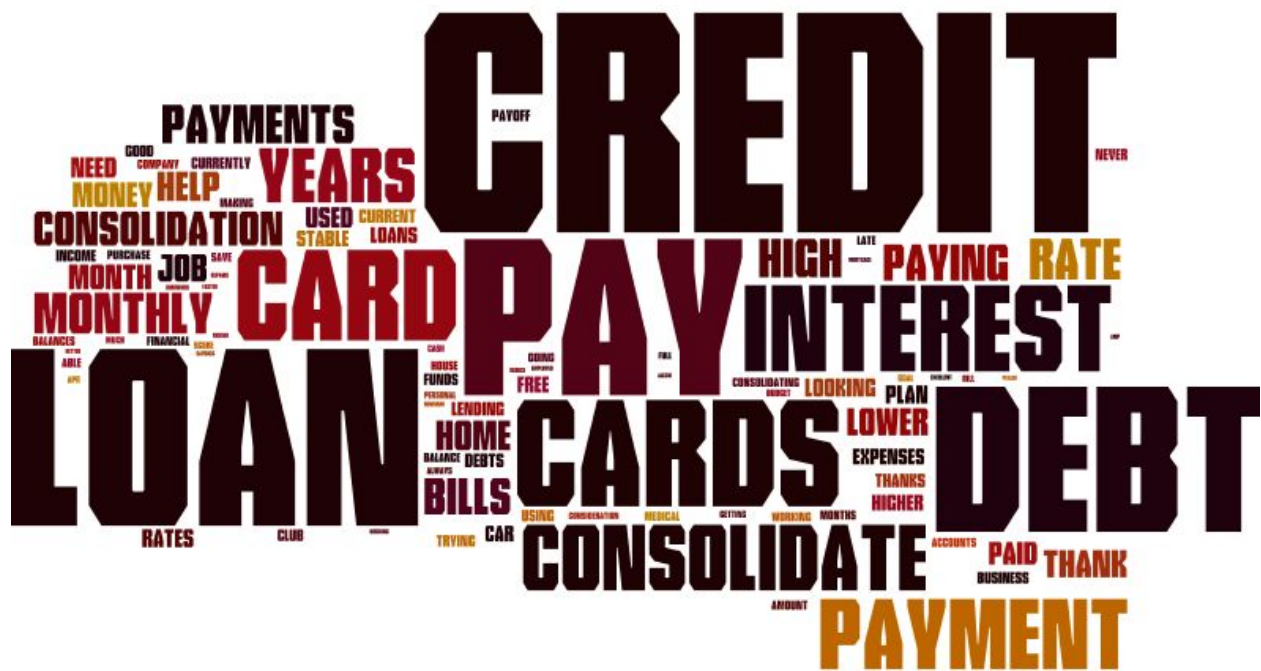


Lending-Club Case Study

March 23, 2019



City-Data.com

Group Members:


1. Dhaval Suthar
2. Prajakta Sumbe
3. Ravi Kiran
4. Veena Iyer

[GitHub Link](#)

Introduction

LendingClub is a US peer-to-peer lending company. The company claims that \$15.98 billion in loans had been originated through its platform up to December 31, 2015.

Lending Club enables borrowers to create unsecured personal loans between \$1,000 and \$40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.



Task 1

1. Research and summarize what your client's needs are going to be.

Ric is the Risk Manager who is concerned about default risk. **Default risk** is the chance that companies or individuals will be unable to make the required payments on their debt obligations. Lenders and investors are exposed to **default risk** in virtually all forms of credit extensions.

1. Fundamentally, Ric needs to make the following decisions:
 - a. Make decisions as to how much money can be invested in the Lending club.
 - b. Need to access all profile and decide profiles in which to invest his money.
 - c. Go through the interest rate that the lending club has provided over the years,
2. Ric's Objectives while making decisions:
 - a. Ric's main objective is to minimize default risks.
 - b. Minimize financial impact when risks occur.
3. Analyzing past data
 - a. can help Ric make better decisions as to the probability of a risk occurring.
 - b. Analyze history of Lending club default risk and financial management plan.

