QUANTILE FORECASTS USING NEURAL NETWORKS FOR RETAIL INVENTORY MANAGEMENT WITH FIXED-TIME REPLENISHMENT INTERVALS

Aluno: Willians Cassiano de Freitas Abreu
Orientador: Marco Antônio Pinheiro de Cristo
Co-orientador: Juan Gabriel Colonna

Summary

- Introduction and Thesis Motivations
- Problem Definition
- Theoretical Foundations
- Research Hypotheses
- Simulator
- Preliminary Results
- Next Steps
- Schedule

Introduction and Thesis Motivations

Thesis Motivations

- Bad inventory management
- Medium-sized retailers
- A1 Curve
- High Opportunity Cost
- Low margins
- Out-of-Stocks: $456.3 Billion (56%).
- Overstocks account for the remaining $362.1 Billion (44%).
- Worldwide nearly $1.5 Trillion of merchandise annually is in an overstock position that creates a loss in revenue.

Inventory Distortion: An $800B Issue for Retailers Worldwide

Source: Tyco Retail Solutions

New study from IHL Group sponsored by Tyco Retail Solutions shines spotlight on growing global issue for retailers across markets.

https://www.retailitinsights.com/doc/inventory-distortion-an-issue-for-retailers-worldwide-0001

Problem Definition

Problem Definition

Retail Inventory Management: Make optimal purchase decisions under stochastic demand and lead times.

Fixed-time replenishment intervals: The inventory management in retail can be described by a Markov (or semi-Markov) decision process, with fixed-time epochs.

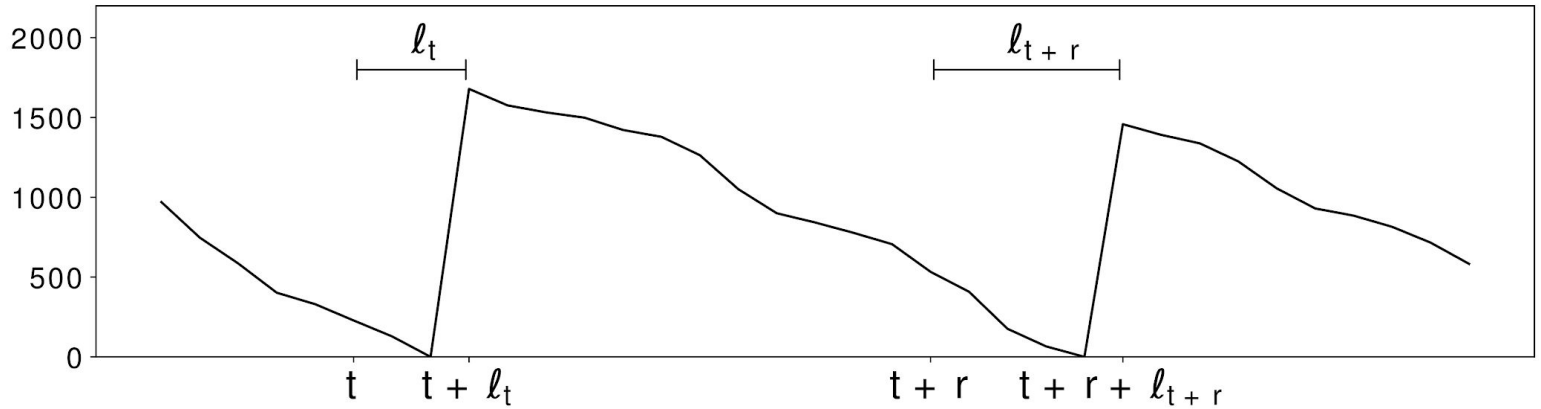
ABC Curve

Stock-out: A stock-out event happens when there is unmet demand for an item in a store.

Overstocking: A retailer store is said to be overstocked, if inventory is carried over between replenishment intervals.

Problem Definition

Make periodic purchase decisions under stochastic demand and lead times based on forecasted aggregate demand, accounting for the asymmetric cost of stock-outs and overstocking.



