Predicting Missing Levels Using Supervised Machine Learning

David H. Oh
Economist
Office of Compensation and Working Conditions

FedCASIC
April 17, 2019

Objective

- The National Compensation Survey (NCS) evaluates each sampled occupation based on a set of factors and determines a “level” of work using the point factor system
- Item nonresponse for the levels has been increasing, but there is no process currently in place to fill in the missing information.
- What is the best approach to imputing these missing values?

Overview

1. **Background and missing levels**
2. Different imputation approaches
3. Summary of the results and the next steps

National Compensation Survey

- Employer-based survey program
- Approximately 11,400 establishments
- Private industry, state/local government
- Provides comprehensive measures of occupational wages, employment cost trends, and benefit incidence and detailed plan provisions
- Various worker characteristics are collected for selected occupations within each establishment

Levels

- Equivalent to the General Schedule (GS) grade levels used in the Federal sector to determine pay
- Reflect varying duties and responsibilities of an occupation
- Range from 1 to 15

Increasing Number of Missing Levels

- Item nonresponse for level information in the NCS data has been increasing by a few percentage points every year
- Consequently, a substantial amount of NCS data are not being utilized in the estimation of products that rely on the level information

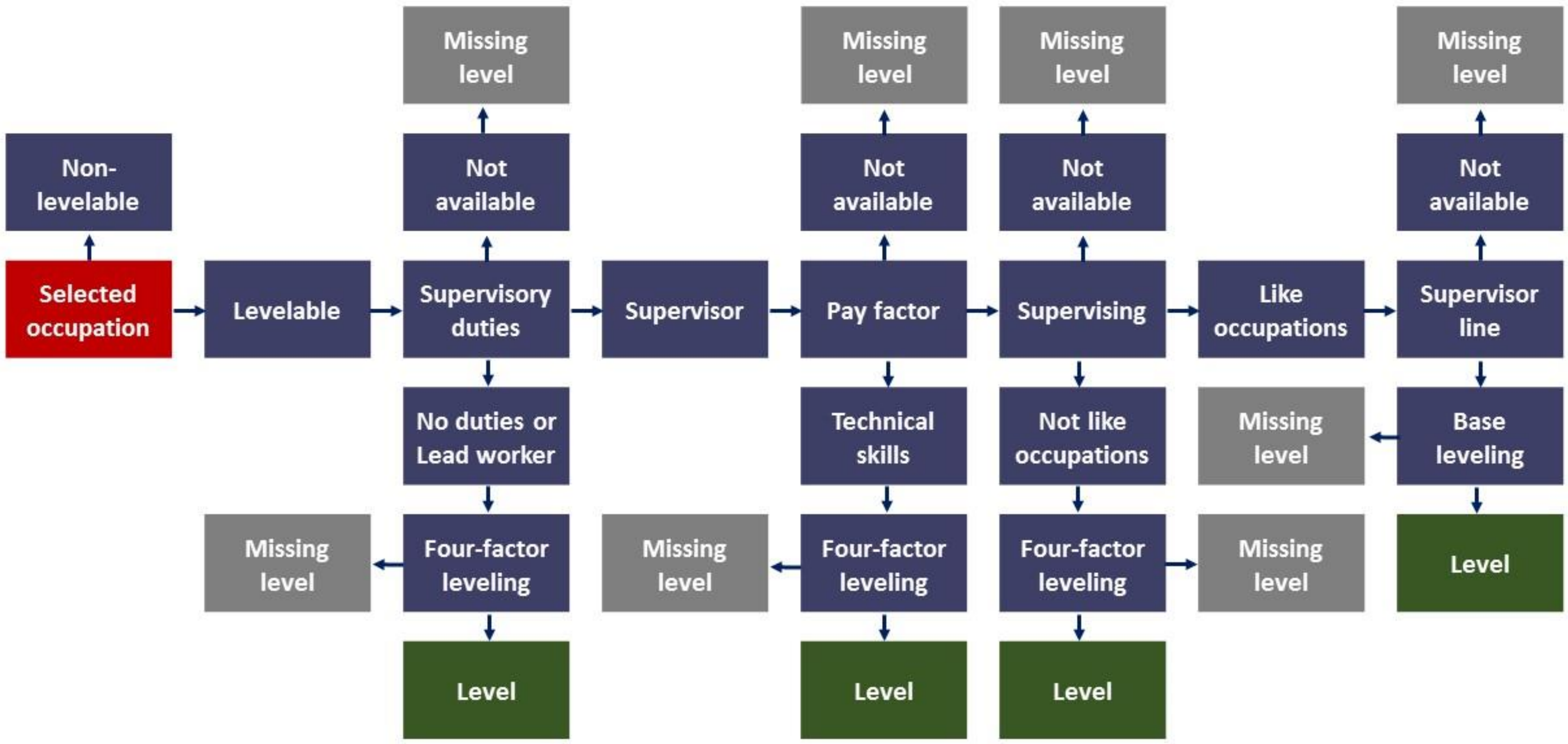
Four-Factor Leveling

- Levels are assigned using the four-factor system provided by the Office of Personnel Management for the purposes of BLS data collection

