TheData Open

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Europe Regional Spring 2022 Datathon

Team No.1

How Can BigSupply Co. Reduce Loss?

Team Member[1]:
Jianlin Chen
Theodor-Mihai Iliant
Peike Sun
Guo Yang

March 13, 2022

[1] Equal contribution and in the order of surname initials

Abstract

Making profits is always a key issue for companies. In this report, we found that BigSupply co. has a large supply chain loss. Starting with Profit, we performed exploratory data analysis through statistical and machine learning methods. We found that the amount of delayed days is a very important factor affecting profits. We then verified the relationship between Delayed Days and Profit from a statistical perspective. We then tested the location distribution of Delayed Days. In addition, Fraud has significant impacts on Delayed Days. Therefore, we then analysed Fraud and found that it is highly correlated with location, prior number of the order cities and some other features. And finally, we presented a machine learning-based solution to help BigSupply Co. solve the supply chain loss problem.

Contents

Chapter 1

Introduction

Making profits is always a key issue for companies. To help BigSupply Co. make profits, we carried out a detailed analysis of the provided supply chain datasets. Firstly, we found that BigSupply Co. was losing significant profits in its supply chain, with a total loss being half of its profits. Also, the Order Profit distribution was heavily skewed to the left, indicating that there were some very large individual order profit losses. Therefore, we think it is essential to help BigSupply Co. solve the problem of large loss.

Our Topic: How to help BigSupply Co. reduce profit loss of orders?

To solve this problem, we need first to understand what factors influence Order Profit. We used LGBMRegressor to learn from the datasets, predict Order Profit and obtain Feature Importance. We found that Delay Days was one of the most important features. In addition, we also used a fixed effects model to regress the panel data. The Delayed Days turns out to be the variable with the lowest p-value. However, the performance of the model as a whole was not satisfactorily. This result led us to turn to a time-series approach - find the relationship between Profit and Delayed Days.

We constructed a VAR model to find the time series relationship between Delayed Days and Order Profit. We found that Delayed Days was highly negatively correlated with Order Profit. And the Delayed Days had an impact on the profitability of future orders. To explore whether they were causally related, we conducted a Granger causality test. The result showed that Delayed Days contributed significantly to the change in Order Profit.

After that, we wanted to take a deeper look at Delay Days. We found that Delay Days were highly correlated with geographic locations. Areas with large Delay Days were mainly concentrated in Africa and the hinterland. Also, we wanted to find out what other characteristics influenced Delay Days. In the VAR model, we found a positive correlation between the number of days delayed and the rate of suspected fraudulent orders on a weekly basis. We found that Fraud Orders are likely to be high in volume and the Order Country can be in a less developed or distant region. This affects the overall supply chain efficiency and leads to higher Delay Days.

Chapter 2

Overview of the Data sets

2.1 Data Collection and Data Processing

2.1.1 Data Collection

- External Datasets Declaration: we believe that the datasets provided do contain sufficient information about orders, customers and items, so no more external datasets are added.

- Geometry Datasets Declaration: shipping distance affects the shipping costs, shipping speeds, and level of potential damage to the goods. Thus, started from the latitude and longitude coordinates for departments contained in the raw data, some data are retrieved from the website GeoNames, shown in below table. Some of the city information in raw datasets does not match the information in any of the external datasets (this could because of incorrectly filling out the form or using a language other than English), The latitude and longitude coordinates obtained by using API package in python of the Order Region are used to replace the missing data.

Data Name	Description
cities5000.txt	Including the city names, ISO country codes for city, city nicknames,latitude and longitude coordinates of cities with a population greater than 5000
country_info.txt	Including ISO country codes, used to distinguish cities with same names in different country

Table 2.1: Geometry Data Description

2.1.2 Data Engineering

After data collection, we merged the provided BigSupply Co. Datasets and removed some columns: 'Customer Email', 'Customer Password', ' Order Zipcode', 'Product Description', 'Product Image' and ' Product Status', because they miss more than 80% of the values or contain only a unique value.

We also removed 'Customer Street', 'Order Item Discount ', 'Sales', 'Order Item Id', 'Order Id', 'Product Name', as they can be interpreted by other columns or there is little business logic that comes with them.

In addition to this, we added some columns through a feature engineering approach. These newly added columns adhere to the business logic and bring significant results in our later research:

- **Distance**: Distance between the Department and the Order City, calculated based on WGS-84 model.

- **Order_City_Count**: The number of different cities in which this customer has placed orders

- **FraudPrevNum**[1]: The number of times this customer has been suspected of fraud before.

- **AbnormalFreq**: Distance between the Department and the Order City, calculated based on WGS-84 model.

- **OrderCount**: The count of total orders of this customer.

- **DelayDays**: Days for shipping (real) - Days for shipment (scheduled).

[1] The FraudPrevNum, AbnormalFreq, OrderCount for each customer are processed with a lag of one, i.e. we use the data from row n to calculate the value of those features in row n+1 to ensure that no future data is included.