Target: $y_t = d[t + \ell_t : t + r + \ell_{t+r}]$

Model: $\hat{y}_t = m(D, L, r, \tau; \theta)$

Theoretical Foundations

Theoretical Foundations

Service Level: The service level indicates the chance that all customers will have their demand met on a randomly chosen t. The service level is used both as a parameter for inventory control models and a performance measure for such models.

Quantile Forecast: Denoted by $\hat{y}_t{}^{\tau}$. An estimator for Y with the property that the observed values of y should fall within the range [0, $\hat{y}_t{}^{\tau}$] with probability τ.

Pinball Loss: The quantile forecast is the solution to the following probabilistic minimization problem:

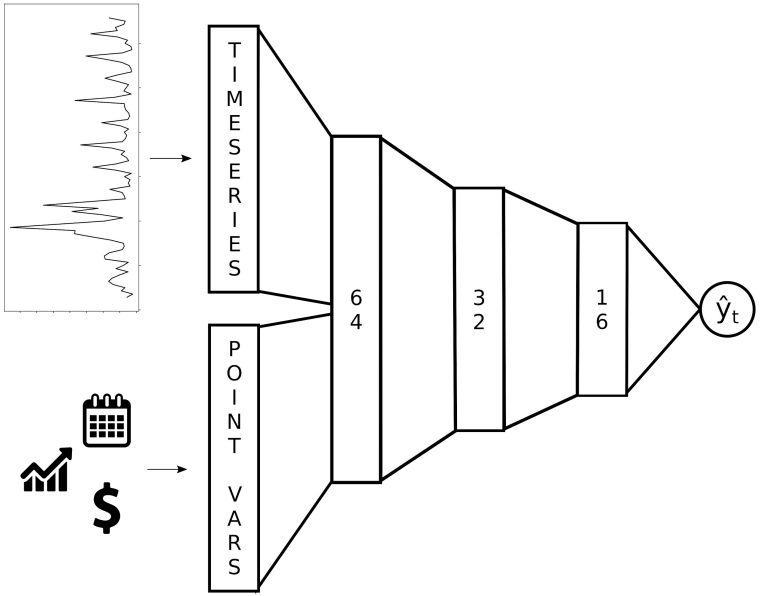
$$C = \sum_{y \sim Y} \tau[(\hat{y}^{\tau} - y)]^{+} + (1 - \tau)[(y - \hat{y}^{\tau})]^{+}$$

Research Hypothesis:

Neural Networks for Quantile Forecasts

Research Hypotheses

An artificial neural network model optimizing the Pinball Loss **learning the distribution of lead times and demand directly from data** performs better than models that assume **known and stationary distributions of lead times and demand**.



$$\hat{y_t}^{\tau} = m(D, L, r, \tau; \theta)$$

VS

$$\hat{y_t}^{\tau} = (r + \mu_L)\mu_D + (r + \mu_L)z_\tau\sigma_D + z_\tau\sigma_L\mu_D.$$

Research aims

Main objective: Propose and validate an effective quantile forecast model able to adhere to service level constraints when applied to a scenario where demand and lead times are stochastic, without making assumptions about the underlying probability distributions.

Specific Objectives:

1. Implement a simulator, able to reproduce the relevant retailer mechanics observed in the real world to evaluate our model and baselines.
2. Develop a baseline model that indirectly optimizes for C, making assumptions about the probability distributions, based on the relevant literature.
3. Develop models that directly optimize for C, using Artificial Intelligence techniques, that perform better than the baseline models, evaluated on the relevant metrics.

Proposed Method

Approaches for Inventory Management using prediction techniques:

- Estimate as a Solution
- Separated Estimation and Optimization
- Empirical Quantiles
- Multi-feature Quantile Forecasts

Our proposed method falls into the Multi-feature Quantile Forecasts category of inventory management using prediction techniques.

We build upon the work of **Bertsimas and Kallus** and generalize the work by **Oroojlooyjadid** et al. to applications in fixed-time replenishment intervals, complementing theirs newsvendor problem application.

Related Work

Bertsimas and Kallus:

Similarities: General Framework. Uses SVMs, DTs and RFs to make quantile forecasts;

Disparities: Do not take into account probability convolutions. Do not applies neural networks, nor related methods.

