Analysis of the factors affecting late deliveries, sales, consumer behaviour and fraud orders in the BigSupply Co. dataset

Junhua Chen*, Chun Hei Yip*, Tunan Shi†, Wai Yeung Ying‡
*Trinity College, University of Cambridge
†Sidney Sussex College, University of Cambridge
‡Emmanuel College, University of Cambridge

***Abstract*—We examine the BigSupply Co. dataset from 4 perspectives: consumer behavior, stochastic trends, supply chain shipping effectiveness and fraud detection. We first show that there is unlikely to be a relation between delayed shipping and consumer dissatisfaction (indicated by abstaining from further ordering), and then find seasonal trends in sales data. Next, we use machine learning methods to identify statistically significant indicators for orders to be delayed or to be fraudulent.**

I. Introduction and Non-technical exposition

As per the abstract, the four perspectives we considered each led to their own discussion questions:

- **Consumer Behavior:** Are customers who receive their orders late less likely to make more purchases?
- **Stochastic trends:** Are there any overall trends in the number of orders or profits? This could be seasonal, or an upward or downward trend (over time).
- **Fraud Detection:** What are some indicators of an order being fraudulent?
- **Lateness prediction:** Are there traits of orders that make them more/less likely to be late?

In (1), we came to the surprising conclusion that consumers statistically do not seem to mind late orders.

In (2), we uncovered a very distinct seasonal spike in October and November in orders for some departments and drops in others departments around the same time, but did not uncover any long term growth trends.

In (3) and (4), we fitted simple machine learning models to predict lateness and fraud, then interpreted them via the permutation importance metric to help discover relevant factors, after which these factors were confirmed by hypothesis testing. It turned out that very few factors were actually relevant to each problem, with fraud being most correlated with just geographic factors and payment method (with some observed product specific behaviour), whereas lateness was determined almost exclusively by Scheduled delivery time and, to a lesser extent, geography.

II. Introduction and Exploratory Analysis

A. Introduction to the Dataset

1) The Orders Dataframe: The columns can be divided up into several themes:

- Geographic information (where to deliver the order). This included Order city, state, country, region, market. It was noted that some of the names were in Spanish.
- Basic Order information, such as customer ID and Item category and department, as well as order date (date and exact time).
- Financial details, including ordered quantity, price, discount, sales etc.
- Delivery Status. This included information whether the order was fraudulent, late, and includes information about scheduled and actual number of delivery Days.

2) Observations about the Dataset:

- Customer country in the customer table had only two distinct values, Puerto Rico and US, even though their requested delivery location might not be in the USA.
- Most countries had data for only subintervals of the time, when investigated during exploratory analysis. For instance, the USA only had orders on about 200 days. We strongly suspect incomplete data for this reason, and thus chose to abandon stochastic analysis of individual country results.
- 2018 saw a sharp decline in orders per day. We suspect this is due to incomplete data.

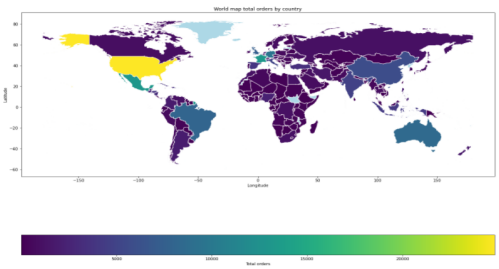
B. Data Cleaning

- For ORDERS table, the order zipcode was mostly blank and thus ignored. Order Item ID serves as another unique order ID (Order cardprod ID is the actual ID of item ordered) so was removed as it was rendered unnecessary by the dataframe indexing.
- For CUSTOMERS table, the email and password columns were removed as they were all blanked out in first place. First and last names were also removed.
- For products, description was completely blank, availability was all set to "available", and the web links were all broken so those columns were removed.

- For each department we were provided their geographic location. After plotting the locations on a map we saw that they were all within a few kilometres of each other on Puerto Rico (see plot in next section), and thus we considered the locations uninformative and removed them.

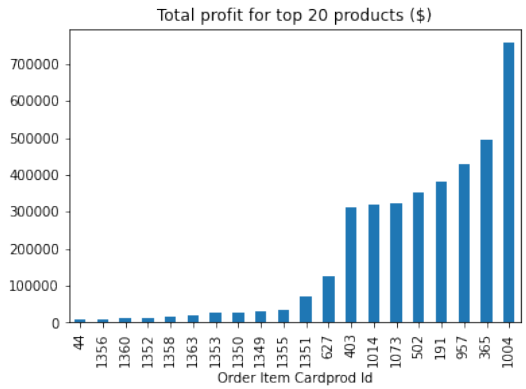
C. Exploratory Data Analysis

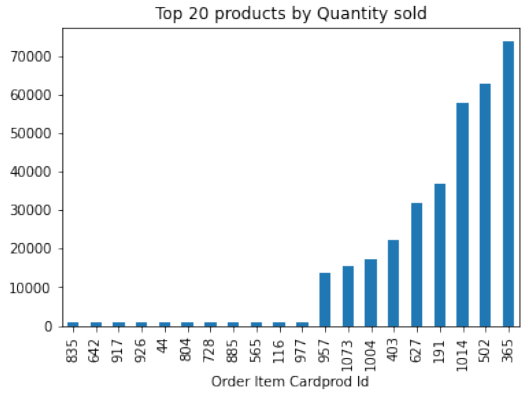
1) Consumer Demographics:



