Default and investment Return Prediction in Online P2P Lending

Chiyuan Cheng (June 2020)
Data Science Career Track, Springboard

Topic Overview

Background

Problem

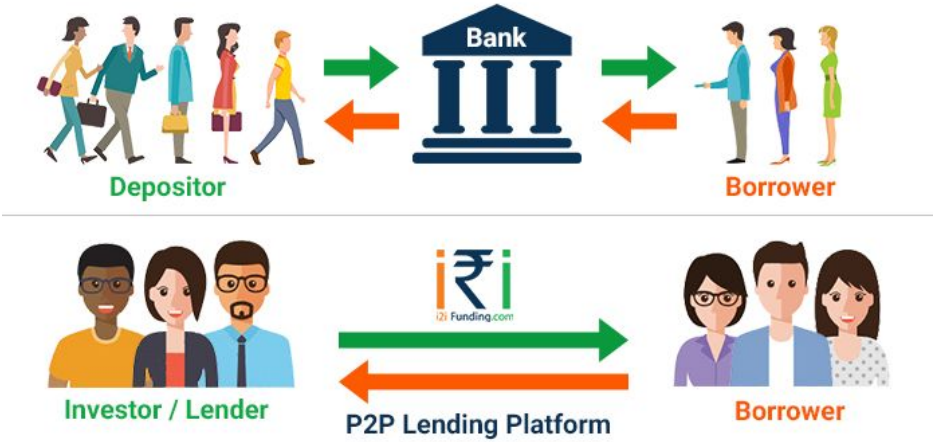
Dataset

Analysis and Result

Machine Learning

Conclusion and next steps

Background



- P2P lending is the fastest-growing investment platforms: lending money to individuals via online service with a much higher return
- Lending Club is the world's largest P2P lender
- However, P2P lending presents a higher investment risk due to the loan default.

Problems

- What are the key attributes for successful P2P loan investment?
- How can we use data to help investors to reduce their investment risk and increase their returns?
- What is the optimal return can investor expect from P2P loan investment ?

Dataset

- Lending Club provides a dataset including all loan from 2007-2019
    - 2,736,278 rows, 150+ features
- 26 features relevant to borrower's application were focused.

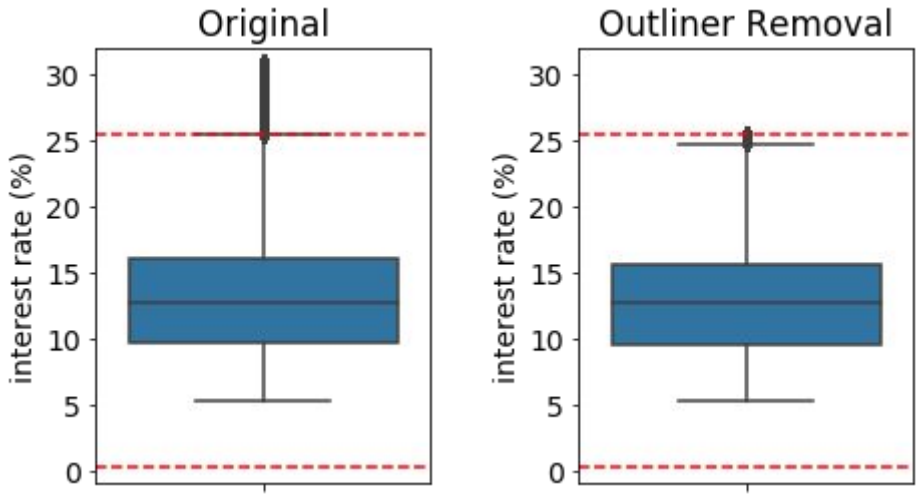
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2730228 entries, 10224583 to 158206429
Data columns (total 28 columns):
issue_d                 datetime64[ns]
earliest_cr_line        datetime64[ns]
last_pymnt_d            datetime64[ns]
last_credit_pull_d      object
int_rate                float64
revol_util              float64
term                    object
grade                   object
emp_length              object
home_ownership          object
verification_status     object
loan_status             object
purpose                 object
addr_state              object
loan_amnt               float64
funded_amnt             float64
installment             float64
annual_inc              float64
int_rate.1              float64
dti                     float64
revol_bal               float64
delinq_2yrs             float64
open_acc                float64
pub_rec                 float64
fico_range_high         float64
fico_range_low          float64
total_pymnt             float64
recoveries              float64
dtypes: datetime64[ns](3), float64(16), object(9)
memory usage: 604.1+ MB
```

Data Wrangling

- Drop the columns with more than 20% missing values

