Lending Club Loan Prediction

DATS 6103

Elizabeth Deschaine
Timur Mukhtarov

DATA SCIENCE PROGRAM
COLUMBIAN COLLEGE OF ARTS AND SCIENCES

LendingClub

What is Lending Club

- Peer-to-peer lending platform
- Pioneer in the rapidly developing fintech industry
- Lower cost than traditional bank loan programs
- Fast loans, interesting investment
- 2006 - Lending Club is born
- 2014 - successful IPO
- 2016 - Difficulty in attracting investors

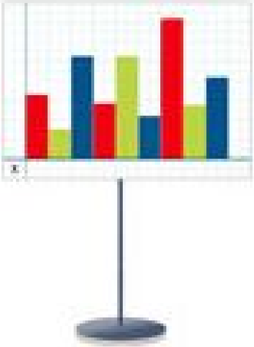
How Peer Lending Works

Borrowers apply for loans.

Investors open an account.

Borrowers get funded.

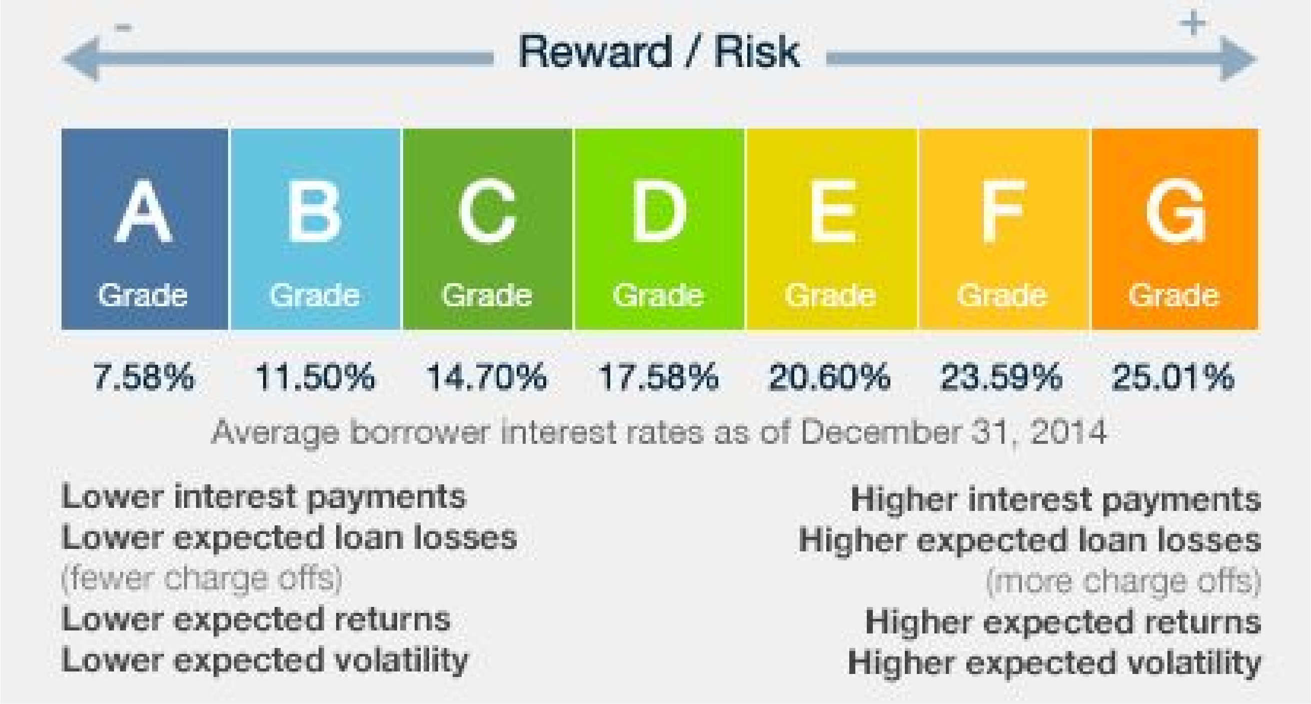
Investors build a portfolio.



Borrowers repay automatically.

Investors earn & reinvest.

Grades and Interest Rate



A Grade	B Grade	C Grade	D Grade	E Grade	F Grade	G Grade
7.58%	11.50%	14.70%	17.58%	20.60%	23.59%	25.01%

Average borrower interest rates as of December 31, 2014

Lower interest payments
Lower expected loan losses
(fewer charge offs)
Lower expected returns
Lower expected volatility

Higher interest payments
Higher expected loan losses
(more charge offs)
Higher expected returns
Higher expected volatility

