**Team 4 Final Project**

Chance Pickett, Clara Smith, and Addison Wilson

University of Arkansas

DASC 1223 – Data Science in Today’s World

Dr. Karl Schubert

4/30/2024

# **Executive Summary**

**Problem**

This report presents the findings of an exploratory data analysis conducted on a dataset of thousands of loans from Lending Club, with a focus on charged off and defaulted loans. The objective was to identify key patterns and predictive factors that lead to loan defaults and charge-offs to help lenders minimize risks. There was also a focus on identifying trends and an overall understanding of the data.

**EDA Process**

The analysis began with data cleaning to fix issues such as missing values, and irrelevant columns. Different visualization techniques, such as bar charts and box plots, were used to create insights into the distribution of the loans.

**Modeling**

A C5.0 decision tree algorithm was utilized to determine the characteristics of loans that result in default, using a 75%-25% testing-training data partition. Three trees are created: a single-trial decision tree without rules, one with rules, and a tree with ten trials of adaptive boosting.

**Results**

The EDA showed socioeconomic considerations of previous debts, lower income, and other pressing concerns that might distract from loan payments. Outstanding principal, last payment amount, total received late fee, and funded amounts were significant predictors of default risk. All three decision trees had an accuracy rate of 99.8%. There is concern that these trees may be significantly overfitted.

**Conclusion**

In conclusion, the completed analysis offers significant insights into factors influencing default and charged off loans as well as aiding lenders in refining risk assessment strategies. Through proper analysis, lenders can have a broader understanding of borrower behavior in the financial industry and implement it to improve business profitability and success.

# **Understanding the Problem**

**LendingClub**

LendingClub is a digital financial services company that focuses primarily on consumer loans. They have no actual physical locations, and instead take advantage of the convenience and availability of technology. As a vital part of their company is loans, it’s important for them to understand trends following loan statuses and how to prevent any negative impacts they might have on the company.

In addition to their work, LendingClub commits to transparency, integrity, and most importantly customer satisfaction. To adapt with changing market dynamics and consumer preferences, LendingClub finds it vital to understand their consumers and their habits.

**Default and Charged Off Loans**

Default and charged off loans represent significant challenges for lenders. To gain a deeper understanding of these challenges, it’s important to know exactly what these loan statuses entail. A loan enters default status when a customer fails to make payments on a loan multiple times. Some provisions that follow along with default status include late fees and a higher interest rate. In comparison, a charged off loan is enacted typically within six missing payments. The debt is then categorized as uncollectable and collateralized debt can be seized. These late payments will stay on a credit file for years and cause significant damage to the targeted customer’s credit.

**Client Goals**

The primary goal of LendingClub is to gain an understanding of their current customer base. This involves identifying patterns in successful loan applications but also delving into the characteristics of applicants who struggle with repayment or even default. By finding these patterns, LendingClub aims to tailor their marketing strategies to effectively target borrowers who have a higher probability of being successful. This report reaches these goals by finding insights and creating recommendations to tackle these pressing issues.

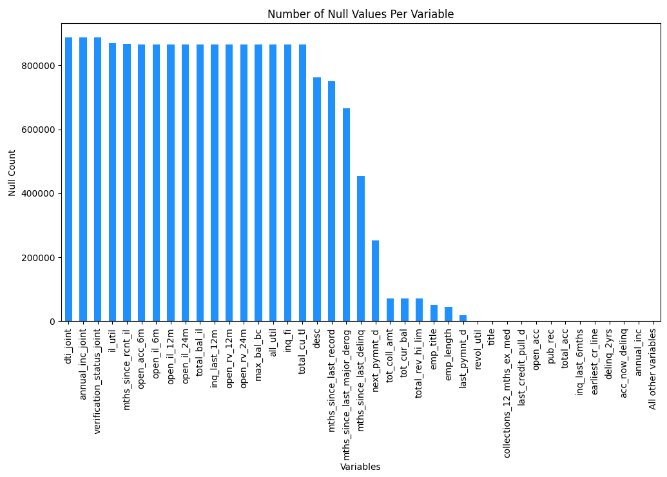
**The Data**

The dataset has a total of 887,379 loans from LendingClub. Of these loans, 1,219 are defaulted and 45,248 are charged off which makes 46,467 loans the target for the data analysis. There are also a total of 1,598,312 null values. Of these values, customers with no joint accounts or open accounts make up for the majority of them.

