BAX 452 –

Machine Learning Final Project report

Credit Risk Analysis for Lending Club using Machine Learning

Group members

Aditya Satpute

Ian Heggen

Meghana Kanthadai

Executive Summary

Through employing machine learning techniques, Lending club can identify at-risk accounts in-advance that are currently active to mitigate and potentially avoid losses due to loan defaults and consequently maintain investor confidence and its leading market position as the top peer-to-peer lender.

Opportunity

While credit scores-based classification may be a great preliminary way to identify risk of lending to a customer, we may have other factors that could determine the risk of loan default. Solely relying on credit scores may limit the opportunity for growth and expansion for a P2P lender like LendingClub.

Hence, by analyzing additional borrower's characteristics like current income, property ownership, default history, earliest credit line and more, LendingClub stands to gain from the additional insights generated using the customer's profile data to expand its lending capabilities and consequently improve revenues and cut potential losses.

About the data

Data consists borrower profiles and the outcomes, weather a loan was fully paid, or defaulted or is current or delayed.

The data set is robust with 887379 rows and 74 columns and contains data on loans issued between Jan 2007 to Jan 2015.

Solution

The testing set contains loans that are current or issued or in delayed in status for which we predict the risk of defaulting in the future using the model trained on past loans which were either fully paid or charged off/ defaulted.

Because we want to predict probability of defaulting, we use classification-based machine learning models like Logistic Regression, Random Forest and Gradient Boosted Trees. We compare model performance by its accuracy in predicting defaults, as the downsides of defaulting are much higher than the upside of loans that are fully repaid.

Finally, we look at the features of the predicted at-risk loans and check for existence of parameters which can potentially act as clear discriminators for risky and ideal loans. We use ensemble approach to increase the reliability of the model's prediction.

Background and Context

The dataset was obtained from Kaggel.com from the following source:
https://www.kaggle.com/datasets/ranadeep/credit-risk-dataset/data
LendingClub is a financial services company headquartered in San Francisco, California

The process of Credit Risk Analysis is central to the finance industry as has everything to do with lending and the borrower's ability to stick to the agreed upon repayment timeline. And because the downside of defaulting is much higher than the upside from profits, it is extremely crucial for lenders to consistently keep a look out for bad loans and take necessary measures if a loan is at a risk of default.

The method used is routinely applied not only in P2P, but B2B and B2C lending in banking industry.

Through this project, we employ machine learning techniques to emulate the Credit Risk Analysis techniques routinely used by big banks and other lenders.

Existing solutions in the industry

As discussed above, the descriptive way to identify borrower's risk of repaying loans is by analyzing their credit history to determine the likelihood of default.

Another way, which is most commonly deployed and is wide spread in the finance industry is to use statistical and machine learning techniques to identify patterns in loans with similar characteristics and their past outcomes to predict default probability for active loans.

The expected margins may vary from company to company which consequently affects the threshold probability of default to classify a loan as risky. This is one of the key factors that determines model's performance and hence brings in an opportunity to improve models.

We build on these existing methods and adopt them to solve our problem.

Analysis

From Fig 1. It can be clearly observed that out outcomes are skewed with majority of the accounts being "Current" in status, followed by "Fully Paid".

This hints towards need for balancing of data especially when using classification models like logistic regression which can be sensitive to this.

Loan grade is a clear discriminator for interest rate from Fig 2. Lower the grade, higher the interest owed. But, from Fig.3 it can be observed that there is no clear indication of grade effecting loan outcomes. Hence, we need a model to go beyond simple descriptive statistics to assess the probability of default.

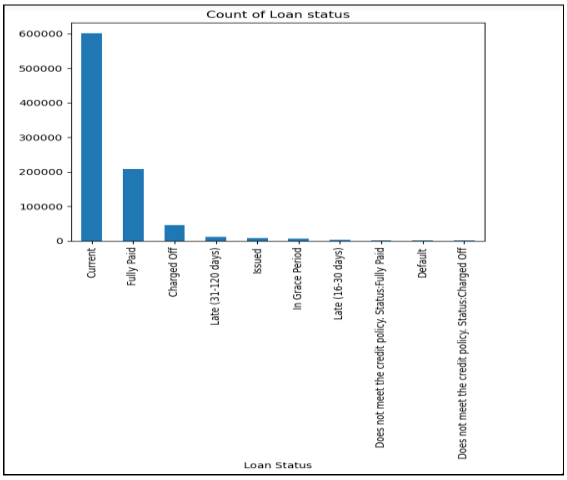

Fig 1.

