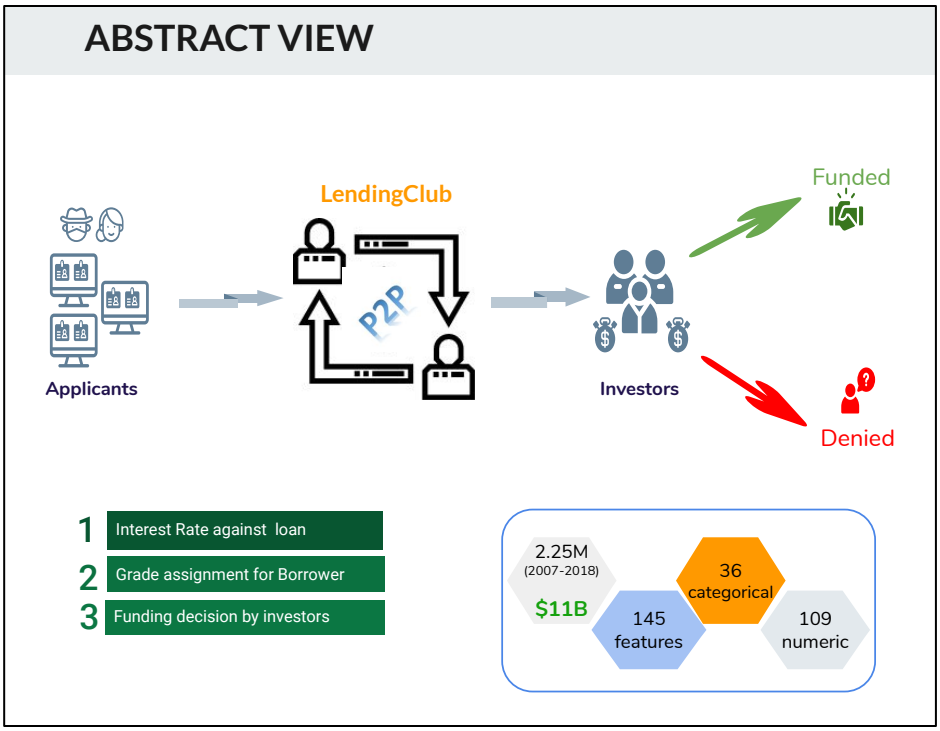
CREDIT RISK ANALYSIS IN SOCIAL LENDING ECOSYSTEM

PROJECT REPORT

Aveek Choudhury - Harshita Ved - Sagar Singh - Sarang Pande



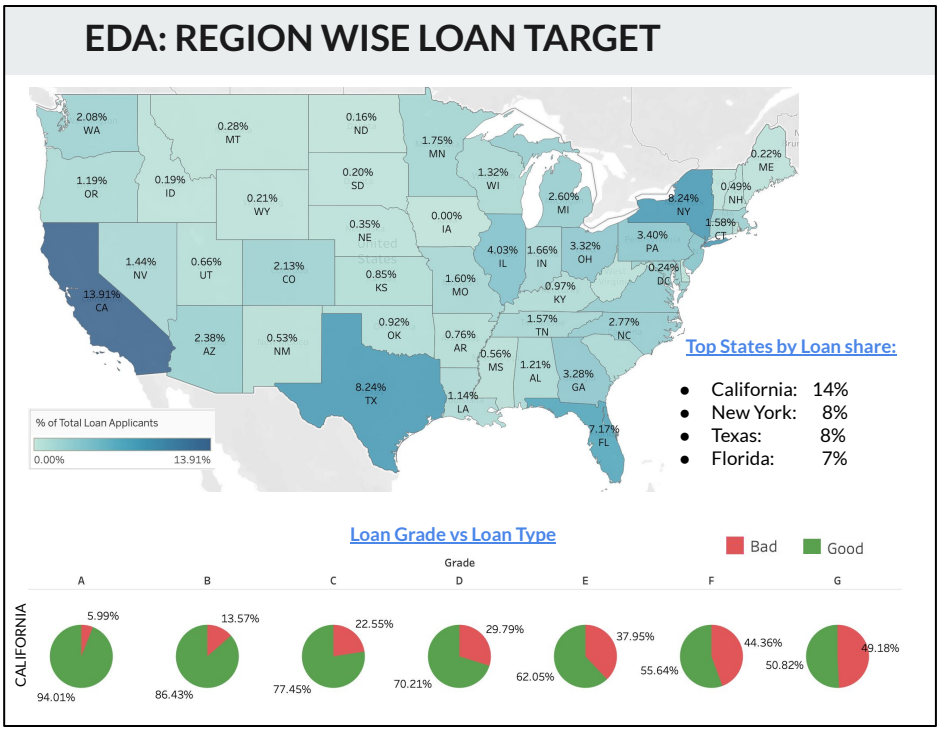
With the emergence of online communities in the past decade, the popularity of social or peer-to-peer(P2P) lending have risen, increasing the accessibility of credit. P2P lending involves the practice of lending money to individuals (or small businesses) via organizations that match anonymous lenders/investors with borrowers bringing new economic capabilities to financing. A key issue in this peer-to-peer loan origination system is balancing credit risk while increasing credit accessibility.

