EFFICIENT SHUTTLE OPERATIONS: HARNESSING GLOBAL TIME SERIES FORECASTING USING TIDE MODEL ARCHITECTURE AND LINEAR PROGRAMMING FOR CAPACITY OPTIMIZATION

Wilhelm Henri R. Alegrado
Jhon Vincent A. Gupo

Operations Excellence
Western Digital Philippines, 109 Technology Ave., SEPZ, Laguna Technopark, Binan, Laguna, Philippines 4024
wilhelmhenri.alegrado@wdc.com, jhon.vincent.gupo@wdc.com

ABSTRACT

Shuttle services are integral to Western Digital Philippines' manufacturing operations, facilitating the daily commute of employees to and from the company. Employees hailing from various locations in Southern Luzon and the Greater Metro Manila Area are transported on over twenty designated shuttle routes and twice daily for both day shift and night shift. Managing such a complex shuttle operation presents challenges in terms of efficiency and cost due to the current reliance on manual analysis for forecasting and anticipating shuttle capacity which was exacerbated by the recent workforce restructuring in Western Digital Philippines.

In this study, the authors employed machine learning techniques, including time-series forecasting and linear programming, to develop a machine learning pipeline. This pipeline utilizes historical shuttle headcount data and various data to determine the recommended shuttle capacity for each route. Model training and validation were conducted using the MASE (mean absolute scaled error) metric to select the optimal time-series forecasting model. Through this process, the chosen model was a global time-series forecasting model utilizing the TiDE (Time Series Dense Encoder) architecture, which incorporates relevant exogenous variables such as holidays, route location, work shift schedule, product volume, and employee overtime count.

The results of the study show that the forecasted shuttle capacity generated by this pipeline aligns with or even improves the actual recommended capacity in 81% of the trips during the validation period while the remaining 19% show either undercapacity or overcapacity. While there remains room for improvement in the pipeline, the significance of the study lies not only in its potential to enhance efficiency and cost optimization but also in the reduction of man-hours spent on manual analysis, showcasing the business value of adopting a machine learning approach in managing shuttle operations.

1. 0 INTRODUCTION

Shuttle operations in Western Digital Philippines facilitate the commute of its employees from and to the company in over twenty established pick-up locations around the Manila area and Southern Luzon area. Given two shifts (day shift, night shift) and two directions (inbound and outbound of the company), the shuttle operation scales rapidly which presents a significant operational cost for Western Digital Philippines and makes the efficient use of shuttle capacity utilization to always be a high-impact operational target.

1.1 Manual Shuttle Capacity Analysis

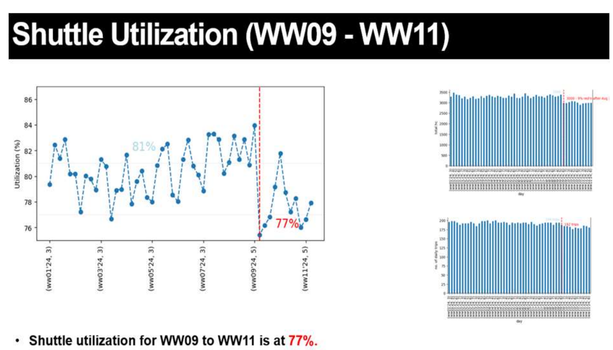

Fig. 1. Sample manual shuttle capacity analysis. Overall shuttle utilization is monitored, and capacity adjustments are recommended based on target shuttle utilization.

Table 1. Recommended Shuttle Capacity

Shift	Route	Current	Recommendation
Day	San Pablo	4 buses	3 buses
Night	San Pablo	3 buses	2 buses
Day	Cabuyao	4 buses	3 buses
Night	Cabuyao	3 buses	2 buses + 1 van
Day	Lipa	2 buses	1 bus + 1 van
Night	Calamba	2 buses	1 bus + 1 van
Night	Silang	2 buses	1 bus + 1 van

Fig. 1 shows a sample shuttle capacity utilization issue encountered. Table 1 shows the recommended shuttle capacity to address low shuttle capacity utilization. These analyses are typically done spontaneously and after the fact, which means shuttle operation losses are already incurred.

1.2 Machine Learning

As a general definition, machine learning is a branch of artificial intelligence (AI) that focuses on using data and algorithms to model the way humans learn while improving its accuracy through training and validation.

Thus, machine learning involves the development of models that can generalize patterns and relationships from input data to make informed predictions or decisions about new, unseen data. This capability to make data-grounded decisions offers the potential to supplement or even replace manual decision-making processes.

