Clustering and Prediction Problem Framing for Sparse Demand Products

Raghav Tamhankar, Sanchit Khattar, Xiangyi Che, Siyu Zhu, Matthew A. Lanham

Purdue University Krannert School of Management

rtamhank@purdue.edu; skhattar@purdue.edu; che4@purdue.edu; zhu554@purdue.edu; lanhamm@purdue.edu

Abstract

In collaboration with a national retailer, this study focused on assessing the impact of sales prediction accuracy when clustering sparse demand products in various ways, while trying to identify scenarios when framing the problem as a regression-problem or classification-problem would lead to the best demand decision-support. This problem is motivated by the fact that modeling very sparse demand products is hard. Some retailers frame the prediction problem as a classification problem, where they obtain the propensity that a product will sell or not sell within a specified planning horizon, or they might model it in a regression setting that is plagued by many zeros in the response. In our study, we clustered products using k-means, SOMs, and HDBSCAN algorithms using lifecycles, failure rates, product usability, and market-type features. We found there was a consistent story behind the clusters generated, which was primarily distinguished by particular demand patterns. Next, we aggregated the clustering results into a single input feature, which led to improved prediction accuracy of the predictive models we examined. When forecasting sales, we investigated a variety of different regression- and classification-type models and report a short list of those models that performed the best in each case. Lastly, we identify certain scenarios we observed when modeling the problem a classification problem versus a regression problem, so that our partner could be more strategic in how they use these forecasts for their assortment decision.

Introduction

The complexity of managing retail assortment has grown due to the increase in product variety and uncertainty in consumer preferences. One of the key inputs into the assortment decision is demand. In this study we focused on estimating demand for products with very sparse demand patterns. Products that often are plagued by the worst case scenario of intermittent (or very sparse demand) patterns are medications and spares.

Business Problem to Analytical Problem Framing

Retailers build demand forecasts to better plan what their customers will want in the upcoming planning horizon. More often than not the demand forecasting problem is framed as a regression-type problem where various times-series and machine learning approaches are used to capture the signal from the noise. However, in the case of products that mostly sell at most one unit or no units in a planning window, a retailer might frame the problem as a classification-type problem, where the target is 0 if no units were sold, and 1 if one or more units are sold. In such a scenario, a propensity of purchase could be estimated. Often in forecasting, clustering of products based on season demand patterns a priori can yield improved predictive performance. However, it is unknown how such clustering would improve prediction in the sparse demand case, where the problem is framed as either a regression or classification-type problem. In either case, propensity (classification-case) or quantity (regression-case) predictions are used in the retailer's assortment recommendation engine to identify the right combination of products to provide to customers. Thus, our research questions are as follows:

Research Questions

- Can informative clusters be generated using popular unsupervised learning algorithms, and is there a business story about those clusters?
- How does clustering of products improve the predictive performance of models in a regression and classification setting?
- In which scenarios is the regression approach preferred to the classification approach when modeling very sparse demand products?

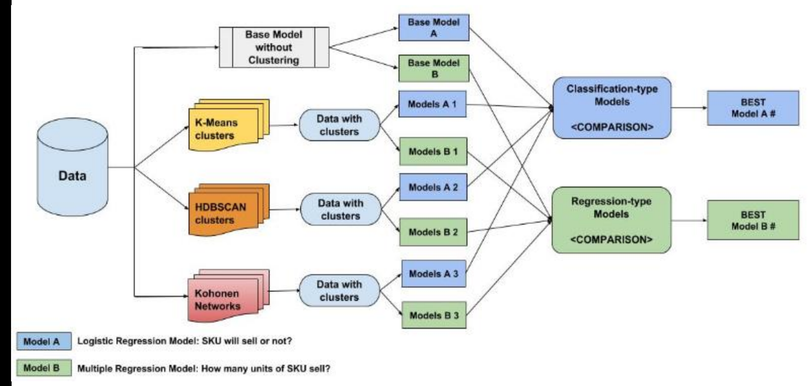
Methodology

Data

The dataset provided by the client contains sale performance of different SKUs at different stores for a timeframe of 2 years. Apart from the sale figures, the data set also contains parameters that describe products' characteristics, such as lifecycle, failure rates, sales coming from other channel, customer look up frequency, etc.