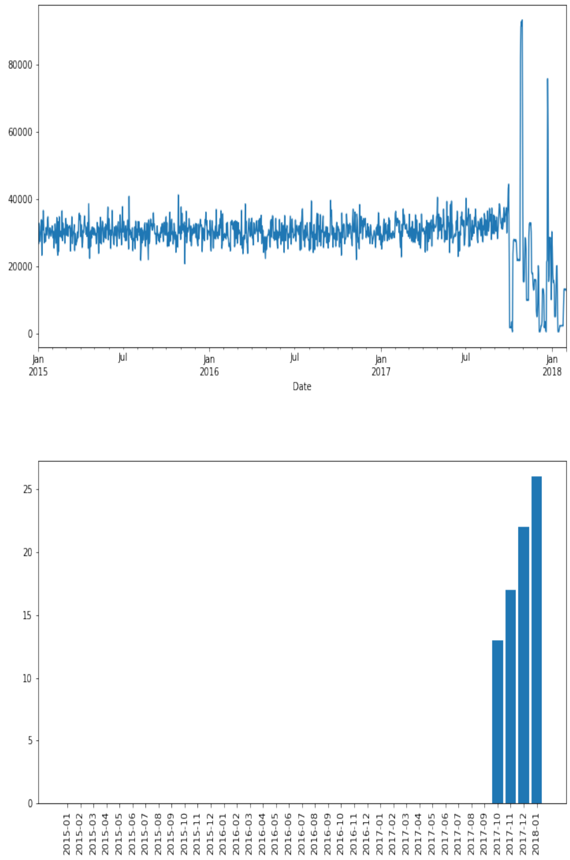


Figure 2.1: EDA

2.1.3 EPS

We have also investigated the pattern of the data in terms of time series. We want to focus on company profits, so we have conducted a time-series study of Order Item Total. This feature shows the total dollar value of the order. As shown in the figure
It is evident that the feature "Order Item Total" fluctuates significantly during the second half of 2017, showing a completely different pattern than before. For the sake of rigour, we applied the Extreme Studentized Deviate test (ESD) algorithm to detect outliers.
The method specifies that, for each iteration, the value with the largest deviation from the mean value needs to be removed from the dataset, and the corresponding t-distribution threshold needs to be updated simultaneously to test whether the null hypothesis holds.(i.e. "the smallest (one-tailed lower) or largest (one-tailed upper) value in the sample is not an outlier.") To implement ESD in python, we first calculate the largest deviation from the mean value after each iteration:

$$R_j = \frac{\max_i |Y_i - \overline{Y'}|}{s}, 1 \leq j \leq k \tag{2.1}$$

Next, we calculate the critical value:

$$\lambda_j = \frac{(n-j) \cdot tp, n-j-1}{\sqrt{(n-j-1+t^2_{p,n-j-1})(n-j+1)}}, 1 \leq j \leq k \tag{2.2}$$

We can then test if the null hypothesis holds by comparing the largest deviation and critical value. If $R_j > \lambda_j$, then the null hypothesis is rejected, and the sample point is an outlier. We repeat the process multiple times until all outliers are removed.
Applying ESD to the feature "Order Item Total", we find that all outliers are concentrated between 2017-09 and 2018-01, and the number of outliers increases over time. We suspect that this may be because some orders have been made, but have not been included in the database. We delete all data after 2017-08-31 to keep our research rigorous.

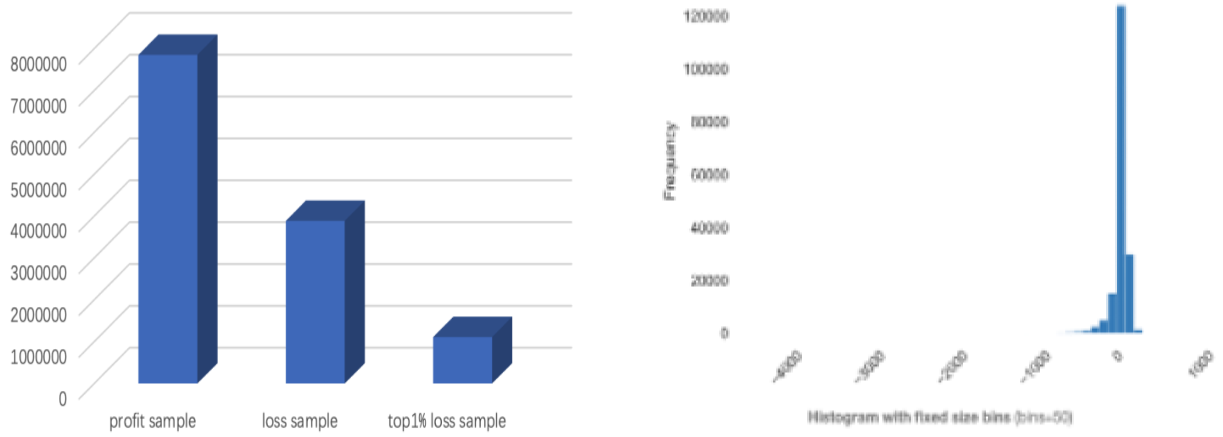


Figure 2.2: Profits

2.2 Exploratory Data Analysis: Decision to Study Delayed Days

Making profits is always a key business objective. Considering this, we analyse and visualise the profit and loss for each order of BigSupply Co.

We calculate the total profit for all profit-making orders and the total loss for all loss-making orders. The chart above shows that the total loss is about 50% of the total profit, indicating that a large proportion of BigSupply's revenue is lost due to loss-making orders.

The chart also shows the frequency of Order Profit in different profit bands. We find that the distribution is highly left-skewed with a Kurtosis of 71.37, indicating that there are some large-loss orders. Further analysis shows that the orders with the largest loss (top 1% of all orders) account for 14% of the total loss.

