# CREDIT RISK ANALYSIS IN SOCIAL LENDING ECOSYSTEM

## Milestone 2

Aveek Choudhury - Harshita Ved - Sagar Singh - Sarang Pande

# MOTIVATION

- Peer-to-peer lending - practice of lending money to individuals (or small businesses) via online services that match anonymous lenders with borrowers, has emerged as a growing e-commerce platform in the financial marketplace bringing new economic capabilities to financing. LendingClub.

- Key issue of loan origination in the financial services industry is balancing credit risk while increasing credit accessibility. Credit risk assessed in terms of the average default rate is the rate at which borrower fails to pay back on loans.

- **Dataset:**
  - 2.25M loans between 2007 and 2018 backed by LendingClub.
  - 145 variables (36 categorical, 109 numeric) - Borrower characteristics and Loan characteristics
  - Grade (A-G) and sub-grade (1-5) reflects LendingClub's assessment of the credit risk.

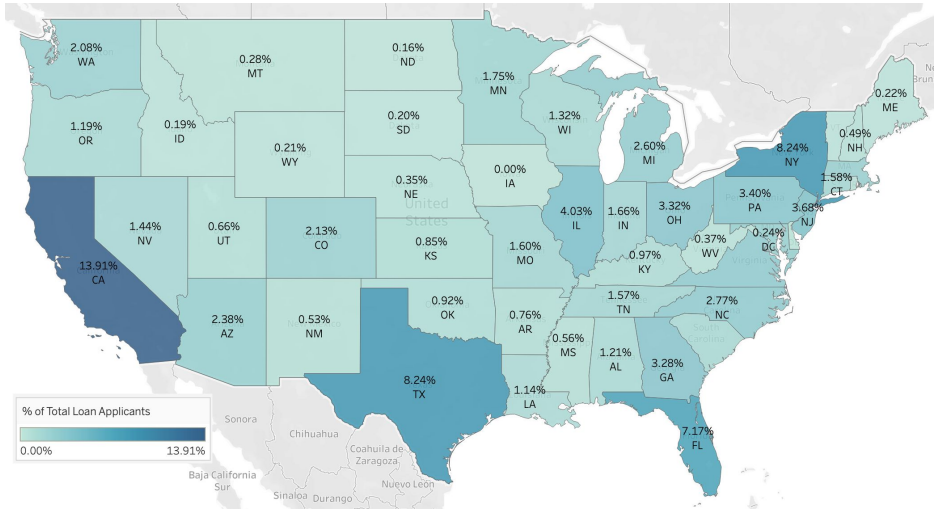
- **Areas of Interest:**
  - Perform EDA to draw conclusions based on region-wise loan targets and understand loan term distribution to help LendingClub understand specific markets for expansion and also design schemes with flexible interest rates and longer terms to attract young borrowers.
  - Explore regression techniques to identify most suited borrower characteristics driving interest rates at which a loan might be offered for a particular borrower and attempt grade classification to aid in quick loan applicant rating.
  - Use of classification techniques such as Logistic, Random Forest and SVM to analyse loan characteristics and make accurate predictions of good vs bad loan, thereby establishing LendingClub's reliability amongst its clients.

# REGION-WISE LOAN TARGETS

- California(CA) alone accounted for 13% of the total loan shares, followed by New York(9%), Texas(8%) and Florida(8%).

- These top states along with many others had only 4-5% default rate for Grade A loans however as the Grade deteriorates especially for G, surprisingly more than 50% turned out to be of bad type i.e, were either defaulted or charged-off, thus indicative of the fact that Grade plays an important role in determining Loan Quality.