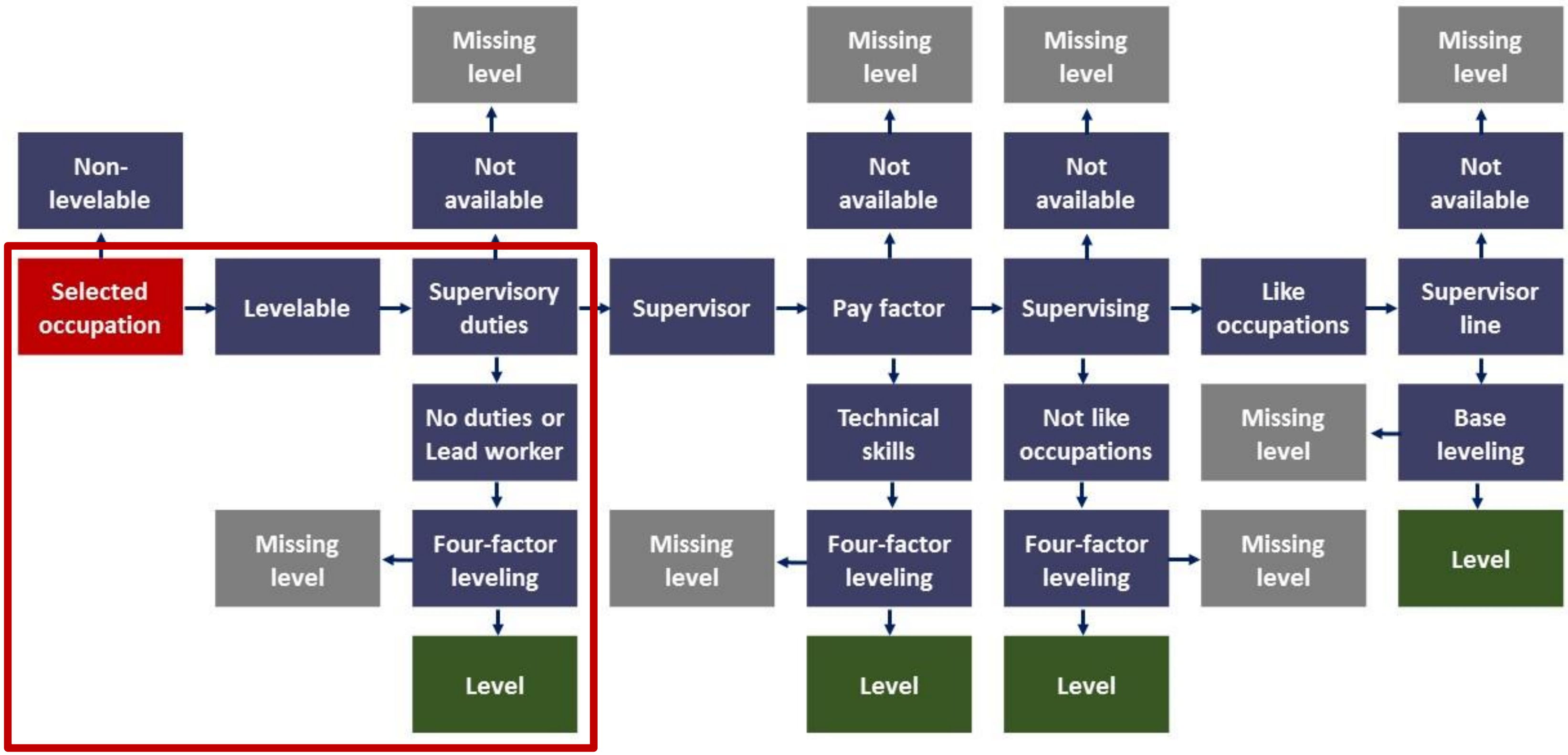
Factor	Points								
Knowledge	50	200	350	550	750	950	1250	1550	1850
Job Controls and Complexity	100	300	475	625	850	1175	1450	1950	X
Contacts	30	75	110	180	280	X	X	X	X
Physical Environment	10	25	40	70	100	X	X	X	X

Level	Min. Points	Max. Points
1	190	254
2	255	454
3	455	654
4	655	854
5	855	1104
6	1105	1354
7	1355	1604
8	1605	1854
9	1855	2104
10	2105	2354
11	2355	2754
12	2755	3154
13	3155	3604
14	3605	4054
15	4055 and up	

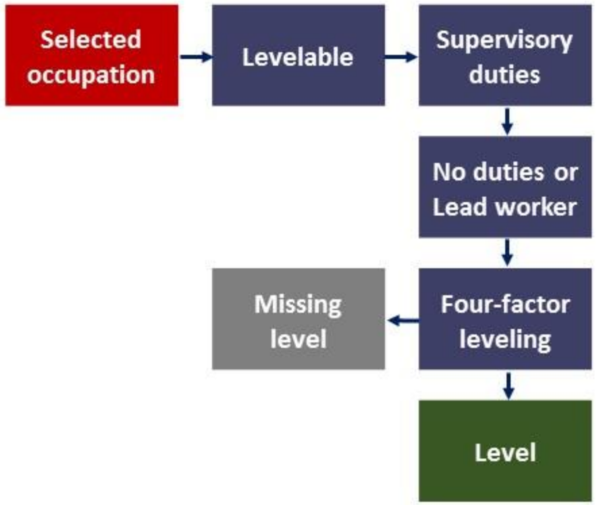
Entire Leveling Process



Major Source of Missing Levels



Scope of the Project



Overview

1. Background and missing levels
2. **Different imputation approaches**
3. Summary of the results and the next steps

Imputation

- Currently, there is no imputation process in place to fill in the missing level information
- Development strategy: Start with the most basic approach and build up
  - Naïvely impute levels
  - Directly impute levels using machine learning methods
  - Indirectly impute levels by imputing the factors first
  - Indirectly impute levels with features that vary with available information

Data

- Limit observations to those that do not have supervisory duties
- A total of 24,312 observations from March 2018 NCS data are randomly split into training (approx. 67%), validation (approx. 16%), and test datasets (approx. 16%)

Performance Measure

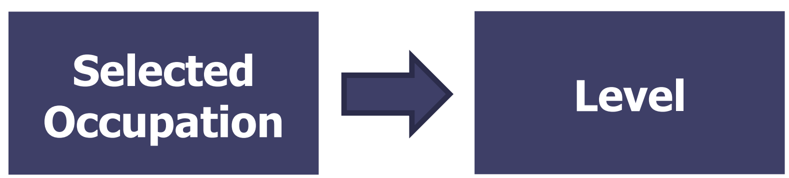
- Accuracy
  - Measures how accurate the method is in predicting the correct level

$$\frac{\text{Count of rows where predicted level} - \text{actual level} = 0}{\text{Total number of rows}}$$