It is evident that the largest amount of orders were delivered to the USA, by a large margin. This is because all of the customers are from the USA (including Puerto Rico), and the departments of the company are located in Puerto Rico.

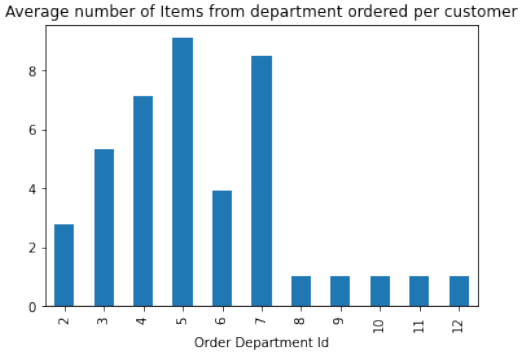
2) On selling behavior of individual products and departments:





It is clear from the graphs that the profits and number of orders from the same 9 products far outshine that of all other orders.

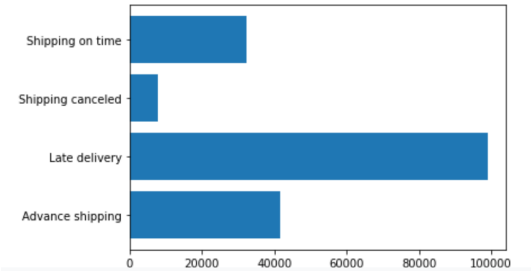
Now we observe another interesting trend:



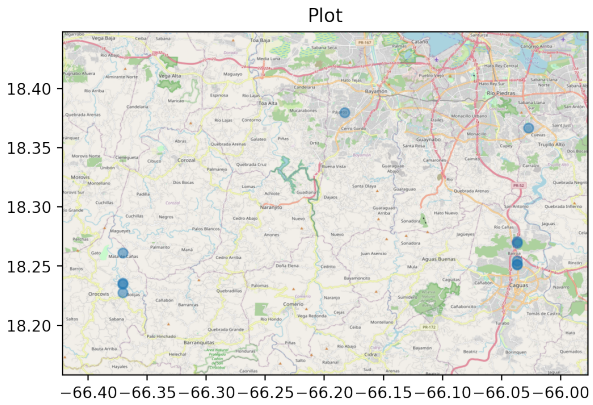
Notably customers only ever brought 1 item per person from departments 8 to 12. We found that the products in departments 8-12 corresponded to names such as ‘DVDs’, ‘Sports Books’, ‘Industrial consumer electronics’, ‘Web Camera’, and ‘Dell Laptop’. Since no detailed description was available in the dataset, we had to infer what these meant from our common sense. It seems that customers who buy these products will only need one copy of them, especially if they are buying these products for their own use.

3) Firm’s overall performance in delivering orders:

Around 54.8% of Orders are delivered late, while 23.0% are delivered early, 17.8% on time and 4.3% cancelled (this can be due to fraud or other reasons). This indicates that late delivery has been a big issue at Big Supply Co. A horizontal bar chart of the overall performance is shown below.



4) Firm’s department centres, mapped geographically:



They are all situated near each other in Puerto Rico, hence why we considered the location data of departments uninformative.

III. Consumer Behaviour

A. *Effect of lateness on consumer behaviour*

We formulate two features for each customer:

- **Order frequency**. This is calculated by

$$\frac{\text{date of last order} - \text{date of first order}}{\text{number of orders} - 1}.$$

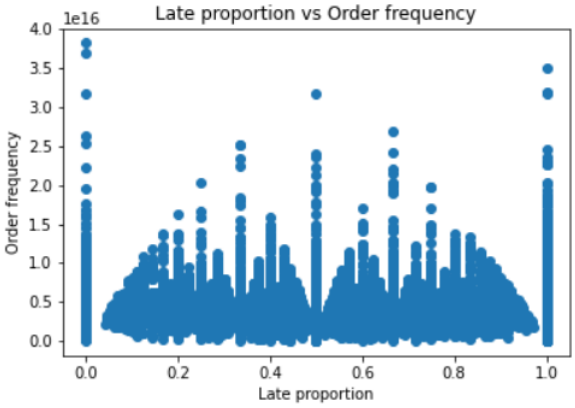
  It should be noted that customers who have only made one order are not considered since there is no meaningful order frequency metric.
- **Proportion of late orders.** This is calculated by

$$\frac{\text{number of late orders}}{\text{number of orders}}$$

  where a late order refers to when *Days for shipping (real)* $>$ *Days for shipment (scheduled)*, as stated in the `Orders` spreadsheet.

