Clustering Model Report On Home Credit Default Risk

Long Fang, Xia Fu, Feiman Li

I. Executive Summary

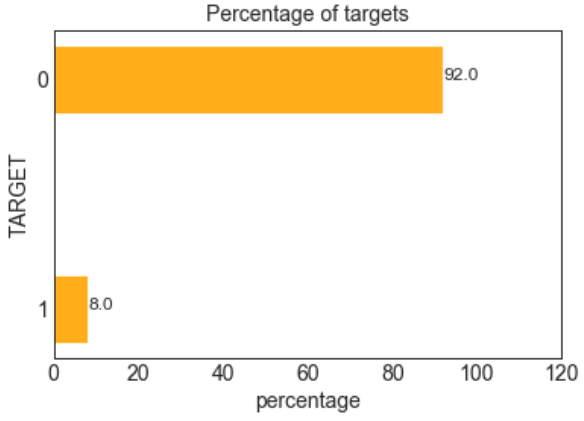
Banks have been relying on scoring models to access credit risk. Today, with machine learning algorithms, the forecasts of future repayments are much more reliable. Previously, we used supervised learning models, based on the bank customer's historical data to build algorithms and train models. Then, we can use these models to evaluate the new customer and predict whether he/she will repay the debt. However, supervised learning only helps us predict the probability of default, and it is difficult to provide more information. Unsupervised learning can make it easy for us to understand the internal structure of data, discover the characteristics of data, and better serve our clients. Thus, in this project, we classified customers through unsupervised learning algorithms and analyzed the weight of variables in identifying whether customers will default. The data are provided by Home Credit, a service dedicated to the unbanked population with lines of credit (loans). Predicting whether customers are having trouble repaying their loans is a key business need.

II. Data Description

In this project, we looked into the training dataset and carefully conducted the data imputation to each feature. This dataset contains all the information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK_ID_CURR. This dataset contains information including the following aspects: personal information (age, gender, family, etc.), housing information, credit score from external sources, as well as borrowers' employment information, etc. There are 122 variables and 307,511 observations in the dataset. Among the 122 variables, 67 of which are float variables, 43 are integer variables and 13 are string variables. As for quality, 67 variables have missing values.

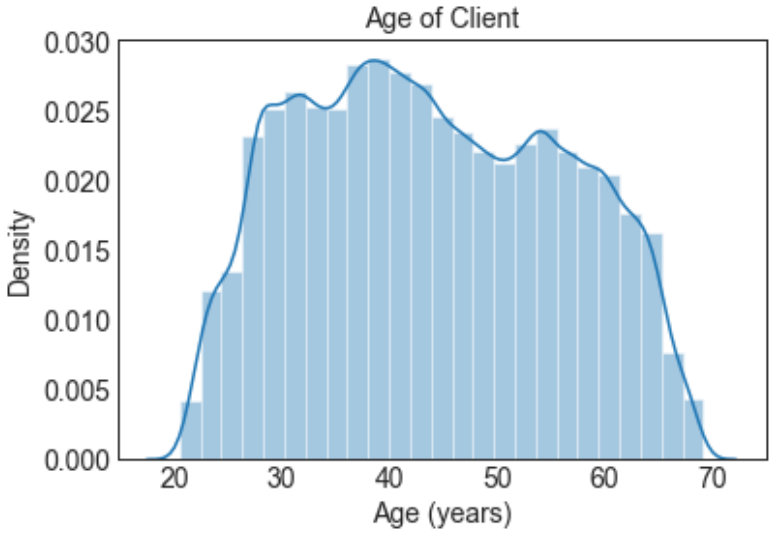
III. Exploratory Data Analysis

Default rate:

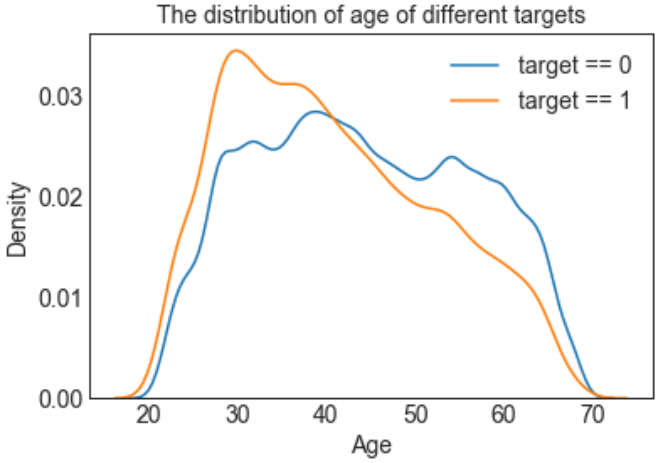


First, we looked at what is the percentage of the default. The variable “TARGET” column in the train dataset represents the result that the loan will be repaid (0) or not (1). 92% of people with "Target=0" repaid for their loan, the other 8% are those who are not able to pay for their loan.

AGE:



Next, we wanted to see the distribution of the lender's age. Because in the dataset, it gives the days from birth rather than age, we divided the days by 365 to estimate their age and plot the figure. The average age for our observations is 43.9. The minimum age is 20, and the maximum is 69.

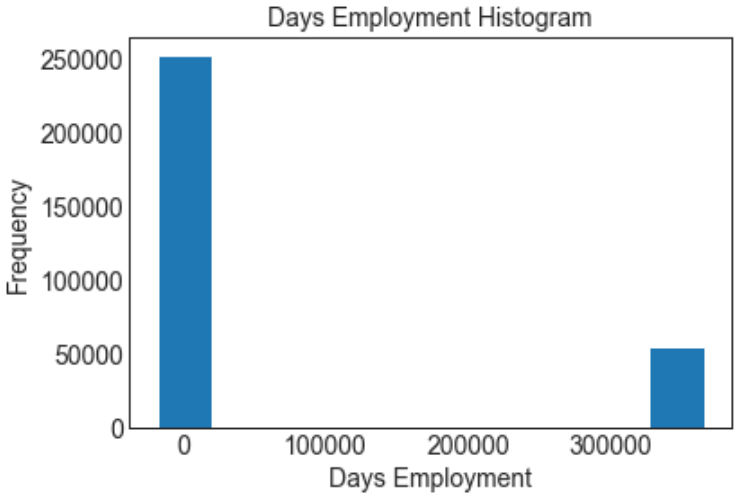


We wanted to know whether the age of borrowers’ contribution to their final result. From the plot, we found that, compared to old people, young people tend not to pay back, which also makes sense. Thus, age is an important factor that we should include in our model.

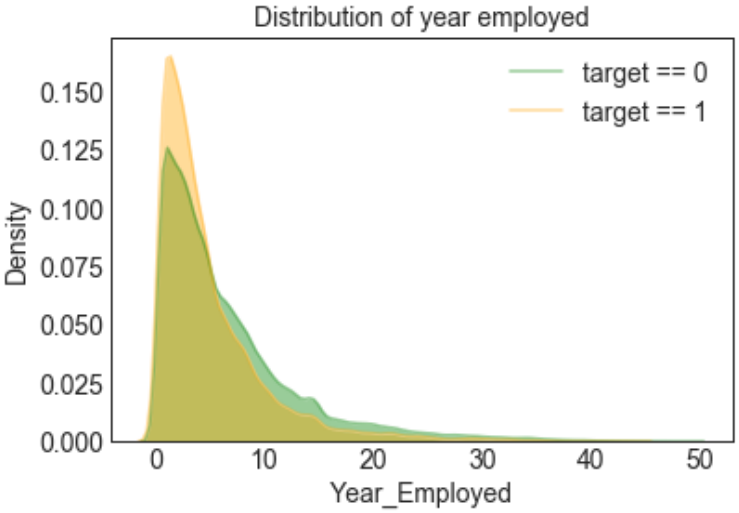
Days Employment:

We then went deep into the days employed, and we found that there are many outliners who have worked 365243 days, which is like errors because that is about 1000

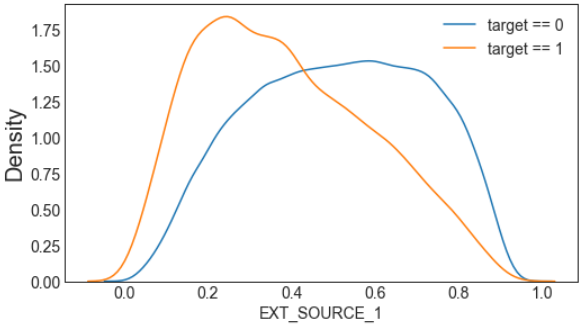
years.

