DataStorm 3.0
Semi Final - Report

Team Crypto
University of Moratuwa

April 8, 2022

Team member details

Team members 1: **Manjitha Kularatne**
Team members 2: **Nuwan Bandara**
Team members 3: **Dasun Prematilake**

Other details

Github repository :

1 Introduction

Many businesses and industries are in a deliberate requirement for accurate forecasting since of the uncertain nature of their business environments due to various factors such as the innovation-based transformations on different verticals. These challenges are critical and decisive in several industries, including retail businesses, where stakeholders are intuitively obligated to make key decisions in a short amount of time. These circumstances lead the business forms to turn towards advanced analytics for better outcomes through data-driven decision making.

Therefore, forecasting sales is one of the most fundamental problems most business chains have which directly result in improving project revenues, adequate preparation for the necessary supply, reduction of wastage, and better manage storage warehousing if the forecast is accurately positioned. In the given challenge, it is required to derive insights from the data and better estimate the sales four weeks for a retail chain which has previously used traditional forecasting methods to estimate projected sales for each item across stores and found those approaches to be inaccurate.

In the state-of-art machine learning approaches for time series forecasting, global forecasting models (GFM) are found to outperform traditional univariate models which work on isolated series. However, one key problem of GFMs is that they are not being localized enough to a particular series even though they share the same set of parameters across all local time series, especially in the context of multi-variate data, where the correlation within same data cluster is high. Therefore, in this stage of the challenge, we compared the performance of using a localized multi-variate long short-time memory (LSTM) model with an ensembled univariate model, cascaded with a fully connected neural network (FCN) to obtain a performance similar to that of GFMs in multi-variate time series forecasting. Our latter approach incorporates two base forecasting models: LSTM model and transformer architecture-based model as univariate models, which were then combined through linear stacking ensembling methodology. As per our evaluations of the challenge, it is found that our ensembling approach with cascaded FCN outperforms the localized multivariate model. We believe that further investigations on the optimizations of both models are needed for obtaining a conclusive remark in this regard.

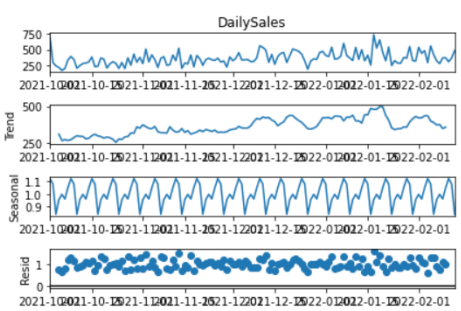
2 Methods

2.1 Pre-processing and feature engineering

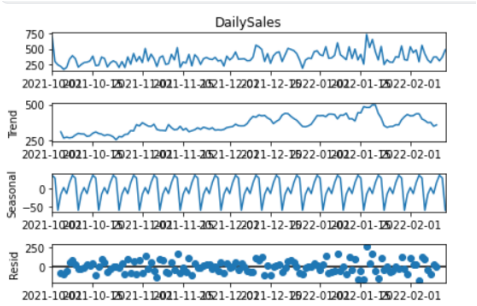
As the first stage of the pre-processing data in the first phase of the competition, we have analyzed the seasonal patterns and the frequency changes of the time-series dataset if any. Through our visual observations and analysis, we realized that there exists no considerate seasonality in a group of time-series data where almost all the patterns were seemingly random due to other variables which are correlated. However we experimentally changed the daily formatted data of the training dataset to the weekly basis with the help of **datetime** python in-build library to consider the promotion variables in account for multi-variate model. Then sorted according to item-Code and the week basis data while adding a unique ID to the help of the model train. Howsoever, we believe that, the abrupt reduction of data from daily to weekly basis resulted in a considerate insufficiency in data, manipulating the multi-variate model to perform weakly in a generic circumstances. For the ensembled-cascaded approach, we have decided to use the daily basis for all the

training, validation, and testing phases of two models to avoid above issue.

For the ensembled-cascaded approach, the data frames have to post-processed as the 2 models run and predicted data receives, it has to arrange according to the submission format by sorting and filtering.



(a) Seasonal decomposition of category 1 using multiplicative model



(b) Seasonal decomposition of category 1 using additive model

2.2 Multi-variate LSTM approach

We used a multivariate LSTM to approach the said problem in the semi-final. Given the problem, we realized that multivariate LSTM would be a suitable option to address the effects of discount type, discount amount, week number, etc. on the weekly sales. The input elements were fed a bidirectional LSTM with 200 units which is directly connected to a 1D convolutional layer with tanh being the activation function of the 32 filters used. The output from the convolution layer is once again fed to a bidirectional LSTM with 150 units. This structure is adjacent to three fully connected dense layers (dropout is employed in the first two dense layers).