It is hypothesised that a lower proportion of late orders would imply a higher order frequency, indicating a higher customer satisfaction.

The following graph was plotted for the features:



We conduct a Spearman's rank correlation coefficient test. This involves taking the ranks of the data and calculating Pearson's product moment correlation coefficient for this. This allows us to test for monotonic relationships between late proportion and order frequency (which is what we would expect). The Spearman's rank correlation coefficient is calculated by

$$r_s = \frac{cov\left(R_X, R_Y\right)}{\sigma_{R_X}\sigma_{R_Y}}$$

where R_X, R_Y are the ranks of the two variables X and Y.

We test the following hypotheses (with significance level 5%):

H_0 No association between proportion of late orders and order frequency

H_1 Some association between proportion of late orders and order frequency

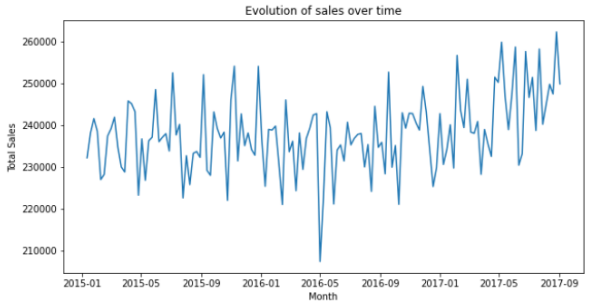
The p-value can be calculated, for this instance, by a permutation test. Using the `scipy` library, this yields a p-value of 0.819. Hence, we do not have sufficient evidence to reject H_0. This shows that late orders may not actually affect customer behaviour. We also note that due to the large size of the data, the probability of a Type II error should be low, hence it is unlikely that H_1 is true.

IV. Stochastic analysis

A. *Seasonal testing for sales/profits*

Aggregating the orders by week, we calculate the total sales for each week. There were some weeks at the start and end of the period that exhibit wild fluctuations, which we attribute to incomplete data entry. Hence, we ignore these periods.

Here is the graph of weekly sales over time:



To test whether there is seasonal or time variation between sales, we carry out an Augmented Dickey-Fuller Test on the weekly sales data. This is used to test whether there is a unit root in the time series, which would indicate non-stationarity (and hence possible time trend/instability).

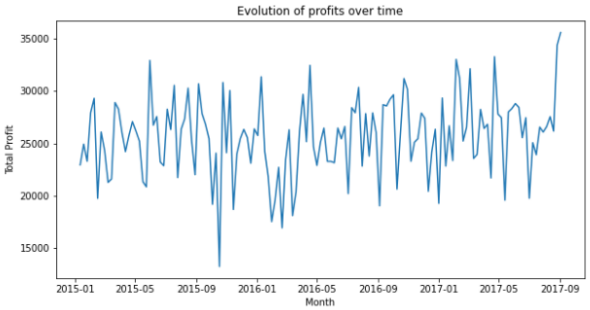
We test the following hypotheses (with significance level 5%):

H_0 Time series is not stationary (unit root exists)

H_1 Time series is (covariance) stationary

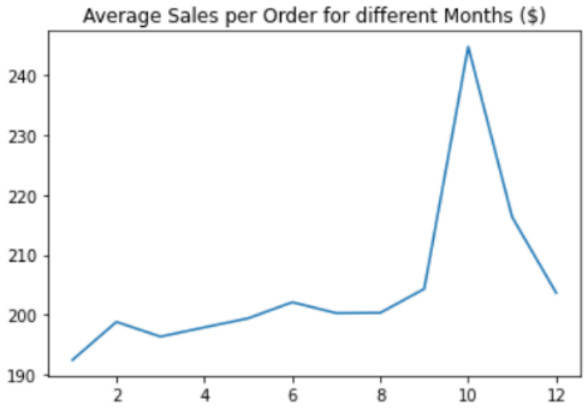
The test yields a p-value of 0.984, which suggests that we cannot reject the hypothesis that the series is non-stationary. It also uses a lag of 12, suggesting that there may be a sales cycle of length approximately 12 weeks. However, note that the mean and standard deviation here are around 240000 and 9000 respectively, suggesting minimal variation in the sales data. Given extra time, this is one of the paths we would explore further (see 'Further Investigation').