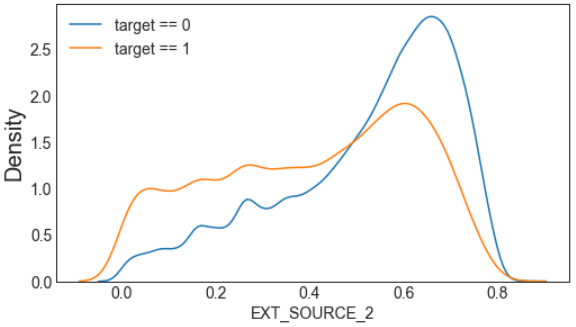


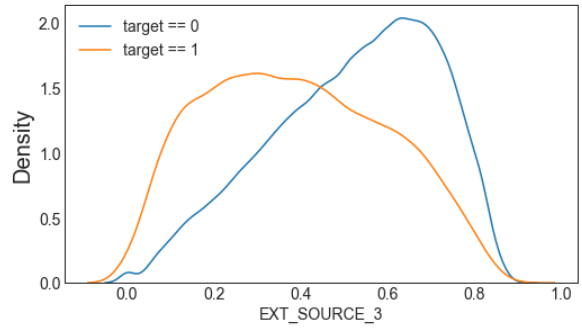
Thus, we plotted only those with employment days less than 365243 for the two target classes and divided the days by 365. Some difference between class 0 and class 1 could be found in the below figure, that the distribution for people who tend to default is a little more right skewed than those who are employed for a longer period of time.



EXT_SOURCES:







EXT_SOURCES are the credit scores derived from other sources. We can see a clear difference on the three scores between people who tend to default and people who don't.

We will see more visualizations after executing unsupervised learning models.

IV. Feature Engineering & Selection

1. Data imputation

Column name	Missing Values	Percentage of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4

As you can see, there are lots of similar variables in the missing value table. Therefore, we could keep one variable for some similar variables. For example, the first three features, COMMONAREA_MEDI, COMMONAREA_AVG, COMMONAREA_MODE, all provide normalized information about spaces where the client lives, such that average (_AVG suffix), modus (_MODE suffix) and median (_MEDI suffix) common area. The others are corresponding average, mode and median for living area, age of building, number of elevators, number of entrances, state of the building, number of floors. We choose to keep one of the three variables because they are duplicated information and at the same time, they contain an amount of missing values, resulting in 29 variables being deleted in total. Also, it is inappropriate to fill the missing values using only mean because you do not know the distribution of the original data. What we try to do is to keep the original distribution. Thus, we fill the missing values with -1, as -1 is different from other values so that machines can recognize it.

Then, we check the missing value table again. Other than above-mentioned missing values such as ATM_REQ_CREDIT_BUREAU_HOUR/DAY/WEEK/MON/QRT/YEAR, providing number of enquiries to Credit Bureau about the client one hour/day/week/month/quarter/year before application (excluding one day before application), EXT_SOURCE_1(Normalized score from external) and OBS_60_CNT_SOCIAL_CIRCLE(How many observation of client's social surroundings with observable 30 DPD (days past due) default).For ATM_REQ_CREDIT_BUREAU,

missing values mean that these person did not enquiry before application, therefore we fill them with zero. For OBS_60_CNT_SOCIAL_CIRCLE, missing values mean that the client's social circle are not the clients in this bank and information cannot be acquired. Fortunately, these variables only have 0.3% missing values, we just delete these observations since we cannot acquire their performance.

2. Dimension Reduction

Before applying our clustering algorithm, we want our models to become simple. Thus, we need to reduce the number of features we have. The basic idea is to use PCA as a tool for feature selection, selecting variables according to the number (from largest to smallest in absolute values) of their coefficients (loadings).

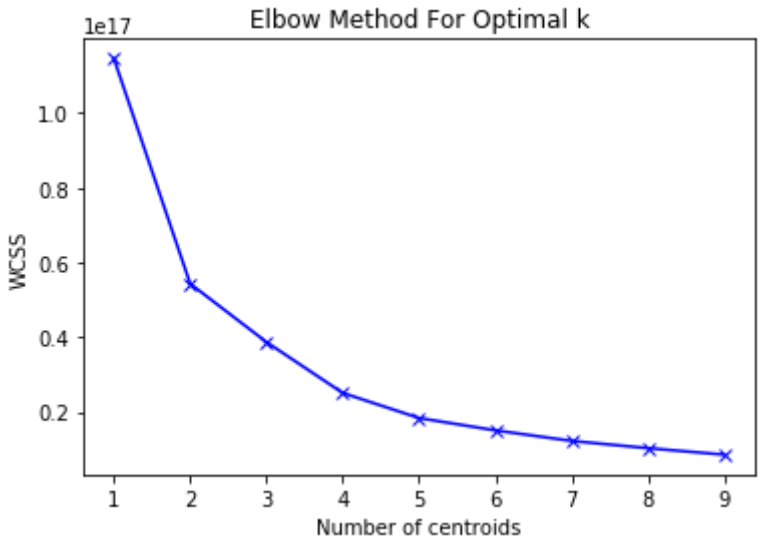
Firstly, We standardized our data. The reason to standardize is that if the value of a certain feature in the data is particularly large, then it has a large proportion in the calculation of the entire error. After the projection to the low-dimensional space, in order to make the decomposition approximate the original data, the entire projection will try to approximate the largest feature and ignore the features with smaller values. Because we did not know the importance of each feature before modeling, it is most likely that a large amount of information is missing.

Then, we choose to conduct Principal Components Analysis. Here, we want to decrease the dimension to 3 for the sake of keeping most of the information.

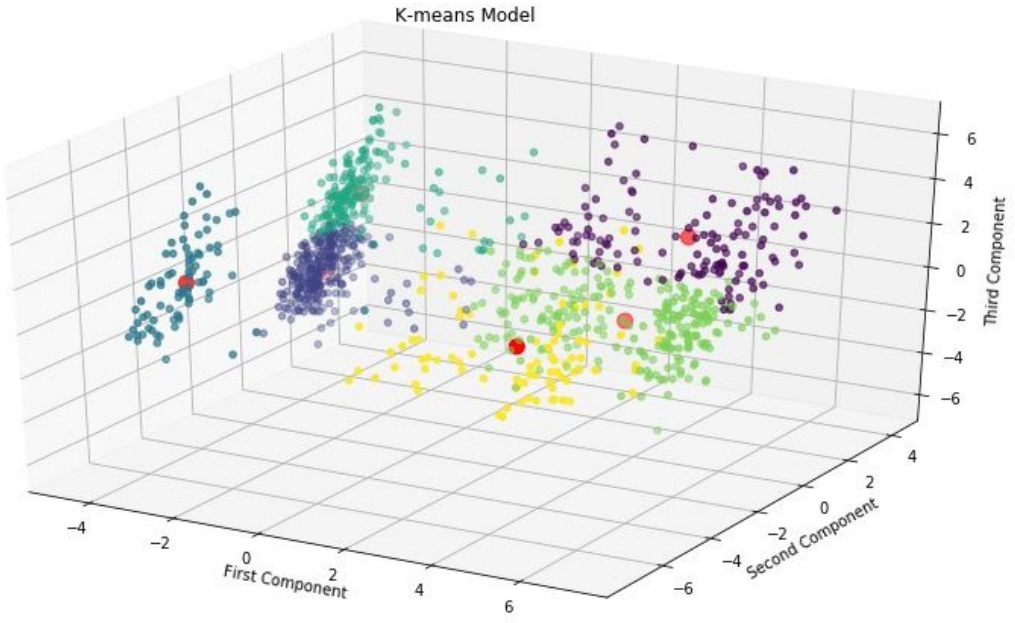
V. Modeling :

K-means:

We first choose K-Means for clustering, which is a powerful unsupervised learning algorithm. To identify the proper number of clusters, we applied the Elbow method. The method consists of plotting the number of clusters using Within-Cluster-Sum-of-Squares(WCSS) as a function.