In the course of this project, public data made available by LendingClub, world's largest P2P organization with over $11B of loan origination, is leveraged to analyze and build an intelligent system to reduce manual effort and time consumed in credit risk analysis process. Once a loan application is filed by a borrower, some key steps in the process include -

- Assigning a proposed interest rate against a loan application by LendingClub
- Categorizing the borrower into specific grades based on his/her profile
- Funding decision made by the investor given the borrower application and relevant features attached by LendingClub

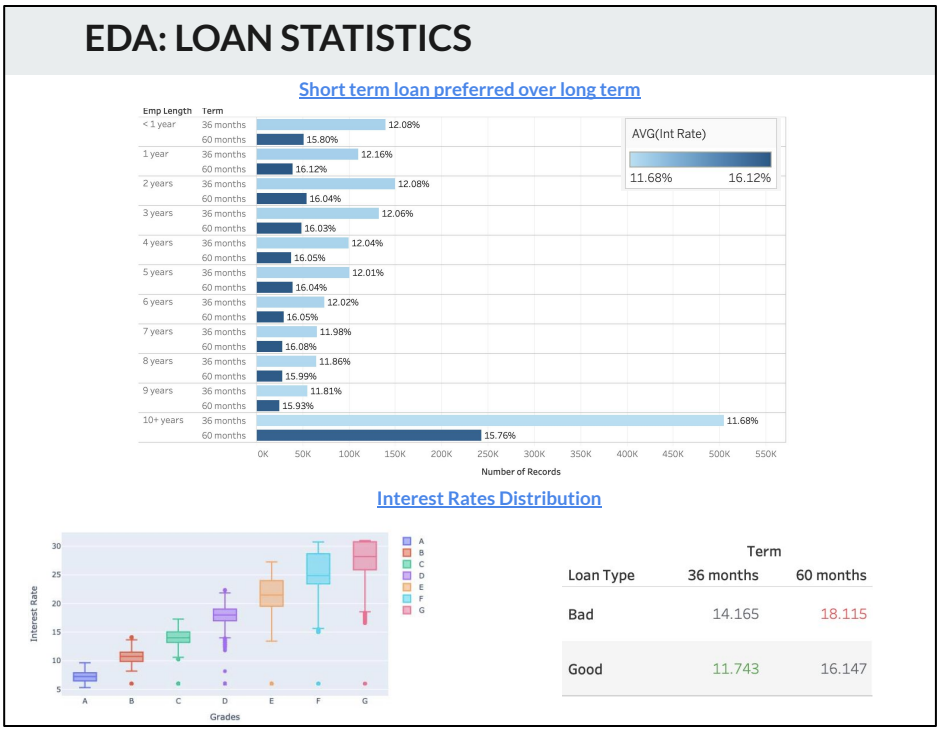
These 3 steps provide an opportunity to integrate data-driven algorithms in order to aid the business process and reduce effort and time. Given this understanding of the business and the dataset, the following approach was decided for the course of analysis -

- Conduct exploratory analysis 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 to analyze loan characteristics and make accurate predictions of good vs bad loan, thereby establishing LendingClub's reliability amongst its clients.