In particular, the study explores two machine learning tools to tackle the analysis of shuttle capacity utilization: global time series forecasting and linear programming.

1.2.1 Global Time Series Forecasting

Time series forecasting is a statistical and computational technique used to predict future values of a time-dependent variable based on historical observations. Particularly for this use case, future values of shuttle head count are predicted given historical values.

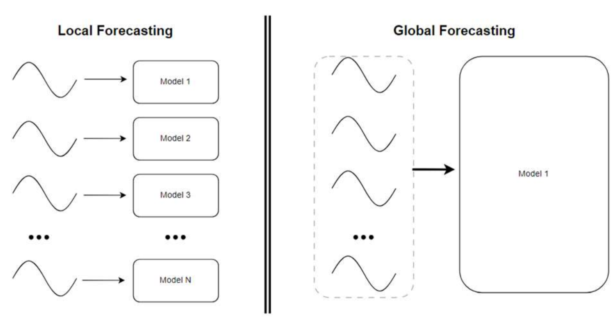

Fig. 2. Comparison of local time series forecasting and global time series forecasting

At a base level, a traditional or local approach to time series forecasting model would involve training one predictive model for each given time series. However, a local approach wouldn't be able to fully capture the commonalities and dependencies involved across the multiple time series instances in addition to the difficulty of maintaining and monitoring separate models for each time series instance.

On the other hand, a global time series forecasting approach involves training only a single global predictive model using multiple time series instances as its training input. Thus, the global predictive model has a larger training set and can leverage shared structures across the targets to learn complex relations leading to better predictions. As shown in Fig. 2, this approach resolves the problem of maintaining multiple time series models because it requires fitting and maintaining only one single model.

Global time series forecasting models also have the added benefit of being able to be trained on exogenous variables which are variables that influence the target time series and can help to produce better predictions.

1.2.2 Linear Programming

Linear programming is a mathematical method used to determine the best outcome or value of a mathematical model given an objective function and a particular set of constraints.

Linear programming is utilized in various fields such as operations research, economics, engineering, and other fields to optimize resource allocation, production planning, scheduling, and decision-making processes.

For this study, linear programming is leveraged to optimize shuttle capacity given the shuttle head count as input and introducing various constraints such as target utilization, maximum unit capacity and cost.

2. 0 REVIEW OF RELATED WORK

There is a rich variety of methods and models for researching passenger transportation demand forecasting, each providing valuable contributions to improving operational efficiency and service quality. Liyanage et al[1]. have shown the power of AI-based neural network models, especially BiLSTM, in precisely predicting bus passenger demands using smart card data, achieving a remarkable accuracy rate above 90%. Cheng et al[2]. explored the details of time-series forecasting, highlighting the importance of lag periods in improving forecast results, with LSTM lag7 being the best option. Halyal et al[3]. offer significant insights into the use of LSTM for forecasting passenger demand in developing countries, with impressive performance metrics.

Moreover, Tang et al[4]. used a combination of methods, including ARIMA, linear regression, and support vector regression techniques, to forecast short-term passenger flow in subway stations, and add external variables such as weather conditions to boost predictive accuracy. Alqatawna et al[5]. widened the scope to consider logistical aspects, applying SARIMAX, ARIMA, AR, and LSTM models to predict order volume, with SARIMAX showing superior predictive accuracy. Xuefent Li et al[6]. suggested a composite model, WT-FCBF-LSTM, to predict passenger demand under hybrid ridesharing service scenarios, demonstrating notable improvements in prediction accuracy. Xue, Sun, and Chen[7] introduced an Interactive Multiple Model (IMM) filter algorithm for short-term bus passenger demand prediction. They analyzed passenger data from a busy route over four months, creating weekly, daily, and 15-minute time series models. By examining heteroscedasticity and using the IMM filter algorithm to merge models, they achieved better prediction accuracy.

Additionally, Nilabhra Banerjee, Alec Morton, and Kerem Akartunali[8] presented a thorough review of demand forecasting in scheduled passenger transportation, focusing on recent research. Their analysis classified existing research by model properties, objectives, and application areas within the planning cycle. They provided insights into various forecasting methods tailored to industry needs and suggest future research directions. They also addressed common issues such as demand unconstraining and the role of expert judgment. The authors pointed out the lack of standardization in method descriptions and testing, proposing open-source testbeds and a checklist for standardizing research reports. They encouraged comparative studies with existing models to promote cumulative knowledge building in the field.