Frequency plot of outcome variable

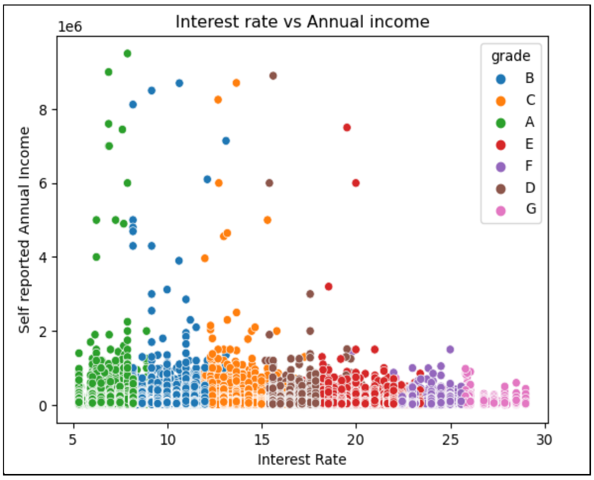

Fig 2.

Frequency plot of Interest rate vs Annual income by loan grade

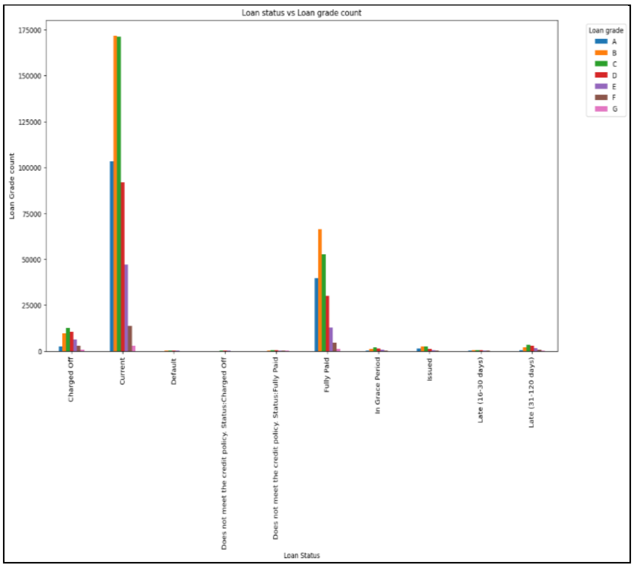

Fig 3. Frequency plot of Loan Grade rate vs by loan grade

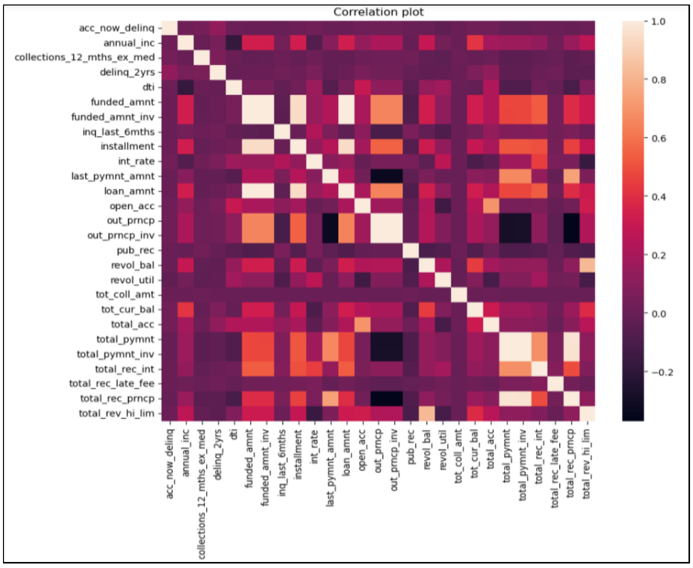

Fig 4. Correlation heatmap of numeric variables

Before modelling, we handle null values as scikit-learn logistic regression and random forests cannot work with missing values.

