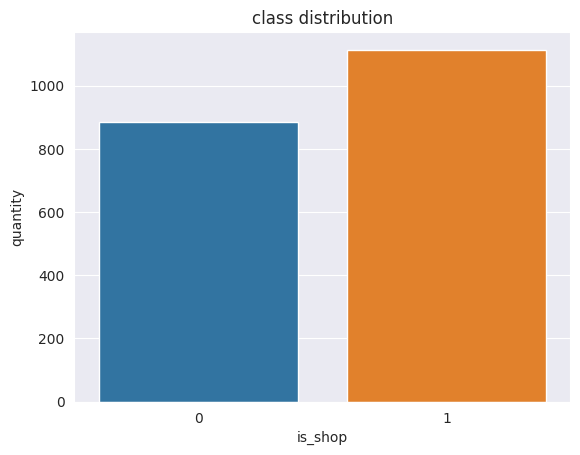
# **Report**

## Initial analysis

You can find the code [here](https://colab.research.google.com/drive/1JwcYbCSJa4-wDlMsb4IPehbf8FKMtGRg?usp=sharing)

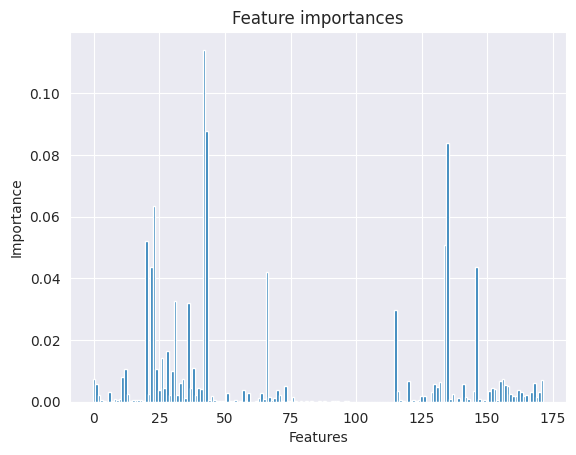
First, let’s see how balanced the is\_shop variable is. As you can see in the plot below is\_shop is almost balanced. So we shouldn’t have a problem with that.

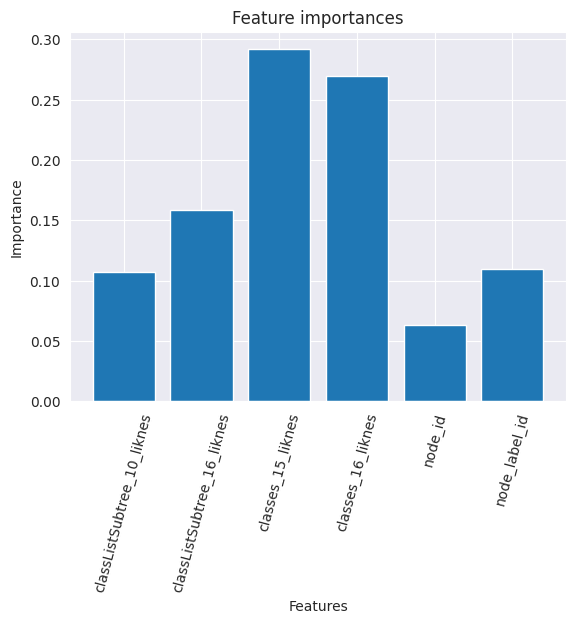


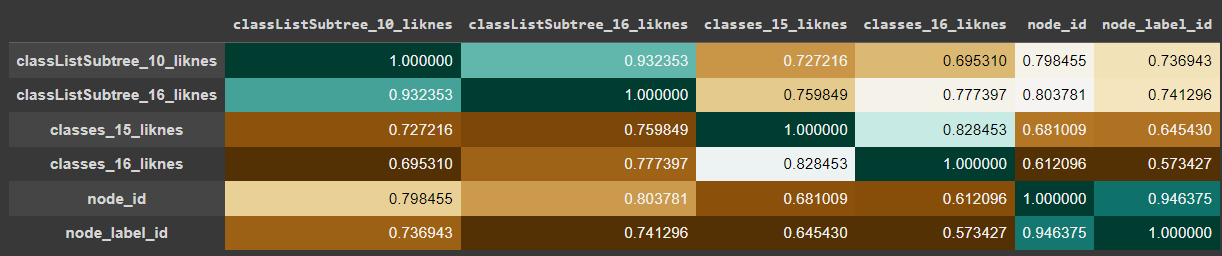
There are 211 columns in the data. Of which 179 are numerical and 32 are not numerical. In general not numerical values are mostly connected to HTML code of the website and we will drop them for the initial analysis part. We will also drop columns that have null values in them.

Classification with only numerical non-null values already gives pretty good results. In the case of logistic regression, we have 0.965 accuracy on the test set, 0.9675 using SVM, 0.99 using random forest, and 0.99 for xgboost. Precision and recall are pretty high too. In the case of xgboost we have 0.989 score for precision and 0.99 for recall. So our initial assumption for not having a problem with imbalance is correct.

Now lets analyze feature importance. We will take feature\_importances\_ from the random forest classifier. You can see the plot below



Let’s select the most important features and do classification only using them. We will take features that have an importance more than 0.05 (6 features). The accuracy on the test set is 0.9825.

Important features are highly correlated. If we will use only 1 feature (most important) for classification accuracy will be 0.945. You can see the correlation matrix of the most important 6 variables below

## Non-numerical values and nans

Non-numerical features either dropped then have a lot of nulls in them, either processed into numerical ones or one hot encoded

Features containing Nans either dropped if they have a lot of Nans or filled with some values. In the case of one hot encoding, they are left untouched

After all processing simple logistic regression achieves 0.9925 accuracy on the test set. While random forest overfits on the training set