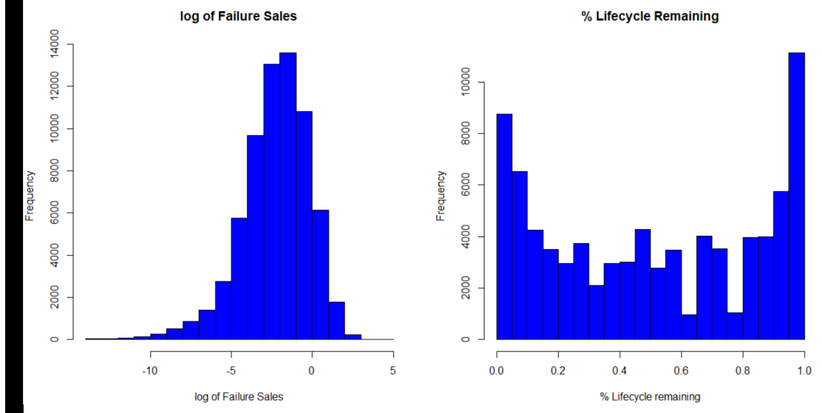
Schematic diagram

Below is a schematic representation of our analytics strategy. Essentially, we bifurcated our analysis by creating 3 analytical models. The grey colored model is the base model:



Data Exploration and Preprocessing

- All the irrelevant (insignificant to the objective) features were removed
- Missing values were imputed using a model-based imputation approach
- Numerical features were normalized to maintain scale-uniformity
- Skewed variables were log-transformed to get tighter distributions



Model Building and Comparison:

a. Data partition

Data was divided into an 80/20 train/test split, where the training set was used to train or build the models. The test set was used to assess, fine-tune, and compare the models for generalizability. Training and testing statistical performance measures were compared to identify and remove any models that might have overfit.

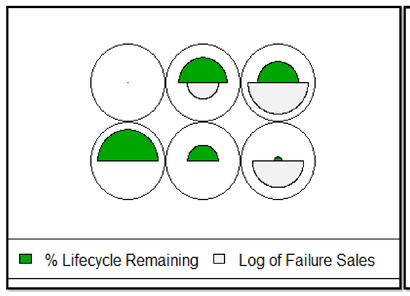
b. Clustering, Regression and Comparison

Unsupervised clustering was performed to group products with similar demand patterns and life cycles and further used as input feature in the supervised models for prediction.

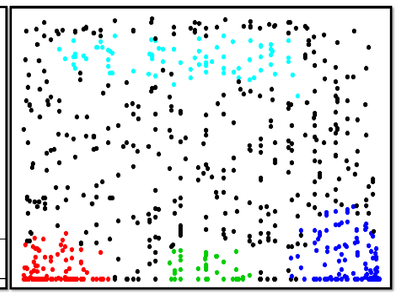
Results

Several clustering algorithms like k-means, SOMs and HDBSCAN were tested to cluster the products based on failure rates and percentage lifecycle remaining. This was carried out to identify products with similar demand patterns. SOMs segregated the products in 6 different clusters whereas HDBSCAN generated 4 different groups leaving out few products which couldn't be associated with any group.

Cluster Output-



Self-Organizing Map/Kohonen Network



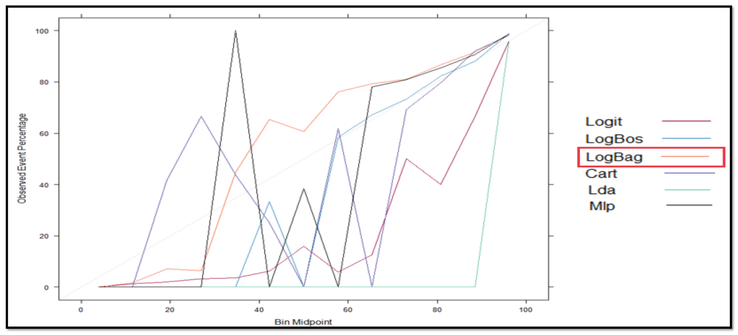
HDBSCAN clusters with noisy data-points

The clusters having high values of failure sales and intermediate value of percentage life cycle remaining correspond to fast moving products. Similarly, clusters containing products with low value of failure sales and extreme values of percentage life cycle remaining represent slow moving products. Different assortment strategies should thus be devised for these clusters to save costs.