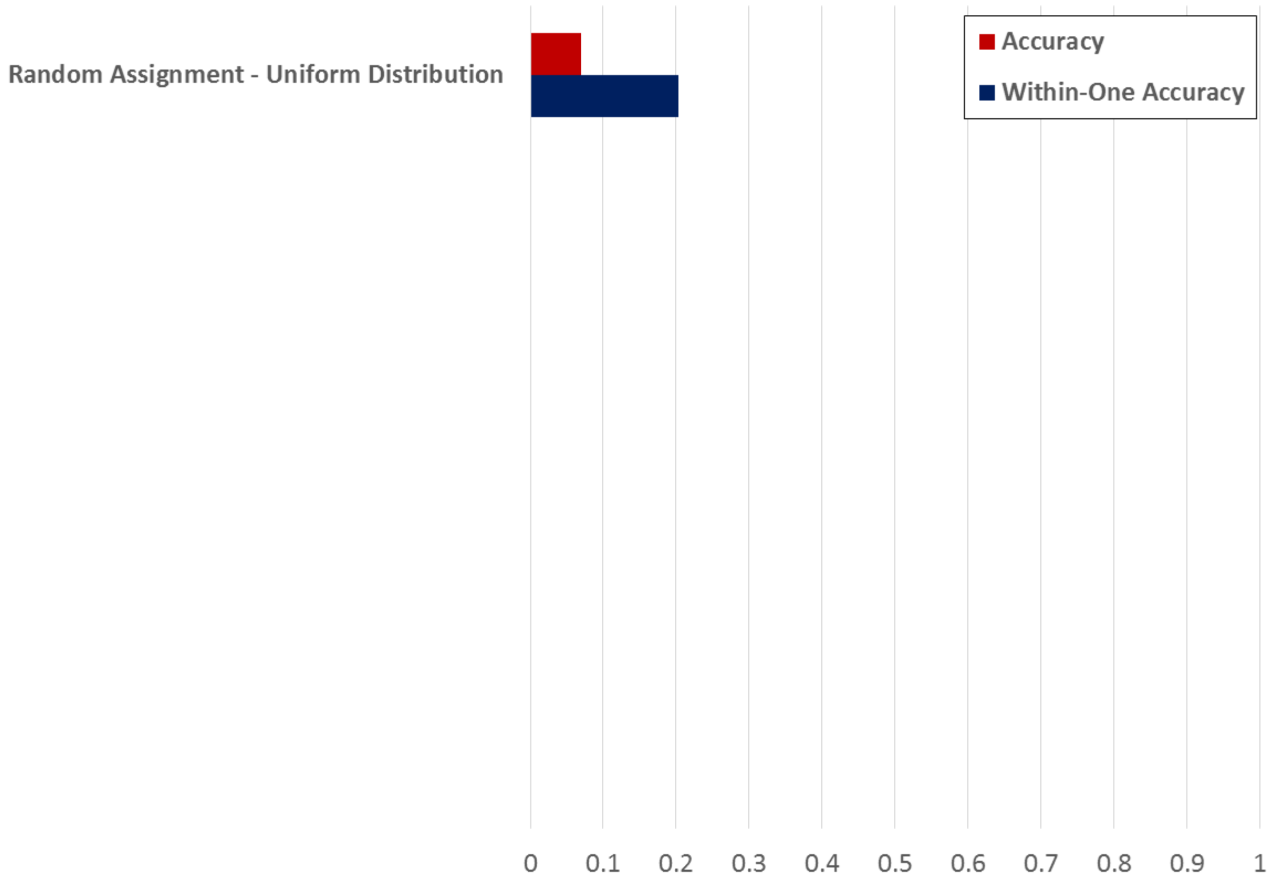
- Within-One Accuracy
  - Measures the precision of the method

$$\frac{\text{Count of rows where } |\text{predicted level} - \text{actual level}| \leq 1}{\text{Total number of rows}}$$