Marching towards our first goal of inferring based on region-wise loan targets and answering to questions related to ideas for business expansion based on regions with a higher level of operating activity, we conducted our preliminary exploration of good and bad loan distribution concerning the state-wise distribution of average income, interest rates being offered and dti ratio. Based on our current observations:

- California(CA) alone accounted for 14% of the total loan shares, followed by New York(8%), Texas(8%) and Florida(7%). These top states along with many others had only a 5-6% default rate of Grade A loans. However, as we go down the Grade level especially for G, surprisingly close to 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.
- Further exploration revealed that 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%.

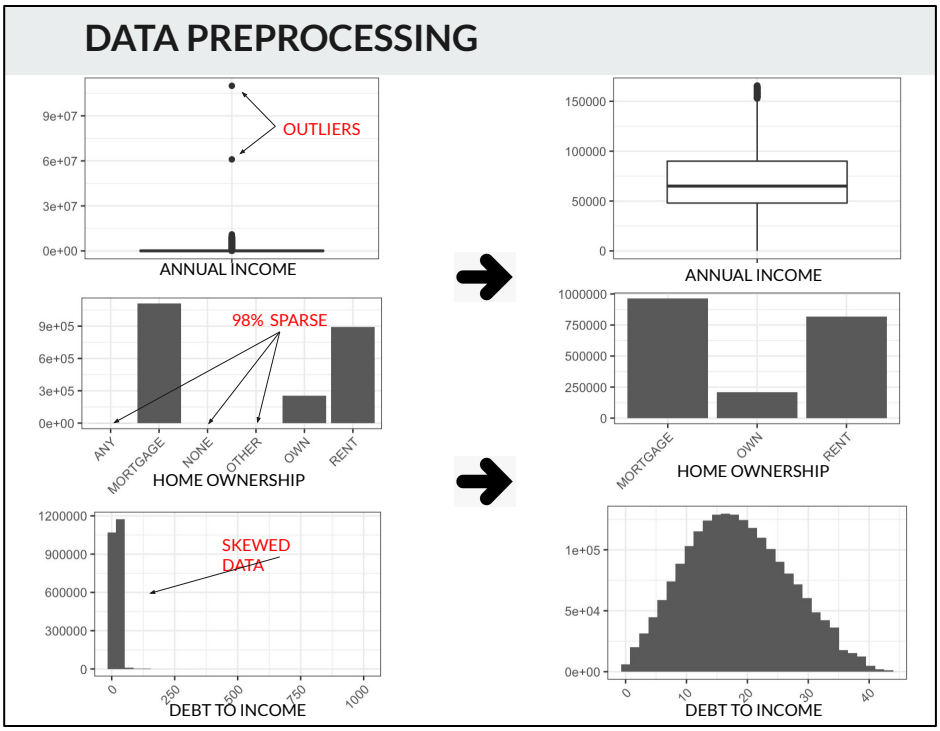


Moving further, patterns across people with varied employment length and loan terms were studied. This was done to test our hypothesis that we can come up with schemes to lure young demographic with better interest rates and longer loan terms.

- A striking variation of 3-4% in interest rates across 36-month and 60-month loan tenure was found and it can be inferred that lower term is preferred across all employment lengths.
- Moreover, in contrast with our initial assumption of interest rate fluctuation with years of employment, we observed no significant difference in terms of interest rate across employment years, except for a slight decrease in the interest rate for 10+ experience level.
- Summarizing the results so obtained, we thus refute our initial claim 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.