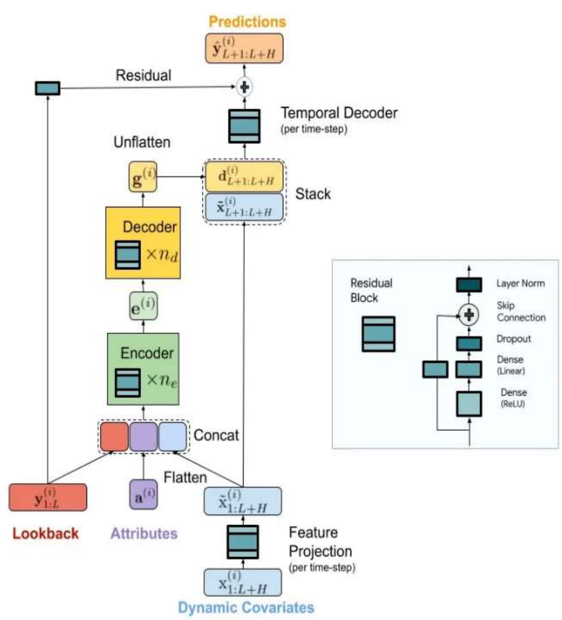

Fig. 3. Overall architecture of TiDE (Time Series Dense Encoder) [9]

Based from research on the diverse range of methods and models available for time series forecasting, the authors decided on the implementation of a relatively newer architecture called TiDE (Time Series Dense Encoder) [9], which is able to achieve state-of-the-art results on numerous datasets when compared to other recent Transformer-based and MLP-based time series models. Fig. 3 shows an overview of the TiDE model wherein dynamic covariates are mapped to a lower dimensional space using a feature projection step. An encoder combines the look-back along with the projected covariates with the static attributes to form an encoding while a decoder maps this encoding to a vector for each time-step on the horizon. Finally, a temporal decoder combines this vector (per time-step) with the projected features of that time-step in the horizon to form the final predictions.

Integrating the time series forecasting results into a linear programming step enables a framework aimed at providing efficient capacity optimization, thereby ensuring the smooth operation of shuttles and the delivery of enhanced service standards.

3.0 METHODOLOGY

3.1 Data Collection

Table 2. Input Data Attributes

ML Task	Data	Variable Type	Description
Time Series Forecasting	Shuttle Headcount	Target Variable	Number of Passengers per shift per route
	Day Type	Exogenous Variable	Whether its Weekday or Weekend
	Production Adjustment	Exogenous Variable	Historical data marked by a company's scheduled production adjustment
	Work shift Schedule	Exogenous Variable	The schedule for operators for the upcoming week
	Product Volume	Exogenous Variable	Historical number of factory builds
	Employee Attendance	Exogenous Variable	Number of employee attendance via attendance tracking system
Linear Programming	Total Capacity	Objective	Total Shuttle Capacity for given route in number of vehicles
	Vehicle Capacity	Constraint	Maximum Passenger Allocation of the Vehicle
	Utilization Target	Constraint	Cost of Trip that is varied per route

Table 2 shows the input variables to the model from various data sources within the company. For time series forecasting, shuttle headcount is the target variable itself to be forecasted while the remaining exogenous variables show variables that intuitively influence the shuttle headcount such as holidays, weekends, employee schedules and product volume. For linear programming, variables are classified into objective or constraint. Objective is the variable itself to be optimized by the linear equation while constraints such as utilization target and vehicle capacity serve to limit the range of possible optimized values.

3.2 Model Training and Validation

Model training and validation involves splitting the collected shuttle head count data into three categories: training data, test data and validation data. Head count data from the last 7 days of the collection period is first set aside as the validation data and will be used to determine the performance of the trained model on unseen data. After the validation data is allocated, the remaining data is divided into training data and validation data. The last 30 days of this remaining data is set as the test data and the rest is set aside as the training data. Both training data and test data is used to train the model across different architectures and configurations.

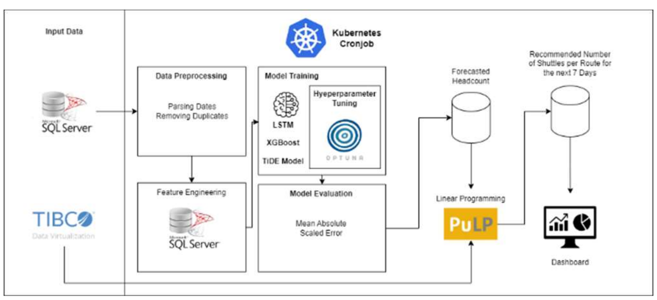

Fig. 4. Model Training and Deployment Flowchart

Fig. 4 describes the process flow used during model training and deployment. Initially, the data undergoes preprocessing, including checks for duplicates and parsing dates into the correct format. Subsequently, the preprocessed data are utilized for feature engineering, generating new independent variables crucial for predicting the forecasted passenger headcount. Three models—LSTM, XGBoost, and TiDE—are tested and compared. Optuna, a Python library, is employed for fine-tuning the models' hyperparameters, with Mean Absolute Scaled Error serving as the evaluation metric.

