ProsperLoan Data Exploration

by Maxwell Ofoegbu

Dataset

There are 113937 Prosperloan data entries with 80 features (ListingNumber, ListingCreationDate, CreditGrade, Term, LoanStatus, ClosedDate, BorrowerAPR, BorrowerRate, LenderYield, EstimatedEffectiveYield, EstimatedLoss, EstimatedReturn, ProsperRating (numeric), ProsperRating (Alpha), ProsperScore, istingCategory (numeric), BorrowerState, Occupation, EmploymentStatus, EmploymentStatusDuration, BorrowerHomeowner, CurrentlyInGroup, GroupKey, DateCreditPulled, CreditScoreRangeLower, CreditScoreRangeUpper, FirstRecordedCreditLine, CurrentCreditLines, OpenCreditLines, TotalCreditLinespast7years, OpenRevolvingAccounts, OpenRevolvingMonthlyPayment, InquiriesLast6Months,TotalInquiries, CurrentDelinquencies, AmountDelinquent, DelinquenciesLast7Years, PublicRecordsLast10Years, PublicRecordsLast12Months,RevolvingCreditBalance, BankcardUtilization, AvailableBankcardCredit,TotalTrades, TradesNeverDelinquent (percentage), TradesOpenedLast6Months,DebtToIncomeRatio, IncomeRange, IncomeVerifiable, StatedMonthlyIncome,LoanKey, TotalProsperLoans, TotalProsperPaymentsBilled, OnTimeProsperPayments, ProsperPaymentsLessThanOneMonthLate,ProsperPaymentsOneMonthPlusLate, ProsperPrincipalBorrowed,ProsperPrincipalOutstanding, ScorexChangeAtTimeOfListing,LoanCurrentDaysDelinquent, LoanFirstDefaultedCycleNumber,LoanMonthsSinceOrigination, LoanNumber, LoanOriginalAmount,LoanOriginationDate, LoanOriginationQuarter, MemberKey,MonthlyLoanPayment, LP_CustomerPayments, LP_CustomerPrincipalPayments, LP_InterestandFees, LP_ServiceFees, LP_CollectionFees, LP_GrossPrincipalLoss,LP_NetPrincipalLoss, LP_NonPrincipalRecoverypayments, PercentFunded, Recommendations ,InvestmentFromFriendsCount , InvestmentFromFriendsAmount and ,Investors). Most variables are numeric in nature, but the variables ProsperRatingAlpha, ProsperRatingNumeric, and IncomeRange are ordered factor variables with the following levels.

(best) ——> (worst)

ProsperRatingAlpha:AA, A, B, C, D, E, HR, NC,

ProsperRatingNumeric: 7,6,5,4,3,2,1

IncomeRange: 100,000+, 75,000-99,999, 50,000−74,999,25,000-49,999, 1−24,999,

Not employed, Not displayed,

Summary of Findings

The typical borrower rate is increasing as the credit rating falls and excluding the non-numeric income ranges, typically the borrower rate is increasing as the income range increases. (Interesting, and not expected)

Key Insights for Presentation

1. A low interest-loan scheme for loan seekers

2. Borrowers and Lenders have direct participation in the credit marketplace.

3. A limited liability company

Main findings

The frequency distribution of BorrowerRate was observed to be unimodal with a centre around 0.2. The distribution also showed some possible extremely small values at points slightly before the 0.1 mark on the left hand and before the 0.4 mark on the right hand. It was observed that the BorrowerRate below the 0.2 mark is more significant than that above the centre.

The distribution for LenderYield was observed to be partly right skewed with a peak half way between 0.1 and .2.A larger portion of the distribution lies toward the left. The far right is observed to have no values. This may suggest the presence of outliers and therefore necessitated for further investigation down the line.

The distribution for the EstimatedLoss was partly right skewed but drew my attention to further investigate the presence of outliers towards the far right of the distribution.

The distribution of the BorrowerAPR was observed to be unimodal and very similar to that observed from the distribution of BorrowerRate with a centre around 0.2. The distribution also showed some possible extremely small values at points slightly before the 0.1 mark on the left hand and before the 0.4 mark on the right hand.

The initial plot of the LoanOriginalAmount showed some immediate points of attention. On the LoanOriginalAmount you would notice sudden spikes and some other

very high figures in the far right which was worth taking a bit of time to identify possible outliers and see if they need to be filtered out of the data.

The MonthlyLoanPayment appeared partly right skewed with a lot of MonthlyLoanPayment in the lower end. It was observed that the smaller the MonthlyLoanPayment the more payment was made.