**Data Cleaning**

**Figure 1**

*Distribution of Null Values*



*Note: “All other variables” are variables with 0 null values. They were grouped for readability.*

The first step to cleaning the data was to visualize the distribution of null values as seen in Figure 1. In this figure, all joint variables and open account variables have the same number of null values as a customer doesn’t have any open accounts, then all variables related to open accounts will also be null. All null values were removed from the dataset to improve balance in the data. The rest of data cleaning was correcting any inconsistencies in formatting as well as checking outliers through individual EDA. As a final step, the consistency and accuracy of relationships between these variables were checked through foreign key references.

**EDA**

**Loan Status by Purpose**

Figure 2.

A graph with blue squares and white text

Description automatically generated

Note. This figure was made using R with a csv file provided by LendingClub. It shows the total number of loan statuses through the purpose of loan.

A bar graph was used to represent what purposes were given to various loans taken out by customers. As Figure 2 showcases, debt consolidation takes overwhelming majority with 28,389 of all loans. The next highest purpose is credit card loans with only 8,059 loans. This matches up with the targeted customer base as it shows that the majority have a history of debt and in turn take out a loan to consolidate. Since they have this history, it is likely that they are not as financially stable and might have other concerns that take over priority than loan payments. It is a problem that is important to highlight to properly take time to educate customers on how to properly repay debt and improve credit.

**Loan Status by Annual Income**

Figure 3.

A graph of a number of blue bars

Description automatically generated

Note. This figure was made using R with a csv file provided by LendingClub. It shows the total number of loan statuses through the annual income of a customer.

Next, a bar graph was made to see the annual income of customers with loan statuses of default and/or charged off. Figure 3 shows a right skewed distribution which indicates that there are extreme values within the higher end of the distribution. The highest number of loan statuses is concentrated between 30k to around 60k yearly income. There is also a slight increase in defaulted loans at annual incomes above 200k. This is likely because it is a general statement of greater than, indicating that there are many more possibilities than the graph breaks of 10k. This does not indicate that individuals who make significantly more than average income have any risk of not paying off loans.

It is important to note the concentration of annual income being in the lower distribution. As the cost of living is varying, it is hard to pinpoint where exactly each customer places in this standard. For an individual, an income from around 40k-50k is livable in most low-cost areas. In comparison, a family will need a much higher annual income to reach the standard of a livable wage. This indicates that many customers are making below or on the brink of a livable wage. This once again infers that these customers are likely to have pressing concerns or struggling paying off loans.

**Loan Status by Home Ownership**

Figure 4.

A graph of a number of blue rectangular bars

Description automatically generated

Note. This figure was made using R with a csv file provided by LendingClub. It shows the total number of loan statuses through the home ownership of a customer.

A bar graph was made to show the home ownership status of customers in default or charged off loan status. Home ownership statuses of rent and mortgage both represent over 20,000 loans each. In comparison, ownership of homes is below 5,000 loans. This is a further indication that customers in this loan status often have other payments that may interfere with these taken out loans. Figure 4 follows along with previous graphs such as Figure 2 and Figure 3 for the fact that there is a history of debt, other payment concerns, etc.

**Loan Amount and Status**

**Figure 5**

A graph of a loan amount

Description automatically generated A Kernel Density Estimation graph is a way to visualize a distribution smoothly. First a bar graph was made to compare the distributions of loan amounts for these two loans, but the amount of other loans was far greater than charged off and default loans that it was hard to visualize the comparison. A KDE graph shows the density of each loan amount instead of count, so in a way both distributions were normalized. For both loans, the distribution of amount is very similar. The amount of the loan most likely will not change if it is charged off or defaulted.

**Loan Term and Status**

**Figure 6**

*Proportional Distribution of Loan Terms for Charged Off, Defaulted, and All Other Loans*