All the analysis above shows that how to reduce the loss is a key problem for BigSupply Co. to make profits.

The following research focuses on which factors are responsible for the loss. We apply machine learning and panel data regression methods respectively in order to get some insights.

Intuitively, the delay should impact the profits. This is the motivation behind engineering this new feature, which is the difference between the real and the scheduled days that an order takes to reach its destination. In order to confirm this business intuition using the available data, we run a model with the target 'Order Profit' and see which are the most important features in the explanation of this target. To this end, a significant issue is the selection of features to use in this model - in this case we provide an analysis made using the LGBMRegressor class. The features related to IDs - such as 'Order Department Id - are removed, whereas features related to categories - such as 'Department Name' - are kept. In addition, features related to the quantity or discount will be very good predictors for the profit - the more quantity you have the bigger the profit, the larger the discount the smaller the profit, so these kinds of features will not allow the model to learn about the underlying factors that determine the profit. Features related to the customer address are removed as well because they clearly do not have an impact on the profit, and related to the order location we only keep 'Order Country', because of its explainability of other features like 'Order Region'. The strategy of removing features can be inspired by an analogy to linear regression, where a group of nearly linearly dependent features increases the variances of the estimators of the coefficients of all explanatory variables in that group. Figure 2.1 shows that our feature, 'DelayDays', is worth being investigated further. We would like to know what causes the delay, and potential ways to prevent this from happening. All of this is done in section 4.

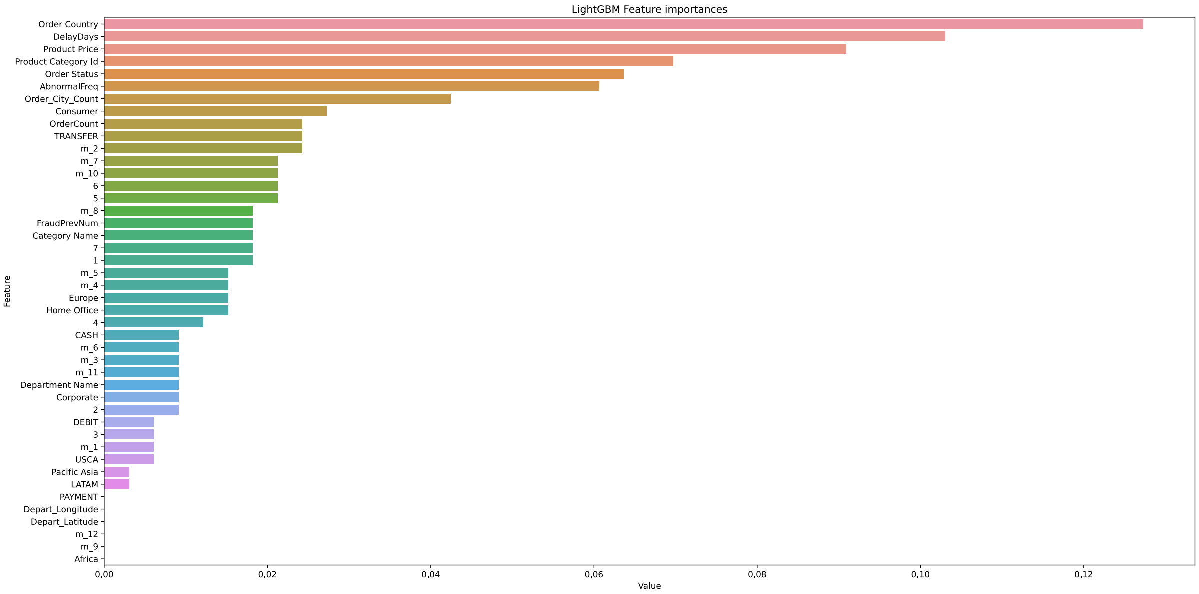


Figure 2.3: Feature importances when attempting to explain the response variable ‘Order Profit‘

The dataset is a relatively typical panel dataset, so we consider applying the panel data regression method. We choose between the fixed effects model and the random effects model as they are both considered effective panel regression models.
It can be seen that the panel regression method does not perform satisfactorily, with a minimum p-value of 0.122 for DelayDays. However, the results still suggest that, firstly, DelayDays has a high probability of affecting Profit; secondly, this dataset may not be well suited for statistical analysis in the panel dimension, so we will next build models in the time series dimension.

Chapter 3

Methodology and Results

3.1 VAR Model

In this section, a traditional econometric model called Vector Auto Regression (VAR) is used to investigate the relationship between profits, fraud status, abnormal status (excluded fraud), and delayed days. The corresponding variable names are "profit", "fraud", "abnormal" and "DelayDays". Being brought up by Christopher Sims in 1980, VAR use all the current variables as dependent variables and lag terms of all the variables as dependent variables. The regression expression can be written as:

$$\begin{bmatrix} Profit_t \\ DelayDays_t \\ Fraud_t \\ Abnormal_t \end{bmatrix} = \begin{bmatrix} c_{Profit} \\ c_{DelayDays} \\ c_{Fraud} \\ c_{Abnormal} \end{bmatrix} + A_1 \begin{bmatrix} Profit_{t-1} \\ DelayDays_{t-1} \\ Fraud_{t-1} \\ Abnormal_{t-1} \end{bmatrix} + A_2 \begin{bmatrix} Profit_{t-2} \\ DelayDays_{t-2} \\ Fraud_{t-2} \\ Abnormal_{t-2} \end{bmatrix} + ... + \epsilon \tag{3.1}$$

