Good Borrower or Bad Borrower?

Predicting Loan Default with Machine Learning

Jillian Goodwyn, DSA 601 Spring 2021

The Big Questions

1. *Can we predict who will and will not be a reliable borrower?*
2. *What are the most important attributes that help predict whether a borrower will default on their loan?*

It is essential for lending companies to identify reliable borrowers with confidence. If too many unreliable borrowers are lent money, loan companies cannot survive. On the other side of the agreement, individual borrowers are often able to change their lives with these loans. Therefore, granting loans to responsible borrowers is a win-win. Accurate predictions of borrower default can directly contribute to a healthy lending ecosystem.

Furthermore, it is important to understand what is happening under the hood of such a predictive model. While it is advantageous to have a highly accurate model, it is equally valuable to have a model that is interpretable. Insight into the factors that signify reliable (and not-so-reliable) borrowers allow lending companies to make strategic business decisions.

About the Data

- All data is from **LendingClub**, a USA-based peer-to-peer lending company that allows individual borrowers to request unsecured personal loans up to $40,000

- Data snapshot from **2014** to ensure enough time has passed for loans to be fully resolved for complete data analysis
- **235,607** unique records (one row per borrower)
- **146** variables that pertain to differing aspects of the borrower, the loan, and the conditions surrounding the loan, for example:

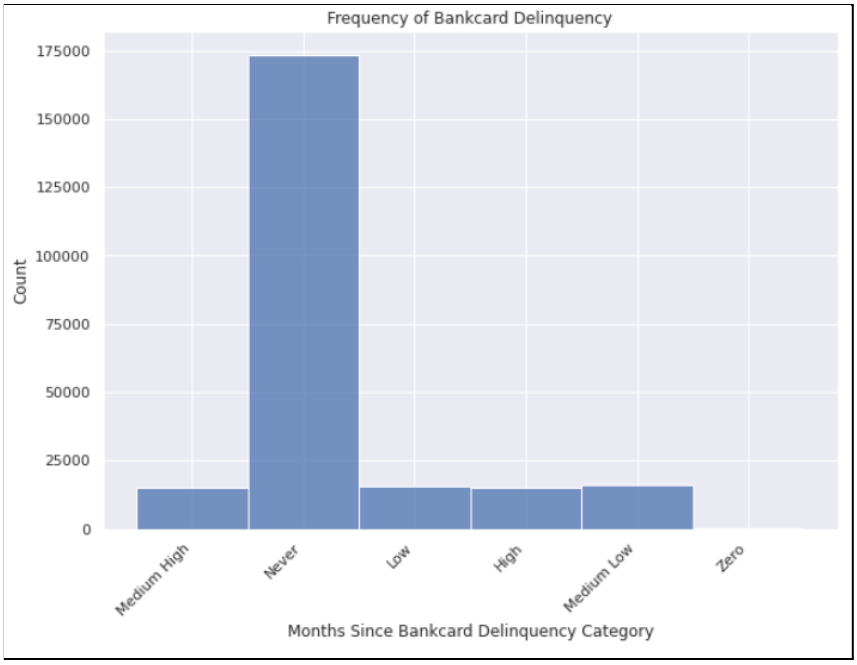
- Borrower: income, employment, home ownership, credit scores, previous debt, credit history, revolving balances
- Loan: amount requested, fees if unpaid, length of time for payment, purpose of loan

---

Methods

A variety of data exploration, cleaning, analysis, and modeling techniques were used to answer our 2 big questions. Of the 146 original input variables, about half (70) were usable for prediction. For example, variables like '*loan issue date*' that are only available after the loan is granted cannot be used to predict who to give a loan to. This would defeat the purpose of predicting good and bad borrowers prior granting loans. Additionally, any variables that were 100% missing (such as any variables about secondary applicants on the loan) were not usable.

