In this dataset, we have 50000 individuals who applied to loans and attempted to pay off. 3305 of them were unable to pay off and thus charged off before the data collection date. On the other hand, 46695 have paid well up to the mentioned check day.

Model Building:

The model used changes the results immensely, most of the time. Here, we are trying to map numerical values, that is days , into binomial values which is whether there is/will be a charge-off or not. Hence, logistic regression will be a good tool for prediction.

A linear/non-linear regression may also be utilized to map days to days in order to predict the charge-off day number. In my opinion, that would be inaccurate because we are less concerned about the date of the charge-off. The Boolean value of the ultimate status is more crucial.

How to interpret the data:

3305 individuals have failed to pay and charged off. Let's introduce a stamp called "Charge off". These individuals are labeled as True for "Charge off".

46695 people have succeeded up to their check-in date. Later on, they may fail or pay-off well, we don't know. We are going to build a logistic regression model to predict them.

A model that is to yield Trues and Falses, should be fed Trues and Falses during the training. We have lots of positive values for charge-off but no negative values, meaning no Falses.

Let's take a step back and interpret the unlabeled 46695 individuals like: Up to their check-in days, say N, they paid their loans. For a day X<=N, they are a False for the charge-off. But what day should we pick among N values to introduce into our model?

When one hand-picks the day, i.e N/2 for all, the model will be biased. Here, I let the computer decide randomly. I picked a random number for charge off up to N. Then I labeled the data as False.

The Input:

We have 3305 rows that are True and 46995 rows that are False. Should we input them all into the model?

Models are imperfect like everything else. Whatever gets in transforms into the outcome. If you add more sugar, the lemonade will be sugary. More lemon would mean sourness.

As False values overwhelm True values, I randomly chose ~3300 False cases for model building. It is more balanced and less biased this way. However, a more prudent analyst may choose to add more Falses to be pessimistic and to be on the safe side.

Randomness-Random Values:

First off, days for False values were selected randomly. Later, ~3300 cases were picked out of a total of 46995 randomly. But we wish to see the pattern within the randomness and make a healthier prediction, right?

Conclusion:

I reran the model from scratch for multiple times, 10000 namely and took the average of the percentage of charge-offs. This way random errors canceled out pretty much and we were able to see the bigger picture.

Approximately, %7 of 46995 individuals will charge off according to my model.