Predicting LendingClub Default Rates and Loan Performance by Borrower Type

Introduction

This paper examines approximately 470,000 LendingClub loans from 2007 to 2014 to evaluate how loan variables other than Loan Grade impact default rates and performance.

The conclusion of this report is that loan variables Term, Loan Purpose, Home Ownership, Income, and Location have a significant impact on default rates and performance and should be factored into lending decisions.

This report is intended for LendingClub Lenders and LendingClub Executives, or any person that might find it interesting.

Data Description

The data is collected from LendingClub's website (https://www.lendingclub.com/info/download-data.action) and consists of 470,000 records and 145 columns. For the Analysis section of this report, loan data is grouped by the variables in Table 1 below

Data Wrangling

Data wrangling for this report included replacing NaN values with 0, grouping income into buckets (<=$50k, >$50k-$100k, >$100k-$200k, >$300k), changing the encoding parameter of the data to ISO-8859-1 and using the skiprows parameter to remove blank rows in the imported Excel files, changing the Loan Term column from object type to string type to filter by rows containing 36 month or 60 month terms, and consolidating the loan purposes of Debt Consolidation and Credit Card to one loan purpose called Debt Consolidation.

Calculations

Performance is calculated as (Total $ Payments) / (Total $ Funded) for each variable group and is not annualized. **Default Rate is calculated as (Count of Charged Off Loans) / (Count of Total Loans)** for each variable group and is not annualized**.** For the purpose of this report, both performance and default rate are used to understand relative performance. LendingClub's official performance calculation is located here: https://www.lendingclub.com/info/statistics-performance.action.

Variables Used to Analyze Loan Data

Table 1

Loan Term
60 months
36 months

Loan Purpose
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding

Borrower Home Ownership
RENT
OWN
MORTGAGE

Borrower Income Group
<$50K
>$50K–$100K
>100K–200K
>$200K

Borrower Location
State

Loan Grade

As see in Figure 1, approximately 70% of the data is allocated across Loan Grade A, B, and C. A-grade loans have a default rate close to 5% and G-grade loans have a default rate close to 35% (Figure 2). Performance is fairly consistent across grade due to the higher interest rates charged on lower grade loans (Table 2). As seen in Table 3, the mean return of an A-grade loan is 7.2% and the mean return of a G-grade loan is 6.6%.

The Data Analysis section starting on the following page (page 3) reviews the LendingClub data by Term, Loan Purpose, Home Ownership, Income, and Location. The trends noticed are consistent across loan grade. An addition to this report could repeat the analysis reviewing only one loan grade such as A or B.