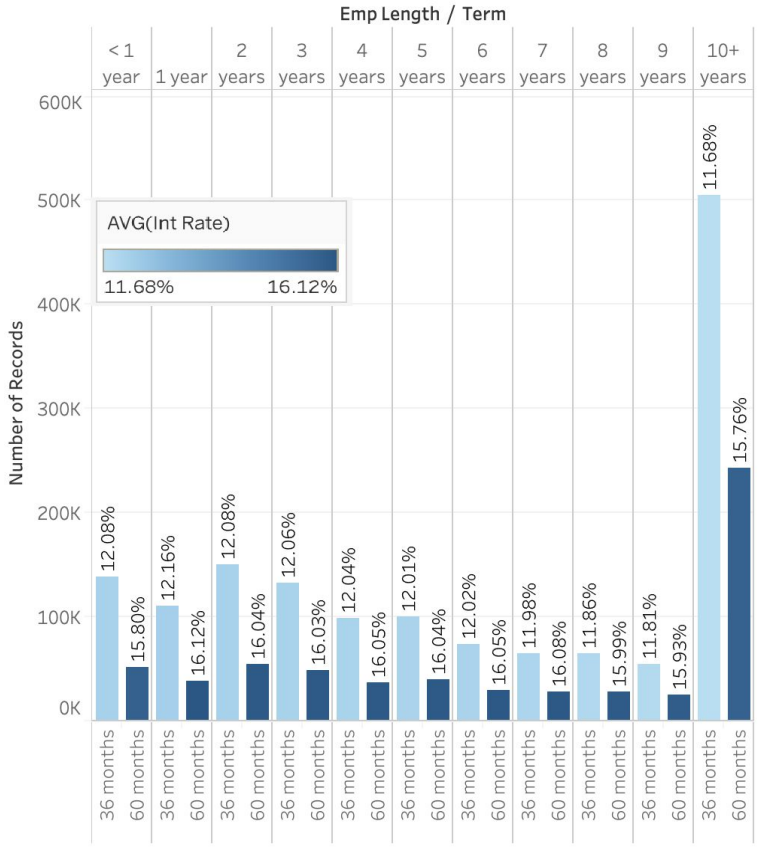
- Idaho(ID) has the highest average dti ratio of 22%, and Washington(DC) had the highest average income of 91.48K with the lowest overall dti ratio of 15%.




| Grade | Addr S.. | Loan Type | % of Total Numbe.. | Avg. Annual Inc | Avg. Int Rate | Avg. Dti |
|---|---|---|---|---|---|---|
| A | CA | Bad | 5.9893457 34% | 89.69K | 7.436756757 | 15.983061425 |
| | | Good | 94.010654266% | 96.81K | 7.124363534 | 14.144419362 |
| | TX | Bad | 6.199192322% | 90.16K | 7.397030457 | 17.579966130 |
| | | Good | 93.800807678% | 94.18K | 7.114286273 | 16.325247442 |
| | NY | Bad | 7.119850822% | 85.82K | 7.430095238 | 15.802230159 |
| | | Good | 92.880149178% | 96.21K | 7.137208736 | 14.242441592 |
| Grade | Addr S.. | Loan Type | % of Total Numbe.. | Avg. Annual Inc | Avg. Int Rate | Avg. Dti |
| G | CA | Bad | 49.180327869% | 79.52K | 27.917133333 | 20.476137124 |
| | | Good | 50.819672131% | 91.80K | 27.267096774 | 18.392080645 |
| | NY | Bad | 52.201257862% | 69.61K | 27.982216867 | 19.902 |
| | | Good | 47.798742138% | 88.71K | 27.209710526 | 17.581157895 |
| | TX | Bad | 47.690217391% | 76.89K | 28.047094017 | 23.827236467 |
| | | Good | 52.309782609% | 88.74K | 27.256857143 | 23.833948052 |