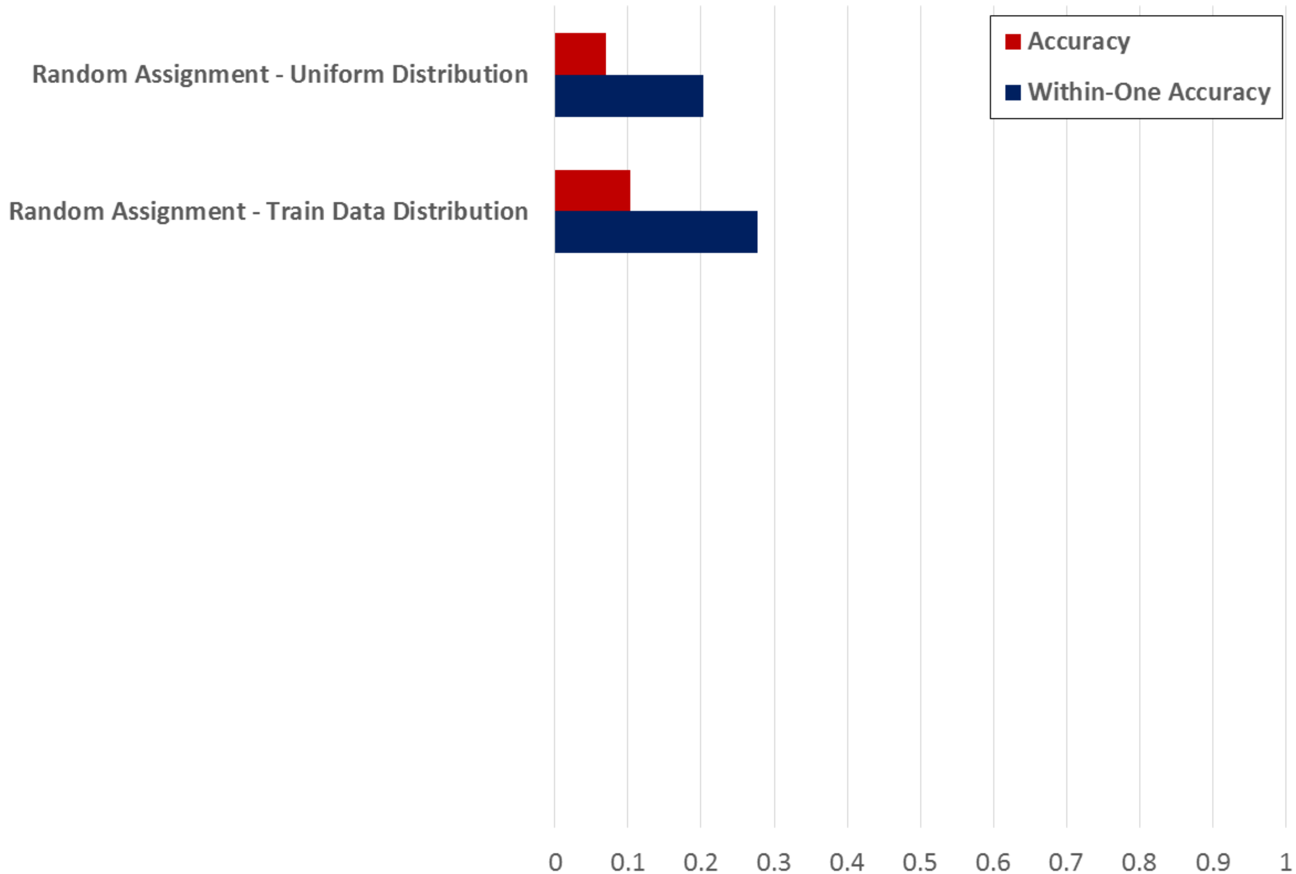
Basic Approach: Naïvely Impute Levels



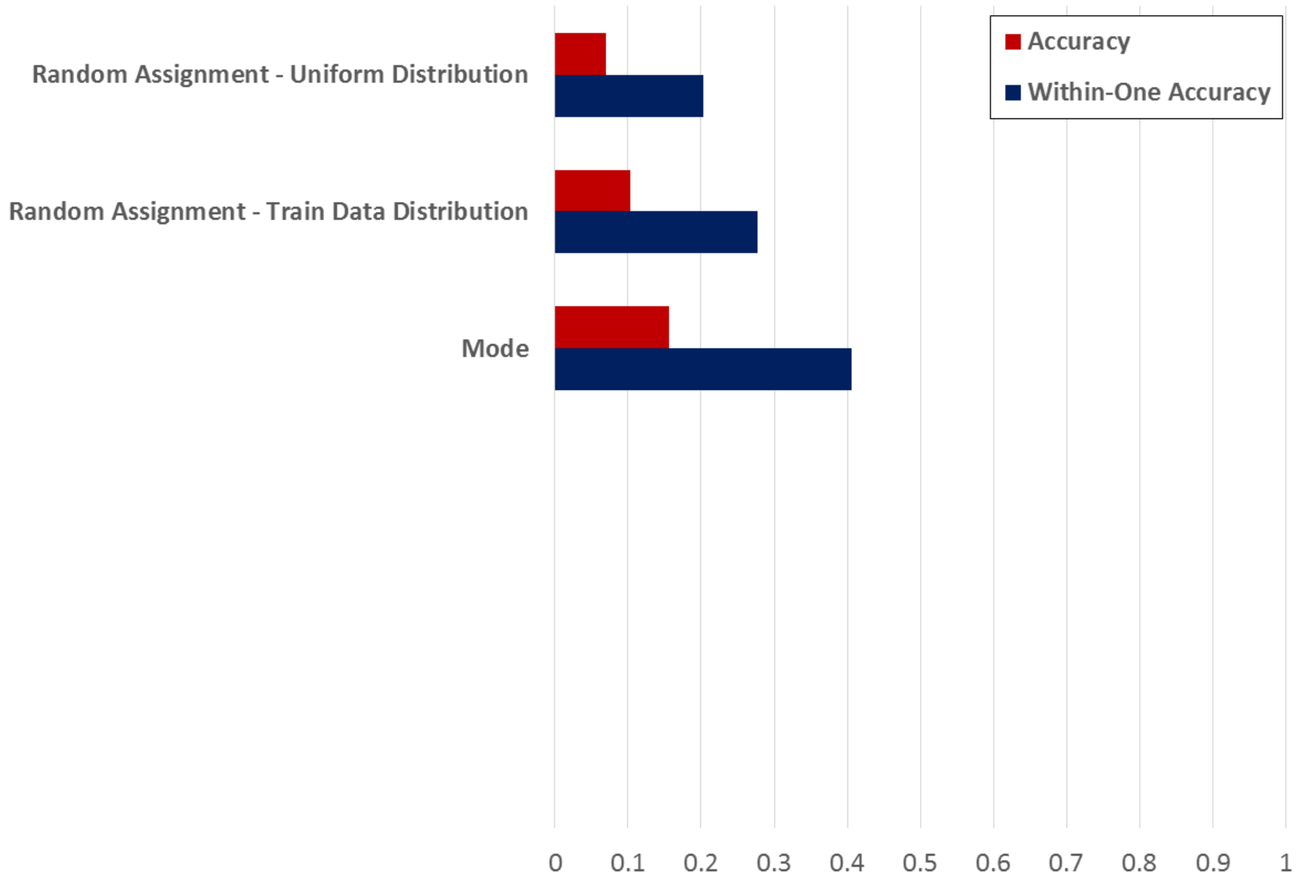
Random Draw from a Uniform Distribution



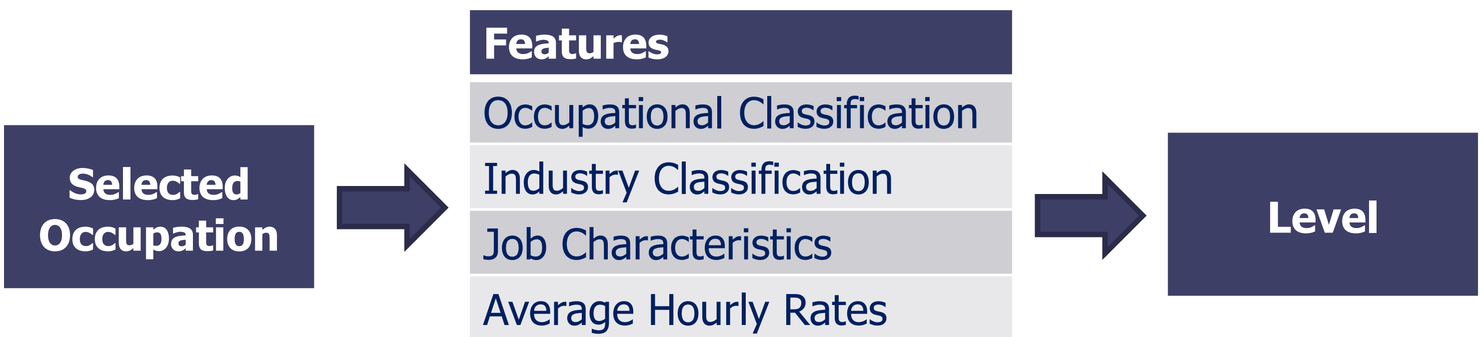
Random Draw
from the Training Data Distribution



