Project Final Report

Data-driven Supply Chain Strategy

by
Team-027 Fall 2022 MGT-6203

Shreya.
Amber Kishore
Vijayshree Chile
Konstantinos Vrachimis
Rachel LaVictoire

Georgia Institute of Technology
Nov 2022

Abstract

In today’s world of express deliveries and hyperlocals, online business is thriving on the web of efficient supply chain network. Optimizing Supply chain is crucial for companies to sustain and expand their business. A poorly designed supply chain could result in loss of productivity, customer complaints, increased costs, loss of revenue and damage to the brand.

Designing a supply chain network in a business involves a lot of factors spanning from product, market, process, technology, costs, external environment and business specific factors like sales, demand forecast etc.

We would like to design supply chain network such that we are focusing on our most profitable markets and are able to maintain optimum stock in that market distribution centers to ensure timely and cost-effective delivery of products to our customers and maximize profitability for our company.

Background / Description of Problem

Supply chain network in simple and basic terms involves determining the following process design:

1. Where are your suppliers?
2. Where will you locate the factories for manufacturing/assembling?
3. Where will you hold inventories, number of warehouses, location of warehouses or stores etc.
4. Transportation and distribution logistics.

Choosing which existing or new market to focus on is also an important criterion in designing the supply chain strategy for a company. There are many factors that help in concluding this.

1. Profitability in market segment
2. Demand forecasting
3. Procurement costs
4. Logistics cost including inventory holding costs
5. Overheads

Factors that are crucial for modeling a supply chain network are: -

1. Government policies
2. Taxation policies
3. Infrastructure status etc.

All the above factors are influenced by one key driver – ***Customer fulfillment.*** You have a happy and satisfied customer if he/ she is getting the desired product in the expected time in a good condition.

Supply chain network designs provide an operating framework that acts as a guiding backbone for management and strategic decisions considering external influences, interdependencies of all processes and factors. It gives management a provision to evaluate these interdependencies and find opportunities to maximize profitability.

Assumptions:

1. Procurement is already optimized and there are minimal delays
2. Manufacturing methodology is already optimized
3. Our stores are designed for omnichannel retail

Research Questions:

1. Identify the most profitable markets based on the maximum sales or maximum profit generated. We would like to focus both on max footfall and max revenue to tailor strategies for these markets.
2. Identify if there are customer fulfilment issues (late deliveries, stockouts) in the markets identified above.
3. For identified customer fulfilment issues, identify if the issues are caused by distance between source and destination or because of operational inefficiencies. Distance of fulfilment will help devise network configurations and Operation inefficiency will help devising contractual negotiations with delivery vendor or onboarding new vendors.
4. Identify the product category mix driving the profit and sales in the focus markets. This along with the information identified in point 3 will help in strategizing which products and categories to focus on per market.
5. Lastly, identify external environmental factors eg: infrastructure status, trade facilitation data (e.g.: Taxes or policies) and use these factors in strategizing the network configuration and 3-5 years plans for market growth & expansion.

Initial Hypothesis

With the initial maneuvering of data and a preliminary data analysis, it seems that company's supply chain network configuration is more centralized (where in merchant locations are concentrated across one region that are shipping products across the world). Our initial hypothesis is that company would benefit more from implementing a decentralized supply chain network configuration (where in merchants/ suppliers are strategically located in several regions).

Method

Data sets

We collected the data from following sources -

1. Kaggle dataset includes a set of retailer orders to various locations across the world. This is a company level dataset [1]
2. The World Bank dataset includes information from all countries in the world regarding country level characteristics across various sections of the economy. This data is a combination of trade facilitation, infrastructure, trade index and ease of doing business data. [2]

