sustainability

MDPI

Article

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, Oak Ridge, TN 37830, USA
* Correspondence: uddinm@ornl.gov; Tel.: +1-865-341-1306

Abstract: According to Bureau of Transportation Statistics, the U.S. transportation system handled 14,329 million ton-miles of freight per day in 2020. Understanding the generation of these freight shipments is crucial for transportation researchers, planners, and policymakers to design and plan for a more efficient and connected freight transportation system. Traditionally, the freight generation modeling has been based on Ordinary Least Square (OLS) regression, although more advanced Machine Learning (ML) algorithms have been evaluated and proven to have excellent performance in various transportation applications in recent years. Furthermore, one modeling approach applied for one industry might not always be applicable for another as their freight generation logics can be quite different. The objective of this study is to apply and evaluate alternative ML algorithms in the estimation of freight generation for each of 45 industry types. Seven alternative ML algorithms, along with the base OLS regression, were evaluated and compared. In addition, the study considered different combinations of variables in both the original and logarithmic form as well as hyperparameters of those ML algorithms in the model selection for each industry type. The results showed statistically significant improvements in the root mean square error reduction by the alternative ML algorithms over the OLS for over 80% of cases. The study suggests utilizing the alternative ML algorithms can reduce the root mean square error by about 30%, depending on industry types.

Keywords: freight generation model; freight production; freight attraction; North American Industry Classification System (NAICS); Commodity Flow Survey (CFS); machine learning algorithms

Citation: Lim, H.; Uddin, M.; Liu, Y.; Chin, S.-M.; Hwang, H.-L. A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model. *Sustainability* **2022**, *14*, 15367. https://doi.org/10.3390/su142215367

Academic Editor: Armando Cartenì

Received: 7 October 2022
Accepted: 16 November 2022
Published: 18 November 2022

1. Introduction

Freight transportation is a critical link in the supply chain of goods. It connects industry productions to demands and directly or indirectly affects national and regional economic productivity and growth. Bureau of Transportation Statistics (BTS) indicates that the U.S. transportation system handled 14,329 million ton-miles of freight per day in 2020 [1]. Understanding the generation of these freights, where they originate from and terminate to, is crucial for freight transportation researchers, planners, and policymakers to design and plan for a more efficient and connected freight transportation system. Note that the term "freight generation", commonly used in the transportation field, includes shipments both originated by (production) and terminated to (attraction) industry in this study.

In view of freight data needs, BTS initiated the quinquennial Commodity Flow Survey (CFS) since 1993 [2]. It is the only publicly available national survey in the U.S. on goods movement which provides national, state, and metropolitan-level data on freight shipments by industry sectors. The CFS data offers a comprehensive overview of the national freight generation and movement. Due to cost and other constraints, the CFS is conducted every five years and published data at state/metropolitan levels. Although CFS filled a large gap in freight data, the transportation communities have been expressing their desires for more timely data with granular geography for over a decade. To this end, freight generation models are frequently adopted by transportation analysts to estimate the quantity or value of goods generated from and/or attracted to a region. These models enable disaggregating the existing CFS data to local levels (e.g., county) and provide freight estimations for intermediate years between the CFS surveys.

This study utilized tonnage and value from the most recently released 2017 CFS data for 45 industry sectors as dependent variables and proposed industry-specific freight generation models based on industry-related factors such as number of establishments, annual payroll, number of employments, and receipt total. Traditionally, freight generation modeling approaches are based on Ordinary Least Square (OLS) regression [3–5]. While more complex Machine Learning (ML) algorithms have been evaluated and proved to have excellent performance in various transportation applications in recent years; based on the best of the authors' knowledge, there has not been any research done on adopting alternative ML models in freight generation estimation. The objective of this study is, therefore, to apply and evaluate the alternative ML algorithms in freight generation estimation.

Seven alternative ML algorithms, along with OLS regression, were evaluated in this study. This research explored various combinations of variables in both original and logarithmic forms, algorithms, and corresponding hyperparameters. Then, the study proposed a selection method to choose the best combination by industry to generate industry-specific models. The selection method considers those alternative algorithms, other than OLS regression (the baseline approach), only when the improvement of model performance is statistically significant. If not significant, the OLS is selected as it has the advantage in terms of interpretability, compared to more complex ML algorithms.

The paper is structured in seven sections. The next section presents a literature review on general approaches and data sources used for freight generation modeling, as well as the application of ML models. Section 3. (Data Sources) summarizes the data used in this study. The ML algorithms adopted in this study and the baseline OSL regression are elaborated in the section after. The following section describes the data processing and model selection procedure. Then, Section 6. (Model Results) discusses the model performance results and summarizes the final model selection for each industry. The final section concludes this study.

2. Literature Review

There exist two major classes of freight generation models, in terms of dependent variables. The classes are freight generation (FG) and freight trip generation (FTG). The FG models often focus on the weight or value of freight (e.g., tons/year) whereas the FTG models focus on the number of freight vehicle trips (e.g., truck trips/year). FG models are a better representation of the regional- or national-level economic activities given the capability to reflect the intensity of production and consumption of goods. Table 1 summarizes past studies where FG is modeled as weight. Due to the scope of the current study, studies on FTG are not included in the review.

As for the scope of the analysis, all the FG models presented in Table 1 were estimated at a regional level. Some models were also estimated at industry- and commodity-specific levels. This aggregate modeling has the advantage of predicting FG from regional economic and other related characteristics. However, this approach may result in certain aggregation biases in the estimated FG data. The alternative is to model using disaggregated data. The estimation of these disaggregated models, however, requires establishment-level freight generation data. These data are often collected through surveys for the freight generating facilities in the study region.

Several explanatory variables were used for FG modeling in past studies. These include employment, establishment size, annual payroll by industry sector, gross floor area, population/population density, port influence, and land use. Among all these, employment is invariably considered as the most preferred explanatory variable. Establishment size and payroll are often considered along with employment. Several studies performed FG modeling on fixed variables without exploring the impact of variable selections on the model output or fit. Furthermore, none of the studies considered receipts total (an important economic characteristic at industry/establishment level) as an explanatory variable.

Additionally, the majority of the studies utilized the OLS regression to model FG due to its ability to explain the relationship between freight activity and explanatory variables, as coefficients of regressions directly represent impacts to model estimates (refer to Section 6.4). The other methods used in the literature are Spatial Regression, Multiple Classification Analysis, One-way ANOVA, and Spatial Autoregressive Model. All these statistical approaches make strict assumptions about the data. Furthermore, a number of the existing models estimated models where explanatory variables affect FG in a linear form which may not always be true [6]. Advanced ML algorithms are often a promising alternative to statistical approaches. The advantage of ML algorithms is that they learn to represent complex relationships in a data-driven manner and are often non-parametric. The usefulness of ML algorithms has already been demonstrated for different areas in transportation research. For instance, ML algorithms are particularly used in modeling travel mode choice [7], freight mode choice [8], crash severity prediction [9], predicting the performance of asphalt mixture [10], and freight demand forecasting [11]. Hagenauer and Helbich [7] conducted a comparative analysis of seven machine learning classifiers for modeling travel mode choice. Uddin et al. [8] explored eight machine learning classifiers, using 2012 Commodity Flow Survey data, for modeling freight mode choice. Iranitalab and Khattak [9] compared four statistical and machine learning methods for prediction of crash severity. Rahman et al. [10] explored machine learning methods to predict two metrics of the performance of the asphalt mixture. Lastly, Salais-Fierro and Martínez [11] applied machine learning methods for demand forecasting in freight transportation.

Table 1. Summary of Studies on the Modeling of Freight Generation.

Study	Study Area	Data Source	Scope of Analysis	Variables Considered	Methods Used	Model Performance
Chin and Hwang [12]	United States	Commodity Flow Survey (CFS)	CFS Area and Industry Sector	Employment and Establishment Size	OLS Regression	Except for four models, $R^2 > 0.70$
	Key Findings: With additional modeling efforts, the developed models could be enhanced to allow transportation analysts to assess regional economic impacts.					
Holguin-Veras et al. [3]	Colombia	Freight Origin-Destination Survey	Region (made up of municipality and 4 countries)	Gross Domestic Product (GDP), Existence of Port	OLS Regression	Adjusted R^2: [0.86, 0.96]
	Key Findings: On average $1600 of GDP is needed to produce a ton of freight.					
Novak et al. [13]	United States	CFS and TranSearch	CFS Area	Population, Number of Employees, Port, Highway Length	OLS and Spatial Regression	R^2: [0.33, 0.63]
	Key Findings: It is recommended to avoid the overuse and addition of highly correlated explanatory variables such as employment and population even when this improves R^2; spatial regression model is the preferred specification for freight generation at the national level.					
Bagighni [14]	United States	Freight Analysis Framework (FAF)	FAF Zone and Commodity	Population, Median Age, Income, Number of Jobs by Industry Sector	OLS Regression	Adjusted R^2: [0.54, 0.81]
	Key Findings: It is possible to develop good freight volume estimating models for individual commodities using regression analysis; however, the level of success for each commodity model varies.					
Oliveira-Neto et al. [5]	United States	CFS	State and Industry	Annual Payroll by Industry Sector	OLS Regression	R^2: [0.40, 0.98]
	Key Findings: Payroll can explain a significant portion of the freight production at the state level for the U.S.					

Table 1. *Cont.*

Study	Study Area	Data Source	Scope of Analysis	Variables Considered	Methods Used	Model Performance
Lim et al. [4]	California	FAF	FAF Zone and Commodity Group	Number of Employees, Population, Farmland Acres, Crop and Livestock Sales, Net Annual Electrical Generation using Coal	OLS Regression	R^2: [0.21, 0.83]
	Key Findings: Models without constant terms have a better fit than models with constant; model fit is dependent on the commodity grouping and the choice of explanatory variables.					
Ha and Combes [15]	France	French Shipper Survey ECHO	Establishment	Employment, Economic Activity, Relations with Economic Agents, Production and Logistics Characteristics	One-way ANOVA and OLS Regression	R^2: [0.16, 0.45]
	Key Findings: The number of employees and the economic sector were identified as very important explanatory variables.					
Mommens et al. [16]	Belgium	Freight volume data compiled from multiple sources	Traffic Analysis Zone and Commodity	Number of Employees, Establishment Size, Gross Floor Space, Population Density	OLS Regression	R^2: [0.31, 0.69]
	Key Findings: It is doubtful that the addition of new explanatory variables will improve the model fit and consequently improvements in model accuracy.					
National Academies of Sciences, Engineering, and Medicine [17]	United States	CFS	Industry	Number of Employees	OLS Regression (linear and non-linear specifications) and Multiple Classification Analysis	Adjusted R^2: [0.01, 0.73]
	Key Findings: The use of the CFS in combination with complementary datasets provides an efficient way to estimate freight generation (FG) models for the entire nation at various levels of geography; non-linear models typically provide the best representation of FG patterns.					
Krisztin [6]	European NUTS-2 regions	Eurostat	Country	Regional Share of Employment, Regional Share of Employment in Agriculture and Manufacturing, Length of Road Network, and Distance to the Closest Seaport	Spatial Autoregressive Model	Adjusted R^2: [0.39, 0.87]
	Key Findings: There are significant non-linearities related to employment rates in manufacturing and infrastructure capabilities in the study regions.					