2. Explore the data and comment on data quality, features and get a feel for the data.

While exploring the data we found the quality of data to be dirty and had a lot of unnecessary columns which made data crowded and unclear. The objective was to find rate of interest and to align with that objective we eliminated fields that indicate time as we did not need time series analysis. Columns contained many null values and invalid columns and we eliminated columns which had more than 50% null values. Features that were needed to align with the objective which we focused on were loan amount, funded amount, interest rate instalments, grade, sub grade, home ownership, annual income and loan status. Methodology used to remove null values, missing values and replace with zero, empty string and mean.

3. Summary highlighting key takeaways from Lending club graphs

Lending Club Statistics:

Total Loan Issuance:

Total loan issuance from Lending Club has been gradually increased from 2011 – 2018. Which shows that it has become reliable platform for online personal loans. Increasing simplicity and functionality of lending club has positive impact on the popularity of platform.

Reported Loan Purpose:

68.18% of Lending club borrowers report using their loans to refinance their existing loans or pay off their credit card bills. Though data shows that purpose of the loan is refinance or credit cards, it doesn't guarantee that money has been used for that purpose.

Investor should focus more on loan grades rather than the purpose of the loan issuance.

Loan Issuance by State:

The chart shows that most of the funds borrowed are by people in New York, Texas, California, Florida, Illinois. It shows that states with larger economies have the more people who tend to borrow more loans.

Also, group of some north American states like Montana, North Dakota, Wyoming, South Dakota, Iowa has least number of borrowers.

Lending Club Statistics Performance:

Investor Account Returns by Average Age of Portfolio:

Client should know the return can change over the life of an investment and different factors can influence the volatility of returns.

Number of notes: owning small number of notes usually leads to more volatile returns.

Concentration of an investment: notes corresponding to small amount of loans and borrowers rather than spreading investment leads to volatility of returns.

Weighted average interest rate: returns are more stable on notes with lower interest rates

Weighted average note age: After past-due status of the loan, returns are declined.

Diversification:

To avoid volatility of the returns, diversification-spreading of investment can lead to more solid returns. Having greater diversification or owning more notes can reduce the volatility of returns.

Diversification chart shows that moving from left to right, returns become more stable as the number of notes increases.

Net Annualized Returns (NAT) chart shows that the accounts having more than 100 notes have been more likely to see positive returns and the accounts with less than 100 notes have been more likely to see negative returns.

Diversification reduces the risk of single investment.

Lending Club Statistics – Demand & Credit Profile

Average Interest Rate

This chart shows the weighted average interest rate for loans issued over time.

For a 3-year term, we see that for grade A the average interest rate was maintained below 10% throughout 2011-2018. For grade B we observe that the interest rate rose from 10% from 2011 and dipped back to 10% during 2015 and is slowly rising. For C and D grade we could slowly see that the rates slowly rise from 15% but finally in 2018 C grade is at 15% and D grade is at 20%. With grade E & F, we see a sudden rise in the interest rate post mid of 2015, 2016 respectively. With grade G we see a rise from 20% over to 30%. Thus, average interest rate for the 3- year term period is between 10 to 15% overall.

For a 5-year term, we see grade A interest rate has been consistently maintained near 10% whereas grade B is between 10-12%. Grade C dwindles near 15% and grade D

reached 20% by the end of 2018. Grade E and F have a rise in interest rate from 18%, 19% reaching 25% 30% respectively till 2018. Grade G has started with an interest rate of 20% and reaches close to 30% by 2018.

Thus, overall the average interest rate for the overall term is 12.84% and for the 36-month loan is at 11.55%, for the 60-month loan period is 14.55%.

