haute école neuchâtel berne jura arc ingénierie www.he-arc.ch

Hes·so Haute Ecole Spécialisée de Suisse occidentale

Data Quality Control

Jonas FREIBURGHAUS

Bachelor's thesis 2020

Computer Science – Software Engineering

Professors: Emmanuel DE SALIS, Hatem GHORBEL

Expert: Boris MATTI

Description

This research project was proposed by the company SIX. They wanted to explore how Machine Learning models could monitor their data and improve their quality.

SIX provides different applications compiling markets' information and is used by financial experts for their respective businesses. Before presenting this information, the data are first aggregated from multiple sources and transformed through defined operations. Consequently, some errors may appear either because wrong information is provided by a source or because of the transformation pipeline. This work aims to build an anomaly detection system using Machine Learning, allowing us to identify and prevent incorrect information from being presented to the end-user.

The objectives are to detect data anomalies and unadjusted prices resulting from a corporate action.

Procedure

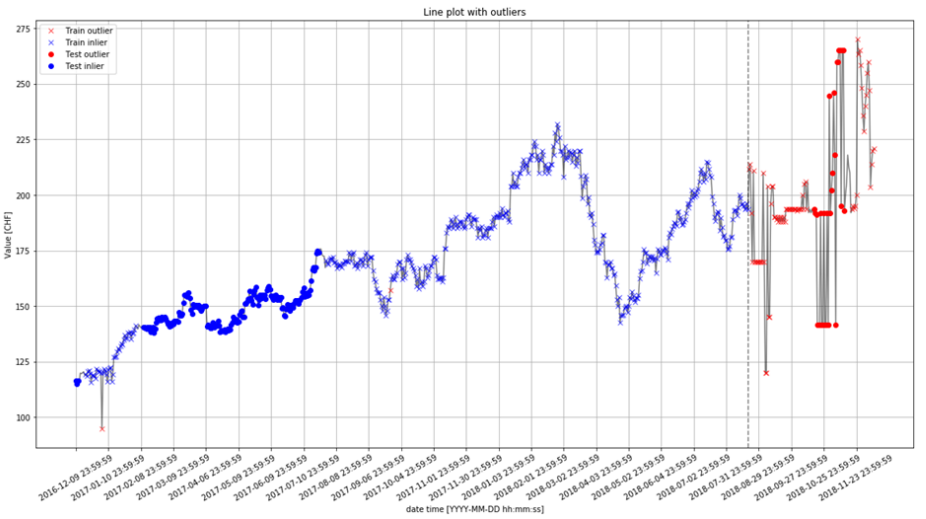
The project is conducted by applying agile development methodology, which consists of executing multiple iterations of similar steps. At the end of each iteration, my thesis supervisors provide feedback to continuously improve the work or to decide that a task has been finalized.

The iterations are summarized by the following steps:

- building of the dataset
- data analysis
- feature engineering
- feature selection
- preprocessing
- modeling
- evaluation
- search for improvements

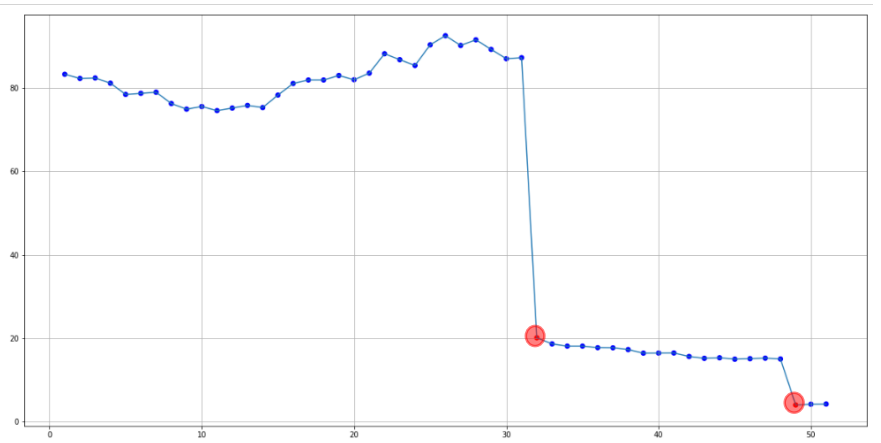
Results

For the first objective, the final solution is a stacking model. The ensemble is composed of binary classifiers. The final estimator uses the predictions made by the ensemble as features to finally classify data points considered as either normal or abnormal. The predicted outliers are the red points.



Anomaly detection on a stock price

The second objective constrained to detecting unadjusted historical prices after a split was achieved with an accuracy of 1.0 on a hand-crafted dataset.



Split detection on prices generated by a Brownian process

Perspectives

The first objective of detecting anomalies on stock prices is achieved with a high accuracy on the test set. This result is to be taken with caution as the training and testing sets are based on the same tickers at different moment in time. The detection of unadjusted prices after a split is working well on simulated prices. The next step would to be to make sure that the models generalize on other financial instruments through a labeled dataset and finally apply them to a stream of prices.