project_draft

September 28, 2021

1 Project Draft

1.1 Introduction

1.1.1 Background

Home Credit is an international consumer finance provider that lends money primarily to those with little or no credit history. They created a kaggle competition where users would use machine learning and statistical methods to predict the loan default risk of individuals based on Home Credit's loan applicant data. Although Home Credit had already used machine learning to project default risk, they hoped to find ways to improve their predictive ability based on what the Kagglers produced.

1.1.2 Significance

Home Credit serves those without credit history or are unbanked, who are likely to be viewed as high risk for a loan default, even if they are financially responsible and will make the necessary repayments. Home Credit needs to avoid making loans to those who will end up unable to complete payments, as this can result in a larger financial cost than the gain of several successful loans. They also want to make their services as accessible as possible through providing the loans to all those who are truly eligible. Machine learning can help us manage these risks and rewards. By building an accurate model, we will be providing Home Credit the ability to provide more loans, since loan applicants identified by our model as at lower risk for a loan default will be less likely to default on the loan.

1.1.3 Research Question(s)

The main research question is whether we can create a model that can estimate the probability of a loan applicant defaulting. Another goal would be to find the features that best predict the default risk.

1.2 Data

1.2.1 Source

The data was retrieved from the Home Credit Defaul Risk Prediction competition on kaggle.com (https://www.kaggle.com/c/home-credit-default-risk/data). The data and supporting information was contained in 10 files, and is 2.68 GB in size.

1.2.2 Original Data Overview

[1]:
```
import numpy as np
import pandas as pd
```

The training dataset is contained in application_train.csv, and contains a total of 307511 rows.

[2]:
```
data = pd.read_csv('./DATA/application_train.csv')
len(data)
```

[2]:
```
307511
```

Below is view of the first 5 rows:

[4]:
```
print(data.head().to_markdown())
```

```
|    |   SK_ID_CURR |   TARGET | NAME_CONTRACT_TYPE   | CODE_GENDER   |
FLAG_OWN_CAR   | FLAG_OWN_REALTY   |   CNT_CHILDREN |   AMT_INCOME_TOTAL |
AMT_CREDIT |   AMT_ANNUITY |   AMT_GOODS_PRICE | NAME_TYPE_SUITE   |
NAME_INCOME_TYPE   | NAME_EDUCATION_TYPE             | NAME_FAMILY_STATUS   |
NAME_HOUSING_TYPE   |   REGION_POPULATION_RELATIVE |   DAYS_BIRTH |
DAYS_EMPLOYED |   DAYS_REGISTRATION |   DAYS_ID_PUBLISH |   OWN_CAR_AGE |
FLAG_MOBIL |   FLAG_EMP_PHONE |   FLAG_WORK_PHONE |   FLAG_CONT_MOBILE |
FLAG_PHONE |   FLAG_EMAIL | OCCUPATION_TYPE   |   CNT_FAM_MEMBERS |
REGION_RATING_CLIENT |   REGION_RATING_CLIENT_W_CITY |
WEEKDAY_APPR_PROCESS_START   |   HOUR_APPR_PROCESS_START |
REG_REGION_NOT_LIVE_REGION |   REG_REGION_NOT_WORK_REGION |
LIVE_REGION_NOT_WORK_REGION |   REG_CITY_NOT_LIVE_CITY |
REG_CITY_NOT_WORK_CITY |   LIVE_CITY_NOT_WORK_CITY | ORGANIZATION_TYPE       |
EXT_SOURCE_1 |   EXT_SOURCE_2 |   EXT_SOURCE_3 |   APARTMENTS_AVG |
BASEMENTAREA_AVG |   YEARS_BEGINEXPLUATATION_AVG |   YEARS_BUILD_AVG |
COMMONAREA_AVG |   ELEVATORS_AVG |   ENTRANCES_AVG |   FLOORSMAX_AVG |
FLOORSMIN_AVG |   LANDAREA_AVG |   LIVINGAPARTMENTS_AVG |   LIVINGAREA_AVG |
NONLIVINGAPARTMENTS_AVG |   NONLIVINGAREA_AVG |   APARTMENTS_MODE |
BASEMENTAREA_MODE |   YEARS_BEGINEXPLUATATION_MODE |   YEARS_BUILD_MODE |
COMMONAREA_MODE |   ELEVATORS_MODE |   ENTRANCES_MODE |   FLOORSMAX_MODE |
FLOORSMIN_MODE |   LANDAREA_MODE |   LIVINGAPARTMENTS_MODE |   LIVINGAREA_MODE |
NONLIVINGAPARTMENTS_MODE |   NONLIVINGAREA_MODE |   APARTMENTS_MEDI |
BASEMENTAREA_MEDI |   YEARS_BEGINEXPLUATATION_MEDI |   YEARS_BUILD_MEDI |
COMMONAREA_MEDI |   ELEVATORS_MEDI |   ENTRANCES_MEDI |   FLOORSMAX_MEDI |
FLOORSMIN_MEDI |   LANDAREA_MEDI |   LIVINGAPARTMENTS_MEDI |   LIVINGAREA_MEDI |
NONLIVINGAPARTMENTS_MEDI |   NONLIVINGAREA_MEDI | FONDKAPREMONT_MODE   |
HOUSETYPE_MODE   |   TOTALAREA_MODE | WALLSMATERIAL_MODE   | EMERGENCYSTATE_MODE
|   OBS_30_CNT_SOCIAL_CIRCLE |   DEF_30_CNT_SOCIAL_CIRCLE |
OBS_60_CNT_SOCIAL_CIRCLE |   DEF_60_CNT_SOCIAL_CIRCLE |   DAYS_LAST_PHONE_CHANGE
|   FLAG_DOCUMENT_2 |   FLAG_DOCUMENT_3 |   FLAG_DOCUMENT_4 |   FLAG_DOCUMENT_5
|   FLAG_DOCUMENT_6 |   FLAG_DOCUMENT_7 |   FLAG_DOCUMENT_8 |   FLAG_DOCUMENT_9
|   FLAG_DOCUMENT_10 |   FLAG_DOCUMENT_11 |   FLAG_DOCUMENT_12 |
FLAG_DOCUMENT_13 |   FLAG_DOCUMENT_14 |   FLAG_DOCUMENT_15 |   FLAG_DOCUMENT_16
```

```
|   FLAG_DOCUMENT_17 |   FLAG_DOCUMENT_18 |   FLAG_DOCUMENT_19 |
FLAG_DOCUMENT_20 |   FLAG_DOCUMENT_21 |   AMT_REQ_CREDIT_BUREAU_HOUR |
AMT_REQ_CREDIT_BUREAU_DAY |   AMT_REQ_CREDIT_BUREAU_WEEK |
AMT_REQ_CREDIT_BUREAU_MON |   AMT_REQ_CREDIT_BUREAU_QRT |
AMT_REQ_CREDIT_BUREAU_YEAR |
|---:|-------------:|---------:|:--------------------|:--------------|:--------
-------|:-----------------|--------------:|------------------:|-------------
--:|--------------:|-----------------:|:-----------------|:------------------
|:-----------------------------|:---------------------|:--------------------|--
--------------------------:|-------------:|---------------:|-----------------
--:|------------------:|-------------:|------------:|----------------:|------
------------:|-------------------:|-------------:|-------------:|:--------------
----|-----------------:|-----------------------:|-----------------------------
:|:-----------------------------|-------------------------:|-----------------
----------:|-----------------------------:|------------------------------:|-----
---------------------:|-------------------------:|--------------------------:|:--
---------------------|--------------:|---------------:|---------------:|-------
----------:|-------------------:|------------------------------:|---------------
---:|-----------------:|----------------:|----------------:|----------------:|--
--------------:|---------------:|-----------------------:|-----------------:|---
-----------------------:|--------------------:|------------------:|------------
-------:|--------------------------------:|-------------------:|-----------------
-:|------------------:|-----------------:|-----------------:|-----------------:|-
----------------:|------------------------:|------------------:|----------------
----------:|----------------------:|------------------:|--------------------:|---
----------------------------:|-------------------:|------------------:|--------
--------:|-----------------:|-----------------:|-----------------:|-------------
---:|------------------------:|------------------:|---------------------------:|
---------------------:|:----------------------|:------------------|---------------
--:|:---------------------|:----------------------|---------------------------:|
---------------------------:|----------------------------:|----------------------
-----:|-------------------------:|------------------:|------------------:|------
------------:|------------------:|------------------:|------------------:|------
------------:|------------------:|-------------------:|-------------------:|----
---------------:|-------------------:|-------------------:|-------------------:|
-------------------:|-------------------:|-------------------:|-----------------
--:|-------------------:|-------------------:|-----------------------------:|---
-------------------------:|-----------------------------:|----------------------
------:|----------------------------:|-----------------------------:|
|  0 |       100002 |        1 | Cash loans           | M              | N
| Y                  |              0 |            202500 | 406598          |