# YEARS OF EMPLOYMENT AND LOAN TERMS

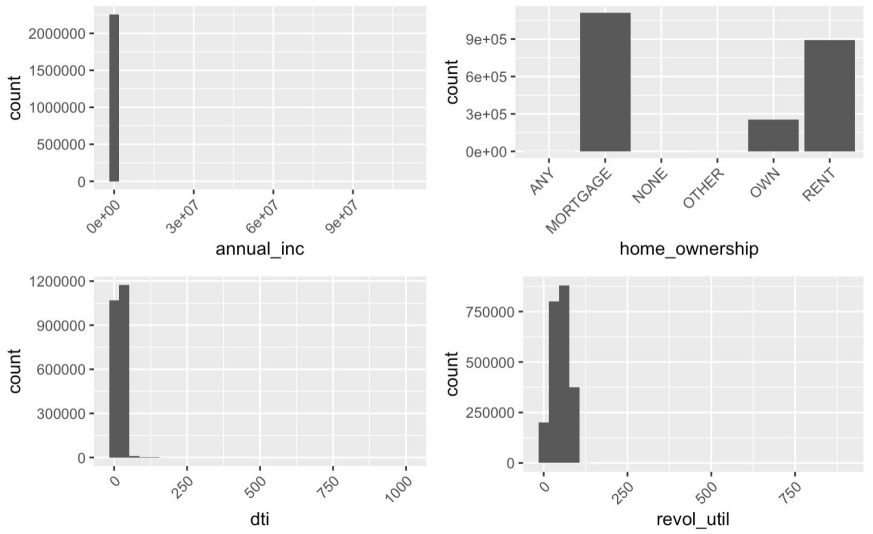
- Striking variation of 3-4% in interest rates is observed across 36-month and 60-month loan tenure for all applicants, and irrespective of work experience everyone preferred taking loans with lesser installment periods.
- Further, no significant difference in terms of interest rate is observed across employment years, except for a slight decrease in the interest rate for 10+ experience level.
- Hence this was a failure in hypothesis that young people being new in the job will tend to take loans of higher installment periods and so will end up paying more interest on average.
- This way focusing on business expansion plans based on attracting young customers will not be much impactful for LendingClub.



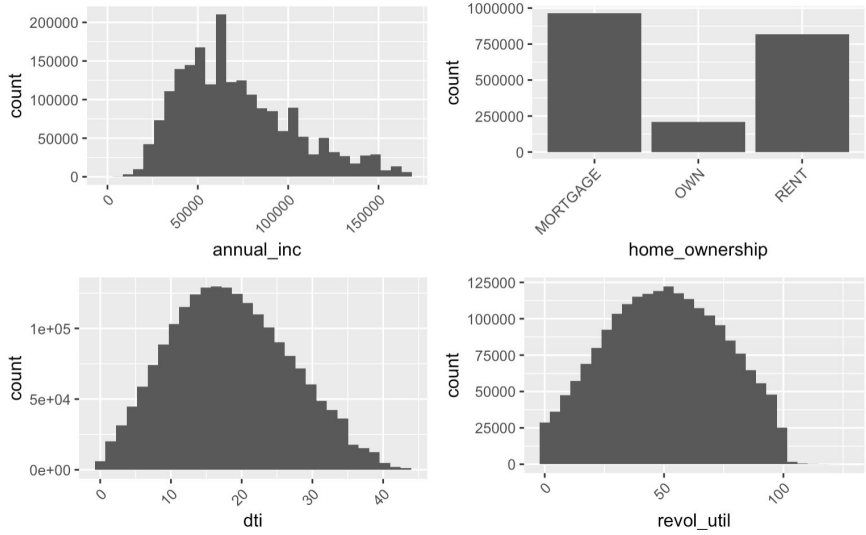
# INTEREST RATE ANALYTICS

- We carefully select 20 out of 145 total characteristics from the dataset which are provided at the time of application submission like loan amount, loan term, home ownership etc.
- To get the best result out of our dataset, we started investigating individual features to detect anomalies like outliers, high leverage points, imbalanced features while still retaining most of the observations.
- Post all anomaly removal, our analysis found that the numerical features were skewed towards the left side of the distribution.
- This lead us to perform normalization on features like annual income and dti using Min Max Scaler, to get the features in the range of 0 to 1. This ensured that our algorithms would treat each feature equally.