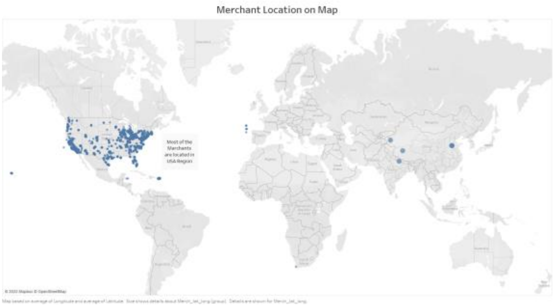
Exploratory data analysis

Data Cleaning and Dataset preparation –

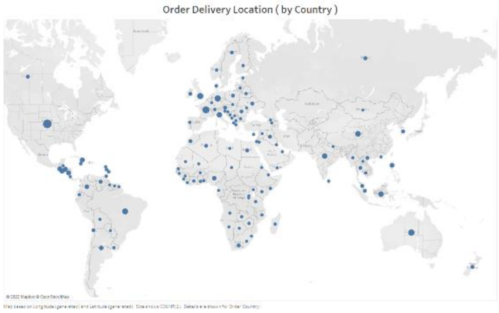
- Two datasets (Supply chain bigdata and WDI (World Development Indicators) data from World bank) are merged.
- The Supply Chain Dataset provide the Country name in Spanish language (as opposed to English in the World Bank Dataset) and the team created a new column of the names in English in order the merge process to proceed.
- Used reverse geo code to fetch country/state for given latitude/longitude
- Removed special characters from column names
- Converted relevant columns to date format
- Imputed NA for all columns having special characters ("..")

Graphical Analysis – we performed graphical analysis to understand the data.

1. <u>Merchant and Customer location to understand supply chain network config</u>:


Merchant Location on Map
Most of the Merchants are located in USA Region
© 2022 Mapbox © OpenStreetMap

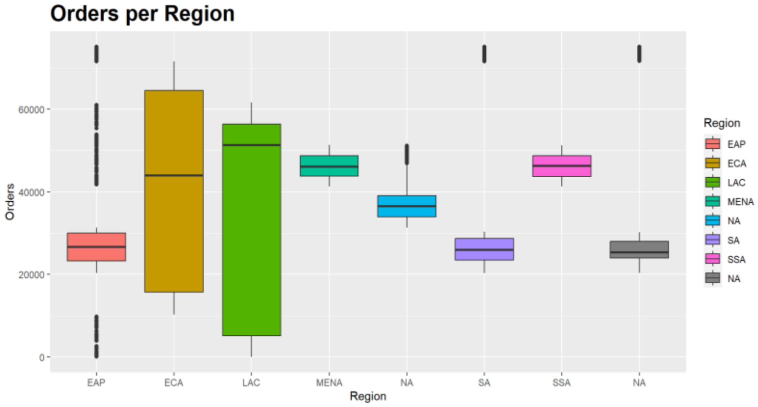
Map based on average of Longitude and average of Latitude. Size shows details about Merch_lat_long (group). Details are shown for Merch_lat_long.

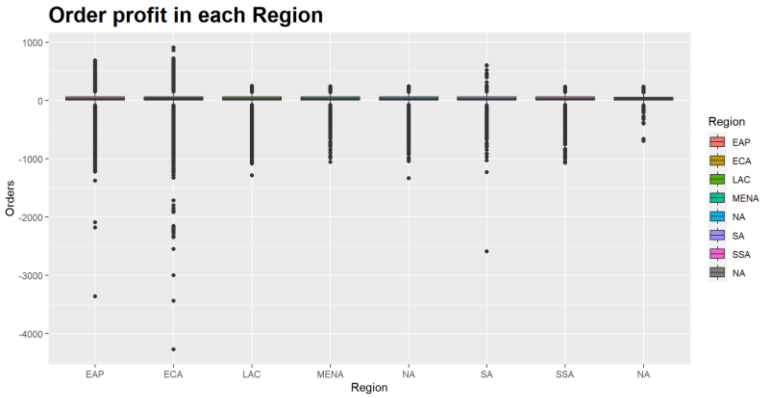

Order Delivery Location (by Country)
© 2022 Mapbox © OpenStreetMap

Map based on Longitude (generated) and Latitude (generated). Size shows COUNT(1). Details are shown for Order Country.

