Data Preparation

First, we need to prepare data to create appropriate numbers for machine learning algorithm.

1) First division was to create 6 groups from tenure variable. I prepared 6 groups with next separations: '0-12', '12-24', '24-36', '36-48', '48-60', '60+'.
2) Then I found attributes which are objects and have 2 unique variables. After I encoded them by LabelEncoder.
3) Also, I encoded by get_dummies attributes with more than 2 variables.
4) And of course, I got X (independent) and y (depended).

Modeling and first evaluation.

Second, I made comparing cross validation for Random Forest and Logistic Regression model with 5 cv.

Logistic model gives a better result in mean by default and evaluate the data better.

With roc-auc metric:

Random Forest: array([0.93023152, 0.93502354, 0.92279772, 0.90450948, 0.91829531]) - 0.9214091275955634 mean.

Logistic Regression: array([0.94110602, 0.94085946, 0.93662781, 0.91852756, 0.92074148]) - 0.9315724665045515 mean.

Giving better results.

What if we use something what can gives us better results for predictions?

I choose the FNN and before predictions made Factor Analysis to decrease quantity of attributes.

I evaluate by Kaiser mark the data and got low result (0.2770658781525998) which tells us that Factor Analysis won't give us a result. But I will use it to decrease not only quantity of attributes but the quantity of epochs for modeling. Let's check the results.

1) I made Factor analysis on Train data exclude Test data.
2) Then check Eigenvalue to choose optimal quantity of Attributes.

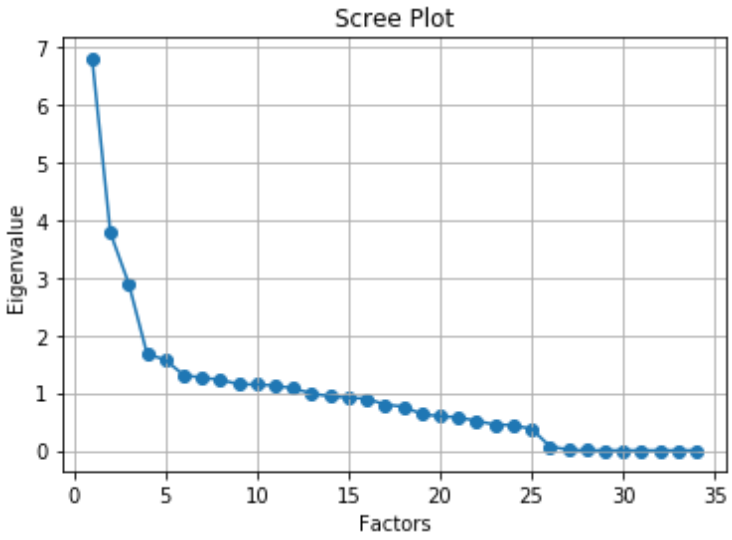

Figure 1 - Eigenvalue, factor analysis

Eigenvalue shows us that 6+ attributes will be ok.

I choose 10 factors with varimax rotation and then prepared data for FNN.

Feed Forward Network.

I created simple architecture with 10-5-2 layers. For input I choose relu activation and softmax for output. For loss categorical_crossentropy is the best with adam optimizer. Metric is auc-roc.

50 epochs and 10 batch-size gives us 94.11% for auc metric.

Discount strategies.

Maximum profit for strategy A I got with 0.9 threshold. But the best result for accuracy_score I got with 0.6 threshold. It is because we have calculation for profit relying to TN TP FN FP directly. Accuracy shows the mean result by all that set. The maximum profit and profit per customer is 68634.389 and 11.358 accordingly.

Strategy B (60733.169 and 10.050) doesn't give better result than A, so I choose strategy is as more profitable.