, annual income, term and debt-to-income

Conclusions

- What can we recommend to the investors of Lending Club?
  - Investing in riskier grades is not worth the higher interest rate, so only invest in grades A and B
  - Grades A and B will give on average a return-on-investment of 4.5%
  - Next to grade investors can look for loans with a short term (36 months) which are from borrowers with a high annual income and who have a low debt-to-income rate.