Pre

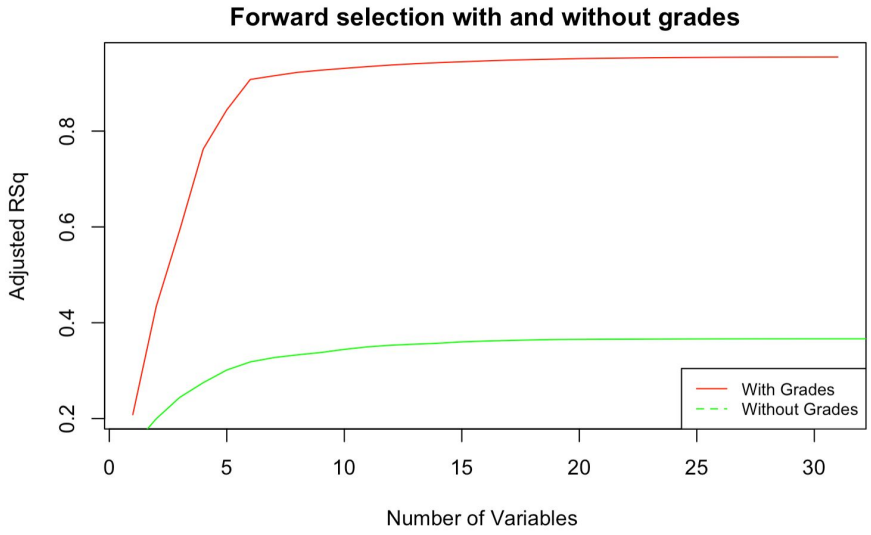


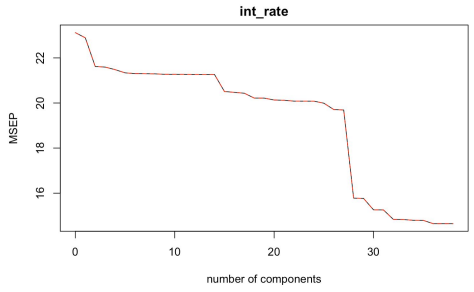

Post



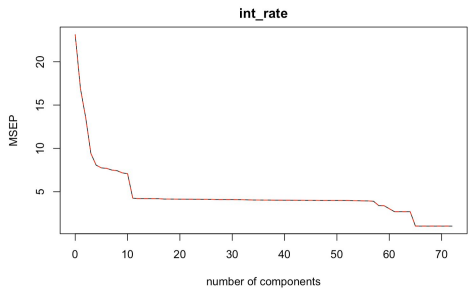
# FEATURE SELECTION

- With 18 features, we started with forward selection on our dataset.
- For lasso regression, since it penalizes the coefficients of the variables, it indicated the importance of variables like term, annual income, public record etc.
- Principal component regression showed that 81% of the variance is explained by only 11 components with grades however it took 29 components without them.
- As Loan Grade is affected by variety of latent factors like credit history and fico scores our feature selection results gave us an opinion that our data of independent variables doesn't provide a good fit to determine interest rates. Thus we move from prediction of interest rates to classification of grades.




PCR Without Grade



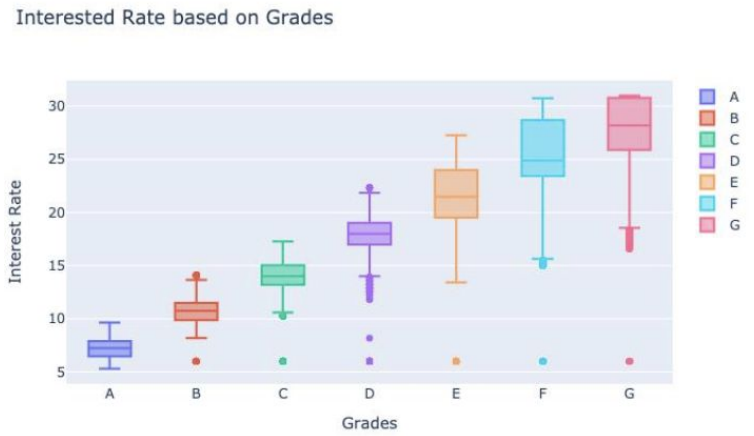

PCR With Grade

# GRADE CLASSIFICATION

- Starting with logistic regression, since this is a multi class classification, we checked for imbalanced classes.
- Next we performed downsampling to have a balanced dataset, which resulted in 78k observations. However, this amounted to only 4% of our data.
- So we decided to upssample grade G observations while at the same time downsample grade A,B,C,D,E observations to create a dataset containing 15% of the data.
- Comparing the results from both the dataset, our downsampled data performed better than the results for upsampled data.
- In absence of latent credit check factors, our grade classification model on an overall basis, performs poorly on certain grades and so making accurate assumptions is not feasible with available features.