24700.5 |   351000           | Unaccompanied     | Working             | Secondary
/ secondary special | Single / not married | House / apartment   |
0.018801 |       -9461 |            -637 |                -3648 |
-2120 |           nan |            1 |                 1 |                  0 |
1 |            1 |            0 | Laborers          |                 1 |
2 |                             2 | WEDNESDAY                   |
10 |                            0 |                            0 |
0 |                       0 |                          0 |
0 | Business Entity Type 3 |       0.083037 |       0.262949 |       0.139376 |
0.0247 |             0.0369 |                        0.9722 |            0.6192
|           0.0143 |             0    |          0.069  |          0.0833 |
0.125  |         0.0369 |                 0.0202 |           0.019  |
0      |             0      |            0.0252 |               0.0383 |
0.9722 |             0.6341 |           0.0144 |            0      |
0.069  |           0.0833 |           0.125  |          0.0377 |
0.022 |            0.0198 |                          0 |                     0 |
0.025  |              0.0369 |                            0.9722 |
0.6243 |            0.0144 |             0    |           0.069  |
0.0833 |           0.125  |          0.0375 |                  0.0205 |
0.0193 |                    0      |                 0    | reg oper account
| block of flats   |           0.0149 | Stone, brick          | No
|                          2 |                            2 |
2 |                          2 |                      -1134 |                 0 |
1 |                 0 |                  0 |                  0 |
0 |                 0 |                  0 |                   0 |
0 |                  0 |                   0 |                   0 |
0 |                  0 |                   0 |                   0 |
0 |                  0 |                   0 |                             0 |
0 |                            0 |                          0 |
0 |                            1 |
|  1 |       100003 |        0 | Cash loans           | F              | N
| N                  |              0 |            270000 |      1.2935e+06 |
35698.5 |        1.1295e+06 | Family             | State servant      | Higher
education             | Married              | House / apartment   |
0.003541 |       -16765 |           -1188 |                 -1186 |
-291 |           nan |            1 |                  1 |                 0 |
1 |            1 |            0 | Core staff        |                2 |
1 |                            1 | MONDAY                          |
11 |                           0 |                            0 |
0 |                       0 |                          0 |
0 | School                 |       0.311267 |       0.622246 |     nan        |
0.0959 |             0.0529 |                        0.9851 |            0.796
|           0.0605 |             0.08 |          0.0345 |          0.2917 |
0.3333 |         0.013  |                 0.0773 |           0.0549 |
0.0039 |              0.0098 |           0.0924 |               0.0538 |
0.9851 |             0.804  |           0.0497 |           0.0806 |
0.0345 |           0.2917 |           0.3333 |          0.0128 |
0.079 |            0.0554 |                          0 |                     0 |
0.0968 |              0.0529 |                            0.9851 |
0.7987 |            0.0608 |             0.08 |           0.0345 |
0.2917 |           0.3333 |          0.0132 |                  0.0787 |
0.0558 |                    0.0039 |                 0.01 | reg oper account
| block of flats   |           0.0714 | Block                 | No
|                          1 |                            0 |
1 |                          0 |                       -828 |                 0 |
1 |                 0 |                 0 |                 0 |
0 |                 0 |                 0 |                  0 |
0 |                  0 |                  0 |                  0 |
0 |                  0 |                  0 |                  0 |
0 |                  0 |                  0 |                           0 |
0 |                           0 |                            0 |
0 |                           0 |
|  2 |       100004 |        0 | Revolving loans      | M              | Y
| Y                  |             0 |              67500 | 135000          |
6750   |   135000          | Unaccompanied     | Working             | Secondary
/ secondary special | Single / not married | House / apartment   |
0.010032 |       -19046 |            -225 |               -4260 |
-2531 |           26 |           1 |               1 |               1 |
1 |           1 |           0 | Laborers          |                1 |
2 |                           2 | MONDAY                          |
9 |                           0 |                              0 |
0 |                        0 |                        0 |
0 | Government            |     nan        |       0.555912 |       0.729567 |
nan       |          nan      |                     nan      |         nan
|        nan      |         nan    |        nan      |        nan      |
nan       |      nan      |              nan      |        nan      |
nan       |          nan      |          nan      |            nan      |
nan       |          nan      |         nan      |        nan      |
nan       |        nan      |         nan      |        nan      |