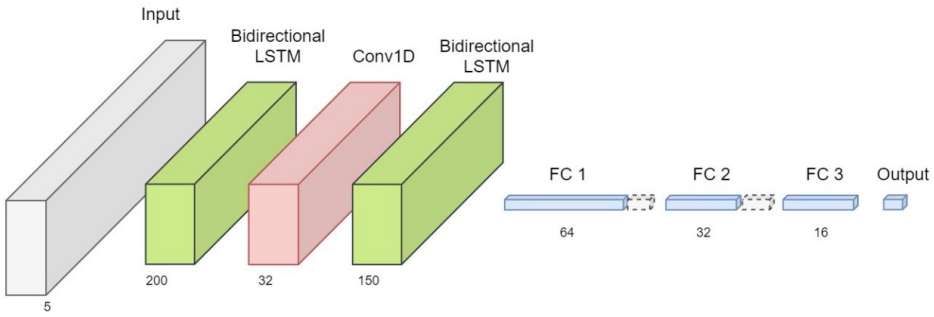


Figure 2: Proposed LSTM model as a localized multi-variate architecture to obtain the final prediction result

2.3 Localized univariate models for ensembled-cascaded approach

2.3.1 Transformer-based model

Transformers have been utilized for time series forecasting tasks in recent years because of their strong characteristic of self-attention in which they retain direct connections to all previous timestamps, thus allowing the information to flow in long sequences. In our approach, we utilized both the single-step transformer model and the multi-step transformer model to validate the best performance with a minimal number of model depth. Both the architectures consist of a positional enocoder, transformer encoder and a decoder layer

while the single-step output window size is one whereas the multi-step output window size is seven. The optimizer for the models was Adam and the utilized loss function was the mean absolute error. Through our experiments in both models, we found that since the input length of the model through the daily sales data is not sufficient enough for a deeper architecture, the single-step transformer performed better than its counterpart. Further, we found that these generic transformers performed poorly in decomposition and thus, looses the sufficient accuracy in the evaluation.

2.3.2 LSTM + Conv1D

In this approach, 1D convolution was first applied to the sales vector to extract the features. 60 such filters were used, and the extracted features were then fed to two conjoined LSTM layers. Here, LSTM layers were used instead of simple RNN layers to avoid the problem of vanishing gradients by allowing the model to remember long-term dependencies. Each LSTM layer consisted of 30 units and the outputs from the second LSTM layer are fed to a fully connected layer which is then directed to a single output node indicating the next term in the sequence or in other words, the predicted sales value. The weights of the model were updated based on the corresponding derivates of the Huber loss function and SGD with momentum was used as the optimizer of the model.

2.3.3 Ensembled-Cascaded Approach

Testing results from two base models; transformer and LSTM + Conv1D were weighted to obtain the result. The intention was to combine distinct behavioral features, including sequential patterns and the localized attentions, extracted from the utilized models. Higher weights were given to LSTM model considering its adaptability to the given dataset through our experience in the previous round.

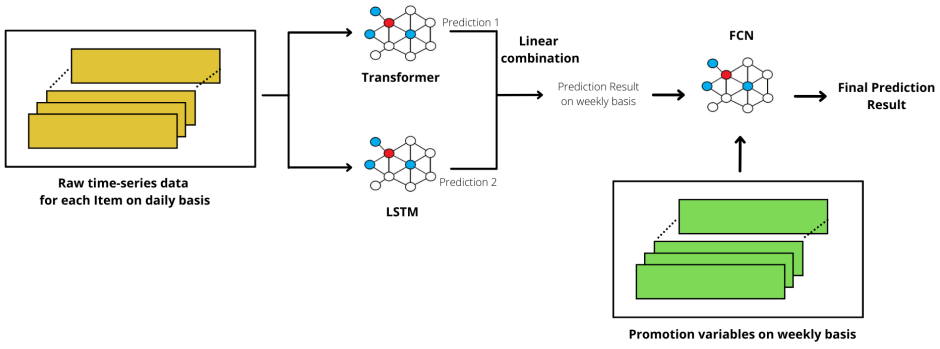


Figure 3: Proposed ensembled-cascaded model with two base uni-variate architectures and the FCN to obtain the final prediction result where the weight for transformer is 0.47 while the LSTM weight is 0.53

$$Output_{univariate} = \alpha * output_{transformer} + \beta * output_{LSTM} \quad (1)$$

$$Output_{FCN} = W^T * (Output_{univariate}, Promo_{Type}, Pr_{selling}, Dis, Days) + b \quad (2)$$

where $Promo_{Type}$, $Pr_{selling}$, Dis and $Days$ refer to discount type, selling price, discount value and the number of days of promotion respectively. Further, W and b are the weight and bias matrices while the output from FCN is then followed by sigmoid activation as the activation function for non-linear transformation. Here, (X, Y) means the concatenation where X and Y are vectors.