```
columns with > 20% missing values: 58 columns
----------------------------
['member_id', 'orig_projected_additional_accrued_interest',
'hardship_loan_status', 'hardship_start_date', 'hardship_en
d_date', 'payment_plan_start_date', 'hardship_length', 'har
dship_dpd', 'hardship_payoff_balance_amount', 'hardship_las
t_payment_amount', 'deferral_term', 'hardship_status', 'har
dship_reason', 'hardship_type', 'hardship_amount', 'settlem
ent_status', 'settlement_amount', 'settlement_date', 'debt_
settlement_flag_date', 'settlement_term', 'settlement_perce
ntage', 'sec_app_mths_since_last_major_derog', 'desc', 'sec
_app_revol_util', 'verification_status_joint', 'revol_bal_j
oint', 'sec_app_collections_12_mths_ex_med', 'sec_app_mort_
acc', 'sec_app_num_rev_accts', 'sec_app_open_act_il', 'sec_
app_open_acc', 'sec_app_fico_range_low', 'sec_app_inq_last_
6mths', 'sec_app_earliest_cr_line', 'sec_app_fico_range_hig
h', 'sec_app_chargeoff_within_12_mths', 'dti_joint', 'annua
l_inc_joint', 'mths_since_last_record', 'mths_since_recent_
bc_dlq', 'mths_since_last_major_derog', 'mths_since_recent_
revol_delinq', 'next_pymnt_d', 'mths_since_last_delinq', 'i
l_util', 'mths_since_rcnt_il', 'all_util', 'total_cu_tl',
'inq_last_12m', 'open_acc_6m', 'inq_fi', 'open_il_24m', 'op
en_rv_24m', 'total_bal_il', 'open_il_12m', 'max_bal_bc', 'o
pen_rv_12m', 'open_act_il']
```

- Remove outliers based on IQR

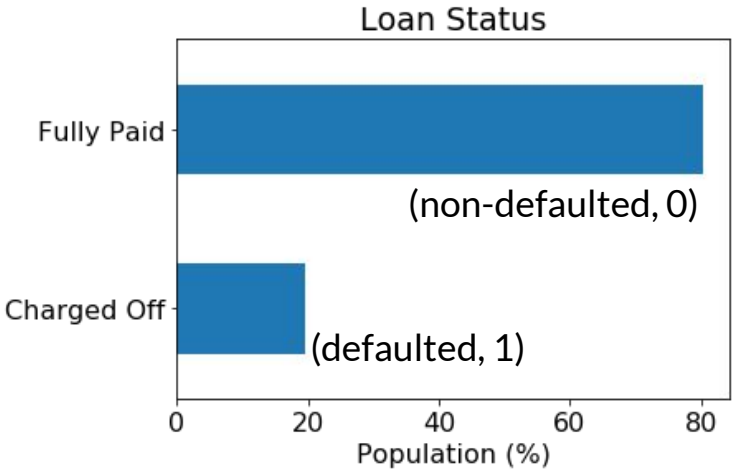


Label Defined

Predict Loan Default
(Classification Problem)

Predict Investment Return
(Regression Problem)

Binary Classification



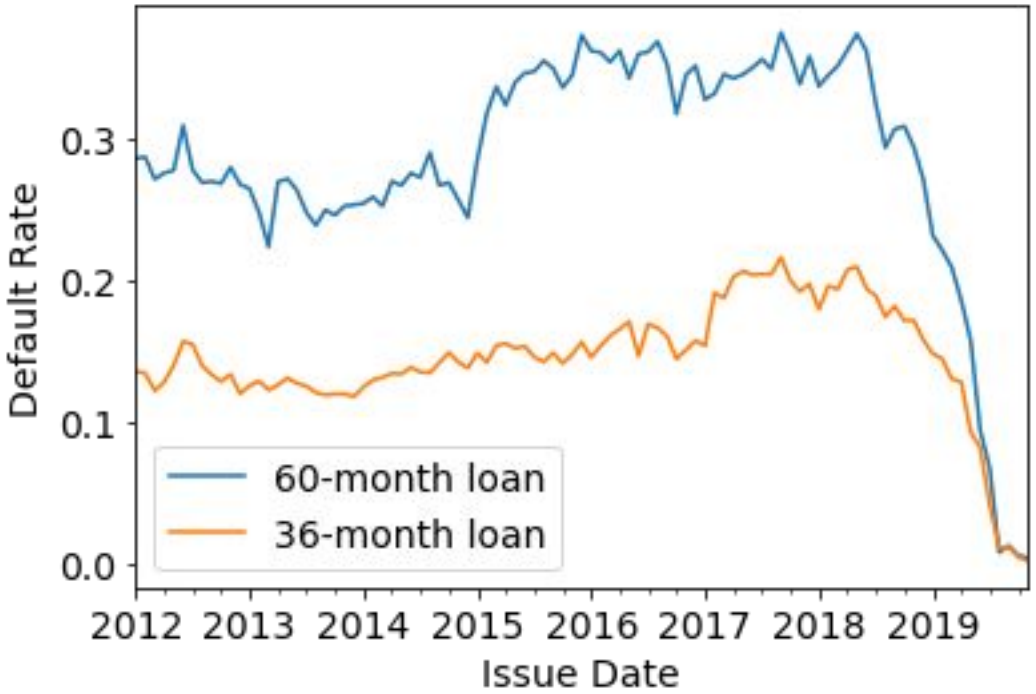
Annualized Return of Investment

$$Annualized\ ROI$$

$$= \frac{Investment\ Gain}{Investment\ Loss} - 1$$

$$= \left(\frac{Total\ Payment}{Funded\ Amount}\right)^{\frac{12}{Loan\ Period}} - 1$$

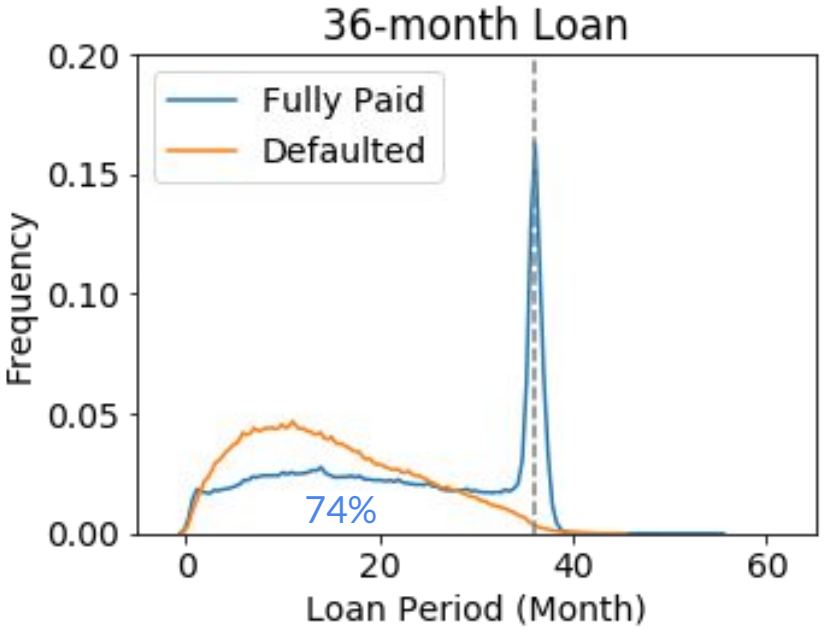
Default Rate

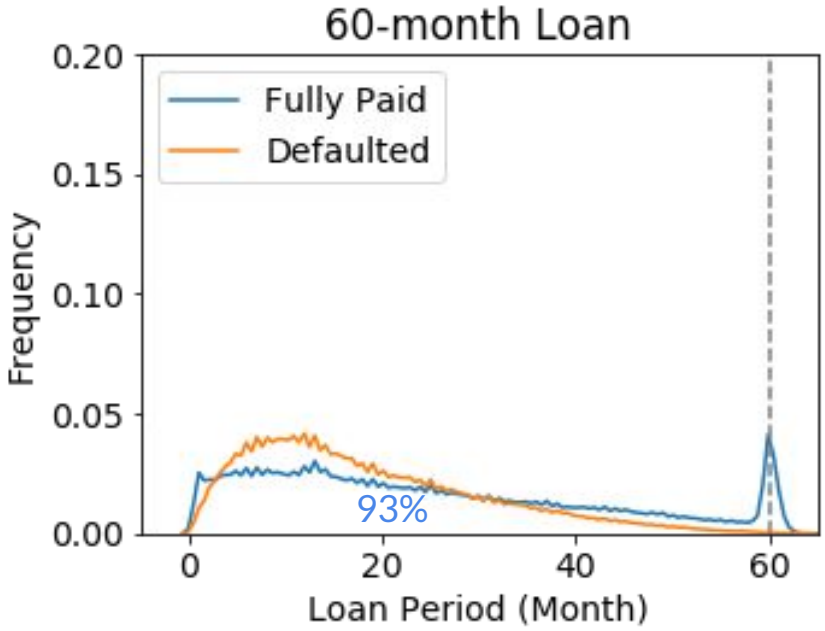