nan      |        nan      |                     nan |              nan
|         nan      |           nan      |                    nan      |
nan       |         nan      |          nan    |        nan      |        nan
|        nan      |       nan      |               nan      |         nan
|                  nan      |              nan    | nan                | nan
|        nan      | nan                | nan                    |
0 |                        0 |                        0 |
0 |                   -815 |                0 |                0 |
0 |                 0 |                 0 |                 0 |
0 |                 0 |                  0 |                  0 |
0 |                  0 |                  0 |                  0 |
0 |                  0 |                  0 |                  0 |
0 |                  0 |                           0 |
0 |                           0 |                            0 |
0 |                           0 |
|  3 |       100006 |        0 | Cash loans           | F              | N
| Y                  |             0 |             135000 | 312682          |
29686.5 |   297000           | Unaccompanied     | Working             | Secondary
/ secondary special | Civil marriage       | House / apartment   |
0.008019 |       -19005 |           -3039 |               -9833 |
-2437 |          nan |           1 |               1 |               0 |
1 |           0 |           0 | Laborers          |                2 |
2 |                           2 | WEDNESDAY                       |
17 |                          0 |                              0 |
0 |                         0 |                         0 |
0 | Business Entity Type 3 |     nan        |       0.650442 |     nan         |
nan      |          nan      |                       nan      |          nan
|        nan      |          nan    |        nan      |        nan      |
nan      |       nan      |               nan      |         nan      |
nan      |           nan      |         nan      |            nan      |
nan      |          nan      |          nan      |         nan      |
nan      |        nan      |          nan      |        nan      |
nan     |        nan      |                        nan |                 nan
|         nan      |            nan      |                       nan      |
nan      |         nan      |           nan    |          nan      |         nan
|        nan      |       nan      |                nan      |          nan
|                  nan      |                nan    | nan                   | nan
|        nan      | nan                    | nan                       |
2 |                           0 |                           2 |
0 |                       -617 |                 0 |                  1 |
0 |                 0 |                   0 |                  0 |
0 |                 0 |                    0 |                   0 |
0 |                  0 |                    0 |                   0 |
0 |                  0 |                    0 |                   0 |
0 |                  0 |                          nan |
nan |                          nan |                          nan |
nan |                          nan |
|  4 |       100007 |        0 | Cash loans           | M              | N
| Y                 |             0 |            121500 | 513000          |
21865.5 |   513000           | Unaccompanied     | Working            | Secondary
/ secondary special | Single / not married | House / apartment   |
0.028663 |       -19932 |          -3038 |                -4311 |
-3458 |          nan |           1 |                 1 |                0 |
1 |           0 |            0 | Core staff        |                1 |
2 |                           2 | THURSDAY                           |
11 |                          0 |                              0 |
0 |                         0 |                         1 |
1 | Religion                |     nan        |       0.322738 |     nan         |
nan      |          nan      |                       nan      |          nan
|        nan      |          nan    |        nan      |        nan      |
nan      |       nan      |               nan      |         nan      |
nan      |           nan      |         nan      |            nan      |
nan      |          nan      |          nan      |         nan      |
nan      |        nan      |          nan      |        nan      |
nan     |        nan      |                        nan |                 nan
|         nan      |            nan      |                       nan      |
nan      |         nan      |           nan    |          nan      |         nan
|        nan      |       nan      |                nan      |          nan
|                  nan      |                nan    | nan                   | nan
|        nan      | nan                    | nan                       |
0 |                           0 |                           0 |
0 |                      -1106 |                 0 |                  0 |
0 |                     0 |                     0 |                     0 |
1 |                     0 |                      0 |                      0 |
0 |                      0 |                      0 |                      0 |
0 |                      0 |                      0 |                      0 |
0 |                      0 |                                0 |
0 |                                0 |                                0 |
0 |                                0 |
```