The above two maps clearly confirms our initial hypothesis that network configuration is more centralized while customers are located worldwide.

2. <u>Identify most profitable / potentially most profitable markets</u>:

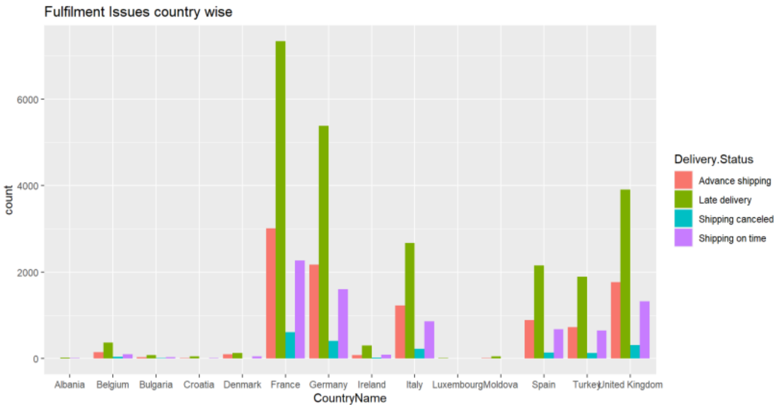


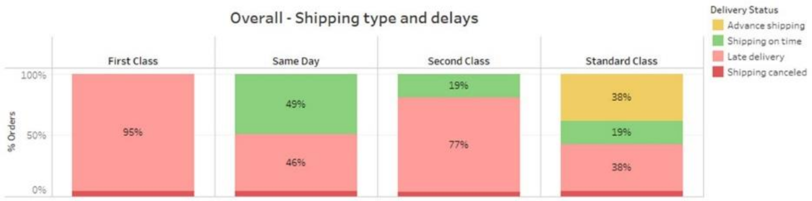


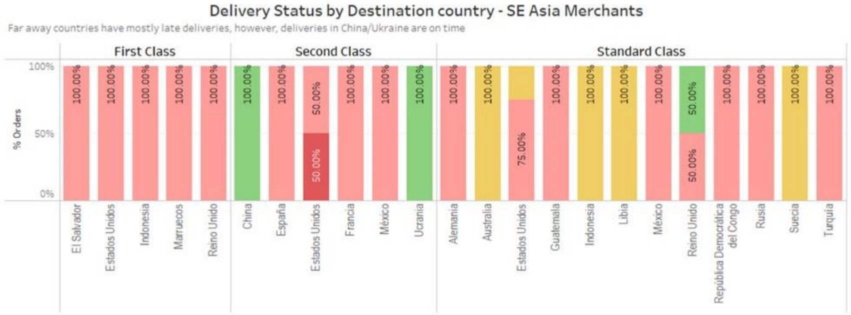
Above graphs helped us in narrowing down on our analysis. ECA region has the maximum profit by revenue and also marks the maximum sales. So company will focus on this market for implementing the strategy first and then will phase out to other regions.

3. <u>Fulfilment Issues – to understand fulfilment type and shipping delays</u> –

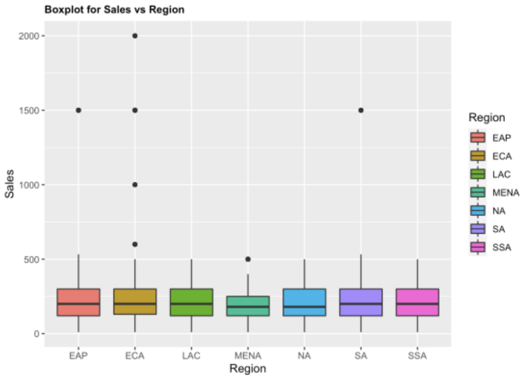
   - Identified fulfilment issues in our focus region.
   - Identified the delivery type (early, late, on-time arrival) split by "Shipping Mode"
   - Also validated our initial hypothesis: - % early / on-time arrival orders are higher **when merchant location & order location is in nearby regions**. Consecutively when this distance is high, orders are generally late.