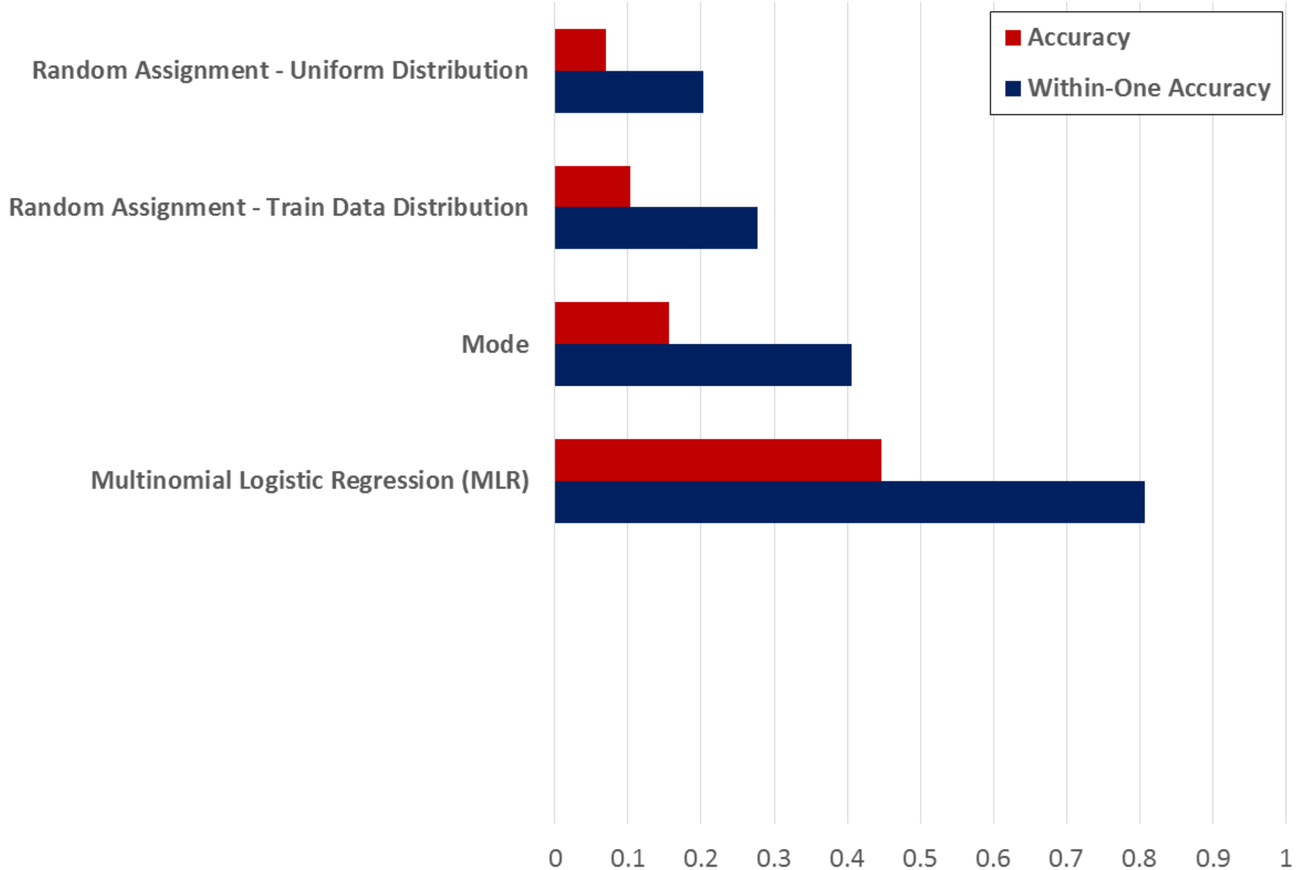
Assign the Mode from Training Data



Machine Learning Approach: Directly Impute Levels

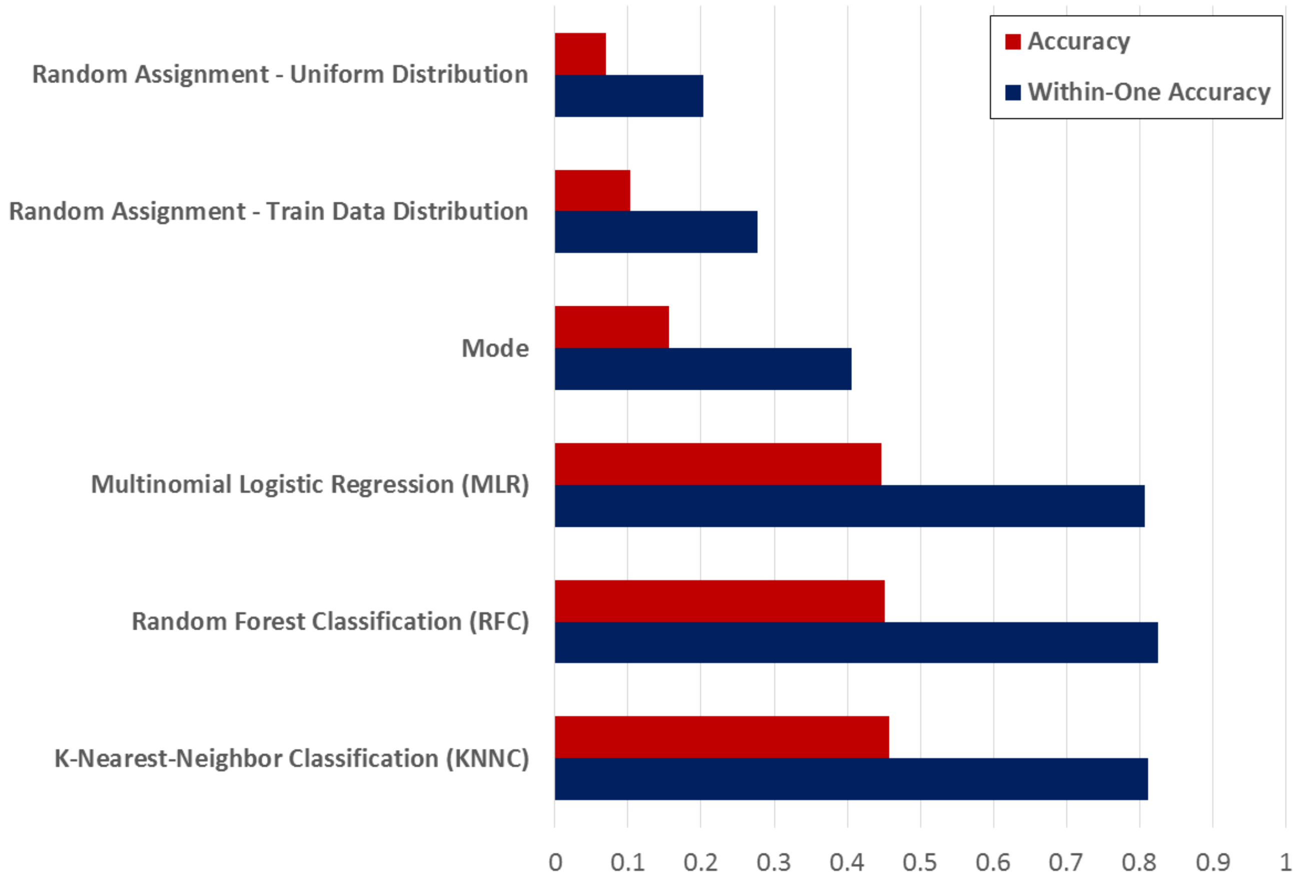


Multinomial Logistic Regression (MLR)