The forecasted headcount from the best-performing model is then utilized as input for a linear programming solver, alongside other variables. For this optimization process, the authors utilize the Python library PuLP. The resulting allocations are stored in a database connected to a dashboard, enabling process owners to make data-driven decisions regarding the number of shuttles per shift per route required in the upcoming days.

3.2.1 Optuna

Optuna is a Python framework used for the automatic optimization of hyperparameters. Unlike traditional methods such as Grid Search, the framework employs a callback function, which intelligently determines iterations to assess whether the results are improving based on the direction of hyperparameters being tested. It does not follow the exhaustive method employed in grid search.

3.2.2 Mean Absolute Scaled Error

$$MASE = \frac{MAE}{MAE_{naive}}$$

$$MAE = \frac{1}{N}\sum_{j=1}^{N} abs(O_i - Y_i)$$

$$MAE_{naive} = \frac{1}{N-m}\sum_{i=2}^{N} abs(y_i - y_u - 1)$$

Mean Absolute Scaled Error is a metric used to evaluate the accuracy of forecasting models, particularly in time series analysis. It measures the average absolute error of forecasts relative to the scale of the data and compares it to the mean absolute error of a naive forecast (such as a simple moving average). The authors used this metric because it provides a robust measure of forecast accuracy that is scale-independent, making it suitable for comparing the performance of models across different datasets and time series with varying scales.

Low MASE values are desired for model performance with MASE values lower than 1 indicating that a model can forecast better than a naïve forecast.

3.3 Model Deployment via Kubernetes

The entire pipeline is executed via a Kubernetes cronjob which is a Docker image deployed on the company's EDGE cluster. Kubernetes is an open-source platform for automating the deployment, scaling, and management of containerized applications. It simplifies container management across a cluster of machines, enabling efficient resource utilization and ensuring high availability. The authors used this tool to manage pipelines to run at specified times every day.

4.0 RESULTS AND DISCUSSION

Table 3. Summary of Experimental Combinations for Time Series Forecasting Model Training and Validation

Experiment	Input Features	Model	Average Validation MASE
1	Headcount	XGBoost	1.83
2	Headcount Holidays Weekends Production Adjustments	XGBoost	1.71
3	Headcount Holidays Weekends Production Adjustments	LSTM	1.37
4	Headcount Holidays Weekends Production Adjustments Location Shift	LSTM	1.35
5	Headcount Holidays Weekends Production Adjustments Location Shift Product Volume Employee Attendance	TiDE	1.22

Table 3 shows the progression of the time series forecasting model across multiple experiments. In each experiment, the MASE is calculated between the validation data and the generated forecast data for each route and shift combination. The average MASE value is then calculated across all route-shift combinations.

For the first two experiments, the researchers used the XGBoost algorithm as the initial global forecasting model. Experiments 1 and 2 show that adding exogenous variables such as information on holidays, weekends, and plant production adjustments improved the MASE value which shows the positive impact of adding related exogenous variables during model training.

In experiment 3, the model architecture was changed from XGBoost to LSTM which is a more complex time series forecasting model leveraging the power of neural networks. The transition from experiment 2 to experiment 3 showed a significant improvement in MASE which shows that a global time series model using complex neural networks better capture and learn the temporal patterns present in the shuttle dataset.

In experiment 4, new exogenous variables such as employee workshift schedule and route location were added to the model training but only represented a small improvement in MASE.

For experiment 5, the authors used a relatively newer neural network time series model architecture called TiDE (Time Series Dense Encoder) along with adding more exogenous variables such as product volume and employee site attendance. With this final iteration, the trained TiDE model was able to achieve an average MASE value of 1.22.

Table 4. Utilization Comparison of Forecasted vs. Actual Capacity

Forecasted Capacity vs. Head Count	Forecasted Capacity vs. Actual Capacity	Percentage Occurrence	Average Utilization Change
Forecasted Capacity >= Head Count	Equal Capacity (Forecasted = Actual)	77.46%	0 %
	Optimized Capacity (Forecasted < Actual)	3.6 %	+ 21.08 %
	Overcapacity (Forecasted > Actual)	8.14 %	- 35.12 %
Forecasted Capacity < Head Count	Undercapacity (Forecasted < Actual)	10.8 %	N/A

The forecasts given by the final trained model was then fed to the linear programming module of the pipeline to get the optimized shuttle capacity for each route-shift combination.