2.3.4 Modification to the loss function for reduced under-estimation

In the description of the challenge, it is necessary to minimize the underestimated error as much as possible while maintaining a lower overall error since, in accordance with the business case, underestimation leads to situations where the retailer would potentially encounter a loss of sales because an adequate amount of stock is not available in the store when it should have been.

In order to address this issue, we modified the implemented loss function in the following manner:

while *batch* **do**:
  $loss_{total} \leftarrow 0$
  $loss_{under} \leftarrow 0$
  $loss_{over} \leftarrow 0$
  $i \leftarrow output_{model}$
  $k \leftarrow target$
  if $k \geq i$ **then**
    $loss_{under} \leftarrow |target - output_{model}|*$
  else
    $loss_{over} \leftarrow |target - output_{model}|*$
  end if
  $loss_{total} \leftarrow (0.7 \times loss_{under} + 0.3 \times loss_{over})$
end while

Here, loss* is calculated through mean absolute error in transformer model for each item in the batch through,

$$l(x, y) = l_1, ..., {l_N}^T, \quad l_n = |x_n - y_n| \tag{3}$$

where N is the batch size. The linear weighted approach is implemented to orient the model more toward reducing the underestimation error.

3 Results

Through our two approaches for items with promotion data, we were able to obtain the following results for validation sub-datasets and the corresponding python implementations are added to the GitHub repository. For the items with no promotion data, we implemented the same uni-variate ensemble model with time series data as in the Storming round of the competition. For the items with promotion data, through exploration of the results, we realized that the ensembled-cascaded is slightly inefficient in learning long-term forecasting, while the multi-variate LSTM model can learn the sequential patterns in a decent manner with fewer layers. However, there is no room for a deeper network in LSTM domain because of the insufficient data with respect to weekly data clusters. The cascaded feature addition to the prediction outputs of the ensembled model through FCN enhanced the performance of that approach through filling the feature gaps

in the uni-variate data and we believe that these concatenated promotion features is the major reason in the superior performance of ensembled-cascaded model over the multi-variate localized model. In contrast, we believe that better interpolation techniques for filling the null data in the promotion dataset may also increase the performance of the multi-variate model. Therefore, we think that there should be further investigations in this regard to build a conclusive remark. In addition, We believe that the success of the ensembled-cascaded model is an obvious result of the different features learned from different modalities in divergent behavioural domains.

Table 1: Here, mean absolute percentage error is considered as the evaluation parameter for testing model performance on validation sub-dataset

Results from validation sub-dataset	
Models	Validation Performance
Ensembled-cascaded approach	48.88
Multi-variate LSTM	51.96

4 Discussion

We believe that the performance of the ensembled-cascaded model could be further improved using optimum linear coefficients for stacking and by exploring other techniques for ensembling such as using a fully connected neural network. Thus, it is obvious to say that ensembling methods of different selected models are a great pathway to study further to obtain better forecasting results which satisfactorily perform in the presence of abrupt technological innovations in the industries, while learning distinct characteristics from the uni-variate data.

We believe that retail management should obtain a defined set of external parameters which affect each item since it could lead to building a better multivariate ensemble model which can sufficiently address the non-promotional data of a particular item with the promotional data. In addition, it is better for the management team to intervene in a deep market analysis to understand the technological and research reasons behind the abrupt changes in the market in parallel with these kinds of forecasting models. Furthermore, the failure of univariate models depicts that the stability of the market is not assured but with long-term sequential dependence and thus should not be expected. Hence, we believe that the forecasting strategy of the company should further investigate deeper ML models which studies the possible market decisive factors which may be placed in collaboration with a ensembled and/or multi-variate sample space. In addition to required business insights, with promotion generally there is a positive impact since weekly sales have increased than it was maintained in other weeks. Its better if management and/or sales team can apply a loyalty card service, in order to track sales in a feasible way with more variables. Analytical methodology can be automated with promotion *percentage/amount* to increase profitability.