Oroojlooyjadid et al.:

Similarities: Applies Neural Networks;

Disparities: Ignores lead times.

Simulator

Simulator

Discrete event simulator implemented in SimPy. Processes:

Daily Sale: This process runs daily and consumes the item inventory by the amount sold that day. If the inventory is — or reaches — zero, a stock-out is recorded;

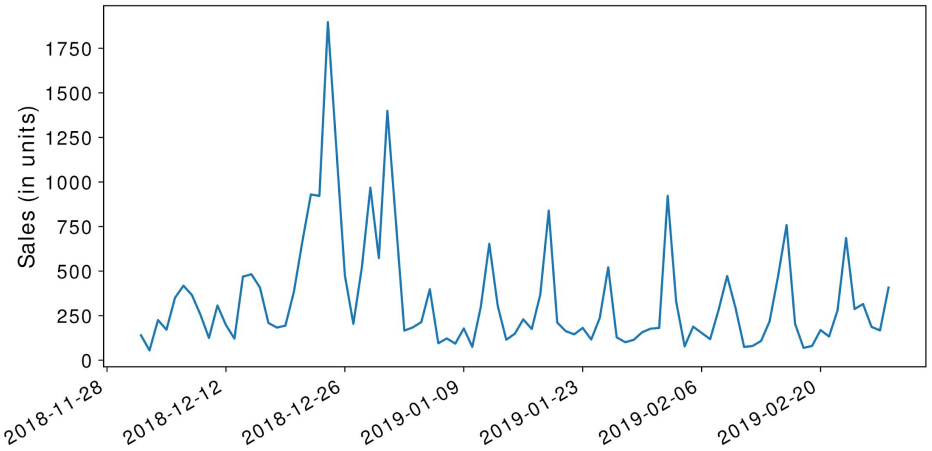
Inventory Management Policy: This process is executed from r to r ticks. It can observe the current inventory level and any information available on that tick — such as a quantile forecast — to make purchase decisions and start an Order Process for a specific quantity.

Order: This process is started from the Inventory Policy Process and is parameterized by the ordered quantity. It samples with replacement a lead time (l) uniformly from the past observed lead times set for that particular item, from our Orders collection (L). After l ticks have passed, it places the ordered quantity into the inventory Container. Hence, we do not need to make assumptions about the underlying distributions of lead times.

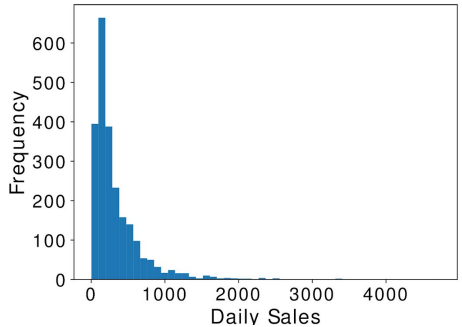
Case Study

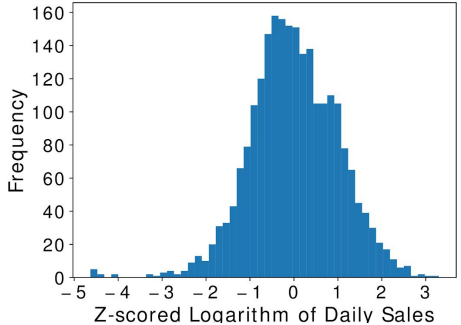
Case Study

Store	City	Annual Revenue
1	1	~R$ 62m
2	2	~R$ 48m
6	2	~R$ 47m
4	3	~R$ 40m
7	4	~R$ 37m
3	2	~R$ 23m
5	3	~R$ 10m
ALL	**ALL**	**~R$ 267m**



- Held inventory takes from 3 to 6 % of annual revenue (R$16m)
- The top 500 best selling items account for 51% of the company sales revenue (R$ 139m)





Preliminary Results

Definitions

Our method: An Artificial Neural Network with 4 Layers;

Baseline method: A Passive Aggressive Regressor;

Optimal method: Oracle forecast.