Loan Performance Details

This is the most useful chart that tells the performance of loans which is important for the risk manager. It gives us information on the Principal amount of loans that is paid, that is in current status, or in grace period status. It tells us that the principal amount of loans is late and not charged off. From the year 2007 to 2012 the late payments were almost always 0. From the year 2013 late payments were negligible. The current which indicates grace period status saw a substantial rise in the year 2014. The increase in loans which were in the grace period was more than the late payments category in the year 2015. The late payments in category C were the highest in 2016. In 2017, the current and late increased substantially compared to the other years. Grade C had again the highest loans which were in their late payment phase and loans in the grace period phase. In the year 2018, the late payments for Grade C and D were the highest.

It can be observed from the above analysis that Grade C had the loans to be maximum in the current phase which is the grace period phase and maximum in the late payment phase.

We can also observe that the number of loans issued has been reducing over the years and is the lowest in 2017 and 2018. In 2018 literally, half of the loans were issued when compared to 2017.

Loan Status Migration Over 9 Months

The chart gives a nine-month recovery rate by loan status. In grace period we have close to 80% inactive status and 22% in net charge offs. Net charge offs are equal to the total principal amount charged off less any funds recovered within three months of the last day of the applicable nine-month period. The default which is more than 120 days has active loans 28% whereas the charge offs is 72%. Thus this can help Ric analyze the grades and loans that more tend to default and the interest rates applicable.


Task 2

observations on featureTools vs manual Feature engineering

While doing manual feature engineering, we removed the features which has more than 50% NA values. We were able to reduce the feature count from 74 to 47. After that, we manually went through each feature and removed some irrelevant features like address, id, zip code, url.

The feature 'Grade' has dependent values so, we applied Labeled Feature Engineering on the Grade. For all other features, we used One Hot Encoding to convert these values from string to numerical format.

We also used FeatureTools Package which provides auto feature engineering approach to find the missing attributes in the dataset which will be more useful for the analysis.



observations on Manual and Auto Feature Engineering :

Manual Feature Engineering	Auto Feature Engineering(FeatureTools)
The traditional process of manual feature engineering requires building one feature at a time by hand informed by domain knowledge	Automated feature engineering with Featuretools allows one to create thousands of features automatically from a set of related tables using a framework that can be easily applied to any problem.
Manual feature engineering requires good domain knowledge and it's a time-consuming process	Compared to Manual Feature Engineering, FeatureTools provided the hidden attributes which are very difficult to get from manual engineering.
If Dataset is not huge, and you have very good knowledge about the domain, it is better to use the Manual Feature Engineering	If you are not well aware of the domain of the problem it is recommended to use Auto Feature Engineering

Task 3

Discuss your Independent and Dependent variables

We designed Linear Regression, Random forest, Neural Networks to predict interest rates.

Our dependent variable is int_rate i.e interest rate.

All other features used to predict interest rates are the independent variables.

MAPE for both Training and Testing data:

Model	Test	Train
Linear Regression	8.27%	7.87%
Random Forest	46.04%	44.56%
Neural Network	40.61	29.75%

5-fold cross validation. How does the model performance change?

Cross-validation is a statistical method used to estimate the skill of machine learning models. Cross-validation is a technique used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. By reducing the training data, we risk losing important patterns/ trends in data set, which in turn increases error induced by bias. So, what we require is a method that provides ample data for training the model and also leaves ample data for validation. In 5 cross validation data is divided into 5 subsets. Now the holdout method is repeated 5 times, such that *each time, one of the 5 subsets is used as the test set/ validation set* and the other 4 subsets are put together to form a training set. The error estimation is averaged over all 5 trials to get total effectiveness of our model. This significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set thus improving performance.

Task 4

Report that discusses the effect of Hyper parameter tuning

On running Linear Regression with default values, we see that R square value given by the model is almost equal to 94%. With hyper tuning with L1 i.e. Lasso on tuning values of fit_intercept is almost equal to 88%, with ridge i.e. L2 we reach a similar value of 94% whereas elastic which is a combination of L1 & L2 gets the value to 85%. MAPE for lasso comes to 42 whereas linear and rigid comes to 9. Elastic MAPE comes to 46.