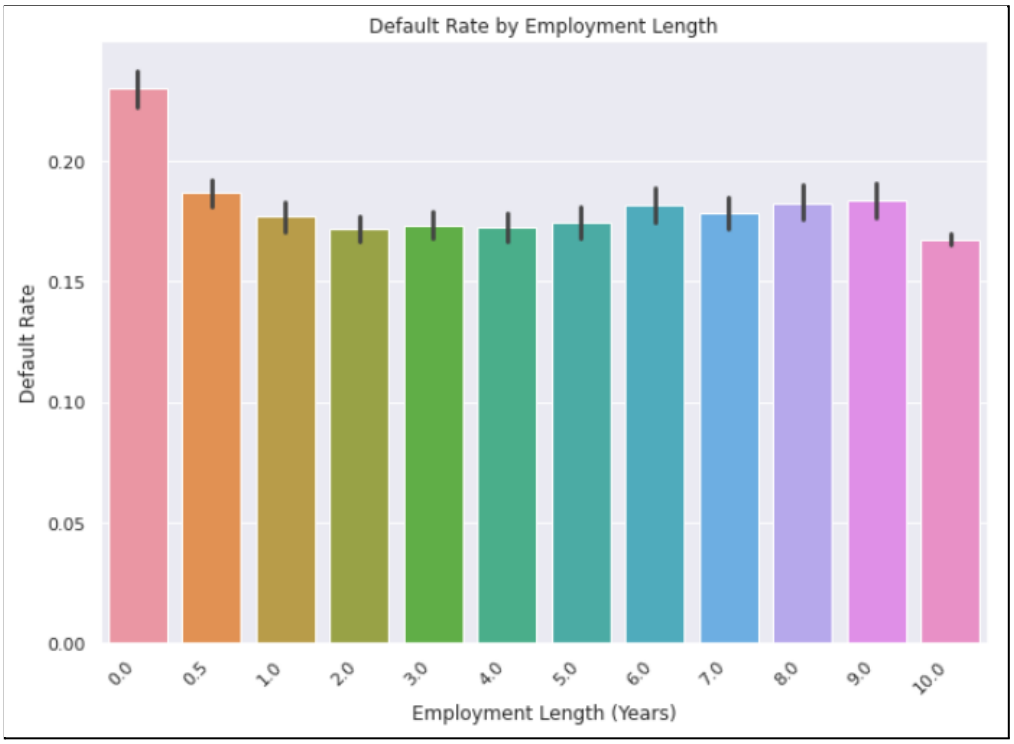
Data Cleaning Summary



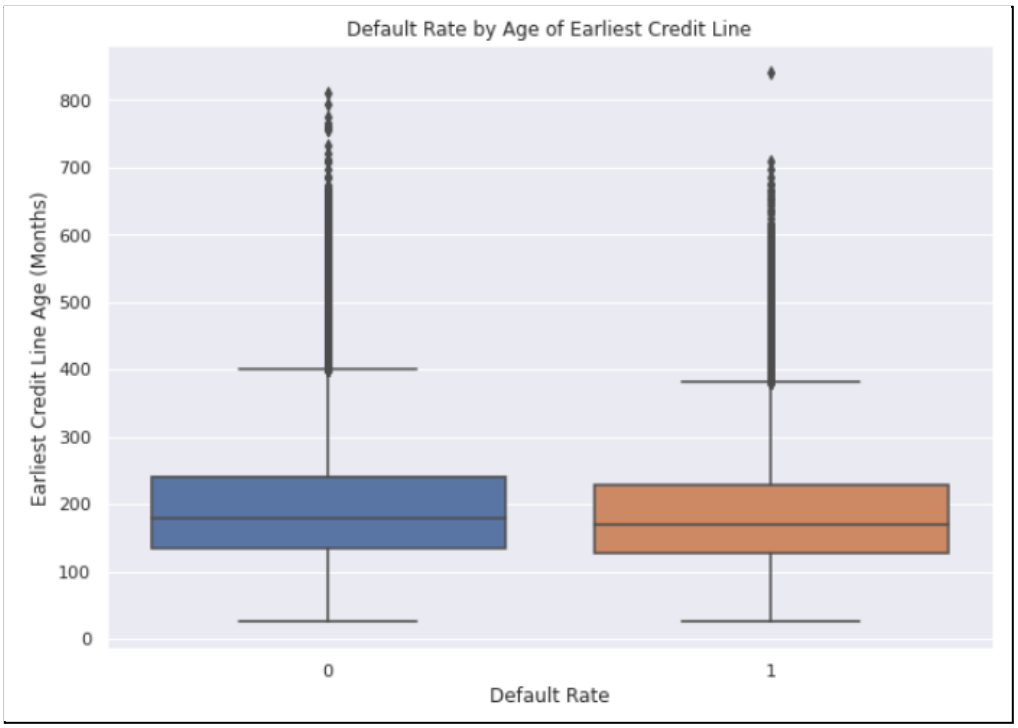
As expected, there was quite a bit of data cleaning! Several fields such as '*employment length*' and '*revolving utilization*' needed to be reformatted so they read in as numbers instead of text. Furthermore, several fields had the tricky issue of missing (blank) values; however, the missing values were actually positive behaviors. For instance, if '*months since last bankcard delinquency*' was missing for a borrower, it actually meant the borrower never had a delinquent bankcard. To address variables like this, the variables were bucketed by their quartiles into "low," "medium-low," "medium-high," and "high" and the missing records were categorized as "never." This helped the models differentiate the underlying behaviors in these variables.