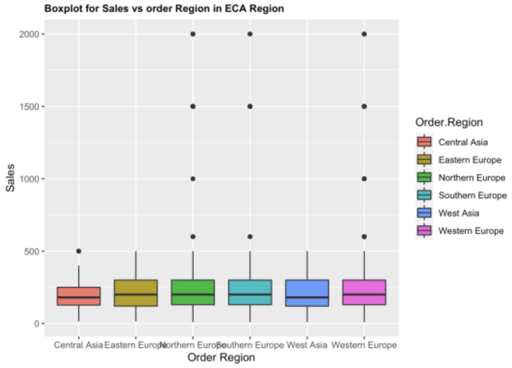






4. Sales





Feature Engineering

1. Feature selection: We have used backward and forward elimination method to identify insignificant predictors. Based on the result we eliminated those insignificant predictors.
2. Feature extraction: Some of the important predictors e.g. distance was not directly available in the dataset. We have extracted it from data using geo tagging on order delivery region provided in the dataset. (we get latitude, longitude for delivery location using geocode package, then we use **haversine dist package** to get distance between merchant and customer)

   geocode (CountryName, method = 'osm', lat = latitude , long = longitude)
   distHaversine(c(x$Longitude, x$Latitude), c(x$longitude_2, x$latitude_2), r = 6378.137)

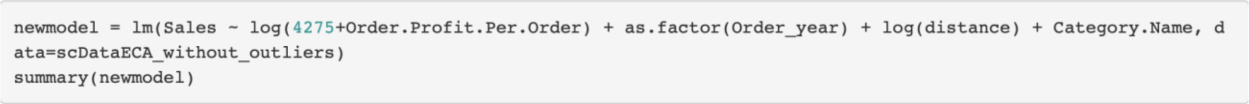
3. Feature transformation: Upon checking the relationship between predictor and response variable, we noticed the non-linear relationship. The values of explanatory variables vary and to reduce the possible heteroskedasticity effects in the model estimation, we have applied log transformation on a few predictors to improve the model fitment. For two variables, we added

to all values the (minimum value of each variable +1) since they contain negative, or zero values and we want to avoid the creation of missing values.

Analytical Modeling

We have used a combination of **multiple linear** and **logistic regression** models to devise our supply chain strategy. For linear regression models we have used **Adjusted R squared** as measure to model fitment and for Logistic we are using **AIC value** as choosing criteria. Detailed **model analysis** is also performed to identify and understand how model stands with outliers, heteroskedasticity etc.

1. **Sales Model:** This model is built to identify product categories driving sales in our focus markets. We started with simple linear regression model and use a mix of variable addition & backward elimination to get the final model. We achieved an adjusted R squared value of 79% in this model. Noticed through our model analysis that there are outliers affecting the model fitment and eventually with outliers removal, we achieved the adjusted R squared value of 2%.

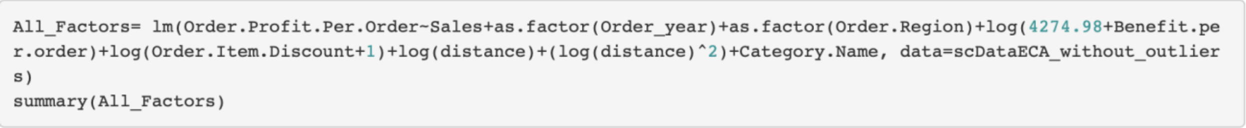


```
newmodel = lm(Sales ~ log(4275+Order.Profit.Per.Order) + as.factor(Order_year) + log(distance) + Category.Name, d
ata=scDataECA_without_outliers)
summary(newmodel)
```

For result see Appendix – Table 1

Product categories for which company should plan local inventory in focus market 'ECA' are: - Computers, Strength training equipment, Cardio equipment, Camping & Hiking equipment, Fishing gears and Crafts materials.

