September 2020

DYNAMIC THRESHOLD OPTIMIZATION FOR REORDERING OF CONSUMABLE USING DEEP LEARNING

HP INC

Dynamic Threshold Optimization for Reordering of Consumables Using Deep Learning

In Managed Print Services (MPS) service delivery space, in order to make sure an uninterrupted printing experience for the customer, it is very important to deliver the supplies (consumables) on time. While delivering supplies on time is an important goal from customer satisfaction perspective, but it has been observed that (many a times) for achieving this, the supplies delivery is made too early, which leads to early swap (of the supplies) in the printers. The ink left in the previous toner/supply (a.k.a. residue), results in wastage of supplies.

The solution provides a “deep learning” approach for reducing the supplies wastage, while maintaining the uninterrupted printing experience intact. This optimization is achieved by virtue of setting up dynamic threshold for ink level, at which delivery of the supplies will be initialized. The solution considers the historic print usage pattern of the given customer and the necessary printer model features, to come up with the appropriate threshold level (at which the new supplies should be ordered). The solution aims at reducing the operational cost of the MPS services by ~$15 million (for the upcoming year).

Managed Print Services (MPS) is a comprehensive suite of hardware, software and services delivered in a contractual engagement. In these engagements it is important to ensure right customer satisfaction (by making on-time delivery of supplies) but at the same time (from a profitability standpoint) we must also ensure that these deliveries are done in a cost-effective manner (by making sure that we control the wastage of supplies). Balancing these two aspects (customer satisfaction and supplies wastage control) at times becomes too tricky and complex task, as we may end up delivering supplies early, which may result into early toner replacement, leading to wastage of supplies.

Currently, supplies consumption rates are not monitored daily, due to which many a times toners do not reach on time. There are scenarios when consumption rates may change suddenly, toners may not be reporting continuously, or toner life cycle may be very less (e.g. 15 days), in such cases toners may not reach at customer premises on time, which causes (unwanted) idle printing durations, resulting into bad customer experience.

Normally in MPS business, supplies fulfillment for a given printer, is driven by setting a static threshold. Static threshold is set higher, due to which toners are shipped early, resulting into early replacement of the toners (again leading to high residue levels in the replaced toners).

Another important point to note is that for a given printer same static threshold is set for all the consumable types (black, yellow, magenta, cyan), which results into higher residue levels. It is also evident that in this entire process manual intervention is required to monitor and change the threshold.

As per our study:

1. High Supplies Cost is primarily due to:
    a. High Residue
    b. Frequent (uncontrolled) Shipments
2. Idle Device time is due to:
    c. Late shipments

Our study also reveals that 95% of the devices (in dMPS customer printer fleets) are configured with static thresholds (supplies ink level at which new order for toner is raised), due to which on many occasions either the toners are being sent too early or too late. Static mechanism of threshold management results into.

- High Residue
- Customer dissatisfaction

Remaining 5% of the devices are using Liner regression algorithm, at toner level, to predict the threshold, but this too does not address the issue. Still toners are being sent either too early or too late, again leading to same issues as with static mode of threshold management.

The solution (based on Deep Learning) attempts to create a perfect balance by satisfying both the parts of the overall equation (Customer satisfaction and Operating cost).

The solution aims at determining the optimum threshold level, to raise a request for toner replacement. In achieving so, it considers the following data points as input to the model:

1. Daily Ink usage timeseries data
2. Features associated to Device usage
    - Type of customer using the printers (for example the industry it belongs to)
    - Location of the device
3. Supplies Delivery mode (e.g. by air, by land etc.)

The above factors play a key role in building a balanced threshold prediction model for achieving the desired result of on-time delivery and control of supplies wastage.

The ink levels and corresponding timestamps are normalized and divided into continuous segments/vectors of window size 10 (~10 days of ink consumption). The information about types of customer and location of the device is transformed as one-hot vector representations. Further, the data is divided into training set (80%) and testing set (20%).

The following figure shows the architecture of the deep learning model:

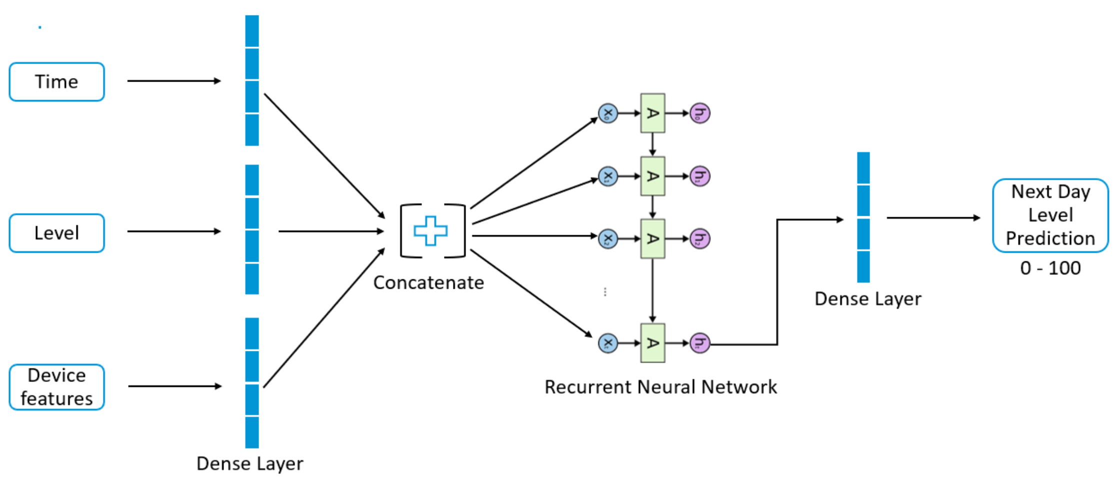

Figure 1: Model Architecture

The vector representations of different input features are passed through a dense layer which act as an embedding layer, as it transforms the input into higher dimensional form. The outputs of the first dense layer are concatenated and passed through a Recurrent Neural Network (RNN). RNN is a generalization of feedforward neural network that has an internal memory. It is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. RNN is very essential part of the architecture as it can capture the dependency of current ink level with the past ink level usage pattern in the sequence. The output from RNN is passed through another dense layer, whose output represents the probability distribution of next day predictions between 0 to 100 levels.

To estimate the expected duration of the shipment, Delivery expectation factors are used. It is a mapping of types of delivery to estimated days, based on the availability of parts in the nearby inventories of the customer site.

Results:

Model Test Accuracy: 88.02%

Field Accuracy: 90.21%

Potential Advantages:

1. The solution is more responsive to “changes in rate of consumption” and captures these changes effectively in comparison with static and it can help eliminate manual interventions for setting of replacement thresholds.
2. This Baseline model (Deep learning algorithm) is showing great improvement (in comparison with Static) and there is further scope for refinement in model, this will help in achieving the target of “reduction in residue” and better “on-time delivery accuracy” and this will continuously learn from new data distributions resulting in better delivery accuracy and less wastage.
3. Static Residue was 12.7% in FY-2019 whereas this solution has reduced it to 6.2%

References:

https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

https://www.tensorflow.org/guide/keras/rnn

Disclosed by Biswajit Ojha, Himanshu Tiwari, Ojha Hemant, Rohit Dokania and Ashish Gupta, HP Inc.