There are a 122 columns in total, so the above dataset view only contains a fraction of these.

```
[61]: len(data.columns)
```

```
[61]: 122
```

56 of these are categorical, while the rest are numerical.

I will also be using bureau.csv and bureau_balance.csv, files containing data on existing and past loan balances at other institutions for the loan applicant. The data I am interested in is in bureau_balance.csv, so while there are several columns in bureau.csv, I am only interesting in using the two id columns it contains. These will allow me to match the data in application_train.csv to bureau_balance.csv. Here is the view of the first 5 rows of bureau_balance:

```
[63]: data = pd.read_csv('./DATA/bureau_balance.csv')
      data.head()
```

```
[63]:    SK_ID_BUREAU  MONTHS_BALANCE STATUS
      0       5715448               0      C
      1       5715448              -1      C
      2       5715448              -2      C
      3       5715448              -3      C
      4       5715448              -4      C
```

This is a large dataset, as it contains a record for every month and every previously existing/closed loans for the applicants.

```
[71]: len(data)
```

```
[71]: 27299925
```

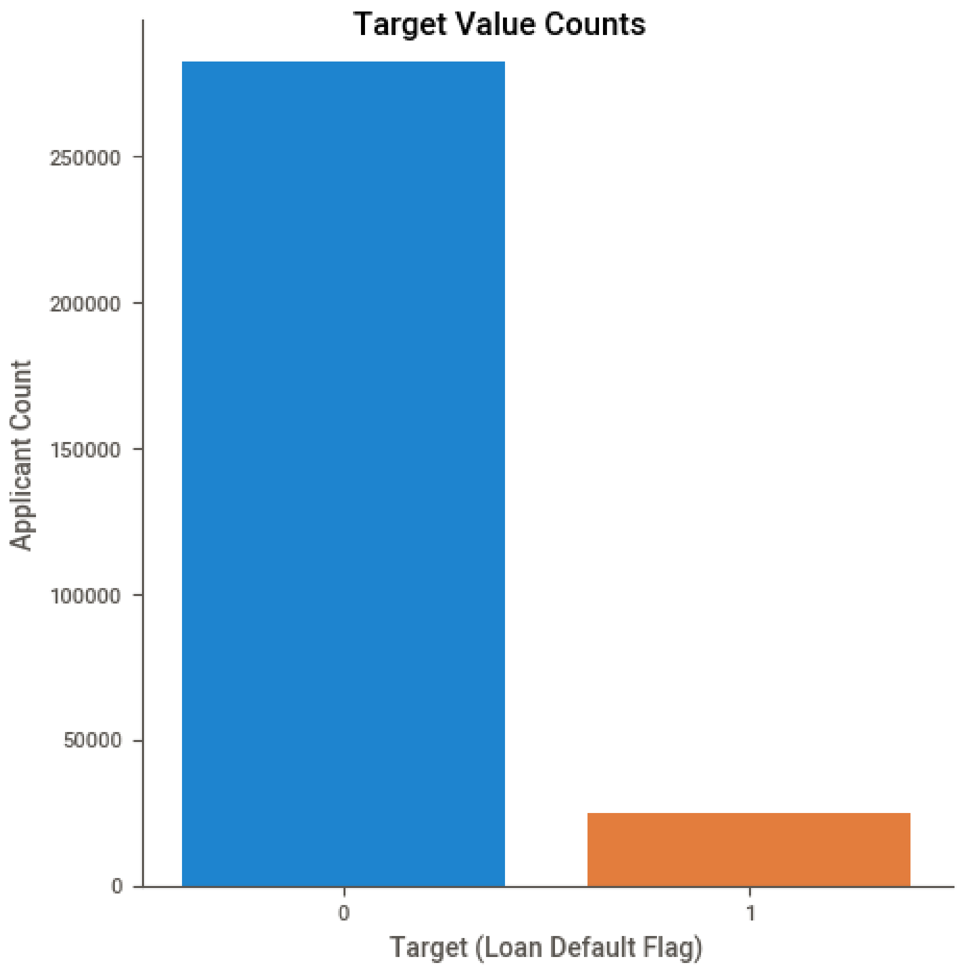
1.2.3 Variables

Target Our target variable is whether the loan applicant defaulted. It is a boolean value, with a value of 0 representing no default, and 1 signaling a loan applicant that ended up defaulting. This is an unbalanced data point, as there are almost 10 times more succesfull loans than defaults

```
[7]: import matplotlib.pyplot as plt
     import seaborn as sns
```

```
[29]: ax = sns.catplot(x="TARGET", kind="count", data=data)
      ax.set_xlabels("Target (Loan Default Flag)")
      ax.set_ylabels("Applicant Count")
      ax.fig.suptitle("Target Value Counts")
```

```
[29]: Text(0.5, 0.98, 'Target Value Counts')
```



Personal Several variables contain information about the individual applying, including demographics like gender or education status, or miscellaneous information like whether the person owns a home or car. These variables are mainly boolean or categorical.

Loan This group of variables describe the loan itself, such as the amount of money being borrowed, the time the approval process began, and what documentation was provided. These variables are of all types, including numerical.