Overall Distribution

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 372348 | 585414 | 576607 | 287344 | 120893 | 37531 | 10817 |



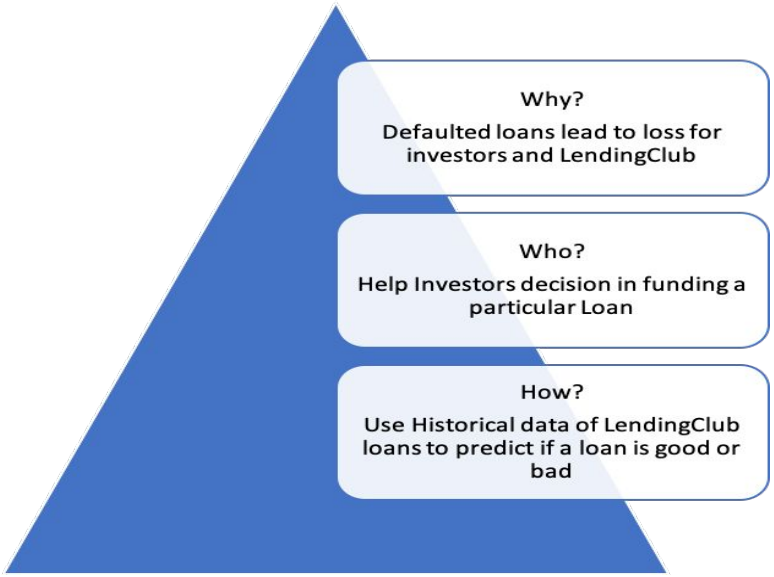

Downsampled Data Results

| predicted_class | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 2400 | 1219 | 318 | 27 | 4 | 2 | 2 |
| B | 282 | 1325 | 740 | 15 | 2 | 0 | 0 |
| C | 0 | 135 | 1200 | 399 | 32 | 3 | 0 |
| D | 0 | 3 | 418 | 1260 | 411 | 31 | 22 |
| E | 0 | 0 | 41 | 973 | 1392 | 489 | 91 |
| F | 0 | 0 | 0 | 73 | 739 | 1252 | 649 |
| G | 0 | 0 | 0 | 1 | 142 | 981 | 1857 |

Upsampled Data Results

| predicted_class | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 7715 | 5309 | 2244 | 371 | 60 | 15 | 13 |
| B | 1676 | 2185 | 1218 | 147 | 13 | 4 | 7 |
| C | 154 | 1849 | 5038 | 2613 | 438 | 39 | 20 |
| D | 1 | 38 | 642 | 2676 | 1155 | 139 | 62 |
| E | 0 | 0 | 151 | 3082 | 4224 | 1598 | 323 |
| F | 0 | 0 | 2 | 397 | 2126 | 2867 | 1359 |
| G | 0 | 0 | 0 | 57 | 1367 | 4677 | 7609 |

# LOAN RISK MODEL - HELPING INVESTORS DECIDE

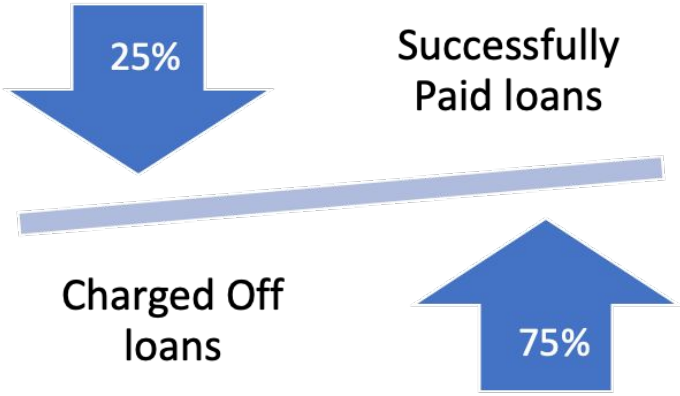



## Initial Data Processing

- Raw feature space – 145
- Remove features with > 90% missing data
- Subset Completed Loans – Fully Paid (Good), Charged Off (Bad)
- Remove all rows with NA values
- Remove features with negligible or no variability
- Refined data dimensions – 8598 observations x 95 features

