Assignment 01. Report

Artem Chernitsa, a.chernitsa@innopolis.university, BAI-01

Preprocessing

- *Which regression model was the most effective for the missing values, and why?*

  I used the MAE and MSE as the indicator that allows to choose between linear regression and the polynomial with some degree. Although polynomial could be easily overfitted, linear regression showed much more bad result related to polynomial of the second degree. I think that's just because the features has slightly more non-linear relation between each other, but the second degree is good enough since this degree doesn't tend to overfitting.

- *What encoding technique did you use for encoding the categorical features, and why?*

  I used the one-hot encoding for the $var3$ means countries and ordinal for the $var6$. Firstly 'yes' and 'no' anyway 0 or 1 or [0,1], [1,0]. It can slightly affect the model, but I just adjust it later, there is no big deal. The main thing is that countries are not necessary related in ordering. So, we can't precisely say what Afghanistan takes after Zambia and so on. So, I one-hot encoded this feature. As I tested later it gives much more better result for some of my models. Partially, these new features will still affect, but if the relation is weak, they will not cause many problems.

Training process

- *Which classification model performed best, and why?*

  Depends on the parameters, e.g. I got better result for Logistic Regression using Grid Search CV, but Naive Bayes and KNN in different preprocessing could give better results with the same data set, e.g. if we encode $var3$ with ordinal encoding KNN becomes better. So, means, that we try to choose optimal approach between them. Anyway, Naive Bayes for the non-reduced data dimensionality gave us awful result, almost random prediction of 54%. This is because one-hot encoding feature. Naive Bayes simply makes an assumption that features are independent to each other. This is the trap, because samples with over 200 additional features with only one '1' in the row between them could

affect a lot on the result. It breaks, since if one feature is presented another one couldn't be presented, easy dependency to see.

- *What were the most critical features with regards to the classification, and why?*

  Most critical features according to the logistic regression coefficients are $var5$, $var1$, $var4$, $var2$, $var6$ in this order. The one-hot encoded countries are less important, but some of them better than others.

- *What features might be redundant or are not useful, and why?*

  I found that date isn't really useful for us. I tried the cyclical encoding for it, but for the case of ordinal encoding categorical features models show relatively good results, however they give more without date. Moreover, a lot of dates have the same year, and it's hard to find the right part like month or day or, maybe hour to improve model. Logistic Regression gave the slightly important weights for hours and minutes, but overall score is worse. Finally, I dropped $var7$. Countries also could be redundant, but I have checked and found that they are actually not. They don't have any order, but they still carrying some dependency.

- *Did the dimensionality reduction by PCA improve the model performance, and why?*

  For the some cases definitely yes, e.g. the best result also best using PCA. But for some are not, because it also projects features that are not really needed, so they're differ in average of 2%. PCA in general simply projects on the lower dimensionality and do it the way that maximum information is kept and the "cells" that are highly correlated gather together. So, bringing PCA allows us to use useful data to improve performance in some cases, but in some cases not, due to reasons of "bad" features or wrong number of components, partially date is mostly denied when we decide to apply PCA.

Additional Research

(a) what is a multi-label learning problem? (b) suggest an example in which you can transform the given problem into a multi-label problem? Will the models work as it is in that case, or would some changes be required? (10 %)

It's simply a problem where we should distinct more than 2 values to predict. We just make a new classes, e.g. using power set of given labels and numerate them. So, we need encode this data as I mentioned before, and probably we should also standardise this values.

Additional data logs

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 2
PCA number of components: 1

F1 score for naive bayes: 0.977880 (non PCA)
F1 score for naive bayes: 0.973868
F1 score for logistic regression: 0.978495 (non PCA)
F1 score for logistic regression: 0.978495
F1 score for KNN: 0.983784 (non PCA)
F1 score for KNN: 0.962567
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 2
PCA number of components: 2

F1 score for naive bayes: 0.977880 (non PCA)
F1 score for naive bayes: 0.973868
F1 score for logistic regression: 0.978495 (non PCA)
F1 score for logistic regression: 0.978495
F1 score for KNN: 0.983784 (non PCA)
F1 score for KNN: 0.968085
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 2
PCA number of components: 3