Compared to the referenced existing studies, the major contribution of this paper includes the followings:

- Evaluation of seven commonly used ML algorithms (i.e., Lasso, Decision Tree, Random Forest, Gradient Boosting, Support Vector, Gaussian Process, and Multi-layer Perceptron regressions), along with Ordinary Least Square (OLS) regression, with statistical tests on model performance
- Comprehensive scope of industry types—covered 45 North American Industry Classification System (NAICS) codes
- Inclusion of receipts total as an exploratory variable

- Industry-specific model selection—extensive model selection process considering model approach (ML algorithms), use of logarithm, and full combination of variable selection for each industry type

3. Data Sources

3.1. Dependent Variables—Freight Generation Data (Tonnage and Value)

The term "freight generation" is used differently in various studies. To clarify the FG modeling used in our study, "freight generation" is defined as the tonnage or value of freight shipments generated in a region associated with their business activities by each industry type. Note that our study does not estimate number of shipments or number of truck trips, which are considered FTG. As discussed in Section 2, FG models, compared to FTG, might better represent the regional- or national-level economic activities since they reflect the intensity of production and consumption of goods.

In addition, the term "freight generation" directly indicates that the study covered the estimation of shipments by both origins (freight production modeling) and destinations (freight attraction modeling). In freight planning, the "freight generation" is a prior process before the next step "freight distribution" (not covered by this study), which combines estimated shipments by origins and destinations and produces estimates of each origin-destination pair. The following two dependent variables are used for both the production (by origins) and the attraction (by destinations) modeling in our study, based on the 2017 CFS data [2].

- tonnage: total weight (in thousand tons) of shipments originated from (terminated to) a region by industry
- value: total value (in million dollars) of shipments originated from (terminated to) a region by industry

Note that there are variations associated with the sampling and other reporting errors that may have been incurred during the survey. Due to data confidentiality and data quality standard, Census suppressed tonnage and/or value for certain records in the public release of CFS data. Although there is another publicly available U.S. nationwide freight data, i.e., Freight Analysis Framework (FAF) [1], it was not considered in this study since the FAF data does not provide industry type information. The descriptions of all the 45 NAICS codes covered in this study are presented in Table A1.

3.2. Independent/Explanatory Variables—Economic/Industry Data

To develop reasonable FG models, many explanatory variables could be obtained from public/private data sources or derived using additional data processing. In this study, we used the indicators that represent economy or business activities, which are commonly used in the FG modeling studies. In addition, to potentially apply the FG models to disaggregate the CFS data, we need the input data at more granular level of geography (e.g., county). With such considerations, the study utilized the two county-level industry data products by Census, i.e., Economic Census (EC) [18] and County Business Pattern (CBP) data [19].

The CBP data, which is a part of the EC data program, are published annually between the five-year interval EC data releases. The main difference is that the EC tables provide additional business/economy information such as the receipt total (sales, revenue, or shipments) by industry, whereas the CBP provides number of employees, number of establishments, and annual payroll. All the in-scope industries in the CBP, as the name indicate, are provided at county level, whereas a few industries in the EC are provided at only state or selected geography level. As the study is to evaluate the FG modeling effort in terms of tonnage and value, only the industry types that were covered in the 2017 CFS data were considered in the EC and CBP tables as well. Among the NAICSs covered in our study, only two industry types, i.e., NAICS 212 (*mining except oil and gas*) and NAICS 551114 (*corporate*, *subsidiary*, *and regional managing offices*), do not have the receipt total at county level and therefore the variable receipt total was not included for the NAICS codes in our model selection process.

Like the 2017 CFS data, there is suppressed information in the CBP and EC tables as well. The imputation process is described in Section 5 (Data Processing and Model Selection). The following is a list of explanatory variables used in our study:

- 2017 CBP: number of establishments (*ESTAB*), number of employments (*EMP*), and annual payroll (*PAYANN*)
- 2017 EC: receipt total (*RCPTOT*) that is the total value of sales, revenue, or shipments

3.3. Shipments by Destinations (Freight Attraction)

The aforementioned explanatory and dependent variables are applied the same way for modeling both freight production (shipments by origins) and freight attraction (shipments by destinations), except for one additional step required for the freight attraction modeling. That is to derive origin-industry-specific input variables in respect to destinations, since the original CBP and EC data are provided by origin industries. The authors utilized the mapping of industry-to-industry share by the U.S. Bureau of Economic Analysis (BEA)'s "Make and Use" tables, following the same procedure as applied by Oliveira-Neto et al. [5]. The below equation represents the industry-to-industry mapping for deriving input variables for each industry's freight attraction model:

$$X'_{di} = \sum_{j} \omega_{ij} X_{dj} \quad (1)$$

where,

X'_{di} is the derived input variables for destination d by a linear combination of the shares of origin (make) industry i to destination (use) industry j;

ω_{ij} is the shares of origin industry i to destination industry j, obtained by the BEA's make and use table.

3.4. Descriptive Statistics of Input Data

Table 2 shows the descriptive statistics of the input data for shipments by origins (freight production modeling), after combining the EC and CBP data with the 2017 CFS data. For each variable, the mean and standard deviation is provided. For tonnage and value, the number of data point (sample size) is also presented as they are different by NAICS. This is because the suppressed tonnage and value in the 2017 CFS data were excluded from this study. As a result, the number of sample size (N) is smaller than 132 (number of the CFS areas) for most of NAICSs. There are suppressed information in the EC/CBP data as well, but the suppressed information in the EC and CBP data were imputed at county level before merging with the 2017 CFS data. More detailed description of data processing is provided in Section 5 (Data Processing and Model Selection).

Similarly, Table 3 shows the descriptive statistics of the input data for shipments by destinations (freight attraction modeling). Note that the number of sample size (N) in Table 3 is 132 (number of the CFS areas) for all NAICSs. This reflects that commodity shipments for each industry could be limited for certain origin areas but can be shipped to any destination zones. Note that the NAICS codes that were not included in the BEA Make and Use table were excluded in this study as the information is required to derive input variables for the freight attraction modeling (estimating shipments by destinations, refer to Section 3.3. (Shipments by Destinations (Freight Attraction)). The excluded NAICS codes for shipments by destinations are NAICS 4233, 4235, 4237, 4239, 4243, 4245, 4246, 4248, 4249, and 45431.

Table 2. Descriptive Statistics of the Input Data for Shipments by Origins.

NAICS	Tonnage (Thousand Tons)			Value (Million $)			Number of Establishments (Count)		Number of Employments (Count)		Annual Payroll (Million $)		Receipt Total (Million $)	
	N	Mean	Std. [1]	N	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
212	119	23,570	34,993	118	710	1308	34	35	1127	1741	76	128	N/A [2]	N/A
311	123	4865	6957	127	6255	7447	199	241	11,466	12,584	507	579	4392	5726
312	107	1296	2083	111	1384	2948	62	110	1628	2689	80	160	666	1339
313	67	89	184	101	284	629	11	28	577	1604	24	64	121	393
314	79	54	195	116	214	656	35	46	687	1604	25	66	114	521
315	43	9	26	98	140	478	42	226	686	3092	18	84	65	364
316	34	16	30	76	57	81	6	12	146	352	5	12	12	36
321	113	1939	2876	125	889	1142	103	108	3047	3593	121	142	610	810
322	112	1411	1827	113	1665	1969	26	34	2103	2873	124	177	686	1076
323	108	164	252	126	675	838	182	221	3469	4140	155	204	535	751
324	98	12,687	24,707	112	4623	11,030	10	13	655	1363	68	154	2067	6675
325	122	5694	11,200	127	5741	9853	93	117	5618	7089	439	636	4296	9013
326	120	497	616	129	1869	2205	86	105	5456	6696	260	316	1391	1868
327	106	6675	6988	129	993	931	103	85	2897	2594	149	138	730	753
331	103	1590	2752	112	1915	2512	27	37	2346	3460	149	239	1011	1881
332	114	899	1284	130	2700	3016	402	463	10,917	12,494	558	663	2423	2860
333	108	308	471	125	2959	3377	169	200	7815	8438	488	550	2268	2774
334	79	31	46	118	2660	4972	89	160	6052	10,593	521	1029	1811	3814
335	94	160	222	115	1098	1256	36	57	2138	2938	135	208	503	881
336	97	1069	2324	112	8093	13,983	81	104	11,616	16,517	749	1144	4736	10,549
337	115	122	178	126	610	889	101	124	2729	4096	113	172	424	785
339	100	66	81	124	1274	1743	203	270	4250	5928	234	386	952	1652
4231	104	691	1783	124	5349	11,377	173	237	3175	4426	169	311	2287	6471
4232	96	192	340	118	788	1430	97	203	1357	2644	78	155	487	1150
4233	96	1703	2095	129	1319	1544	125	128	1912	2082	108	128	759	1063
4234	87	241	564	119	4183	8084	255	375	4756	7634	418	877	3123	7721
4235	117	1058	1669	124	1573	2458	67	101	1124	1714	70	117	737	2215
4236	107	275	527	125	4325	8658	208	349	3931	7293	377	986	2829	6764
4237	113	209	246	132	1300	1518	143	157	1974	2280	121	151	747	1021
4238	95	563	951	128	3879	4449	434	466	6081	6513	388	458	2589	3641
4239	99	2240	4068	125	1786	3464	234	495	2590	4369	136	245	440	1947
4241	103	362	587	121	1062	1819	67	115	1115	1951	64	119	472	1230
4242	86	196	782	112	6299	12,078	69	158	2188	5438	254	807	2054	6738
4243	86	102	382	103	1360	4409	116	513	1615	6106	98	403	935	4719
4244	125	3060	4220	131	6580	9345	257	455	6383	8787	338	484	3781	7630
4245	82	10,628	18,113	89	2314	3524	45	93	531	1074	28	54	786	1789
4246	110	1230	2332	122	1506	2792	84	113	1131	1667	83	147	646	1848
4247	116	11,937	22,794	122	6847	12,943	36	36	664	932	55	140	2986	24,341
4248	127	526	579	129	1262	1820	28	49	1376	1887	86	148	600	1519
4249	89	2214	4538	129	2410	3124	209	302	2759	3388	134	167	235	763
4541	99	207	433	123	4314	8864	307	538	4291	6698	194	353	3129	7868
4931	119	2235	2416	121	9207	11,193	120	144	6848	9242	297	398	208	462
5111	77	23	24	93	146	211	107	129	2661	4461	164	413	508	1794
45431	120	433	722	124	272	393	50	68	535	859	24	43	148	344
551114	56	362	475	69	1324	1891	375	400	26,406	34,602	2823	4615	N/A	N/A

[1] Std.: Standard Deviation; [2] N/A: Not Available.

