Fintech Project

A Robotic Investor for Lending Club

WEEK 2

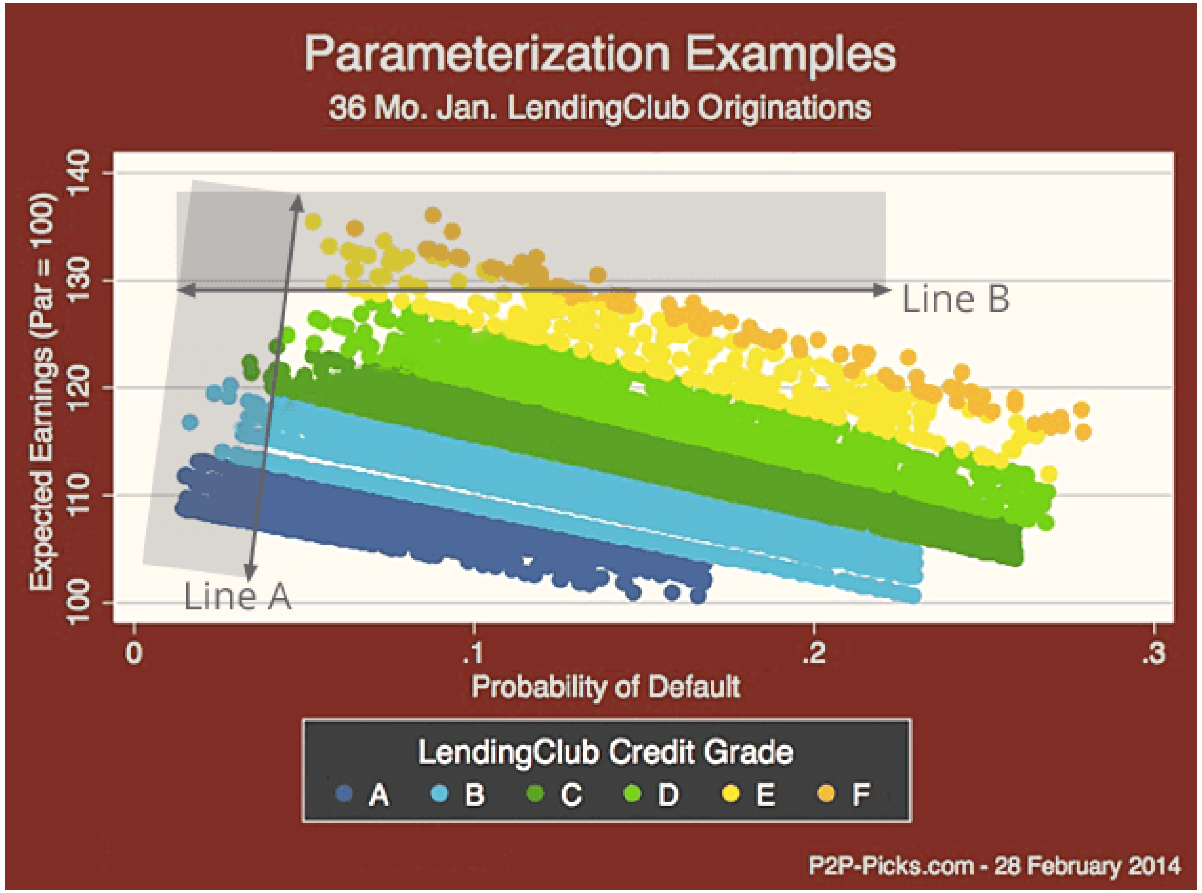
- Questions from week1

- Homework?
  - Data Fetching
  - Feature name unifying
  - Current data and historical data comparison

Data Preparation （interacting with EDA）

- Understand the data and the features
- Unify historical data features and the currently listed loan features
  - Different format
  - Common features (why different features exist?)
- Useless features
- Missing value handling
- Data types: numerical and object (string)
  - Object: intrinsically numerical; date time; ordinal; high cardinality

Modeling Target: ROI or others?