The variation of interest rates across different grades, loan type and term was then analyzed.

- Boxplot of interest rate vs grade revealed a linear increase in interest rates as the grades worsened.
- The plot reveals the presence of extreme outliers across all of lower grades. This raises the question - why are certain loans issued with interest rates way outside their respective brackets? Could it be because of some internal contacts? Or could it just be a data anomaly? On investigating the outliers it was recorded that majority were borrowers with experience more than 7 years and many open credit lines. This could suggest a possibility of lower interest rates for repeat customers or internal connections allowing them to secure out of the box interest rates.
- However, in absence of masked personal informations, making such strong assumption of collusion certainly requires internal investigation at Lendingclub.



With EDA, an approximate intuition of strong relationship between interest rates and loan grades could be inferred. However in order to further reinforce this hypothesis through statistical learning, we started off with data preprocessing phase. Since, the project objective lies in interest rate estimates and grade classification, the focus was mainly on features related to parameters provided by the client at the time of loan application i.e., borrower's characteristic with features like debt to income ratio, home_ownership, revolving balance, annual income, employment years and purpose of loan along with intended loan term.

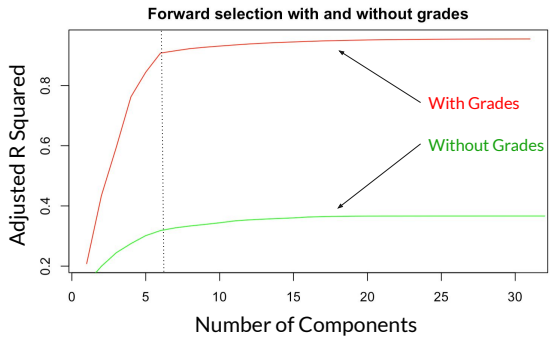
Our processing steps involved -

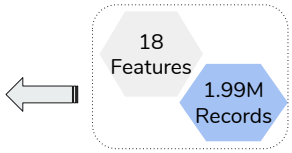
- Normalization for features like annual income
- Outlier removal for features like home ownership
- Removal of skewness for features like dti ratio i.e. debt-to-income ratio
- Elimination of rows with nulls: There were a number of records for which certain variables has no data. This could be because they are not applicable for the specific record and/or the variable was introduced at a later date and hence earlier records have missing data and/or data was simply not available or not recorded.

For the task of Interest rate prediction, 18 features and 1.9M observations corresponding to borrower's characteristics were remaining.

However for Good and Bad Loan Risk Classification, the data comprised of 8598 clean records across 109 features designated as combination of borrower and loan characteristics with a distribution of 75% good vs. 25% bad loans.

