Optimizing supply chains through predicting late deliveries & Customer Segmentation with Machine Learning

Jason Liu Doyle

Introduction

We now know that predicting late deliveries can **significantly enhance operational efficiency** and **reduce costs** associated with delayed shipments (Aljohani, 2023).

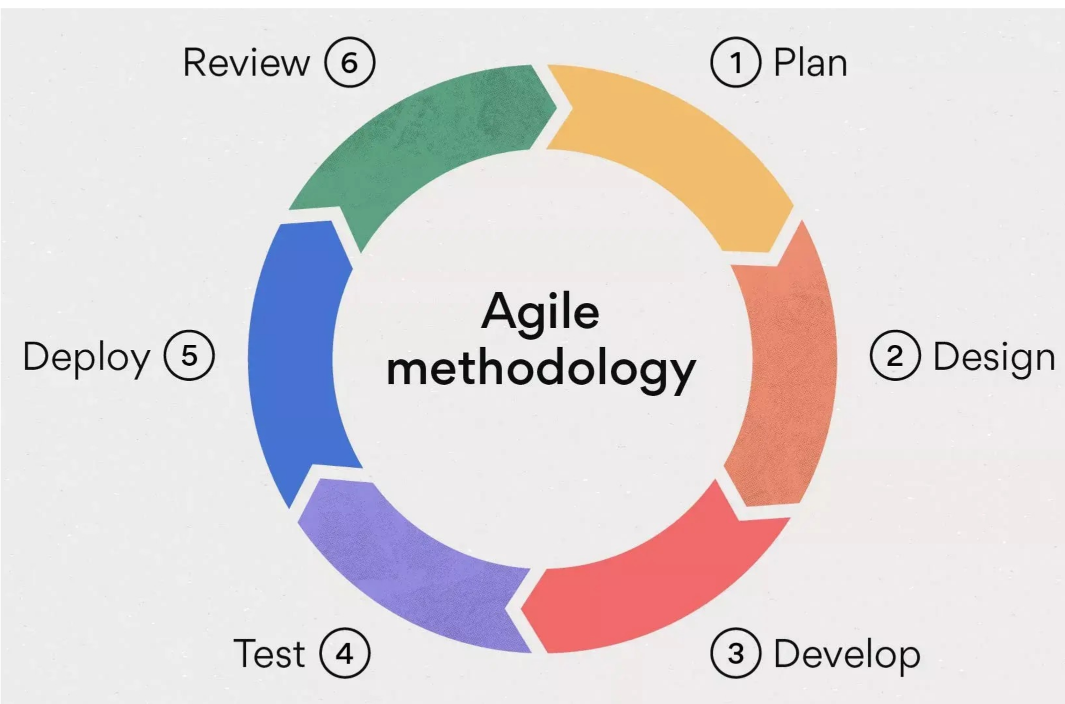
Similarly, having an understanding customer segments can be useful in **guiding marketing strategies** and **improving upon customer service**, which will eventually lead to **customer retention and acquisition** (SupplyChainBrain, 2023). For these reasons, I have chosen to focus on these two areas in this project.

Objectives

1. Implement machine learning models on data given to us by the business to predict whether orders will arrive late or on time.
2. Identify and analyze distinct customer segments to better understand their characteristics.

Methodology

- **Sprint 1** : Review of CA2 + Feedback, Brainstorm ideas, Define project scope, and Conduct initial research
- **Sprint 2** : Develop new models and perform exploratory data analysis (EDA)
- **Sprint 3 :** Test model efficiency and compare with previous CA results
- **Sprint 4:** Finalize models, prepare the report and presentation (focus on finalizing the deliverables, wrapping up the project).



