IDS 572 Assignment 1 – Loan default prediction and investment strategies (LendingClub)
Due date: Sept 20th .

In this assignment, we will analyze data from an online lending platform, Lending Club (www.lendingclub.com). The goal is to develop models to predict which loans are at risk of default. Such models can then be used to devise investment strategies.

Background

Lending Club (LC) offers an online platform for matching borrowers seeking loans and lenders looking to make an investment. With lower operational costs than traditional lenders (banks), such online lending platforms leverage technology, data and analytics to bring quicker and more convenient financing for individual and small business borrowers from investors looking for attractive investment yields. With increasing volumes, what started as peer-to-peer platforms for connecting individual borrowers and individual investors has today evolved to include institutional investors, hedge funds, etc. Also called marketplace lending or alternate lending, such fintech platforms have seen significant growth in recent years. It is estimated that in 2018, 38% of all personal loans in the US were issued through fintech firms, growing from 5% in 2013 [11]. Some estimate the global online lending market to grow from ~$42B in 2018 to ~$460B in 2022 2[2]. Lending Club, a pioneer in fintech, is one of the largest online lending platforms, with over $50B in total loans issued till date 3[3].

"LendingClub uses technology to operate its online credit marketplace at a lower cost than traditional lending programs, passing the savings on to borrowers in the form of lower rates, and offering investors the potential for competitive returns"[4] Further information is detailed in their website, which you should examine to understand how borrowers apply for loans and the information available for investors to decide on loans to finance.

LC issues personal loans between $1000 and $40,000 for 36 to 60 month durations. Interest rates on these loans are determined based on a variety of information, including credit rating, credit history, income, etc. Based on this, LC assigns a grade for each loan, ranging from A for safest loans to G for highest risk; subgrades are also assigned within each grade. Loans are split into $25 notes, which investors can purchase. Interested investors can browse different loans the LC website, which shows the assigned loan grade and other information.

The online lending business model and how Lending Club operates is described in various web resources. Having an understanding of this is important, to appreciate the role of data and analytics, and the future potential of this rapidly developing area of fintech.

An introduction to alternative lending. Morgan Stanley Investment Insights, May, 2019
https://www.morganstanley.com/im/en-us/financial-advisor/insights/investment-insights/an-introduction-to-alternative-lending.html
https://en.wikipedia.org/wiki/LendingClub
A Trillion Dollar Market By the People, For the People – How Marketplace Lending Will Remake Banking As We Know https://foundationcapital.com/wp-content/uploads/2020/04/FC_CharlesMoldow_TrillionDollarMarket.pdf

[1] https://www.cnbc.com/2019/02/21/personal-loans-surge-to-a-record-138-billion-in-us-as-fintechs-lead-new-lending-charge.html
[2] https://www.alliedmarketresearch.com/peer-to-peer-lending-market
[3] https://www.lendingclub.com/info/statistics.action
[4] https://www.lendingclub.com/public/how-peer-lending-works.action

You may find this interesting - "Theorem uses data science and machine learning to invest in marketplace lending loans" https://www.theoremlp.com/
"LendingClub (A): Data Analytic Thinking" Harvard Business School Case, 2018. Our work in this assignment, though sharing some aspects of this case, takes a different approach in analyses.

To facilitate investment, LC provides access to their data (https://www.lendingclub.com/info/download-data.action - sign up required). Large sets of data are provided in different files. For the purpose of this assignment, we will use a data extract on loans issued in a 5-month period starting January, 2015. The data carries information on 36 month loans, which will all have completed their term by now. Some loans were fully paid back, while others were "charged off" (defaulted).

Assignment

The data on loans is in the file lcData5m.csv and the LCDataDictionalry.xls file describes the variables.

In this assignment, we will first explore the data on loans, to develop an understanding of loan grades and subgrades and how they may relate to default and returns performance, loan purpose and any relation to performance, analyses of returns from loans, etc. We also need to look into missing data, and how to address this. While the data carries information on over 100 variables, we need to determine *which data will be available when looking to invest in a loan* — since our goal is to develop a model to predict loan default and then decide which loans to invest in; such a model will thus be only able to consider variables available *before* a loan is issued. The subsequent task is to develop models to identify god/bad loans, and evaluate these. We will also consider investment performance corresponding to these models and identify the best model.

Questions:

1. Describe the business model for Lending Club. Consider the stakeholders and their roles, and what advantages Lending Club offers. How does the platform make money? (Not more than 1.5 pages, single spaced, 11 pt font. Please cite your sources).

2. Data exploration
   (a) some questions to consider,
   (i) What is the proportion of defaults ('charged off' vs 'fully paid' loans) in the data?
   How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?