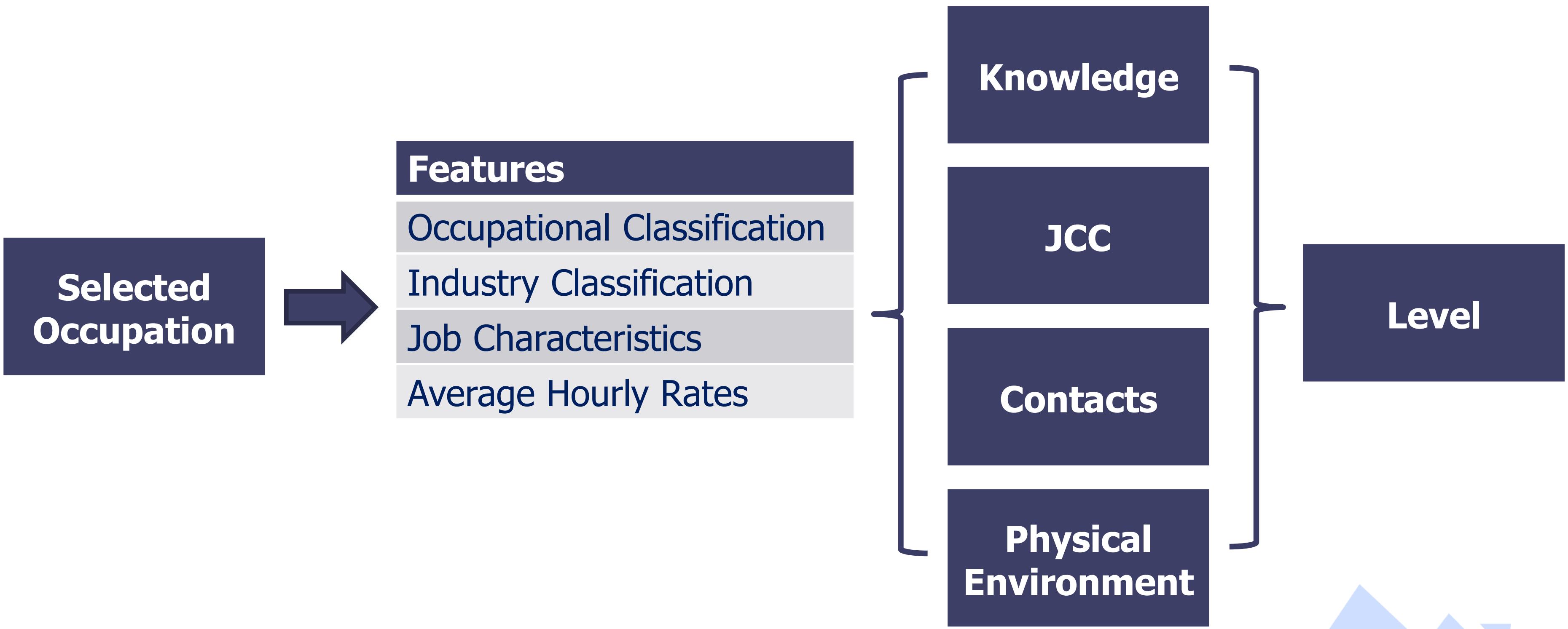


Random Forest Classification (RFC)

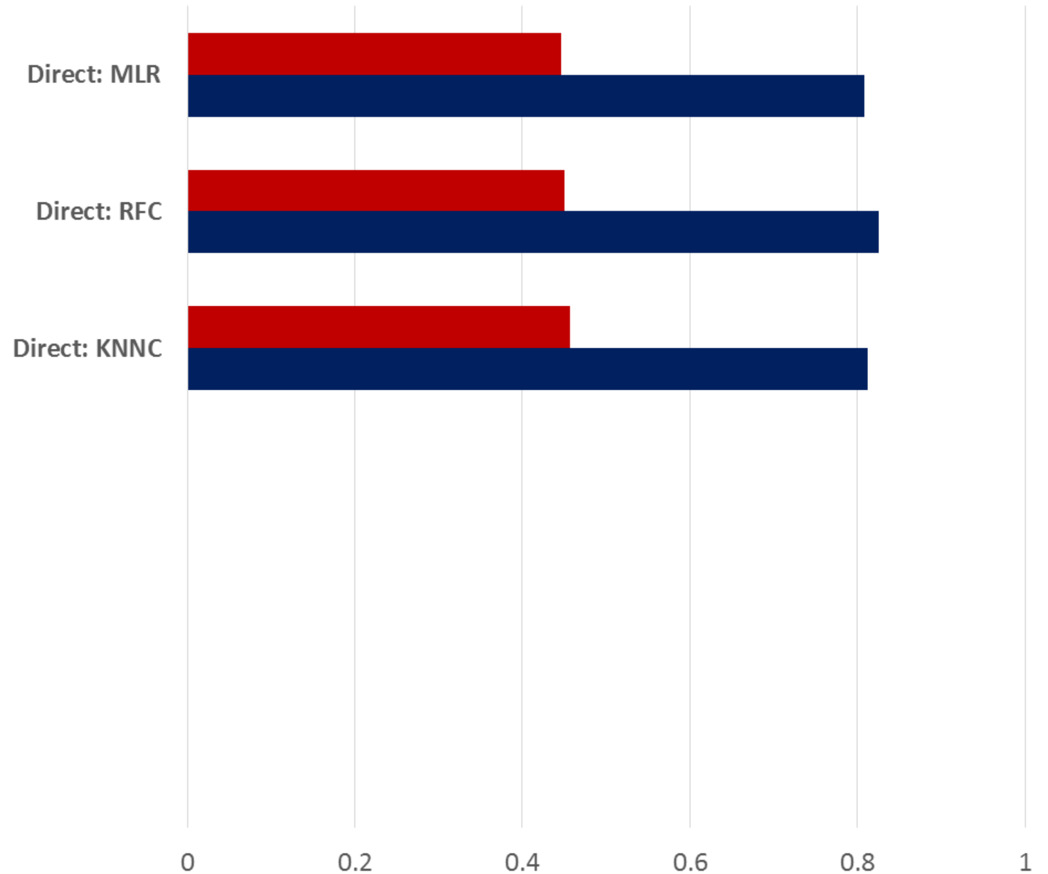
K-Nearest-Neighbor Classification (KNNC)



