**Summary**:

I created these variables to grasp the gist of the data and for exploratory analysis. In addition to that features are crucial to a model’s performance. Following are some of the variables I created and will point out the one which were crucial model building. I will start by stating approach and the reason for it and finally which variables were statistically significant in driving the model performance.

**Ratio** : The ratio of amount due to av. bank balance in account. This give us an idea about the potential to pay the loan.

**Percent\_loan\_paid**: Fraction of loan paid by the customer.

**delinquency\_rate**:factor by which the delinquency has increased w.r.t to as\_of\_date

**movement**: Tell about loans whose delinquency has increased or decreased/unchanged

**days\_since\_last\_payment**: No of days since the last payment

**target**: variable made for the prediction task. 1- increase 0- Unchanged/Decreased

**Approach**: I chose logistic regression, since it generates probabilistic model which can be used to find the odds of a particular variable containing two states. In this we are trying to predict which the loan is likely to worsen and which loan delinquency is like to improve. So I created a target variable and restructure the problem as a binary classification problem. The target variable is 1 or 0 based on whether the delinquency increased as of december 1st.

So I started building model one variable at a time and if a variable was degrading the performance I removed them. After the model building I found these variables to be statistically significant and crucial in lowering the negative log likelihood.

* Percent\_loan\_paid
* days\_since\_last\_payment
* type of the loan

The logistic regression generates the likelihood of a target variable and hence I choose 0.33 as the threshold and if the value was less than 0.33 I put it under the 0 category( ratings unchanged) and otherwise.

I used 90:10 split for training and testing the model. I found the accuracy to be around 77.12 percent.

**Exploratory Analysis:** we learn **Referral channel** is the most effective of all other than **promontory** since it has only **one entry**. Referral has the **lowest principal amount due**, l**owest delinquency rate**, has better **bank\_balance/ amount\_due ratio** , **highest loan percentage paid.** And In terms of **collection** method **ACH** performs better than its counterpart in terms of **percent\_loan paid**, **lower outstanding principal**.In addition to that we see that **Direct channel and FAP** have **higher deliquency rate**, whereas, **Referreal and Promontory** have **lower deliquency rate**. Further, I observed that **higher fico score** has less **delinquency\_rate.**

**Recommendation: Certain Channels such as Referral, and collection method such as ACH and split have better delinquency rate. In addition to that higher percent\_loan\_paid and lower days\_since\_last payment will further check the delinquency rate from falling. Finally, more data and experimentation will allow us to build better models.**