F1 score for naive bayes: 0.977880 (non PCA)
F1 score for naive bayes: 0.972541
F1 score for logistic regression: 0.978495 (non PCA)
F1 score for logistic regression: 0.968085
F1 score for KNN: 0.983784 (non PCA)
F1 score for KNN: 0.973262
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 3
PCA number of components: 1

F1 score for naive bayes: 0.980873 (non PCA)
F1 score for naive bayes: 0.919751
```

```
F1 score for logistic regression: 0.978723 (non PCA)
F1 score for logistic regression: 0.940541
F1 score for KNN: 0.983957 (non PCA)
F1 score for KNN: 0.931217
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 3
PCA number of components: 2

F1 score for naive bayes: 0.980873 (non PCA)
F1 score for naive bayes: 0.965771
F1 score for logistic regression: 0.978723 (non PCA)
F1 score for logistic regression: 0.968421
F1 score for KNN: 0.983957 (non PCA)
F1 score for KNN: 0.968750
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 3
PCA number of components: 3

F1 score for naive bayes: 0.980873 (non PCA)
F1 score for naive bayes: 0.965771
F1 score for logistic regression: 0.978723 (non PCA)
F1 score for logistic regression: 0.968421
F1 score for KNN: 0.983957 (non PCA)
F1 score for KNN: 0.973822
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 4
PCA number of components: 1

F1 score for naive bayes: 0.971119 (non PCA)
F1 score for naive bayes: 0.214749
F1 score for logistic regression: 0.957895 (non PCA)
F1 score for logistic regression: 0.701149
F1 score for KNN: 0.951872 (non PCA)
F1 score for KNN: 0.736264
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 4
```

```
PCA number of components: 2

F1 score for naive bayes: 0.971119 (non PCA)
F1 score for naive bayes: 0.957012
F1 score for logistic regression: 0.957895 (non PCA)
F1 score for logistic regression: 0.947368
F1 score for KNN: 0.951872 (non PCA)
F1 score for KNN: 0.936842
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: no
Polynomial features: 4
PCA number of components: 3

F1 score for naive bayes: 0.971119 (non PCA)
F1 score for naive bayes: 0.958441
F1 score for logistic regression: 0.957895 (non PCA)
F1 score for logistic regression: 0.952880
F1 score for KNN: 0.951872 (non PCA)
F1 score for KNN: 0.953368
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 2
PCA number of components: 1

F1 score for naive bayes: 0.976758 (non PCA)
F1 score for naive bayes: 0.932891
F1 score for logistic regression: 0.978495 (non PCA)
F1 score for logistic regression: 0.946809
F1 score for KNN: 0.961326 (non PCA)
F1 score for KNN: 0.940541
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 2
PCA number of components: 2

F1 score for naive bayes: 0.976758 (non PCA)
F1 score for naive bayes: 0.930118
F1 score for logistic regression: 0.978495 (non PCA)
F1 score for logistic regression: 0.941799
F1 score for KNN: 0.961326 (non PCA)
F1 score for KNN: 0.936170
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 2
PCA number of components: 3

F1 score for naive bayes: 0.976758 (non PCA)
F1 score for naive bayes: 0.959165
F1 score for logistic regression: 0.978495 (non PCA)
F1 score for logistic regression: 0.973545
F1 score for KNN: 0.961326 (non PCA)
F1 score for KNN: 0.983957
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 3
PCA number of components: 1

F1 score for naive bayes: 0.971008 (non PCA)
F1 score for naive bayes: 0.951223
F1 score for logistic regression: 0.951872 (non PCA)
F1 score for logistic regression: 0.951872
F1 score for KNN: 0.945652 (non PCA)
F1 score for KNN: 0.945652
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 3
PCA number of components: 2

F1 score for naive bayes: 0.971008 (non PCA)
F1 score for naive bayes: 0.949758
F1 score for logistic regression: 0.951872 (non PCA)
F1 score for logistic regression: 0.952381
F1 score for KNN: 0.945652 (non PCA)
F1 score for KNN: 0.962162
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 3
PCA number of components: 3

F1 score for naive bayes: 0.971008 (non PCA)
F1 score for naive bayes: 0.951298
F1 score for logistic regression: 0.951872 (non PCA)
F1 score for logistic regression: 0.957447
```