FEATURE SELECTIONS: INTEREST RATE





Borrower's Characteristic

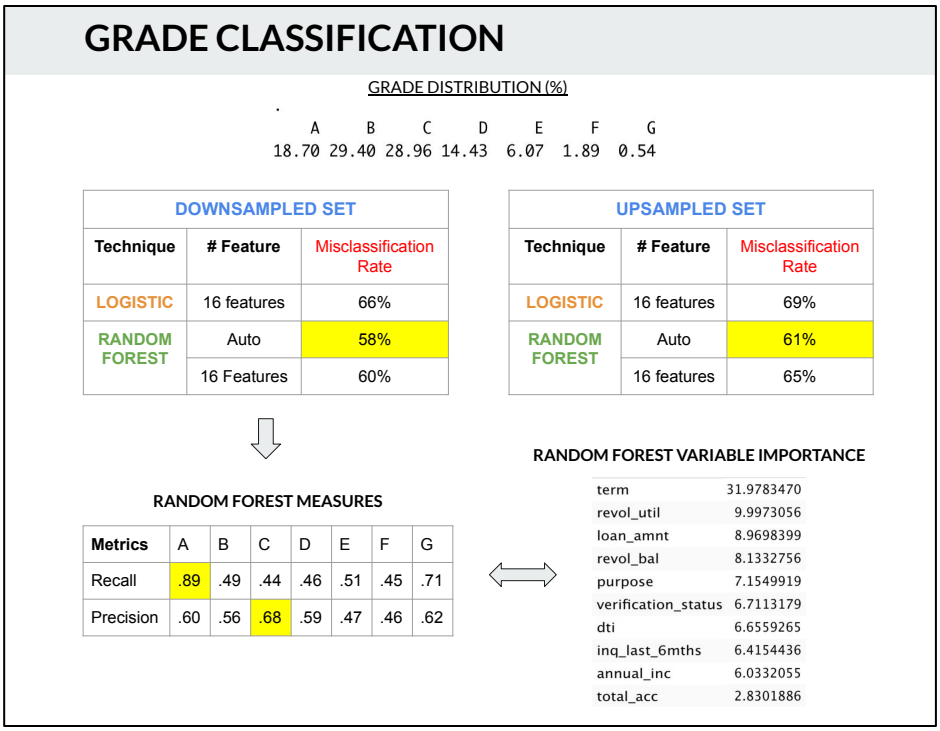
LINEAR REGRESSION				
	WITH GRADES		WITHOUT GRADES	
	RIDGE	LASSO	RIDGE	LASSO
RMSE	1.15	1.03	3.87	3.83
LAMBDA	-0.8	-3.43	-1.2	-3.75

PRINCIPAL COMPONENT REGRESSION		
	WITH GRADES	WITHOUT GRADES
	PC1	PC1
Variance	15.48	7.1468
Adj CV	3.2	4.81

The interest rate assigned to a particular loan application presents a useful view of the return an investor can expect on funding a certain loan. Also, based on our initial analysis through visualization methods, it is known that interest rate brackets are influenced by the borrower grades strongly. The problem for predicting interest rate presents a regression problem.

- Forward subset selection was applied on the processed data with a max number of variables as 18 out of which we had 7 categorical variables. Forward selection was conducted both in presence and absence of grades to study other statistically significant variables capable of predicting interest rate.
- Based on the validation plots, it is observed that in absence of grade, the other variables aren't capable of explaining much variability. The RSS and adjusted R^2 score generated for this model was not so significant.
- Since Lasso regression penalizes the coefficients of the variables, it indicated the importance of variables like term, annual income, public record etc. but the evaluation metrics of interest rate prediction weren't significant.
- To base our hypothesis that grades are an essential component in interest rate determination, we found that our linear regression and PCR model resulted in much better results when feature set involved grade parameters.
- Principal component regression showed that 81% of the variance is explained by only 11 components with grades however it took 29 components without them.