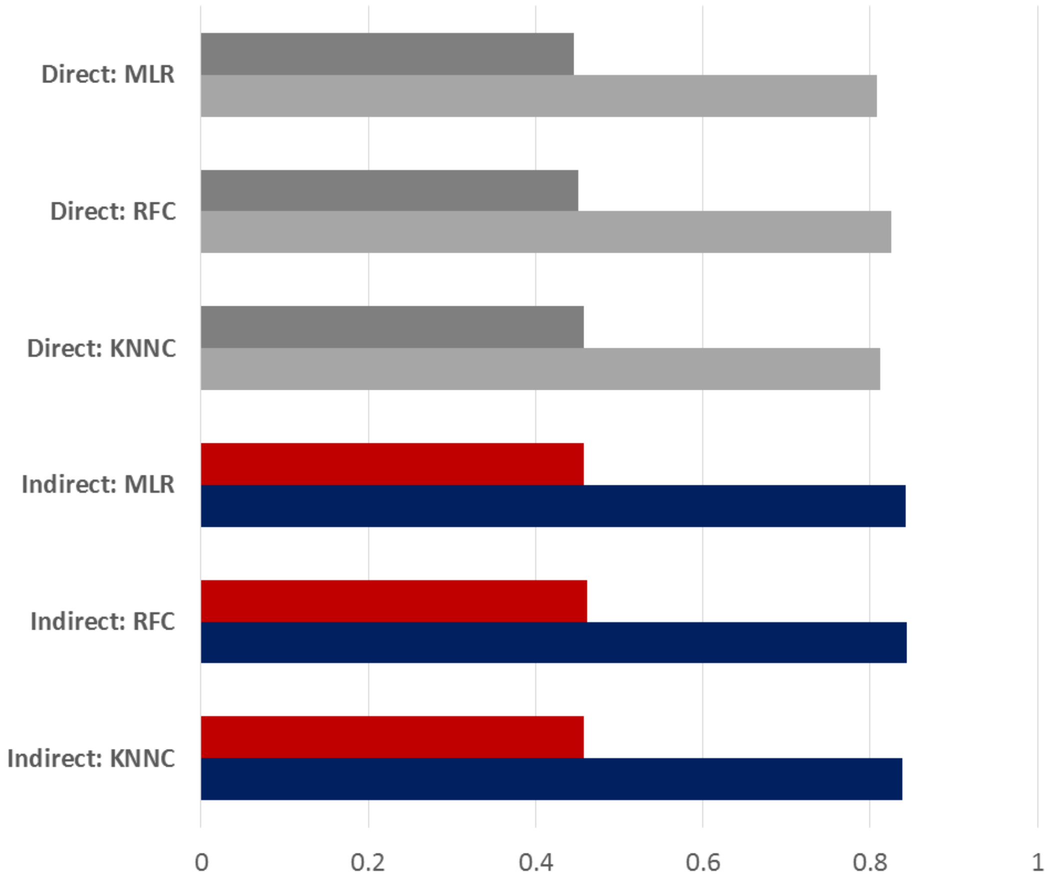
Machine Learning Approach: Indirectly Impute Levels



Directly Impute Levels

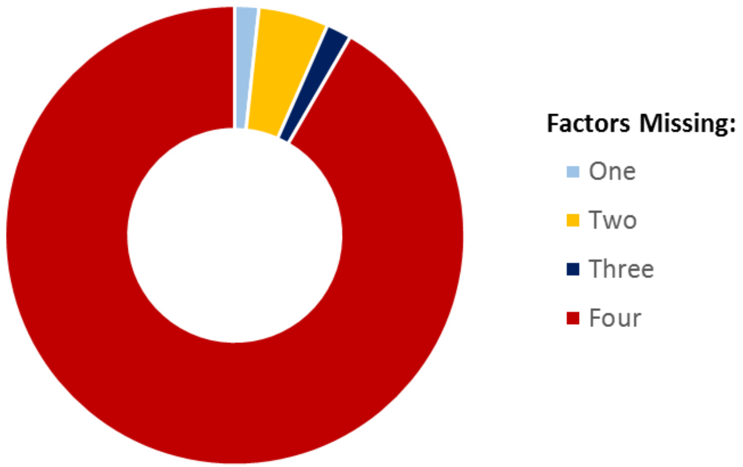


Indirectly Impute Levels



Why is there a Lack of Gain in Performance?

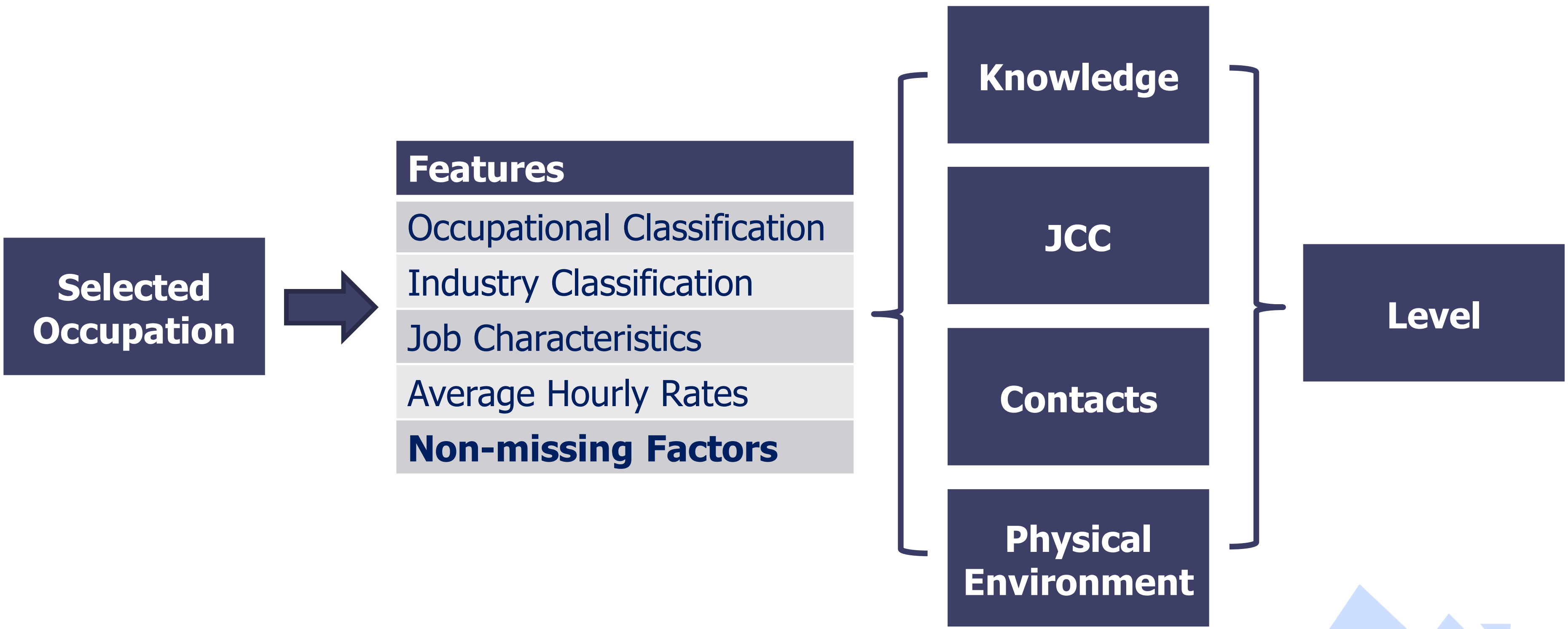
- The majority of missing levels have none of the four factors coded
- The few that are partially coded tend to be the less impactful factors (i.e., Contacts and Physical Environment)



