Machine learning for classification with big data in price statistics production pipelines

The UK Office for National Statistics (ONS) are developing a prototype pipeline for automated production of price indices from big data sources such as web-scraped and scanner data. This talk will cover solutions for two challenges in using machine learning to classify big data sources in automated pipelines. Firstly, it is impossible to manually group the hundreds of thousands of unique price quotes from these sources into the defined COICOP classification hierarchy every month. Therefore, we must develop automated processes to perform this task without adversely affecting quality in our price statistics. The second problem is how to measure classification quality in the context of a price index.

To solve the labelling problem, we have developed a process incorporating natural language processing techniques such as fuzzy string matching and sophisticated word vectorisation techniques to create features. These are used with a semi-supervised label spreading algorithm to populate class labels across a large dataset based upon an initial, much smaller set of manual labels. We can use this to train and assess a range of classifiers. These range from basic rules-based approaches to Decision Trees, Logistic and Random Forest classifiers. Initial findings using this innovative approach means that we can classify clothing datasets of tens of thousands of unique products from over a dozen COICOP categories with over 90% accuracy, using only a text description and just tens of labelled products for each category.

We will discuss appropriate metrics with weightings and visualisations to build, test and tune classifiers in production pipelines. These depend on statistical properties of the data, such as distribution and category prevalence and what the impact of making false positive (type I) and false negative (type II) errors is on the index. We will also discuss the potential for constructing classification cost functions, to assess the impact of misclassifying individual price quotes on the index so we can assess the impact of different methods on the output as well as monitoring performance over time, crucial for accurate price statistics.