Thus, for Linear Regression, we find that rigid tuning provides best effect when compared to Lasso and Rigid.

For Neural Networks, with the default values, MAPE comes to 40 whereas when we try to hyper tune it using optimization. Optimization can be done by tuning the solver value which has a by default value of adam. We changed that to sgd which does not scale the value to appropriate value. When we tune the learning rate to invscaling we reach the MAPE value of 65. The learning rate of adaptive does not tune the model well. Epochs can be changed only when solver is at adam or sgd.

Thus, for Neural Networks, hyper tuning does not provide good models other than a decent improvement when using learning rate as invscaling.

For Random Forest, with default values, MAPE comes to 46 whereas with hyper tuning parameters like no of trees only improved the model. The attribute n_estimators indicates the number of trees and when we tune it gives a value of 96.77 which is a tremendous improvement. The depth of the tree can also be tuned and it can be done using the max_depth attribute. This tuning helped improve the model to 95%.

Thus, for Random Forest, we observe that hyper tuning of parameters improves the accuracy of the model.

Overall hyper tuning of parameters drastically improves the model's accuracy from 46% to almost of 97%, whereas for Linear Regression, tuning of L2 i.e. rigid maintains the accuracy of the model at 94%.

Summarize the MAPE for Auto ML model

Auto ML	MAPE
H2O ai	2.95%
TPOT	3%

Discuss manual model vs AutoML approaches with respect to :

- a. Interpretability - The easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models. A linear regression model predicts the target as a weighted sum of the feature inputs. The linearity of the learned relationship makes the interpretation easy. With Manual Model interpretability is less as we have to physically choose algorithms and run our model whereas with Auto ML interpretability is high as it computes various algorithms and tries various permutations and combinations which makes for a robust model and improves interpretability. Interpretability is vital as it tells how the decision is being made and how the machine is learning.

- b. Reproducibility - replicability is the ability of another person to produce the same results using the same tools and the same data. As in Manual model approach, it is recommended that a person should have good domain knowledge about the problem and with that he/she can decide the best algorithm to produce similar result that of the training model. Also in Manual approach, manual data manipulation happens which causes issues in the generating same result in a production environment. Whereas in Automated model approach, even if a person is not familiar with a domain field, using the automated tool it is easy to generate the result. Although, it is recommended that the person is aware of the tools more and able to use them to reproduce the results.

Task 5

use cases that are important for our client :

1. Interest rate

As Ric is Risk Manager and he is concerned about taking default risks. As we learned loan with high interest rates are tend to be more volatile. It is recommended for him to invest in those loans where interest rates are low.

2. Diversification


To avoid the volatility of returns we would suggest our client that rather than investing on a single loan, he should invest on number of loans. The Rate of Return will be more by using this strategy.

3. Loan Status Prediction

We have designed one model which predicts the status of loans like default and charged off. By using that we can determine in which loans our client should not invest in.

4. Average age of the loan

Long terms loans usually have higher interest rates. Which increases the decreases chances of returns and increases default risks. We may suggest our client to invest in short term goals.



References

<https://medium.com/@jayeshbahire/lasso-ridge-and-elastic-net-regularization-4807897cb722>

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

https://scikit-learn.org/stable/auto_examples/model_selection/plot_multi_metric_evaluation.html#sphx-glr-auto-examples-model-selection-plot-multi-metric-evaluation-py

<https://github.com/EpistasisLab/tpot>

<https://github.com/BirgerK/ft-automl-tpot-example/blob/master/2b-birgerk-processByTPOT.ipynb>

<http://docs.h2o.ai/h2o-tutorials/latest-stable/>

<https://github.com/h2oai/h2o-tutorials/blob/master/h2o-world-2017/automl/README.md>



<https://towardsdatascience.com/data-sciences-reproducibility-crisis-b87792d88513>

