Report 19229028 Khashbat Enkhbat

When analyzing this data set, phishing instances, we could see three variables in each column. These variables include Type, Subject and Text. As Orange can use Subject and Text features to predict Type, we can that the implemented results are in picture 1. However, because the Logistic Regression could not see any correlation, we can see that Precision is not the best, which is 0.055. It would make sense because all the instances of feature Subject and Text are different in some way, as they are different form of texts. What can give us better results, or how can we make correlation in these two features. We can add corpus function to give us bag of words. It would make our prediction models re-work and give us better results. For reference, we can see picture 3. We need to add a few different functions to our previous workflow. These new features can be found from Orange add-on “Educational”. With it we can see our precision dramatically increase to 0.839. It would make sense because the models now correlate with the instances of specific words. When we see Confusion Matrices related to our new workflow. We can see that Logistic Regression gives us the best result, with a total of 8 false assumptions in picture 4. The other models such as Random Forest and Neural Network give us 10 or more errors.









