Problem Statement and Approach

This case study required us to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default, so that the company can utilise this knowledge for its portfolio and risk assessment. Our approach towards finding these factors was to first clean the dataset, impute values and then perform Univariate and Bivariate analysis on the different variables.

Univariate Analysis - Summary

1. 60 months loans have 15% greater probability of default as compared to 36 month loans
2. Surprisingly, loans that are verified have highest default changes(16.7%) and loans which are not verified have lowest default changes(12.9%).
3. Max no of loans are funded towards end of year(Q3) which may be to holiday season and people buying new things. There is not significant variance in default rate across months. Default rate ranges from 12% to 15%.
4. There is no significant difference among MORTGAGE, RENT and OWN type ownership. In terms of probability, OTHER type has highest default rate but there are significantly less data points in this category as compared to other hence it is not conclusive.

Bivariate Analysis - Summary

1. Interest Rate - Higher the interest rates, higher the chances of default
2. Installments - Higher the number of installments, higher the chances of default
3. Annual Income - Although not significant, annual income for defaults is lesser, but we cannot conclude anything from this data
4. DTI (Debt to Income Ratio) - There is no significant difference in dti for charged off vs defaults. Therefore it is not conclusive. open_acc - no significant difference.
5. Principal received to date, Total Payment - These two can be a good factor for default prediction. As very brute-force approach we can use intersection as a threshold to predict whether a loan will be defaulted or charged off.
6. Number of Open Accounts - This has a strong negative correlation with Revolving Utilization