Modelling

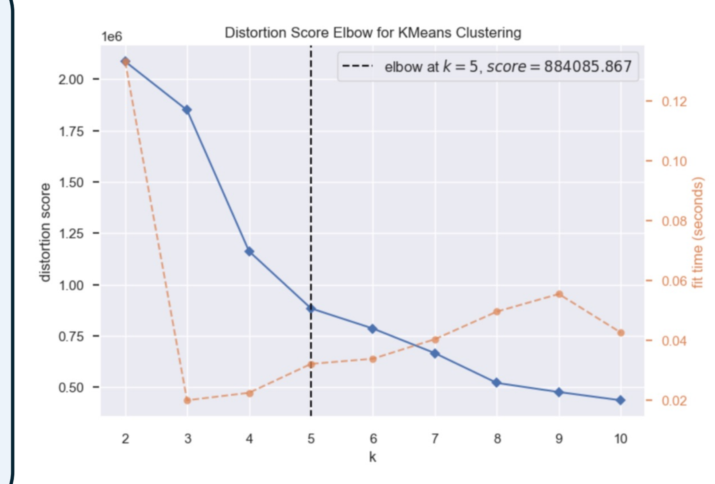
Objective 1

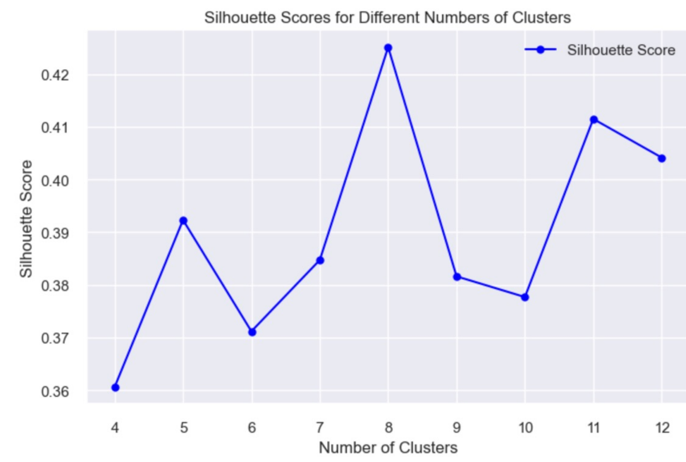
I implemented **8 classification models** (Supervised) on **50% of my data** and found the **top three models.**
I decided to swap out Decision Tree for Random Forest, they more often provide more robust and generalized predictions by combining the results of multiple trees which reduce the overall risk of overfitting (Breiman, 2001). I then swapped gradient boosting for the faster and more flexible XGBoost model (Bentéjac et al., 2020).

	Classifier	Accuracy	Recall	Precision	F1	ROC_AUC
2	Decision Tree	0.761932	0.727142	0.823604	0.772373	0.766273
3	Random Forest	0.752493	0.674725	0.848670	0.751767	0.851244
0	Nearest Neighbors	0.701706	0.560874	0.851399	0.676254	0.774921
5	Gradient Boosting	0.667051	0.486596	0.849798	0.618842	0.730039
4	AdaBoost	0.643696	0.428834	0.859175	0.572113	0.676973
6	Quadratic DA	0.593220	0.543641	0.663292	0.597536	0.647057
7	Neural Net	0.567826	0.948939	0.566219	0.709243	0.570356
1	Support Vectors	0.555462	1.000000	0.555462	0.714208	0.520173

Objective 2

To begin the clustering process I first conducted the dimension reduction technique **PCA** (Chandra, 2013) , on 100% of dataset and then then used the **elbow method** to find the optimal cluster amount





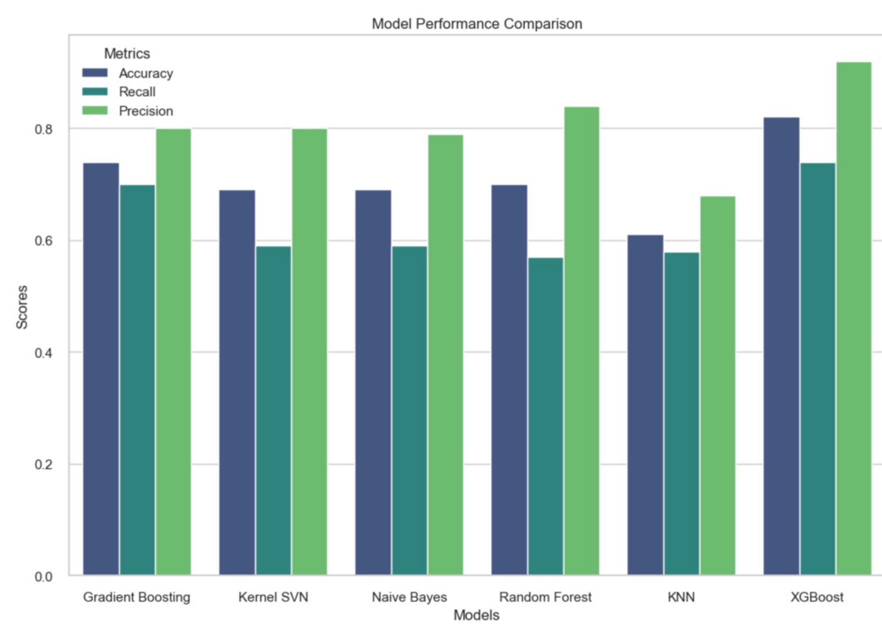
I then decided to run a **silhouette score** on a series of clusters that were 1) around the elbow score 2) didn't enter into a redundant amount of clusters. This led to a **range of 5-12**, and then I found **8 clusters** to be the peak score.

