Lending Club Loan Data

**Brief**:

In the attached Jupyter notebook, you’ll see I’ve started data exploration from the lens of predicting what factors weigh in heavily towards deciding successful completion of loan payments.

**Dependent variable:**

For starters, I’ve limited the analysis to just historical loans (could be fully paid, completed payments or charged off). I’m going to be using Logistic Regression and KNN cross-fold validation to classify current loans into two buckets – “High probability of paying the full loan amount” and “High risk of charging off”.

**Independent variables:**

* Purpose of the loan
* Grade of the loan
* New features to be added:
  + Debt to income ratio
  + Loan amount
  + Annual income
  + Interest Rate (APR%).
  + Maturity of loans (short vs long term).

If time permits, I would also like to build a regression model to gauge what percentage of the principal amount will be recovered in cases of a charge-off.