Table 4 shows the utilization comparison between forecasted shuttle capacity and the actual shuttle capacity over the validation period. In 77.46% of the capacity forecasts, the pipeline was able to forecast the same capacity as the actual shuttle capacity giving no relative utilization improvement while 3.6% show that using the forecasted shuttle capacity resulted in an average utilization gain of 21.08%.

However, 8.14% of the capacity forecasts show overcapacity wherein the forecasted capacity is much higher than the actual capacity resulting in lower utilization. If used, utilization would incur an average loss of 35.12%. Undercapacity is also observed in 10.8% of the capacity forecasts wherein the forecast capacity is even lower than the shuttle head count which is not desirable even with the perceived utilization increase.

5.0 CONCLUSION

In this study, the authors demonstrated that a machine learning pipeline consisting of global time series forecasting and linear programming can forecast shuttle capacity with similar or better efficiency in 81.06% of capacity recommendations. Overcapacity and undercapacity are observed in the remaining 18.94% of recommendations which represent room for improvement in the pipeline.

Even with the observed instances of undercapacity and overcapacity, this machine learning pipeline can reduce the man-hours needed to conduct manual shuttle capacity analysis. Manual analysis would only then be mostly relegated to addressing routes where overcapacity or undercapacity is routinely observed.

6.0 RECOMMENDATIONS

Reducing the occurrence of undercapacity and overcapacity is the foremost recommendation for the continuation of this study which is directly tied to further improving the global time series forecasting model. This improvement can involve using other state-of-the-art model architectures or including more exogenous variables that influence shuttle head count.

The linear programming module can also be expanded to add more constraints such as shuttle trip cost, travel time and even CO_2 emissions as a function of distance traveled. Adding these constraints would involve the collection of additional shuttle data.

On a broader note, this machine learning pipeline of forecasting and optimization can be replicated in similar areas of operations research such as material consumption to predict purchase order quantity or production volume to predict tool and manpower allocation, as demonstrated by Alqatawna et al., albeit with a different methodology.

Utilizing a combination of linear programming techniques applied to forecasted variables can offer a promising approach and serve as a template for various use cases.

7.0 ACKNOWLEDGMENT

The authors would like to thank Ma'am Janoah Delos Santos, Sir Lito Villanueva, Real Estate Office team members and Data View developers for their valuable contributions to this project.

8.0 REFERENCES

1. Sohani Liyanage et al., AI-based neural network models for bus passenger demand forecasting using smart card data, Journal of Urban of Urban Management, **2022**, pp 365-380.

2. C. H. Cheng et al., An Intelligent Time-Series Model for Forecasting Bus, Appl. Sci., 12, **2022**, 4763.

3. Shivaraj Halyal et al., Forecasting public transit passenger demand: With neural networks using APC data, Case Studies on Transport Policy, **2022**, pp 965-975.

4. Liyang Tang et al., Forecasting Short-Term Passenger Flow: An Empirical Study on Shenzhen Metro, IEEE Transactions on intelligent transportation systems, **2019**, pp 3619-3622.

5. Ahmad Alqatawna et al., Incorporating Time-Series Forecasting Techniques to Predict Logistics Companies' Staffing Needs and Order Volume. Computation **2023,** 11, 141.

6. Xuefeng Li et al., The forecasting of passenger demand under hybrid ridesharing service modes - A combined model based on WT-FCBF-LSTM, Sustainable Cities and Society, **2020**, 102419

7. Rui Xue, Daniel Sun, and Shukai Chen, Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach, Discrete Dynamics in Nature and Society, **2015**, Article ID 682390, 11 pages.

8. Nilabhra Banerjee, Alec Morton, and Kerem Akartunali, Passenger demand forecasting in scheduled transportation, European Journal of Operational Research, 286, **2020**, pp 797-810.

9. Abhimanyu Das et al., Long-term Forecasting with TiDE: Time-series Dense Encoder, Transactions on Machine Learning Research, **2023**, 21 pages.

9.0 ABOUT THE AUTHORS

Wilhelm Henri R. Alegrado is a Sr. Data Scientist at Western Digital Storage Technologies for Operation Excellence. He holds a master's degree in data science at Asian Institute of Management. He has been with the company for 12 years with experience from previous departments with an Engineering and Development background.

Jhon Vincent A. Gupo is an Associate Data Scientist at Western Digital Storage Technologies for Operation Excellence. He holds a bachelor's degree in computer science at Laguna State Polytechnic University – Los Baños.

10.0 APPENDIX

10.1 Dashboard Example