Imputation of Industry and Occupation Categories

BLS Imputation Group

Marin Lolic, Casey Nguyen, Ahmed Soliman

Project Goal

The goal of this project is to provide the U.S. Bureau of Labor Statistics (BLS) with completed data for industry and occupation across different decades (1970s to present)

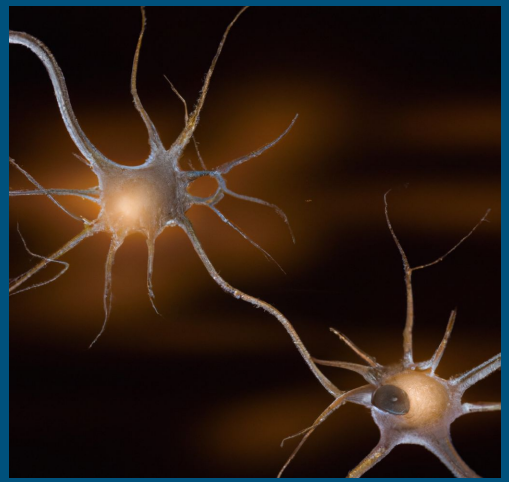
There are two primary resources:

- Dual-coded files, which have a very large number of samples, but are only available for a few of the time periods
- Crosswalks, which can collectively cover all the time periods, but have far fewer samples

Project History

Prior team created 240 different random forest models to predict every combination of industry and occupation categories, mostly utilizing two “universal crosswalks” that they stitched together

Our idea was to greatly simplify this processs, creating two neural network-based models (one for industry, one for occupation), hopefully improving accuracy in the process

