Plan for Data Cleaning/Pre-processing

1. Removed the features which has more than 75% null values

2. Check missing value count and percent

```
# Check missing values count and percent total = df.isnull().sum().sort_values(ascending=False) percent= (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)*100 missing_data= pd.concat([total, percent],axis=1, keys=["Total", "Percent"])
            missing_data.head(54)
Out[48]:
                  mths_since_last_delinq 454312 51.197065
                            next_pymnt_d 252971 28.507661
                            tot_cur_bal 70276 7.919502
                          total_rev_hi_lim 70276 7.919502
                        tot_coll_amt 70276 7.919502
                                 emp title 51462 5.799326
                             emp length 44825 5.051393
                             last_pymnt_d 17659
                      revol_util 502 0.056571
                                     title
                                               152 0.017129
             collections_12_mths_ex_med 145 0.016340
                        last_credit_pull_d 53 0.005973
                        total_acc 29 0.003268
                                              29 0.003268
                               delinq_2yrs
```

3. Calculating the NaN values for all the columns

```
In [56]: 
#Let's see the data shape and NaN values
print(df.shape)
print(type(df))
print(df.head())
               print(df.isnull().sum())
print(df.isnull().sum().value_counts())
               (887379, 16)
<class 'pandas.core.frame.DataFrame'>
loan_amnt funded_amnt_inv
                                                                 int_rate sub_grade home_ownership
                                                     36 months
                      5000 0
                                          4975.0
                                                                     10.65
15.27
                                                                                                    RENT
                                          2500.0
                                                     60 months
                      2500.0
                      2400.0
                                          2400.0
                                                     36 months
                                                                      15.96
                                                                                     C5
                                                                                                    RENT
                                                                                  B5
                                          3000.0 60 months
                      3000.0
                                                                     12.69
                                                                                                    RENT
                   annual_inc loan_status
24000.0 Fully Paid
30000.0 Charged Off
                                                         purpose addr_state
                                                                                  dti delinq_2yrs \
                                                                           AZ 27.65
GA 1.00
IL 8.72
                                                  credit_card
car
                                                                                                  0.0
                                 Fully Paid small_business
Fully Paid other
                      12252.0
                                                                                                  0.0
                      80000.0
                                     Current
                                                           other
                                                                           OR 17.94
                  1008.71
                                2.0
                                           10.0
                                                   3003.653644
                                                                             3003.65
                                0.0
                                           38.0
                                                   3242.170000
                                                                            3242.17
               loan_amnt
funded_amnt_inv
                                       0
```

4. Filling NaN values with Mean

```
In [57]: ▶ #Filling the numeric columns with missing values by mean
             for cols in df.columns:
                 if df[cols].isnull().sum() != 0:
                     df[cols].fillna((df[cols].mean()), inplace=True)
             df.isnull().sum()
   Out[57]: loan amnt
                               0
             funded amnt inv
                               0
             term
                               Θ
             int rate
             sub_grade
             home ownership
             annual inc
             loan status
             purpose
             addr state
             dti
             delinq_2yrs
             ing last 6mths
             total acc
             total pymnt
             total_pymnt_inv
             dtype: int64
```

Feature Engineering

Feature Engineering can simply be defined as the process of creating new features from the existing features in a dataset.

The performance of a predictive model is heavily dependent on the quality of the features in the dataset used to train that model. If you are able to create new features which help in providing more information to the model about the target variable, it's performance will go up. Hence, when we don't have enough quality features in our dataset, we have to lean on feature engineering.

Manual Feature Engineering

Manual feature engineering can be a tedious process (which is why we use automated feature engineering with featuretools!) and often relies on domain expertise. Since we have limited domain knowledge of loans and what makes a person likely to default, we will instead concentrate of getting as much info as possible into the final training dataframe. The idea is that the model will then pick up on which features are important rather than us having to decide that. Basically, our approach is to make as many

features as possible and then give them all to the model to use! Later, we can perform feature reduction using the feature importances from the model.

The process of manual feature engineering will involve plenty of Pandas code, a little patience, and a lot of great practice manipulation data. Even though automated feature engineering tools are starting to be made available, feature engineering will still have to be done using plenty of data wrangling for a little longer while.

We initially had 74 columns in the Lending Club dataset, we narrowed it down first to 34 columns by eliminating the ones with more than 75% null values, then we understood the dataset properly with the help of LCDataDictionary and then removed further columns manually on the basis of our understanding.

Feature Tools

Featuretools is an open source library for performing automated feature engineering. It is a great tool designed to fast-forward the feature generation process, thereby giving more time to focus on other aspects of machine learning model building. In other words, it makes your data "machine learning ready".

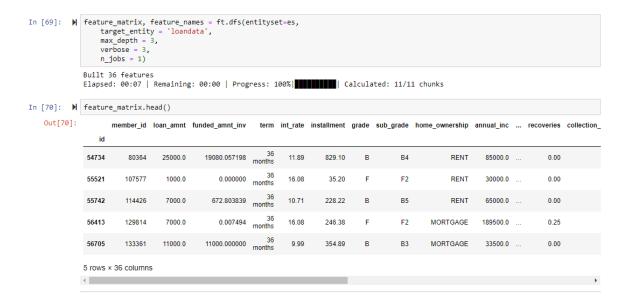
Before taking Featuretools for a spin, there are three major components of the package that we should be aware of:

Entities

Creating entity set for feature tools

• Deep Feature Synthesis (DFS)

Creating feature matrix



• Feature primitives