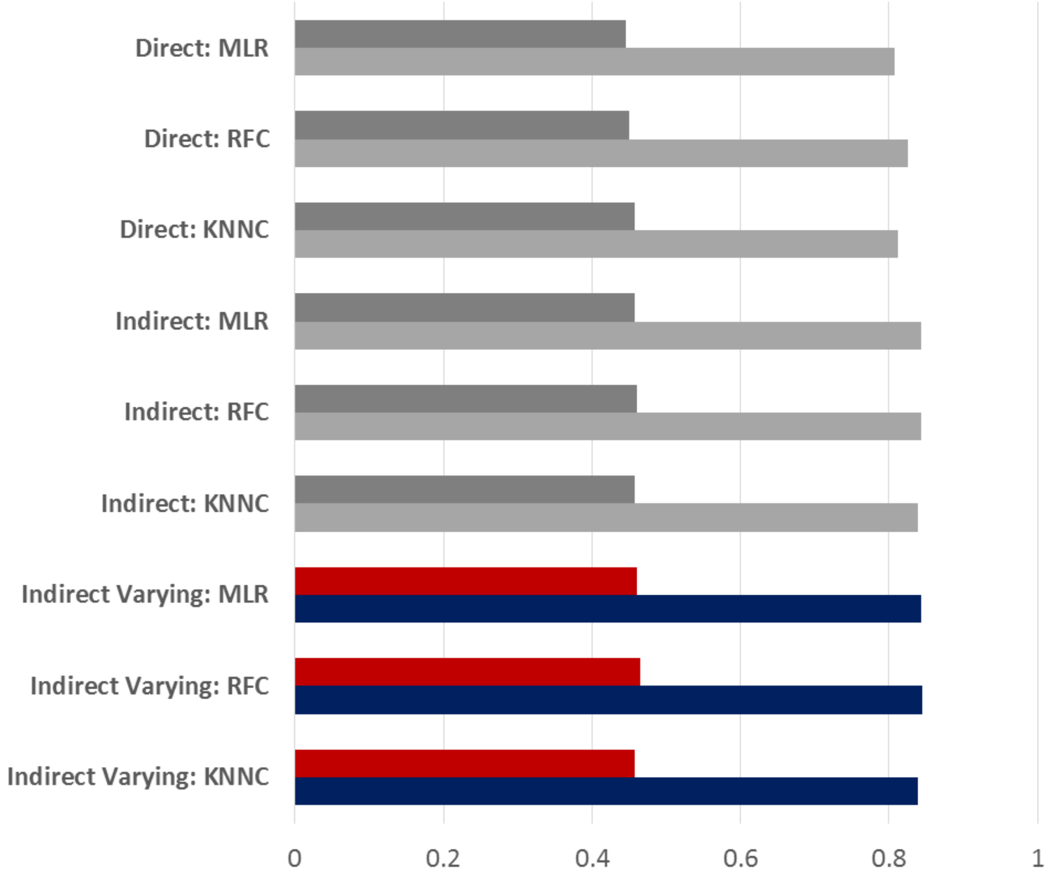
Procedural Recommendation : Coding one is better than coding none

		Percent of Occupations Missing All Four Factors Given One Factor Information				
		0%	25%	50%	75%	100%
Accuracy	Knowledge	0.46	0.50	0.53	0.58	0.62
	JCC	0.46	0.50	0.54	0.59	0.63
	Contacts	0.46	0.46	0.46	0.46	0.47
	Phy. Env.	0.46	0.46	0.46	0.46	0.46
Within-One Accuracy	Knowledge	0.84	0.87	0.90	0.93	0.96
	JCC	0.84	0.87	0.90	0.93	0.96
	Contacts	0.84	0.84	0.85	0.85	0.85
	Phy. Env.	0.84	0.84	0.84	0.84	0.84

Machine Learning Approach: Indirectly Impute Levels with Varying Features



Indirectly Impute Levels with Varying Features



Procedural Recommendation : Coding one is better than coding none

		Percent of Occupations Missing All Four Factors Given One Factor Information				
		0%	25%	50%	75%	100%
Accuracy	Knowledge	0.46	~~0.50~~ 0.52	~~0.53~~ 0.57	~~0.58~~ 0.63	~~0.62~~ 0.69
	JCC	0.46	~~0.50~~ 0.52	~~0.54~~ 0.58	~~0.59~~ 0.64	~~0.63~~ 0.69
	Contacts	0.46	0.46	~~0.46~~ 0.47	~~0.46~~ 0.47	0.47
	Phy. Env.	0.46	0.46	0.46	0.46	0.46
Within-One Accuracy	Knowledge	0.84	~~0.87~~ 0.88	~~0.90~~ 0.91	~~0.93~~ 0.95	~~0.96~~ 0.99
	JCC	0.84	~~0.87~~ 0.88	~~0.90~~ 0.91	~~0.93~~ 0.95	~~0.96~~ 0.98
	Contacts	0.84	~~0.84~~ 0.85	0.85	0.85	~~0.85~~ 0.86
	Phy. Env.	0.84	0.84	0.84	0.84	0.84

Overview

1. Background and missing levels
2. Different imputation approaches
3. **Summary of the results and the next steps**

Summary of the (Preliminary) Results

- Machine learning approach performs much better than the most basic imputation approaches
- In practice, current method correctly predicts the actual level 47 percent of the time and within plus-or-minus one of the actual level, 84 percent of the time
- Simulation shows promising performance of the current model with increases in the number of partially coded factors, especially Knowledge and JCC

Next Steps

- Machine side:
  - Introduce additional variation in features that are optimized for each factor
  - Increase the number of training data
  - Explore other models/methods
- Human side:
  - Increase the effort to collect even partial factor information, especially Knowledge and JCC

Contact Information

David H. Oh
Economist
Office of Compensation and Working Conditions
www.bls.gov/ect
202-691-5985
oh.david@bls.gov