Table 3. Descriptive Statistics of the Input Data for Shipments by Destinations.

NAICS	Tonnage (Thousand Tons)			Value (Million $)			Number of Establishments (Count)		Number of Employments (Count)		Annual Payroll (Million $)		Receipt Total (Million $)	
	N	Mean	Std. [1]	N	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
212	132	21,951	22,729	132	702	919	33	34	1090	1682	73	124	44	137
311	132	4744	5189	132	6041	7218	198	242	11,175	12,308	498	569	4328	5619
312	132	1188	1597	132	1212	2504	65	108	1899	2793	98	167	844	1459
313	132	49	100	132	224	394	12	27	580	1525	25	62	130	375
314	132	40	120	132	195	448	40	51	942	1617	39	68	181	486
315	132	4	6	132	106	198	37	208	606	2848	16	78	58	335
316	132	5	20	132	38	79	6	12	176	344	6	12	30	49

Table 3. *Cont.*

NAICS	Tonnage (Thousand Tons)			Value (Million $)			Number of Establishments (Count)		Number of Employments (Count)		Annual Payroll (Million $)		Receipt Total (Million $)	
	N	Mean	Std. [1]	N	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
321	132	1708	1824	132	847	732	102	108	2998	3577	119	142	602	802
322	132	1236	1454	132	1454	1716	26	35	1995	2839	117	174	647	1053
323	132	154	237	132	650	819	171	207	3293	3913	149	195	515	722
324	132	10,438	21,857	132	4185	9811	10	13	620	1304	63	147	1915	6321
325	132	5412	9044	132	5645	8116	88	111	5305	6698	413	600	4106	8648
326	132	499	560	132	1826	2076	86	107	5338	6592	260	317	1433	1910
327	132	6484	6447	132	982	985	102	85	2899	2597	149	139	733	756
331	132	1326	1959	132	1721	2138	29	40	2279	3435	144	235	958	1823
332	132	864	1006	132	2684	2707	394	453	10,749	12,276	550	653	2401	2830
333	132	320	468	132	2845	2969	168	201	7697	8426	480	548	2249	2783
334	132	41	55	132	2437	4239	88	158	5905	10,349	504	1002	1788	3735
335	132	124	149	132	981	1059	37	58	2139	2967	135	210	518	897
336	132	858	1941	132	7276	12,126	80	104	11,087	16,183	715	1119	4506	10,261
337	132	115	113	132	585	623	102	125	2739	4075	114	172	435	785
339	132	58	73	132	1261	1574	214	284	4431	6095	250	404	1100	1844
4231	132	684	1202	132	5086	6652	165	228	3042	4260	162	299	2189	6221
4232	132	189	260	132	729	844	1626	2244	25,377	33,348	1758	2784	13,137	24,316
4234	132	208	308	132	3942	5207	229	338	4274	6882	376	790	2803	6957
4236	132	289	399	132	4168	5741	201	337	3800	7050	364	953	2735	6539
4238	132	702	1129	132	3791	3683	388	417	5444	5831	347	410	2317	3259
4241	132	364	574	132	992	1350	852	1571	16,783	25,333	1074	1892	11,693	32,900
4242	132	162	294	132	5964	7160	64	149	2025	5121	235	759	1900	6339
4244	132	3028	3764	132	6537	8251	248	440	6164	8486	326	468	3651	7368
4247	132	10,905	21,166	132	6467	12,033	32	33	596	850	49	127	2677	22,080
4541	132	185	232	132	4095	5009	353	559	4773	6974	216	368	3238	7905
4931	132	2083	1899	132	8955	8550	114	137	6466	8827	281	380	198	443
5111	132	31	64	132	181	354	79	96	1990	3311	124	307	383	1331
551114	132	540	1235	132	1125	1548	361	395	25,336	34,027	2696	4504	N/A [2]	N/A

[1] Std.: Standard Deviation; [2] N/A: Not Available.

4. Machine Learning Algorithms

Seven commonly used machine learning algorithms (i.e., Lasso, Decision Tree, Random Forest, Gradient Boosting, Support Vector, Gaussian Process, and Multi-layer Perceptron regressions), along with Ordinary Least Square (OLS) regression, were considered for the comparison.

4.1. Ordinary Least Squares Regression (OLS, the Baseline)

As discussed previously, the OLS regression is the most often used method in the FG models. Therefore, OLS method was used as the baseline to be compared with other ML algorithms. As the OLS has advantage of simplicity and interpretability over other ML algorithms, the OLS is still suggested as the final model if the model performance improvements by the ML algorithms are not found to be statistically significant.

$$Y_{oi} = \alpha_i + \beta_i X_{oi} + \varepsilon_{oi} \tag{2}$$

where,

Y_{oi} is the tonnage/value of freight shipments originated from origin o and industry i;
X_{oi} is the set of explanatory variables for origin o and industry i;
α_i and β_i are the parameter estimates of linear regression model for industry i.

4.2. Least Absolute Shrinkage and Selection Operator (Lasso)

Least absolute shrinkage and selection operator (Lasso) is a regression technique that performs both variable selection and regularization, by "shrinking" coefficients of regression models, to enhance the estimation accuracy while providing the interpretability of typical linear regression models. The Lasso shrinks coefficients by adding the penalty term to the residual sum of squares (RSS) that is to be minimized in the OLS.

$$Minimize\ [RSS + \lambda \sum |\beta_i|] \tag{3}$$

The model flexibility decreases as λ increases, leading to smaller variance but larger bias. This is especially useful to mitigate overfitting, which is frequently observed for small sample size data. In the Lasso module of the Python scikit-learn library—version 0.24.2 [20], used in this study, the λ can be controlled with the hyperparameter named "*alpha*". The Lasso method evaluated for the FG model in our study used the "*alpha*" ranging from 0 to 0.02 with an increment of 0.002.

4.3. Decision Tree Regression (DTR)

The Decision Tree is one of the popular non-parametric supervised learning methods used for classification and regression, depending on whether the dependent variable is categorical or continuous (like tonnage/value in this study). The main advantages of the Decision Tree regression (DTR) are: (1) the splits of nodes are unbiased; (2) each node contains a single model fit, relatively easier to interpret the model result; and (3) there are less limitations for applying the residuals, including general least squares.

Using the *DecisionTreeRegressor* module in the Python scikit-learn library [20], the following hyperparameter settings were evaluated and the hyperparameter setting with the lowest Root Mean Square Error (RMSE) was used for each model selection by DTR.

- Maximum depth of the tree (*max_depth*): [1, 2, 3, 4, 5]
- Complexity parameter used for the minimal cost-complexity pruning (*ccp_alpha*): [0, 0.002, . . . , 0.018, 0.02]

4.4. Random Forest Regression (RFR)

The Random Forest is an ensemble learning method for supervised learning, designed to improve model accuracy by randomly constructing multiple decision trees, rather than just one tree. Random Forest regression (RFR) is simply an ensemble of multiple random regression trees for the continuous dependent variables. The Random Forest is known to produce highly accurate estimation results for large sample sizes. Once the model is trained, the prediction process is relatively efficient, significantly faster than the training speed.

Using the *RandomForestRegressor* module in the Python scikit-learn library [20], the following hyperparameter settings were evaluated and the hyperparameter setting with the lowest RMSE was used for each model selection by RFR.

- Maximum depth of the tree (*max_depth*): [1, 2, 3, 4, 5]
- Complexity parameter used for the minimal cost-complexity pruning (*ccp_alpha*): [0, 0.002, . . . , 0.018, 0.02]
- Number of trees in the forest (*n_estimators*): 10

4.5. Gradient Boosting Regression (GBR)

Gradient Boosting is another ensemble learning technique that forms multiple decision trees sequentially accounting for weak predictions of the previous decision trees. Specifically, the next trees are trained on the weighted data where more weights are assigned for the observations that were more difficult to estimate or classify in the previous iteration. If the sample size is sufficient for the training, the Gradient Boosting can outperform the Random Forest.

Using the *GradientBoostingRegressor* module in the Python scikit-learn library [20], the following hyperparameter settings were evaluated and the hyperparameter setting with the lowest RMSE was used for each model selection by GBR.

- Maximum depth of the tree (*max_depth*): [1, 2, 3, 4, 5]

- Complexity parameter used for the minimal cost-complexity pruning (*ccp_alpha*): [0, 0.002, . . . , 0.018, 0.02]
- Learning rate to control the contribution of each tree (*learning_rate*): [0.01, 0.1, 1]
- Number of the estimators (trees) (*n_estimators*): 10

4.6. Support Vector Regression (SVR)

Support Vector Machines (SVMs), which are more often used in classification, refer to a set of supervised learning for classification and regression, using a subset of training data as for the decision points, also called "support vectors". Support Vector regression (SVR) is generally advantageous for high dimensional data, possibly effective even when number of dimensions is greater than the sample size. The choice of kernel functions and regularization parameters can be critical to avoid over-fitting in the SVR.

Using the *SVR* module in the Python scikit-learn library [20], the following hyperparameter settings were evaluated and the hyperparameter setting with the lowest RMSE was used for each model selection by SVR.

- Margin of tolerance where no penalty is given to errors (*epsilon*): [0, 0.002, . . . , 0.018, 0.02]
- Regularization parameter (C): [0.1, 0.3, . . . , 1.9, 2.1]
- Kernel distribution type to be used in the algorithm (*kernel*): [Linear, Polynomial, Gaussian (RBF), Sigmoid]

4.7. Gaussian Process Regression (GPR)

Gaussian Process regression (GPR) is an extension of linear regression, where "Gaussian Process" represents finite linear combinations of random variables that are normally distributed. During the model fitting of GPR, the hyperparameters of the kernel are optimized to maximize the log-marginal-likelihood based on the passed optimizer. One of the main advantages by Gaussian Process is that the estimates can be provided in probabilistic forms where their empirical confidence interval can also be obtained.

Using the *GaussianProcessRegressor* module in the Python scikit-learn library [20], the following hyperparameter settings were evaluated and the hyperparameter setting with the lowest RMSE was used for each model selection by GPR.