Data Exploration

To acquire a better understanding of the data and which variables might be good predictors of reliable borrowers, visualizations and aggregations were used to summarize and dig into the data. Promising variables like '*FICO score*,' '*Employment Length*,' '*Age of Earliest Credit Line*,' and many others were graphed against default rates to see if any patterns arose.



'*Employment length*' appeared promising: default rates seem higher for those employed for less time



'*Age of Earliest Credit Line*' was less promising: the distribution is similar for borrowers who defaulted and those who did not default

Identify Variables for Modeling

Pairs of independent variables that were highly correlated to one another were identified, and the variable that was least correlated with borrower default was removed. This was necessary to reduce noise from these co-correlated variables that would otherwise decrease model accuracy.

Then a simple linear regression model was applied to identify variables that contributed most to predicting loan default. These either increased or decreased the likelihood of a borrower to default on their loan. For example, '*average FICO score'* decreased the likelihood of default, meaning the higher the '*average FICO score,'* the less likely a borrower was to default. The final list of input variables were as follows:

Increase Likelihood of Default:

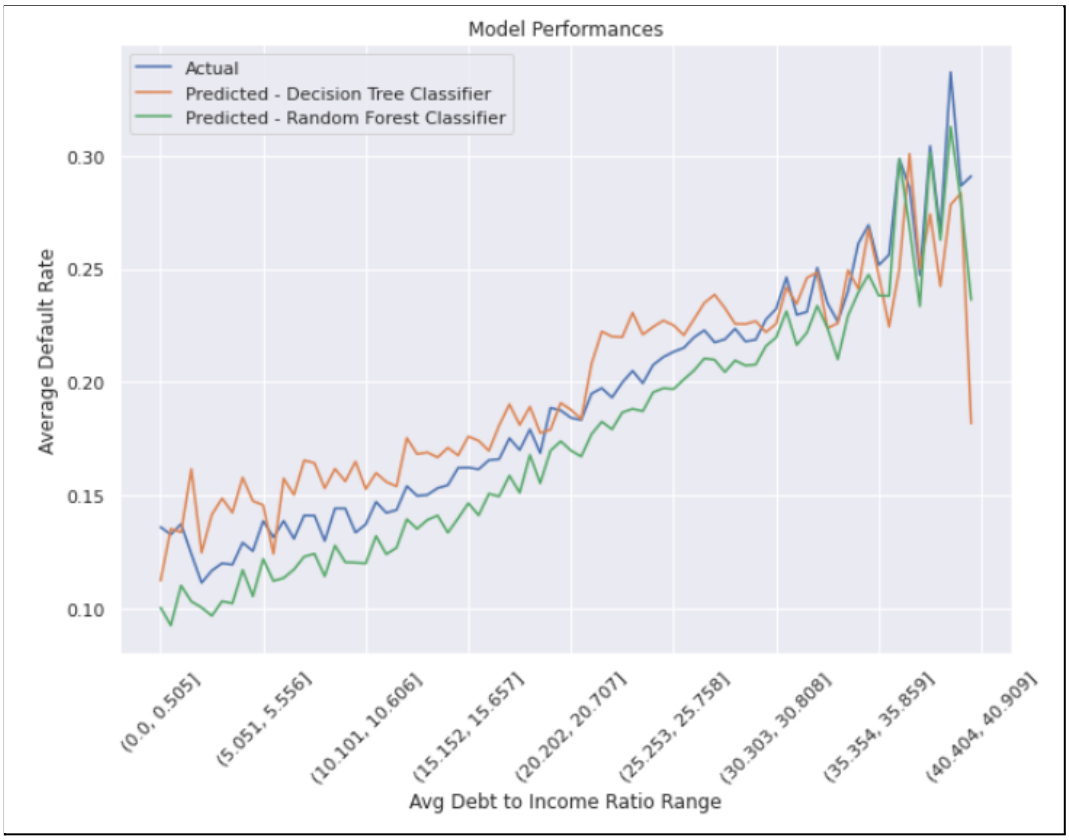
- Higher number of open credit accounts
- Higher loan amount requested
- Higher debt-to-income ratio
- Higher number of inquiries in past 6 months
- Higher number of trades opened in the past 24 months
- Higher months since oldest revolving account opened **(surprising!)**

Decrease Likelihood of Default:

- Higher number of delinquencies more than 30 days in the past 2 years **(surprising!)**
- Higher Number of revolving trades with balance greater than 0
- Higher number of public record bankruptcies **(surprising!)**
- Longer employment title text
- Higher average FICO score
- Higher annual income

