# Corpus preparation

## Data Cleaning

## Random Forest

Checked whether the dataset has missing values. None was found.

Checked whether the dataset has any duplicates. The 12 duplicates found in the dataset were removed.

The “duration” feature was dropped from the dataset to avoid data leakage. Data leakage occurs when information that wouldn't be available at prediction time is used during model training. The purpose of the prediction is to identify customers that are likely to subscribe to a term deposit before making a call, “duration” of a call is obtained only after a call is made.

Checked for structural errors in categorical features by obtaining the unique values of each feature.

Boxplots of numerical features were used to identify outliers in the dataset. The boxplots gave outliers for age, campaign, pdays, previous, cons.conf.idx.

The outliers in age are valid outliers, the age values greater than 69.5 up to the maximum age of 98 were displayed as outliers, but it is possible for clients to have such ages, therefore the outliers were not handled.

To address outliers in the 'campaign' feature, quantile-based flooring and capping was applied by computing the 10th and 90th percentiles. Approximately 6% of the values exceeded the upper bound and therefore were capped.

The pdays feature had very low variability because 96% of the values were 999 indicating no contact by a previous campaign. This was left to be handled by the model.

Similarly, with previous, about 86% of the values were 0 indicating the number of contacts performed before the campaign is zero. This feature was also left unaltered to be handled by the model.

Finally, cons.conf.idx (consumer confidence index) had 1% outliers. This was also left to be handled by the model because of the low number of outliers.

According to the documentation the missing values were encoded with the “unknown” label. The categorical features with this label were identified and processed.

Job, marital, education, housing, loan rows were dropped as all these features had less than 5% rows with the “unknown” label.

The default column had 21% of those labels, therefore those rows were imputed with the mode.

The feature poutcome had a label called “nonexistent” for 86% rows of the dataset. So, feature dropped.

The class imbalance ratio was 9:1, therefore class weight was set to balanced when creating the model.

## Neural Network

# Solution methodology

# Evaluation criteria

# Model evaluation

# Experimental results

# Limitations

# Further enhancements

Need appendix with all source code as text