We could see from the plot that California is the most BorrowerState followed by three other States namely, Texas, New York and Florida which were just half way of California.

Seeing that the ListingCategory represents various loan category, the category number 1 which, according to the Google Docs, represents Dept Consolidation was ranked highest followed by category zero.

The distribution of Investors values appeared right skewed (long tailed), with relatively few points above 400 in value but there isn't a lot of detail beyond the 400 value. In addition, the bin boundaries were not particularly aligned with the tick marks which made interpretation trickier. However, in the log transform plot it was seen that there are more investors in small loan region. There are basically few investors willing to take the risk of loaning money beyond $500.

I saw a unimodal distribution for the ProsperRating (numeric) with a center at around 4, with the upper half, from rating number 4, dominating the range. However, we could notice a sudden count drop in the prosper rating of 7. This implied there are only a few numbers of persons with a highest ProsperRating. There are generally more people with average ProsperRating of 4.

The ProsperRatingAlpha is defined as the ProsperRating assigned at the time the listing was created between AA-HR. We can notice a slightly opposite distribution between the ProsperRatingNumeric and ProsperRatingAlpha with the greater number of people in the c-rating. We noticed that the NC is zero which suggests the absence of this class of rating within the ProsperRatingAlpha.

The distribution showed that a greater number of the people have some kind of employment with the fulltime-workers kind greater than that of the self-employed kind.

The distribution for the IncomeRange showed that the greater part of the population lies within the $(25,000 -49,999) to (50,000 -79,999) range. There were also a small number of unemployed people in the population. However, there is also a moderate size group whose employment status was not ascertained.

The distribution for the CreditGrade showed a unimodal characteristic with the C grade being the highest.

Just as said earlier, these distributions are in conformity that the greater part of loan seekers fall into the C-group. However, there is also a significant number of people in the AA, A,B,E,HR group. The distribution showed a steep drop in the number of loan seekers occupying the NC group.

For better visibility I didn't have a good reason to associate each feature with a different color. It was better to plot everything in one color to avoid being distracted!

The correlation plot showed that LoanOriginalAmount is highly correlated with MontlyLoanPayment and this should be expected. However, the former is not correlated with any other feature in our distribution. This is in fact a confirmation of the ideology for the ProsperLoan. The LenderYield is observed to be highly correlated with the BorrowerAPR, seeing that the BorrowerAPR is the Borrower's annual percentage rate, it is expected that the LenderYield correlate with it. That is, the more the BorrowerAPR the more yield for the lender. The EestimatedLoss is highly correlated with the BorrowerAPR, the LenderYield, and the BorrowerRate.

The correlation matrix is very sticking as it helps us see the variables that are strongly correlated with our response variable of choice--BorrowerRate. We can see that the BorrowerRate is strongly correlated with BorrowerAPR, LenderYield and the EstimatedLoss. That implies that of the numerical variables we are using for this exploration we can henceforth limit ourselves to only those numerical variables that are correlated with the BorrowerRte. Interestingly, these variables are also strongly correlated with each other.

The listing category and investors showed no correlation with any of the numeric variables in our correlation matrix. However, MonthlyLoanPayment and LoanOriginalAmount show correlation with each other but with no other numeric variable.

The ProsperRatingNumeric shows a strong negative correlation with all of the response variables because it showed same with the BorrowerRate under the multivariate case . We could see that as the ProsperRatingNumeric drops (increases in the negative) the response variables increase in the positive.

We can see that as the ProsperRatingAlpha decreases (increasing in the negative towards NC), the BorrowerRate increases. the BorrowerRate for the Retired class of the EmploymentStatus variable is notice to be one of the least.

Interestingly, the final analyses showed that the BorrowerRate appears to rise with a rise in income range

The ProsperRatingAlpha,ProsperRatingNumeric, IncomeRange, and CreditGrade show very strong correlation with BorrowerRate, as theses predictor variables increase in the positive, the response--BorrowerRate—decrease.

When the income range was measured against the BorrowersRate in our final plot we noticed that with a rise in the income range resulted to increment in BorrowerRate. This is not what was expected at this point one could simply conclude that ProsperLoan is just one of those profit making ventures out there!

Conclusions'

1. The typical borrower rate is increasing as the credit rating falls (interesting and expected), and
2. Excluding the non-numeric Income ranges, typically the borrower rate is increasing as the income range increases (Interesting, and not expected.
3. As ProsperRatingNumeric drops, the BorrowerRate increases. At the same PropsperRatingNumeric of 7, the self employed have the highest BorrowerRate while at the drop of ProsperRatingNumeric below 7, the self employed, employed and full-time have the least BorrowerRate.