A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model

Hyeonsup Lim, Majbah Uddin, Yuandong Liu, Shih-Miao Chin, and Ho-Ling Hwang

Oak Ridge National Laboratory (ORNL)

Introduction & Data Source

Motivation

- Traditionally, freight generation (FG) modeling approaches are based on OLS regression.
- This study proposed industry-specific freight generation models based on industry-related factors such as number of establishments.

Data Source

- This study utilized tonnage and value from the most recently released 2017 CFS data for 24 industry sectors as dependent variables.
- For the independent variables, the study utilized the two county-level industry data products by Census, i.e., Economic Census (EC) and County Business Pattern (CBP) data.

Summary Statistics of the Input Data

NAICS	*tonnage (1,000 tons)*			*value (million $)*			*number of establishments*		*number of employments*		*annual payroll (million $)*		*receipt total (million $)*	
	N	mean	std.	N	mean	std.	mean	std.	mean	std.	mean	std.	mean	std.
212	119	23,570	34,993	118	710	1,308	34	35	1,127	1,741	76	128	N/A	N/A
311	123	4,865	6,957	127	6,255	7,447	199	241	11,466	12,584	507	579	4,392	5,726
312	108	1,284	2,077	112	1,372	2,938	62	110	1,628	2,689	80	160	666	1,339
313	86	69	166	101	284	629	11	28	577	1,604	24	64	121	393
314	99	43	176	116	214	656	35	46	687	1,604	25	66	114	521
315	81	5	19	101	135	472	42	226	686	3,092	18	84	65	364
316	74	7	22	83	52	79	6	12	146	352	5	12	12	36
321	113	1,939	2,876	125	889	1,142	103	108	3,047	3,593	121	142	610	810
322	113	1,399	1,824	113	1,665	1,969	26	34	2,103	2,873	124	177	686	1,076
323	112	158	249	126	675	838	182	221	3,469	4,140	155	204	535	751
324	98	12,687	24,707	112	4,623	11,030	10	13	655	1,363	68	154	2,067	6,675
325	122	5,694	11,200	127	5,741	9,853	93	117	5,618	7,089	439	636	4,296	9,013
326	121	493	615	129	1,869	2,205	86	105	5,456	6,696	260	316	1,391	1,868
327	106	6,675	6,988	129	993	931	103	85	2,897	2,594	149	138	730	753
331	106	1,545	2,726	112	1,915	2,512	27	37	2,346	3,460	149	239	1,011	1,881
332	114	899	1,284	130	2,700	3,016	402	463	10,917	12,494	558	663	2,423	2,860
333	108	308	471	125	2,959	3,377	169	200	7,815	8,438	488	550	2,268	2,774
334	87	28	45	118	2,660	4,972	89	160	6,052	10,593	521	1,029	1,811	3,814
335	100	150	218	115	1,098	1,256	36	57	2,138	2,938	135	208	503	881
336	98	1,058	2,315	112	8,093	13,983	81	104	11,616	16,517	749	1,144	4,736	10,549
337	120	117	176	126	610	889	101	124	2,729	4,096	113	172	424	785
339	104	63	80	124	1,274	1,743	203	270	4,250	5,928	234	386	952	1,652
423	118	7,665	10,109	132	23,600	36,419	1,791	2,381	27,453	35,257	1,882	2,947	19,660	33,376
424	128	25,620	33,416	130	28,256	36,715	961	1,651	18,405	26,483	1,178	2,047	20,662	39,837

Data Processing and Model Selection

1. Imputation of missing data (for CBP/EC)

Imputed the number of employments based on employment size range (*EMPFLAG*).

2. Data Transformation

No transform vs Log-transform

3. Normalization

A simple min-max normalization

4. Variables and Hyperparameters

Find the best variable selection and hyperparameter settings for each model:

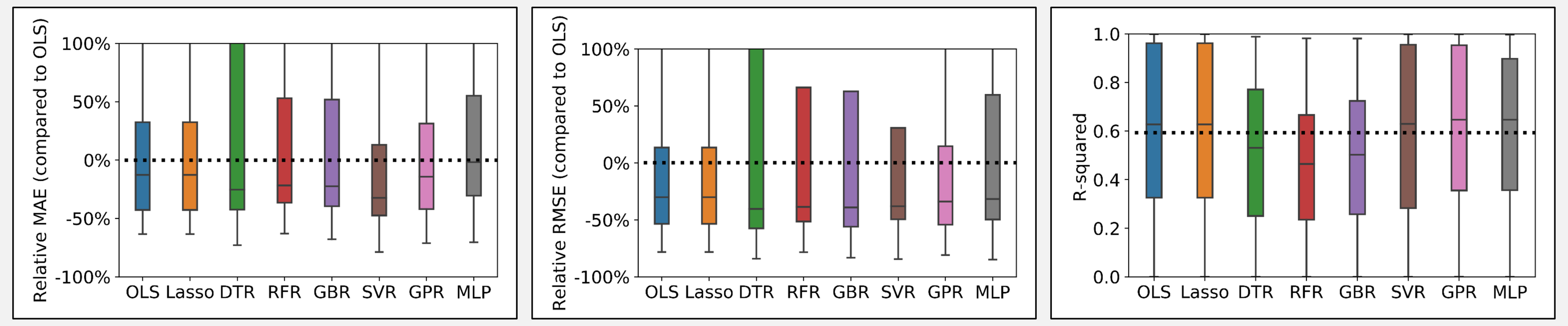
- Ordinary Least Squares (OLS, the baseline)
- Least Absolute Shrinkage and Selection Operator (Lasso)
- Decision Tree Regression (DTR)
- Random Forest Regression (RFR)
- Gradient Boosting Regression (GBR)
- Support Vector Regression (SVR)
- Gaussian Process Regression (GPR)
- Multi-layer Perceptron Regression (MLP)

5. Model Performance and Final Model Selection

Since the MAEs and RMSEs may not be directly comparable across different NAICS, the relative differences of MAE and RMSE were compared to the OLS.

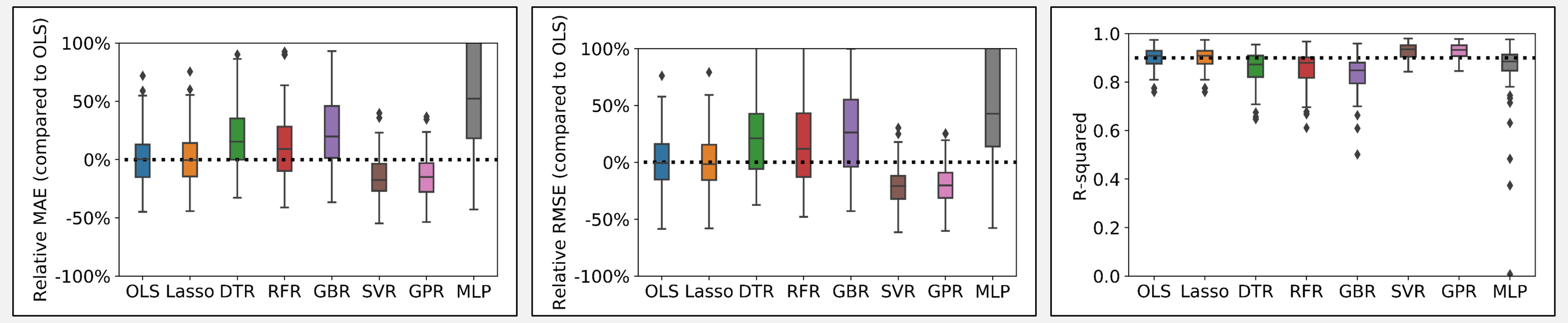
Example of Not Significantly Improved by ML Methods – NAICS 315 Tonnage

For the NAICS 315 tonnage estimation, all the ML algorithms have the third quartile of their MAE higher than the average MAE by OLS.



Example of Significantly Improved by ML Methods – NAICS 321 Value

Comparably, for the NAICS 321 value estimation, the SVR and GPR algorithms clearly show that the third quartile for MAE/RMSE are lower than the average MAE/RMSE by OLS.



Final Model Selection by Industry

The below tables summarize the final model suggestion for each NAICS code, which was determined based on two statistical tests, paired T-test and Wilcoxon, for the difference of MAE between the OLS and the alternative method for each NAICS.