where c is the constant , A_i is the $i-th$ lag 4×4 coefficient matrix and ϵ is the error term

$$A_i = \begin{bmatrix} A_{i,1,1} & A_{i,1,2} & A_{i,1,3} & A_{i,1,4} \\ A_{i,2,1} & A_{i,2,2} & A_{i,2,3} & A_{i,2,4} \\ A_{i,3,1} & A_{i,3,2} & A_{i,3,3} & A_{i,3,4} \\ A_{i,4,1} & A_{i,4,2} & A_{i,4,3} & A_{i,4,4} \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_{Profit} \\ \epsilon_{DelayDays} \\ \epsilon_{Fraud} \\ \epsilon_{Abnormal} \end{bmatrix} \tag{3.2}$$

However, when use Stata to do the analysis, the software automatically takes the time term 0 into consideration and thus reflects the relationship between variables with no lag. In such case, the regression relationship can be improved to the following form:

$$\begin{bmatrix} Profit_t \\ DelayDays_t \\ Fraud_t \\ Abnormal_t \end{bmatrix} = \begin{bmatrix} c_{Profit} \\ c_{DelayDays} \\ c_{Fraud} \\ c_{Abnormal} \end{bmatrix} + A_0 \begin{bmatrix} Profit_{t-0} \\ DelayDays_{t-0} \\ Fraud_{t-0} \\ Fraud_{t-0} \end{bmatrix} + A_1 \begin{bmatrix} Profit_{t-1} \\ DelayDays_{t-1} \\ Fraud_{t-1} \\ Abnormal_{t-1} \end{bmatrix} + A_2 \begin{bmatrix} Profit_{t-2} \\ DelayDays_{t-2} \\ Fraud_{t-2} \\ Abnormal_{t-2} \end{bmatrix} + ... + \epsilon \tag{3.3}$$

As a reminder here, the coefficient $A_{0,1,1}$, $A_{0,2,2}$, $A_{0,3,3}$ and $A_{0,4,4}$ are not compulsorily set to be 1.
All the daily data, weekly data and monthly data are used to run the regression, and significant results with respect to profit can only be found in monthly data. The monthly regression results are shown in the below impulse-response figure, followed by the robustness check and Granger Causality test (shown in Appendix).
All the regression relations pass the robust check because all the eigenvalues lie inside the unit circle. And all result p-values for Grander Causality Test is smaller than 0.05, indicating that all of Delay, abnormal status and fraud are the reason for profit.

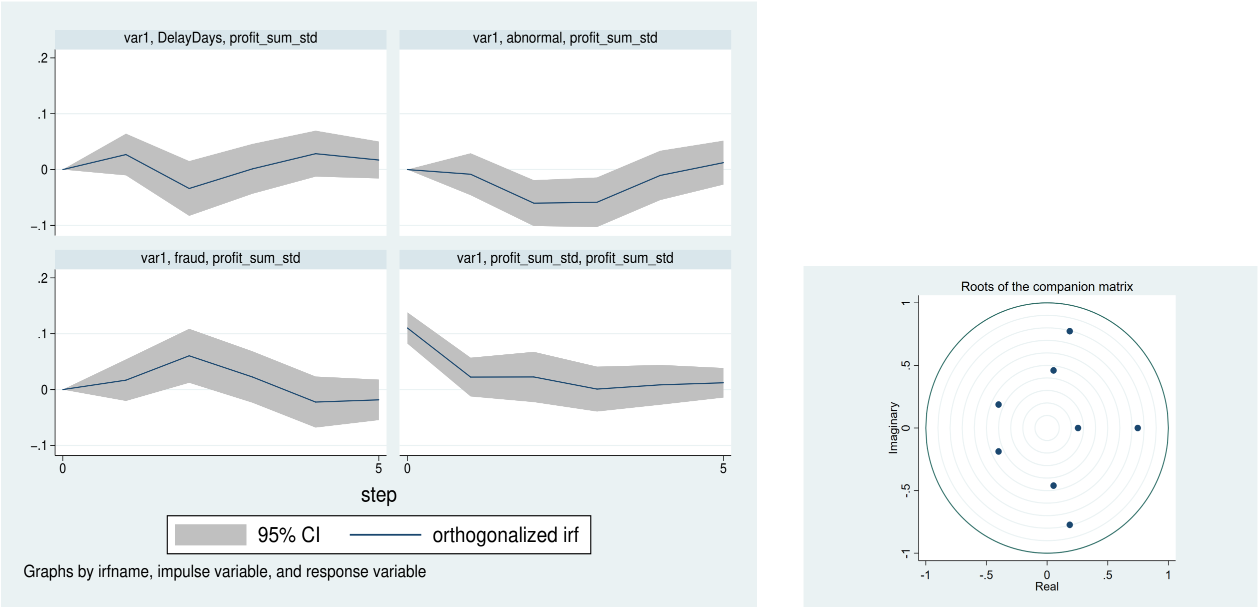


Figure 3.1: Influence factors of profits and Robustness Check

The result can be concluded in the following table:

Factors	What kind of influence on profit	Which month (from now)
Delay	decrease	second
Abnormal Status	decrease	second and third
Fraud	increase	second

Table 3.1: Monthly Data VAR results

3.2 Characteristics of Delayed Days

Since Delayed Days affects Profit, we intend to take a deeper look at the characteristics of Delayed Days. We first analyse the correlation between Delayed Days and some numerical features by applying two non-parametric test methods, Spearman and Kendall, as they do not require a specific sample distribution. The test results are given in the following table:

	Feature	Kendall Correlation	Kendall P-value	Spearman Correlation	Spearman P-value
1	Order Item Discount Rate	0.000741	0.685597	0.000991	0.685552
2	Order Item Quantity	0.000690	0.731264	0.000841	0.731307
3	Order Item Total	-0.002628	0.141037	-0.003604	0.141090
4	Order Profit	-0.002967	0.096001	-0.004077	0.095943
5	Product Price	-0.002574	0.168214	-0.003375	0.168155
6	Distance	0.004352	0.014960	0.005956	0.014999
7	Order_City_Count	-0.000331	0.863345	-0.000418	0.864318
8	FraudPrevNum	0.000508	0.815906	0.000570	0.815991
9	OrderCount	-0.000227	0.905991	-0.000287	0.906581

We found that a large number of features failed the tests. Features "Order Profit" and "Distance" pass the tests, with "Order Profit" negatively correlated with Delayed Days and "Distance" positively correlated with it. This matches the results of the VAR model, and is to some extent logical, i.e. the more distant the city, the greater the likelihood of delays. As Distance is an exogenous variable, it means that it is possibly the region where the order was placed that causes Delayed Days. Hence we next investigated the effect of geographic location on Delayed Days.
For convenience, we transformed "Delayed Days" into a binary feature called "Delay", where the feature value is 1 when a delay occurs and 0 if no delay occurs. We then do two chi-squared tests. One on "Delay" and "Order Region" and another on "Delay" and "Order Country". Both features are categorical variables and the chi-square test verifies that there is an independent relationship between the two features. We constructed the chi-square statistic from the actual and expected frequencies, and determined the acceptance and rejection domains. The results are shown in the following table:

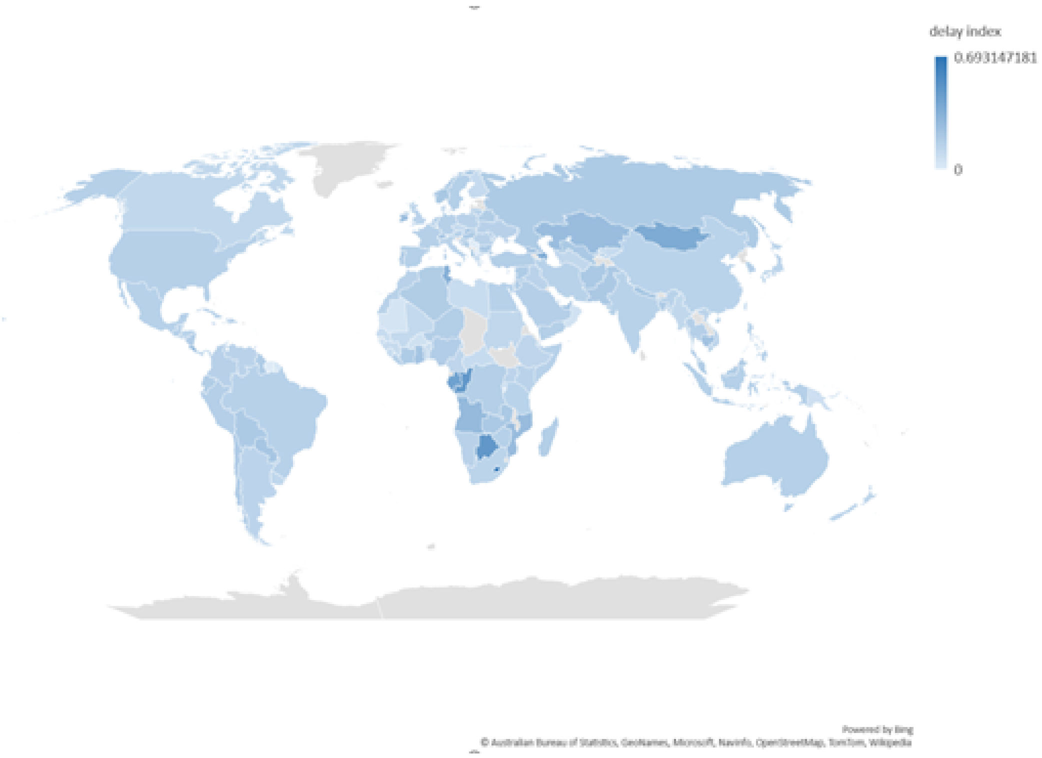
	Feature	Chi-2	P-value	df
1	Order Country	604.432	0.000	163
2	Order Region	72.324	0.000	22

To have a more intuitive idea of which regions have the most significant Delay Delivery, we use Delay Index to measure the extent of Delay Delivery in each region. Its formula is Delay Index = $\log(\text{MinMaxScaler}((F-E)/F)+1)$. Where F is the average number of actual Delay Days in a region, and E is the average number of theoretical Delay Days in a region. MinMaxScaler and log are performed to ensure "Delay Index" fits a normal distribution between 0 and 1. (Note that many of the countries' names in the dataset are written in Spanish and need to be translated using a translation API.) Finally, a distribution map of the "Delay Index" for different regions is plotted:

	Feature	Chi-2	P-value	df
1	Order Country	497.864	0.000	163
2	Order Region	71.489	0.000	22
3	Order Market	8.814	0.066	4

Table 3.2: Details concerning models where the target is the delay

Three of all the categorical features are significant: "Market", "Order Region" and "Order Country". This suggests that there is also a significant relationship between Fraud and geographic location. Similar to Delay Days, we use Fraud Index to measure the extent of Fraud in each region. Its formula is Fraud Index = $\log(\text{MinMaxScaler}((F-E)/F)+1)$, where F is the actual regional average Fraud rate, and E is the theoretical regional average Fraud rate. For the earth data map, MinMaxScaler and log are performed so that Fraud Index fits a normal distribution between 0 and 1. Below we plot the top 10 largest indexes:



Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom, Wikipedia

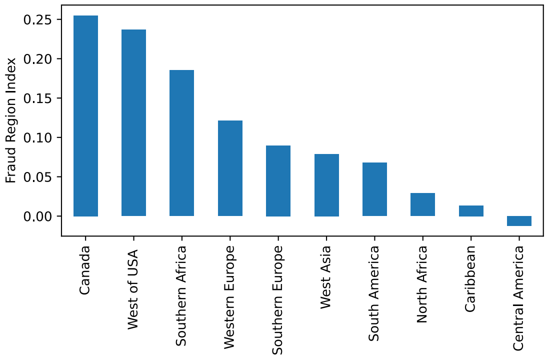


Figure 3.2: Feature importances when attempting to explain the response variable ‘Order Profit‘

We find that many Delay occurs in Africa and other landlocked countries. This result is logical, as Africa has insufficient transport infrastructures and inland areas are far from the sea. The results have great implications for BigSupply Co., as they can predict the approximate Delay Days based on the different shipping locations and hence estimate a more appropriate Scheduled Shipment Days. They can also optimise the supply chain efficiency by better laying out their supply chains and warehouses based on the geographical distribution of delay days.

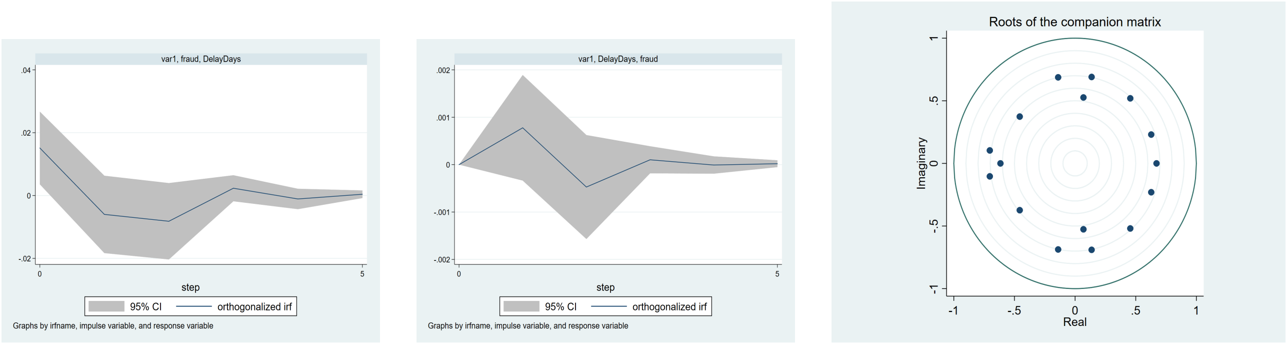


Figure 3.3: Influence factors of profits and Robustness Check

3.3 More about Delay Days: Fraud

When running the Var model using the weekly data, an interesting relationship between Delayed days and Fraud emerges.

In the time step 0, Fraud has a significantly positive effect on delay days, while the inverse is not true. This points out that more fraud in this week is more likely to induce more delays in the same time period. Both regression pass the robust check for all the eigenvalues for the whole VAR model are in the unit circle, however, both p-value Granger Test are bigger than 0.05 (results shown in the Appendix), indicating there is no Granger causal relationship between Delay and Fraud. However, the principle of Granger Test is to see the sequential order of two variables, and since the impulse effect is happened in time 0, meaning there is no lag term, so it is reasonable that the Granger test result is not significant.

A possible explanation for this result is with the increasing of fraud orders, more goods are are delivered fraud orders and time is wasted when waiting for sales return, this turnover process induces short of stock and caused delaying in the same time. And since the time period for each process are usually just few days, so the impulse relationship will appear in term 0. Further discussion with respect to "Delay Days" and "Fraud" is carried out in the following chapter.

There is a significant relationship between Fraud and Delay Days. A better understanding of Fraud will help us to reduce delays. We construct one-hot features of Order Status, and then extract the binary features of Fraud to do further research. As Fraud is a binary feature, we use a two-sample t-test to verify the relationship between Fraud and other numerical features. The standard Student t-test requires that the two samples satisfy the homogeneity of variance. In cases where homogeneity of variance does not meet, we use the Welch t-test. Otherwise, we use the Levene test.