The clusters obtained through the algorithms were used as predictor variables to predict the sales in the regression and classification models. Four different data sets were created as shown in the image on the left.

	Baseline Model				With Clusters				
Model	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC	
Logistic Regression	0.95241	0.98397	0.93889	0.95843	0.88535	0.95422	0.85584	0.89847	▼ -6.3%
Logistic Boosted	0.92294	0.88446	0.93943	0.91561	0.92552	0.92462	0.92590	0.92535	▲ 1.1%
Logistic Bagged	0.98517	0.97958	0.98756	0.98410	0.96308	0.97917	0.95618	0.96615	▬ -1.8%
CART	0.93942	0.95344	0.93342	0.94209	0.95826	0.96357	0.95599	0.95927	▲ 1.8%
LDA	0.93885	0.96511	0.92759	0.94385	0.98660	0.96511	0.99580	0.98250	▲ 4.1%
Neural Network	0.95868	0.93217	0.97004	0.95363	0.97756	0.93503	0.99578	0.96946	▲ 1.7%

Probability Calibration Plot



While predicting probability of sparse demand products, the KPI should be sensitivity, and how well the models perform based on a probability calibration plot. The reason being that propensities are used to rank products to be included in the assortment. While the bagged logistic regression and logit classification models performed essentially the same based on sensitivity and AUC, we recommend using bagged logistic model based on noticeable differentiation in these models based in the calibration plot.

Regression Model Comparison:

	Baseline Model		With Clusters		
Model	RMSE	Adjusted R-Squared	RMSE	Adjusted R-Squared	
MLR_Forward	0.56785	0.58642	0.52345	0.64960	▲ 10.8%
Poisson Regression	0.46397	0.86643	0.43014	0.90592	▬ 4.6%
Neural Network	0.67834	0.56827	0.72970	0.55075	▼ -3.1%
Decision Tree	0.72543	0.47462	0.73080	0.46838	▼ -1.3%
Decision Tree Bagged	0.65633	0.50946	0.61970	0.51980	▬ 2.0%

Performance KPI in this model is highest accuracy and amount improvement in the model. Based on these metrics, Zero-inflated Poisson Regression model is the best for predicting sales quantity. The R-squared obtained is 90.59% and has a 2 % lift in accuracy after addition of clusters.

Conclusions

In the retail industry, higher prediction accuracy is important for assortment planning, as it directly influences store sales, profitability, and customer satisfaction. Our study has focused on developing an analytical-based assortment planning framework to help retailers that face spare demand pattern to better predict future sales. After performing preliminary descriptive analysis, we consolidate the SKUs with similar life cycles and failure rate using multiple clustering methods. In the model comparison process, we observe that adding cluster to the model improves the prediction accuracy for LDA classification and Poisson regression. We would recommend the regression type forecasting for the given problem as the adjusted R-square obtained using Poisson Regression is very high (90.5%). The model takes into account the sparse nature of response variable and at the same time provides better interpretation. The probabilistic prediction generates products SKU rankings that can support retailers to prioritize production planning; demand is also aggregated on store level to help managers optimize shelf space. Quantity prediction results will facilitate the retailers to allocate resources and streamline supply chain efficiency to reduce supply chain costs and inventory holding costs. We plan to continue our study by further evaluating how seasonality can influence model performance.

Acknowledgements

We would like to thank our industry partner and professor for providing us continuous guidance, support, and feedback. This was a fantastic experience to collaborate with retail analytics professionals, be part of their team, and go through the process of designing and delivering results to decision-makers. The Purdue BIAC partially funded this project.