Final Freight Generation Model Selection by Tonnage

NAICS	Model	Log-transform	*ESTAB*	*EMP*	*PAYANN*	*RCPTOT*	Regression parameters / Hyperparameters
212	SVR	No	√	√	√		epsilon: 0, kernel: sigmoid, C: 1.9
311	SVR	No	√	√		√	epsilon: 0, kernel: rbf, C: 1.1
312	SVR	Yes	√		√		epsilon: 0.06, kernel: poly, C: 2.1
313	SVR	No		√	√	√	epsilon: 0, kernel: rbf, C: 2.1
314	SVR	Yes	√	√		√	epsilon: 0.1, kernel: linear, C: 1.5
315	OLS	No		√			0.008 + 0.938 × *EMP*
316	SVR	Yes	√		√	√	epsilon: 0.12, kernel: rbf, C: 2.1
321	SVR	Yes	√	√	√	√	epsilon: 0.2, kernel: rbf, C: 1.9
322	SVR	Yes	√	√			epsilon: 0.2, kernel: linear, C: 1.9
323	GPR	Yes	√	√			sigma:0.5, noise level:1, alpha: 1e-9
324	SVR	No		√		√	epsilon: 0, kernel: sigmoid, C: 2.1
325	SVR	No			√	√	epsilon: 0, kernel: poly, C: 2.1
326	SVR	Yes	√	√			epsilon: 0, kernel: linear, C: 1.1
327	OLS	Yes	√				exp(-0.457) × *ESTAB*^0.805
331	GPR	Yes	√	√	√	√	sigma:0.5, noise level:1, alpha: 1e-11
332	SVR	Yes			√	√	epsilon: 0.08, kernel: linear, C: 2.1
333	SVR	No	√	√	√		epsilon: 0.02, kernel: linear, C: 2.1
334	SVR	No			√		epsilon: 0.02, kernel: sigmoid, C: 1.7
335	SVR	Yes		√	√		epsilon: 0.08, kernel: linear, C: 2.1
336	GPR	Yes	√	√	√	√	sigma:1, noise level:1, alpha: 1e-9
337	SVR	Yes	√	√		√	epsilon: 0.06, kernel: linear, C: 2.1
339	SVR	Yes		√			epsilon: 0.04, kernel: poly, C: 0.1
423	SVR	Yes		√	√		epsilon: 0, kernel: linear, C: 1.9
424	SVR	Yes	√			√	epsilon: 0.2, kernel: poly, C: 1.7

Final Freight Generation Model Selection by Value

NAICS	Model	Log-transform	*ESTAB*	*EMP*	*PAYANN*	*RCPTOT*	Regression parameters / Hyperparameters
212	SVR	Yes			√		epsilon: 0.04, kernel: linear, C: 0.3
311	GPR	Yes	√		√	√	sigma:1.5, noise level:1.5, alpha: 1e-9
312	SVR	Yes	√	√	√		epsilon: 0.08, kernel: linear, C: 1.1
313	SVR	Yes	√	√			epsilon: 0.14, kernel: linear, C: 2.1
314	GPR	No	√	√		√	sigma:0.5, noise level:0.5, alpha: 1e-9
315	SVR	No	√		√	√	epsilon: 0, kernel: linear, C: 1.9
316	SVR	Yes			√		epsilon: 0.2, kernel: linear, C: 0.5
321	SVR	Yes	√	√	√	√	epsilon: 0, kernel: linear, C: 2.1
322	SVR	Yes	√		√		epsilon: 0.16, kernel: linear, C: 1.7
323	GPR	Yes	√	√			sigma:1.5, noise level:1.5, alpha: 1e-9
324	GPR	Yes	√		√		sigma:1.5, noise level:1.5, alpha: 1e-11
325	SVR	No	√			√	epsilon: 0, kernel: linear, C: 1.5
326	GPR	Yes	√	√	√		sigma:1, noise level:1.5, alpha: 1e-10
327	SVR	No		√	√		epsilon: 0, kernel: rbf, C: 0.7
331	SVR	No	√	√	√	√	epsilon: 0.02, kernel: linear, C: 1.3
332	OLS	Yes				√	exp(0.062) × *RCPTOT*^1.009
333	SVR	No	√	√	√		epsilon: 0, kernel: linear, C: 1.3
334	GPR	No			√		sigma:1.5, noise level:1, alpha: 1e-9
335	SVR	Yes	√	√		√	epsilon: 0.18, kernel: linear, C: 0.7
336	SVR	No	√	√	√	√	epsilon: 0, kernel: linear, C: 0.9
337	SVR	No	√		√		epsilon: 0, kernel: sigmoid, C: 2.1
339	SVR	No	√	√	√		epsilon: 0, kernel: rbf, C: 2.1
423	SVR	No		√			epsilon: 0, kernel: rbf, C: 2.1
424	SVR	No	√	√		√	epsilon: 0, kernel: sigmoid, C: 2.1

Conclusions

Key Contributions:

- Built a framework to conduct the industry-specific model selection.
- Evaluated the significance of model improvements when using the ML algorithms over the OLS for the freight generation modeling.
- Suggested the use of OLS regression for certain industry sectors when the MAE reductions by the ML algorithms are not statistically significant.
- Utilized all combinations of available variables in the CBP & EC data tables for model selections.

Challenges:

- Limited to the freight shipments originated from the CFS areas.
- The model selection results might be quite different if one applies the same framework to estimate different dependent variables, such as truck volume and number of shipments.

Acknowledgement

This research effort was sponsored by the Federal Highway Administration (FHWA) and the Bureau of Transportation Statistics (BTS), under U.S. Department of Transportation.

Contact Information

Hyeonsup Lim, ORNL

limh@ornl.gov