	Feature	Method	Fraud	NotFraud	T Statistic	P-value
1	Order Item Discount Rate	Student T	0.103214	0.101648	1.347872	0.177701
2	Order Item Quantity	Student T	2.178286	2.187523	-0.381378	0.702923
3	Order Item Total	Student T	176.747942	178.406303	-0.993093	0.320666
4	Order Profit	Student T	20.223986	21.393865	-0.730679	0.464976
5	Days for shipping (real)	Student T	3.504790	3.498193	0.246055	0.805640
6	Days for shipment (scheduled)	Student T	2.914848	2.932399	-0.773888	0.438998
7	Product Price	Student T	132.296807	132.798771	-0.260688	0.794333
8	Distance	Student T	5404.533461	5414.999417	-0.193804	0.846329
9	Order_City_Count	Welch T	2.303619	2.217986	2.532992	0.011348
10	FraudPrevNum	Welch T	0.044971	0.051051	-1.684712	0.092123
11	OrderCount	Welch T	2.318520	2.234291	2.478736	0.013227
12	DelayDays	Welch T	0.589941	0.565795	1.012152	0.311528

Table 3.3: Statistical significance tests

We found significant relationships between Fraud and "Order_City_Count" or "FraudPrevNum" or "Order Count".

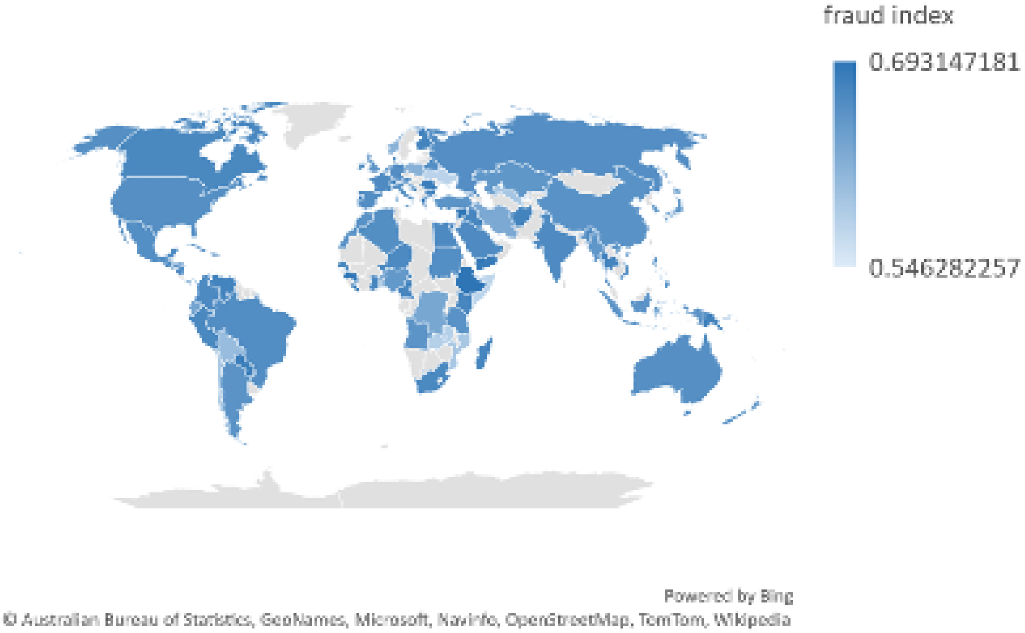
- The larger the "Order City Count", the more likely the order is a Fraud, which is in line with common sense as some people who place Fraudulent orders often change the location to avoid being detected.
- The smaller the "FraudPrevNum", the more likely the order is a Fraud. The explanation we give is that an account may be blocked after it has been found to have placed a fraudulent order.
- The larger the "Order Count", the more likely the order is a Fraud. This is logical as many large Fraud orders are split into smaller orders, and therefore, the number of previous orders is relatively high.

The significant relationships shown in this table confirm that the three features above can help us to identify whether an order is fraud, while the other features are not particularly useful in fraud detection.
In addition to the numerical features, we also investigate the relationship between the categorical features and Fraud. The chi-squared test is used as there are two categorical features. The results are as follows: (only significant results are shown)

	Feature	Chi-2	P-value	df
1	Order Country	497.864	0.000	163
2	Order Region	71.489	0.000	22
3	Order Market	8.814	0.066	4

Table 3.4: Details concerning models where the target is the delay

Three of all the categorical features are significant: "Market", "Order Region" and "Order Country". This suggests that there is also a significant relationship between Fraud and geographic location. Similar to Delay Days, we use Fraud Index to measure the extent of Fraud in each region.
Fraud Index = log(MinMaxScaler((F-E)/F) + 1)
Where F is the actual regional average Fraud rate, and E is the theoretical regional average Fraud rate. For the earth data map, MinMaxScaler and log are performed so that Fraud Index fits a normal distribution between 0 and 1.


fraud index
0.693147181
0.546282257
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom, Wikipedia

We find that Canada, West of USA and Southern Africa are the regions with the largest Fraud index (the Fraud index here is not yet processed by MinMaxed and logged). Based on this result, BigSupply Co. may need to focus on monitoring these regions to prevent fraudulent transactions.