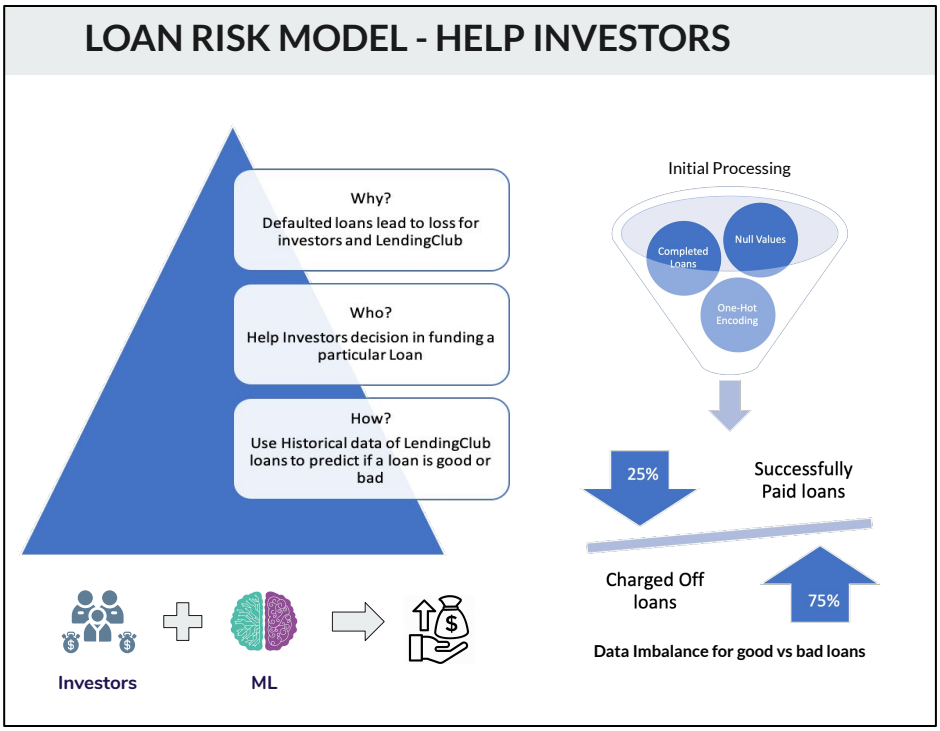
Given LendingClub's business process where grades to borrowers and interest rates to loan applications are assigned by officials after the borrower files an application, it is not beneficial to the process if grade is required as a predictor for interest rate. Without the presence of 'grade', the interest rate prediction doesn't show promising results. Hence, it would be more meaningful to concentrate on identifying the grade that a borrower belongs to, given his loan application. Since, from prior analysis it has been inferred that the interest rates are strongly correlated with the grades/sub-grade, the problem of interest rate prediction then becomes trivial.



Our observation of the strong relationship between the grades and the interest rate also motivated us to analyze a new perspective of using borrower characteristics to assign grades to loan applicants and then use their conjunction to make predictions of interest rates.

The problem of assigning grades to borrowers is a multi-class classification problem with 7 labels. It was observed that the distribution of the dataset across the grade categories was imbalanced. In order to balance the data before fitting a classification method, the following steps were performed -

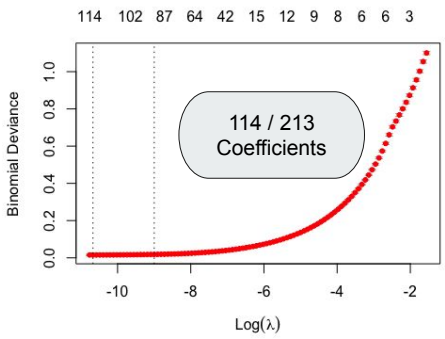
- First run - balance data by downsampling to minority class size. This resulted in 78k observations, 4% of entire dataset.
- Alternate approach - grade G was upsampled, and simultaneously other grades were downsampled. This created a dataset comprising of 15% of original observations.
- The comparison of models built on both the datasets, it is seen that the downsampled data performs better.
- 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.



A key issue of loan origination is balancing the credit risk, as we have talked about repeatedly. A loan default leads to loss of capital for the investors. In the P2P setting, lending investors operate with a high level of information asymmetry and must carefully consider credit risk when making funding decisions. Even with the availability of information like borrower grade in the application notes, investors might not know how to extract useful knowledge as assessing credit risk requires expertise. Hence, the approach to build a classification engine to distinguish between good and bad loans to support funding decisions given the historical patterns.

The dataset after raw processing of null values and considering only completed loans, comprises of 8598 clean records with 96 features. The distribution is imbalanced with 75% good loans and 25% defaulted or bad loans.