- A long-term loan has a higher risk than a short-term loan.
- The default rate drops significantly after 2018, because many loans had not yet reached full maturity.
- The loan data after 2015 should not include into the ML models to avoid bias.

Loan Period

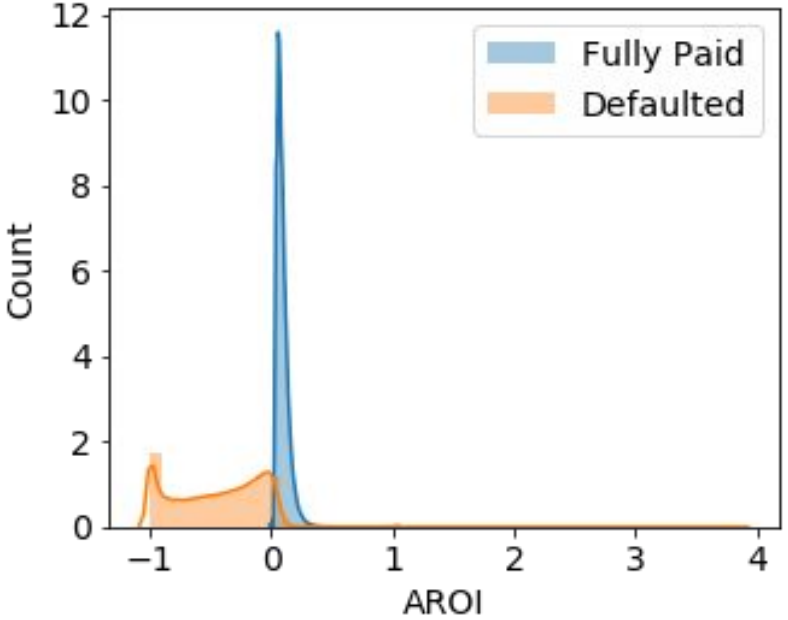
- 93% of borrowers with a longer loan term pay off their loans early, before loans expire.
- Most of the defaulted loans occurred in the 10th month.

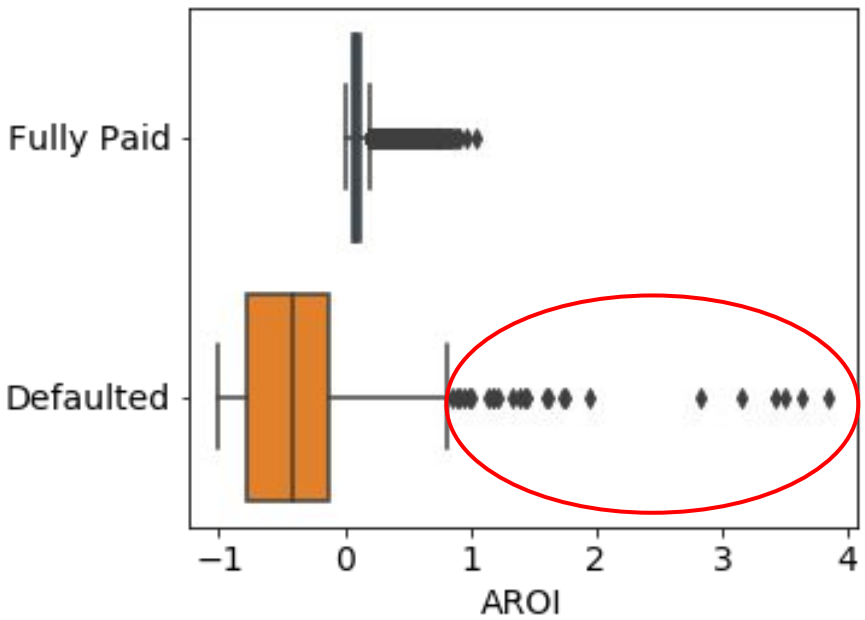




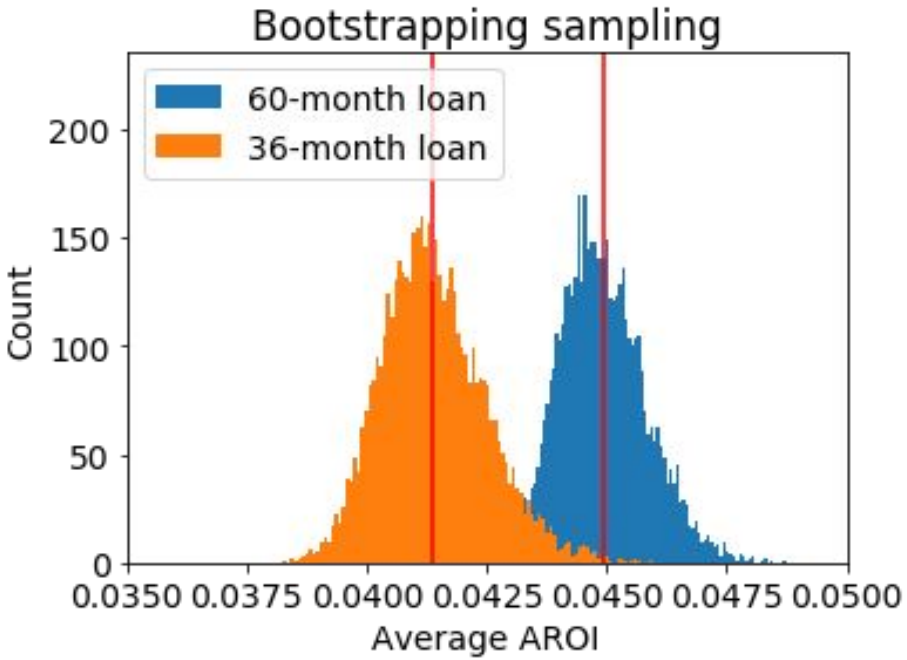
Annualized Return of Investment (AROI)

- AROI of the non-defaulted and defaulted loans have completely different distributions.
- The defaulted loan has a wider distribution, with many outliers at the higher end.
- Opportunity to obtain a much higher return from default loans using anomaly detection





Default Loan:
Loan Term Does Not Affect Return



two-sample bootstrapping hypothesis test

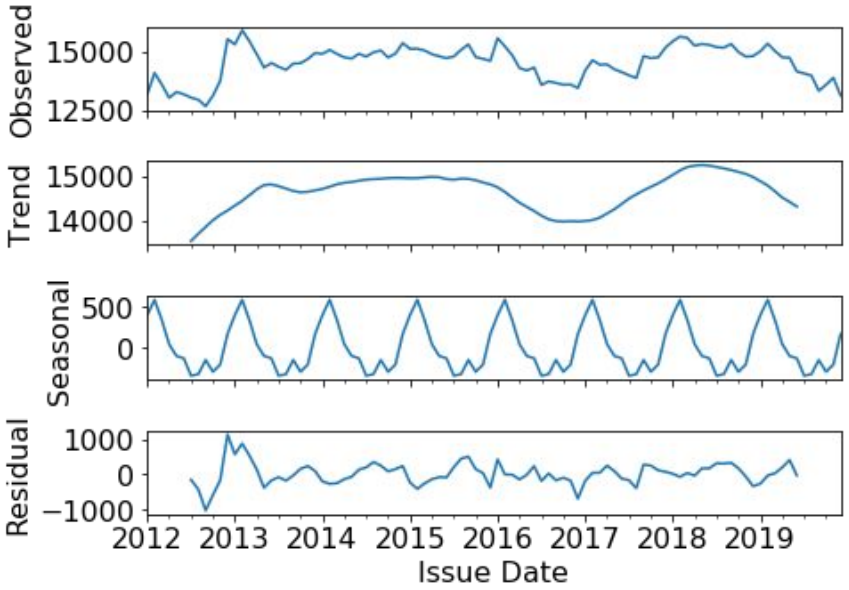