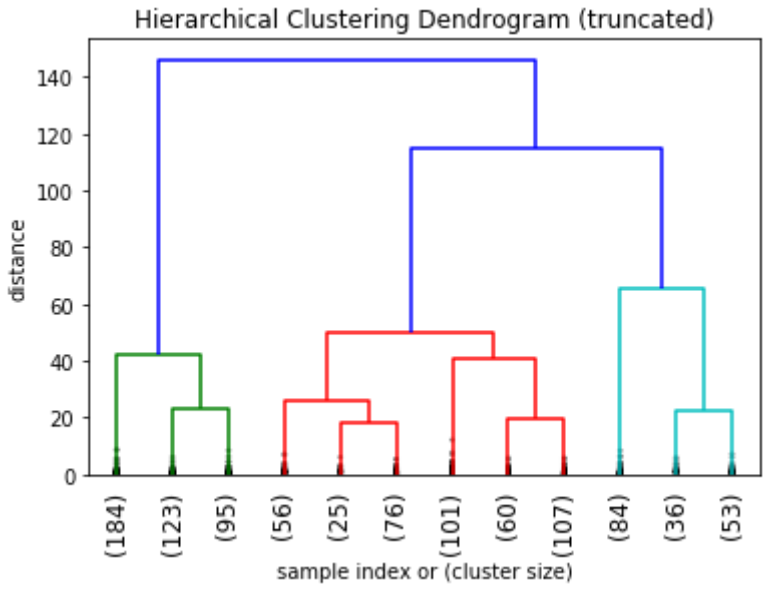


As shown in the figure, imagine this line as a curved arm, then the best value is in the position of the elbow. We choose the K value of 6 in this case in order to get more information about our observation. Then we sub-sampled 1,000 points to plot the 3D figure. The red points are the center of each cluster.

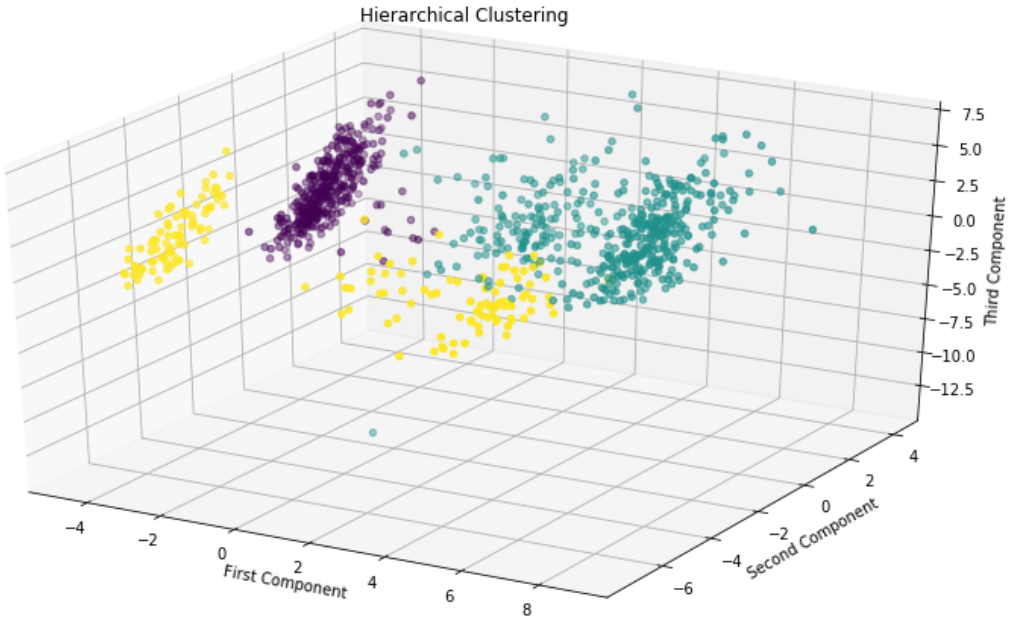


Hierarchical Clustering:

We also tried the Hierarchical Clustering for this project. We imported the linkage package and applied the ward method to calculate the distance between clusters. Then we plotted the dendrogram. (subsample=1000)

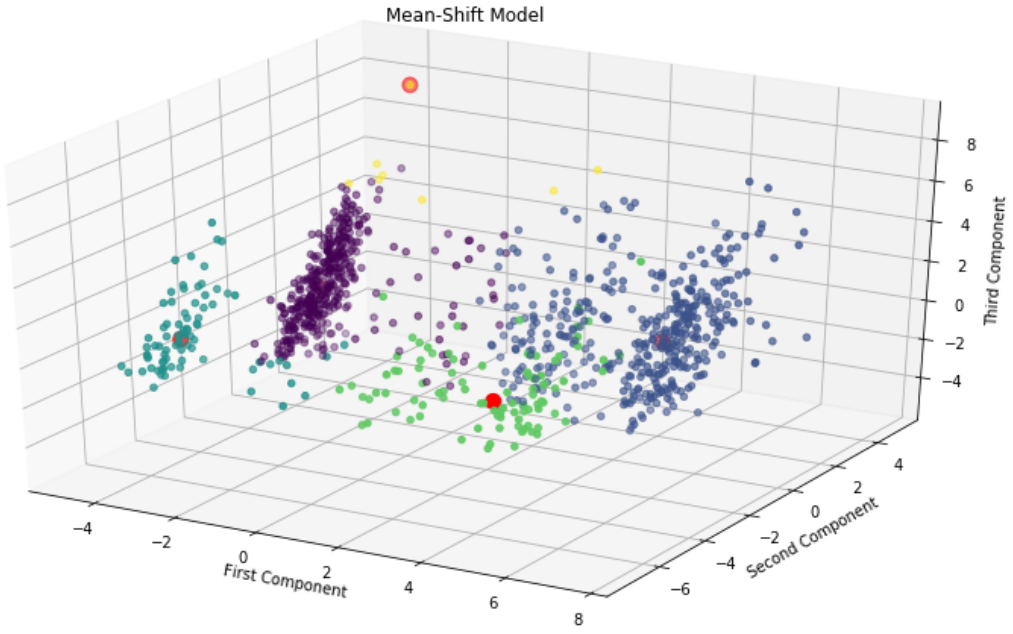


Large distance spans are usually the places we're interested in, so we thought if when distance equals 100 the clustering is good. Similarly, we draw the 3D figure. The clearly different positioning of those yellow points made us doubt this result. And we saw a similar partition to k-means model when distance is 42 and the number of clusters is 6. However, because we are not able to run this model for all points on our computer, which requires our device with strong computing capabilities, we didn't choose it as our final model.



Mean-shift:

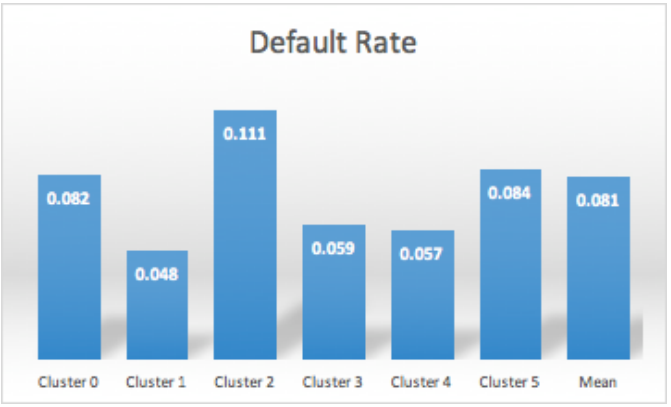
In order to compare with K-means clustering, we conduct Mean-shift clustering. We calibrated our model with grid search for best bandwidth. We set quantile equals to 0.1 and n_sample equals to 1,000 and the best bandwidth is 2.53 also with 5 clusters.



This clustering result was a little different from k-means. Later we calculated the default rate and we thought the clustering result is not better than k-means. Thus, we chose k-means to classify all the data and concatenate with our original cleaned dataset to execute the next step.

VI. Interpretation:

We classified all the records into six clusters and then concatenated with the clean data. And then we calculated the default rate for each cluster.



As the default rate graph shows, cluster 0 and cluster 5 have the default rate that is close to the average 0.08. And the average default rates of cluster 1 & 3 & 4 are lower than the average. However, we find that the people in cluster 2 are more likely to default, whose default rate is 0.11.

Next, we tried to interpret the results. We look at some features that contribute the five most to each of the principle components, such as ages, area size, AMT_credit, work city, etc.

```
PC_1= PCs.nlargest(5, 'PC-1')
PC_1.index.tolist()
```

```
['FLOORSMAX_AVG',
 'ENTRANCES_AVG',
 'APARTMENTS_AVG',
 'LIVINGAREA_AVG',
 'TOTALAREA_MODE']
```

```
PC_2= PCs.nlargest(5, 'PC-2')
PC_2.index.tolist()
```

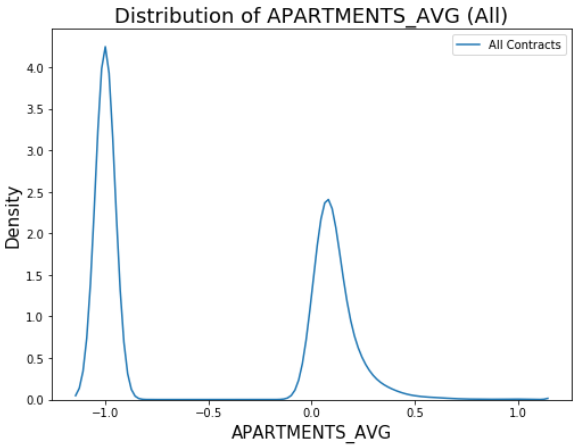
```
['FLAG_EMP_PHONE',
 'DAYS_BIRTH',
 'NAME_INCOME_TYPE_Working',
 'REG_CITY_NOT_WORK_CITY',
 'CNT_FAM_MEMBERS']
```

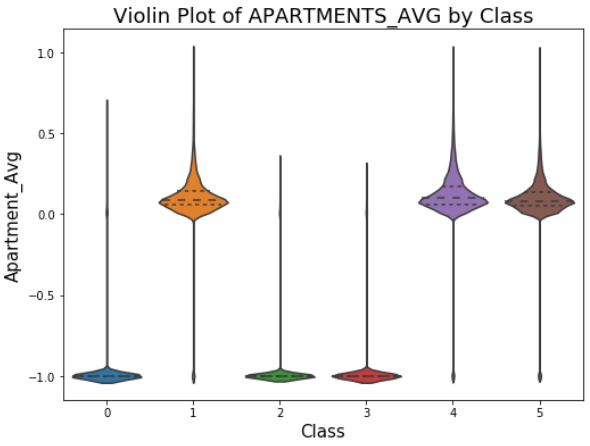
```
PC_3= PCs.nlargest(5, 'PC-3')
PC_3.index.tolist()
```

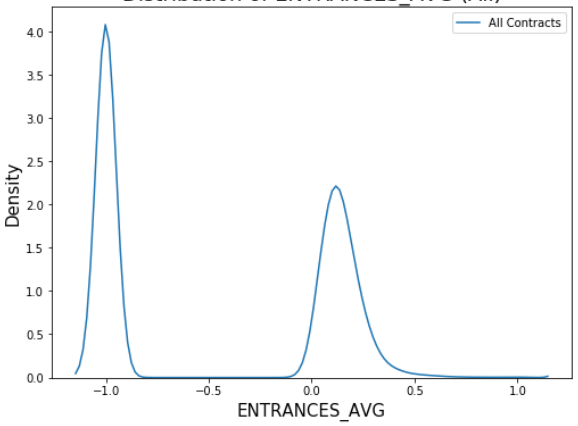
```
['AMT_GOODS_PRICE',
 'AMT_CREDIT',
 'AMT_ANNUITY',
 'FLAG_OWN_CAR_Y',
 'CODE_GENDER_M']
```

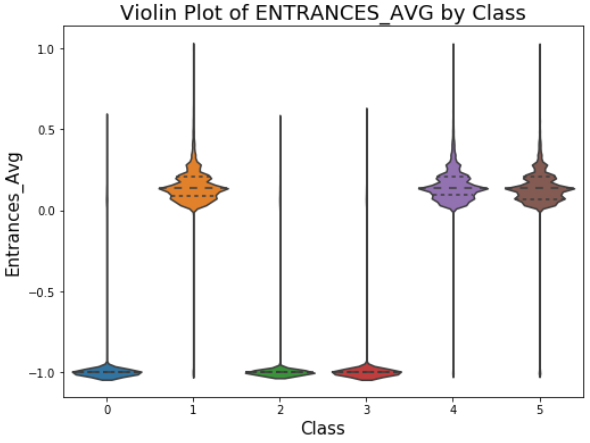
Principle Component 1:

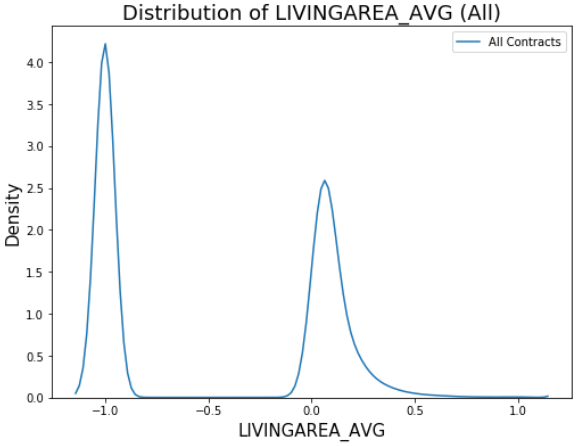
For principle component 1, we can see that the five features that contribute to principal component 1 are all the normalized information about building where the client lives. There are many records showing the value of "-1", which indicate that they were missing in the original dataset and were filled with "-1" in the data processing step.

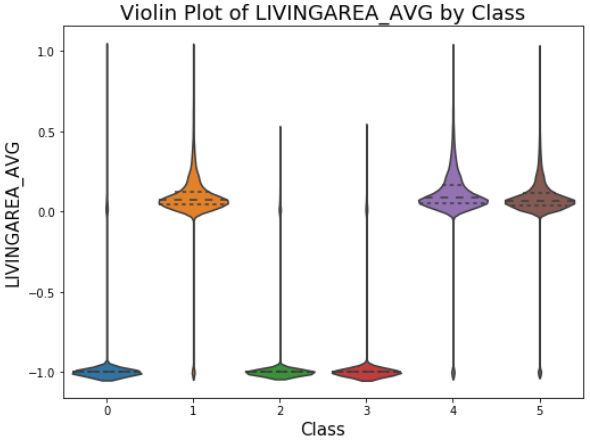

Distribution of APARTMENTS_AVG (All)
All Contracts
Density
APARTMENTS_AVG

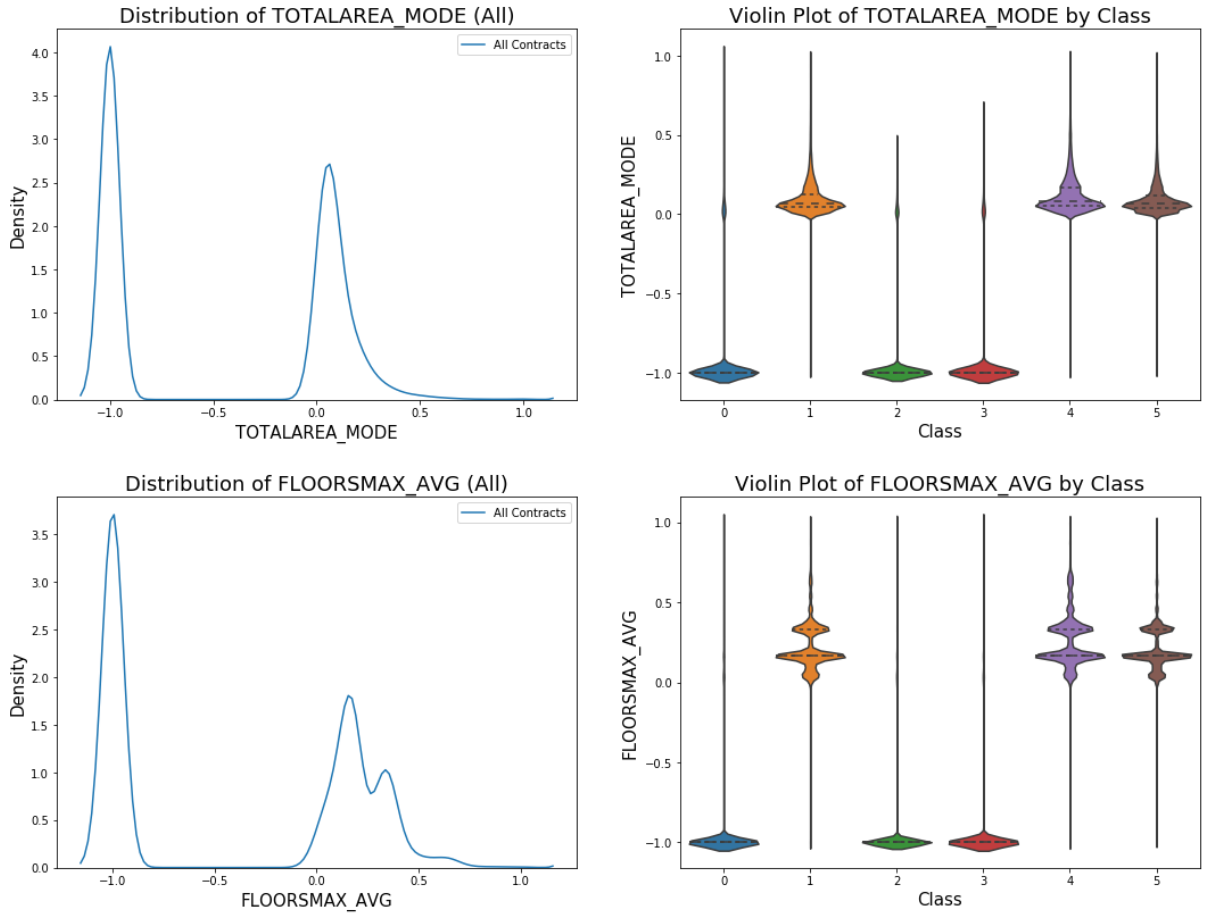

Violin Plot of APARTMENTS_AVG by Class
Apartment_Avg
Class


Distribution of ENTRANCES_AVG (All)
All Contracts
Density
ENTRANCES_AVG


Violin Plot of ENTRANCES_AVG by Class
Entrances_Avg
Class


Distribution of LIVINGAREA_AVG (All)
All Contracts
Density
LIVINGAREA_AVG


Violin Plot of LIVINGAREA_AVG by Class
LIVINGAREA_AVG
Class

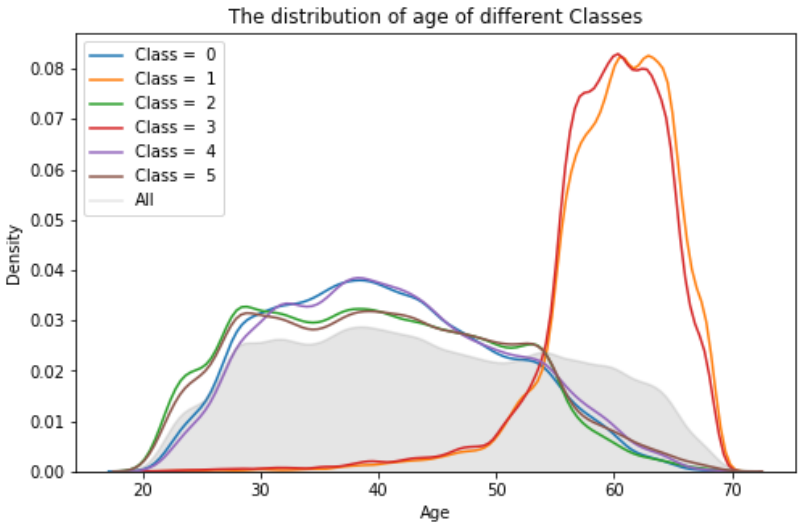


We found that their characteristics are very similar in the same cluster. We can conclude that clusters 1 & 4 & 5 have values of these property features. However, the other clusters 0 & 2 & 3 are most likely to have missing values in these variables. It is notably that cluster 2 has a higher default rate and less likely to fill this information.

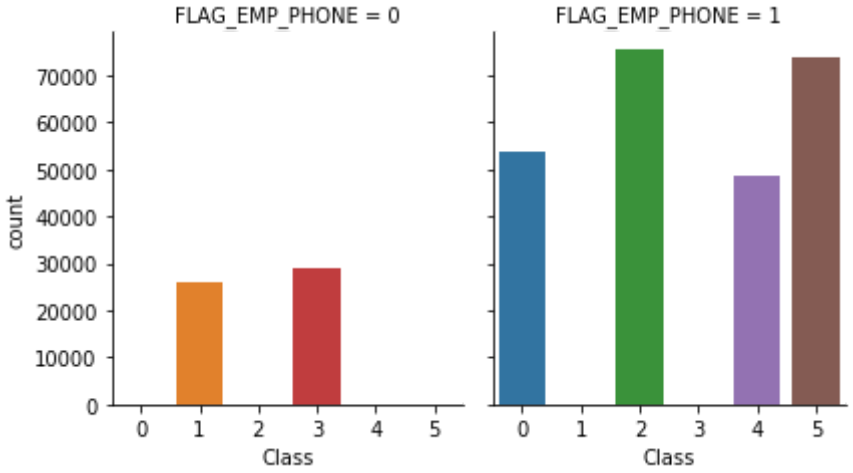
Principle Component 1						
Cluster number	default rate (%)	Apartment_avg	Entrances_avg	Livingarea_avg	TotalArea_mode	FloorsMax_avg
0	8.2	NaN	NaN	NaN	NaN	NaN
1	4.8					
2	11.1	NaN	NaN	NaN	NaN	NaN
3	5.9	NaN	NaN	NaN	NaN	NaN
4	5.7					
5	8.4					

Principle Component 2:

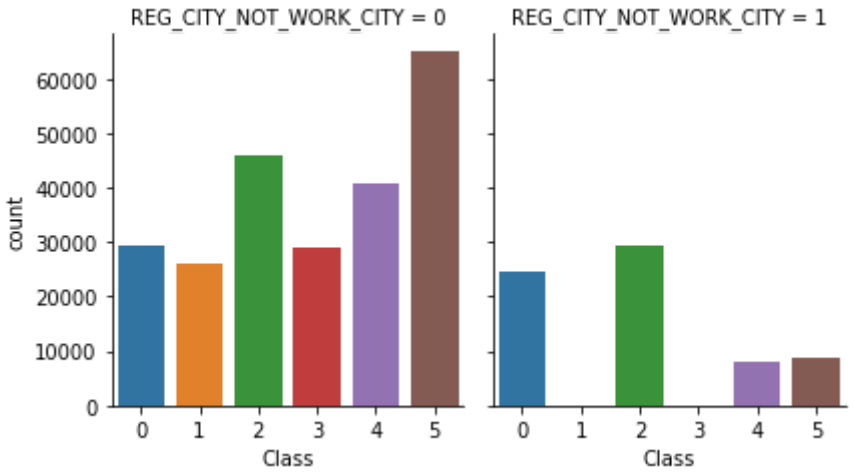
Personal information contributes to this component the most, including age, have work phone information or not, and whether peoples' work addresses match their permanent addresses. Here we can see the details:



Here we can see the distribution of contractors' age of different classes. We find that cluster 1 & 3 have more old people. As they have lower average default rates, we believe older people have lower average default rates than younger people.

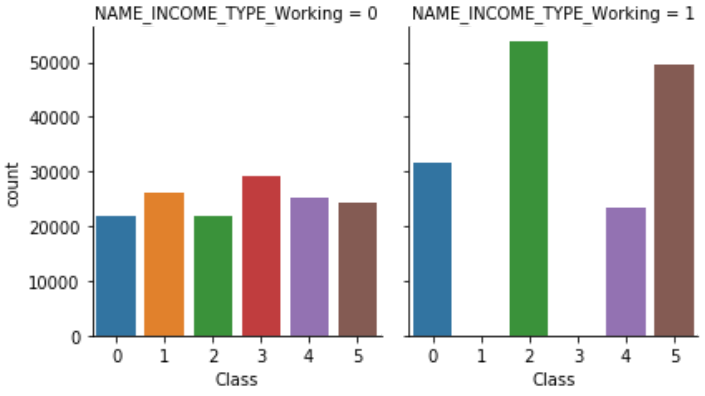


For FLAG_EMP_PHONE we can see that people in clusters 1 & 3 are less likely to have their work phone information, and this is correlated with age variable. We think it is meaningful because they are older and most likely not employed.

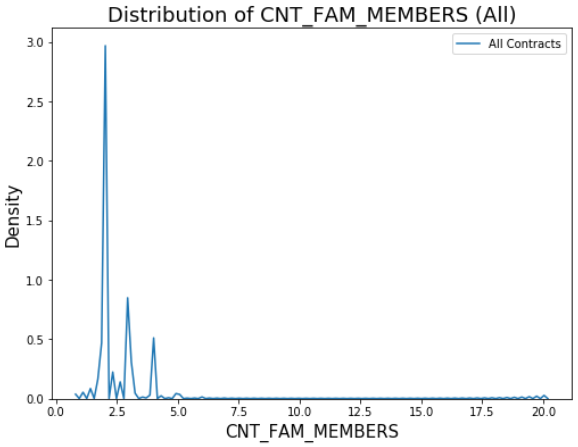


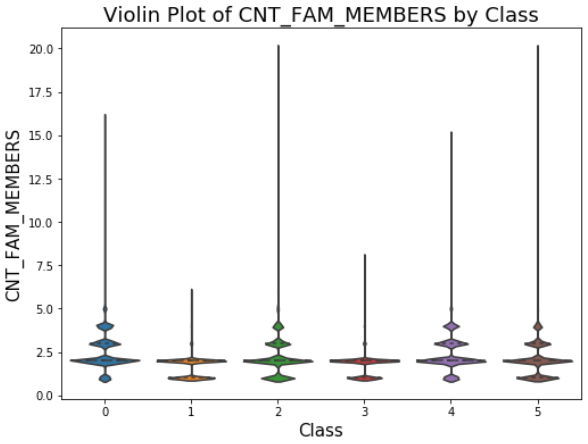
REG_CITY_WORK_CITY: flag if client's permanent address does not match work address (1=different, 0=same, at city level). For this graph, we can see a higher percentage of contractors in cluster 0 & 2 work in a different city to their permanent address.

Then, in order to have more knowledge of this feature, we calculate the default rate of people whose permanent address match work address and whose not match in cluster 2. We find 10.1% of those who have the same work address as their permanent address and 12.6% of those different in cluster 2. It is interesting that cluster 2 has the default rate of 11.1%. Thus, we think this feature is important for predicting the default behavior.



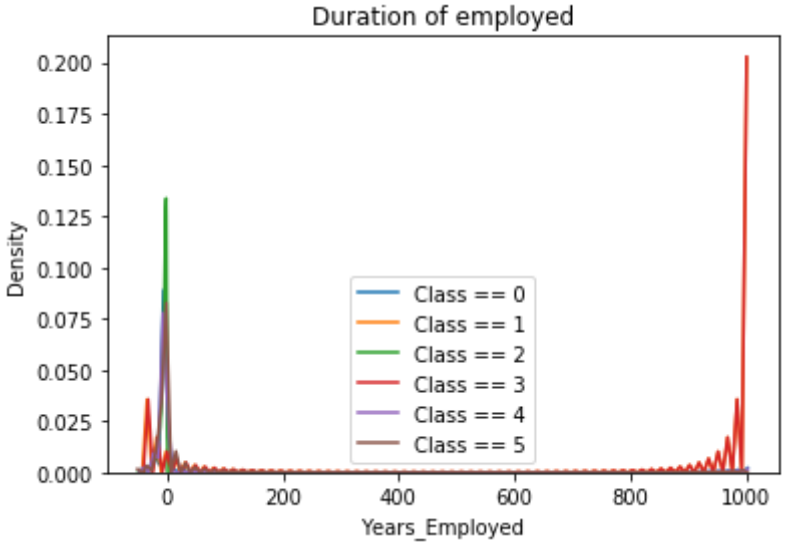
This indicates whether their income is from working or other types. We can see that people in cluster 1 and 3 tend to have incomes not from working. This makes sense because they are older so that they may not be employed at this stage.



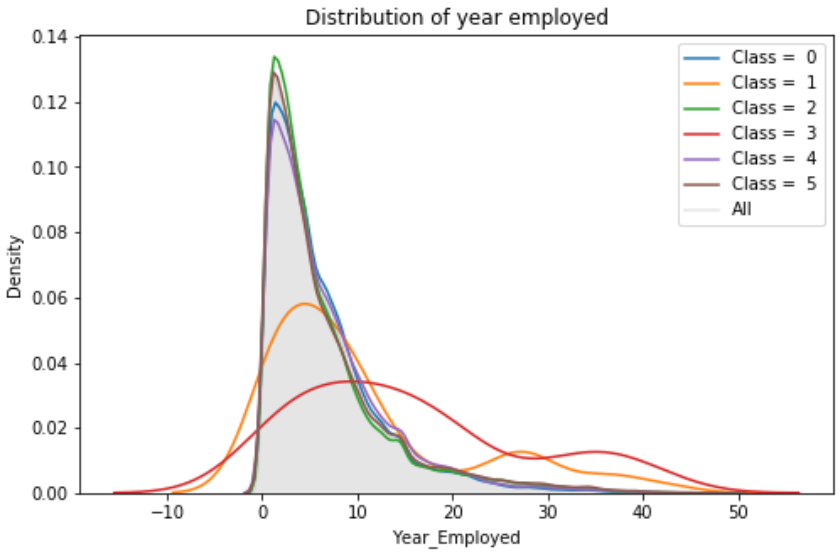


CNT_FAM_MENBERS: indicates how many family members the client has. The value distribution for most of the records in clusters are similar, but the only difference is the range of the distribution. cluster 2 & 5 tend to have a broader range of numbers, from 1 to 20. Compared to that, the range of cluster 1 & 3 is more meaningful, having from 1 to 8 family members.

Besides, we wanted to see the years people have been employed. We find that there are some outliers that are very weird as 1000 years or close to 1000 years.



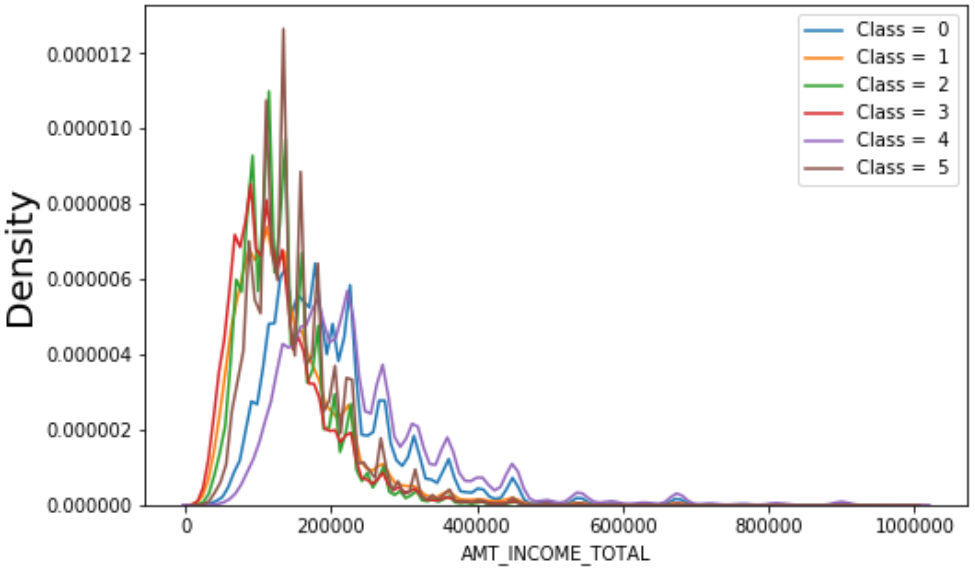
Then we eliminated those outliers and see how it looks in each cluster.



We can see the lines of clusters are closed to the line of mean except cluster 1 & 3. It is obvious as these two clusters have more older people so that in these two clusters they have more people who have worked more than 5 years.

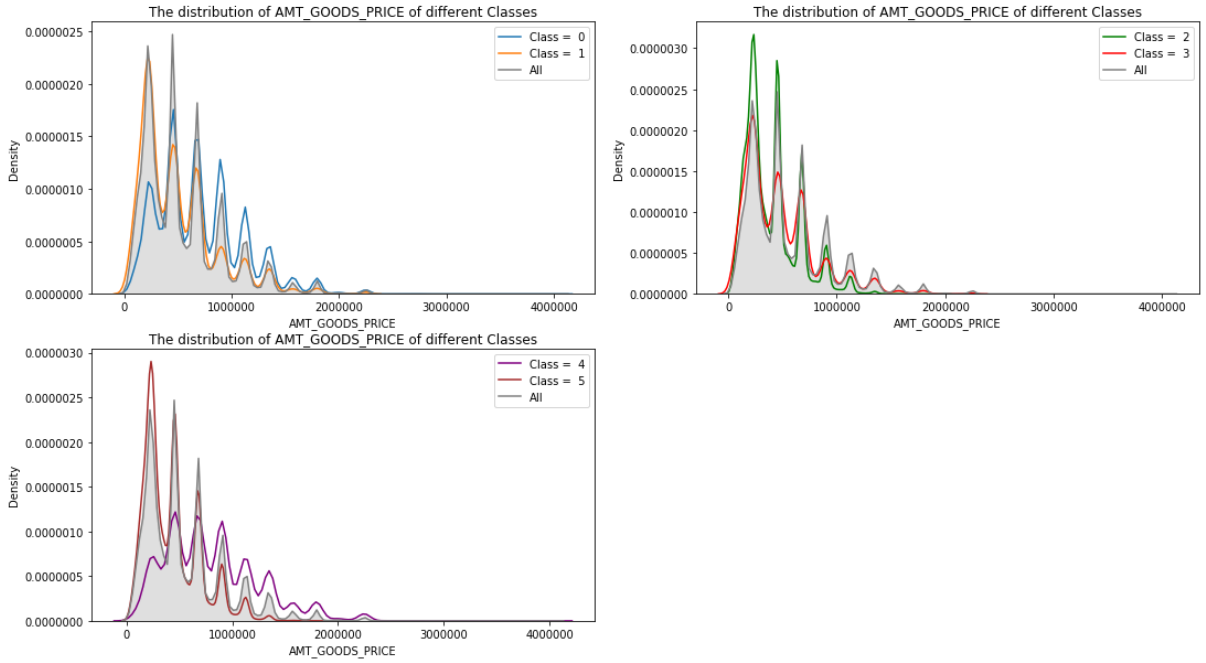
Principle Component 2							
Cluster number	default rate (%)	Age	FLAG_EMP_PHONE	REG_CITY_NOT_WORK_CITY	CNT_FAM_MEMBERS	NAME_INCOME_TYPE_working	Year_Employed
0	8.2		1	more different	higher	less high	
1	4.8	higher	0	same	normal	none	longer
2	11.1		1	more different	higher	higher	
3	5.9	higher	0	same	normal	none	longer
4	5.7		1	less different	less higher	less high	
5	8.4		1	less different	less higher	higher	

Principle Component 3:

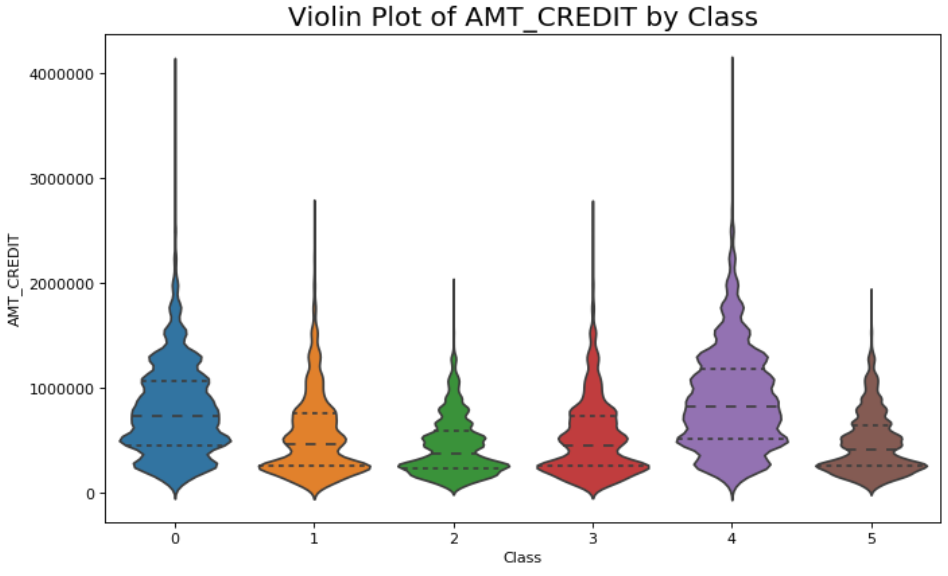


Comparing the income distribution of different clusters, we find that cluster 4, which is one of the clusters with the lowest default, has a higher amount of total income. However,

cluster 0, which has a normal average default rate, also has a high amount of total income and, cluster 3, which has low average default rate, has a low amount of total income. Thus, income cannot be an indicator of default rate.

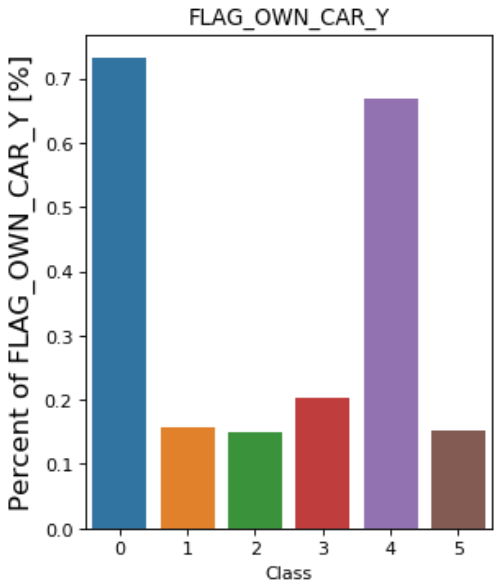


AMT_GOODS_PRICE: For consumer loans it is the price of the goods for which the loan is given. We want to see how this feature is distributed in each cluster. We set clusters into 3 groups, two of which in order to clearly show the trends. We can see that clusters 0 & 4 tend to have higher prices of goods. Other clusters show similar fluctuation to all data points.