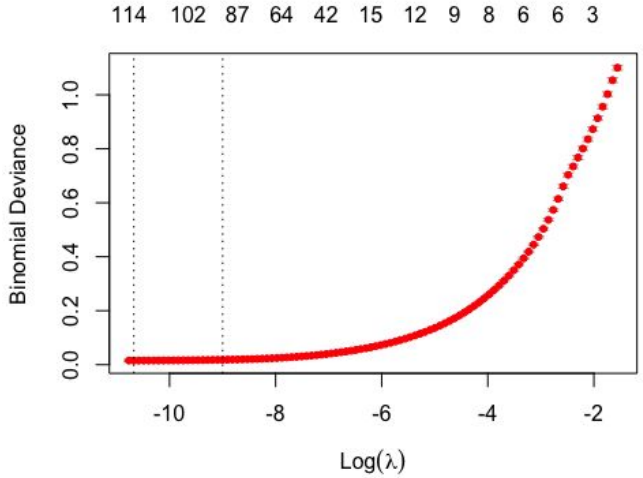
Data Transformation -

- 14 factor variables, 81 numerical variables
- Numerical Variables like income amount, loan amount etc. normalized
- Factor variables converted to numerical one-hot encoded features




Data Imbalance for good vs bad loans

# FEATURE SELECTION & MODELING - IMBALANCED DATASET




Penalized Logistic Regression (Lasso)

- Dimensionality reduction performed using L1 regularized logistic regression (cross validated lambda -
- 114 coefficients selected from 213 features (after one-hot encoding)
- High overall accuracy and sensitivity (for minority class) on unseen test set - >98%

| | predicted.classes | |
|---|---|---|
| Y_test | Charged Off | Fully Paid |
| Charged Off | 441 | 3 |
| Fully Paid | 0 | 1276 |

Test Set Confusion Matrix (L1 regularized Logistic Regression)

| | pred.rf.orig | |
|---|---|---|
| | Charged Off | Fully Paid |
| Charged Off | 429 | 15 |
| Fully Paid | 0 | 1276 |

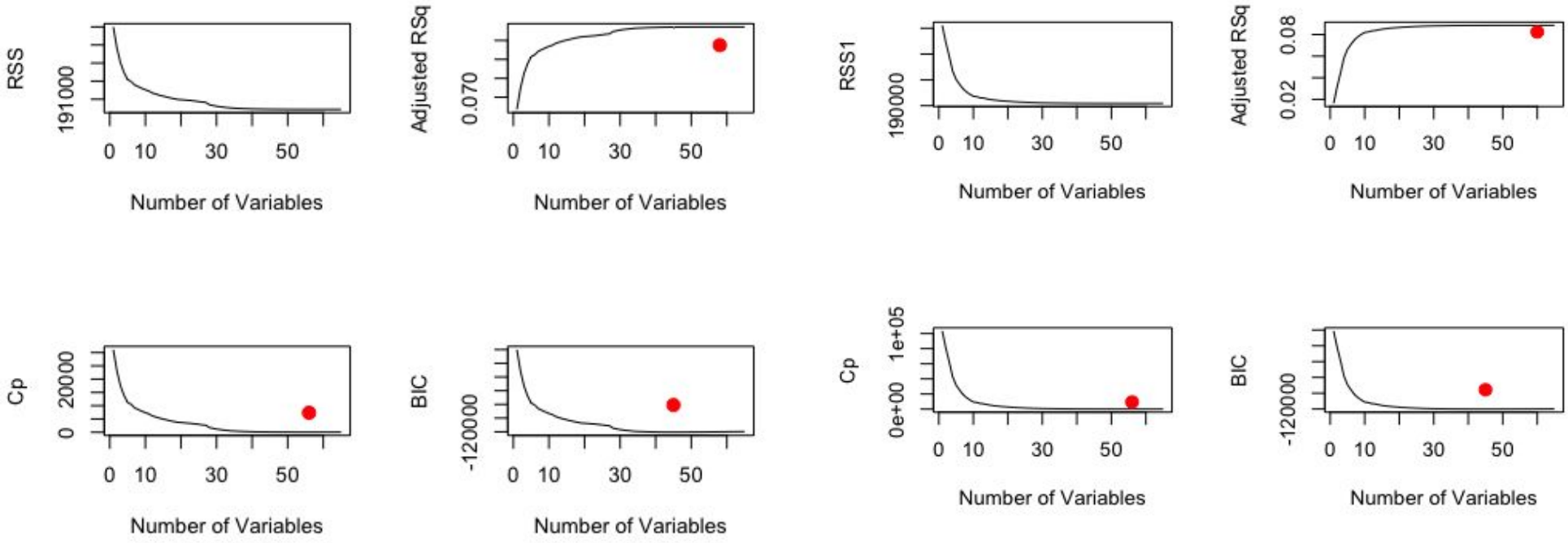
Test Set Confusion Matrix - Random Forest