The correlation heatmap in Fig.4 above which shows the existence of certain highly redundant/ correlated columns. These columns are essentially derived columns and do not add meaningfully contribute to the model and hence, are dropped.

Additionally, we also create our own derived columns like number of months since first credit line using the earliest credit issue date and the loan issue date.

Model formulation

Given task of identifying borrowers who are most likely to default, we use past data on borrower profiles to identify patterns to predict current/ active borrower’s probability of defaulting. Hence, it’s a classic case of supervised classification problem.

Here, active accounts are the borrowers with loan status “Current”, “Issued”, “In Grace Period”, “Late (31-120 days)”, “Late (16-30 days)”.

Inactive, past data i.e. the model training and validation dataset is accounts with loan status as “Default”, “Charged Off”, “Fully Paid”.

We assume the following economic model for our client:

Assumptions: Investors earn 5% to 30% as interest on loans. While LendingClub adds 1-8% as service charge for borrowers and 1-2 % in commission fee from investors.

On average, investors make 20 cents if a loan is fully paid and loose a dollar in case of default. Assuming that our client wants the net average margin to be 5% from the service and commission fee, we get the following expression for profitability:

$(1-p) \times 1/5 - p \times 1 > 1/20$

$p < 1/8$

So, for our client to have a comfortable margin, we need the probability of default to be less than 12.5% each borrower.

Hence, we use this p-value as our classification cut-off for loans to classify accounts as likely or unlikely to default and base our models on this logic in opposition to the out of box model behavior where the classification cut-off is 50%.

Models and Results

We begin with the simplest widely-used classification model Logistic regression with Lasso Regularization and cross validation for best model selection. This is done to ensure a parsimonious model with only variables that are important in deciding whether a borrower in the past defaulted or not.

Then we test Random Forest and extreme gradient boosted trees (Gradient Boosted Trees) and gather and compare metrics like OOS MSE, recall, accuracy and f1 scores. The following tables summarize the results

Summary of results for classification threshold of 0.5

Model	**Recall**	**Accuracy**	**F1 score**	**MSE**
L1 Regularized Logistic regression	0.9925	0.9950	0.9949	0.0049
Random Forest	0.9554	0.9923	0.9738	0.0076
Gradient Boosted Trees	0.9752	0.9952	0.9838	0.0047

Table 1.1

Summary of results for default classification threshold of 0.125

Model	Recall	Accuracy	F1 score	MSE
L1 Regularized Logistic regression	0.9966	0.9840	0.9841	0.0159
Random Forest	0.9957	0.9472	0.8490	0.0527
Gradient Boosted Trees	0.9915	0.9859	0.9547	0.0140

Table 1.2

The following figures provide the confusion matrix results for the models with a “default” classification threshold of 0.5 and 0.125