A blue and orange pie chart

Description automatically generatedA blue and orange pie chart

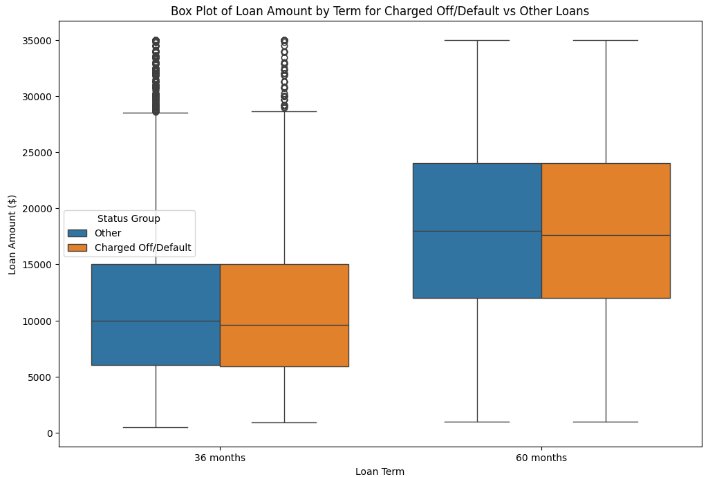
Description automatically generated

The difference between the top and bottom pie charts is that the top two aggregate charged off and default loans. The proportions are very similar as they are both about a 65/30 split. For the bottom two pie charts, charged off and default loans were distributed separately. The split for charged off loans is again 65/30, but for default loans it is about 60/40. This means that for charged off loans and all other loans, about 65% of them have a three-year term, but only 60% of default loans have a three-year term. The difference is subtle, but interesting, nonetheless. It is possible that five-year term loans are more likely to be defaulted.

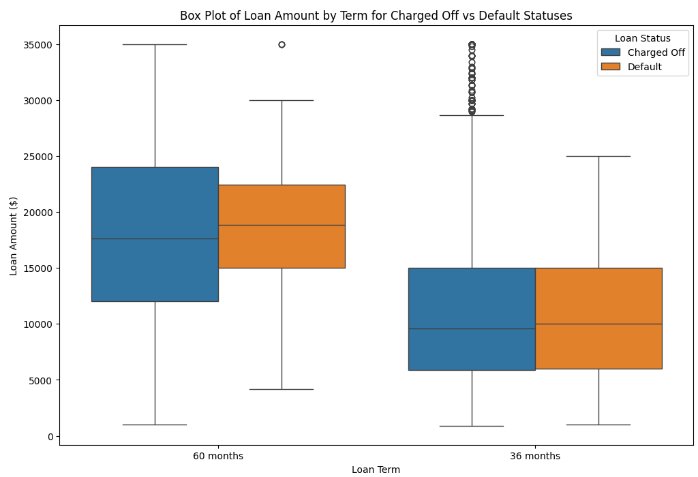
**Loan Amount by Term for Status**

**Figure 7a, 7b**

*Box Plot of Loan Amount by Term for Charged Off, Default, and All Other Loans*



**7a.**



**7b.**

Again, the top plot shows charged off and default loans aggregated, while the bottom plot shows them separately. Aggregated, the charged off and default loans show the same variance and outliers as all other loans. But when separated, for both term lengths, default loans have less variance than charged off loans. This is most likely because the title charged off is assigned to loans more aggressively than default loans, so loans that may be paid back might be mis labeled as charged off, causing a wider spread of different loans with that assignment.

**Modeling**

**Introduction**

The decision tree is a form of supervised learning commonly used in machine learning. Utilizing a simple algorithm that splits categories of variables based shared attributes, decision trees can serve as incredibly powerful algorithms for classification tasks (IBM, 2022).

Decision trees follow the shape of a regular tree, where each branch splits based upon a decision that is decided by through a recursive partitioning algorithm. The initial decision is marked by a root node, where cascading decision nodes further refine the variables used for classification into distinct subsets that are further divided. Different values of these variables are assigned branches that descend the tree, all the way until all the predictive nodes have the same class, or there are no more predictive features to split subsets with (Lantz, 2023, p. 148-150).

Decision trees remain a popular simple classification algorithm due to their simplicity of interpretation, easy implementation in popular programming languages, and their lack of requirement for normalization of data. Decision trees are also very flexible with the type of data that they can handle, freely accepting categorical and numerical variables within the same model, and allowing the incorporation of missing values (Duggal, 2023). Despite singular decision trees not being near as robust as their ensemble, the random forest (Liberman, 2020), decision trees on their own can still be very powerful classifiers for tasks such as movie suggestions, disease diagnosis, and customer behavior (Hogarty, 2022). Because decision tree algorithms are relatively forgiving to the kinds of data that are given to them, this form of predictive modeling was utilized for our project.

One of the most famous uses of decision trees is predicting the risk of default in those who receive loans in the finance industry, in which classification algorithms can be used to determine the rules that increase the likelihood of default in borrowers (FasterCapital, 2024). Due to our group’s focus on those who have defaulted on their loans, the usage of a decision tree would allow us to uncover major decision rules that increase the likelihood of default, and create targeted recommendations based upon the findings of the model.