```
F1 score for KNN: 0.945652 (non PCA)
F1 score for KNN: 0.939227
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 4
PCA number of components: 1

F1 score for naive bayes: 0.966719 (non PCA)
F1 score for naive bayes: 0.955555
F1 score for logistic regression: 0.952381 (non PCA)
F1 score for logistic regression: 0.957447
F1 score for KNN: 0.944444 (non PCA)
F1 score for KNN: 0.957447
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 4
PCA number of components: 2

F1 score for naive bayes: 0.966719 (non PCA)
F1 score for naive bayes: 0.955555
F1 score for logistic regression: 0.952381 (non PCA)
F1 score for logistic regression: 0.947368
F1 score for KNN: 0.944444 (non PCA)
F1 score for KNN: 0.945652
```

```
Encoding var3: ordinal
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 4
PCA number of components: 3

F1 score for naive bayes: 0.966719 (non PCA)
F1 score for naive bayes: 0.955555
F1 score for logistic regression: 0.952381 (non PCA)
F1 score for logistic regression: 0.941799
F1 score for KNN: 0.944444 (non PCA)
F1 score for KNN: 0.956522
```

```
Encoding var3: one-hot
Encoding var6: ordinal
Existing var7: no
Polynomial features: 2
PCA number of components: 1
```

```
F1 score for naive bayes: 0.530524 (non PCA)
F1 score for naive bayes: 0.978208
F1 score for logistic regression: 0.983957 (non PCA)
F1 score for logistic regression: 0.984127
F1 score for KNN: 0.972973 (non PCA)
F1 score for KNN: 0.984127
```

```
Encoding var3: one-hot
Encoding var6: ordinal
Existing var7: no
Polynomial features: 2
PCA number of components: 2

F1 score for naive bayes: 0.530524 (non PCA)
F1 score for naive bayes: 0.976870
F1 score for logistic regression: 0.983957 (non PCA)
F1 score for logistic regression: 0.994709
F1 score for KNN: 0.972973 (non PCA)
F1 score for KNN: 0.973822
```

```
Encoding var3: one-hot
Encoding var6: ordinal
Existing var7: no
Polynomial features: 2
PCA number of components: 3

F1 score for naive bayes: 0.530524 (non PCA)
F1 score for naive bayes: 0.976936
F1 score for logistic regression: 0.983957 (non PCA)
F1 score for logistic regression: 0.973545
F1 score for KNN: 0.972973 (non PCA)
F1 score for KNN: 0.973262
```

```
Encoding var3: one-hot
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 2
PCA number of components: 1

F1 score for naive bayes: 0.530524 (non PCA)
F1 score for naive bayes: 0.954345
F1 score for logistic regression: 0.967033 (non PCA)
F1 score for logistic regression: 0.957447
F1 score for KNN: 0.956044 (non PCA)
F1 score for KNN: 0.957447
```

```
Encoding var3: one-hot
Encoding var6: ordinal
```

```
Existing var7: yes
Polynomial features: 2
PCA number of components: 2

F1 score for naive bayes: 0.530524 (non PCA)
F1 score for naive bayes: 0.954345
F1 score for logistic regression: 0.967033 (non PCA)
F1 score for logistic regression: 0.957447
F1 score for KNN: 0.956044 (non PCA)
F1 score for KNN: 0.957895
```

```
Encoding var3: one-hot
Encoding var6: ordinal
Existing var7: yes
Polynomial features: 2
PCA number of components: 3

F1 score for naive bayes: 0.530524 (non PCA)
F1 score for naive bayes: 0.955898
F1 score for logistic regression: 0.967033 (non PCA)
F1 score for logistic regression: 0.967742
F1 score for KNN: 0.956044 (non PCA)
F1 score for KNN: 0.962567
```

The rest of the combinations are slightly worse. Also for the one-hot-encoding the feature $var3$ we hardly can compute the polynomial features of 3rd and 4th degrees. The best result is used is the *Logistic Regression*, applied on one-hot encoded feature $var3$ with *Polynomial Features* of the *second degree* with the *PCA of 2 components* and no *date* is presented.