Findings

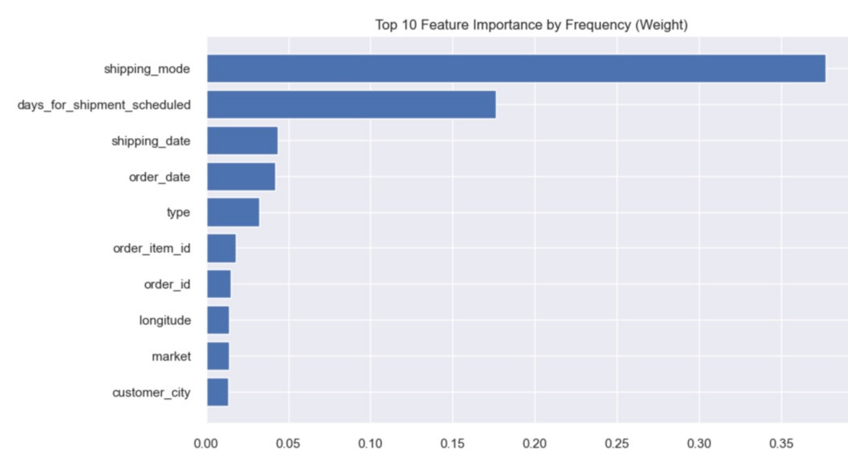
Objective 1

After running grid searches, adjusting parameters and validating the findings I found **XGBoost** to the best overall performer for predicting late deliveries

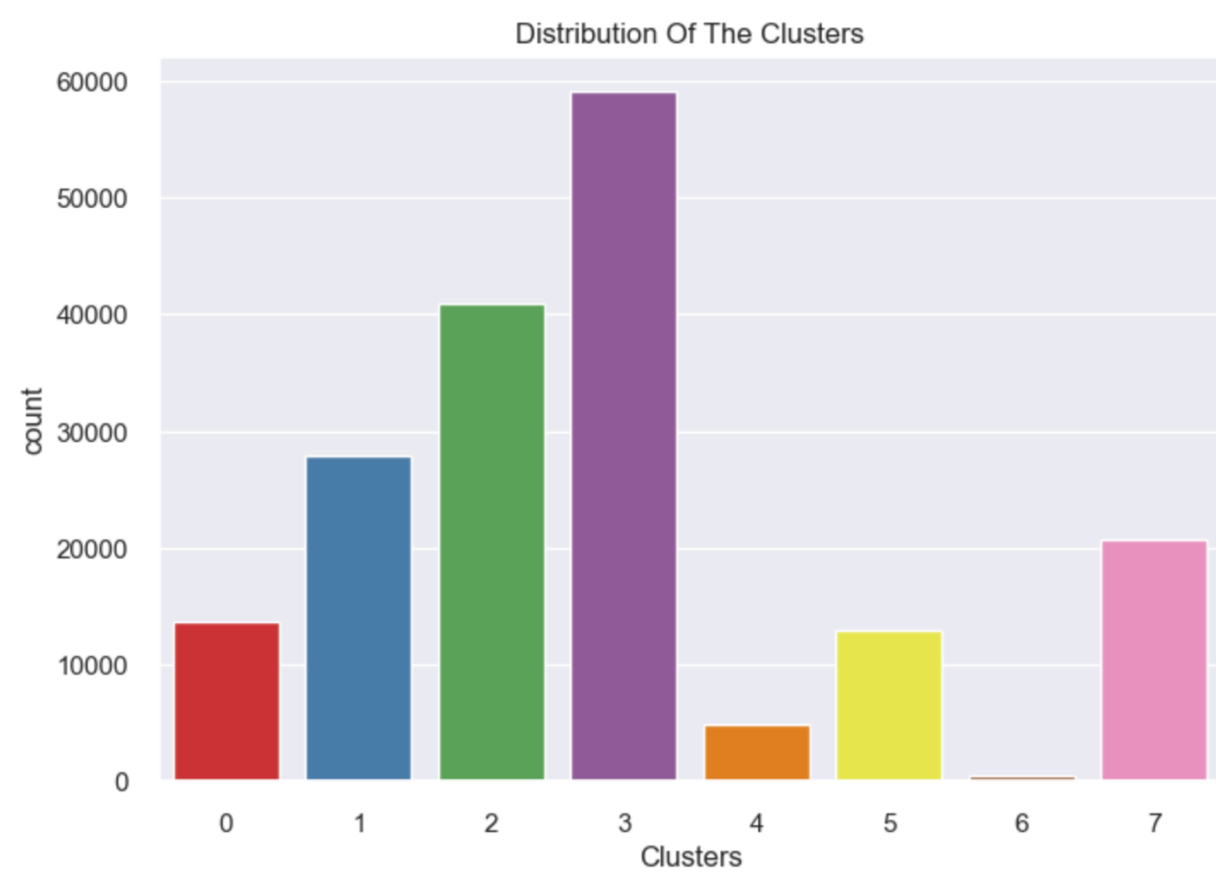
With **a recall of .74**, **precision of .92** and .an **accuracy of 82.**