**Methodology**

For our decision tree, a C5.0 algorithm was chosen. The C5.0 algorithm is one form of decision tree algorithm, in which decision-splits are created with the goal to reduce the datasets entropy. Entropy refers to the amount of “disorder” in a dataset, characterized by the proportions of observations within a subset that have different classes. Splits can either be homogenous, or non-homogenous; a homogenous split, for example, would refer to the breaking of a subset into two groups such that each group had instances of a single, unique class. A non-homogenous split, on the other hand, would refer to a split that would have a mix of classes in its results. Homogenous splits are the most desirable, with the goal of creating subsets that are of a single class, or “pure” (Lantz, 2023, p. 155).

Another reason for the choice of C5.0 was its relatively easy, beginner-friendly implementation in R using the C5.0 package (Kuhn, n.d.). The choice of R, despite not being a language known for machine learning models, was motivated partially due to the desire for seamless integration tidyverse, and partially because of the professional motivations of Clara Smith, who ran the model. The following R packages were utilized for this project:

* Tidyverse – Primary utilized the dplyr package for data wrangling, and transforming the dataset to a form that was suitable for the decision tree.
* VIM - “Visualization and Imputation of Missing Values” - an R package that provides several different methods of visualizing and imputing missing values in the dataset.
* C50 - The package utilized the implement the C5.0 decision tree algorithm.
* Gmodels – A small R package that contains several model fitting functions. For this project, CrossTable() was utilized to create a cross tabulation.

**Cleaning the Data for the Model**

The first step in creating our model involved the manual inspection of the values contained within the LendingClub dataset, utilizing the LendingClub dictionary excel sheet. Given that 74 variables were included in the original dataset, for the model to work, it was crucial that this dataset was significantly filtered.

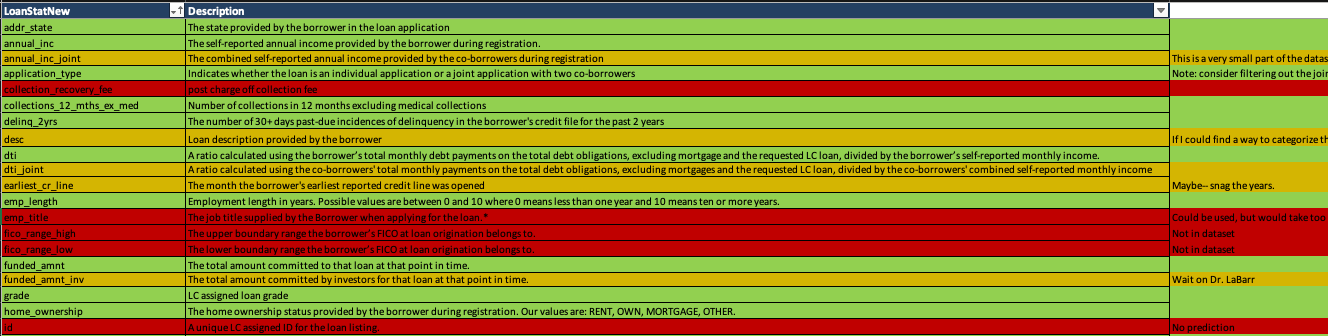
Clara went through the excel file, and manually picked variables that possibly could have high predictive power for the final model (green), and variables that could’ve had the potential to be predictive given additional data wrangling (yellow). The red variables, and most yellow variables, were filtered from the final dataset for several reasons, including:

1. Variables that had zero predictive power and would have instead resulted in severe overfitting of the data (id, member\_id, url).
2. Variables that could have had predictive power, but may have required additional lengthy wrangling beyond the scope of time for this project (emp\_title, title, zip\_code)
3. Variables that had a proportion of missing values that were greater than 50% of the dataset, and could introduce noise into the model (mnths\_since\_last\_delinq, total\_bal\_il)
4. Variables that appeared in the dictionary, but were not present in the dataset (is\_inc\_v, last\_fico\_range\_high, last\_fico\_range\_low)

Ultimately, 32 different variables were used to test and train the machine learning model.

**Figure 8**

*Screenshot of excel file of lending terms, color coded based upon their inclusion and feasibility in the final model.*

 The target class of the dataset is the ”loan\_status” variable. Because our group wanted to understand the risks and contributors for defaulting, it was important to filter our dataset between those that defaulted (“Default“) , and those that did not default (”Not Default”).

The first step in achieving this included wrangling the “Does not meet credit policy” values of the loan\_status variable to shorter versions. The “Does not meet credit policy” tag of these observations refers to the fact that with today’s standards, these loans would not have been approved. These values were wrangled into “Charged Off” or “Fully Paid” as appropriate for the dataset.