- Random Forest algorithm fitted using imbalanced data - high overall accuracy and sensitivity (> 96%)
- Variable selection - top 25 variables selected based on Gini Impurity - no significant change in classifier metrics

**Failures -**

- Best subset selection failed due to computational limitations (Exhaustive search leads to vector memory overflow)
- L1 regularization preferred over backward and forward subset selection as binary classification task - cannot interpret RMSE, R-squared

# SUBSET SELECTION FOR DIMENSIONALITY REDUCTION




**Forward selection for loan_status**

**Backward selection for loan_status**

Variables selected by subset selection methods appear to be not correct according to business processes laid by LendingClub

# RESAMPLING & MODELING PERFORMANCE

Results of various resampling techniques in conjunction with Random Forest & Logistic Regression algorithms presented below -

| Oversampling (bootstrap) | | Undersampling | | Mixed-sampling | | ROSE sampling | | SMOTE | |
|---|---|---|---|---|---|---|---|---|---|
| **Logistc Reg. (L1)** | **Random Forest** | **Logistc Reg. (L1)** | **Random Forest** | **Logistc Reg. (L1)** | **Random Forest** | **Logistc Reg. (L1)** | **Random Forest** | **Logistc Reg. (L1)** | **Random Forest** |
| 98.6% | 97.9% | 99.1% | 98.6% | 98.8% | 98.4% | 90.1% | 62.1% | 98.8% | 97.5% |

- The Random Forest algorithm shows equivalent metrics with all predictors as well as top 25 features selected by gini impurity
- The high accuracy and sensitivity across all models indicates a “too-perfect” scenario
- On analyzing lending club business and domain knowledge, not all variables in the dataset will be present at the start of the process.
- The feature set needs to be reduced to a set of variable that are available at the start of business process

# RESULTS

- This project applies evaluation of classification models in the context of LendingClub's P2P lending business. We find significant opportunity for LendingClub to increase their profitability with the methods we lay out, and find our random forest model to be the most profitable choice combined with SMOTE resampling technique.

- The two main excellent failures for us are
    - In case of loan classification, balanced and unbalanced dataset performed equally well which begs researching on optimal tuning for hyper parameters or trying more complex approaches
    - Since grade classification and interest rates depend on lot many latent factors such as credit scores, debts across multiple assets, we were unable to define optimal technique for grade and interest rate prediction using the given feature set

## Future Scope:

- Regarding the grade classification problem, we plan to apply techniques like ROSE sampling and SMOTE sampling to balance the classes. Post balancing, we intend to apply support vector machines and random forest. Finally, we would conclude by comparing all the models using one vs all technique to determine the most efficient model.

- Regarding the loan classification problem, we intend to put more effort into feature selection to further increase our accuracy based on business specifications from Lending club.