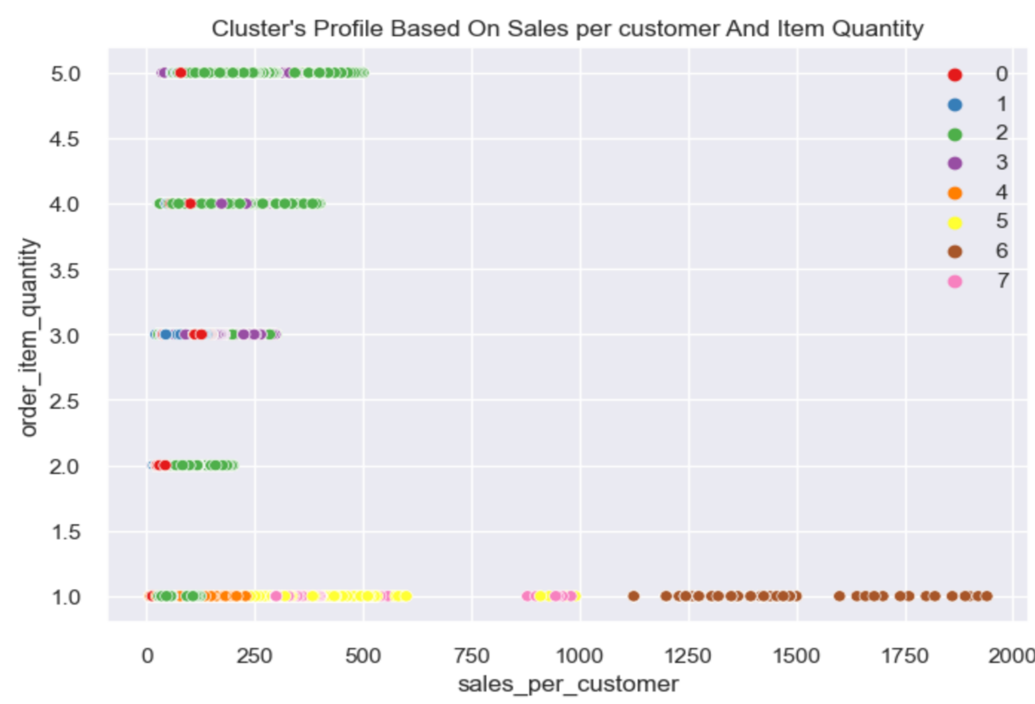
I also found the models' **most important features** in predicting late deliveries



Objective 2



We identified **8 unbalanced clusters**, which we were able to categorize based on **location, market, spending behaviour,** and **time-based spending habits.**



Insights

Top Predictors

1. **Shipping mode**
2. **Days for shipping scheduled**
3. **Shipping date**
4. **Order Date**
5. **Payment Type**

Clusters

Cluster 6 is the clear **high-value group**, with **high spending and profitability**, making it a key segment for business focus.

Clusters 0, 2, and 3 represent **low-value** customers who could benefit from strategies to increase profitability or minimize costs.

Cluster 7 **shows promise** with **moderate sales** and profitability, making it a potential growth segment.

Actionable insights

No.1 Action you can take today to drop late deliveries is increase the estimated delivery time to at least 3 days until supply chain team can improve the logistics

Priority areas for logistics to look into is shipping mode, followed by seasonality of orders (shipping date & order date features)

Conclusion

- XGBoost was the best classifier
- I identified eight customer segments and their characteristics
- Achieved a moderate silhouette score and conducted deeper EDA uncovering insights beyond CA2.

In future experiments I would like to
- Use 100% of the dataset
- try out slightly costlier models
- Collaborate with domain experts for further feature engineering.

Acknowledgements

- Breiman, Leo. "Random Forests." Machine Learning, vol. 45, no. 1, 2001, pp. 5–32, link.springer.com/article/10.1023/a:1010933404324, https://doi.org/10.1023/a:1010933404324. Accessed 24 Sept. 2024.
- Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2020) 'A comparative analysis of gradient boosting algorithms', Artificial Intelligence Review, 54(3), pp. 1937–1967. doi:10.1007/s10462-020-09896-5.
- Chandra, P.L. (2013) 'Methodological Analysis of Principal Component Analysis (PCA) Method', M International Journal of Computational Engineering & Management, 16(2), p. 32. Available at:https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6f5c37562374b9053d8212d7c97bbdd68cee2133 (Accessed: 23 May 2024).
- Aljohani, A. (2023). Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility. Sustainability, 15(20), 15088. Available at: https://doi.org/10.3390/su152015088.
- SupplyChainBrain. (2023). How Predictive Analytics Can Help Supply Chains to Thrive. SupplyChainBrain. Available at: https://www.supplychainbrain.com/articles/35396-how-predictive-analytics-can-help-supply-chains-to-thrive.