Machine Learning Classifier Accepted Criteria: application to price statistics

Daniel Ma, Serge Goussev

March 2021

Abstract

As part of Statistics Canada's initiative to modernize price indices such as the Canadian Consumer Price Index (CPI), traditional price data collection in the field is being augmented with alternative sources, such as scanner, Application Programming Interface, or web scraped data, to enhance quality and timeliness and lessen collection costs. The utilization of these data sources requires supervised machine learning methods to accurately classify products to applicable categories prior to aggregation through the applicable taxonomy. In a production setting, the utilization of a highly accurate classifier also minimizes the human effort required by National Statistical Institute (NSI) officers to quality assure the classified data prior price index publication. A framework tailored to prices statistics is therefore required in order to evaluate, monitor, and select optimal models for production purposes, an under-researched topic within the literature.

This paper proposes a series of precisely defined evaluation criteria organized as a systematic framework for classification model selection and monitoring, focusing on the needs of the Canadian CPI. Rigorous scoring criteria utilized for traditional flat classification methods are combined with novel hierarchical metrics, applicable to NSI adoption of taxonomies utilized in the calculation of price statistics. Furthermore, the framework includes an assessment of dependencies of classification, specifically on the type of price index and data that will be utilized. Specifically, the evaluation framework pays close attention to the distortions of bilateral or multilateral index methods from misclassification, as well as the presence or absence of weight data, identifying the trade-offs faced.

The research combines applicable methods and metrics into a holistic framework, allowing it to be utilized to weigh trade-offs faced in utilizing a classifier to calculate price statistics. The proposed framework is evaluated using a publicly available dataset applicable to calculation of price statistics and to support NSI replicability on their own data.