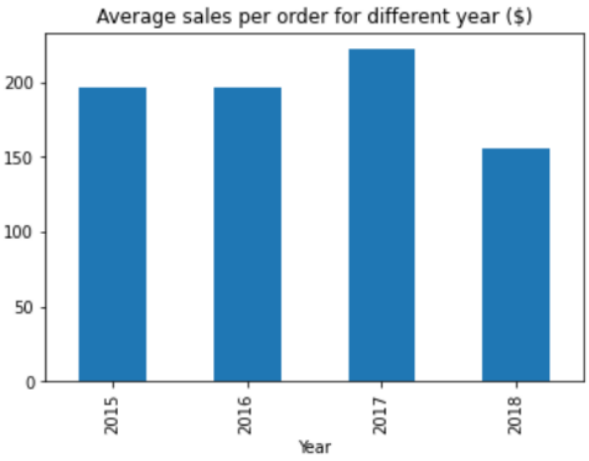
Running the test on weekly profits, however, yielded a p-value of 2.63×10^{-19}, suggesting that the profits were covariance stationary and yielded no time-dependent trend. The mean and standard deviation, around 26000 and 4000 respectively, also suggests that profits were quite constant over time. Here is a graph of profits over time:



B. Department seasonality

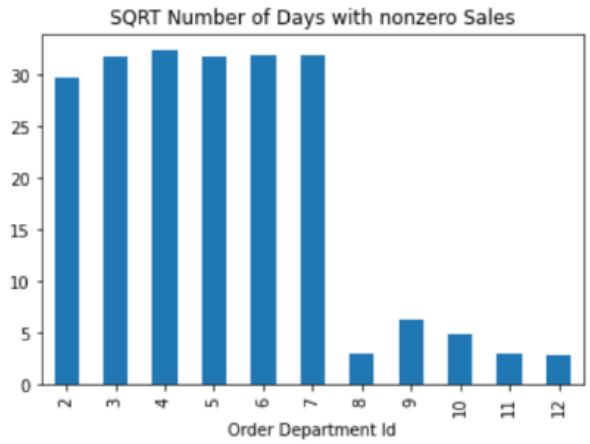
With the suspicion that there may be some sort of underlying seasonality, we analysed the average sales for each month of the year and also for each year. The visualisations are shown below:



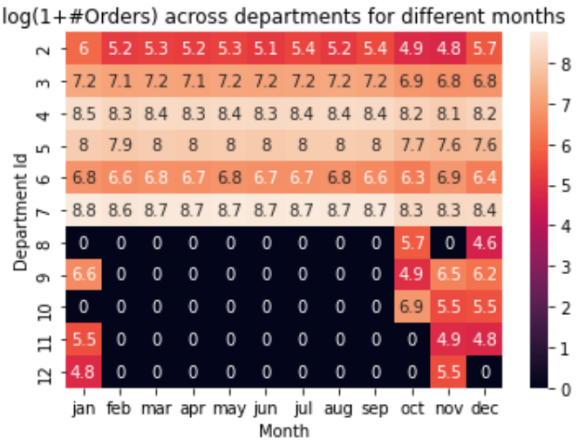


Note the dip from 2017 to 2018 in average sales per order from the above graph. This is possibly due to a small sample size, since data in 2018 was only included up to January. Also, data from October, where average sales is the highest (elaborated further below), is not included in 2018, leading to the smaller value for average sales per order. For monthly sales, there is a significant spike in average sales per order in October and November, implying that people are more likely to spend in large amounts in the festive period. We explain this spike by considering

- the orders within each department; and
- the average sales price per department.



Department 10 (technology) is understandably much more expensive than the other departments, as the average electronic product is of a much higher price range. Further, across all of 3 years, the departments numbered 2 through 7 have about 900 days of "nonzero sales" (says in which more than 1 product was sold), while departments numbered 8 through 12 only have about 10 to 50 such days. Furthermore, these sales all occur in the months of October, November, December and January. The below graph well illustrates the seasonality and departmental trends.



From the visualisation, we can also deduce that

- Departments 2 to 7 have sales orders of magnitude above the rest.
- Departments 8 to 12 show evident seasonality. This, combined with the high sales per order of department 10, accounts for the spike in sales volume in October and November. This is dampened somewhat in November and December by slight reductions in sales in the other departments.
- It is extremely surprising that departments 8 to 12 had no orders at all for the other months. This could indicate incomplete data entry, but due to the fact that there is no clear date cut-off for the departments not having orders, it does not appear to be the case.

V. Factors contributing to late orders

A. Methodology

One of our aims was to investigate what factors contribute to the late delivery of orders. Upon analysis of the data, a 1 in the column "Late Delivery Risk" corresponded perfectly to a Late Delivery Status and a 0 corresponded to a cancellation, on-time or early delivery. Thus, we chose the "Late delivery Risk" column to be the definition of whether an order was late. On a high level, our approach consisted of 3 steps.