2. **Order Profit Model:** This multiple linear regression model is built to identify product categories driving maximum revenue generation in our focus market. With outlier's removal, we achieved adjusted R squared value of 97%.

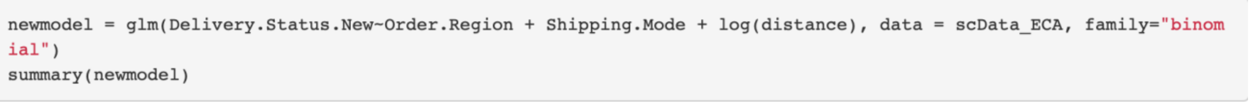


```
All_Factors= lm(Order.Profit.Per.Order~Sales+as.factor(Order_year)+as.factor(Order.Region)+log(4274.98+Benefit.pe
r.order)+log(Order.Item.Discount+1)+log(distance)+(log(distance)^2)+Category.Name, data=scDataECA_without_outlier
s)
summary(All_Factors)
```

For result see Appendix - Table 2

This model helped us in identify an additional product category i.e. Golf equipment that company could benefit by maintaining local inventory in ECA region.

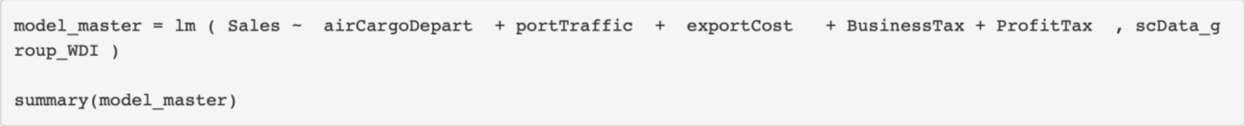
3. **Identify customer fulfilment issues by logistic regression:** This model helps us in confirm that distance has a direct bearing on the fulfilment issues. Region that have greater distance from merchant locations are more likely to receive deliveries late. We tried logistic regression models with different AIC values and chose the one with least AIC value (better model fitment).



```
newmodel = glm(Delivery.Status.New~Order.Region + Shipping.Mode + log(distance), data = scData_ECA, family="binom
ial")
summary(newmodel)
```

For result see Appendix – Table 3

4. **Identify macroeconomic factors impacting supply chain:** Lastly, we built a multiple linear regression model from our merged dataset to identify macro-economic factors that might impact supply chain strategy in our focus market and countries. We used backward elimination method to discard the insignificant predictors and eventually identified Business tax in the country, airCargo traffic and portTraffic as major significant contributing factors enabling ease of doing business in a country.



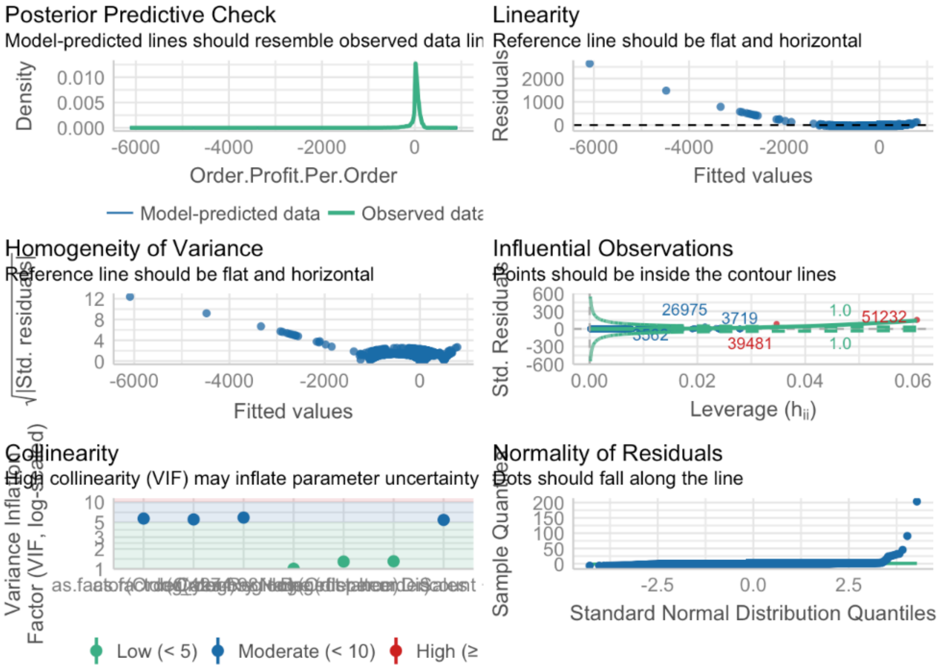
```
model_master = lm ( Sales ~  airCargoDepart  + portTraffic  +  exportCost   + BusinessTax + ProfitTax  , scData_g
roup_WDI )

summary(model_master)
```