- Constant value added to the diagonal of the kernel matrix (*alpha*): [1×10^{-11}, 1×10^{-10}, 1×10^{-9}]
- Kernel specifying the covariance function of the model (kernel):
- Combined two kernels, Dot-Product kernel and White kernel
- For the Dot-Product kernel (*DotProduct*), the parameter *sigma* to control the inhomogeneity of the kernel: [0.5, 1.0, 1.5]
- For the White kernel (*WhiteKernel*), the parameter *noise_level* to control the noise level of the kernel: [0.5, 1.0, 1.5]

4.8. Multi-Layer Perceptron Regression (MLP)

Multi-layer Perceptron (MLP) is a class of feedforward artificial neural network, where the "multi-layer" refers to consisting of at least three layers: input layer, hidden layer, and output layer. The MLP utilizes backpropagation for training, and different activation functions can be used for the training of hidden layers. The MLP is known to require relatively large data size for the training.

Using the *MLPRegressor* module in the Python scikit-learn library [20], the following hyperparameter settings were evaluated and the hyperparameter setting with the lowest RMSE was used for each model selection by MLP.

- Hidden layer size and number of neurons in each hidden layer (*hidden_layer_sizes*)
- Number of hidden layers: [1, 2, 3]
- Number of neurons in each hidden layer: [3, 4, 5]
- L_2 penalty parameter (*alpha*): [0.00001, 0.0001, 0.001]

- Activation function for the hidden layer (*activation*): [Identity, Logistic, Rectified Linear Unit (ReLU)]

5. Data Processing and Model Selection

5.1. Imputation of Missing Data (for CBP/EC)

For the records by origin (for all the 132 CFS areas) and industry (for the NAICSs covered in this study) in the published 2017 CFS tables, about 19% of tonnage and 8% of value are suppressed due to the sampling variability. These suppressed tonnage and value were excluded from evaluation in our study because different imputing methods could affect the modeling results inadvertently, especially for evaluating different modeling approaches.

The county level CBP and EC data also have suppressed information for the number of employments, annual payroll, number of establishments, and receipt total. Unlike the rest of the variables, the group for the range of number of employments are provided in the CBP data where the exact number of employments are suppressed. Therefore, in the first step, we imputed the number of employments by using the mid-point of the employment size range (*EMPFLAG*).

After the number of employments is imputed, the suppressed values for the rest of variables were imputed based on the ratio of the attribute value over the number of employments for known values. This imputation process was conducted at the county level data for each NAICS code. Once the imputation is completed, county-level data were aggregated to the CFS area-level to be merged with the 2017 CFS data.

5.2. Data Transformation

The FG modeling may require transformation of the input data to improve the accuracy since their relationships may not be linear as in the original units. In this study, we evaluated the model performance either in the original input data units or log-transformed values. The following equation indicates the case where both explanatory variables and dependent variables are transformed with natural logarithm.

$$log(Y_{oi}) = \alpha'_i + \beta'_i log\ (X_{oi}) + \varepsilon'_{oi} \tag{4}$$

$$Y_{oi} = exp\ (\alpha'_i + \varepsilon'_{oi}) \cdot X_{oi}^{\beta'_i} \tag{5}$$

For more comparable model selection evaluation, the final model results were converted to the original units of tonnage and value if log-transformed.

5.3. Normalization

The normalization is the process of rescaling the input data to a similar range or distribution across different attributes. In our study, this is an essential process to improve the model stability and performance especially for more complex models, such as MLP. The normalization could be also helpful to interpret the importance of variables based on regression parameters that have different units if not normalized. In our study, a simple min-max normalization was used, where the min value was set to be zero for all cases. As such, the normalized value in our study is obtained simply by dividing the original value with the max value of each attribute.

5.4. Variable Selection

There are many different techniques for the variable selection, such as forward, backward, and stepwise selection. However, these techniques are heuristic approaches in that they choose or change subset of possible variable selection based on the previous variable selection result. Under our study, the number of explanatory variables is only four, except for NAICSs 212 and 551114 (only three without the receipt total). Therefore, instead of applying such variable selection techniques, this study evaluated all possible variable combinations among the four (or three) independent variables. The maximum number of possible combinations is $\mathbf{15}\ (= \mathbf{2^4} - \mathbf{1})$ with four independent variables, excluding the one case that none of independent variables are selected.

5.5. Optimization of Hyperparameters

For each modeling approach, different hyperparameter setting could yield substantially different model performance results. Therefore, the authors attempt to test many different hyperparameters discussed in Section 4 (Machine Learning Algorithms). Then, the hyperparameter setting with the lowest RMSE among all tested settings was selected for each model approach and variable selection. Note that not all possible hyperparameters, such as minimum number of required samples at a leaf node in Decision Tree regression, were tested in the study.

5.6. Model Performance—Error Measurements

Three metrics were used for the model performance evaluation in this study:

- Root Mean Square Error (RMSE): the square root of arithmetic mean of the squared difference between the 2017 CFS tonnage/value and the estimated tonnage/value
- Mean Absolute Error (MAE): the arithmetic mean of the absolute difference between the 2017 CFS tonnage/value and the estimated tonnage/value
- R-squared: the R-squared (or coefficient of determination) statistic between the 2017 CFS tonnage/value and the estimated tonnage/value

Both the RMSE and the MAE are commonly used to measure accuracy of continuous variables (i.e., tonnage and value). The study used the RMSE as the primary metric to determine the final model selection by NAICS, because the OLS regression (baseline model) fits to minimize the sum of squared error. In other words, using the MAE as the primary metric for the OLS could bias the final model selection toward preferring one of the alternative ML models over the OLS.

All the three metrics were evaluated based on only the validation sets. Note that the validation sets are based on K-fold cross validation where K is 4, with 25 times of repeats. Therefore, there are 100 validation sets and associated performance metrics observed for each model selection. The K-fold cross validation is useful especially when dealing with small dataset ($N \leq 132$ for each industry), since the data is split into K number of folds making all parts of the data being equally used as part of the validation sets.

Furthermore, unadjusted R-squared was used, instead of adjusted R-squared which is used as a correction to the unadjusted R-squared for the case with multiple predictors. This is because the R-squared was obtained only based on the validation sets, not the training set, where the model complexity is already accounted in the estimates of validation sets. Finally, alternative models, other than OLS, are selected only when the reduction of RMSE appears to be statistically significant by paired T-test and Wilcoxon statistics with the *p*-value of 0.05.

6. Model Results

To better understand the model selection process, we present an example of model selection process with the case of estimating tonnage of shipments by origins for NAICS 212. Then, the following sections will discuss the significance of model improvement and summarize the final model for each industry.

6.1. Example of Model Selection—Tonnage of Shipments by Origins for NAICS 212

The model selection was considered with different aspects, i.e., variable selection, log-transform, and ML algorithms, as well as the hyperparameter settings of each ML algorithm. For an easier understanding of the model selection choices considered for each NAICS, Table 4 presents an example of model selection with associated RMSEs for estimating tonnage of shipments by origins for NAICS 212.

Table 4. RMSE by Model Selection—Tonnage of Shipments by Origins for NAICS 212.

	Variable Selection	OLS [1]	Lasso [2]	DTR [3]	RFR [4]	GBR [5]	**SVR [6]**	GPR [7]	MLP [8]
No Log-Transform	*EMP*	26,911	30,419	32,947	32,672	30,163	28,093	26,983	45,787
	PAYANN	29,914	32,346	32,898	32,687	27,125	26,705	29,997	51,983
	ESTAB	27,221	28,141	32,898	29,658	29,481	27,508	27,100	43,834
	EMP, PAYANN	28,814	30,419	32,877	30,184	28,469	27,328	30,583	41,621
	EMP, ESTAB	27,775	30,419	32,898	30,345	31,922	**25,341**	28,378	50,355
	PAYANN, ESTAB	30,015	32,346	32,898	30,785	31,769	26,215	30,306	45,076
	EMP, PAYANN, ESTAB	29,708	30,419	36,093	31,680	31,605	26,755	31,384	48,811
Log-Transform	*EMP*	25,608	25,742	29,348	28,221	30,282	27,645	25,436	86,158
	PAYANN	26,083	26,181	30,701	28,383	30,350	28,650	25,952	40,286
	ESTAB	29,122	29,248	29,318	28,726	30,826	27,478	28,980	36,600
	EMP, PAYANN	25,989	26,102	29,515	28,123	30,254	28,563	25,833	98,109
	EMP, ESTAB	25,733	25,841	30,217	28,278	30,344	27,092	26,363	45,223
	PAYANN, ESTAB	26,281	26,354	30,721	28,485	30,366	27,141	26,741	415,283
	EMP, PAYANN, ESTAB	26,134	26,219	30,054	28,304	30,300	27,092	26,679	33,675

[1] OLS: Ordinary Least Squares Regression, [2] Lasso: Least Absolute Shrinkage and Selection Operator, [3] DTR: Decision Tree Regression, [4] RFR: Random Forest Regression, [5] GBR: Gradient Boosting Regression, [6] SVR: Support Vector Regression, [7] GPR: Gaussian Process Regression, [8] MLP: Multi-layer Perceptron.

As presented in Table 4, there is a total of 112 choices for the model selection of NAICS 212: 2 choices for the log-transform, 2 choices for each variable (excluding the case with no explanatory variable), and 8 different algorithms. In fact, there is no receipt total information at county-level (available for only state level) from the EC table for the NAICS 212 and 551114, but all the other industry types in this study have the county-level receipt total information. Therefore, the total number of possible model selections was 240 for industries other than NAICS 212 and 551114.

In addition to the summarized model selection in Table 4, the different hyperparameter settings were tested as well and then only the hyperparameter settings with the lowest RMSEs were presented in Table 4. For the NAICS 212 tonnage estimation, the SVR model with number of employees and number of establishments was selected as the best alternative model since it yielded the lowest RMSE among all the options.

Finally, the alternative ML algorithm was suggested as the final model only when the reduction of RMSE over OLS is statistically significant with paired T-test and Wilcoxon statistics, as shown in Table 5.

6.2. Significance of Model Performance Improvement by Industry

Figure 1 shows the box plots of three model performance measurements based on 100 validation sets (K-fold cross validation where K is 4, with 25 times of repeats) for two dependent variables: (a) tonnage of shipments by origins for NAICS 333 and (b) tonnage of shipments by destinations for NAICS 337. The two cases were chosen intentionally to provide two distinguishable examples of “with” versus “without” significant improvement by the alternative ML algorithms. More specifically, the first example on the left side (Figure 1a) shows the case where the alternative ML algorithm does improve the model performance significantly, whereas none of the alternative ML models showed statistically significant improvement over OLS for the example on the right side (Figure 1b).