Figure 1

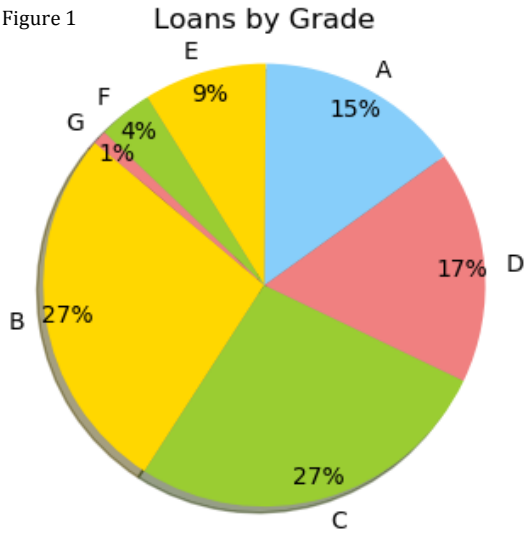


Figure 2

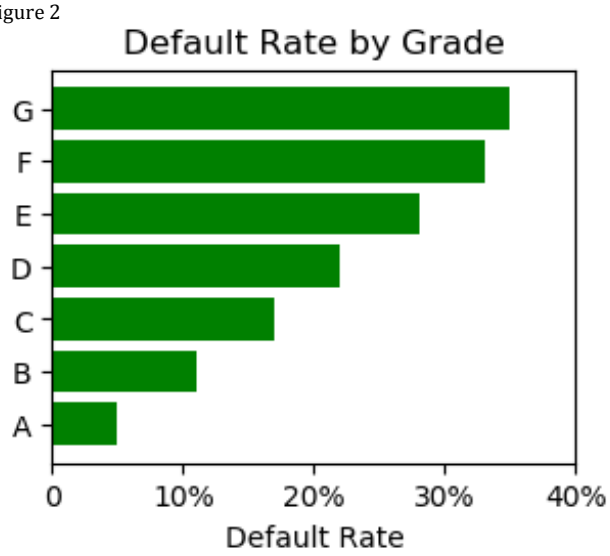


Table 2

Grade	Average Interest
A	7.3%
B	10.9%
C	13.4%
D	15.2%
E	16.4%
F	17.7%
G	19.0%

Table 3

Grade	Mean Performance
A	7.2%
B	9.1%
C	8.7%
D	7.9%
E	8.1%
F	9.7%
G	6.6%

Data Analysis

Term: Approximately 62% of the LendingClub loans from 2007-2014 were 36-month term loans (Figure 3). The shorter term of 36-months sharply reduces default rates. Approximately 13% of 36-month term loans were charged off, compared to 23% for 60-month term loans (Figure 4). The mean return of 36-month term loans from 2007-2014 is 9.5%, compared to 6.0% for 60-month term loans. A t-test indicated that the performance of these two populations is significantly different with 99% confidence.

Figure 3

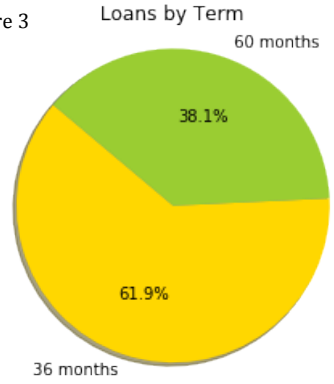
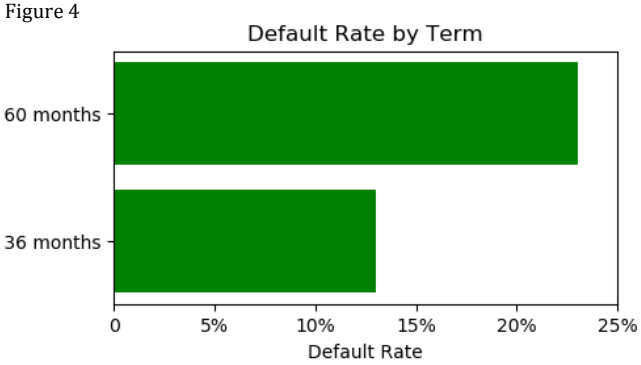


Figure 4



Purpose: Approximately 80% of the LendingClub loans from 2007-2014 were for the purpose of debt consolidation (Figure 5), with 16% of those loans defaulting (Figure 6). The default rate of 16% is in the middle range - loans with the purpose of small business have the highest defaults at 25%, while loans for cars and weddings are closer to 11% (Figure 6).

Figure 5

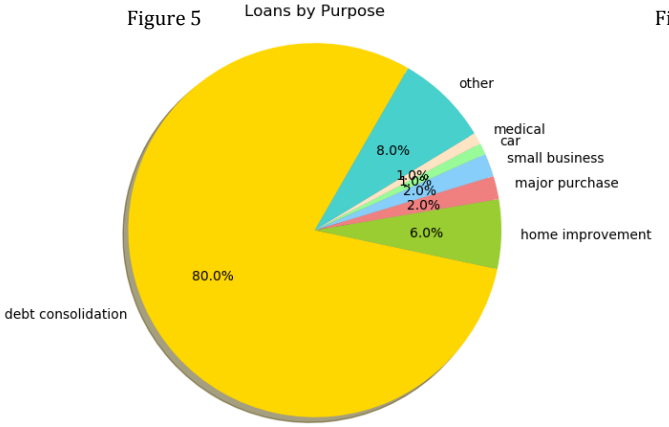
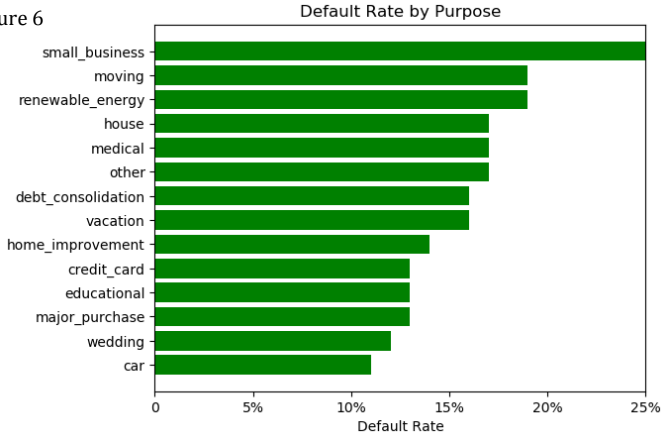