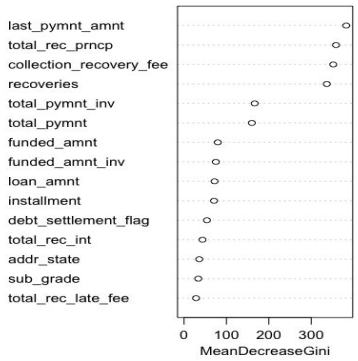
FEATURE SELECTION & MODELING - IMBALANCED



Penalized Logistic Regression (Lasso)

	Predicted Class	
Actual	Charged Off	Fully Paid
Charged Off	441	3
Fully Paid	0	1276

Confusion Matrix (L1 regularized Logistic Regression)



Variable Importance - Random Forest

	Predicted Class	
Actual	Charged Off	Fully Paid
Charged Off	429	15
Fully Paid	0	1276

Confusion Matrix - Random Forest

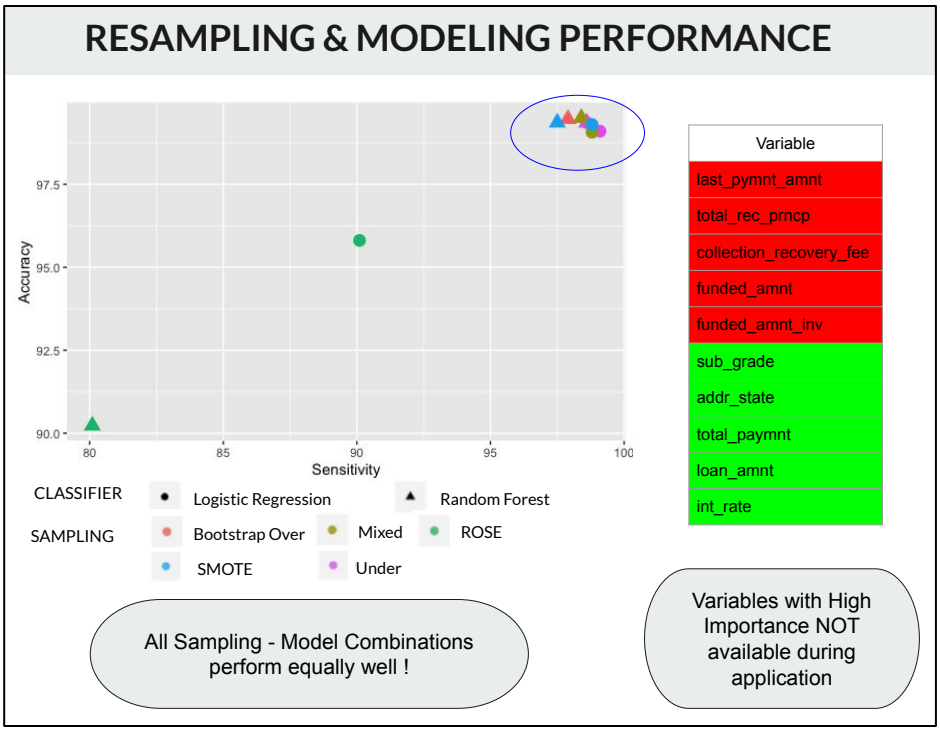
Since the filtered data was imbalanced in terms of class distribution, and the number of features was very high (95 - 18 categorical, 77 numerical), there was a need to conduct dimensionality reduction and variable selection.

Given that the response variable is binary, results from backward and forward selection methods weren't very helpful as the mentioned methods fit a linear model to calculate RMSE and R-squared values which are more useful for regression tasks and are not the correct measures for a classification task. Another hindrance for exhaustive best subset selection method was the lack of computational power.

Thereby, variable selection was made through L1-regularized logistic regression. The best penalizing factor was chosen via cross-validation, whose logarithmic value can be seen to be around -10.5 from the chart. The penalized logistic regression allowed selection of 114 coefficients out of a total of 213 variables (includes one hot-encoded categorical features).

The resulting classifier performance on the unseen test set is very high with 99.8% overall accuracy and 99.3% sensitivity score (with respect to Charged Off loans, the minority class).

The results on the right represent a random forest classifier fit on the training set with imbalanced class distribution. The graphic represents the top 15 important variables in decreasing gini impurity measure. The random forest classifier fit using the top 25 important variables only (as given by the gini impurity measure) doesn't show much difference in results as compared to the L1-regularized logistic regression model, with overall accuracy of 99.12% and minority class sensitivity of 96.7%.