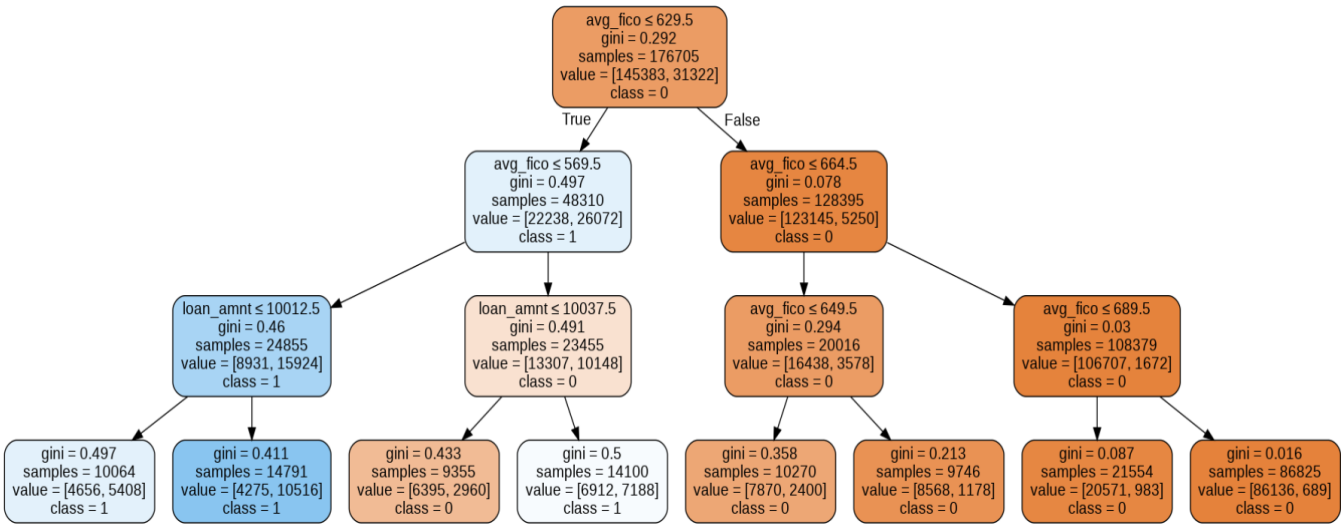
Modeling

Once the most influential predictors were identified, the dataset was split into training and test datasets and a Decision Tree Classifier and a Random Forest Classifier were applied to predict borrower default. With approximately 87% accuracy on both the training and test sets, the Decision Tree Classifier produced reliable results. The Random Forest Classifier achieved 99% accuracy on the training dataset but 86% on the test data, indicating overfitting. With more reliable results, the Decision Tree Classifier was chosen as the "winning" model.



We can see the performance of the Decision Tree Classifier (orange) versus that of the Random Forest Classifier (green) compared with the original default data (blue): the '*debt to income ratio*' variable is used as a baseline to display these differences

Using the Decision Tree Classifier, a visual decision tree was then created to highlight the most important variables that the model uses to determine probability of default.



The visual decision tree breaks the variables down into nodes with decision rules. In this case, it is saying:

- If the borrower's '*average FICO score'* is greater than 630, the borrower is **unlikely** to default
- If the borrower's '*average FICO score'* is less than or equal to 630, the borrower **is likel**y to default
- If the borrower's score is less than 630 but more than 570 and the requested loan amount is under approximately $10,000, the borrower is **unlikely** to default

---

Discussion

While most of the model results were as expected (higher annual income and FICO score decrease the likelihood of borrower default), there were some surprising results. For example, it was surprising that having higher months since the oldest revolving account opened actually increased the likelihood of default. In most credit score calculations, it is generally noted that having older accounts actually helps prove reputability.

Additionally, having a higher number of delinquencies in the past 2 years and/or a higher number of public record bankruptcies actually decreased the likelihood of default. It would be common sense if this were the opposite; however, perhaps either of these behaviors indicates a stronger need for the loan requested for the borrower.

Limitations

- Only 2014 data was used for this analysis and it is not certain that it could be extrapolated to current or future data. Economics, policies, and companies change drastically over time. In fact, LendingClub did experience issues from a scandal from 2016-2017, and this would greatly impact any current predictions made on new data
- Because secondary applicant data, among many other data points, were not available for this analysis, the results and conclusions are limited only to the complete variables that were available in the dataset
- This dataset was built on data points from one company, and it therefore may not be appropriate to apply any findings or conclusions to other companies.

Future Considerations

- ❏ Because most borrowers do not default, it may be useful to employ oversampling techniques to increase the number of examples of default (this could also help with overfitting)
- ❏ Addition of any other data points, such as secondary applicant data and borrower demographics