Before we start analyzing the data, we should initially look at them. By describe() method we can make initial descriptive statistics.

In our case, It tells us, that we have 1.3e5 records. Also, we can get some insights into the number of transactions. It is obvious, that since the median (50% percentile) is negative, more transactions lead to cash outflows. As expected from the insights in the Descriptive statistics and analysis section, most transactions are money outflow. The overall distribution of transaction sums can be viewed using a boxplot. We have compared several models, as a result, boosting algorithms (LGMBoost) have shown the best roc-auc score. However, we have tested logistic regression, applied Cross-validation and grid search to this model and improved roc-auc score from 0.64 to roughly (with the best threshold) 0.7. F-score did not change much.

Since we have no imbalance in data (0s and 1s are in the roughtly same proportions), evaluating the model using Roc-Auc curve is more reasonable.

To conclude, our model is good at recognizing gender using transactions history.