Project Objective

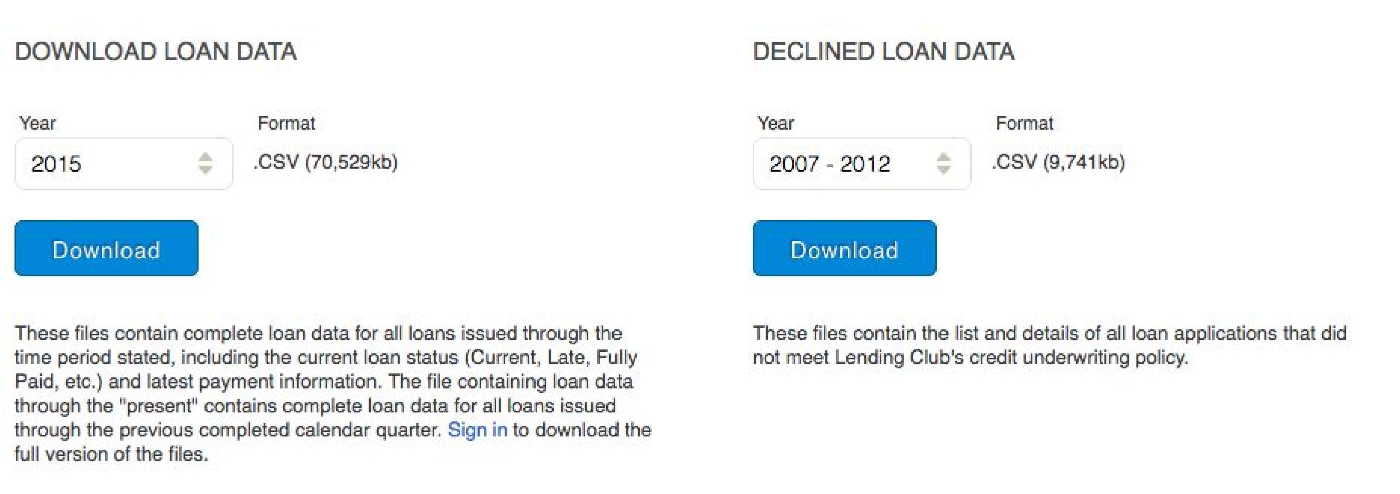
- Provide a tool for potential investors to predict the probability that a loan will “succeed”
- Success is defined as the loan either paid off or current

Strategies

- Categorize the data based on the borrower’s description
- Classify the loan status into Success (1) or Failure (0)
- Use interest rate, dti, and loan amount to predict loan status
- Build logistic regression and SVM models

The Data

Lending Club public data/Kaggle



DOWNLOAD LOAN DATA

Year: 2015

Format: .CSV (70,529kb)

These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Sign in to download the full version of the files.

DECLINED LOAN DATA

Year: 2007 - 2012

Format: .CSV (9,741kb)

These files contain the list and details of all loan applications that did not meet Lending Club's credit underwriting policy.

887k	Rows in the original data set
50+	Variables
2	Confused Data Science students

2 models

Logistic Regression

Support Vector Machine

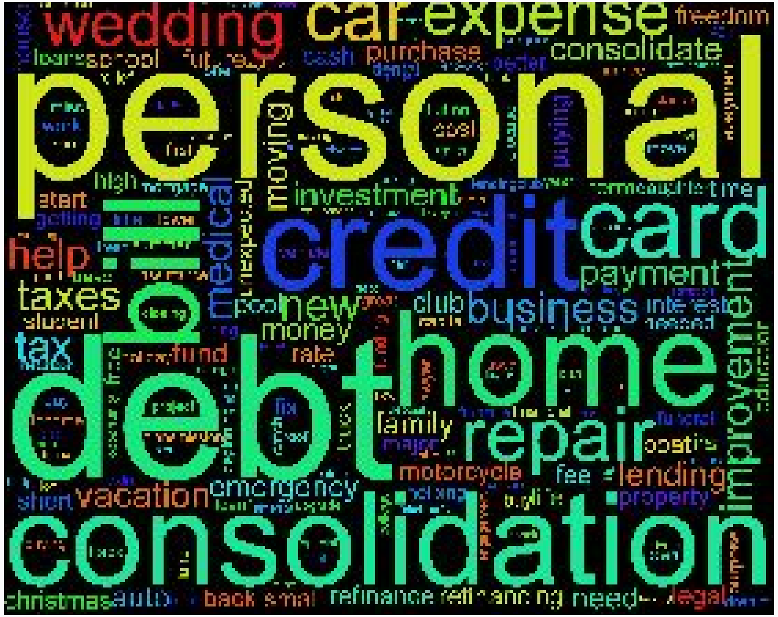
Exploratory Data Analysis (EDA)

[Tableau]

Tools

- Python
    - Pandas
    - Numpy
    - Word Cloud
    - Natural Language Toolkit (NLTK)
        - Word Net
    - Sklearn
        - Logistic Regression
        - SVM
- Tableau
    - Visualizations

Generate Word Cloud



Python Library

Requires “list of words”

Way to check progress

Natural Language Toolkit

> “[NLTK] provides easy-to-use interfaces… such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries”

```
for syn in wordnet.synsets(target_word):
    for l in syn.lemmas():
        local_list.append(l.name())
```

Synset: a set of synonyms that share a common meaning.

