Guillaume Babo

Capstone Project

**Statistical Analysis Report**

Part backorders is a common supply chain problem. The main goal in this project is to identify parts at risk of backorder before the event occurs so the business has time to react.

After cleaning and wrangling the dataset to get it into a shape that was easy to analyze, I started my statistical analysis. My main focus was trying to get any insights from the data and see if there were any correlations. From that. I would be able to make a hypothesis.

**Count**

Since my target is to predict how many parts could possibly go on backorder, using the training dataset,t I made a pie chart to count how many parts went on backorder or did not.

Yes: 1676567 or 99.7%

NO: 11293 or 0.7%

From this result, I can tell that the dataset is **Imbalanced**.

Imbalanced classification is a supervised learning problem where one class outnumbers other class by a large proportion. This problem is faced more frequently in binary classification problems than multi-level classification problems.

**Correlation**

None of the data is in correlation with the backorder information.

**Insights**

With imbalanced data sets, an algorithm doesn’t get the necessary information about the minority class (in our case 0.7%) to make an accurate prediction. Using ML algorithms with this unbalanced dataset would probably lead to a biased and inaccurate prediction. The methods to deal with this type of datasets are widely known as ‘Sampling Methods’. Generally, these methods aim to modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by changing the size of original data set and provides the same proportion of balance.

**Conclusion**

I am dealing with an imbalanced dataset that is a binary classification where there are no correlations between the data and the backorder information. At this point, it is difficult to make a hypothesis based on the data. On the other hand many questions arose from my analysis:

- What strategy should I use to deal with this imbalanced dataset?

- No correlation at this point would make my dataset challenging?