**Figure 9**

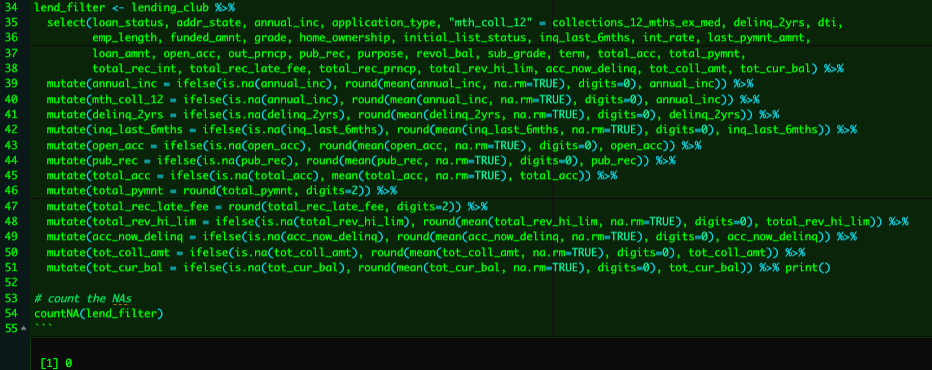
*Transforming the legacy loans into their respective classifications.*



Next, using the chosen variables, Clara selected the lending\_club variables that were deemed appropriate during the dictionary phase of the cleaning. Several of these variables had small amounts of NA values to fill and were imputed with the means of the available data values of those variables. Dollar amounts were also rounded to two digits. A series of if-else statements were used to achieve this.

**Figure 10**

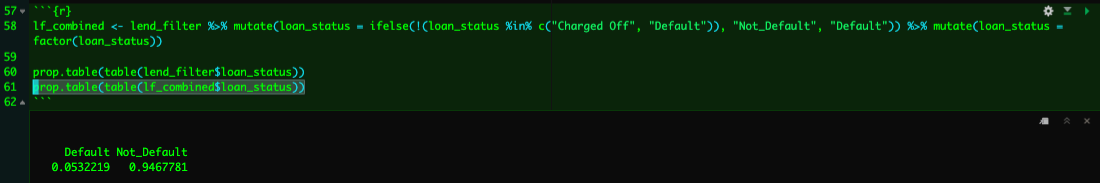
*Filtering the dataset by the 32 candidate variables. Missing values were imputed by means, and dollar amounts were rounded to two decimal places.*



Finally, every observation was categorized in a binary loan classification, in which the values of “Charged Off” and “Default” were combined under the umbrella of “Default”, while everything else was categorized as “Not Default”.

**Figure 11**

*Creating the binary target classification variable observations, “Default” and “Not\_Default”*

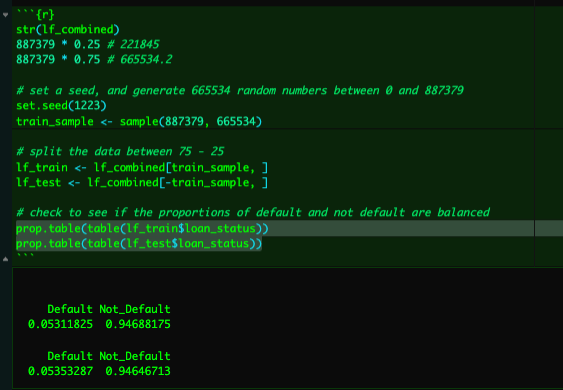


**Splitting the Dataset**

The lf\_combined dataset was then split into training and testing data, with 75% of the data being used to train the dataset, and 25% of the data being used to test the predictive power of the model. To ensure fairness, it is very important to make sure that the proportions of Default and Not\_Default were equal in both splits.

**Figure 12**

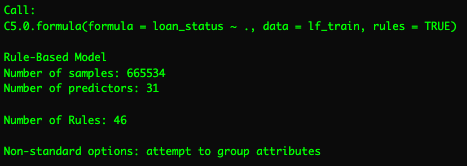
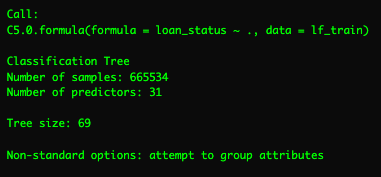
*Using random sampling to split the dataset into a 75%-25% training-testing pair of datasets.*



**Training and Testing the Model**

The C.50() function was utilized to train the decision tree model. Two distinct models were trained. The first model containing a singular decision tree, in which decisions are reported as either if-else statements or lists of rules. The second model that utilizes adaptive boosting, in which a set number of decision trees are built based upon the ‘’weaknesses of the prior trees, and the classes made by each tree are tallied. This second model used ten trees, which is a common industry standard (Lantz, 2023, p. 170-171).