- H0: average return between 60- v.s. 36-month default loans are the same
- H1: average return between 60- v.s. 36-month default loan are different

p > 0.05 (p=0.4882, a=0.05)
→ reject H0,
→ The return for 36-month and 60-month loan are the same for the default loan.

Time-series Analysis of Mean Loan Amount

Understand loan demand with loan amount

- The loan amount relates to the load demand.
- Understand loan demand can allow the LC to have sufficient funds to supply the requested loan upon approval.



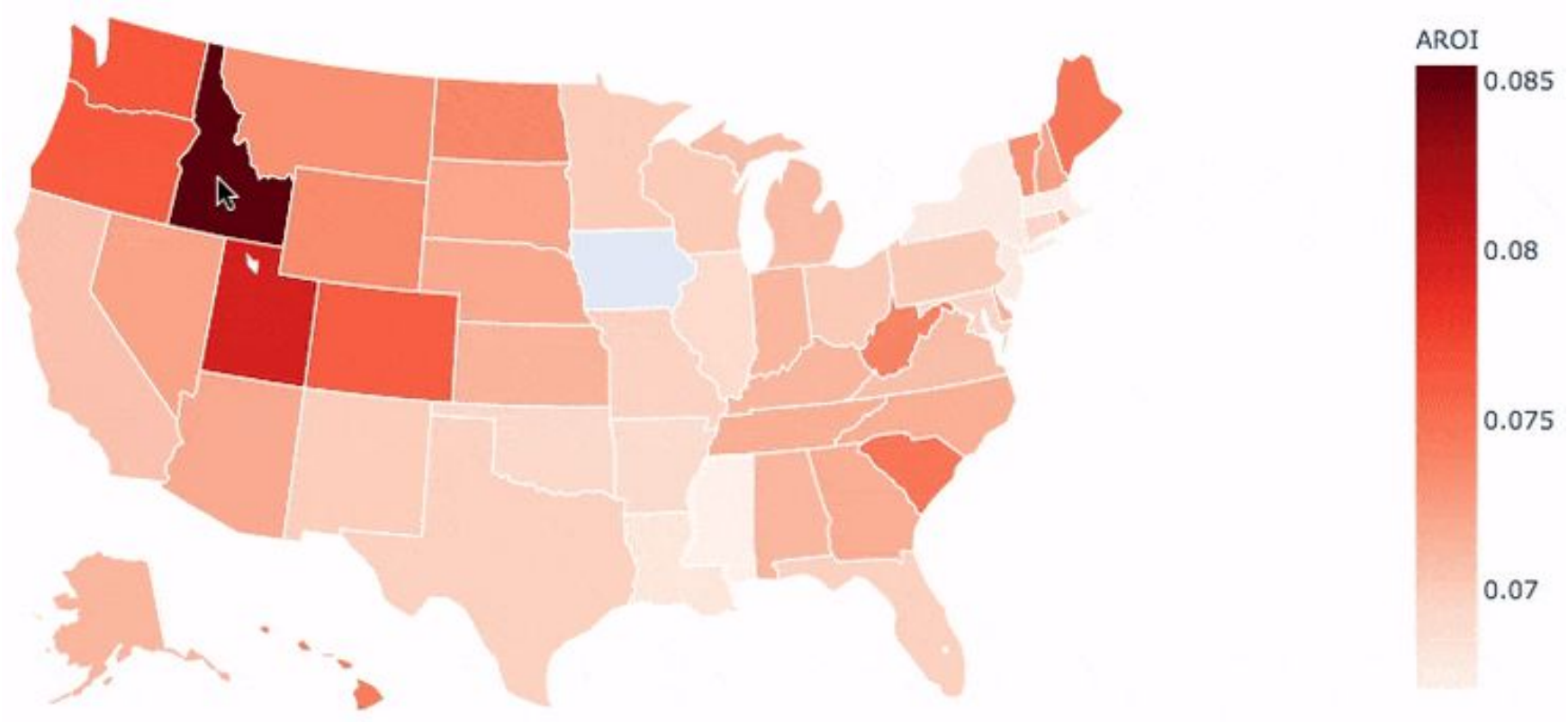
Augmented Dickey-Fuller test

- H0 : Time series is not stationary
- H1 : Time series is stationary

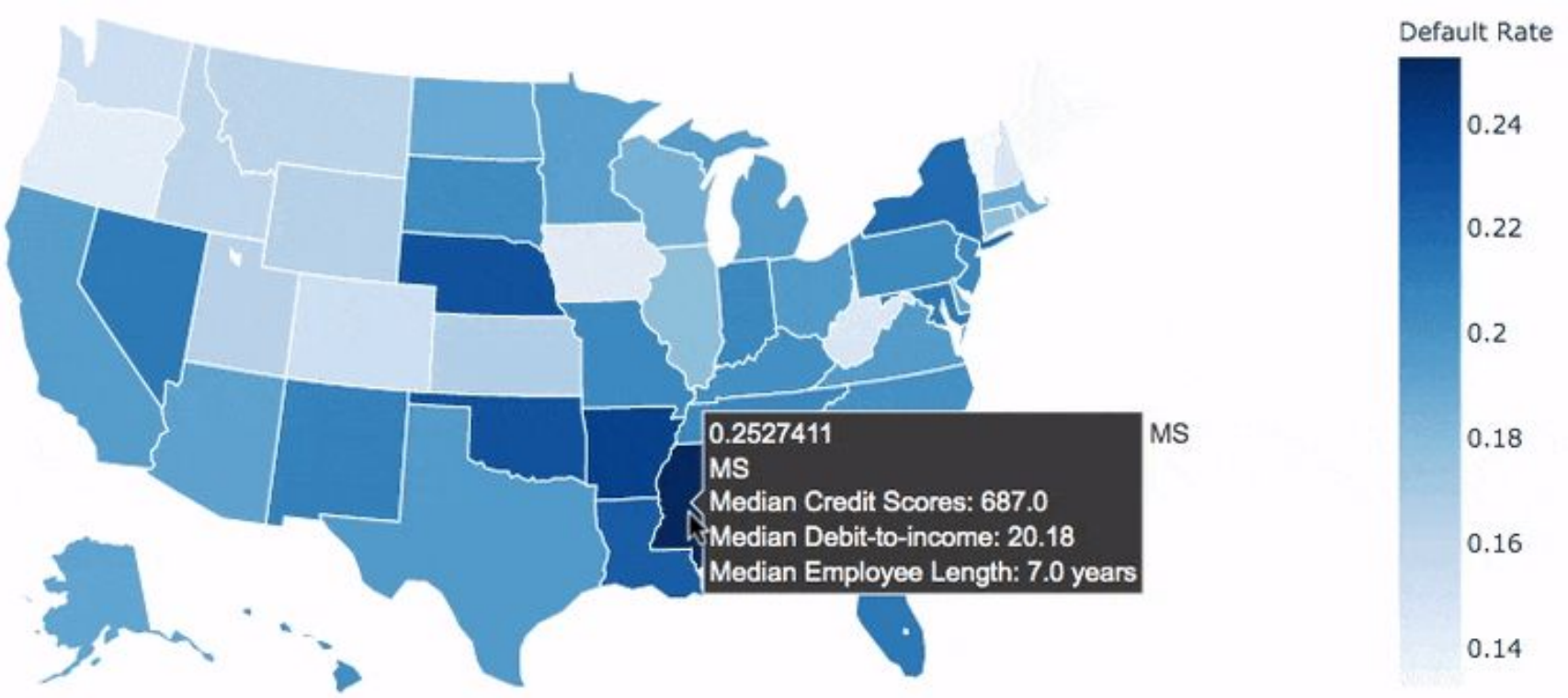
p>0.05 (p=0.078, α=0.05),
→ eject H0
→ **Time-series data is not stationary**.

- ❏ Loan amount drops between 2016-2018
- ❏ The seasonal frequency is about 1 year.

Annualized ROI by States

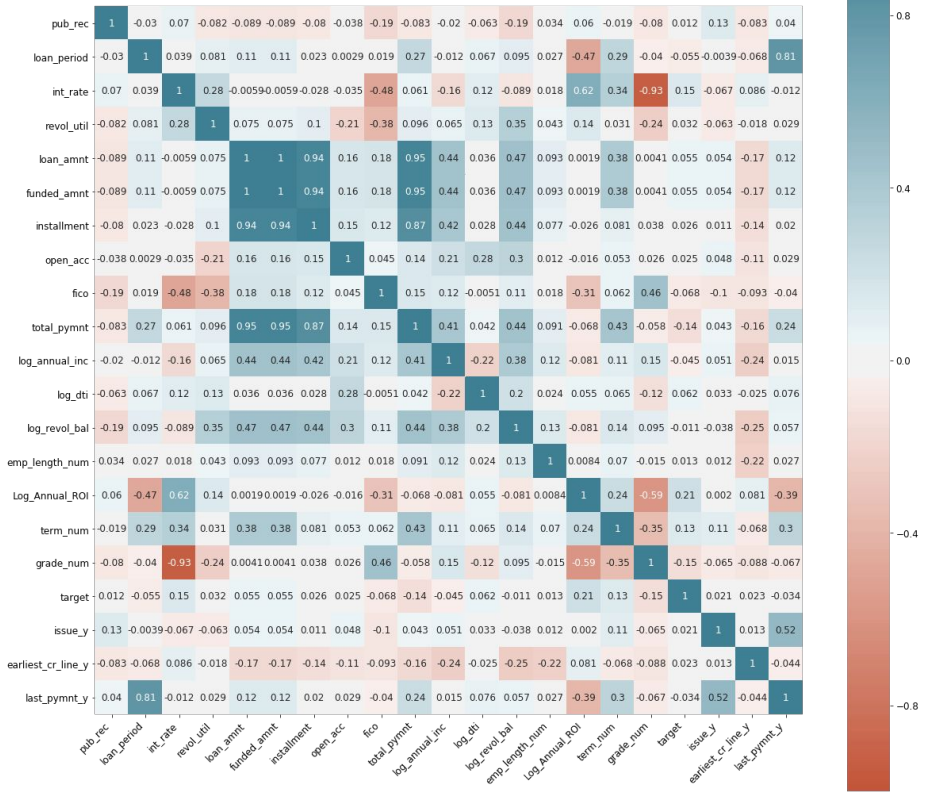


Default Rate by States



Machine Learning: Feature Selection

