EDA report

December 7, 2017

1 Product backorders analysis

In Data analysis part a list of questions about the data set were answered. Questions to answer using the data:

1. How coomn are backorders?
2. Given that,how likelet are backorders based on the part risk flags?and how prevalent are they?
3. whats the relation ship between sales and forecast?
4. What's the relationship between potential issue and pieces past due are each represented by part risk flags or are thet unrelated concepts?
5. what's the relationship between lead time and backorders?
6. What aspects of supply chain represent the biggest risks?
7. Based on that risks what would I recommand improving first?

1.1 Data Wrangling

Dataset was acquired from https://www.kaggle.com/tiredgeek/predict-bo-trial.Dataset composed of one one file named Training_Dataset.the training file was opened and stored in a dataframe using python. the dataset contains the historical dataand it has 23 columns and 1687860 entries. the dataset totally have 100894 missing data also some entries of two columns include -99 values.The missing data is an example of Missing at Random (MAR) data mechanism where missing data is related to observed data.
Goal: Prepare the backorder dataset for EDA and Modeling
Tasks performed:

- Handling missing Data
- convert to binary
- Handling the outliers
- How common are backorders?

- Write the clean data into a new data frame further steps
  In this dataset, Every Categorical Feature includes only two values: 'Yes' and 'No.'for reducing memory usage binaries were converted from strings ('Yes' and 'No') to 1 and 0.
  Missing values in all columns of the footer in the dataset were represented as NaN, so I dropped the footer row.columns perf_12_month_avg and perf_6_month_avg have missing valu as -99.There is a strong correlation between perf_6_month_avg and perf_12_month_avg. So, linear regression would use to filing missing values. However another interesting point to note here is that many observations have both perf_12_month_avg and perf_6_month_avg as null, so linear regression cannot fill such values, and we need to see another approach there. Probably we would like to check for the central tendency of the data and replace the null accordingly. The data did not distribute normally. Therefore picking median to fill remaining values is a good choice.

1.2 Inferential Statistics

The calculation below shows how to handle missing data in lead time: 1.Proportion of orders that "went_on_backorder" for missing lead_time records. 2.Proportion of orders that "went_on_backorder" for non-null lead_time records. Went on backorder ratio for orders that they went on backorder is 0.66%.
The proportion of backordered products with missing lead time is 50% less than those without missing lead time. The proportion of backordered products with missing lead time is half of the products with no missing values.

Since the dataset was massive, I decided to reduce data by capturing data from the total sales volume which is a significant reduction in data for not much loss of fidelity. How I captured the total sales values is I used the cumulative sum of total sales volume.For data reduction, I captured 60% total sales volume, which is data was reduced to 7397 rows. Using data reduction may save some computing time and also presenting a cleaner dataset for the predictive model.