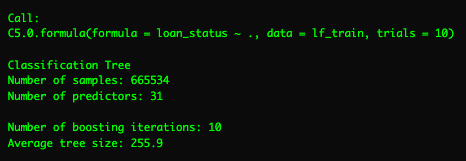
**Figure 13a and 13b**

*Print of the main attributes of the resulting if-else and rule-based decision trees.*

**13a. 13b.**

**Figure 14**

*Ten-trial decision tree with adaptive boosting.*



Even though creating a model out of training model is important for examining the predictive power of different variables on that data set, it is important to generalize to unseen datasets. Thus, the testing dataset for both the regular and adaptive boosted decision trees is combined with the predict() function.

**Figure 15**

*Example syntax to test the model on unseen training data.*



**Understanding the Results**

**Interpreting The Models**

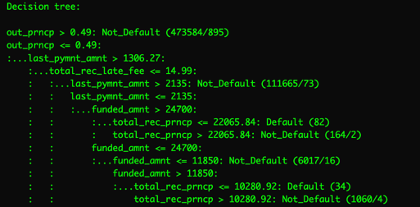
The summary() function, when applied to each model, outputs a detailed series of decisions that make up the decision tree. For the regular C5.0 tree, the non-rule tree creates a tree size of 69. For the rule-based C5.0 tree, 46 rules are created.

The output of summary(loan\_model) can be translated in English as follows:

* If the outstanding principal is greater than 0.49 (i.e. the borrower has repaid more than 51% of the original loan), classify as “not\_default.”
* Else, if the outstanding principal is less than 51%...
  + ...And the last total payment amount is greater than 1306.27...
  + ...and the total recorded late fee is less than 14.99...
  + ...And the last total payment amount is greater than 2135...
  + ...classify as “not\_default.”

**Figure 16**

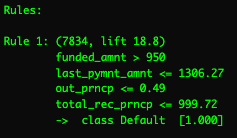
*A small snippet of the single-trial decision tree.*



A similar structure also applies to the rule tree:

**Figure 17**

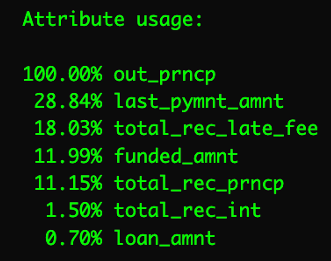
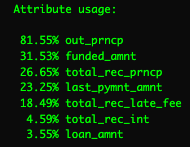
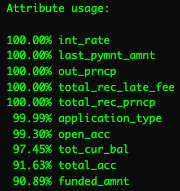
*A small snippet of the single-trial decision tree with rules.*



The attributes usage section of each tree summary provides a summary of the most important predictor variables in each model. The percentages refer to the percentage of rows that factor in a specific variable when making a split decision. In the decision tree without rules (figure xa), for instance, outstanding principle is used for 100% of the decisions, last payment amount if used for 28.84% of the decisions, and total received late fee is used for 18.03% of the decisions. A similar principle applies to the rules tree, and the adaptive boosting tree.

**Figures 18a, 18b, 18c**

*A small of the single-trial decision tree with without rules (18a), with rules (18b), and the first ten attributes of the adaptive boosting tree 18c).*

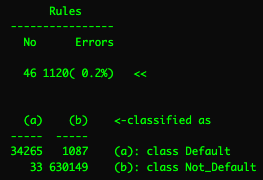
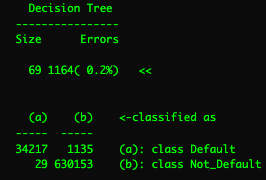
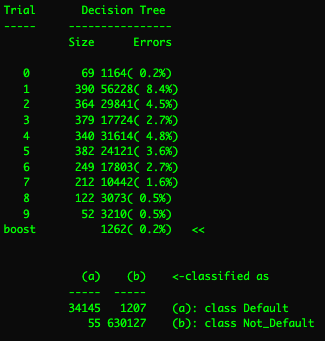
**18a.**  **18b. 18c.**

The most important part of the model is examining its predictive accuracy. The following graphs represent the confusion matrix of the training data for the three trees. Important observations include:

* Of the 665534 cases in the training data, 1120 of cases were misclassified in the single-trial decision tree, 1164 in the rules tree, and 1262 in the adaptive boosting tree.
* The predictive accuracy of all three models is about 99.8%, but each successive model does worse than the previous. It is likely that each model is severely overfitted to the data.