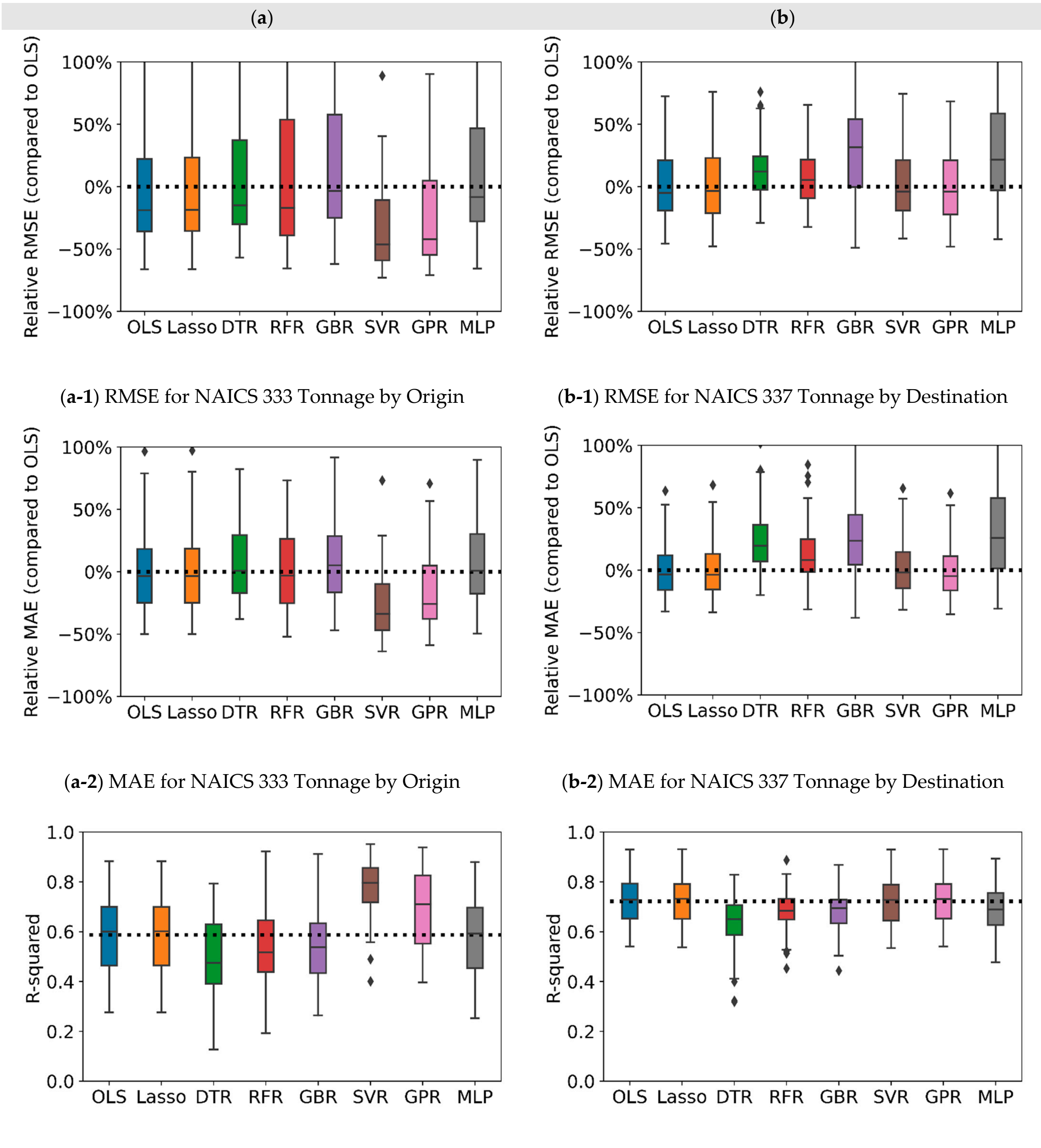
Table 5. Significance of Improvement by ML algorithms over OLS—Shipments by Origins.

NAICS	Measure	Alternative	RMSE			*t*-Test		Wilcoxon	
			OLS	**Alternative**	**% Dif.**	**Stat.**	***p*-Value**	**Stat.**	***p*-Value**
212	tons	SVR	25,733	25,341	−1.5%	2.41	0.018 *	1780	0.01 *
	value	SVR	466	449	−3.8%	4.22	<0.0005 *	1566	0.001 *
311	tons	GPR	4447	3963	−10.9%	11.31	<0.0005 *	131	<0.0005 *
	value	SVR	3020	2576	−14.7%	8.75	<0.0005 *	541	<0.0005 *
312	tons	SVR	1667	1414	−15.2%	5.79	<0.0005 *	891	<0.0005 *
	value	SVR	1985	1894	−4.6%	5.84	<0.0005 *	894	<0.0005 *
313	tons	OLS	79	-	-	-	-	-	-
	value	SVR	239	226	−5.5%	4.80	<0.0005 *	1288	<0.0005 *
314	tons	GPR	67	47	−29.1%	7.74	<0.0005 *	318	<0.0005 *
	value	SVR	158	144	−9.2%	2.78	0.006 *	1413	<0.0005 *
315	tons	OLS	11	-	-	-	-	-	-
	value	OLS	135	-	-	-	-	-	-
316	tons	GPR	26	26	−0.1%	1.20	0.232	1680	0.004 *
	value	SVR	68	66	−2.7%	3.57	0.001 *	1713	0.005 *
321	tons	RFR	1946	1605	−17.5%	7.76	<0.0005 *	721	<0.0005 *
	value	SVR	384	311	−19.2%	11.19	<0.0005 *	247	<0.0005 *
322	tons	SVR	1627	1492	−8.3%	12.85	<0.0005 *	161	<0.0005 *
	value	SVR	1167	1128	−3.4%	4.66	<0.0005 *	1148	<0.0005 *
323	tons	SVR	179	150	−16.1%	11.12	<0.0005 *	183	<0.0005 *
	value	SVR	244	241	−1.4%	0.98	0.331	2186	0.244
324	tons	SVR	11,290	10,730	−5.0%	3.86	<0.0005 *	1523	0.001 *
	value	SVR	5009	4389	−12.4%	6.45	<0.0005 *	506	<0.0005 *
325	tons	GPR	8887	8783	−1.2%	8.09	<0.0005 *	298	<0.0005 *
	value	SVR	3592	3409	−5.1%	3.36	0.001 *	1945	0.046 *
326	tons	SVR	319	279	−12.4%	10.15	<0.0005 *	390	<0.0005 *
	value	SVR	771	732	−5.1%	7.72	<0.0005 *	569	<0.0005 *
327	tons	SVR	4768	4441	−6.9%	9.39	<0.0005 *	383	<0.0005 *
	value	SVR	361	355	−1.5%	1.51	0.133	1909	0.034 *
331	tons	SVR	1525	1341	−12.1%	5.78	<0.0005 *	875	<0.0005 *
	value	SVR	966	926	−4.1%	7.04	<0.0005 *	727	<0.0005 *
332	tons	SVR	925	909	−1.7%	1.07	0.286	2473	0.858
	value	SVR	627	622	−0.8%	1.69	0.095	1826	0.016 *
333	tons	SVR	333	231	−30.6%	18.83	<0.0005 *	9	<0.0005 *
	value	SVR	1391	1186	−14.7%	14.25	<0.0005 *	51	<0.0005 *
334	tons	SVR	42	37	−13.4%	14.22	<0.0005 *	-	<0.0005 *
	value	OLS	1085	-	-	-	-	-	-
335	tons	SVR	216	201	−6.8%	10.93	<0.0005 *	347	<0.0005 *
	value	OLS	730	-	-	-	-	-	-
336	tons	OLS	1662	-	-	-	-	-	-
	value	SVR	4478	4437	−0.9%	1.16	0.249	1918	0.037 *
337	tons	SVR	108	96	−11.2%	10.16	<0.0005 *	442	<0.0005 *
	value	OLS	225	-	-	-	-	-	-
339	tons	DTR	69	66	−4.3%	3.02	0.003 *	1687	0.004 *
	value	SVR	871	642	−26.3%	6.39	<0.0005 *	739	<0.0005 *
4231	tons	OLS	876	-	-	-	-	-	-
	value	OLS	4492	-	-	-	-	-	-

Table 5. *Cont.*

NAICS	Measure	Alternative	RMSE			*t*-Test		Wilcoxon	
			OLS	Alternative	% Dif.	Stat.	*p*-Value	Stat.	*p*-Value
4232	tons	SVR	131	118	−9.9%	4.91	<0.0005 *	1418	<0.0005 *
	value	SVR	296	289	−2.4%	2.31	0.023 *	2432	0.749
4233	tons	GBR	1295	1110	−14.3%	6.92	<0.0005 *	801	<0.0005 *
	value	SVR	446	419	−6.1%	5.63	<0.0005 *	963	<0.0005 *
4234	tons	SVR	333	327	−1.7%	1.04	0.299	1968	0.055
	value	OLS	2746	-	-	-	-	-	-
4235	tons	SVR	969	878	−9.4%	3.47	0.001 *	1520	0.001 *
	value	OLS	1031	-	-	-	-	-	-
4236	tons	Lasso	372	371	−0.3%	1.54	0.126	2312	0.464
	value	SVR	3905	3686	−5.6%	4.04	<0.0005 *	1291	<0.0005 *
4237	tons	OLS	118	-	-	-	-	-	-
	value	SVR	543	513	−5.5%	6.31	<0.0005 *	924	<0.0005 *
4238	tons	SVR	544	519	−4.7%	1.03	0.305	2048	0.101
	value	Lasso	1391	1382	−0.6%	0.92	0.36	2297	0.433
4239	tons	Lasso	1494	1477	−1.1%	0.75	0.452	2356	0.561
	value	Lasso	793	743	−6.4%	1.72	0.089	2275	0.39
4241	tons	OLS	382	-	-	-	-	-	-
	value	OLS	843	-	-	-	-	-	-
4242	tons	SVR	574	568	−1.0%	3.33	0.001 *	1642	0.002 *
	value	OLS	8976	-	-	-	-	-	-
4243	tons	OLS	96	-	-	-	-	-	-
	value	SVR	1175	845	−28.1%	5.69	<0.0005 *	971	<0.0005 *
4244	tons	OLS	1154	-	-	-	-	-	-
	value	OLS	1949	-	-	-	-	-	-
4245	tons	RFR	8771	8485	−3.3%	1.10	0.275	2367	0.587
	value	DTR	1809	1695	−6.3%	2.20	0.03 *	1700	0.005 *
4246	tons	OLS	1525	-	-	-	-	-	-
	value	OLS	1075	-	-	-	-	-	-
4247	tons	OLS	16,491	-	-	-	-	-	-
	value	SVR	8490	8245	−2.9%	1.78	0.078	2171	0.224
4248	tons	OLS	314	-	-	-	-	-	-
	value	SVR	614	571	−6.9%	3.69	<0.0005 *	1753	0.008 *
4249	tons	SVR	2608	2235	−14.3%	4.93	<0.0005 *	1515	0.001 *
	value	SVR	1685	1652	−2.0%	3.13	0.002 *	1575	0.001 *
4541	tons	SVR	282	276	−2.2%	2.95	0.004 *	1661	0.003 *
	value	SVR	5102	4833	−5.3%	4.54	<0.0005 *	1367	<0.0005 *
45431	tons	SVR	324	303	−6.6%	3.82	<0.0005 *	1520	0.001 *
	value	OLS	119	-	-	-	-	-	-
4931	tons	SVR	1494	1394	−6.7%	5.97	<0.0005 *	231	<0.0005 *
	value	SVR	5854	5505	−6.0%	3.98	<0.0005 *	794	<0.0005 *
5111	tons	SVR	21	20	−6.3%	8.79	<0.0005 *	493	<0.0005 *
	value	SVR	183	182	−0.4%	0.94	0.352	1852	0.021 *
551114	tons	RFR	493	480	−2.7%	2.39	0.019 *	1680	0.004 *
	value	DTR	1733	1653	−4.7%	3.80	<0.0005 *	1503	<0.0005 *
									(* p-value < 0.05)