Each synset contains one or more lemmas, which represent a specific sense of a specific word.

Categories of Loans

```
category_list =

["home", "wedding", "medical", "business", "car", "vacation"]
```

Building a Logistic Regression model

Y = Loan Status (Success vs. Failure)

Features (Xi) = Interest Rate, Debt-to-Income ratio (DTI), Loan Amount

Data = Categorized Data

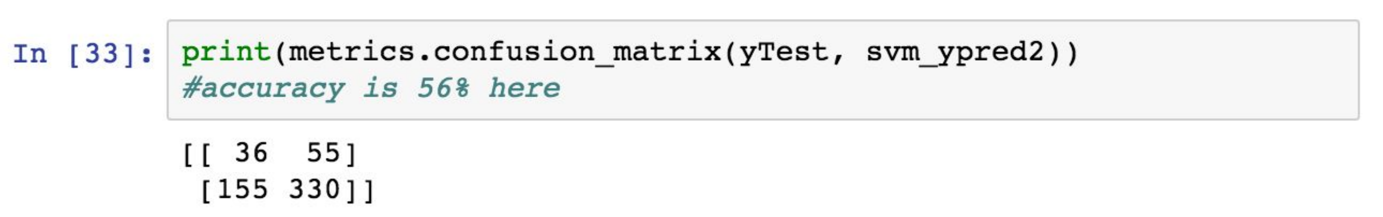
class_weight='balanced'

Predicting Probabilities

- Subset data based on category
- User selects interest rate, dti, loan amount
- Accuracy is based on Probability of class 1 or class 0

Support Vector Machines (SVM) model

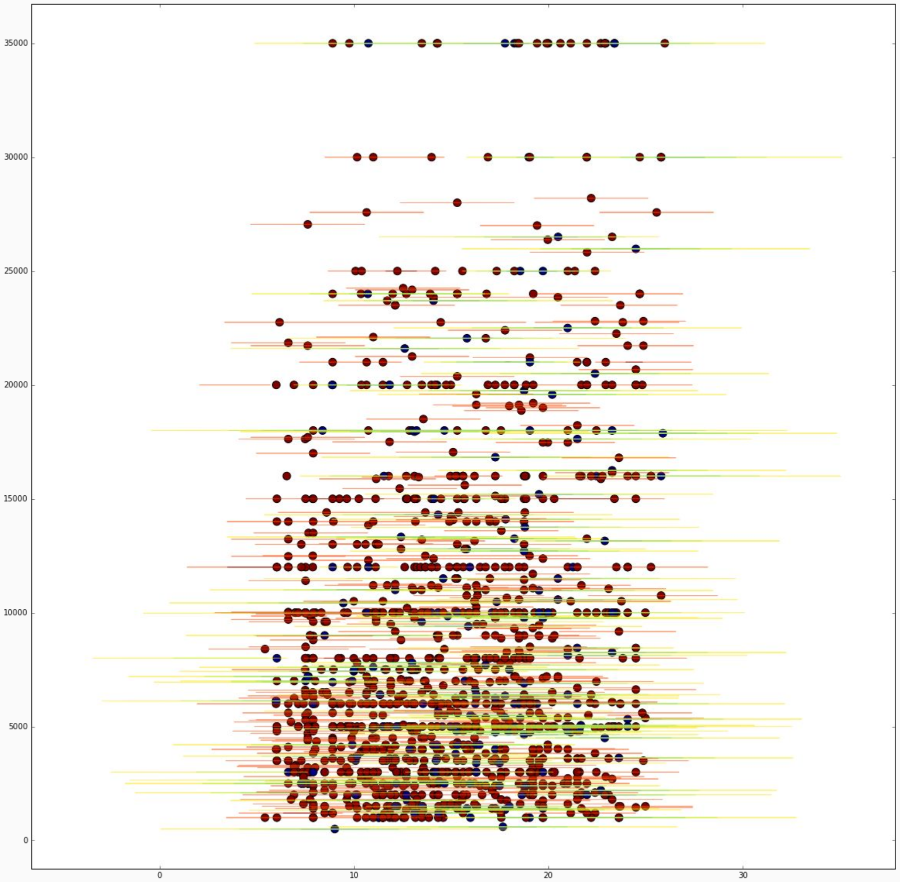
- 50+ percent accuracy
- kernel='rbf'
- gamma=0.05
- C=1
- class_weight='balanced'
- Features: Interest Rate, Loan Amount



```
In [33]: print(metrics.confusion_matrix(yTest, svm_ypred2))
         #accuracy is 56% here

         [[ 36  55]
          [155 330]]
```

SVM graph



Conclusions

- Interactive visualizations on Tableau Public
- Successfully categorized “Other” loans
- A logistic regression model, 3 features, good accuracy
- A prototype of a program prospective lenders can use
- SVM model

Challenges

- SVM models with better accuracy
- Decision Tree models
- Neural Networks

Q&A