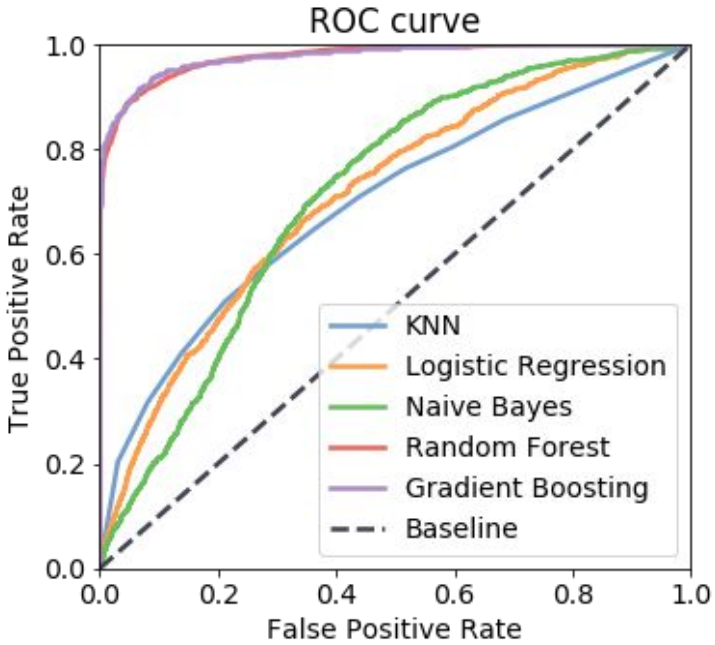
- Remove features that is uncorrelated with the target variable
  - Pearson correlation < 10%
- Remove features that are highly correlated with each other to avoid multicollinearity

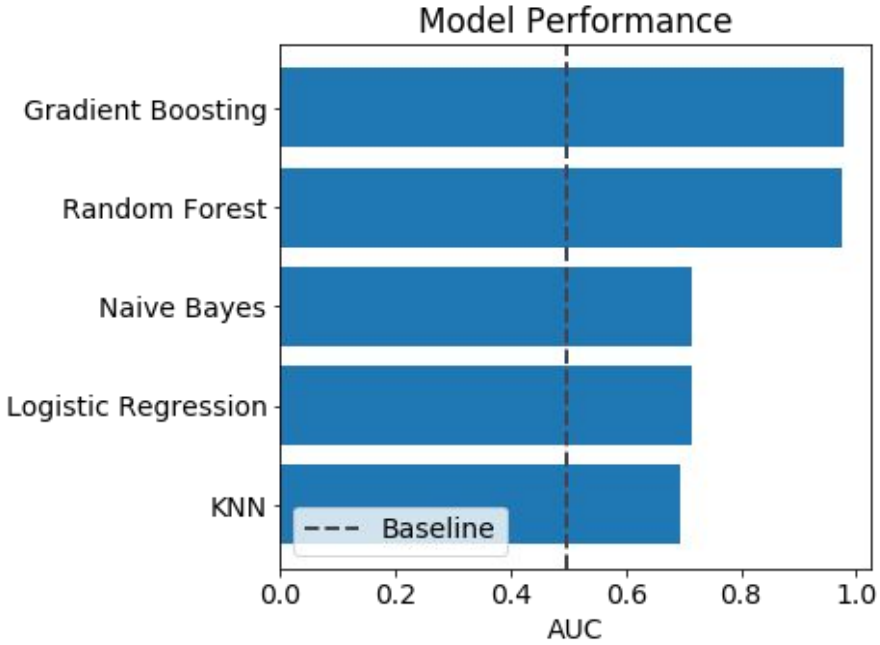


Deal with Imbalance Data:

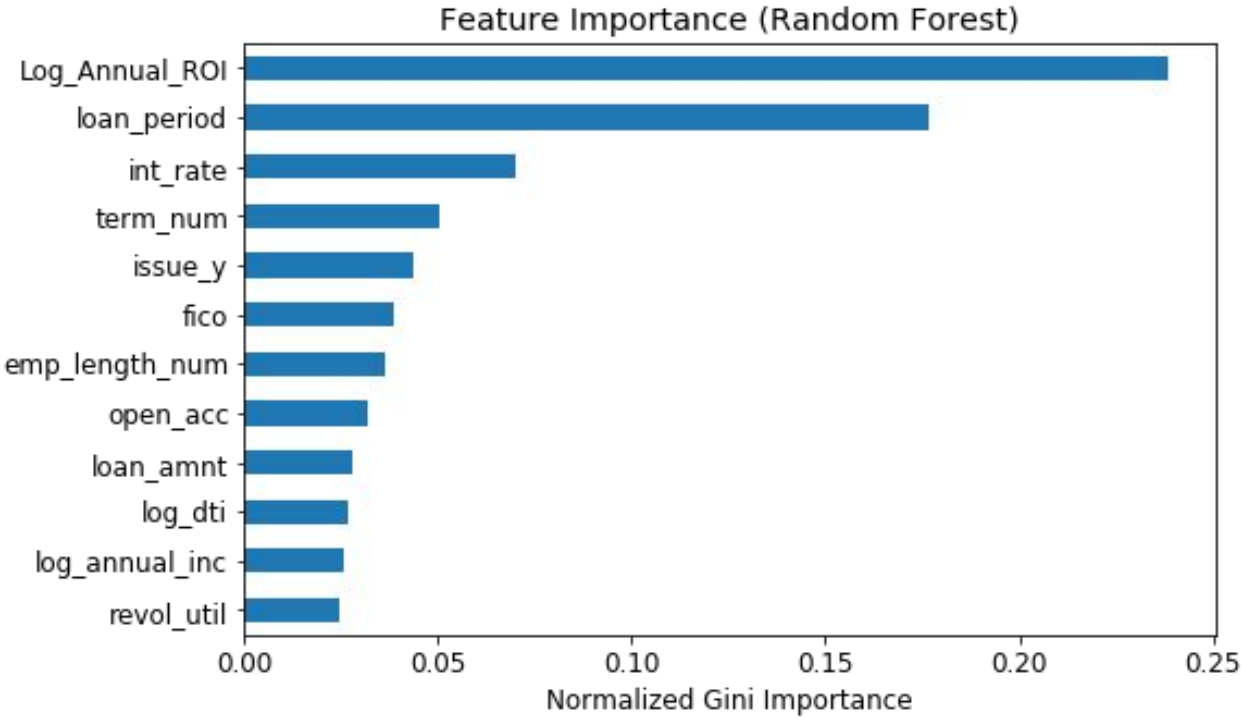
Synthetic Minority Oversampling TEchnique (SMOTE)

Machine Learning:
Classification Models to Predict Loan Default



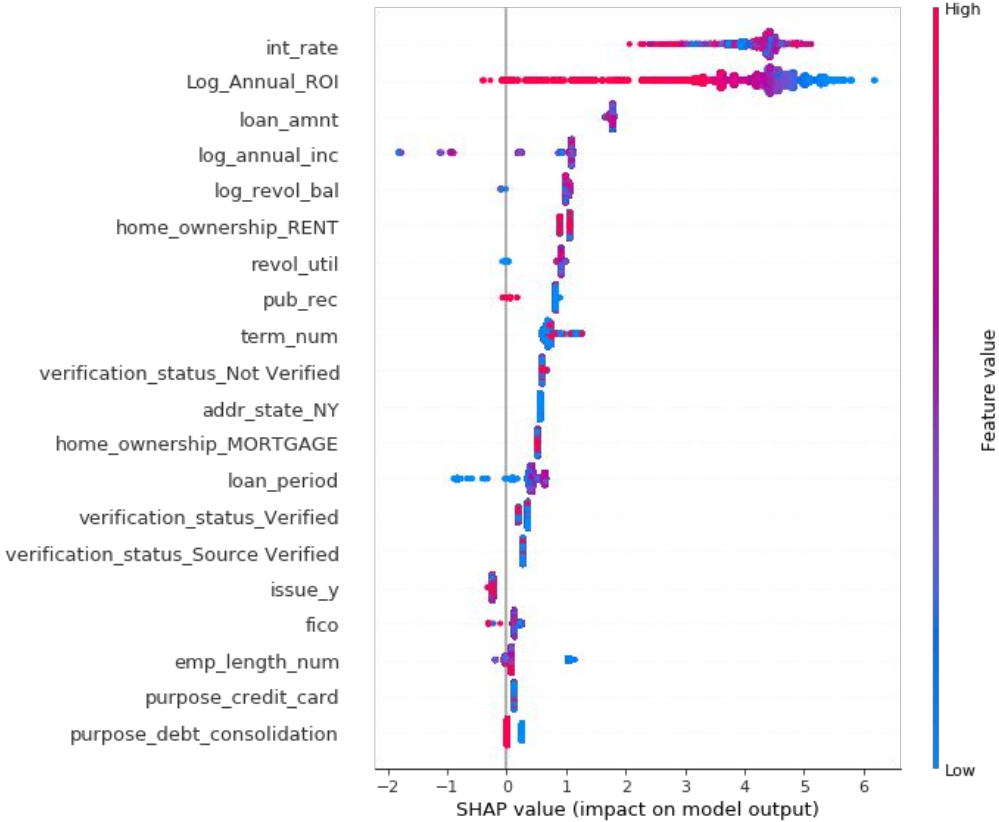


Feature Importance for Default Prediction

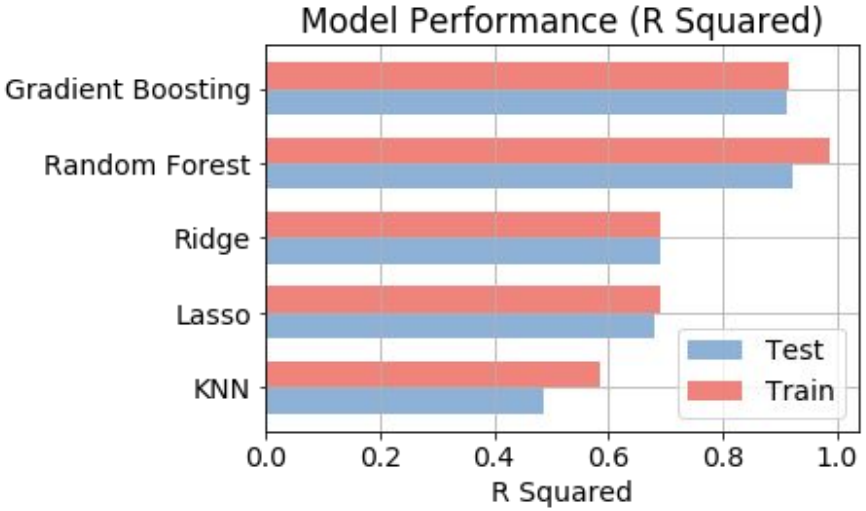


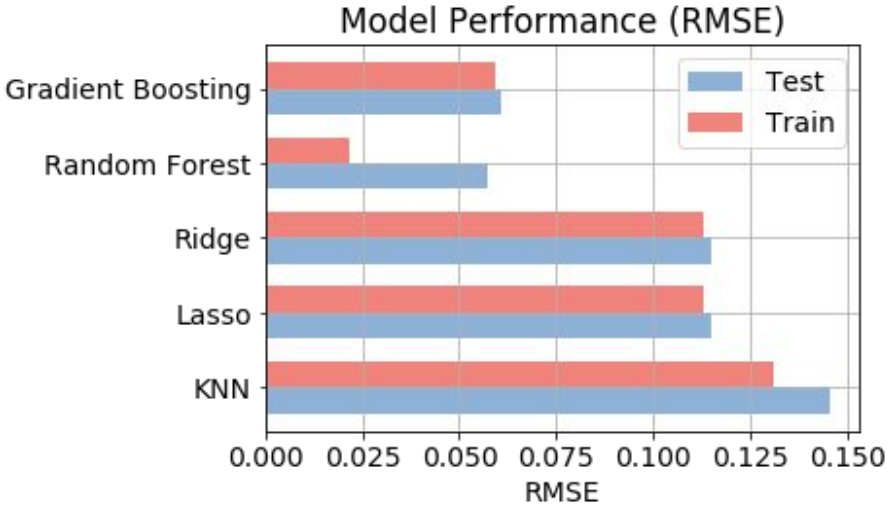
Interpretation:
Shapley Additive Explanations (SHAP) Plot

- Interest rate, AROI and loan amount are strong predictors for loan default.
- A lower AROI (blue) increases the chance to default, while a shorter loan period (blue) has a lower chance to default.

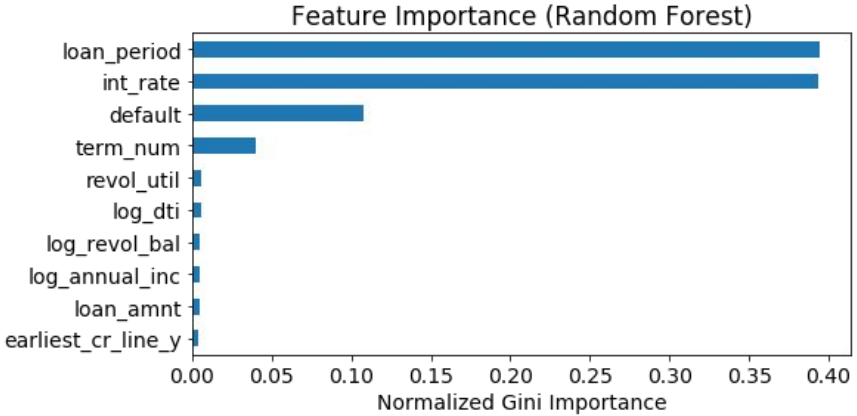


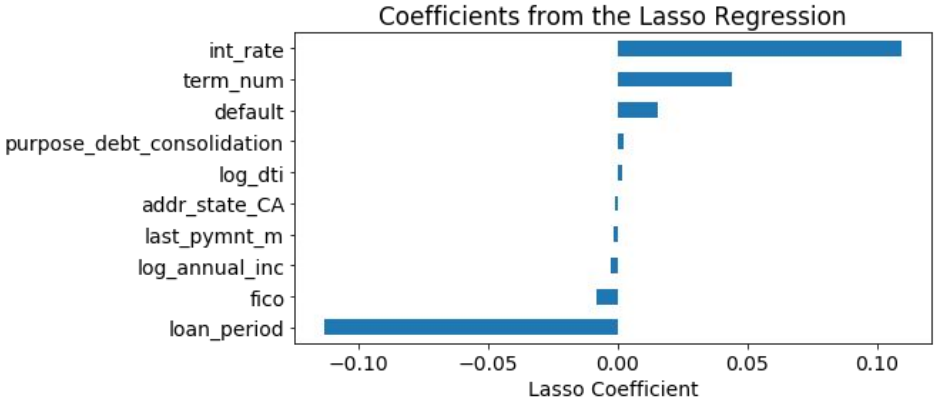