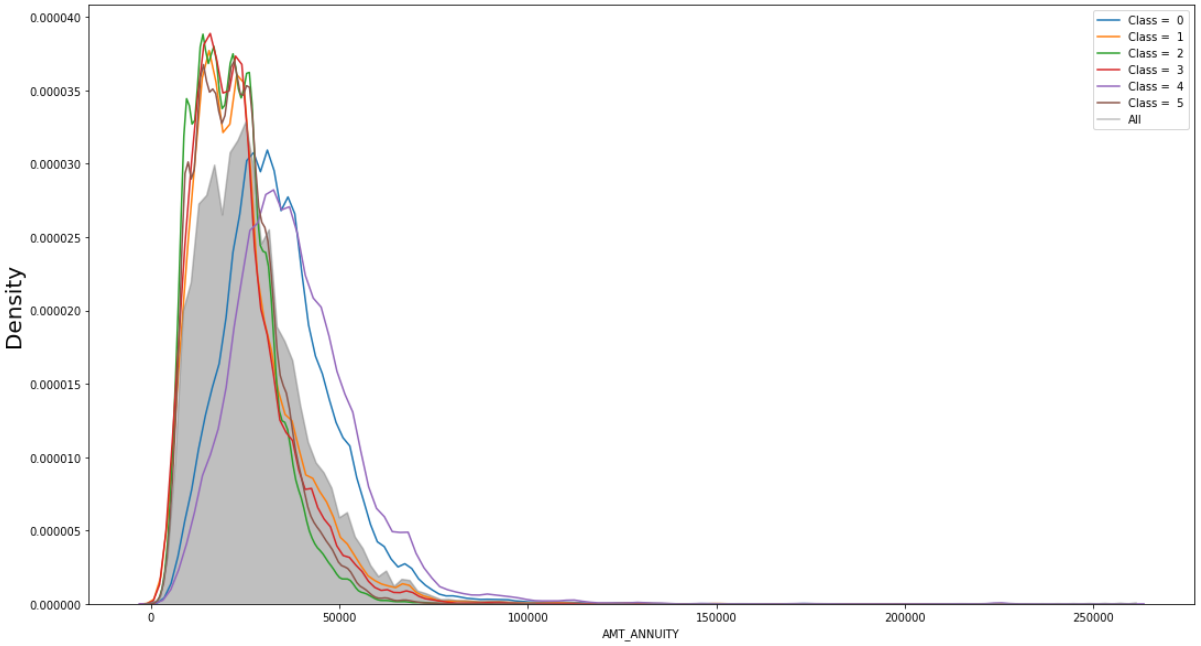
AMT_CREDIT: is the credit amount of the loan. We can see that cluster 0 & 4 have higher credit than other clusters.

Our insights: the interesting thing we find here is that, even clusters 1 & 3 & 4 have the lower default risk, only cluster 4 has a higher credit amount. We think it is because people in cluster 4 have higher annual income and credit amounts are more related to income but not default rate. Our suggestion is to reconsider the credit amount of those in cluster 0 to avoid single default with a large amount of money.

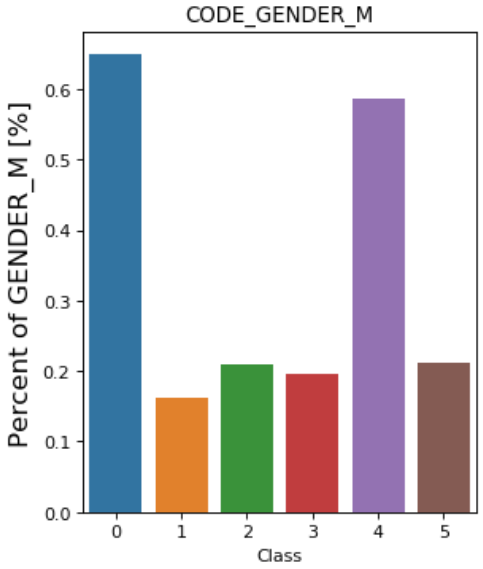


FLAG_OWN_CAR_Y: indicates the client owns a car. We can see that clusters 0 & 4 have a higher percentage of owning a car.

Our insights: here we find that the higher income, the more likely own a car. We cannot say if people have higher income or own a car, they have less default rate, as cluster 0 has the highest rate of owning a car but the average default rate is higher than overall average default rate. However, if people do not have a car or the income is not relatively high, they have more likelihood to default, as the people in clusters 2 & 5 the least likely own a car and have the highest average default rate among all the clusters.



AMT_ANNUITY is the loan annuity. It is obvious that cluster 0 & 4 has higher loan annuity.

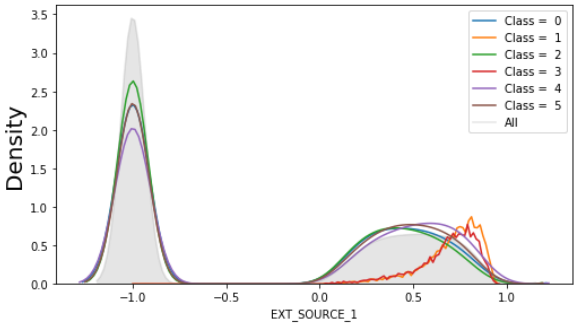


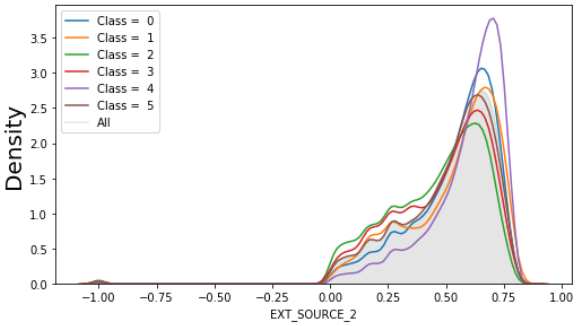
This is the percentage of male in each cluster. The overall percentage of Male is 0.34, so that we can see that people in cluster 0 & 4 are most likely male. The other clusters have a higher percentage of females than total.

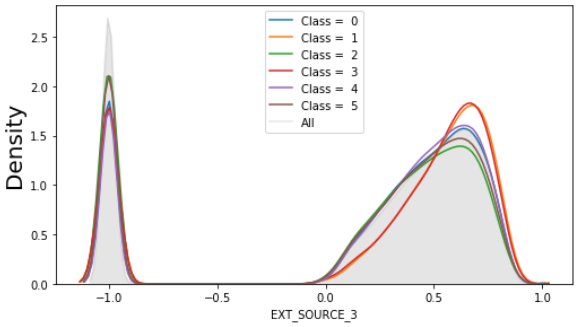
Principle Component 3							
Cluster number	default rate (%)	CODE_GENDER_M	AMT_GOODS_PRICE	AMT_CREDIT	FLAG_OWN_CAR_Y	AMT_ANNUITY	AMT_INCOME_TOTAL
0	8.2	higher	higher	higher	higher	higher	higher
1	4.8						
2	11.1						
3	5.9						
4	5.7	higher	higher	higher	higher	higher	higher
5	8.4						

Other:

External Sources: Normalized score from external data sources







As there are many missing values in external credit sources, thus we label them as -1 to distinguish from other results. It is hard to see large differences for EXT_SOURCE_2 among all the clusters, while people in cluster 4 perform slightly better. However, we can readily see that cluster 1 & 3 have better average credit in EXT_SOURCE 1 & 3. Also, the cluster 2 & 4 & 5 have more missing values in EXT_SOURCE 1 and only cluster 1 have a few missing values in EXT_SOURCE 3. We need more information to get deeper insights why this happens. What we can conclude is that, the higher credit they earned from external sources, the lower default rate in Home Credit.

Summary Characteristic of Each Cluster:

Cluster 0 (8.2% default rate): higher amount of income; higher percentage of owning a car; higher percentage of male; more likely to have missing value of housing information; higher average credit amount; higher amount of goods price; having the work phone; different working city to permanent address; abnormal amount of family members.

Cluster 1 (4.8% default rate): higher percentage of old people (aged 60 or higher); less likely to have information about work phones; longer years of employment; higher credit score at external source 1 & 3.

Cluster 2 (11.1% default rate): more likely to have missing value of housing information; different working city to permanent address; unnormal amount of family members; lower average credit amount.

Cluster 3 (5.9% default rate): higher percentage of old people (aged 60 or higher); more likely to have missing value of housing information; less likely to have information about work phones; higher credit at external source 1 & 3; longer year of employment.

Cluster 4 (5.7% default rate): high average income; high percentage of owning a car; high average credit amount; better credit at external source 2; higher percentage of male; higher amount of goods price.

Cluster 5 (8.4% default rate): normal average default rate; low percentage of owning a car; low average credit amount. Similar characteristics as cluster 2 except for that it has filled property information.

Finally , we suggest that our client attach great importance to these variables. Financial Institutes can classify future borrowers according to these characteristics, and evaluate their default risk.