Figure 6



Income: Approximately 53% of the LendingClub loans from 2007-2014 were to individuals with income from $50k to $100k (Figure 7). Loans to individuals making less than $50k a year had the highest default rates at 18% (Figure 8). Defaults drop sharply as income rises. The mean performance of loans to individuals with annual income below $50k is 7.3%, compared to a return of 8.9% for loans to individuals making more than $50k. A t-test indicated that the performance of these two populations is significantly different with 99% confidence.

Figure 7

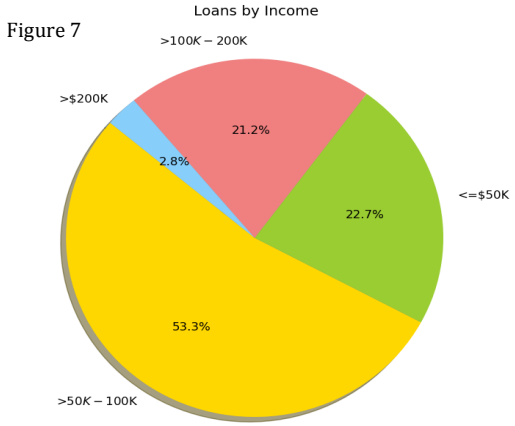
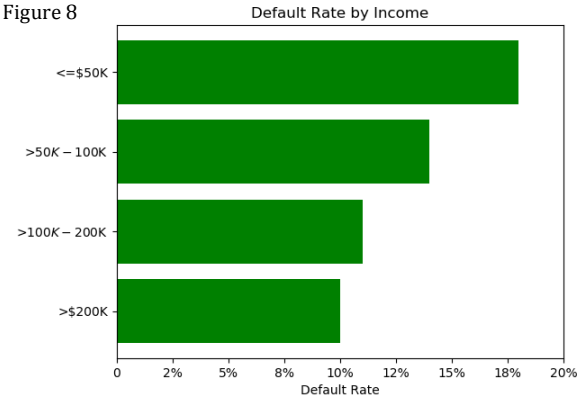


Figure 8



Home Ownership: Approximately 56% of the LendingClub loans from 2007-2014 were to individuals paying a mortgage (Figure 9). Renters have slightly higher default rates of 18%, compared to 16% for owners and 14% for individuals still paying a mortgage (Figure 10). The mean performance of loans to renters is 7.8%, while the mean return of loans to non-renters is 8.8%. A t-test indicated that the performance of these two populations is significantly different with 99% confidence.