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



(**a-1**) RMSE for NAICS 333 Tonnage by Origin (**b-1**) RMSE for NAICS 337 Tonnage by Destination

(**a-2**) MAE for NAICS 333 Tonnage by Origin (**b-2**) MAE for NAICS 337 Tonnage by Destination

(**a-3**) R-squared for NAICS 333 Tonnage by Origin (**b-3**) R-squared for NAICS 337 Tonnage by Destination

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

Figure 1. Box Plot with Model Performance: (**a**) Tonnage of Shipments by Origins for NAICS 333; (**b**) Tonnage of Shipments by Destinations for NAICS 337: the box extends from the first quartile (Q1) to the third quartile (Q3) and the whiskers extend at the farthest data points within the interval, no more than 1.5× the interquartile (Q3-Q1) from the edges of the box; the rhombus marks (⧫) represent the data points outside this range of the whiskers.

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

RMSE/MAE is calculated as the difference of RMSE/MAE between each ML algorithm and the OLS divided by the RMSE/MAE of OLS. The dotted line in Figure 1 represents the arithmetic mean of each performance metric for the baseline algorithm, OLS.

For estimating tonnage of shipments by origins for NAICS 333 (Figure 1a), the SVR algorithm clearly shows that the third quartiles (i.e., upper bound of the colored box) for both RMSE and MAE are lower than the average RMSE/MAE by OLS. In addition, the first quartiles (i.e., lower bound of the colored box) for the R-squared values are also higher than the average R-squared by OLS method.

Comparably, for estimating tonnage of shipments by destinations for NAICS 337 (Figure 1b), all the ML algorithms have the third quartile of both RMSE and MAE higher than the average RMSE/MAE by OLS. For the R-squared value of the NAICS 337 tonnage estimation, the median (i.e., the mid-line inside of the box) of R-squared values by DTR, RFR, GBR, and MLP are even lower than the average R-squared value by OLS.

In addition to the visual investigation of the box plots in Figure 1, statistical tests were conducted to evaluate the significance of model performance improvements. Specifically, as shown in Tables 5 and 6, two statistical tests, paired t-test and Wilcoxon, for the difference of RMSE between the OLS and the alternative best ML method for each NAICS were evaluated with a significance level of p-value 0.05. The alternative method was suggested as the final model only when both of the test statistics show significant improvements of RMSE, as compared to the RMSE by OLS. Note that no alternative ML methods are provided in Tables 5 and 6, where OLS performed better than all the seven alternative ML methods. Overall, both t-test and Wilcoxon statistics yield fairly consistent conclusions in terms of which industry types were improved significantly over the OLS by applying the alternative ML method.

Table 6. Significance of Improvement by ML algorithms over OLS—Shipments by Destinations.

NAICS	Measure	Alternative	RMSE			*t*-Test		Wilcoxon	
			OLS	Alternative	%Dif.	Stat.	*p*-Value	Stat.	*p*-Value
212	tons	RFR	18,452	17,962	−2.7%	2.67	0.009 *	1795	0.012 *
	value	SVR	770	756	−1.8%	5.20	<0.0005 *	1042	<0.0005 *
311	tons	GPR	2083	2078	−0.3%	0.54	0.59	2095	0.139
	value	SVR	3537	3482	−1.6%	1.87	0.065	1843	0.019 *
312	tons	SVR	1077	1050	−2.6%	4.86	<0.0005 *	1038	<0.0005 *
	value	SVR	1827	1796	−1.7%	4.08	<0.0005 *	1538	0.001 *
313	tons	GPR	73	73	−0.3%	0.47	0.638	1643	0.002 *
	value	SVR	261	254	−2.9%	4.28	<0.0005 *	1332	<0.0005 *
314	tons	GPR	47	43	−7.6%	6.17	<0.0005 *	776	<0.0005 *
	value	GPR	121	120	−0.2%	2.13	0.036 *	2036	0.093
315	tons	GPR	4	4	−3.8%	3.82	<0.0005 *	1428	<0.0005 *
	value	GPR	106	91	−13.8%	7.13	<0.0005 *	728	<0.0005 *
316	tons	MLP	15	14	−6.6%	1.32	0.189	2513	0.967
	value	SVR	61	49	−20.2%	7.25	<0.0005 *	657	<0.0005 *
321	tons	SVR	1154	961	−16.8%	18.37	<0.0005 *	14	<0.0005 *
	value	RFR	443	419	−5.4%	4.67	<0.0005 *	1269	<0.0005 *
322	tons	SVR	709	693	−2.2%	3.28	0.001 *	1883	0.027 *
	value	GPR	691	691	−0.1%	1.09	0.278	2263	0.368
323	tons	GPR	135	133	−2.0%	5.03	<0.0005 *	852	<0.0005 *
	value	SVR	253	244	−3.6%	1.58	0.118	2243	0.332
324	tons	GPR	10,031	10,017	−0.1%	1.38	0.17	2305	0.449
	value	GPR	4476	3775	−15.7%	11.84	<0.0005 *	175	<0.0005 *

Table 6. *Cont.*

NAICS	Measure	Alternative	RMSE			*t*-Test		Wilcoxon	
			OLS	Alternative	%Dif.	Stat.	*p*-Value	Stat.	*p*-Value
325	tons	GPR	5442	5437	−0.1%	1.22	0.226	2194	0.255
	value	GPR	3960	3957	−0.1%	2.34	0.021 *	1952	0.049 *
326	tons	SVR	260	257	−1.0%	1.82	0.072	2016	0.08
	value	SVR	868	864	−0.4%	0.81	0.418	2321	0.483
327	tons	DTR	4436	4119	−7.1%	4.22	<0.0005 *	1347	<0.0005 *
	value	SVR	364	355	−2.4%	4.55	<0.0005 *	1037	<0.0005 *
331	tons	GPR	1276	1244	−2.5%	7.74	<0.0005 *	701	<0.0005 *
	value	GPR	1298	1270	−2.2%	7.75	<0.0005 *	701	<0.0005 *
332	tons	SVR	692	615	−11.1%	5.62	<0.0005 *	989	<0.0005 *
	value	SVR	1150	1114	−3.1%	2.01	0.047 *	2323	0.487
333	tons	SVR	353	284	−19.6%	9.52	<0.0005 *	159	<0.0005 *
	value	SVR	1801	1782	−1.0%	2.17	0.032 *	1860	0.022 *
334	tons	GPR	43	42	−3.3%	6.20	<0.0005 *	889	<0.0005 *
	value	OLS	2181	-	-	-	-	-	-
335	tons	GBR	132	127	−3.6%	1.94	0.055	1910	0.034 *
	value	SVR	656	639	−2.6%	3.24	0.002 *	2418	0.713
336	tons	GPR	1483	1049	−29.3%	6.61	<0.0005 *	1018	<0.0005 *
	value	SVR	5910	5813	−1.6%	0.87	0.387	2143	0.189
337	tons	GPR	61	61	−0.3%	0.89	0.376	2508	0.953
	value	SVR	304	269	−11.6%	10.23	<0.0005 *	261	<0.0005 *
339	tons	GPR	39	38	−2.4%	6.02	<0.0005 *	856	<0.0005 *
	value	RFR	853	717	−16.0%	7.52	<0.0005 *	696	<0.0005 *
4231	tons	SVR	612	576	−5.9%	4.14	<0.0005 *	1496	<0.0005 *
	value	Lasso	2019	2018	0.0%	2.16	0.033 *	2233	0.315
4232	tons	OLS	178	-	-	-	-	-	-
	value	SVR	371	335	−9.8%	3.45	0.001 *	2323	0.487
4234	tons	GPR	168	165	−1.6%	2.28	0.025 *	1476	<0.0005 *
	value	SVR	1765	1697	−3.9%	3.98	<0.0005 *	1463	<0.0005 *
4236	tons	GPR	264	261	−1.4%	3.95	<0.0005 *	1536	0.001 *
	value	OLS	2458	-	-	-	-	-	-
4238	tons	SVR	861	849	−1.4%	2.34	0.021 *	1502	<0.0005 *
	value	SVR	1892	1817	−4.0%	4.53	<0.0005 *	1040	<0.0005 *
4241	tons	GPR	311	306	−1.9%	2.84	0.005 *	1502	<0.0005 *
	value	GPR	580	560	−3.5%	6.63	<0.0005 *	646	<0.0005 *
4242	tons	SVR	218	212	−3.1%	3.50	0.001 *	1819	0.015 *
	value	SVR	4406	3845	−12.7%	6.13	<0.0005 *	940	<0.0005 *
4244	tons	GPR	1151	1056	−8.3%	7.66	<0.0005 *	513	<0.0005 *
	value	GPR	1737	1639	−5.6%	6.49	<0.0005 *	628	<0.0005 *
4247	tons	GPR	15,045	14,784	−1.7%	2.26	0.026 *	1594	0.001 *
	value	GPR	7666	7344	−4.2%	4.15	<0.0005 *	1416	<0.0005 *
4541	tons	SVR	131	121	−7.4%	5.39	<0.0005 *	1185	<0.0005 *
	value	SVR	1596	1552	−2.8%	2.11	0.037 *	2210	0.279
4931	tons	GPR	1131	1014	−10.3%	6.12	<0.0005 *	863	<0.0005 *
	value	GPR	4358	3964	−9.0%	8.70	<0.0005 *	586	<0.0005 *

Table 6. *Cont.*

NAICS	Measure	Alternative	RMSE			t-Test		Wilcoxon	
			OLS	Alternative	%Dif.	Stat.	p-Value	Stat.	p-Value
5111	tons	GBR	43	40	−6.5%	1.98	0.051	2314	0.468
	value	GPR	204	201	−1.8%	4.63	<0.0005 *	1871	0.025 *
551114	tons	RFR	1012	984	−2.8%	3.93	<0.0005 *	1302	<0.0005 *
	value	SVR	1194	1157	−3.1%	7.04	<0.0005 *	609	<0.0005 *

(* p-value < 0.05)

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

As shown in Tables 5 and 6, about 57% of cases for estimating shipments by origins show a reduction of RMSE that are statistically significant, while 67% of estimating shipments by destinations show a statistically significant improvement. Overall, for the cases where the alternative ML methods bring a statistically significant improvement, the RMSE reduction is ranged from 0.1% to 30.6%.