1) Targeted data cleaning: Remove any orders with status "Shipping canceled" (as predicting their lateness does not make sense). We only consider factors that are interesting, rather than those directly encoding answer or those that could have not possibly been known when the order is dispatched.
2) Fit simple machine learning models to predict Late Delivery Risk and use the permutation importance metric to identify features are used in the prediction. We consider a feature significant precisely when not using it reduces the accuracy by at least 1%.
3) Perform statistical tests to confirm the significance of the causation.

1) Elaboration on methodology of data cleaning: We found it difficult to avoid feeding in some form of ground truth to our models.

For instance, we could not include both the scheduled and actual delivery times into the data. Since the actual delivery time being greater than the scheduled delivery time obviously indicates that an order must be late, so we would be encoding the answer into a combination of these two features. Since we would not know the actual delivery time when the order has just been dispatched, we elected to remove the actual delivery time and keep the scheduled delivery time.

Additionally, we discovered that when looking up orders under some customer and timestamp, multiple entries of orders would appear with the exact same order delivery details and status, except with possibly different department details which will all have the same delivery status. This is possibly explained by the order system identifying each distinct item in a basket as a separate order.

Thus, if just one of those orders was placed in the training set, then the model, through memorisation, would seem to be able to "predict" the lateness of the other orders. We mitigated this hazard by splitting test/train sets by date so that each group of orders only ever belong in one of the train/validation sets.

Lastly, the customer table was joined to the orders table so their values could be used too.

2) Elaboration on model construction: We used 4 simple models; XGBoost, Random Forest, Logistic Regression, and Multinomial Naive Bayes, and examined each one with an F1 score exceeding 0.6. Random forests was fixed at 500 estimators and depth limited to 10. XGBoost's parameters were selected by Grid Search Cross Validation to be 500 estimators, learning rate 0.1 and max depth 5. For Naive Bayes, numerical features were not considered for the ease of programming, and for all models categorical features were encoded by the one-hot method.

We use models to detect relationships between features and guide further investigation. This is more useful than simple correlation matrices because we are able to detect if a combination of features is a high indicator for a certain label and extract more complicated relationships, especially when using XGBoost which can model relationships with multiple variables rather than treating them as independent.

Models were examined by calculating the permutation importance for each column, which permutes the features of that column in the validation set and calculates the decrease in accuracy as the result of removing the signals from that feature. Any feature that received a importance of at least 1% by any model was considered significant and earmarked for statistical testing, with the cutoff chosen by consideration of our limited time.

3) Elaboration on statistical testing: Statistical tests were performed to test the hypotheses that a certain factor is independent to whether the order was late or marked as fraud. This is to provide evidence that our models are not over-fitting the data, or that we have not just stumbled upon a model (with the right parameters) that had high validation set F1-score. The test used was the chi-squared test (for independence).

B. Model performance and feature importance

We summarize the performance of models by their F1-scores and accuracy. We did not consider building good models to be a high priority as we are using them as tools to identify complex correlations.

Model	F1-score	Accuracy
Random Forest	~ 0	0.570
XGBoost	0.694	0.710
Multinomial Naive Bayes	0.718	0.641
Logistic Regression	0.705	0.679

It should be noted that the Random Forest model ended up labelling all of the data as late, which explains why its F1-score is close to 0 and accuracy is close to 50%.

The imbalance of the F1 and accuracy is due to the slightly skewed nature (57% to 43% ratio) of the data. The models with good performance all concurred on what features were significant. "Scheduled Days for Shipment" was by far the most significant feature, with a decrease of 16.2%. The feature 'Order City' was the only other important feature, being assigned an importance of 1.6%. The top 4 feature importances (ranked by maximum across all models) are shown below.

1) **Days for shipment (scheduled)**: 0.162
2) **Order City**: 0.0165
3) **Order Country**: 0.00477
4) **Order State**: 0.00382

C. Statistical tests

We carry out a chi-squared test for independence between lateness and "Order City" and "Scheduled days for shipping".

We test the following hypotheses (with significance level 5%):

H_0 Factor and lateness are independent
H_1 Factor and lateness are not independent

Under H_0, we have that $p_{ij} = p_i \times q_j$, where p_i, q_j are probabilities of taking values X_i, Y_i respectively. Under H_1, there are no restrictions on p_{ij} except that they sum to 1. We observe that H_0 is a subset of H_1, with $\dim(H_1) - \dim(H_0) = (mn-1) - (m-1+n-1) = (m-1)(n-1)$.