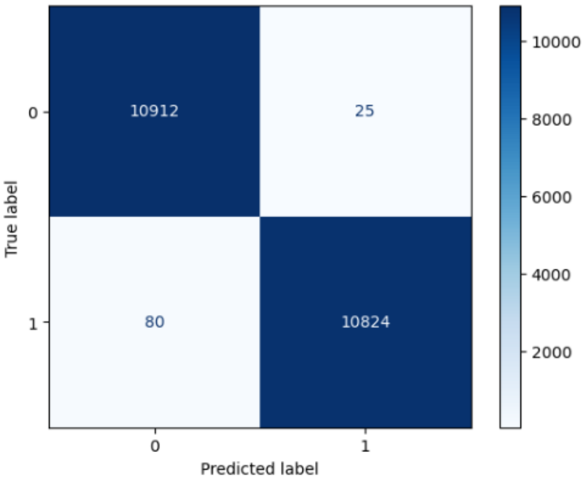

Figure 5.1 Confusion matrix for logistic regression with p = 0.5

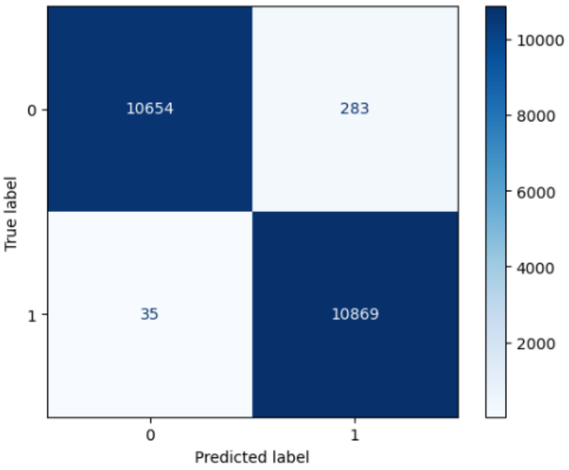

Figure 5.1 Confusion matrix for logistic regression with p = 0.125

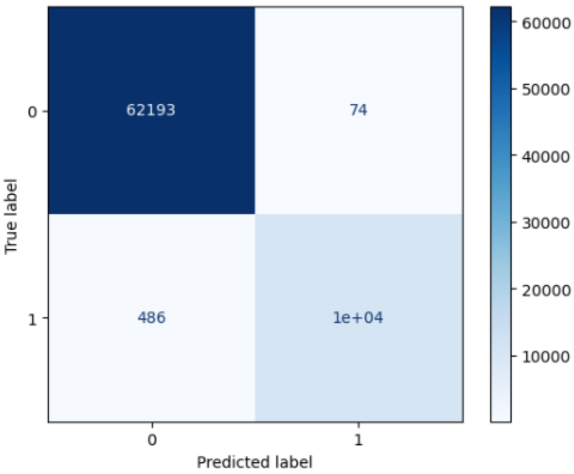

Figure 5.3 Confusion matrix for Random Forest with p = 0.5

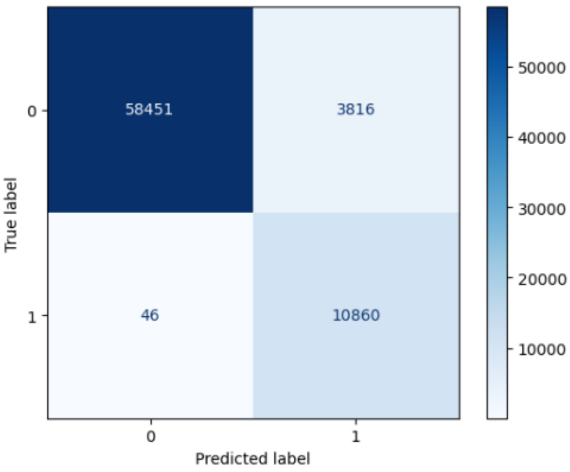

Figure 5.4 Confusion matrix for Random Forest with p = 0.125

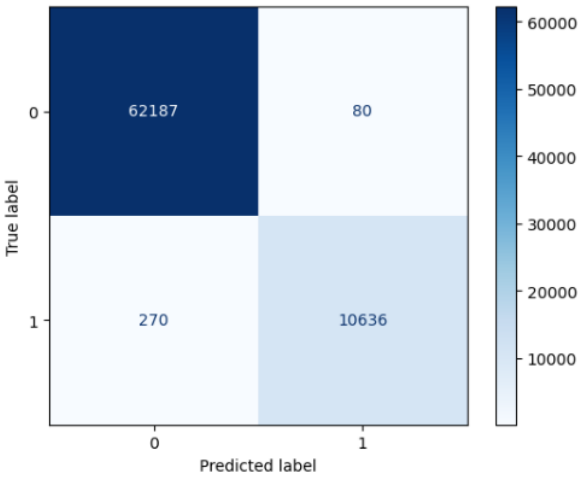

Figure 5.5 Confusion matrix for Gradient Boosted Trees with p = 0.5

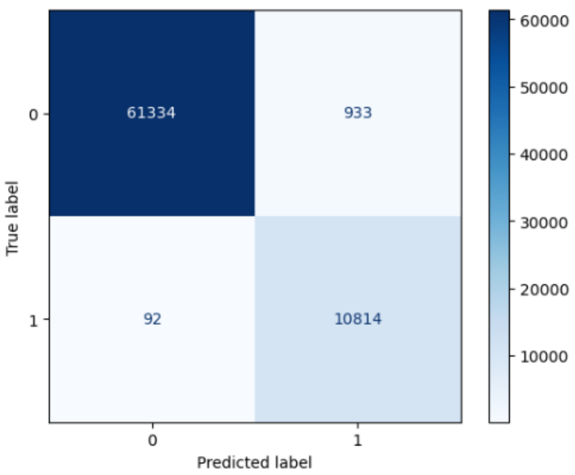

Figure 5.6 Confusion matrix for Gradient Boosted Trees with p = 0.125

Figures 6.1 and 6.2 indicate the obtained OOS precision-recall curve. It clearly indicates that the logistic regression and Gradient Boosted Trees models are the best models for the task at hand. The performance of the model will not be impacted as much due to change in probability threshold. This is extremely ideal for our application and hence promotes confidence in the models.