Chapter 4

Solutions to Improve Profits

- Loss eats up the general revenue, so reducing delay days and detecting fraud in advance will help the company to reduce losses and gain more revenue.
- Model for predicting Delay days and put the results and process: 1. SARIMA model 2. different ML model sets
- Fraud Detection model, using ML model set and show the impact of sample imbalance. Perform SMOTE to improve the good

Delay

As explained in section 2.2, it is worth considering the feature 'DelayDays'. Based on the features that cause the delay, we can inform the company so that they make the right decisions to maximize their profits. We would like to predict the delay of an order at the current time, so we must eliminate all the features whose calculation depends on the future. This way, all the necessary features that are required to predict the delay are known at the current time, and hence also the prediction function. We split the dataset into train-test, in the proportion 80-20%. In figure 4.1 we display several models with performance metrics MSE and RMSE.

	Regression Model	MAE for Delay Detection	RMSE for Delay Detection
1	Lasso	1.1977	1.4875
2	Ridge	1.1977	1.4855
3	LGBMRegressor	1.1947	1.4886
4	Random Forest	1.1967	1.4865
5	XGBRegressor	1.2191	1.5288
6	Decision Tree	1.7093	2.1913
7	Linear Regression	1.1977	1.4855

Table 4.1: Details concerning models where the target is the delay

Fraud

The fact that fraud has a negative impact on the supply chain as a whole is obvious. Less obvious though is why it would negatively impact the delay, which is where the business insight comes in. Usually fraudulent transactions occur in large numbers, and especially if more of them focalize in a local area, the impact on the stores that have to ship those orders will be significant. That is, fraudulent orders impact other non-fraudulent orders, causing them delay.

Similarly to delay case, we now want to be able to predict whether an order represents a fraud or not. In this case, however, we treat it as a binary classification problem. As we want to do this at the current time, we remove all the features whose calculation depends on the future. Training is done on first 80% of the data, whereas the test data is the last 20% of the data.

We give details and motivation for using the SMOTE method here. Because only about 2.2% of the rows of the initial dataset are interpreted as fraud, there is a high imbalance between positive and negative samples. In this case, the model will ignore the minority class, although it is that class that we are most interested in. Before applying SMOTE on the model for 'FRAUD' target, the accuracy - for all 7 models tried - on the last unseen 20% of the data is about 97%, whereas the recall and f1 score vary between 0 and 5%. SMOTE works by oversampling the minority class - and it is only applied on the training data - by adding virtually no information to the already existing data. Before SMOTE is applied, most models produce confusion matrices where all frauds are detected as non-frauds, i.e. we have the maximum possible number of false negatives. Recall the confusion matrix has, in row i and column j, the number of samples of class i that are predicted to be of class j. Ideally, we would like to have as small a number of false negatives as possible. Such types of data augmentation, like the SMOTE method, are also performed in Reinforcement Learning as well as Computer Vision, but for different purposes. Table 4.2 shows the results of several models, where SMOTE is applied only for the model concerning fraud.

	Classification Model	Accuracy Score for Fraud Detection	Recall Score for Fraud Detection	F1 Score for Fraud Detection	(%)#False Negatives/Total #Negatives
1	Logistic	90.24	8.23	12.89	70.24
2	Extra Trees	97.55	25	0.48	99.75
3	XGB	97.36	6.32	1.12	99.38
4	Decision Tree	94.44	9.48	11.64	84.93
5	SVM	90.07	8.13	12.79	70.00
6	Gaussian Naive Bayes	73.64	8.43	15.56	0
7	Random Forest	90.69	8.12	12.51	72.59

Table 4.2: Details concerning models where the target is fraud

Based on the results, we can pick Linear Regression for delay prediction and Logistic Regression or Gaussian Naive Bayes for the fraud classification.

Appendix A

Fixed Effects Model Results

	Features	Parameter	Std. Err.	T-stat.	P-value
1	const	0.1507	0.0196	7.6849	0.0000
2	Market.Europe	-0.0068	0.0082	-0.8275	0.4080
3	Market.LATAM	-0.0893	0.0599	-1.4905	0.1361
4	Market.Pacific Asia	-0.048	0.0069	-0.6918	0.4891
5	Market.USCA	0.0171	0.0146	1.1702	0.2419
6	Order_Item_Discount_Rate	-0.0012	0.0012	-1.0129	0.3111
7	Order_Item_Total	-0.0006	0.0016	-0.3832	0.7016
8	Type.DEBIT	-0.0059	0.0039	-0.3832	0.1358
9	Type.PAYMENT	-0.0051	0.0042	-1.2042	0.2285
10	Type.TRANSFER	-0.0030	0.0041	-0.7342	0.4628
11	Product_Price	-0.0004	0.0016	-0.2416	0.8091
12	Distance	0.0008	0.0013	0.6065	0.5442
13	OrderCount	0.0007	0.0008	0.9244	0.3553
14	DelayDays	-0.0018	0.0012	-1.5464	0.1220

Appendix B

Granger Causality Wald Tests Result

Equation	Excluded	chi2	df	Prob>Chi2
profit	abnormal	15.051	2	0.001
profit	fraud	19.510	2	0.000
profit	DelayDays	9.206	2	0.010
profit	ALL	45.106	6	0.000

Table B.1: Granger Causality Wald Tests Result for weekly data

Equation	Excluded	chi2	df	Prob>Chi2
fraud	DelayDays	4.992	4	0.288
DelayDays	fraud	5.209	4	0.267

Table B.2: Granger Causality Wald Tests Result for weekly data