Since the data used for training the classifier is imbalanced, there is a need to balance the dataset so as to ensure that the classifier is able to learn patterns for both the categories efficiently. There is no best technique for resampling or classification that suits all imbalanced datasets and hence we aim to compare the performance of a combination of sampling techniques and classifiers to decide on the best strategy for classification in this case.

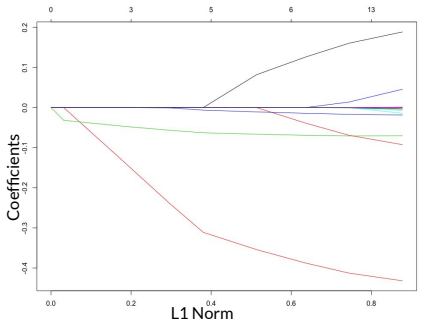
The 5 types of resampling methods used here - bootstrap over-sampling, under-sampling, a hybrid of the two i.e. mixed-sampling, ROSE sampling and SMOTE sampling

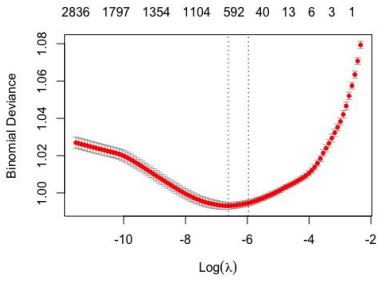
All sampling methods were evaluated in combination with logistic regression and random forest classifiers.

Form the accuracy vs sensitivity (for minority class) plot, it is seen that almost all combinations of sampling and classification approach perform equally well on the unseen test set, with metrics above 95%.

On analyzing the cause for high performance metrics of all combinations, it is observed that - variables like last_payment, total recovery participation, collection recovery fee etc are given high importance by the algorithms. Many of the highly important variables in this case are not available to LendingClub at the time loan application is submitted, hence there is a need to reselect variables given the business understanding of variables available at the time of application.

MODELING - CONTD.





L1-regularized Logistic regression using subset of variables

	Predicted Class	
Actual	Charged Off	Fully Paid
Charged Off	395	49
Fully	123	1153

Best Performance - SMOTE resampling, Random Forest Classifier

Since, the variables from previous random forest classifiers were identified to be not present during the loan application for viewing by the investors, a subset of variables is taken. These variables include borrower characteristics like - income, dti ratio, etc along with some loan characteristics like loan amount, and purpose. Additionally, the grade assigned to a particular borrower and the interest rate calculated by LendingClub against the loan application was also considered as they can be important variables contributing towards a funding decision by an investor.

The new dataset was still imbalanced but with around 90k observations in total.

- The L1-regularized logistic regression selected around 53 coefficient when the best lambda is selected with 1 std. Error from the minimum lambda value. The resulting logistic regression model was evaluated with an overall accuracy of 77%.

- Different combinations of sampling techniques with classification algorithms were evaluated against the unseen test set. The best results were obtained from mixed sampling or SMOTE sampling along with Random Forest algorithm as the choice for classification. The SMOTE sampling with Random Forest classifier combination showed an overall accuracy of 90% and sensitivity (minority class) of 88%.

RESULTS

GOALS	OUTCOME	LIMITATIONS	EXPLANATION
INTEREST RATE PREDICTION		• Computational Power • Absence of latent credit history features like fico scores.	Sub-optimal performance on grade classification
GRADE CLASSIFICATION			
LOAN CLASSIFICATION			Borrower characteristics & loan characteristics successfully capture the trends

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.

- In case of loan classification, the imbalanced dataset could be modeled using the subset of features available during application submission to perform almost 90% accurately.
- Amongst other learnings, the highlights are -
    - The interest rate was found to have highly correlated to grade assigned to a borrower and in the absence of grade, other factors couldn't contribute to its prediction accurately
    - 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