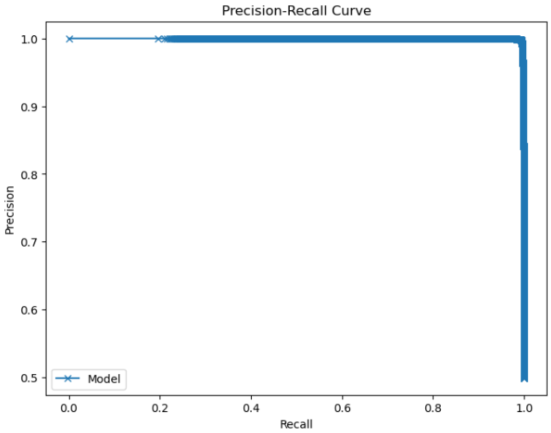

Figure 6.1 Precision-Recall curve for L1 regularized logistic regression model

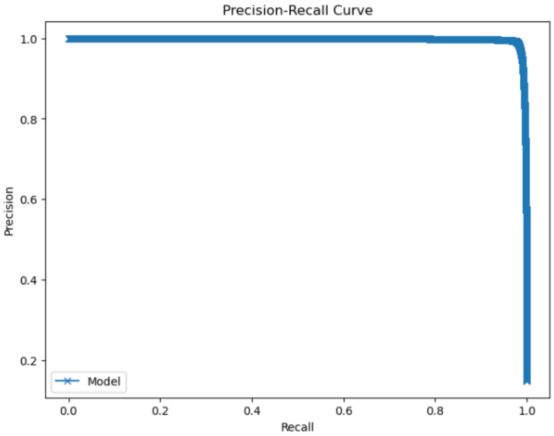

Figure 6.2 Precision-Recall curve for Gradient Boosted Trees model

Results indicate the logistic regression model is the best performing model. But, the biggest drawback of this model is that it is trained on limited data, and its OOS MSE is comparable with the Gradient Boosted Trees model. Hence, we take a model average of the predicted default probability from both L1 regularized logistic regression model and Gradient Boosted Trees to further improve our confidence and ensure generalizability.

Variable importances in the gradient boosted tree method gives the following relative feature importance as displayed in Figure 7 below

