Vladimir Zelenokor
keksikoz.hs@yandex.ru
Which regression model was the most effective for the missing values, and why?
I applied Linear Regression, and the score for it was: 0.31854379960373524
I used this model because var4 does not have classified value.

What encoding technique did you use for encoding the categorical features, and why?
I used both One-Hot encoding and Ordinal encoder.
Because by using one hot encoding on var6 it will be almost like ordinal encoding
And I used ordinal encoding on var 3 to reduce the number of features to only var 3.
I know that is not the most efficient way because var3 is look like country names and, as it is country names, the distances between countries might have an effect on the ML_models, so it might be better for me to use one hot encoding for it because that will reduce the probability that two countries with big distance between them will be after each other in the ordinal encoding.

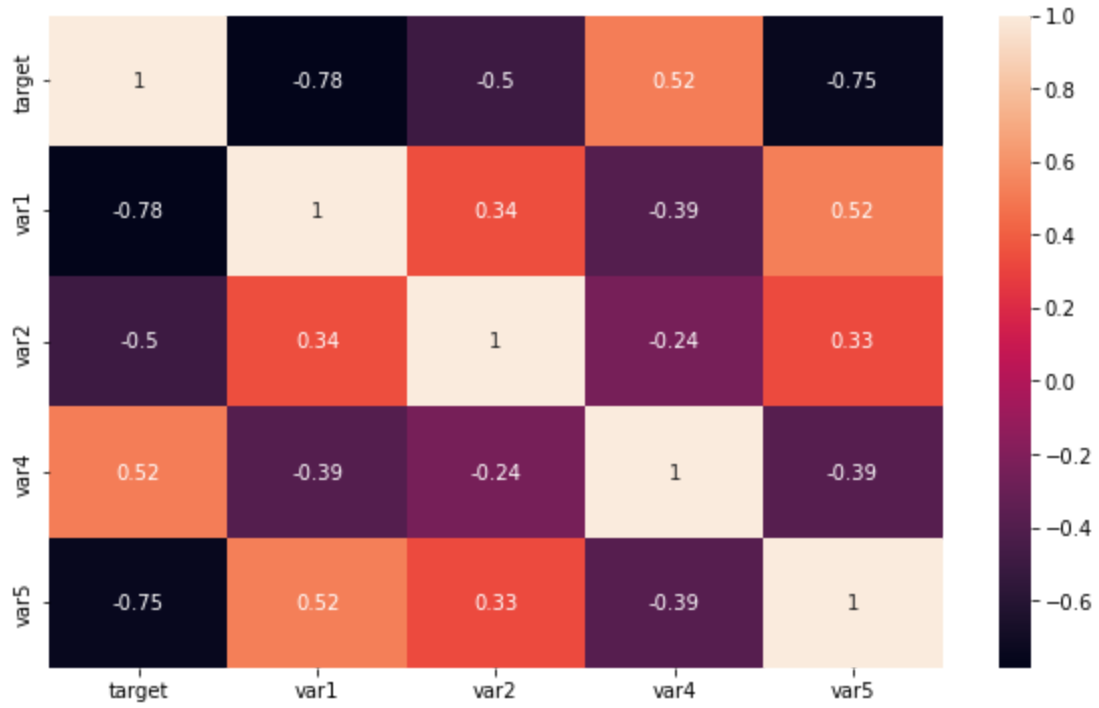
Which classification model performed best, and why?
For me, after applying one hot encoding to var3 and filling the gabs in var4 by linear regression, I noticed:

Score of Linear Regression before applying PCA is = 0.8012264236106603
Logistic Regression score = 0.9538461538461539
GaussianNB score = 0.9692307692307692
grid_search_clf score = 0.9692307692307692

As we see here the GNB and GridSearchCV have the highest score because:
Interpretation, Since we know the contribution of each feature in the prediction, it is easy to understand their influence on the prediction, and it does not require and is not affected by feature scaling.

And also:
Testing precision of GaussianNB= 0.9696969696969697
Testing recall of GaussianNB= 0.9696969696969697

What were the most critical features with regard to the classification, and why?



Correlation measures the linear link between two or more variables. We can predict one variable from the other using correlation. The reasoning for utilizing correlation to pick features is that desirable variables are significantly associated with the aim. Furthermore, variables should be linked with the aim but uncorrelated with one another.
We can anticipate one variable from the other if they are associated. As a result, if two characteristics are associated, the model only requires one of them because the second one adds no more information. The Pearson Correlation will be used in this case.
As the threshold for picking variables, we need to set an absolute value, say 0.5. If we discover that the predictor variables are associated, we can eliminate the variable with the lowest correlation coefficient value with the target variable. We may also compute multiple correlation coefficients to see if more than two variables are connected. This is referred to as multicollinearity.
From the schema var 4 and var1 are the most critical feature.

Did the dimensionality reduction by the PCA improve the model performance, and why?
After PCA:
The score of GNB model when LogisticRegression model were applied 0.9693877551020408
The score of GNB model when GaussianNB model were applied 0.9897959183673469
The score of GNB model when KNN model were applied 0.46938775510204084

AS we see here the GaussianNB model is the best after PCA = 0.9897959183673469
Before PCA: GaussianNB has also the highest score = 0.9692307692307692

After applying PCA, the performance improves.
PCA (and indeed any other compression approach) decreases dimension but not necessarily information: it may just eliminate noise or represent the essence more compactly.