Financial Information These variables contain the financial information of the applicant. They include normalized scores (I believe these scores act like a credit score, but Home Credit does not state this in the column description they provided) and the amount of enquiries to Credit Bureau about the client in the time prior to the application. These variables are mainly numerical.

Employment Home Credit has provided columns containing the applicant's annual income and field of work. These variables are of all types.

Property info A fair amount of variables contain data about the property the client lives in. These are numeric, and have been normalized.

Regional Info This group of variables describe the region where the client lives, including the normalized population and Home Credit's rating of the region.

1.2.4 Missing Data and Imputation

Most columns are missing less than 1% of their data, outside of the property info columns. Many of the property info columns are missing the majority of their values, which may prevent me from using their values. Other variables with a large share of missing values are the applicant's car age, occupation type, two of their scores (normalized financial scores), and the Credit Bureau enquiry counts.

Records with null property values make up 75% of the data. Almost every record has a null record when you also include where the other frequently missing columns. However, ignoring these columns, only 1% of our data would have a null value.

```
[52]: prop_cols = [col for col in data.columns if 'AVG' in str(col) or 'MEDI' in␣
       ↪str(col) or 'MODE' in str(col)]
      add_mis_cols = ['OWN_CAR_AGE','OCCUPATION_TYPE','EXT_SOURCE_1','EXT_SOURCE_3']
      add_mis_cols.extend([col for col in data.columns if 'AMT_REQ_CREDIT' in␣
       ↪str(col)])
      other_cols = [col for col in data.columns if str(col) not in prop_cols and␣
       ↪str(col) not in add_mis_cols]
```

```
[57]: len(data['DAYS_BIRTH'].dropna())
```

```
[57]: 307511
```

```
[58]: len(data[prop_cols].dropna())/len(data)
```

```
[58]: 0.2555843530800524
```

```
[59]: len(data[prop_cols+add_mis_cols].dropna())/len(data)
```

```
[59]: 0.028197365297501553
```

```
[60]: len(data[other_cols].dropna())/len(data)
```

```
[60]: 0.9903092897489846
```

I am planning to drop the property columns from the dataset, and may revisit them later to do some additional feature engineering and data cleansing on them. I will also ignore the applicant's car age, as 2/3 of its values are missing. The remain columns have a significantly lower amount of missing values.