Figure 9

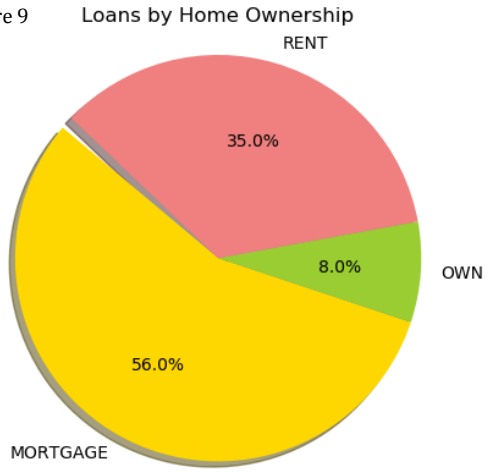
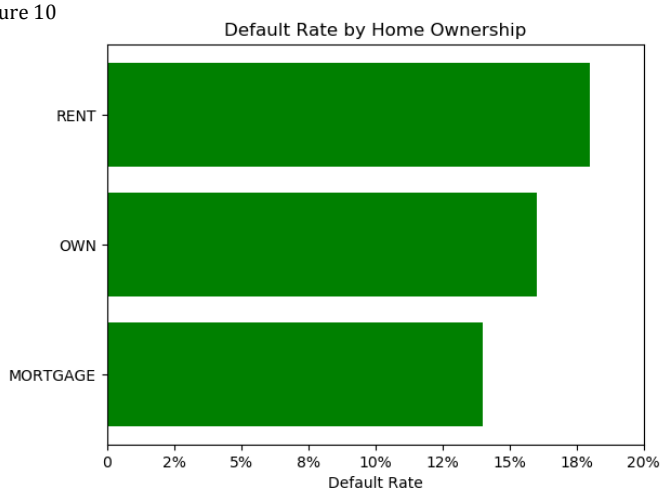


Figure 10



State: Figure 11 gives a breakdown of loans by state. The blank squares in the upper right corner represent states with less than .5% allocation (NH, RI, DC, AK, MT, MS, DE, WY, SD, VT, MS, NE, IA, ID and ME). We can see from the graph that roughly 35% of loans are allocated to CA, NY and TX.