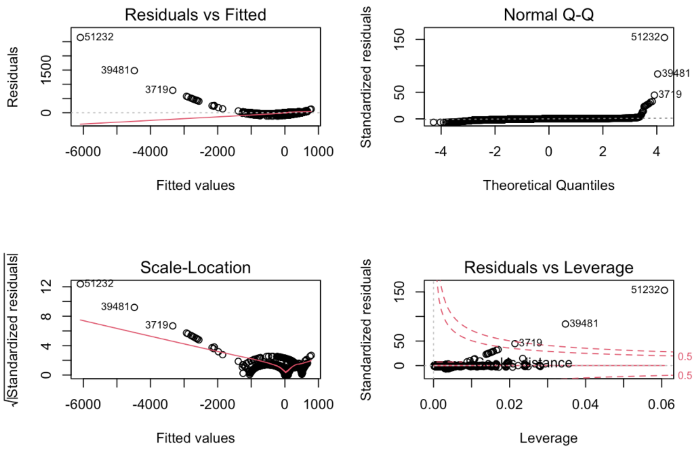
This model helped us in identifying that out of 5 focus countries in our focus market ECA, tax is Spain is quite conducive to consider it as a market where company should tie up with 3PL warehouses and distribution centers to enable local inventory for nearby countries as well.
For result see Appendix – Table 4

Challenges faced:

We faced major issues while data cleaning process because both our datasets were big and were procured from different sources. However, we tried multiple non-linear transformations and arrived at the ones (log transformation in predictors) that improved model predictability.
One of the major challenges we faced was to build a model that holds the linear regression assumptions. We were not successful with achieving model assumptions.

Our model analysis gave us a good understanding of how our models performed and how our data is distributed. As an example, see below the analysis for Order profit model. Similar analysis has been done for all the 4 models.





We have **improved the model predictability** substantially from removing the outliers.

Overall Conclusion:

It is important for an organization to execute any change in strategy in a phased manner such that utilization of resources and capital can be optimized and our models in this project have significantly helped us in determining a strategy based on cost-benefit analysis. Since converting to a decentralized configuration is a significant investment, it makes sense to operate in a market which has potential growth and is driving maximum revenue for the company. Our models have helped in narrowing down to our focus market (**ECA region – countries France, Germany, Italy, Spain and UK**), identifying customer fulfilment issues (**Late deliveries because of distance between merchant and customer**) and the macro-economic factor analysis which indicated that company should make its first move in Spain (since ease of doing business is better there owing to its low business tax and high air cargo traffic).

Data analysis is used by a lot of companies to come up with their strategies. The novelty of our project is that we merged the micro-factors and macro-economic factors to come to a joint conclusion. Had we not considered world bank data; we would not be able to narrow it down to Spain. We would have 5 countries in focus with no evidential data to support choosing a country **for 3 PL warehouse and distribution-center setup for more regionally localized deliveries.**

We are not only harnessing the power of big data but also moving in Top down fashion to put emphasis on “what is required now” than “what all options are available”. Having a close collaboration of Business and technology makes our idea and project novel and interesting. Moreover, these models could be used in analysis of any generic online retailer because the focus in more on identifying and understanding the problem via data analysis and for solving it we have purely used business and domain understanding.