For categorical columns with missing values, I will either choose 1 category as the default category (a category that shouldn't have effect on default probability), or create a new "unknown" category. For numeric columns, I will use KNN imputation with 5 neighbors to fill the values or choose a default value. In order to do the imputation, I selected a subset of columns to reduce computation time and scaled the values. With these changes, I no longer have any missing values in my training dataset.

1.2.5 Data Re-structuring and Wrangling

Home Credit has also provided a file with past loan balances at other institutions for the loan applicant, titled bureau_balance.csv. I will be aggregating this file by loan to get the count of months where the loan was late on payment. I will than match up these loans to the applicants using the data in bureau.csv. Applicants can have multiple loans, so for each applicant, I will take the total number of defaults accross all their loans. Final, I will then add the total default count as a new variable in our training dataset by doing a left join on the applicant id, where the left dataset is our original training dataset.

1.2.6 New Variables

So far, I have created 3 new variables. The first is the social circle default ratio columns. These 2 columns our based on 4 other columns in the original dataset that contain the amount of observations along with the amount of defaults in the client's social circle. By creating a ratio, we should be able to know the true propensity of defaults in the client's social circle as a ratio, instead of only going off the raw default count. These values are generated for a 30 day default and 60 day default.

The other new column was the amount of requests to the Credit Bureau over the year prior to the application. In the original table, the total number of requests are split between the last hour, day, week, month, quarter, and year, excluding the observations added to another column. For example, the year request count excludes the count from the last quarter. I will be adding these columns together for one final annual total, as I don't believe we have enough observations for the smaller periods for them to be useful. Finally, the column from the bureau aggregation will also be added to the training dataset.

1.2.7 Clean Data Overview

After our column drops and additions, we have 78 columns in our cleaned data, 52 of which are categorical and 26 numerical. Below is a views of the first 5 rows of this data.

```
[3]: final_dataset = pd.read_csv('./DATA/final_dataset.csv')
     final_dataset.head()
```

```
[3]:    SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
     0      100002       1         Cash loans           M            N
     1      100003       0         Cash loans           F            N
     2      100004       0    Revolving loans           M            Y
```

```
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

   …  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
0  …                        0.0                        0.0
1  …                        0.0                        0.0
2  …                        0.0                        0.0
3  …                        0.0                        0.0
4  …                        0.0                        0.0

  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON  \
0                       0.0                        0.0
1                       0.0                        0.0
2                       0.0                        0.0
3                       0.0                        0.0
4                       0.0                        0.0

  AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR  \
0                      0.0                        1.0
1                      0.0                        0.0
2                      0.0                        0.0
3                      0.0                        0.0
4                      0.0                        0.0

   DEF_30_RATIO_SOCIAL_CIRCLE  DEF_60_RATIO_SOCIAL_CIRCLE  \
0                         1.0                         1.0
1                         0.0                         0.0
2                         0.0                         0.0
3                         0.0                         0.0
4                         0.0                         0.0

   AMT_REQ_CREDIT_BUREAU  DEFAULT_COUNT
0                    1.0           27.0
1                    0.0            0.0
2                    0.0            0.0
3                    0.0            0.0
4                    0.0            0.0

[5 rows x 78 columns]
```

```
[4]: len(final_dataset.columns)
```

```
[4]: 78
```

```
[6]: # Find columns with missing values
     final_dataset.columns[final_dataset.isna().any()].tolist()
```

```
[6]: []
```

```
[ ]:
```