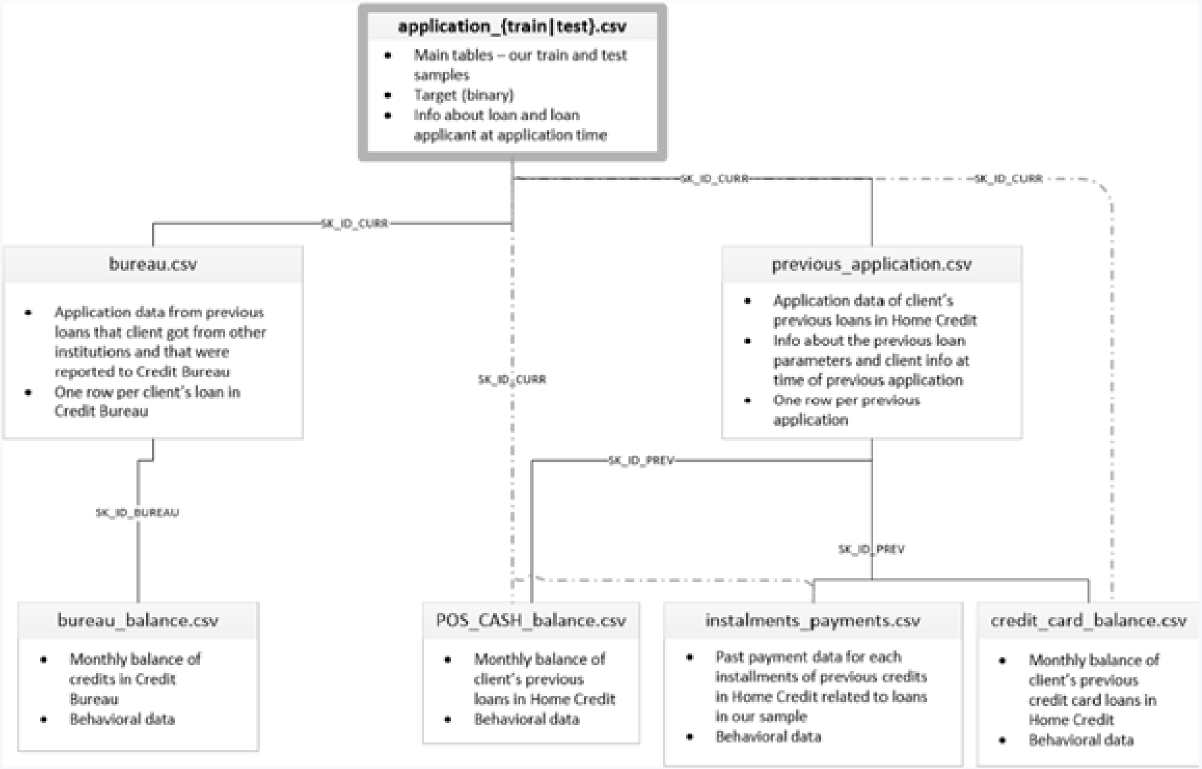
Data Wrangling

Data Description: There are **seven** main files that are available as part of this problem and I have tried using information from all these files as part of the solution.



application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

I will explain the data wrangling / feature engineering steps that were performed on each of these files before building a combined single file that could be directly used for model building.

- **Feature Engineering**

  - **Previous_application.csv (**Application data of client's previous loans in Home Credit**)**

- The **WEEKDAY_APPR_PROCESS_START** column is a string column that needed to be changed to a numeric column. The string days were mapped sequentially to numbers with Sunday being 0 and Saturday being 6.

- The following new features were built from this data set

    - # of **approved** previous loans, a current customer has from home credit.
    - # of **cancelled** previous loans, a current customer has from home credit.
    - # of **refused** previous loans, a current customer has from home credit.
    - # of **unused offers** of previous loans, a current customer took from home credit.
    - # of **each type of previous loans (cash/revolving),** a current customer took from home credit
    - # **of previous loans by each purpose** (Education/Buying a home/Furniture) a current customer has from home credit.
    - # of **previous goods loans by each category** (mobile/jewelry/computers) a current applicant took from home credit.
    - # of **previous loans as client type (repeat customer, new customer etc.)**, a current applicant took from home credit.
    - # of **previous loans by each channel type**, a current applicant took from home credit.
    - # of **previous loans by the seller industry**, the current applicant took from home credit.
    - # of **previous loans by loan portfolio type** (car, pos, cash etc.), a current applicant took from home credit.
    - # **of previous loans** by the person who accompanied the applicant (Spouse, Children etc.) at the time of applying for the loan.
    - # of **previous loan by product type** (cross sell or walk in), the current applicant took from home credit.
    - T**otal # of previous loans**, a current applicant took from home credit.

