README

August 14, 2020

1 Loan Data Exploration

1.1 Dataset

The dataset consisted of 61 attributes of 113,937 loans. The attributes included (e.g. Credit Grade, LoanStatus, Estimated return, Estimated Effective Yield , Estimated Loss Rate P_CustomerPayments, LP_CustomerPrincipalPayments, LP_InterestandFees, LP_ServiceFees, LP_CollectionFees, LP_GrossPrincipalLoss, LP_NetPrincipalLoss and LP_NonPrincipalRecoverypayments) and the key attributes are ListingKey, ListingNumber and ListingCreationDate.

1.2 Summary of Findings

In the exploration, it is found that the income range distribution is normal distribution, with one peak at 32192 that represnts 28.3% of the total samples are classiefied in the income range between 25000 USD and 49999 USD, followed by a 27.3% represents the income range between 50000 USD and 74999 USD.

Also, it is found that the estimated return for almost 40000 client is around 0.08-0.09. The majority of client has a loan with original amont near to 5000 USD.The distribution of LP_CustomerPayments, LP_CustomerPrincipalPayments, LP_NonPrincipalRecoverypayments and LP_InterestandFees are left skwed while LP_ServiceFees and LP_CollectionFees are right skwed. LP_GrossPrincipalLoss and LP_NetPrincipalLoss are non uniform distribution.It seems that the income range has a strong correlation with credit card.

The variables have strong correlation are: 1- EstimatedEffectiveYield has negative correlation with EstimatedLoss, and positive with EstimatedReturn
2- EstimatedLoss has negative correlation with EstimatedReturn
3- EstimatedReturn has positive correlation with EstimatedEffectiveYield, and negative with EstimatedLoss
4- LP_CustomerPayments has positive correlation with LP_CustomerPrincipalPayments
5- LP_CustomerPrincipalPayments has positive correlation with LP_CustomerPayments
6- LP_ServiceFees has negative correlation with LP_CustomerPayments , LP_CustomerPrincipalPayments and LP_InterestandFees
7- LP_GrossPrincipalLoss has positive correlation with LoanOriginalAmount and LP_NetPrincipalLoss
8- LP_NetPrincipalLoss has positive correlation with LoanOriginalAmount and

LP_GrossPrincipalLoss
The variables have no correlation are:
1-LoanOriginalAmount
2- LP_InterestandFees
3- LP_CollectionFees
4- LP_NonPrincipalRecoverypayments So, the 4 columns those have no any correlation with any other variables will not be included in further analysis.

Outside of the main variables of interest, I extended my investigation of income range against all the variable that showed certain degree of corellation by looking at the impact of the three categorical quality features on each other then visulize the impact of the main variable, the income range, on the rest of the numerical variables. The multivariate exploration showed that there indeed is an effect of the income grade on the other loans variables, but in the dataset, this is initially hidden by the fact that higher income were more prevalent in smaller return. Controlling for the estimated loss and return shows the effect of the other LP's of clients on the whole pattern pf data. This effect was clearest for the estimated return and estimated loss variables, with less systematic trends for yield.

1.3 Key Insights for Presentation

For the presentation, I just focus on features that could be affected by the income range, EstimatedEffectiveYield with EstimatedLoss, and with EstimatedReturn. EstimatedLoss with EstimatedReturn. EstimatedReturn with EstimatedEffectiveYield, and with EstimatedLoss. LP_CustomerPayments with LP_CustomerPrincipalPayments. LP_CustomerPrincipalPayments with LP_CustomerPayments. LP_ServiceFees with LP_CustomerPayments, with LP_CustomerPrincipalPayments and with LP_InterestandFees. LP_GrossPrincipalLoss with LoanOriginalAmount and with LP_NetPrincipalLoss. LP_NetPrincipalLoss with LoanOriginalAmount and with LP_GrossPrincipalLos.

There is an interaction effect visible between Estimated Loss , Estimated Return, and the categorical measures of income: Income Range. This is most evident for the return measure.It is clear that the not employed have the highst Estimated Loss and lowest Estimated Return. But what woundring is there is a portion of the people those have income more than 100K are in the zone of the highst Estimated Loss and lowest Estimated Return. We can see how Income Range affected the Estimated EffectiveYield and Estimated Loss.most of the sample have yield in the range between 0 and 0.2. but the still the not employed people tend to have high loss estimated up to 0.3, while most of the sample have loss estimated to be below 0.2 .

We can see how Income Range affected the Estimated EffectiveYield and Estimated return.most of the sample have return estimated in the range between .05 and 0.3. Except some client have income range between 50K and 74K, have estimated return below than zero.

Reproducing the same plots but with the LP_Customer Payments and LP_Customer Principal Payments by Income Range parameters by the income range shows that the majority of the clients have a payment , whatever the priciple of the normal, up to 10K.

Reproducing the same plots but with the LP_Customer Payments and LP_Service Fees Payments by Income Range parameters by the income range shows that when the fees increase the payment delayed, espicially with the client have high income range.

Reproducing the same plots but with the LP_P_Interestand Fees and LP_Service Fees Payments by Income Range parameters by the income range shows that Interestand Fees are recorded by negative numbers, that means it is pending and not payed yet. Even with low service fees.

Reproducing the same plots but with the LP_Gross Principal Loss and LP_Loan Original Amount by Income Range parameters by the income range shows that Gross Principal Loss is high at the low Loan Original Amount. and the Gross Principal Loss is higher in the client have income range more than 100k.

Reproducing the same plots but with the LP_Gross Principal Loss and LP_Net Principal Loss by Income Range parameters by the income range shows that the relation between the two parameters are directly proportional except for som clients with income range more than 100k are lower than the datum line of the relation between Gross Principal Loss and Net Principal Loss.

Reproducing the same plots but with the Loan Original Amount and LP_Net Principal Loss by Income Range parameters by the income range shows that the net Principal Loss is high at the low Loan Original Amount. and the Gross Principal Loss is higher in the client have income range more than 100k. ”

[ ]: