**G1 is divided into 3 classes: 0-9, 10-15, 16-20**

**Dataset used for training: por 70%**

**Distribution of G1: 0-9: 24.2%, 10-15:68.7%, 16-20:7.1%**

**The accuracy of the model prediction with: multinomial Naive Bayes, Random Forest, XGBoost Classifier**

|  |  |
| --- | --- |
| **MNB** | **68.2%** |
| **RF** | **68.2%** |
| **XGBC** | **70.7%** |

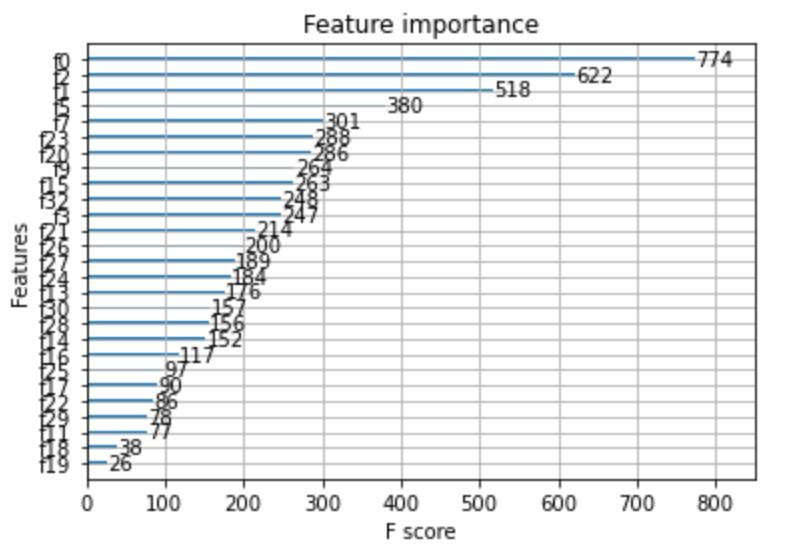
**Problem: a very dominate group of G1 graded within 10-15**

**Feature importance, based on XGBoost classifier (due to its better performance)**

**But it should not be much different from the other 2 classifiers**

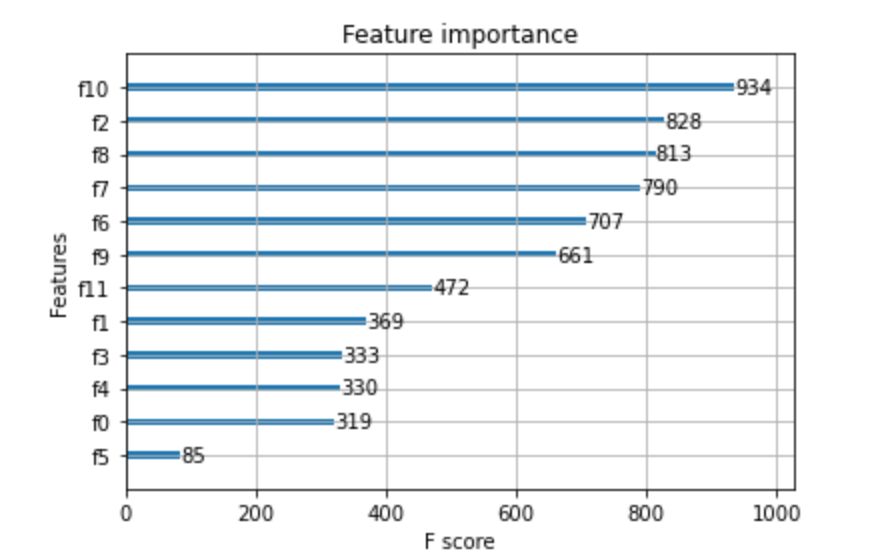
**If take features as one-hot input:**

|  |  |  |
| --- | --- | --- |
| **F0** | **Age** | **774** |
| **F2** | **Fedu** | **622** |
| **F1** | **Medu** | **518** |
| **F5** | **Sex\_F** | **380** |
| **F7** | **Address\_R** | **301** |
| **F23** | **Reason\_course** | **288** |
| **F20** | **Fjob\_other** | **286** |
| **F9** | **Famsize\_GT3** | **264** |
| **F15** | **Mjob\_other** | **263** |
| **F32** | **Internet\_no** | **248** |
| **Table 1** | |  |



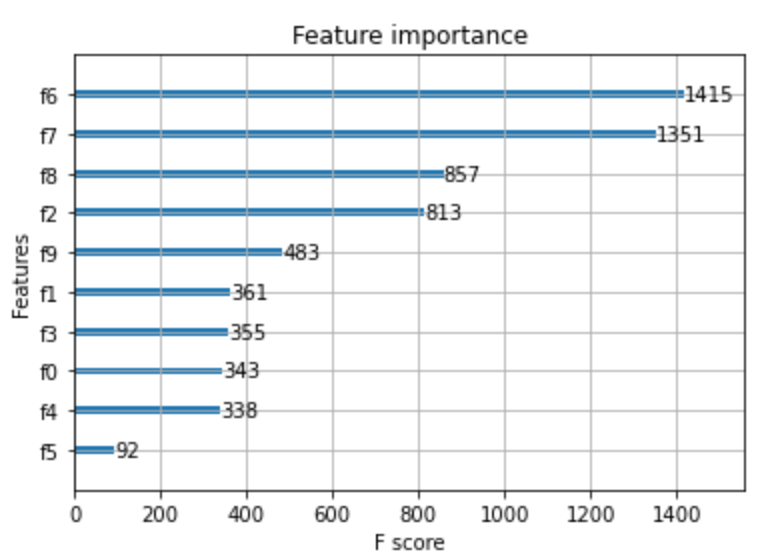
**If take features as numeric:**

|  |  |  |
| --- | --- | --- |
| **F10** | **Reason** | **934** |
| **F2** | **Age** | **828** |
| **F8** | **Mjob** | **813** |
| **F7** | **Pstatus** | **790** |
| **F6** | **Medu** | **707** |
| **F9** | **Fjob** | **661** |
| **F11** | **Reason** | **471** |
| **F1** | **Sex** | **369** |
| **F3** | **Address** | **333** |
| **F4** | **famsize** | **330** |
| **Table 2** | | |



**If merge Fjob,Mjob and Fedu,Medu:**

|  |  |  |
| --- | --- | --- |
| **F6** | **Merge\_edu** | **1415** |
| **F7** | **Merge\_job** | **1351** |
| **F8** | **Reason** | **857** |
| **F2** | **Age** | **813** |
| **F9** | **Guardian** | **483** |
| **F1** | **Sex** | **361** |
| **F3** | **Address** | **355** |
| **F0** | **School** | **343** |
| **F4** | **Famsize** | **338** |
| **F5** | **Pstatus** | **92** |
| **Table 3** | | |



**Findings:**

1. **The importances of Mjob/Fjob/Medu/Fedu are randomly more important than other features as can be seen from the first and second table, when merged job and edu together (by multiplying with +1 smoothing), the importances of merge\_edu and merge\_job has become the top 2 in the table 3**

**\*this might because of the multiplication results in a larger value than other features? Not sure**

**\*can also see from the comparison between table 2 and 3 that the Pstatus importance has decrease**

1. **Features in Table 3 generally have higher F-scores which indicating the feature importances. In table 3, the difference of the F-scores between the first 2 and the second 2 features; and the fifth feature are big, but the differences between the rest are small, especially features of [address, school, famsize, Pstatus]**

**\*within the 4 features in the bracket, feature importance of address, school and famsize are very similar to those in table 2, despite of Pstatus (which is compliant to findings in 1)**

**Concerns:**

1. **The distribution of the G1 scores, make the classifier tends to predict more (10-15), less fail and almost 0 (16-20), its very hard to predict (16-20), an out-standing score for G1 based on only demographic data.**

**Further improvement:**

1. **Find out more about the relationships between family \parents related features**
2. **In order to come up with the Introduce of new variables**