Figure 11

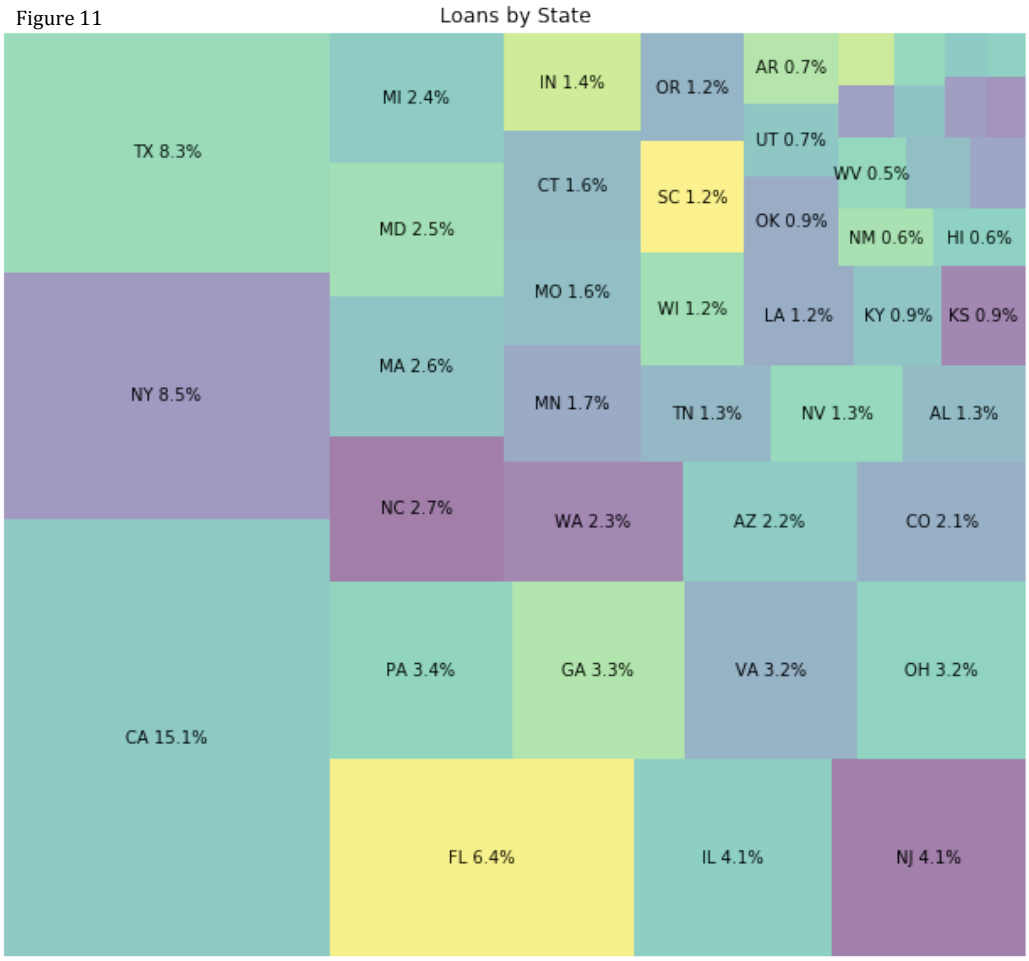
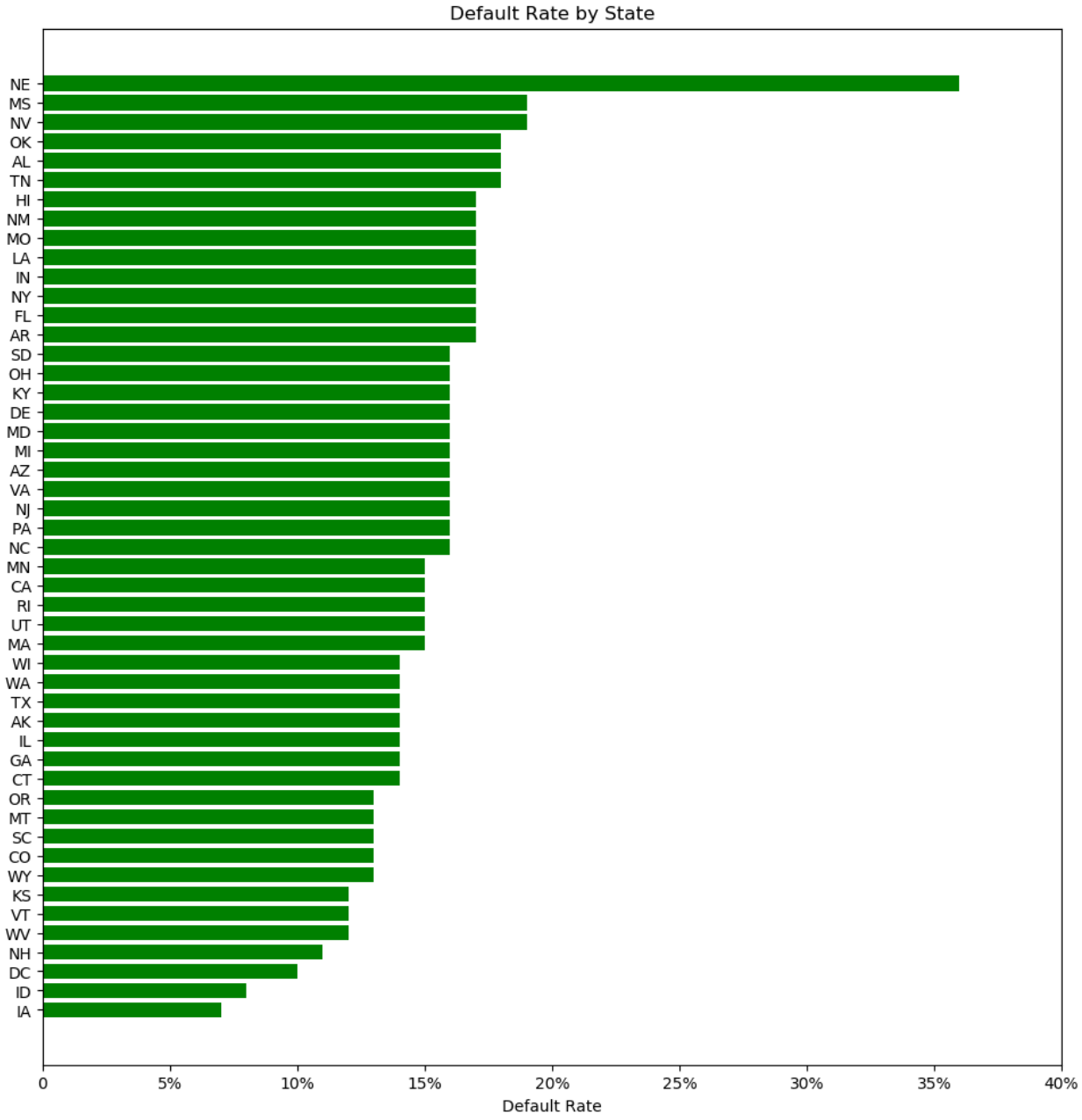


