Result: I applied XGB Classifier on the data given to me by classifying my output in 3 classes making it a multiclass classification and the test accuracy that model gives me was 99.5% and train accuracy is 99.7%.

Description of data: The original dataset contains 111 features and in which many values are Nan (Not defined) so a lot of data cleaning needs to be done. Hence first columns such as member_id, id, purpose, etc which have no relation with predictions and features whose mean are 0 i.e all values 0 and then features with all values Nan are deleted. By doing this my number of features are reduced to 33 having an effect on our prediction. Then dividing of the data into 3 parts is done i.e training set on which I will train my algorithm, development set on which the parameters of the classifier is set, the test set on which I will test the accuracy of my algorithm.

Choice of Algorithm: I have chosen algorithm XGB Classifier because it is solely based on the principle of run->test->optimize.

Q1.Which parameters according to me is important?
Ans-> The parameters such as the grade of the borrower, his employment status, the place where he is going to invest, the amount of money funded, the previous remaining money to be paid if taken any, borrower home ownership and the interest charged by the company for the borrower, etc.

EFFECT OF EACH FEATURE TAKEN ON Y

1. Loan amount <15000 are the most defaulter cases

2.No special trend in the funded amount and funded amount investment.

3. Most defaulters are having int_rate >1 and less than 2.

4.Most defaulters are having installments less than 600.

5. Defaulters can have any grade but are having last 3 sub-grades mostly.

6.No special trend in home ownership and emp_length.

7. The borrowers with less annual income<200000 are the most defaulters.

8.No special trend in verification and purpose found by seeing the graph.

9.Persons with dti>=10 & dti<=20 have defaulters very often.

10.delinq_2yrs <=2 and inq_last_6_months<=4 are where defaulters are there.

11.open_acc <=17 and revol_bal <=40000 are the defaulter zone

12.total_acc <=32 or 35 and out_pmcp and out_pmcp_inv=0 are the defaulter cases

13.total_pymt and total_pymt_invst <=10000-12000 and total_rec_pmcp<=7000-10000 are the defaulter zone.

14.recoveries and collection_recovery_fee >0 are defaulters only

15.Persons having lat_pymt_amt are very fewer defaulters.

These are the trends i can make out from the given data but analyzing this data more properly by someone more experienced in this field can give more insights.

Accordingly, I cannot find out a strong trend in between any feature and the type of borrower but my XGB Classifier is surely combining these trends only to get good predictions.

Some points can be missed because of not being possible to plot all data on a graph for 3000 points i have seen the trend with making 3 graphs of each feature i.e 3*33 graphs to analyze. For 32000 data points, it would have become a huge number of graphs and confusing to analyze the trends.