P2P-Picks.com - 28 February 2014

For interest of i%, and default rate p
Theoretical ROI (%)

100+ROI = (100+I)*(1-p)
(if default ones receive no payments at all)

(http://blog.lendingrobot.com/research/predicting-the-number-of-payments-in-peer-lending/)

Target? Loan Status

As of 05/30/2017

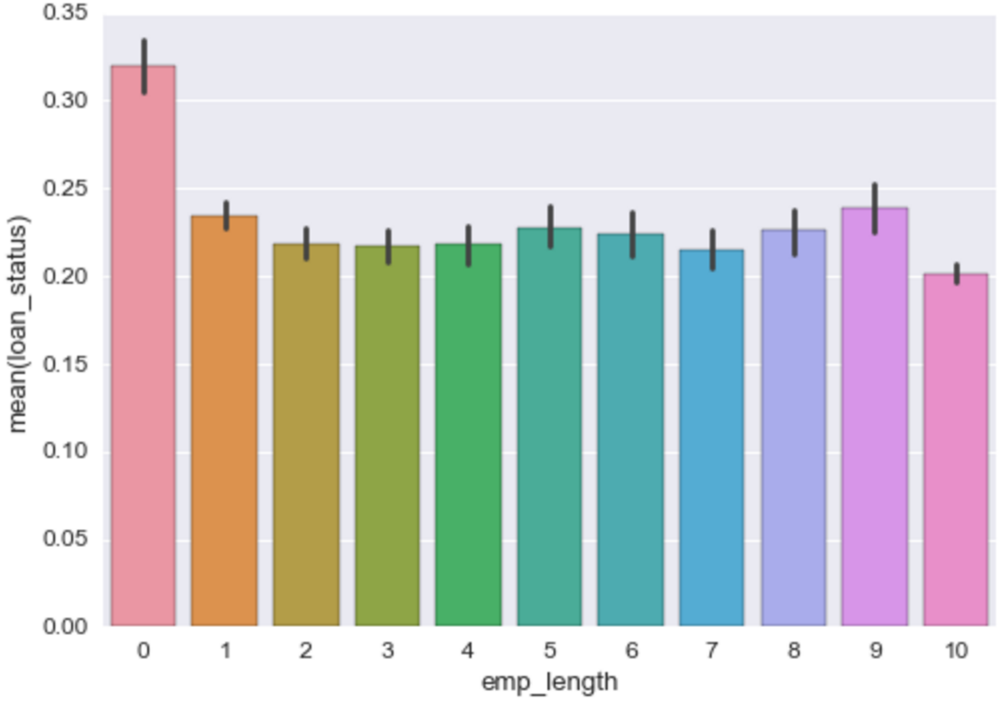
Fully Paid	119363
Current	79396
Charged Off	32634
Late (31-120 days)	2576
In Grace Period	794
Late (16-30 days)	630
Default	236

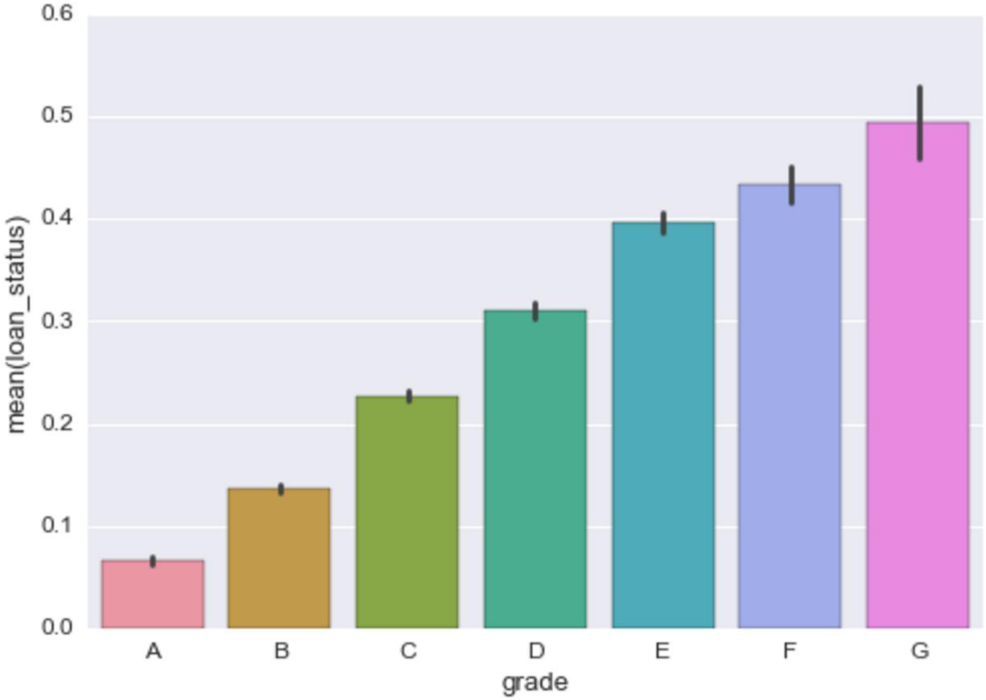
Questions:

1. Loan status varies over time
2. When to become default
3. Default rate v.s. missing payments

(http://blog.lendingrobot.com/research/predicting-the-number-of-payments-in-peer-lending/)

EDA and Visualization





Data Selection

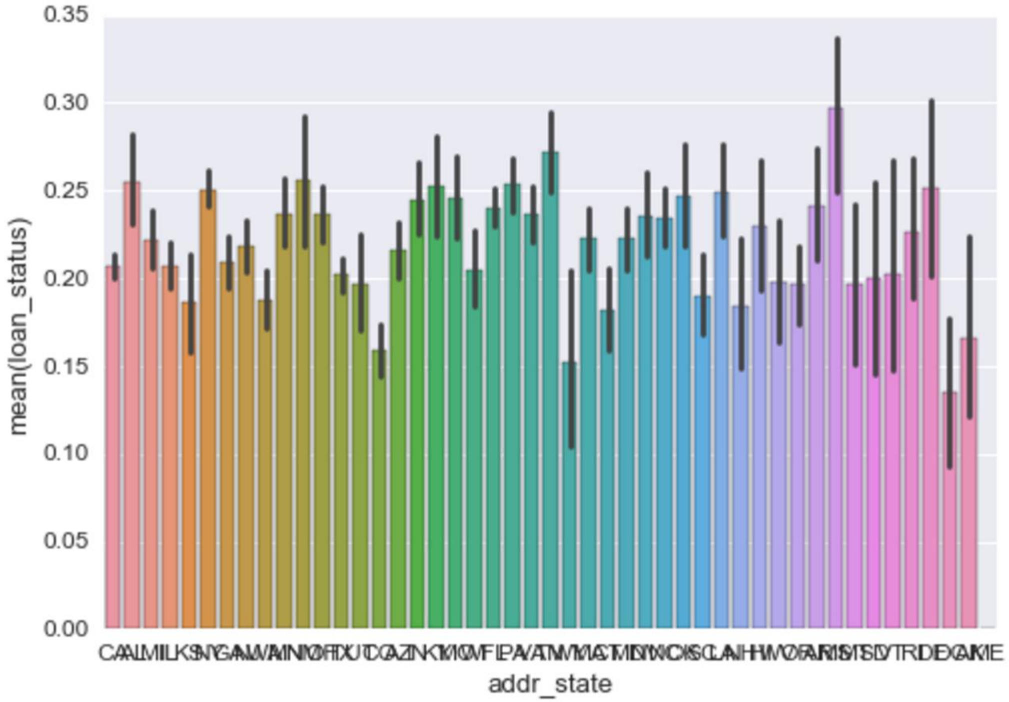
- Only 36 months loans (term=36 months)
- Binary classification: Charged off + Default →1, Fully paid → 0
- How about 2015 data instead of 2014 data?