We calculate the likelihood ratio for this test. By computing the maximum likelihood estimator for p_{ij} under H_0 and H_1, we get

$$2\log\Lambda = 2\sum_{i,j} o_{ij} \times \log\left(\frac{o_{ij}}{e_{ij}}\right)$$

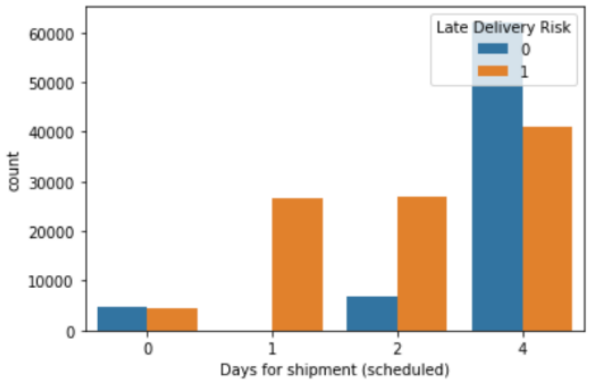
where o_{ij}, e_{ij} are the observed and expected values for each pair of categorical values. For $o_{ij} \approx e_{ij}$, we can use the Taylor expansion for $\log$ to get the Pearson statistic

$$\sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

By Wilks' Theorem, for n large, this approximately follows the chi-squared distribution with $(m-1)(n-1)$ degrees of freedom. This allows us to compute a p-value for the observed data.

1) For scheduled delivery time: We get a chi-squared test statistic on the order of 4×10^4, and there are 3 degrees of freedom. This yields a very small probability, meaning that we reject H_0 and state that it is likely that scheduled delivery time and lateness are not independent.

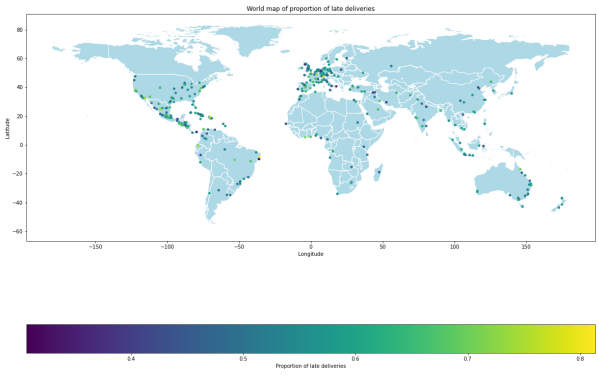
Note since the observed value for (1 days of shipment, not late) is 0, which deviates significantly from the expected value, the Pearson statistic and indeed the log likelihood ratio may not be appropriate anymore. However, in any case there is clear evidence showing the two variables are not independent.



Indeed, we see that if days for shipment is equal to 1, then the order is always late. The proportion of late orders for days with shipment = 2 is also extremely high. Clearly, BigSupplyCo often underestimates the number of days it takes to ship an item - it cannot be done in 1 day.

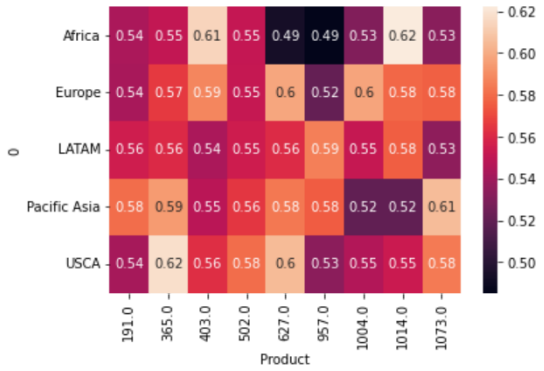
2) For Order City: We first remove cities which have had under 100 orders, in order to make the Pearson test statistic more accurate. This results in 342 remaining cities, with 341 degrees of freedom and a chi-squared test statistic of 1452. The p-value is still very small 8.8×10^{-137}, showing that there is sufficient evidence to reject H_0 and it is likely that some cities are more likely to yield late deliveries than others.The reason will be further explored in the section below.

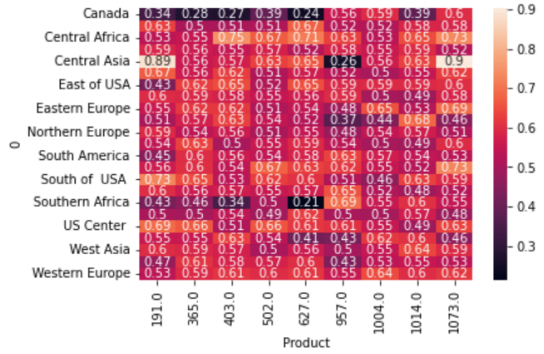
Below is a map showing the proportion of late orders in these 342 cities:



Hence, as our model suggests, the days for scheduled shipment and the destination of the order have an impact on the lateness of the order.

