Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
Paper No. 117
DOI: 10.11159/icsta25.117

Enhancing data collection through predictive modelling in Labour Force Surveys

Jeremy Heng
Ministry of Manpower
Singapore
jeremy_heng@mom.gov.sg

Extended Abstract

Labour Force Surveys (LFS) are commonly used by many countries around the world to compile official labour statistics. The LFS in Singapore is conducted monthly over a period of six months for each household. In the survey, each member of the household will have to provide information on whether they are employed, unemployed or outside the labour force. This information is used to compile labour statistics such as employment rate, unemployment rate and labour force participation rate. Based on past trends and historical data, most respondents do not change their labour force status on a month-to-month basis.

To reduce survey burden on respondents and to save time and resources, the Singapore Ministry of Manpower seeks to develop a predictive model to predict the labour force status of individuals. Since the LFS have been conducted for many years, there is a wealth of survey data that can be tapped on. A machine learning model called CatBoost is used to determine which individual has a change in labour force status from the previous month. Before training the model, the first step of data preprocessing is to define and group the relevant variables: employed, unemployed and outside the labour force. "Status changed" and "month" variables are also created and these categorial variables are converted to numerical variables with min-max scaling.

To train the model, we first use Stratified Shuffle Split to split 80% of the dataset into the train set and 20% into the test set. Without a test set, the model might perform very well on the training data but poorly on new data. Hence, by splitting the dataset, we can evaluate the model on the test set to check its generalization ability. Next, sample weights are used whereby higher weights are assigned to samples from the minority class to ensure that they are not underrepresented. Bayesian Optimization is then used to find the optimal parameters for the model. After training the model with these parameters, we obtain the feature importance score of each feature used, which quantifies how much each input feature contributes to the model's predictions.

A few metrics such as accuracy, precision and recall are used to measure the performance of the model. The model is found to have an accuracy of 97%. Another machine learning model, LightGBM, is also trained as an alternative for comparison. While both models boast equally high accuracy, the CatBoost model is chosen as its F1-score and PR-AUC score is higher. F1 and PR-AUC scores can be used to evaluate the balance between competing precision and recall scores of the model, which implies that CatBoost achieves a better balance between false positives and false negatives.

With the predictive model, we are able to identify survey respondents who do not change their labour force status based on their demographic profile and initial survey response. For example, retired individuals are likely to stay retired and employees who embarked on a new job are likely to stay in their job for the next 6 months. This reduces the number of survey touchpoints for each round of the LFS, and the time spent on each survey. Over time, we seek to expand the model to cover other data items of the survey, thereby improving cost efficiency and reducing respondent burden.