If we had more time, we could also determine the optimum inventory levels that should be maintained in each market for regionally localized deliveries. We could use Time Series/ ARIMA models for this short term demand planning and analysis through our Supply chain bigdata.

Citations & References :

- https://www.accessengineeringlibrary.com/binary/mheaeworks/52b0dfcc1cd50117/f83cbe3062ecdf94f8e5df542c84af0945284dfc068065142bc65a1dabaf4e61/book-summary.pdf
- Chiles, C. R., & Dau, M. T. (2005). *An analysis of current supply chain best practices in the retail industry with case studies of Wal-Mart and Amazon. com* (Doctoral dissertation, Massachusetts Institute of Technology). (Case study)
- Qi Li, Ang Liu, Big Data Driven Supply Chain Management, Procedia CIRP, Volume 81, 2019, Pages 1089-1094, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2019.03.258. (https://www.sciencedirect.com/science/article/pii/S2212827119305633)
- Croissant, Yves, and Giovanni Millo. 2008. "Panel Data Econometrics in : The Package." *Journal of Statistical Software* 27 (2): 1–43.
- Wooldridge, J. M., 2011. "*Econometric Analysis of Cross Section and Panel Data.*" 2nd ed., MIT Press

Inspiration and Research:

- MGT 8803 course in GT OMSA program – Supply Chain module
- https://www.managementstudyguide.com/supply-chain-network-design.htm#google_vignette
- https://ware2go.co/the-5-data-sets-you-need-to-effectively-leverage-3rd-party-logistics-fulfillment-in-2020/
- https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis
- For macro factors data - https://data.worldbank.org/indicator

GitHub Repository :
https://github.gatech.edu/MGT-6203-Fall-2022-Canvas/Team-27.git

1. Code Files – folder contains all codes for data clean and model analysis
2. Data Files - folder contains data files around which the analysis is positioned

Appendix – Model result

Table 1 - Sales order model :

	Dependent variable:	
	Sales	
log(4275 + Order.Profit.Per.Order)	106.442***	-9.525
as.factor(Order_year)2016	1.297	-0.934
as.factor(Order_year)2017	4.364***	-0.639
log(distance)	-2.915*	-1.621
Category.NameAs Seen on TV!	219.401***	-13.137
Category.NameBaby	-20.982***	-7.864
Category.NameBasketball	219.299***	-9.187
Category.NameBooks	-48.829***	-4.912
Category.NameCameras	371.135***	-4.216
Category.NameCardio Equipment	217.082***	-3.354
Category.NameCDs	-68.585***	-6.491
Category.NameChildren's Clothing	276.365***	-4.305
Category.NameCleats	103.360***	-3.272
Category.NameComputers	1,418.277***	-4.744
Category.NameConsumer Electronics	172.419***	-5.028
Category.NameCrafts	380.304***	-5.357
Category.NameDVDs	84.484***	-5.482
Category.NameElectronics	43.451***	-3.99
Category.NameFishing	321.780***	-3.308
Category.NameFitness Accessories	34.462***	-8.775
Category.NameGirls' Apparel	47.244***	-4.952
Category.NameGolf Apparel	17.485***	-6.058
Category.NameGolf Balls	-24.311***	-4.73
Category.NameGolf Gloves	74.014***	-4.847
Category.NameGolf Shoes	126.610***	-5.746
Category.NameHockey	35.608***	-5.997
Category.NameIndoor/Outdoor Games	73.429***	-3.298
Category.NameKids' Golf Clubs	190.945***	-5.11

Category.NameLacrosse	23.922***	-7.765
Category.NameMen's Footwear	52.473***	-3.28
Category.NameMen's Golf Clubs	72.413***	-6.107
Category.NameShop By Sport	42.199***	-3.38
Category.NameSoccer	89.843***	-8.074
Category.NameStrength Training	320.268***	-11.924
Category.NameTrade-In	16.203***	-5.295
Category.NameWater Sports	123.545***	-3.323
Category.NameWomen's Apparel	71.890***	-3.285
Category.NameWomen's Golf Clubs	169.042***	-7.195
Constant	-788.499***	-80.943
Observations	54,098	
R^2	0.824	
Adjusted R^2	0.824	
Residual Std. Error	64.929 (df = 54053)	
F Statistic	5,753.836*** (df = 44; 54053)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2 - Profit order model :