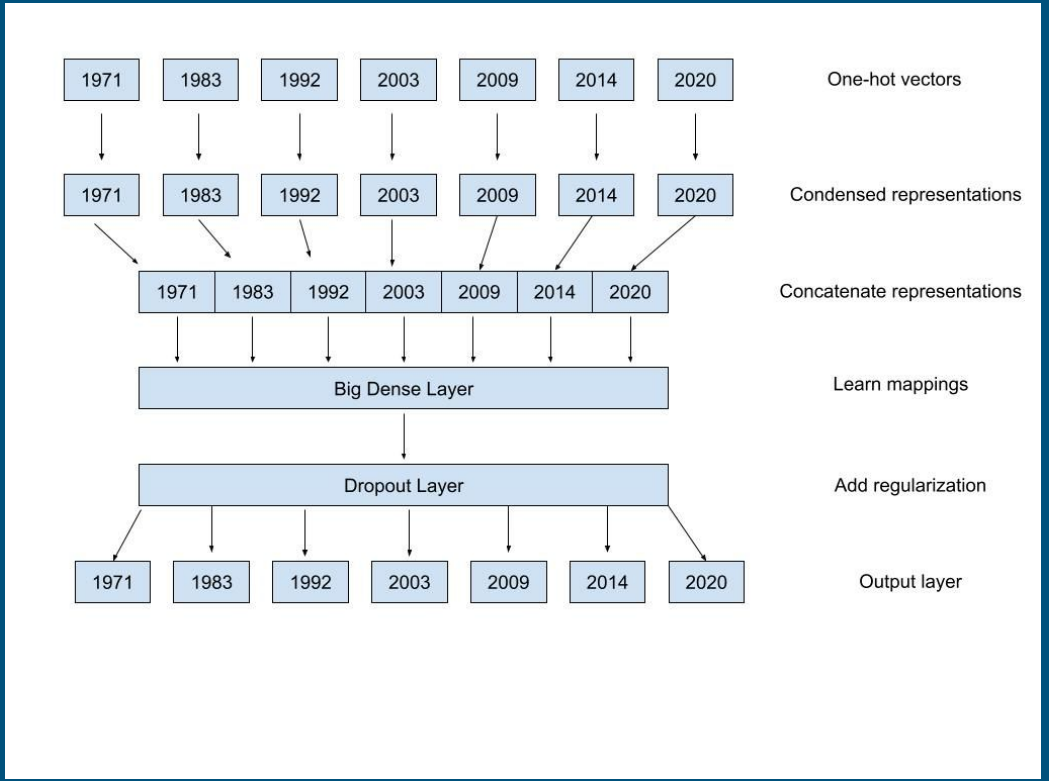


Credit: Dall-E

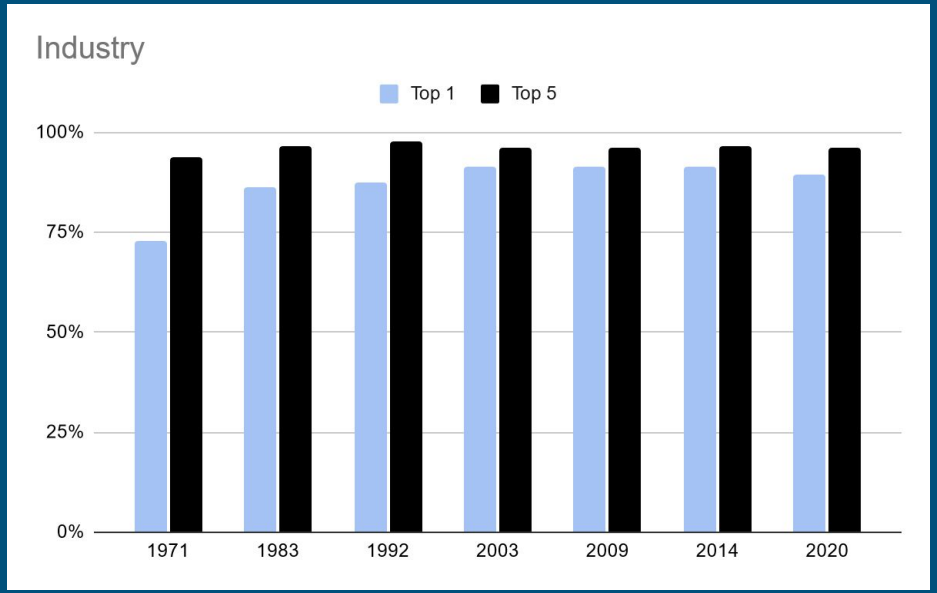
Model Construction Process

Industry Model Example

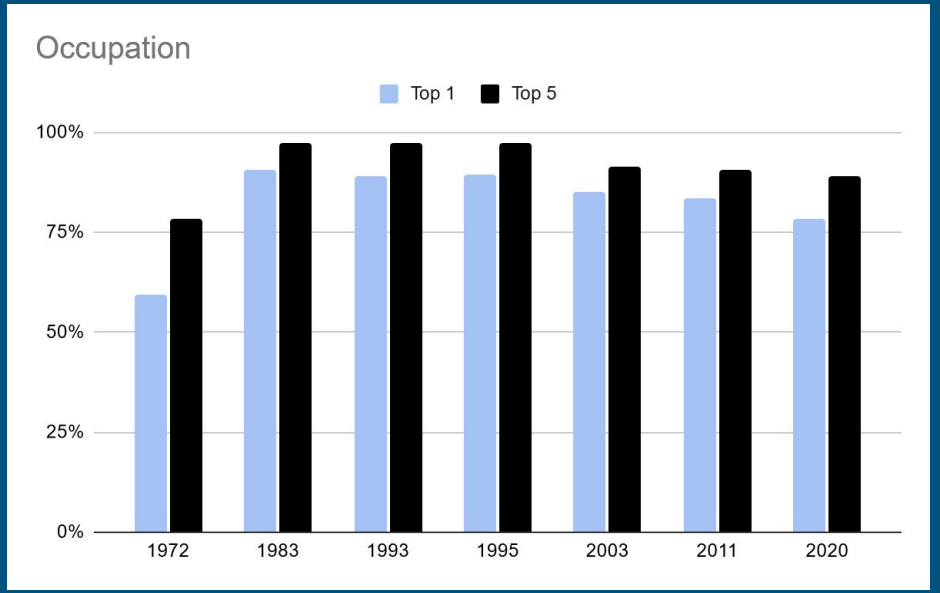
Same Process for Occupations



Holdout Data Accuracy



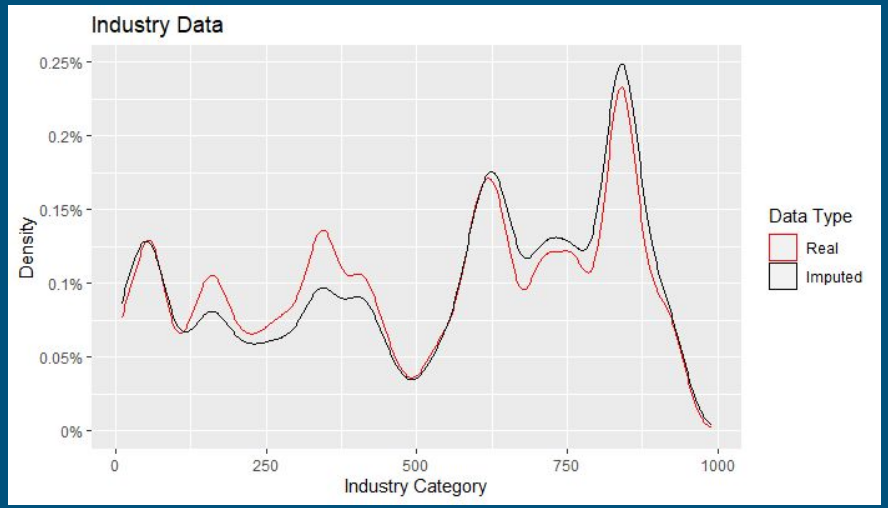
Categories per Period ~ 250

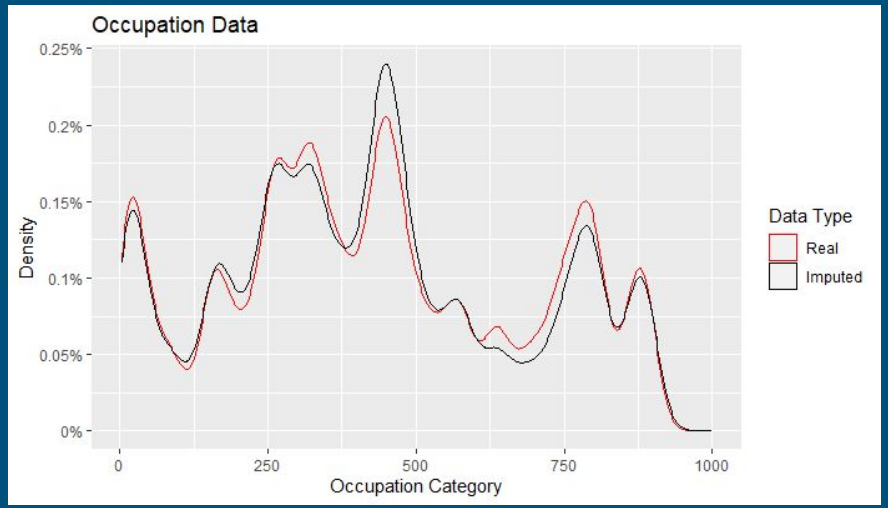


Categories per Period ~ 500

Improving The Project

- Our sponsor had an additional request - to see how our methodology would work using a dual-coded data set
- This provided both industry and occupation information together, and with a very large sample size
- However, only available for two periods (1970/1980)
- Results below show close correspondence between real data for 1980 and our imputed data for 1980





Conclusions

- Neural networks can be used for complex imputation problems
- Very high accuracy, simpler to use, and far fewer models needed for this method than for alternative methods
- More dual-coded data (and from different periods) make the results even better