6.3. Summary of Best Model by Industry

Table 7 summarizes the final model suggestion for each NAICS code, which was determined based on the significance tests on Tables 5 and 6. For each NAICS code and measurement, the final model algorithm along with its variable selection and use of log-transformation is provided.

Table 7. Final Freight Generation Model Selection.

NAICS	Measure	Shipments by Origins (Freight Production)						Shipments by Destinations (Freight Attraction)					
		Model	Log	*ESTAB*	*EMP*	*PAYANN*	*RCPTOT*	Model	Log	*ESTAB*	*EMP*	*PAYANN*	*RCPTOT*
212	tons	SVR	No	√	√			RFR	Yes	√	√		
	value	SVR	No	√	√	√		SVR	Yes			√	√
311	tons	GPR	Yes	√		√	√	GPR	No	√	√		
	value	SVR	Yes	√		√	√	SVR	No	√	√	√	√
312	tons	SVR	No		√		√	SVR	Yes			√	
	value	SVR	Yes	√	√	√	√	SVR	Yes	√	√		√
313	tons	OLS	No		√			GPR	Yes				√
	value	SVR	No	√		√		SVR	Yes		√		
314	tons	GPR	No	√		√	√	GPR	No	√			√
	value	SVR	No	√	√		√	GPR	No				√
315	tons	OLS	No				√	GPR	Yes	√			√
	value	OLS	No				√	GPR	Yes	√	√	√	√
316	tons	GPR	Yes	√				MLP	Yes	√	√		
	value	SVR	No			√	√	SVR	No		√	√	√
321	tons	RFR	No			√		SVR	Yes			√	√
	value	SVR	Yes	√		√		RFR	Yes	√			√
322	tons	SVR	Yes	√	√			SVR	No	√	√		
	value	SVR	No	√	√	√	√	GPR	No	√	√		
323	tons	SVR	Yes	√	√			GPR	Yes	√		√	
	value	SVR	Yes	√	√		√	SVR	Yes	√		√	
324	tons	SVR	No	√	√	√	√	GPR	No		√		
	value	SVR	No	√	√	√	√	GPR	No	√	√		

Table 7. *Cont.*

NAICS	Measure	Shipments by Origins (Freight Production)						Shipments by Destinations (Freight Attraction)					
		Model	**Log**	***ESTAB***	***EMP***	***PAYANN***	***RCPTOT***	**Model**	**Log**	***ESTAB***	***EMP***	***PAYANN***	***RCPTOT***
325	tons	GPR	Yes				√	GPR	No				√
	value	SVR	No	√			√	GPR	No				√
326	tons	SVR	No	√	√		√	SVR	No	√	√	√	√
	value	SVR	No	√	√			SVR	No	√	√		√
327	tons	SVR	No	√				DTR	Yes	√	√		
	value	SVR	Yes		√	√		SVR	No	√	√	√	√
331	tons	SVR	Yes		√	√	√	GPR	Yes			√	√
	value	SVR	No	√	√	√	√	GPR	Yes	√		√	√
332	tons	SVR	No			√	√	SVR	No		√	√	
	value	SVR	Yes				√	SVR	No	√			√
333	tons	SVR	No	√	√	√		SVR	No		√	√	
	value	SVR	No	√	√	√		SVR	No	√			√
334	tons	SVR	No			√		GPR	No	√	√	√	
	value	OLS	No			√		OLS	No	√			
335	tons	SVR	Yes	√	√		√	GBR	Yes			√	√
	value	OLS	No		√	√		SVR	No	√			
336	tons	OLS	Yes		√		√	GPR	No		√	√	√
	value	SVR	No	√	√	√	√	SVR	No	√	√		√
337	tons	SVR	No	√	√		√	GPR	Yes	√			√
	value	OLS	No		√	√	√	SVR	Yes	√	√	√	
339	tons	DTR	Yes		√	√		GPR	No		√	√	
	value	SVR	No	√		√	√	RFR	Yes	√			√
4231	tons	OLS	Yes		√			SVR	Yes	√	√	√	√
	value	OLS	Yes	√		√		Lasso	No	√	√		
4232	tons	SVR	No		√	√	√	OLS	Yes		√	√	
	value	SVR	No		√	√		SVR	No	√			√
4233	tons	GBR	No	√	√		√	N/A					
	value	SVR	No		√	√	√						
4234	tons	SVR	Yes		√	√		GPR	Yes	√	√	√	
	value	OLS	Yes	√	√			SVR	No	√	√		√
4235	tons	SVR	No		√	√		N/A					
	value	OLS	No			√							
4236	tons	Lasso	Yes	√				GPR	Yes	√			√
	value	SVR	No	√	√	√	√	OLS	Yes	√			
4237	tons	OLS	Yes		√		√	N/A					
	value	SVR	No	√	√		√						
4238	tons	SVR	Yes	√			√	SVR	Yes	√	√	√	
	value	Lasso	Yes		√			SVR	Yes	√	√	√	
4239	tons	Lasso	Yes	√			√	N/A					
	value	Lasso	No	√	√								
4241	tons	OLS	Yes			√		GPR	Yes	√	√		
	value	OLS	Yes			√	√	GPR	Yes		√		√
4242	tons	SVR	No	√				SVR	Yes			√	√
	value	OLS	Yes	√	√			SVR	Yes	√	√		√
4243	tons	OLS	No			√	√	N/A					
	value	SVR	No		√		√						
4244	tons	OLS	No		√			GPR	Yes	√	√		
	value	OLS	No			√	√	GPR	Yes	√		√	
4245	tons	RFR	Yes	√	√			N/A					
	value	DTR	No	√	√								
4246	tons	OLS	Yes			√		N/A					
	value	OLS	No			√							

Table 7. *Cont.*

NAICS	Measure	Shipments by Origins (Freight Production)						Shipments by Destinations (Freight Attraction)					
		Model	**Log**	***ESTAB***	***EMP***	***PAYANN***	***RCPTOT***	**Model**	**Log**	***ESTAB***	***EMP***	***PAYANN***	***RCPTOT***
4247	tons	OLS	Yes		✓			GPR	Yes			✓	✓
	value	SVR	Yes			✓	✓	GPR	Yes			✓	✓
4248	tons	OLS	No		✓			N/A					
	value	SVR	No	✓	✓	✓							
4249	tons	SVR	Yes	✓	✓	✓	✓	N/A					
	value	SVR	Yes			✓	✓						
4541	tons	SVR	Yes	✓	✓			SVR	No	✓	✓		
	value	SVR	Yes	✓	✓		✓	SVR	No	✓	✓	✓	
4931	tons	SVR	Yes		✓			GPR	No	✓			✓
	value	SVR	Yes		✓			GPR	No	✓			✓
5111	tons	SVR	Yes	✓		✓		GBR	Yes	✓			
	value	SVR	Yes	✓		✓		GPR	Yes	✓		✓	✓
45431	tons	SVR	No		✓	✓		N/A					
	value	OLS	No		✓								
551114	tons	RFR	Yes			✓		RFR	Yes			✓	
	value	DTR	Yes	✓				SVR	Yes	✓	✓		

(✓: the variable is included in the final model)

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

Overall, as shown in Table 7, the SVR was selected as the best model for 52 NAICS tonnage/value cases (58%) for the estimation of shipments by origins (freight generation). For the estimation of shipments by destinations (freight attraction), both the SVR and GPR were selected as the best model for 29 cases (41%) each.

The OLS, selected only when none of the seven alternative ML algorithms showed the significant reduction in RMSE, was selected for 23 cases (26%) for the estimation of shipments by origins (freight generation) and for only 3 cases (4%) for the estimation of shipments by destinations (freight attraction). The MLP, which can be arguably considered as the most complex model among the eight models, was selected for only one case for estimating tonnage of shipments by destinations for NAICS 316.

In terms of the variable selection, the number of employee (*EMP*) was included most, 54 cases (60%), for estimating shipments by origins (freight generation). For the estimation of shipments by destinations (freight attraction), the number of establishment (*ESTAB*) was included most, 48 cases (69%), in the final model selection. In addition, the results show that the log-transform would improve the overall model performance for 41 cases (46%) in the estimation of shipments by origins and for 39 cases (56%) in the estimation of shipments by destinations. In addition, the receipt total (*RCPTOT*), which was not considered in any of the referenced study, was included in 49% of the final models. Note that this is only a summary of how many times each variable is selected for all the 45 NAICS codes. As discussed in Sections 5.4 and 5.6, the variable selection was determined by RMSE of validation sets, considering all possible combinations.

6.4. Discussions in Model Interpretability

Oftentimes, a regression-based modeling approach can be explained with explicit equational forms to represent the relationship between the dependent variable and the independent variables. The following two equations, by OLS and Lasso, show the example of final model for estimating values of shipments by origins for NAICS 4239. Note that the estimated coefficient for the number of employee variable (*EMP*) is smaller in the Lasso regression (1.071) than the same coefficient estimate in the OLS regression (1.096).

$$\text{OLS Regression}: \widehat{Value_{4238}} = \mathbf{exp(-0.058)} \cdot \boldsymbol{EMP_{4238}}^{\mathbf{1.096}} \tag{6}$$

$$\text{Lasso}: \widehat{Value_{4238}} = \mathbf{exp(-0.111)} \cdot \boldsymbol{EMP_{4238}}^{\mathbf{1.071}} \tag{7}$$

This straightforward interpretability is one of the clear advantages for utilizing simple regression-based modeling approaches, such as OLS and Lasso. However, one can choose more complex models with higher model performance if the model performance (the focus of this study) is more important for their applications. In addition, such complex models could still provide insights of which factors are affecting more on the tonnage and value of shipments by exploring the variable importance. For example, Figure 2 shows the variable importance for the Support Vector Regression (SVR) model that estimates value of shipments by origins for NAICS 322. The variable importance was calculated by permutation feature imputation technique, where we measured the decreased R-squared value by randomly shuffling a single feature value. In this case, the annual payroll (*PAYANN*) appears to be impacting the most to the model estimates.

