Bank of England

Technical Assessment - Data Scientist - 010030

Task II - ML Insights

Anomaly Detection Algorithms

It became quite clear in the early stages of the data exploration process that numerous outliers were scattered throughout the provided datasets. There exists various ways to detect and deal with outliers, and the optimal method often depends on the nature of the data in question. Initial actions and intuitions will be described, followed by a deep dive into one potential method of using machine learning to for outlier detection.

A quick and efficient way of visualising unfamiliar data is by using Seaborn's pairplot function. It does not require many inputs, and although it can be computationally expensive for datasets with many features, the ability to quickly scan dozens of plots for potential correlations can be worth it. A potential avenue for breaking down this particular dataset before feeding it to the pairplot function is to separate absolute values (in £) and ratios (dimensionless). This reduces the number of plots to generate by half.

When the pairplots are generated, the histograms located along the diagonal can be used to visualise individual distributions, and the scatter plots on either side to identify correlated features. In the below example, the GWP vs NWP relationship is focused on, with various anomaly detection methods tested.

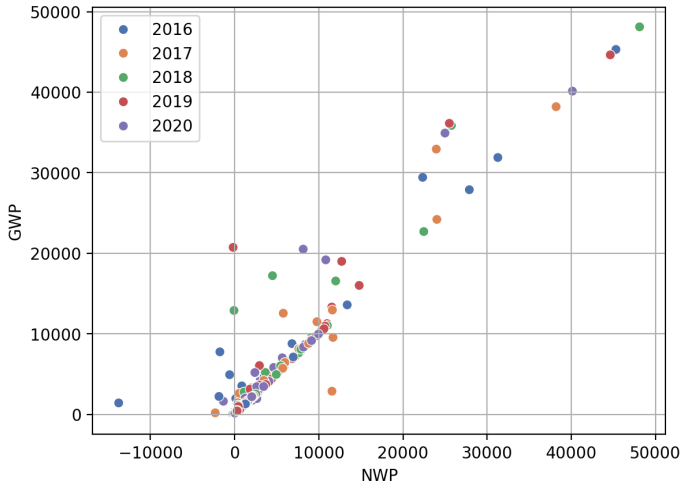

Figure 1: Gross Written Premium vs Net Written Premium

From a first glance at Fig. 1, one can assume that linear regression seems appropriate due to the fact that we are trying to predict a quantity, the data seems linear, and the underlying ML model is simple and efficient. Once a line is fit to the data, a distribution of residuals can be obtained, representing the distance between the predicted y-value

and the actual y-value. The 1.5-IQR rule can then be applied to this distribution to identify the outliers. However, due to the nature of the data, the base factor of 1.5 was not deemed adequate to sufficiently separate the inliers from outliers, thus a range of factors was tested (see Fig. 2).