D. Regional variation of delivery speeds

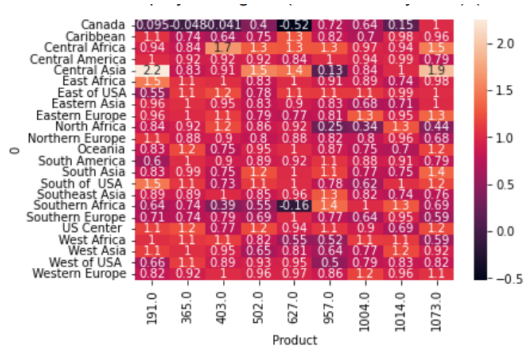




We calculated the average number of late days for major products in the various markets (first) and regions (second) by the groups that have more than 30 orders in each product. It is noted that all products in each group have more than 30 track records of delivery.

It can be seen that Central Asia performs very poorly in product 191 and 1073 while Canada and South Africa perform better than average in most categories.

Previously it has been found that the number of days scheduled would be a major influencing factor for the lateness. In order to identify region-product pairs with higher number of days late and higher number of days scheduled, we calculate the indexes inside the heatmap as the average of (number of days late) $\times$ (number of days scheduled):



However, it is noted that the two products 191 and 1073 still have some serious delivery delay issues with average number of days late being at 2.2 and 1.9 respectively.

This may be explained by the land-locked nature of Central Asia where transport is slower, compared to the highly developed transport infrastructure and ports in Canada. This can also be seen from South Africa, whose average number of days late is generally lower than Central Africa's.

This also explains why the 'Order City' would influence the lateness in delivery in the previous section.

VI. Factors contributing to fraudulent orders

A. Methodology

Like section V, we will investigate statistically significant predictors for fraudulent orders through model construction. We will assume that the label 'suspected fraud' will serve as the 'ground truth' for fraud detection, despite it only being 'suspected' of fraud. A quick check shows that there are 4066 suspected fraudulent orders out of the roughly 180 000 orders, a minuscule proportion. Thus it is essential that we balance the dataset.

Our methodology framework was same as Section V, consisting of task specific cleaning and feature engineering, prediction and statistical testing.

1) Elaboration on methodology of data cleaning: We operated on the same principles as we did in section V and avoided features that directly encoded the ground truth. This meant removing the delivery status class, since all fraudulent orders would be cancelled, and furthermore the final status would not be known at dispatch time. Also, as the orders are canceled it would not make sense to look at if they were late so those columns were omitted also.

To reduce correlation between columns, of the financial columns, only 'Sales' and 'Profit' was kept.

As the fraud class is disproportionately small, it was decided to undersample and produce a dataset with more comparable numbers of fraud and non-fraud samples. We ultimately used a sub-dataset of the 4000 fraudulent orders plus 8000 non-fraud orders, split into train/validation in 4:1 ratio by taking the 80th time percentile. This was for reasons similar to described in section V. The choice of of the 4000:8000 ratio was to create some form of imbalance within the dataset to make the permutation importance more useful.

2) Elaboration on model construction: The same models were fitted as in section V, except random forests, and as with before numeric features were ignored by naive Bayes. The parameters chosen by gridsearch CV for XGBoost was 500 estimators, max tree depth 2 and learning rate 0.05.

We chose to measure the permutation importance on the cross validation set so that both positive and negative samples

have a fair say on the importance, as opposed to the whole dataset where the importance is almost completely determined by non-fraud samples.

3) Elaboration on statistical testing: The same methodology and ideas from section V were reused here. Since the tests here include tests on continuous data as well as tests on categorical data, we use both the chi-squared test for independence and Welch's t-test.

B. Model Performance and feature importance

We summarize the performance of models by their F1-scores and accuracy. We did not consider building good models to be a high priority as they are merely mediums for us to identify correlations. Due to the intended imbalance in the train/test sets, the F1 scores are better indications of accuracy.

Model	F1-score	Accuracy
Multinomial Naive Bayes	0.496	0.720
Logistic Regression	0.670	0.770
XGBoost	0.802	0.829

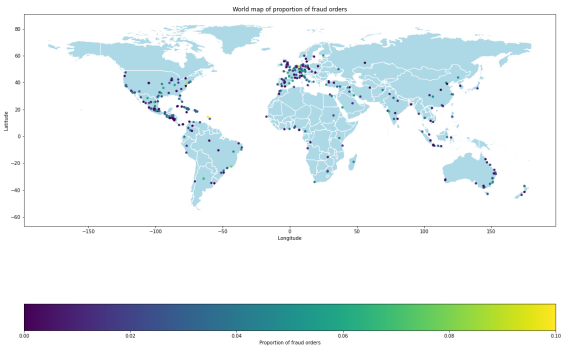
Three features were identified as being significant, Payment Type which had an importance of 36%, Profit of 2.8% and Order city which had an importance of 2.7%. Several other geographical features were also indicated to be relevant, we just consider order city due to close correlations between the categories. All features with importances exceeding 1% are shown below:

1) **Type**: 0.359
2) **Order Profit**: 0.0277
3) **Order City**: 0.0275
4) **Customer City**: 0.0192
5) **Order State**: 0.0166

C. Statistical tests

It was unnecessary to apply hypothesis test to Payment Type, for it was discovered that all fraudulent orders was executed by the 'TRANSFER' payment method, thus showing that the company should be more wary of transfer payments. For Order City, we can again use a chi-squared test for independence. We filter cities in a similar way as Section V. Also, since fraud is a highly unlikely feature, we also filtered cities with fewer than 5 suspected fraud orders, in order to maintain the validity of the chi-squared test. This resulted in 171 cities remaining, with 170 degrees of freedom and a chi-squared test statistic of 589. The p-value is miniscule (1.4×10^{-47}), suggesting that the order city and the fraud proportion are not independent.

