Tolmachev Anton a.tolmachyov@innopolis.university

Theoretical Part.

2.1 Regarding Preprocessing.

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

  Polynomial regression model with 3rd degree happened to perform the best. I iterated degrees starting with 1st to overvue Linear Regression as well. MSE was falling with dropping the degrees down to the 3rd after which models started overfitting drastically.

  Since the dataset provided didn't have any description along with features, I can only wonder if some of them were correlating to some degree, that meant a certain level of improvement as that degree was reflected in polynomial features.

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

  I used OHE (one-hot encoding).

  Since I was aware of only ordinal and one-hot encoding techniques, Those were the only ones to choose from. Albeit both would work, I decided to use OHE, since ordinal encoding has a downside - it makes some features' values matter more than the other. OHE makes them equipotent.

2.2 Regarding the training process.

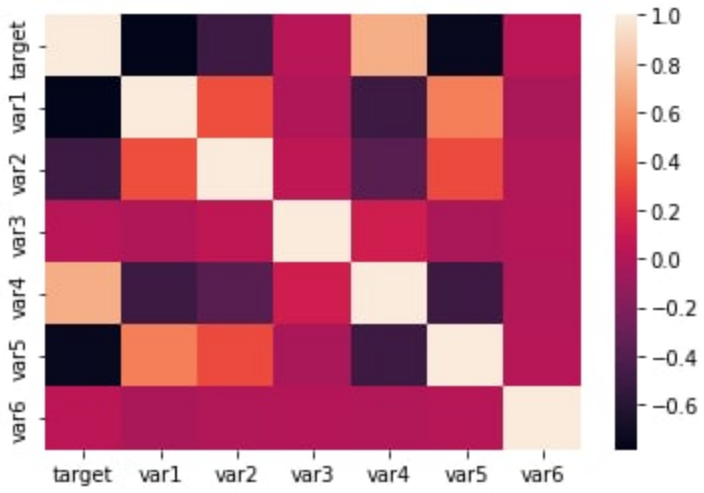
- **Which classification model performed best, and why?**

  The Logistic Regression did.

  I presume that it overperformed Naive Bayes due to the possibility that Laplace estimation in this particular case couldn't fully improve Zero Frequency error, which was pretty likely to be met due to the fact that ~⅔ of the var4 column was predicted as float .

  And regarding KNN, I think that it was inferior to the logistic regression because I didn't scale the data, since doing so downgraded the overall performance.

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



According to the heatmap, var4 feature correlates with the target the most. It might be so because var4 is mostly predicted based on the other features, which so to say comprises them and makes var4 strongly correlate with target in the end. Although var4 is the feature that correlates with the target positively the most, var1 and var5 strongly correlate with it negatively, which simply means a negative sign in models' coefficients.

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

  var3 and var6 correlate with target the least, so they might be redundant, although their combinations in polynomial features might be useful, so such a judgment should be confirmed on a bigger heatmap with all the polynomial features included.

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

  PCA didn't improve performance, moreover it downgraded it. Such an outcome was predictable, since although PCA chooses such eigenvectors that maximize variance in features, the variance's maximum is limited by the original. Moreover, the less degrees there're in PCA's projection, the more that restriction on variance becomes, and PCA ends up losing data and performance all along.