   (ii) How many loans are there in each grade? And do loan amounts vary by grade?
   Does interest rate for loans vary with grade, subgrade? And is this what you expect, and why?

   (iii)What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? And within grade? Do defaults vary by purpose?

   (iv) Calculate the annual return. Show how you calculate the percentage annual return. Compare the average return values with the average interest_rate on loans – do you notice any differences, and how do you explain this?

How do returns vary by grade, and by sub-grade.
If you wanted to invest in loans based on this data exploration, which loans would you invest in?

(v) Generate some new derived attributes which you think may be useful for predicting default., and explain what these are.

(b) Are there missing values? What is the proportion of missing values in different variables? Explain how you will handle missing values for different variables. You should consider what he variable is about, and what missing values may arise from – for example, a variable monthsSinceLastDeliquency may have no value for someone who has not yet had a delinquency; what is a sensible value to replace the missing values in this case?
Are there some variables you will exclude from your model due to missing values?

3. Consider the potential for data leakage. You do not want to include variables in your model which may not be available when applying the model; that is, some data may not be available for new loans before they are funded. Leakage may also arise from variables in the data which may have been updated during the loan period (ie., after the loan is funded). For example, it has been noted that the FICO scores on loan applicants are updated periodically, and the data can carry thus FICO scores from after the loan issue_date. So, even though FICO score can be useful, the values in the data may not be usable. Identify and explain which variables will you exclude from the model.

4. Develop decision tree models to predict default.
   (a) Split the data into training and validation sets. What proportions do you consider, why?
   (b) Train decision tree models (use both rpart, c50)
   [If something looks too good, it may be due to leakage – make sure you address this]
   What parameters do you experiment with, and what performance do you obtain (on training and validation sets)? Clearly tabulate your results and briefly describe your findings.
   How do you evaluate performance – which measure do you consider, and why?
   (c) Identify the best tree model. Why do you consider it best?
   Describe this model – in terms of complexity (size).
   Examine variable importance. Briefly describe how variable importance is obtained in your best model.

5. Develop a random forest model.
   What parameters do you experiment with, and does this affect performance?
   Describe the best model in terms of number of trees, performance, variable importance.
   Compare the random forest and best decision tree model from Q 4 above. Do you find the importance of variables to be different? Which model would you prefer, and why.

For evaluation of models, you should include confusion matrix related measures, as well as ROC analyses and lifts. Explain which performance measures you consider, and why.

6. The purpose of the model is to help make investment decisions on loans. How will you evaluate the models on this business objective? Consider a simplified scenario - for example, that you have $100 to invest in each loan, based on the model's prediction. So, you will invest in all loans that are predicted to be 'Fully Paid'. Key questions here are: how much, on average, can you expect to earn after 3 years from a loan that is paid off, and what is your potential loss from a loan that has to be charged off ?
   One can consider the average interest rate on loans for expected profit – is this a good estimate of your profit from a loan? For example, the average int_rate in the data is 11.2%; so after 3 years, the $100 will be worth (100 + 3*11.2) = 133.6, i.e a profit of $33.6. Now, is 11.2% a reasonable value to expect – what is the return you calculate from the data? Explain what *value of profit* you use.
   For a loan that is charged off, will the loss be the entire invested amount of $100? The data shows that such loans have do show some partial returned amount. Looking at the returned amount for charged off loans, what proportion of invested amount can you expect to recover? Is this overly optimistic? Explain which *value of loss* you use.
   You can also consider the alternate option of investing in, say in bank CDs (certificate of deposit); let's assume that this provides an interest rate of 2%. Then, if you invest $100, you will receive $106 after 3 years (not considering reinvestments, etc), for a profit of $6.
   Considering a confusion matrix, we can then have profit/loss amounts with each cell, as follows:

		Predicted	
		FullyPaid	ChargedOff
Actual	FullyPaid	*profitValue*	$6
	ChargedOff	*lossValue*	$6

   (a) Compare the performance of your models from Qs 4 and 5 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Evaluate different thresholds and analyze performance. Which model do you think will be best, and why.

   (b) Another approach is to directly consider how the model will be used – you can order the loans in descending order of prob(fully-paid). Then, you can consider starting with the loans which are most likely to be fully-paid and go down this list till the point where overall profits begin to decline (as discussed in class). Conduct an analyses to determine what threshold/cutoff value of prob(fully-paid) you will use and what is the total profit from different models. Also compare the total profits from using a model to that from investing in the safe CDs. Explain your analyses and calculations.
   Which model do you find to be best and why. And how does this compare with what you found to be best in part (a).