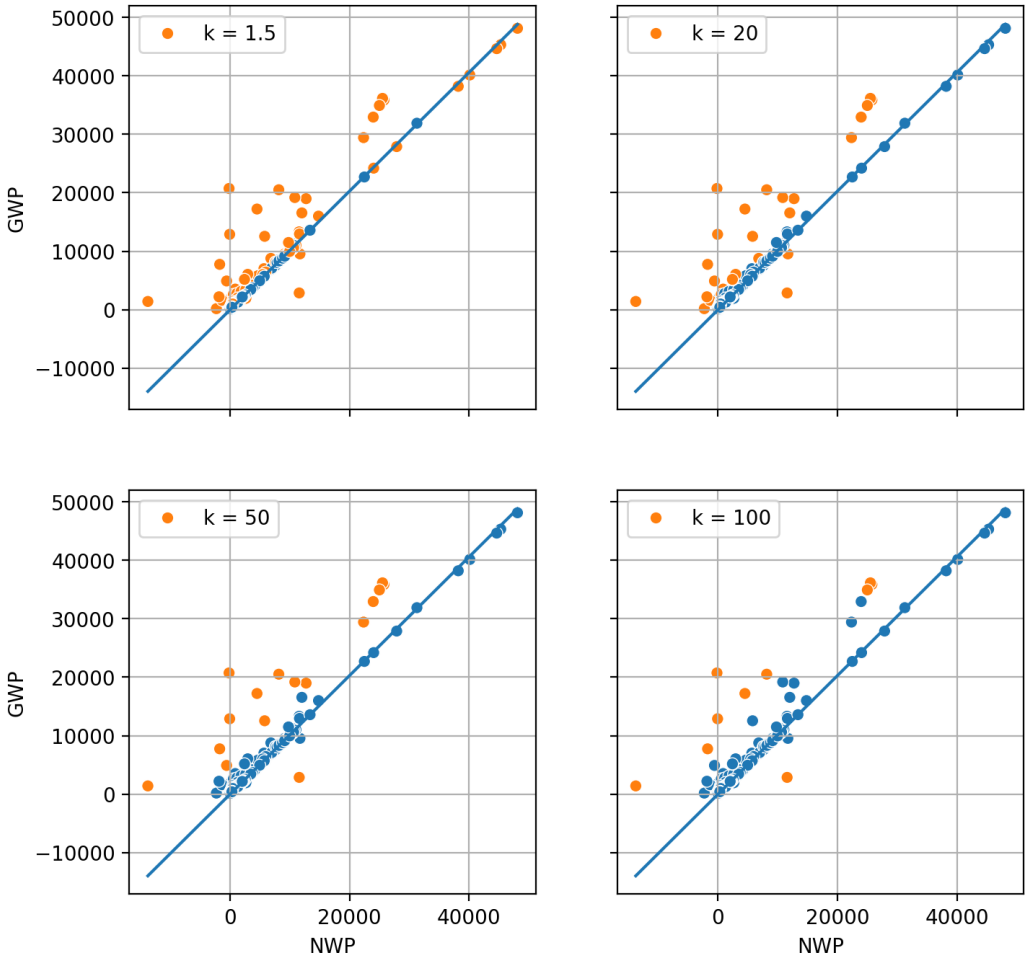

Figure 2: Outliers detected using a range of IQR factors

A potential avenue for further exploration starting from Fig. 2 is to identify whether there are any trends between the outliers, especially in the k = 100 case. Fig. 3 groups the residuals by firm, and takes an average across the available years. The adjusted IQR method is then once again applied.

Using k=40 appears to simulate an adequate contamination rate, however further work should be carried out to incorporate sample sizes for each firm into the above assessment (Firm 1, for example, only has data from one year).

Performance Comparison

Quantile Regression can often provide a good counterpoint to the more popular Linear Regression model. Quantile Regression seeks to minimise the mean absolute error (MAE), as opposed to Linear Regression which minimises the mean squared error (MSE). The result is that outliers have an outsized impact on the fit provided by Linear Regression, as the square of their errors is such that it "pulls" the best line fit more than in

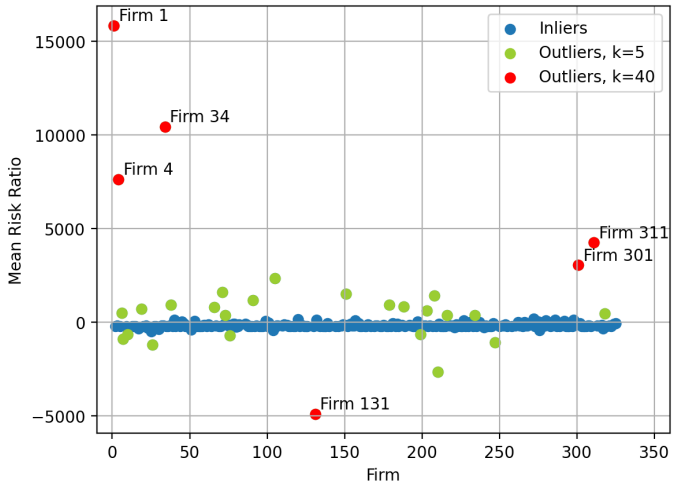

Figure 3: Outlier Firms by mean risk ratio

Quantile Regression. This is evidenced by the discrepency in fits between the algorithms in Fig. 4. The fundamental difference in each method's cost function makes it difficult to quantitatively compare the two.

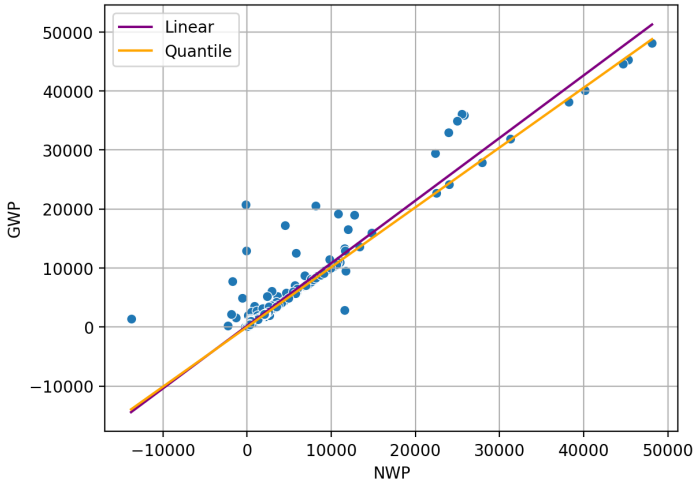

Figure 4: Linear Regression vs Quantile Regression

Road to Production

Automating anomaly detection in the context of continued analysis requires a combination of advanced techniques and systematic processes. Once the baseline for normal behavior within the dataset is validated, these algorithms can be integrated with streaming processing frameworks such as Apache Kafka or Spark Streaming to enable real-time monitoring and rapid response to anomalies. Threshold-based detection could be a suitable starting point for flagging anomalies, although more advanced unsupervised and ensemble learning algorithms should be tested as well. Regular retraining of models with new data and continuous evaluation will help ensure the model is robust to evolving patterns in real-world data and maintain the effectiveness of the anomaly detection system over time.