# **CAPSTONE PROJECT2 PROPOSAL**

# **Product Backorder Prediction-Part 2: Machine Learning**

# **By: Hediyeh Ledbetter**

**Introduction**

This section includes model training and validation

1. Model selection
2. Train-test split of data
3. Model training
4. Model tuning and cross-validation

The training data set the amount of backoredered products are less than 1% of the whole prod- ucts, so the data is very imbalanced. Therefore I add up the data that "went on backorder" to this

Sample. For model selection, I want to use supervised learning models to predict “went on backorder” product according to what they have reordered.

I want to use supervised models logistic regression, and random forest(Bagging-based

Ensemble).

I used supervised learning because the process of an algorithm is learning from the training dataset and can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance and it is grouped into regression and classification problems.

Comparing logistic regression and decision tree:

* Both algorithms are really fast. There isn't much to distinguish them regarding run-time.
* Logistic regression will work better if there's a single decision boundary, not necessarily parallel to the axis.
* Decision trees can be applied to situations where there's not just one underlying decision boundary, but many, and will work best if the class labels roughly lie in hyper-rectangular regions.
* Logistic regression is intrinsically simple; it has low variance and so is less prone to over-fitting. Decision trees can be scaled up to be very complex, are more liable to over-fit. Pruning is applied to avoid this.

This problem of over-fitting is overcome to a large extent by using Random Forests, which are nothing but a very clever extension of decision trees.