Train/Test Split

- Out of time train/test split
- In time train/test split

Feature Engineering

1. Natural Language Processing (NLP) needed,
2. Geographical Information System (GIS) needed,



Methods to deal with Categorical Variables

1. Convert to Number
2. Dummy Coding
3. Combine Levels
4. Leave-one-out encoding

Key challenges with categorical variable

1. A categorical variable has too many levels. This pulls down performance level of the model. For example, a cat. variable “zip code” would have numerous levels.
2. A categorical variable has levels which rarely occur. Many of these levels have minimal chance of making a real impact on model fit. For example, a variable ‘disease’ might have some levels which would rarely occur.
3. There is one level which always occurs i.e. for most of the observations in data set there is only one level. Variables with such levels fail to make a positive impact on model performance due to very low variation.
4. If the categorical variable is masked, it becomes a laborious task to decipher its meaning. Such situations are commonly found in kaggle competitions.
5. You can’t fit categorical variables into a regression equation in their raw form.

Logistic regression v.s. XGBoost

- Data normalization and other preprocessing requirements
- Efficiency and performance
- How to explain your results?

→ XGBoost

- Set up model
- Tune parameters: cross validation and grid search
- Evaluation metrics: ROC, AUC
- Feature importance

Pickle

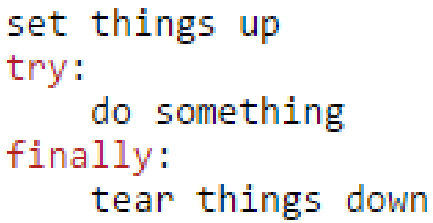
1) saving a program's state data to disk so that it can carry on where it left off when restarted (persistence)

2) sending python data over a TCP connection in a multi-core or distributed system (marshalling)

3) storing python objects in a database

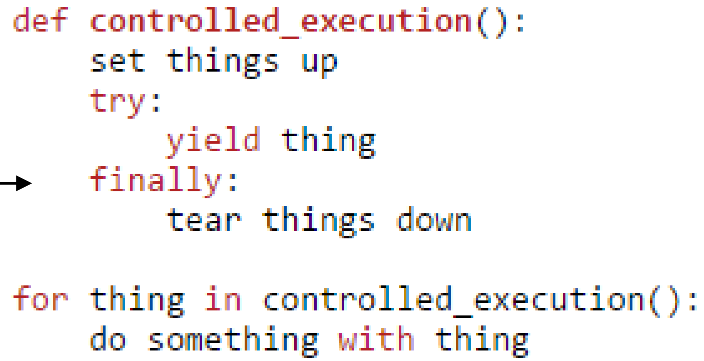
4) converting an arbitrary python object to a string so that it can be used as a dictionary key (e.g. for caching & memorization).

With As



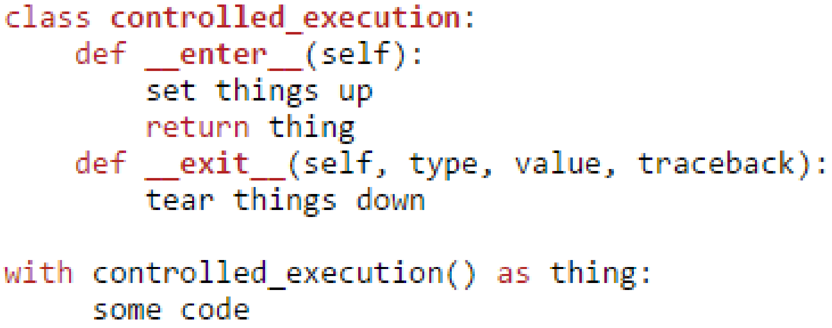
```
set things up
try:
    do something
finally:
    tear things down
```

This image cannot currently be displayed.



```
def controlled_execution():
    set things up
    try:
        yield thing
    finally:
        tear things down

for thing in controlled_execution():
    do something with thing
```



```
class controlled_execution:
    def __enter__(self):
        set things up
        return thing
    def __exit__(self, type, value, traceback):
        tear things down

with controlled_execution() as thing:
     some code
```

with open(   ) as

Home Work

- Build a model to predict the default rate of a loan
- Study Pickle and save your trained model for future usage