**Figures 19a, 19b, 19c**

*Confusion matrices of the single-trial decision tree with without rules (19a), with rules (19b), and the first ten attributes of the adaptive boosting tree (xc).*

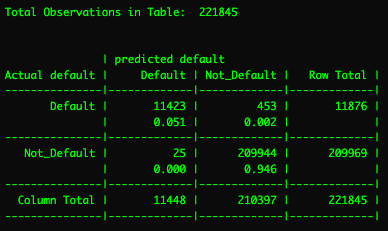
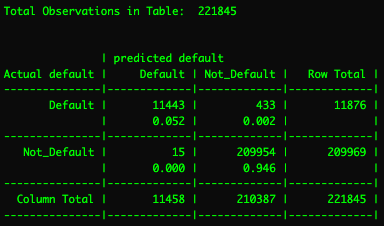
  

**19a. 19b. 19c.**

The cross-tabulation tables of the single-trial decision tree and adaptive boosting tree applied to the testing data are below. It is again significant that both models are overfit, and that the adaptive boosting performed slightly worse than the single-trial decision tree.

**Figures 20a, 20b**

*Cross tabulation tables of the single-trial decision tree without rules (20a) and the adaptive boosting tree (20b).*



**20a. 20b.**

**Interpreting The EDA**

The results of the exploratory data analysis reveal key insights into loan characteristics. The KDE graph shows that the loan amounts for charged off and default loans have similar distributions, suggesting that loan amount alone is not a strong predictor of loan status. The proportional distribution of loan terms indicates that most loans have a three-year term, but default loans have a slightly higher proportion of five-year terms, hinting at increased risk for longer-term loans.

The box plots highlight that default loans have less variance than charged off loans, which may indicate inconsistencies in labeling. This inconsistency could be due to more aggressive labeling of charged off loans, possibly including loans that might be repaid. These findings suggest that closer monitoring of longer-term loans and improved consistency in labeling could enhance risk management strategies. Overall, the analysis provides useful insights for refining lending practices.

**Recommendations**

The results of the models point to a few key attributes of the dataset that may be important indicators of the risk of defaulting:

1. Outstanding principal is the largest predictor of default status. Most specifically, if the outstanding principal for a borrower’s loan falls below 0.49, then this significantly increases the likelihood that the borrower will default on their loans. This makes sense, given that if less than half of the loan is repaid, then it increases the likelihood that the full loan is never fully paid off.
2. The last payment amount for a loan is another significant predictor of default status. For the single-trial decision tree, if the last payment amount if below 1306.27 USD, then that significantly increases the risk of defaulting on one’s loan. However, the last payment amount can only be firmly interpreted in the context of the borrower’s behavior, and further detailed analysis of the link between last payment amount, the frequency of repayment, and the outstanding balance is required to make a more accurate assessment.
3. Total received principal to date is also a significant contributor to default status. Total received principal is also difficult to interpret in isolation, and further analysis would need to be made to payment behavior and larger loan amounts to make a more accurate assessment.

To encourage successful loan paying behavior, strategies that LendingClub may enact include:

1. Encourage borrowers to more frequent methods of loan payments, such as on a monthly basis. This is especially beneficial for the borrower, as it can reduce the amount of interest that the borrower must pay, and increase the borrower’s credit score (RegionsBank, n.d.). It may also be a good idea to encourage or offer bonuses for those borrowers who make extra loan payments.
2. Small loan payments may be a sign of financial hardship on the borrower. Small loan payments may not be desirable for the borrower, as a lack of decrease in the outstanding loan balance can result in higher interest rates over time (Suknanan, 2023). Banks can undertake initiatives such as financial education and early intervention systems to identify at-risk borrowers and offer assistance to encourage loan repayments (Everfi, 2024).
3. Outstanding principal can extend the repayment period, which can increase the risk of default for borrowers. Offering flexible repayment plans that are specifically designed for the borrower’s financial situation may help maximize chances of repayment (Barboni & Theys, 2023).

**Retrospective**

**Chance**

There were a few challenges that faced our group, including a missing member and a tight schedule. Our biggest success was that we were able to adapt to these challenges and create a presentation that we can all be proud of. I enjoyed making my part of the EDA in python and learning about what a KDE plot is. Something different that I decided to do with this project compared to my midterm project was the approach. This time I talked less and listened more to what my group members had to say. I believe this helped the ball get rolling faster than my last project. Other than improving my EDA skills and getting the opportunity to work with my peers, the biggest takeaway for me is to use this same approach in my future team projects.

