For the given data set I have decided to use three different models such as Logistic Regression, Random Forest and Deep Leaning (ANN). Logistic Regression predicted “y” feature with 87% accuracy rate, Random Forest algorithm predicted 89%. The highest accuracy rate of 97% has been achieved by using Deep Learning model with two inner layers.

The advantages of Logistic Regression model is the implementation simplicity, the extension to multinomial regression and it is less incline to overfitting. In my solution, I have used regularization technic with L2 penalty to avoid overfitting. The disadvantage of this model is that non-linear problems cannot be solved meaning that there is an assumption of linearity between dependent variable and independent variables.

The advantages of Random Forest algorithm is the ability to handle both linear and non-linear relationship and it is not influenced by outliers. The disadvantages of this model is the difficulty to interpret and control and it can be very computationally intensive.

The tf.keras.Sequential model produced the best result and it is also my preferred choice for this dataset. It was implemented with two inner layers. The advantages of this model is the ability of embedding pre-processing layer that means that accepts raw data and all features are pre-processed inside of the model before prediction. The data can be read from file with specified batch size which allows to create data pipeline. To answer the last question, what can be changed in implementation to meet the business requirement is to add pre-processing layer to Keras model and implement data generator for large dataset.