Below is a map showing the proportion of suspected fraud orders in cities with a high amount of orders:



We carry out Welch's t-test to test whether the mean profit for non-fraud and fraud-orders are equal. This assumes the population means are normally distributed (which is true in this case, since the size of the population is large for both fraud and non-fraud, and we can use the Central Limit Theorem). Note, however, that since some orders exhibit striking similarity, the samples may not be independent, hence compromising the size/power of the t-test.

Let the sample mean and sample standard error for fraud and non-fraud orders be m_f, m_n, s_f, s_n respectively. Let the number of fraud and non-fraud orders be n_f and n_n respectively.
We test the following hypotheses (with significance level 5%):

H_0 $m_f = m_n$
H_1 $m_f \neq m_n$

Then, the t-statistic is defined as $\frac{m_f - m_n}{(\frac{s_f^2}{n_f} + \frac{s_n^2}{n_n})^{\frac{1}{2}}}$. This follows, approximately, a t-distribution with degrees of freedom approximated by $\frac{(\frac{s_f^2}{n_f} + \frac{s_n^2}{n_n})^2}{\frac{s_f^4}{n_f^2(n_f-1)} + \frac{s_n^4}{n_n^2(n_n-1)}}$.

This gives a t-statistic of -0.639 and a p-value of 0.522. So we fail to reject H_0, suggesting that order profits for fraud orders do not differ from order profits for non-fraud orders.

Thus despite being seen as significant by the model, statistical testing has refuted the notion that profits for individual orders and fraudulent activity are related.

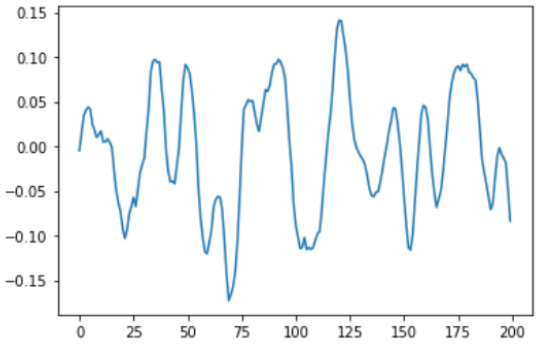
VII. Further Investigation

There are a range of further directions of study we could have undertaken on the dataset. We list some of them below.

- Conduct time series analysis on the sales and profit data, potentially grouped according to countries or regions of delivery, and incorporate these features into our machine learning models (as mentioned in Section III).
- Analysing the relationship between sales of individual products and geographical regions. We did do it to an extent in section V, but the effort could be greatly expanded.
- There was a 0.15 cross correlation between US sales and US lateness proportion when lag=10. Due to lack

of complete time data and lack of global trend, we did not pursue this further (see below diagram)

- It seems that in some specific countries like Canada, some items may have sold better. We had no time to investigate this.
- In the domain of customer behaviour, we wanted to investigate if that giving more discounts could incentivise more purchases. We did not have time to perform any statistical tests, but the tests would be done in a fashion similar to section III.
- One of our avenues for research was the correlation between orders per day and proportion of late orders that day for different product departments. No clear correlation was identified for departments 2 to 7 but for 8 through 12 correlations of magnitude 0.1 to 0.3 were seen. As within these departments we found each customer only ever placed one order per department (see section II-C) we could not explain this by "customers placing multiple orders". We also conducted statistical tests but found the results insignificant due to the weak correlation and also the small number of days within those departments with nonzero amounts of orders (around 10-50). However this could be subject for further analysis.



VIII. Difficulties Encountered

A variety of difficulties were encountered.

- It was very difficult to identify nontrivial relations within the data. Many indicators, e.g. sales, profits, scheduled dates for delivery, etc. exhibited surprising uniformity across the different regions.
- It was also difficult to construct the models without accidentally implying the answer within the features, or conducting the train/test split inadequately (something that would let the model memorise the answer). The "actual delivery time" within the lateness prediction was a key example of this, and preparing training and validation sets had to be done carefully, as can be seen in the "Methodology" subsection of Section V.

IX. Contribution

All authors contributed equally and are listed in no particular order.

References

[1] Buitink et. al *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research.