**Clara**

I did this project with very little directive, outside of the textbook resource that I utilized to learn how to make a decision tree (Lantz, 2023, p. 148-150). I was tasked within my group to handle all of the data processing and modeling as it relates to machine learning, for I was the only group member who had any knowledge of ML.

Overall, I am both very proud, and very frustrated with the results of my work. I am proud insofar as I had begun studying machine learning techniques in R a month prior to the start of the project, in the desire to explore applying machine learning techniques to a biology lab I am part of (in which all tools of analysis are written in R). I picked a decision tree primarily as I was not familiar with categorical variable re-encoding, and I did not feel like I had any targeted and knowledgeable human resource to get into the nitty-gritty of my methodology. I am frustrated insofar as I am entirely aware that the resulting decision tree, while very cool to look at, is also extremely overfitted, which made interpreting the results much more difficult than anticipated.

Regardless of the outcome, I am at least proud that I tried to challenge myself to learn advanced material, even if the result was not as great as I would’ve liked it to be. The amount I learned about machine learning through doing this hands-on project was staggering, and I’m eager to apply new techniques I've learned to future projects.

**Addison**

As my concentration is in social data analytics, I didn’t find myself with much interest in banks, loans, etc. I was originally worried about going into a field that is unfamiliar to me, but I found the differentiation very fun to explore. It was interesting to learn about how loans work as its something I will run into in the near future as an adult. In terms of data analysis, I focused a lot of my time on making presentable visualizations through R as my previous visualizations have been hard to read for a general audience. I did run into personal problems with trying to create visualizations of the varying date variables. They were extremely convoluted and hard to work with as it ranged from such a long period of time. It’s something I especially want to expand on as I think it’s important to notice trends through time periods. I also enjoyed my team as they were always available for communication and collaboration.

**Citations**

Abdou, H. A., Pointon, J., & El‐Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems With Applications*, *35*(3), 1275–1292. https://doi.org/10.1016/j.eswa.2007.08.030

Barboni, G., & Theys, N. (2023, July 3). *The impacts of flexible repayment schedules: Evidence from borrowers and lenders in India*. VoxDev. https://voxdev.org/topic/finance/impacts-flexible-repayment-schedules-evidence-borrowers-and-lenders-india

Duggal, N. (2023, February 20). *Advantages of decision trees*. Simplilearn.com. https://www.simplilearn.com/advantages-of-decision-tree-article

Everfi. (2024, May 1). *7 Reasons financial education programs are important for your bank*. EVERFI. https://everfi.com/blog/financial-education/5-reasons-your-bank-needs-to-be-thinking-about-financial-education/

FasterCapital. (2024, March 25). *Loan Decision trees: How to use decision trees to represent and analyze your loan performance - FasterCapital*. https://fastercapital.com/content/Loan-Decision-Trees--How-to-Use-Decision-Trees-to-Represent-and-Analyze-Your-Loan-Performance.html

Hogarty, S. (2022, November 9). *Decision trees: Definition, analysis, and examples*. Ideas. https://www.wework.com/ideas/professional-development/business-solutions/decision-trees-definition-analysis-and-examples

IBM. (2022, April). *What is a Decision Tree? | IBM*. IBM Think. https://www.ibm.com/topics/decision-trees

Kuhn, M. (n.d.). *C5.0 Decision Trees and Rule-Based Models*. GitHub. https://topepo.github.io/C5.0/

Lakeview Law Group. (2020, November 30). *Understanding the Difference Between Delinquency and Charge Off*. https://lakeviewlawgroup.com/blog/delinquency-and-charge-off/

Lantz, B. (2023). *Machine Learning with R - Fourth Edition* (4th ed.). Packt Publishing. https://www.packtpub.com/product/machine-learning-with-r-fourth-edition/9781801071321

Liberman, N. (2020, May 21). Decision trees and random forests - towards data science. *Medium*. https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

*Living Wage Calculator*. (n.d.). https://livingwage.mit.edu/

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series. Materials Science and Engineering*, *1022*(1), 012042. https://doi.org/10.1088/1757-899x/1022/1/012042

RegionsBank. (n.d.). *What are the advantages of increasing monthly payments?* https://www.regions.com/insights/personal/debt-calculators/what-are-the-advantages-of-increasing-monthly-payments

Suknanan, J. (2023, March 22). What happens if you pay off a personal loan early? *CNBC*. https://www.cnbc.com/select/can-you-pay-off-a-personal-loan-early/

USNews. (n.d.). *LendingClub*. https://www.usnews.com/banking/reviews/lendingclub