

# COURSEWORK

EMFSS ST3189 Machine Learning

STUDENT NUMBER:

200693280

Table of contents:

<b>Part 1: (Unsupervised learning)</b>	<b>1</b>
<b>Part 2: (Regression)</b>	<b>4</b>
<b>Part 3: (Classification)</b>	<b>6</b>

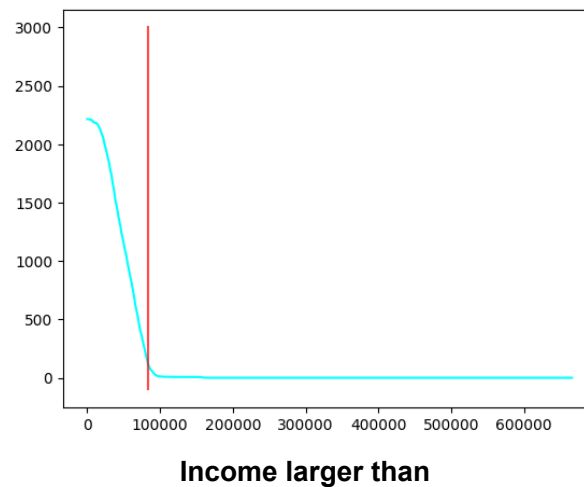
## Part 1:

### Description:

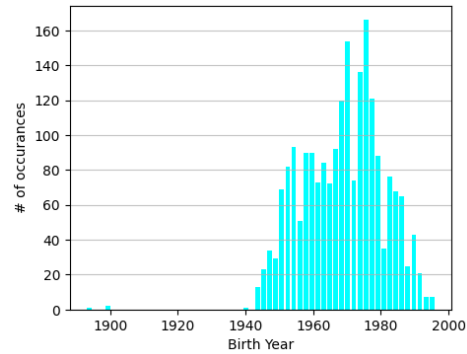
We are going to be looking into a dataset of customers of some company. We have some binary columns containing reaction of each person to each of the previous campaigns (“AcceptedCmp1”, “AcceptedCmp2”, “AcceptedCmp3”, “AcceptedCmp4”) and some general data about the customer such as date of birth, buying habits (“MntWine”, “MntFishProducts”, etc.), income and others which we will discuss more in-depth further.

### Feature engineering:

First, let us look at the number of people who have each type of income:

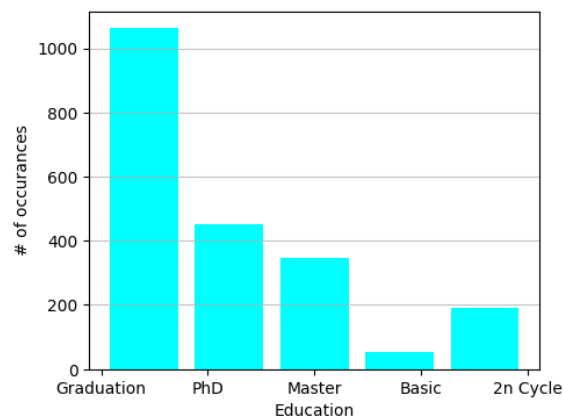


I have highlighted with a red line the 95% quantile for incomes, which turned out to be 84130. It is easy to observe that there are almost no people with incomes higher than that mark. Moreover, ones that are, are extremely high earners, which makes them outliers for this dataset. We need to remove this data from the dataset to prevent clusters of outliers from forming and therefore improve the quality of our analysis.



Taking note of the birth years from our data, we can see that there are some obvious outliers as well, and people from that period would be more than a 100 years old, which is highly unlikely and means that these observations are probably invalid due to reasons like dishonest answers. Based on that, we can also remove all the observations where the birth year is below 1923 (100 years from now).

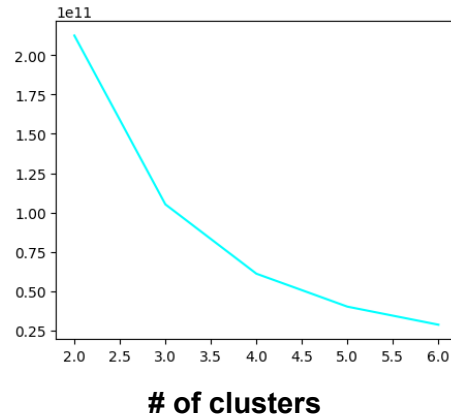
It should be noted that while the last two deletions affected only a handful of rows, and this amount probably would not affect the predictive performance of the regular models, it can have a considerable effect on the clustering models as obvious outliers will be forming a cluster of their own, thus reducing the amount of meaningful clusters.



If we look at the frequency of each of the Education statuses we can see that “Master” and “2n Cycle” degrees are the most uncommon, while also being very similar in nature. This means that this can be an opportunity to reduce dimensionality of the data by combining these two into one.

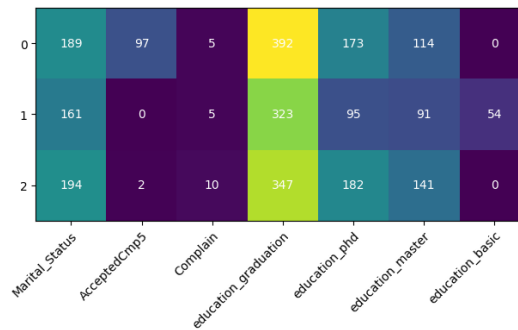
## Clustering:

I have decided to use k-means clustering algorithm for this task, but this technique requires you to input the appropriate amount of clusters in which the data should be separated. To do so there are several methods, but I have decided to use the “Elbow Method”. Let us look at the resulting plot:

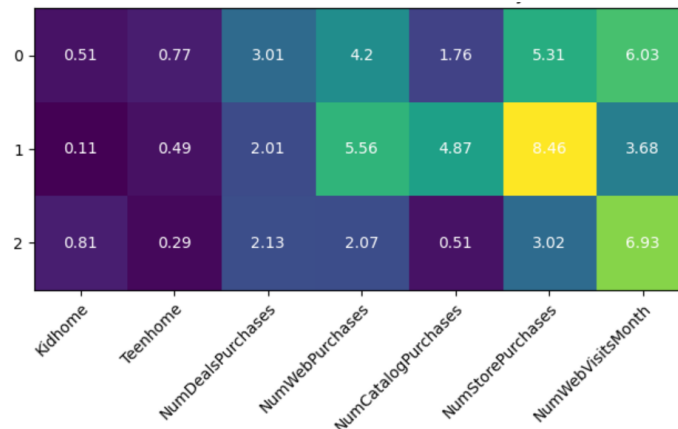


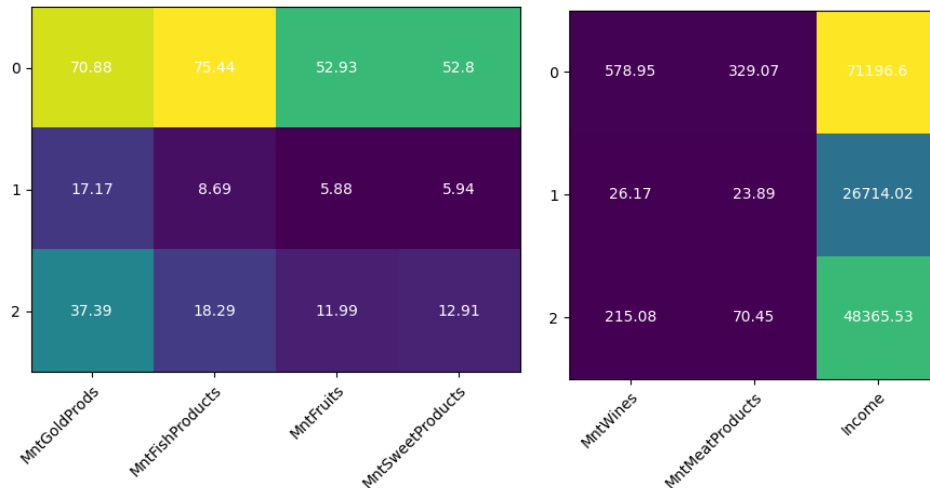
We can see that this plot has a steady decline until the number of clusters reaches 3, but after that, there is a clear “elbow”, which means that 3 seems to be the appropriate number of clusters for our purposes.

After fitting the model on 3 clusters we can construct a series of heatmaps to better understand the data. First of all, let us look at the categorical variables:



Here we can see the number of “TRUE” values for each of the clusters. For example, education\_basic has a value of 54 for cluster 1, this means that 54 people from this cluster have basic education, but for clusters 0 and 2 this value is 0, meaning that people from clusters 0 and 2 have more advanced levels of education. Before making any further conclusions we should look at other columns in our dataset:





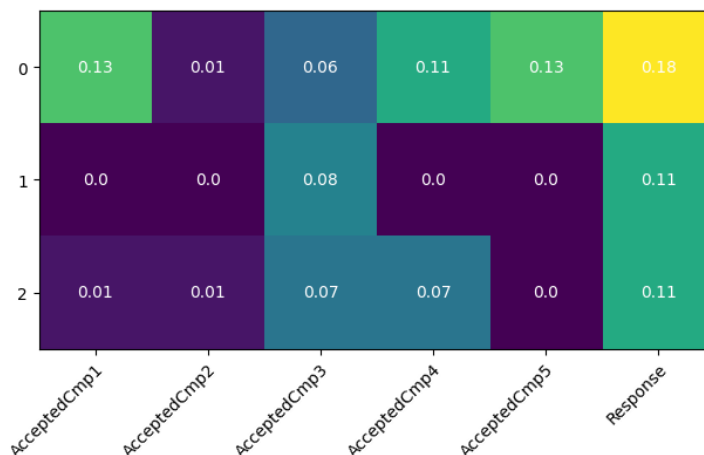
\*For the cells in these plots we will use the averages of the values for each of the clusters.

We can see that consumers from cluster 0 have more teens in the household, make more purchases with special offers, buy more gold and fish products. This means that they make more, buy more expensive items and have a large family. This suggests that they have high incomes, which is further proved by the fact that the average income for these people is significantly higher than that of other clusters.

For cluster 1 we observe that these people have lower level of education, shop online and in Catalog more, as well as buys much less expensive products like wines and meats than the other two groups. I would consider these people to be low-earners which is further proved by the income average as well.

Cluster 2 has people of the “middle class” who hover in between the other two classes.

We can also see what kind of ad campaign is better each of the clusters from the following heatmap:



So, for future campaigns we should use for each cluster campaigns which are similar:

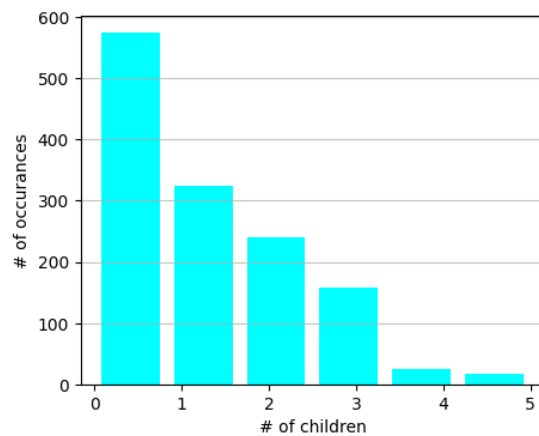
for 0: to 1, 4, 5, Response (the latest campaign)  
for 1: to 3 and the latest  
for 2: to 3, 4 and the latest

## Part 2:

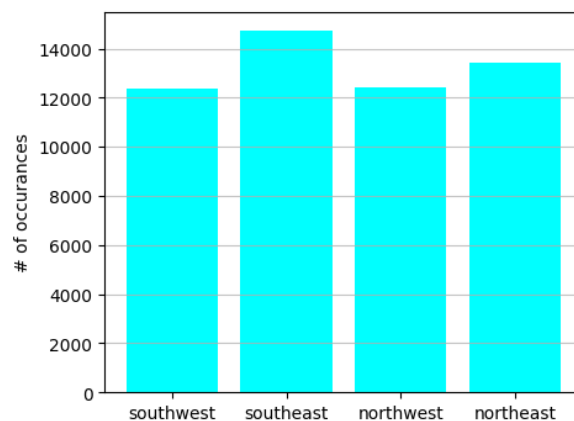
### Description:

This DataSet is about insurance information about people who smoke or not, their age, sex, region, and insurance company charges.

### Feature engineering:



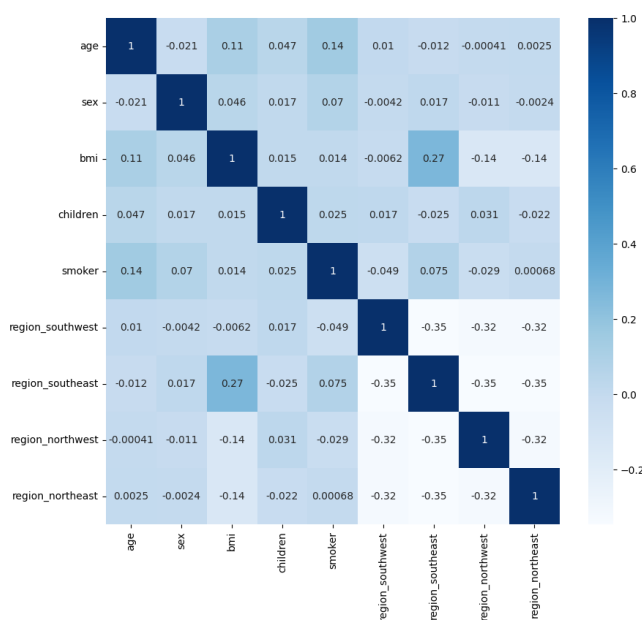
We can see that there are almost no observations where the number of children is 4 or five, so it makes sense to combine them to reduce the dimensionality of the data. Also, from the pairplot we can see that there isn't much of a difference in charges for 4 and 5 children



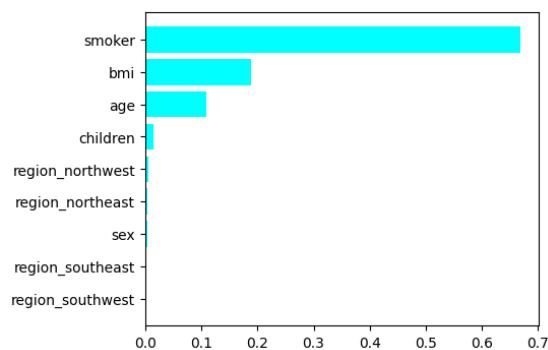
Regions are represented equally, so no reason to modify these classes or worry about representability

We can also add an interaction effect between "smoker" and "age" variables because people of old age who also smoke are worse than people who do only one of these things.

Also we split regions into categorical variables.



There are no meaningful correlations between variables which means that based on this plot alone we cannot think of excluding any variables.

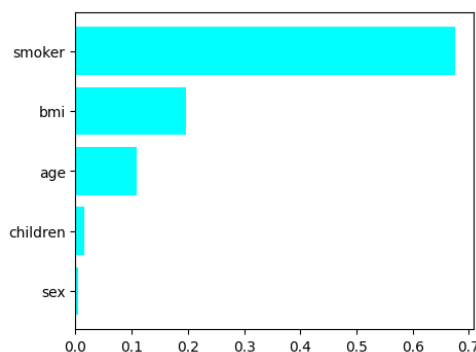


We can see that Random Forest deems "smoker" variable the most important, which makes



sense because "charges" in our dataset are based on Individual medical costs billed by health insurance, and smokers generally receive worse rates for insurance because they have higher health risks. The same can be said about the BMI and age. People who are obese run higher risks of heart issues, people who are too skinny - malnutrition, osteoporosis, decreased muscle strength, hypothermia and lowered immunity. When aging people also become more susceptible to illnesses and suffer from them in more severe forms, many illnesses are also associated with aging, meaning that such a people will spend more money on their health, therefore, insurance for them is generally considerably higher.

As we saw in the feature importance plot that regions don't have much meaning for the random forest. Therefore we will now try to drop them to potentially make the forest focus on more important features.



As we can see, MAE has improved.

By using bootstrap and LGBM model we can further Improve the result to the following MAE value:

MAE: 2254.8707291562996

### Part 3:

## Description

We are given a dataset, containing information about people who have mortgage and credit loans and also about their monthly income and the ratio of debt to income.

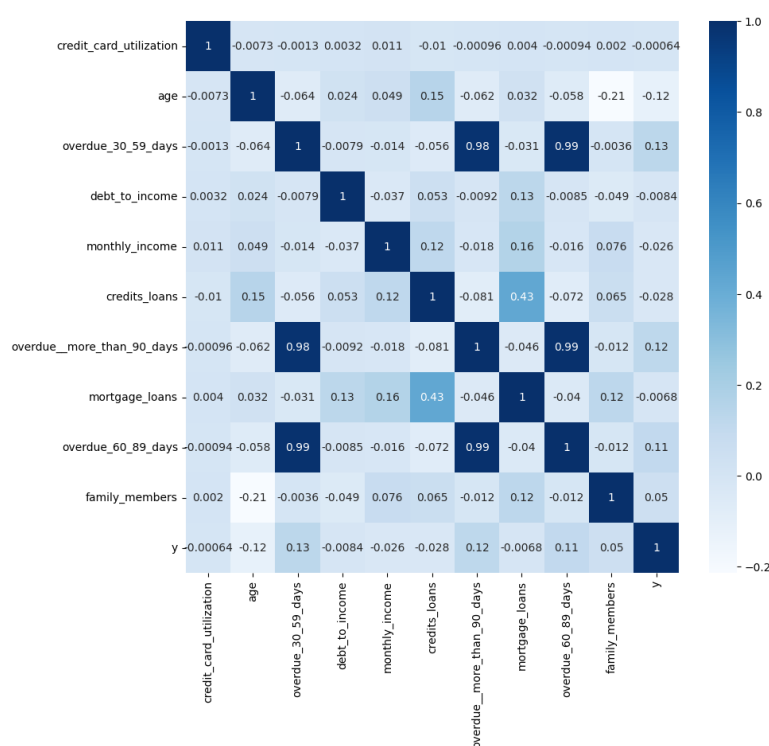
For us, the most important thing is to minimize the number of real defaults that were classified as non-default, that is, after the selection of clients by the model, the bank would receive as few clients who would not be able to repay the loan, as possible. However, it is also important that the number of issued loans does not decrease significantly after the model has been run. That

is why we will use two metrics to assess the model's performance: FOR(False Ommission Rate) will show how many clients, of which the model offered to issue a loan, will be defaulters, and FPR(False Positive Rate) will show how many of all non-defaulters were denied a loan by the model. By minimizing both metrics we will achieve the best model.

## Feature engineering

By the provided research we can see that columns 'monthly\_income' and 'family\_members' have lots of nan's in them, so we should replace them.

By correlation matrix we can see that family members have correlation with age



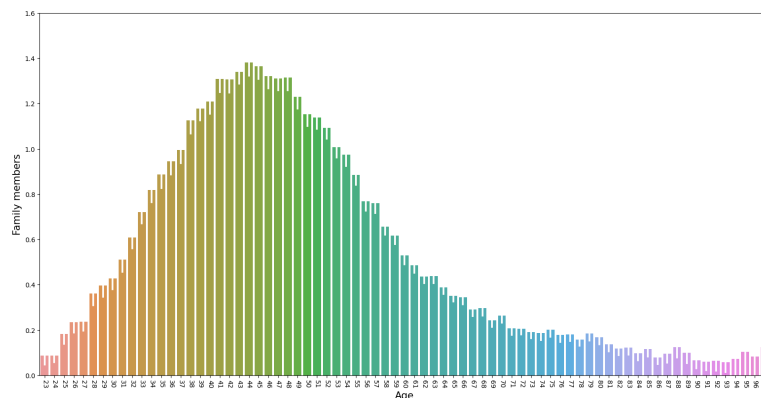
So we can replace nan's with average family members for people of this age, rather than replace it with a simple mean value from the whole dataset.

Also by correlation matrix, we can see that monthly income correlates with mortgage loans and credit loans, so we can make regression that will predict monthly income in missing points.

Again using the correlation matrix we can see a high correlation between three overdue columns. To avoid multicollinearity we can sum all these columns with coefficients 1,2,3 (the more days of overdue - the larger the coefficient is ).

Also, we can deal with family members. This variable seems to be useless on its own, but we can try to divide monthly income by it.

Farther we can deal with debt/income. As we already have a monthly income in the columns, then to avoid multicollinearity we can multiply mon



thly income by debt/income to have net debt in columns.

## Dealing with imbalanced target

I checked Random undersampling, Random oversampling and Balanced Random Forest.

Balanced Random Forest and Random undersampling turned out to be the best ways to deal with the imbalanced target so we will work with them further.

## Profitability of the model

I find out the Operating profit without model.

We will test them in terms of the best approximate operating profit that they can give us. We will assume that in case of default we lose all the money that the client could spend (i.e his/her credit limit) and we will use an interest rate of 20% because it is the nearest value to the real credit card interest in the US dollars now. We considered the credit limit as half of the income of

a person because such restrictions are very popular in banks right now.

I find out the Operating profit with different models.

To calculate operating profit with model we will count it only for those clients that are predicted to be non-defaulters by the model. For different models we may have different optimal thresholds may be optimal in terms of operating profit so we should try several options.





Now, as one of the two best resampling strategies is Random Undersampling we can use it not only with Random Forest, but try other models.

The next step was Stacking with Random Forest

```
Stacking 1: 3 772 312
Stacking 2: 3 769 629
Stacking 3: 3 779 963
Stacking 4: 3 779 627
Stacking 5: 3 784 322
```

As we can see, the best models in terms of operating profit are Stackings with Balanced Random Forest but no resampling.

## Results

	Model object 	Increase in oper... 	Accepted defau... 	Denied non-def... 
0	Stacking BRF & RF & CAT	12.33%	4.09%	4.83%
1	Stacking BRF & XGB & LGBM	12.29%	4.06%	4.96%
2	Stacking BRF & LGBM	12.28%	4.06%	4.97%
3	Balanced Random Forest	12.06%	3.75%	6.3%
4	Random Forest	11.82%	4.19%	4.55%

We can see that with the Stacking of Balanced Random Forest, Catboost, and Random Forest we again get the best result by operating profit. This model gives us a 12% increase in

operating profit, which is a significant improvement for the bank. So we can conclude that the bank should implement this model to check clients before issuing a credit card for them.