Comparison of anomaly detection techniques applied to different problems in the telecom industry

Nowadays, with the growth of digital transformation in companies, a huge amount of data is generated every second as a result of various processes. Often this data contains important information which, when properly analyzed, can help a company gain a competitive advantage. One data processing task common to many different applications is detection of anomalies, that is, data points or groups of data points that stand out from most of the others. Since it is not feasible to have an operator constantly analyzing the data to find anomalous values, due to the generally large volumes of data, the focus of this dissertation is the exploration of a Data Mining area called anomaly detection. In this dissertation we first develop an anomaly detection software in Python, that applies 10 different anomaly detection algorithms, after automatically optimizing their parameters, to an arbitrary dataset. Before applying these algorithms, the software also performs the task of data scaling and imputation of missing values. It outputs the results of the performance metrics of each algorithm, the values of the optimized parameters and the graphics for the results visualization generated using the method t-SNE. This software was then applied to three case studies to compare the performance of different anomaly detection approaches using real-world datasets. These datasets have an increasing level of difficulty associated with them: the amount of missing data and the uncertainty associated with the ground truth regarding the anomalies. In the first case study, we detected fraudulent bank transactions using a public dataset. Then, in the second case we identified clients of a telecommunication company who were likely to miss their payment, leading to contract termination. For this case we used a dataset from a telecommunications company. In the third case, we detected low quality of internet service, again using a large dataset with real measurements from a telecommunications company. Finally, we implemented a state of the art, neural network model, specially applicable to the task of identifying anomalies in time-series data. We optimized the parameters of the network, and applied it to address the problem of low quality of service.