Machine Learning:
Regression Models to Predict Investment Return





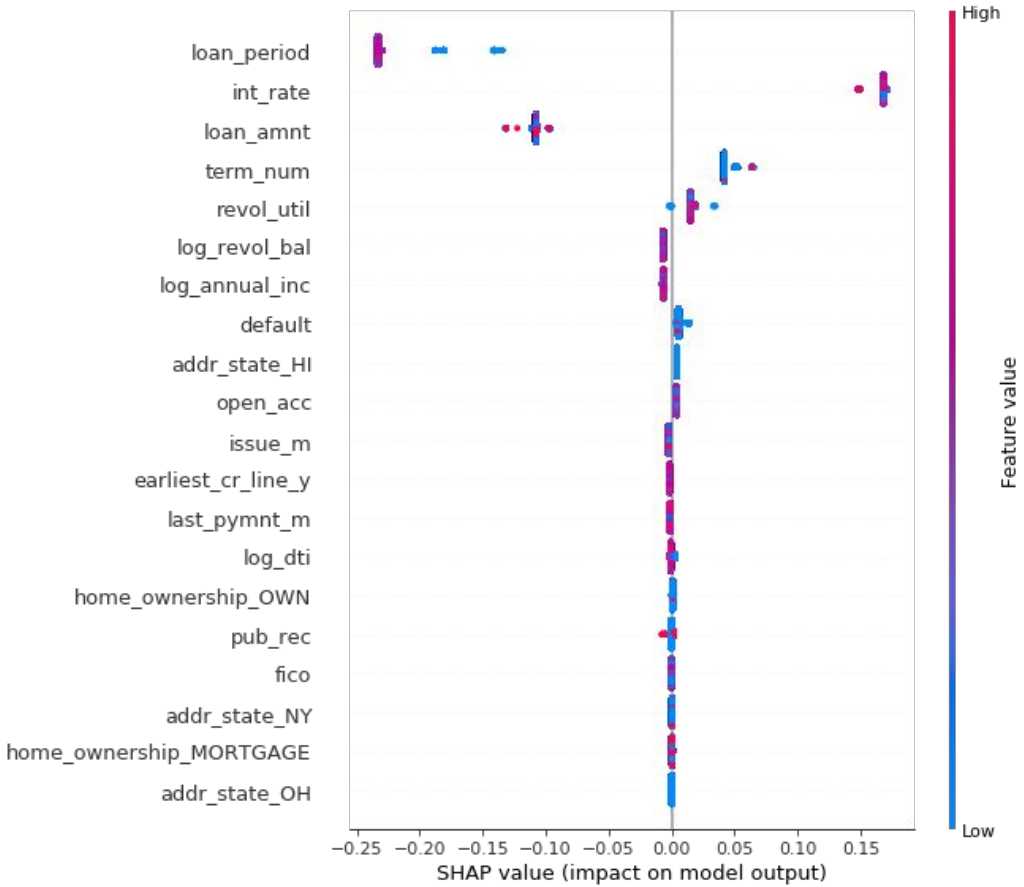
Feature Importance for Investment Return





Interpretation: SHAP Plot

- A shorter loan period, a higher interest rate and a shorter loan term can increase the AROI.
- The loan period and loan amount have a negative impact on the return, while the interest rate has a positive impact on the return.



Select Investment Strategies

Investment Strategy	Mean Return (AROI)
LC (benchmark)	5%
Randomly choose 100 loans from the predicted non-defaulted loans	9.5%
Apply regression model on the predicted non-defaulted loan, and choose 100 loans with the highest AROI	24%
Choose 100 loans with the highest AROI from predicted non-defaulted loans	35%
Apply regression model on the predicted defaulted loans, and choose 100 loans with the highest AROI	40%

Limitation and Future work

- To improve the model performance, the trained models should be validated using the loan data after 2015 or updated data, where the loans have not matured yet.
- Add external features, such as demographic data (Census), macroeconomic metrics, or include the text in the “description” by NLP.
- Deep learning or PCA can be used to predict return with less feature engineering processes.
- Build a model to predict average loan amount to forcase the future loan demand.

Conclusion

- Random Forest and Gradient Boosting models perform the best with default prediction and investment return prediction, while Gradient Boosting performs slightly better than Random Forest.
- P2P lenders can take advantage of the predictive models to help investors to make smart decisions when evaluating loan application.
- Although identifying defaulted borrowers in advance can help investors lower their investment risk, developing investment strategies to accurately assess and predict the return can help investors choose the right loans with optimal returns.