Definitions

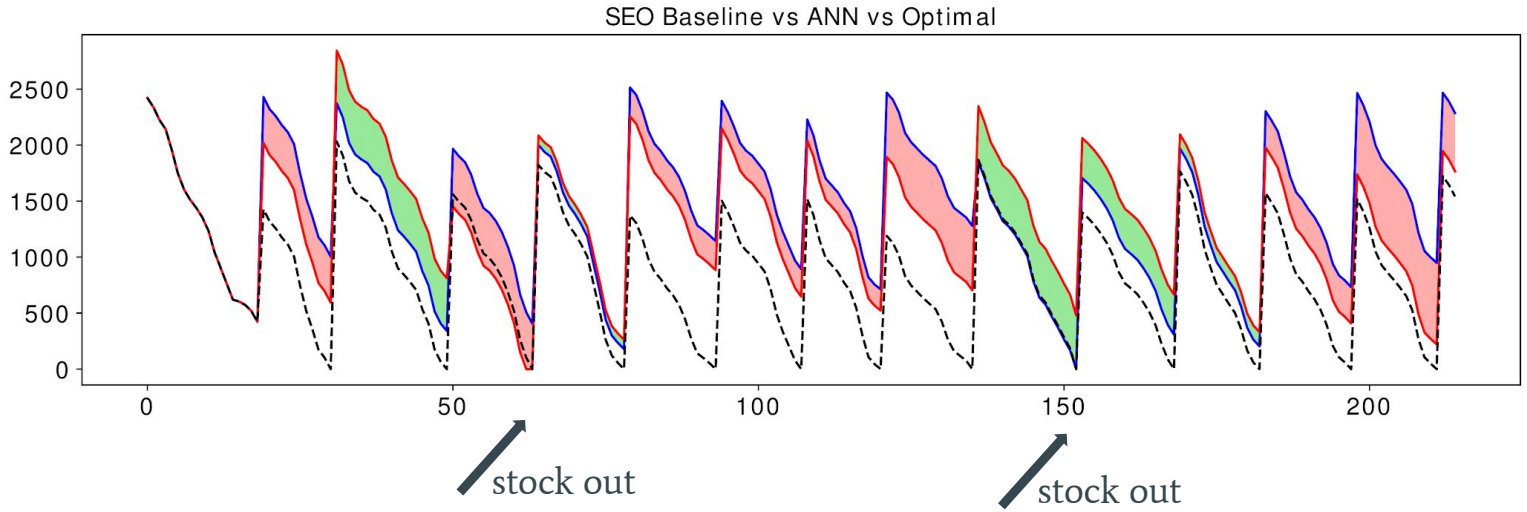
Our method: An Artificial Neural Network with 4 Layers;

Baseline method: A Passive Aggressive Regressor;

Optimal method: Oracle forecast.

Preliminary Results

Simulation results for a popular brand of beer, from January 15th, 2018 to September 1st 2018:



Our method Loss: 863.24
Our method stock outs: 1

Baseline Loss: 615.25
Baseline stock outs: 2

Next Steps

Next Steps

- Aggregated Pinball Loss
- B and C segments
- Temporal Connectionist Methods
  - Recurrent Neural Networks
  - Long Short-Term Memory Networks
  - Gated Recurrent Units Networks
  - Dilated Causal Convolutions
  - Residual Model

Schedule

Schedule

Goal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Solve the aggregate Pinball Loss problem																				
Experiment with B and C segments																				
Research Recurrent Networks as alternative architecture																				
Evaluation and Hyper Parameter Tunning (RNNs)																				
Research Long Short-Term Memory Networks as alternative architecture																				
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Evaluation and Hyper Parameter Tunning (LSTMs)																				
Research Gated Recurrent Units Networks as alternative architecture																				
Evaluation and Hyper Parameter Tunning (GRUs)																				
Research Dilated Causa Convolutions Networks as alternative architecture																				
Evaluation and Hyper Parameter Tunning (TCNNs)																				
Research Stochastic Temporal Convolutional Networks as alternative architecture																				
Evaluation and Hyper Parameter Tunning (STCNNs)																				

Thank you