Figure 12 gives a breakdown of defaults by state. Nebraska (NE) has a very high default rate above 35%. There were only 14 loans to borrowers in NE from 2007-2014 and 70% of those loans were to loan grade D and E. Nevada (NV) consistently has higher default rates. The District of Columbia (DC) and states like Kansas (KS), West Virginia (WV), Wyoming (WY) and Vermont (VT) consistently have lower default rates.

Figure 12



Logistic Regression

A machine-learning model using logistic regression was fit to the 2007-2013 data. The input variables were Term, Purpose, Home Ownership, and Income. The target variable was Default in a binary format (1 for No Default and 0 for Default). The model was trained on LendingClub data from 2007-2013, and then the model was used to select loans in the 2014 dataset that, based on the 2007-2013 data, had a probability of defaulting that is less than 10%.

[The model is from the scikit-learn machine learning Python Package (Code: from sklearn.linear_model import LogisticRegression). The model fit accuracy is .85. Please see Jupyter notebook for code.]

Table 4

	2014 All Loans	2014 Loans Selected by Model	Variance
Loan Count	235,627	42,807	-192,820
Default Rate	15.98%	13.89%	2.09%
Mean Return	5.60%	7.50%	1.90%

As seen in Table 4, the model selected 42,807 loans from the 235,627 loans issued in 2014. The loans selected by the model show an improvement in default rate and performance. The default rate for the portfolio selected by the model is 13.89%, compared to a default rate of 15.98% for the entire portfolio. The mean performance for loans in the portfolio selected by the model is 7.50%, compared to a mean return of 5.60% for loans in the entire portfolio.

Logistic Regression Adjusted for Class Imbalance

To support the above analysis, a second machine-learning model using logistic regression, but adjusting for class imbalance, was fit to the 2007-2013 data.

Class imbalance occurs in logistic regression when the target variable, in this case 1 for No Default and 0 for Default, is skewed. Since roughly 80% of the LendingClub data is class 1 (No Default), there is a class imbalance: the model could predict No Default 100% of the time and still have an 80% accuracy score.

To adjust for class imbalance class 1 (No Default) was randomly reduced to 34,692 records, making it equal to the 34,692 class 0 (Default) records. The accuracy of the logistic regression model is reduced to 60%. The model still does well selecting loans. The default rate for the portfolio selected by the model is 12.14% and the mean performance for loans in the portfolio selected by the model is 8.32% as seen in Table 5.

Table 5

	2014 All Loans	2014 Loans Selected by Model	Variance
Loan Count	235,627	7,387	-228,240
Default Rate	15.98%	12.14%	3.84%
Mean Return	5.62%	8.32%	2.70%

Conclusions & Recommendations

The conclusion of this report is that considering past default rates for certain loan variables is beneficial to future lending strategies. From 2007-2014, there are clear discrepancies in default rate and performance by Term, Purpose, State, Home Ownership, and Income. Further supporting this analysis, two logistic regression models considering these variables were shown to outperform. From 2007-2014, loans for the purpose of Small Business, Renewable Energy, and Moving have higher default rates; loans to certain states such as Nevada (NV) and Mississippi (MS) have higher default rates; loans to renters (myself included) and individuals making less than $50k have higher default rates; and, finally, loans with longer durations of 60 months have higher default rates. The recommendation of this report for investors is to consider more variables than Loan Grade and ensure that you are being compensated for taking additional risk.