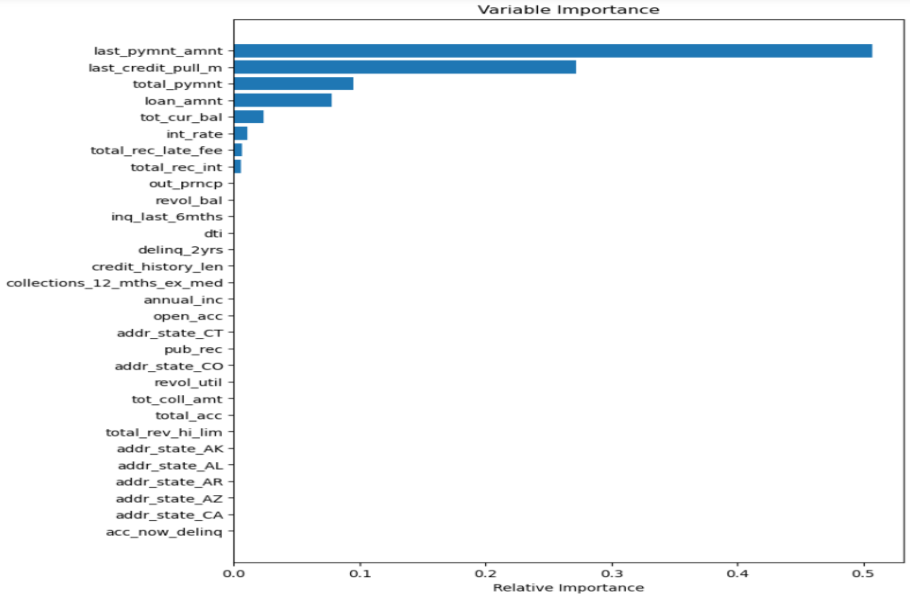

Figure 7. Relative importance of variables for gradient boosting

Additionally, we conducted a post prediction analysis of variables that are markedly different for the predicted ideal and non-ideal borrowers. Figures 7.1 to 7.4 below display some of the key features which can be scrutinized closely when approving future loans