Results

	Dependent variable:			
	Order.Profit.Per.Order			
	(1)	(2)	(3)	(4)
as.factor(Order_year)2016	-0.459	5.374		0.322
	(1.514)	(3.297)		(0.517)
as.factor(Order_year)2017	4.365***	4.634***		0.003
	(1.055)	(1.064)		(0.177)
log(distance)		2.984		0.257
		(2.969)		(0.464)
as.factor(Order.Region)Eastern Europe		-1.837		-0.245
		(5.812)		(0.908)
as.factor(Order.Region)Northern Europe		5.965		0.191

		(6.388)		(0.999)
as.factor(Order.Region)Southern Europe		5.838		0.392
		(6.330)		(0.990)
as.factor(Order.Region)West Asia		-1.972		-0.199
		(5.820)		(0.910)
as.factor(Order.Region)Western Europe		4.544		0.246
		(6.288)		(0.984)
Sales			0.027***	0.027***
			(0.001)	(0.001)
log(4274.98 + Benefit.per.order)			3,772.506***	3,772.482***
			(2.589)	(2.589)
log(Order.Item.Discount + 1)			-0.490***	-0.491***
			(0.077)	(0.077)
Category.NameAs Seen on TV!			-1.008	-1.037
			(2.893)	(2.898)
Category.NameBasketball			4.763**	4.737**
			(2.370)	(2.375)
Category.NameComputers			45.366***	45.334***
			(2.021)	(2.024)
Category.NameWomen's Golf Clubs			-1.415	-1.440
			(1.940)	(1.946)
Constant	21.451***	-10.223	-31,540.050***	-31,542.380***
	(0.722)	(28.210)	(21.665)	(22.087)
Observations	54,242	54,242	54,242	54,242
R^2	0.0004	0.0005	0.976	0.976
Adjusted R^2	0.0003	0.0004	0.976	0.976
Residual Std. Error	113.971 (df = 54239)	113.970 (df = 54233)	17.808 (df = 54198)	17.809 (df = 54190)
F Statistic	10.102*** (df = 2; 54239)	3.387*** (df = 8; 54233)	50,426.960*** (df = 43; 54198)	42,512.280*** (df = 51; 54190)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3 - Customer fulfillment issue model

Results	
	Dependent variable:

	Delivery.Status.New
Order.RegionEastern Europe	-0.065
	(0.113)
Order.RegionNorthern Europe	-0.055
	(0.112)
Order.RegionSouthern Europe	-0.059
	(0.110)
Order.RegionWest Asia	-0.062
	(0.113)
Order.RegionWestern Europe	-0.017
	(0.109)
Shipping.ModeSame Day	-3.332***
	(0.066)
Shipping.ModeSecond Class	-1.960***
	(0.060)
Shipping.ModeStandard Class	-3.671***
	(0.056)
log(distance)	0.131**
	(0.061)
Constant	2.042***
	(0.578)
Observations	54,243
Log Likelihood	-30,489.680
Akaike Inf. Crit.	60,999.360
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4 – Macroeconomic factors model

Results

	Dependent variable:
	Sales
airCargoDepart	0.021***
	(0.008)
portTraffic	0.001*
	(0.0004)
exportCost	-29.719*
	(15.277)
BusinessTax	6,128.918***

	(845.305)
ProfitTax	563.890
	(349.668)
Constant	-12,468.240**
	(4,936.114)
Observations	675
R^2	0.168
Adjusted R^2	0.162
Residual Std. Error	42,663.220 (df = 669)
F Statistic	27.096*** (df = 5; 669)
Note:	*p<0.1; **p<0.05; ***p<0.01