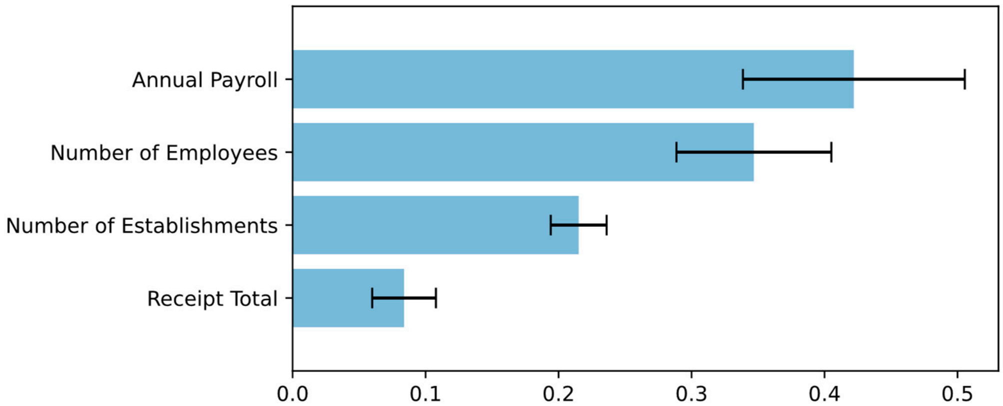


Figure 2. Example of Variable Importance in SVR—Value of Shipments by Origins for NAICS 322.

7. Conclusions

This study explored eight models, i.e., Ordinary Least Square (OLS) regression, Lasso, Decision Tree, Random Forest, Gradient Boosting, Support Vector, Gaussian Process, and Multi-layer Perceptron regressions, applied for the FG models by industry type (NAICS code). The seven alternative ML algorithms, which have been commonly used for regression but not often in FG modeling, were evaluated whether the model performance improvement is significant over the OLS. Overall, the Support Vector regression was selected most as the best model approach for the estimation of shipments by origins, while both the Support Vector regression and the Gaussian Process regression were equally selected most as the best model approach for the estimation of shipments by destinations. Combining all the cases of shipments by origins and destinations, the RMSE reductions (compared to OLS) for 134 cases (84%) are, ranged from 0.1% to 30.6%, statistically significant with both paired t-test and Wilcoxon statistics.

The following summarizes the key contributions of this study:

- Built a framework to conduct the industry-specific model selection, i.e., the variation selection, log-transform, and algorithm.
- Evaluated the significance of model improvements when using the alternative ML algorithms over the OLS for the FG modeling.
- Suggested the use of OLS regression for certain NAICSs if the RMSE reductions by the alternative ML algorithms are not statistically significant.

- Considered all possible variable combinations from the four variables in the CBP and EC data tables.
- Covered all the NAICS codes from the 2017 CFS data and estimated tonnage/value of freight shipments by both origins (generation) and destinations (attraction).

Although the study focused on model performance in applying ML algorithms for the FG models, simplicity and interpretability of model approaches could be more important depending on their applications. This is one of the main reasons why alternative ML algorithm is being selected over OLS only when the improvement is "statistically significant". Note that most of complex ML models may not be provided with explicit equational forms, but their variable importance can be still obtained, as discussed in Section 6.4. (Discussions in Model Interpretability).

The scope of this study is limited to estimating tonnage and value of the freight shipments by industry type (NAICS codes). The proposed model selection results could be quite different when different dependent variables, such as truck volume and number of shipments, are to be estimated.

Furthermore, there can be more variables, such as population, GDP, access to ports, network access/length by mode, land use, etc., to be considered to improve model performance depending on industry types and data availability. Additionally, note that not all hyperparameters were evaluated for each ML algorithm, meaning that there may be potential further improvements with hyperparameter settings not considered in this study. The authors expect that more complex algorithms, such as Random Forest, Gradient Boosting, and Multi-layer Perceptron regressions, are more likely to outperform the OLS with larger size of training data (e.g., the data at the establishment level or more granular level of geography). With all, the authors believe that the future research in FG modeling can be focused on the following areas:

- Applying the proposed framework with use case of disaggregating freight data into more granular level of geography (e.g., county-level freight data).
- Using other external/private data sources to reveal the relationship between economy activity and associated freight shipments at individual business level.
- Expanding the model framework to forecasting future freight demand by industry type.

Author Contributions: Conceptualization, H.L., M.U., S.-M.C. and H.-L.H.; Data curation, H.L.; Formal analysis, H.L. and M.U.; Methodology, H.L.; Writing—original draft, H.L., M.U. and Y.L.; Writing—review and editing, H.L., M.U. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research effort was sponsored by the Federal Highway Administration (FHWA) and the Bureau of Transportation Statistics (BTS), under U.S. Department of Transportation, through the project titled "Design and Development of Statistical Models and Freight Data", grant number 2116-Z239-18.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1 summarizes the description of 45 North American Industry Classification System (NAICS) codes used in this study.

Table A1. Description of 45 NAICS Codes Used in the Study.

NAICS Code	Description
212	Mining (except oil and gas)
311	Food manufacturing
312	Beverage and tobacco product manufacturing
313	Textile mills
314	Textile product mills
315	Apparel manufacturing
316	Leather and allied product manufacturing
321	Wood product manufacturing
322	Paper manufacturing
323	Printing and related support activities
324	Petroleum and coal products manufacturing
325	Chemical manufacturing
326	Plastics and rubber products manufacturing
327	Nonmetallic mineral product manufacturing
331	Primary metal manufacturing
332	Fabricated metal product manufacturing
333	Machinery manufacturing
334	Computer and electronic product manufacturing
335	Electrical equipment, appliance, and component manufacturing
336	Transportation equipment manufacturing
337	Furniture and related product manufacturing
339	Miscellaneous manufacturing
4231	Motor vehicle and motor vehicle parts and supplies merchant wholesalers
4232	Furniture and home furnishing merchant wholesalers
4233	Lumber and other construction materials merchant wholesalers
4234	Professional and commercial equipment and supplies merchant wholesalers
4235	Metal and mineral (except petroleum) merchant wholesalers
4236	Household appliances and electrical and electronic goods merchant wholesalers
4237	Hardware, plumbing and heating equipment and supplies merchant wholesalers
4238	Machinery, equipment, and supplies merchant wholesalers
4239	Miscellaneous durable goods merchant wholesalers
4241	Paper and paper product merchant wholesalers
4242	Drugs and druggists' sundries merchant wholesalers
4243	Apparel, piece goods, and notions merchant wholesalers
4244	Grocery and related product merchant wholesalers
4245	Farm product raw material merchant wholesalers
4246	Chemical and allied products merchant wholesalers
4247	Petroleum and petroleum products merchant wholesalers
4248	Beer, wine, and distilled alcoholic beverage merchant wholesalers
4249	Miscellaneous nondurable goods merchant wholesalers
4541	Electronic shopping and mail-order houses
45431	Fuel dealers
4931	Warehousing and storage
5111	Newspaper, periodical, book, and directory publishers
551114	Corporate, subsidiary, and regional managing offices

References

1. U.S. Department of Transportation. Bureau of Transportation Statistics and Federal Highway Administration, Freight Analysis Framework Version 5.4 (FAF5). Available online: https://www.bts.gov/faf (accessed on 25 October 2022).
2. U.S. Department of Transportation. Bureau of Transportation Statistics and U.S. Department of Commerce, U.S. Census Bureau. 2017 Commodity Flow Survey. Available online: https://www2.census.gov/programs-surveys/cfs/data/2017 (accessed on 1 August 2022).
3. Holguin-Veras, J.; Sarmiento, I.; Gonzalez-Calderon, C.A. Parameter Stability in Freight Generation and Distribution Demand Models in Colombia. *Dyna* **2011**, *78*, 16–20.
4. Lim, R.; Qian, Z.S.; Zhang, H.M. Development of a Freight Demand Model with an Application to California. *Int. J. Transp. Sci. Technol.* **2014**, *3*, 19–38. [CrossRef]
5. Oliveira-Neto, F.M.; Chin, S.M.; Hwang, H.L. Aggregate Freight Generation Modeling: Assessing Temporal Effect of Economic Activity on Freight Volumes with Two-Period Cross-Sectional Data. *Transp. Res. Rec.* **2012**, *2285*, 145–154. [CrossRef]

6. Krisztin, T. Semi-Parametric Spatial Autoregressive Models in Freight Generation Modeling. *Transp. Res. Part E Logist. Transp. Rev.* **2018**, *114*, 121–143. [CrossRef]
7. Hagenauer, J.; Helbich, M. A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Syst. Appl.* **2017**, *78*, 273–282. [CrossRef]
8. Uddin, M.; Anowar, S.; Eluru, N. Modeling Freight Mode Choice Using Machine Learning Classifiers: A Comparative Study Using Commodity Flow Survey (CFS) Data. *Transp. Plan. Technol.* **2021**, *44*, 543–559. [CrossRef]
9. Iranitalab, A.; Khattak, A. Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction. *Accid. Anal. Prev.* **2017**, *108*, 27–36. [CrossRef] [PubMed]
10. Rahman, S.; Bhasin, A.; Smit, A. Exploring the Use of Machine Learning to Predict Metrics Related to Asphalt Mixture Performance. *Constr. Build. Mater.* **2021**, *295*, 123585. [CrossRef]
11. Salais-Fierro, T.; Martínez, A. Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution. *Mob. Netw. Appl.* **2022**, *27*, 2172–2181. [CrossRef]
12. Chin, S.M.; Hwang, H.L. National Freight Demand Modeling: Bridging the Gap Between Freight Flow Statistics and US Economic Patterns. In Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 21–25 January 2007.
13. Novak, D.C.; Hodgdon, C.; Guo, F.; Aultman-Hall, L. Nationwide Freight Generation Models: A Spatial Regression Approach. *Netw. Spat. Econ.* **2011**, *11*, 23–41. [CrossRef]
14. Bagighni, S. Volume Estimation Models for Generation and Attraction of Freight Commodity Groups Using Regression Analysis. Ph.D. Dissertation, The University of Alabama in Huntsville, Huntsville, AL, USA, 2012.
15. Ha, D.H.; Combes, F. Building a Model of Freight Generation with a Commodity Flow Survey. In *Commercial Transport*; Springer International Publishing: Cham, Switzerland, 2016; pp. 23–37.
16. Mommens, K.; Van Lier, T.; Macharis, C. Freight Demand Generation on Commodity and Loading Unit Level. *Eur. J. Transp. Infrastruct. Res.* **2017**, *17*, 1. [CrossRef]
17. National Academies of Sciences, Engineering, and Medicine. *NCFRP Report 37: Using Commodity Flow Survey Microdata and Other Establishment Data to Estimate the Generation of Freight, Freight Trips, and Service Trips: Guidebook*; Transportation Research Board: Washington, DC, USA, 2016.
18. U.S. Department of Commerce, U.S. Census Bureau. 2017 Economic Census Data. Available online: https://www.census.gov/programs-surveys/economic-census/year/2017/economic-census-2017/data.html (accessed on 1 August 2022).
19. U.S. Department of Commerce, U.S. Census Bureau. 2017 County Business Patterns. Available online: https://www.census.gov/data/datasets/2017/econ/cbp/2017-cbp.html (accessed on 1 August 2022).
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *JMLR* **2011**, *12*, 2825–2830.