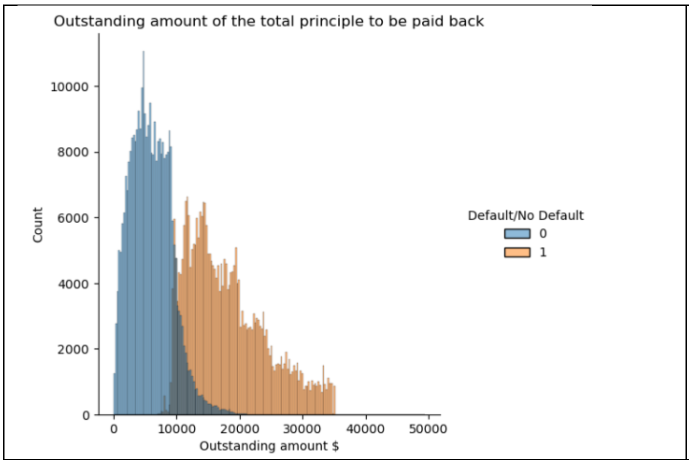

Figure 7.1 Outstanding amount in reference to principle

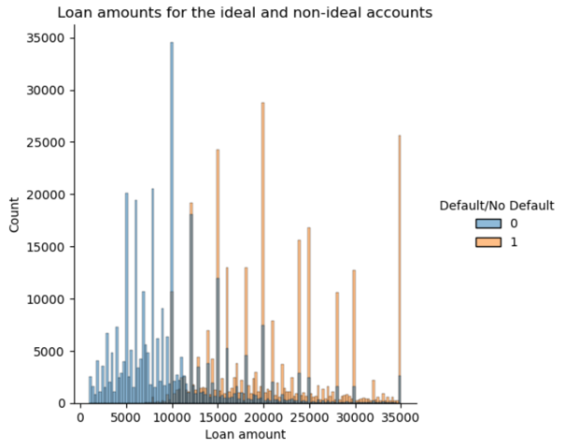

Figure 7.2 Total principal vs predicted default/risky accounts

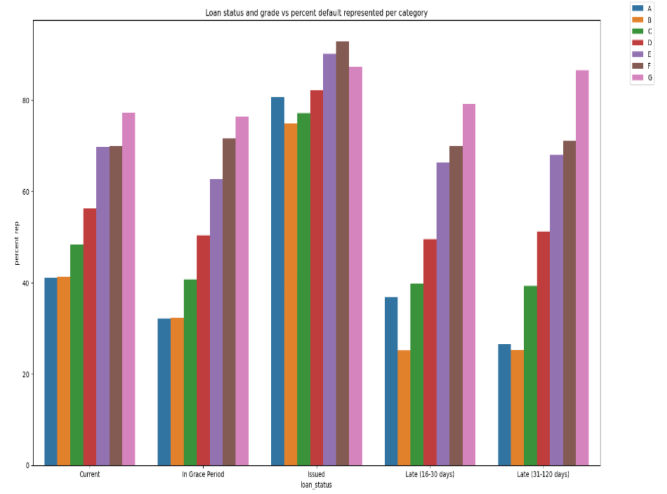

Figure 7.3 Loan status and loan grade counts for Risky accounts

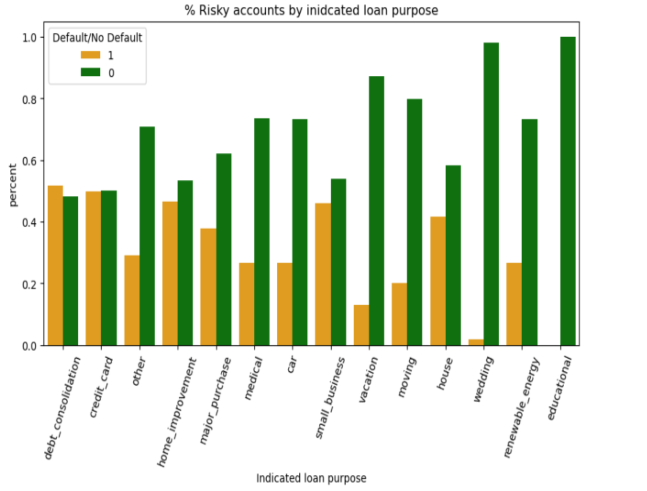

Figure 7.4 Indicated loan purpose vs predicted risky/safe account counts

We observe the following from the models:

1. The more outstanding amount is left to be repaid since the loan was issued, the higher the probability of default.
2. As the loan amount increases, the risk of default increases.
3. Surprisingly, the most recent accounts for which the loans were been issued are risky. Additional research indicates that LendingClub got into legal troubles in 2016.The observed pattern of increasingly risky loans starting 2013 maybe be hinting towards this. Refer to figure 8.
4. Educational and wedding loans are the safest loans with almost zero default probability

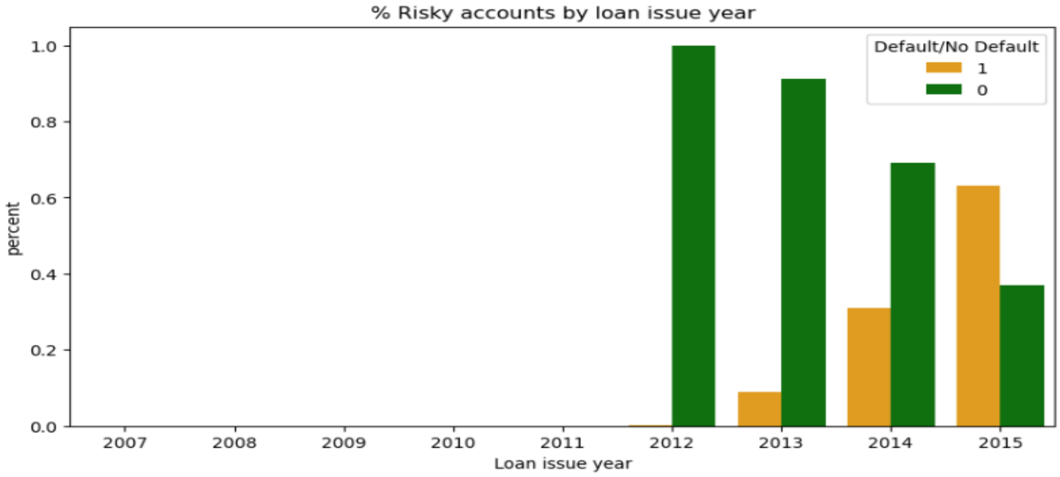

Figure 8. Fraction of risky loan counts by year

Recommendations and Business value

Based on our models, we have created a list of member ID’s who are currently our ideal customers and are have a low risk of defaulting i.e. within the calculated default threshold.

We have also further classified risky customers based on the default probability into moderately risky, high risk, and extremely risky accounts.

The following are the top 3 recommendations based on our analysis:

1. There is an immediate need for account clean up as we observe a rise in the number of risky accounts. We recommend lending club to use the predicted default probability to take immediate action on high-risk accounts by closely monitoring them and/or request collateral, raise interest rate
2. Target borrowers looking to borrow for education and wedding
3. Encourage customers to sign up for a payment plan

An additional benefit from our analysis is the customer demographic and model result analysis indicate certain key parameters like loan purpose, state data etc. that can be used to guide decisions when approving loans and targeting customers through online marketing.

Summary and Conclusions

Thus, through implementing our models we are confident that LendingClub can hugely benefit in three distinct ways including risk analysis, data driven loan approval process and ideal customer targeting.

References

Wikipedia, LendingClub Homepage, Stack Overflow, Kaggle, Official python library documentation, Investopedia