Lending Club Case Study Documentation

Part 1: Exploratory Data Analysis

1. Section 1A: Checking for Missing Values

    - It is identified that missing values comprise less than 0.01% of the data and thus are too insignificant of a chunk to merit imputing. Thus, they are dropped.

2. Section 1B: Exploring Categorical Variables – Transformations

    - Among the variables in scope, grade, term, loan status and issue date are identified as non-numeric
    - Grade, term and loan status are good candidates for conversion to dummy variable for our predictive model as they take a limited number of values
    - Given the many possible values for issue date, this variable is coded as months since issue date so it is numerical and captures the same information
    - Visualizations of distributions of categorical variables reveal:
        i. A, B, C grades and 36-month loans are most frequent in the data set
        ii. The most common loan statuses are fully paid or current
        iii. Most loans in the data set have been issued in the past 5 years, with some far back reaching outliers

3. Section 1C: Exploring Numerical Variables – Managing Outliers

    - Visualizations of distributions of numerical variables reveal generally skewed distributions, with revolving balance, dti, annual income and total payment showing the greatest skewness.
        i. This result is logical given that these covariates are likely highly sensitive to outlier customer situations
    - Removing all outliers using the outside 1.5 IQR rule would remove ~16% of observations
    - Given the high-incidence of outliers, the following configurations will be tested in the predictive model:
        i. A model with outliers
        ii. A model excluding all outliers
        iii. A model with outliers winsorized 5% on each side

4. Section 1D: Cross Variable Exploration

    - First this section examines collinearity among covariates and using an exhaustive correlation heat map finds:

        i. Strong correlation between grade and interest rate, which is validated using a boxplot of interest rates by grade visualization. Both co-variates should likely not be in the model to avoid multi-collinearity.

        ii. Strong correlation between funded amount and loan amount. While a scatter plot of these covariates does confirm this correlation, but also shows that for a non-trivial number of observations the loan amount is adjusted upward from the funded amount. In order to capture the potential impact of this upward adjustment on default, both are included in the predictive model.

    - Second, for loans of interest (36-month term with >= 36 months of data), this section examines the relationship between defaults and each of the co-variates:

        i. The distribution of continuous co-variates by defaults and non-defaults show that interest rate, revolving balance, total payments and months since issue to be more promising predictors of default as these factors show the least overlap in the distribution for defaults and non-defaults.

        ii. Additionally, visualization also shows a clear relationship between default and grades; lower grade loans are more likely to default.

Part 2: Business Analysis

1. Section 2A: Percentage of Loans Fully Paid

    - For the loans of interest (36-month term with >= 36 months of data), ~75% are fully paid, with fully paid classified as having a status of:

        i. Fully paid

        ii. Does not meet the credit policy. Status: Fully Paid

    - This calculation was conducted using the Pandas group by function

2. Section 2B: Cohort Analysis – Defaults

    - For loans of interest, Grade G loans issued in 2016 have highest rate of default

    - This calculation was conducted using the Pandas cross tab function

3. Section 2C: Cohort Analysis – Rate of Return

- The bucketing of rate of return by cohort shows that returns generally improve with higher grade (e.g. going from G to A) but are variable over time

- This calculation was conducted using the Pandas cross tab function

Part 3: Modeling

1. Section 3A: Generate train and test set and standardize data

    - 75% of the data is randomly assigned as the training set – on which the predictive model will be calibrated and calculated and 25% is assigned as the test set on which on the model will be tested for prediction accuracy

        i. This train-test split will prevent over-fitting (e.g. good model performance on a subset of data but not generalizing well).

    - Continuous variables are standardized so they are all in similar units. All variables are standardized using training set means and standard deviations to avoid any spillover of test data features on to the training set.

2. Section 3B: Cross-validate and run logistic regression

    - Logistic Regression CV runs 5-fold cross validation to optimize its parameters for prediction accuracy and then runs the cross-validated model on the training data

    - The classifier is run on the model with outliers, model without outliers and model with winsorized outliers

    - In terms of prediction accuracy, our model does quite well, around 91% for the model with outliers, 92% for the model with no outliers and 88% for the model with winsorized outliers

    - All models beat the naïve baseline of assuming no defaults (around 75% accuracy) or randomly guessing defaults $1/4^{th}$ of the time (around 63% accuracy)

3. Section 3C: Analyze effectiveness using ROC curves

    - ROC curves show that each calibration of the model also has and ROC AUC > 0.9 and beats the two baselines of assuming no defaults and randomly guessing defaults $1/4^{th}$ of the time in terms of AUC as well.

    - Thus, the model is validated by prediction accuracy as well as in terms of maximizing the true positive rate while minimizing the false positive rate

    - Of course, with the AUC < 1, this model is not perfect and can be improved