- **POS_CASH_balance.csv:** (Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit)
    - The following new features were created from this file
        - **# of previous POS and cash loans** a current applicant took from Home credit.
        - **Mean of credit term for each current loan applicant** (SK_ID_CURR) across all POS and cash loans.
        - **Count of days past due (DPD) i.e. times a customer has not paid on time**, for a current customer across all previous POS and cash loans he took from home credit.
        - **Maximum of DPD for a current customer** across all previous POS and cash loans he took from home credit.
        - **Mean of DPD for a current customer** across all previous POS and cash loans he took from home credit.

- **credit_card_balance.csv**: (Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.)
    - The following new features were created from this file
        - **The mean amount balance** of a current applicant across all previous credit card loans he took from home credit.
        - **The # of each type(active/signed/refused) of previous credit card accounts/loans** a current applicant had with home credit.
        - **The mean credit limit** of a current applicant across all previous credit card loans he took from home credit.
        - **The mean drawn amount** of a current applicant across all previous credit loans he took from home credit.
        - **The mean # of times a current applicant withdrew using a credit card** across all previous credit loans he took from home credit.
        - **The mean amount a current applicant withdrew using his credit card** across all previous credit loans he took from home credit.
        - **The mean of exceedslimitby flag** (This flag = 1 for any payment in which the balance amount is greater than the credit

        - limit) across all previous credit loans applicant took from home credit.
        - **The mean of fullpaymentflag** (This flag = 1 for any payment in which the payment made is greater than or equal to the balance) across all previous credit loans the applicant took from home credit.
        - **The mean of minpaymentflag** (This flag = 1 for any payment in which the payment made is greater than or equal to the minimum payment required) across all previous credit loans the applicant took from home credit.

- **Installments_payments.csv (**Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.)

    - The following new features were created from this file

        - **The sum of installments made** by the current applicant across all previous loans he took from home credit.
        - **The max # of installments made** by the current applicant across all previous loans he took from home credit.
        - **The mean days past due (DPD)** for the current applicant across all loans he took from home credit.
        - **The maximum days past due (DPD)** for the current applicant across all loans he took from home credit.

- **bureau_balance.csv**:

    - A new feature **STATUS** has been derived which shows **the maximum number of continuous months a customer didn’t pay his due amount** for each previous non-home credit loan he took for which the data is present in the bureau_balance.csv data set.

After the above steps were executed the below steps were done in sequence

- **Step 1** - **bureau.csv** and the **bureau_balance.csv (with the new feature)** was joined on SK_ID_BUREAU.

- **Step 2 -** The below new features were created on the above joined dataset, most of whose columns are from bureau.csv (Application data from previous loans that client got from other institutions), except the column STATUS which comes from bureau_balance.csv.

    - **Maximum credit overdue across any non-home credit loan** taken by the current applicant.
    - **Mean of amount credit across all non- home credit loans** taken by the current applicant.
    - **Maximum current debt across all non-home credit loans** taken by the current applicant.
    - **# of ACTIVE non-home credit previous loans** that a customer took
    - **# of CLOSED non-home credit previous loans** that a customer took
    - **# of BAD non-home credit previous loans** that a customer took
    - **The maximum number of continuous months a customer didn’t pay his due amount** across any previous non-home credit loan the current applicant took.

- **Step 3** - **The above merged dataset** was **merged** with **the application_train.csv dataset on SK_ID_CURR**.

- **Step 4 – The feature engineered previous_application.csv dataset** was combined with **updated POS_CASH.csv, installments_payments.csv and credit_card_balance.csv on SK_ID_CURR**.

- **Step 5 – The above merged dataset** was then **merged** with the **data set created at step 2 on SK_ID_CURR**.

The merged dataset obtained after the above steps was named dfinal.csv.

The below steps were performed on dfinal.csv before passing it to the model.

- The logically **numerical and categorical** columns were **segregated as some of the numerical columns were originally assigned an object data type and vice-versa.**
- Next the data types were checked. The numerical columns were assigned the data type **float64 or int64** if the original data types differed and similarly the categorical variables were assigned **category** data types.

Handling Missing Values

- Missing values were handled broadly in 2 steps:

  - There were many numerical and categorical columns that had a huge amount of missing values. I decided on a threshold of 45%, assuming columns having more than 45% of missing values will not be able to provide useful information and thus will not prove useful to be part of the predictive model.

    Therefore, all columns (numerical/categorical) having more than 45% missing values were discarded.

  - After the above step, there were still a few columns with some missing values. All these missing values were imputed with their respective mean values.

Converting the categorical variables to dummies

- Dummy variables were created for each of the variables categorized above as a categorical variable and after converting them to dummies their data type was changed to int64 as the predictive model will expect the values as numbers.

Merging the Numerical and Categorical Variables

- In the end the numerical and categorical columns were merged to create the final data set

Selecting the best features using Random Forest

- The merging of numerical and categorical columns in the above step resulted in a dataset having 364 features, but I wanted to choose the optimum number of features to reduce the model training/running time without compromising the model accuracy. Hence, I used the random forest model to select the best features to be passed to the predictive model.