Predicting Marino's bustling hours; A time series analysis

Harishraj Udaya Bhaskar Siddhant Ashay Shah

Abstract

The goal of this project is to create a prediction model that can forecast how many individuals will go to the gym on different days and at different times of the day. This study builds and tests machine learning models using historical data on gym attendance, demographic data, and environmental factors. To find patterns and predict future attendance rates, the models are trained using time-series analysis. The findings of this study will assist Marino managers in making wise choices regarding staffing, facility utilization, and equipment in order to maximize gym operations and enhance the general client experience.

1. Introduction

For gym owners and managers, estimating the number of visitors at various times of the day and week is crucial. The customer experience may be enhanced overall and gym operations can be optimized with the aid of accurate projections. Various machine learning models have been created to predict future gym attendance in order to achieve this.

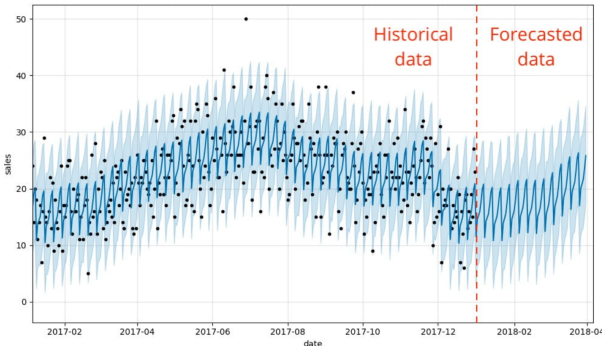

Fig 1: Forecasting time series

The Autoregressive Integrated Moving Average (ARIMA) model is one of the most well-liked time-series forecasting models. For predicting non-stationary time-series data, ARIMA models are frequently utilized because they can handle a variety of trends, seasonality, and cyclical patterns. Previous research has shown that ARIMA models can predict gym attendance rather well.

The Seasonal ARIMA (SARIMA) model is an additional time-series forecasting model that has been gaining prominence.SARIMA is a seasonal component-added version of the ARIMA model that handles data with seasonal trends. In situations where the data exhibit seasonal trends, SARIMA models have been proven to perform better than ARIMA models.

Long Short-Term Memory (LSTM) and other deep learning models have recently been used for time-series forecasting problems. Recurrent neural networks (RNNs) with the ability to handle long-term dependencies, such as LSTM models, have been demonstrated to be effective in a number of time-series prediction applications.

In this investigation, we will evaluate how well the ARIMA, SARIMA, and LSTM models predict gym attendance. To train and test the models, we will use historical information on gym attendance, demographic data, and environmental conditions. The goal of this project is to create a UI where students and faculties using the gym can get accurate predictions for the upcoming days.

2. Technical Approach

2.1 Method 1: Time series analysis using ARIMA

ARIMA (Autoregressive Integrated Moving Average) is a time-series analysis technique that models the future values of a dependent variable based on its past values and the errors or residuals from previous predictions. ARIMA models are commonly used to forecast trends in data that exhibit a certain degree of predictability. In the context of student entry logs of a gym, ARIMA is used to predict the number of students who will visit the gym in the future, based on historical data of gym entry logs. This would be useful for the managers at Marino Recreational Center to plan for staffing and equipment needs, and to optimize the gym experience for students.

2.2 Method 2: Time series analysis using SARIMA

SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) is an extension of the ARIMA model that includes the influence of exogenous variables in the time-series analysis. Exogenous variables are external factors that can affect the dependent variable, but are not necessarily part of the time-series being analyzed. In the context of student entry logs of a gym, an example of an exogenous variable could be the weather or the academic calendar. SARIMAX could be used to forecast the number of students who will visit the gym while taking into account these external factors. This would enable gym managers to plan for contingencies and make more accurate predictions, ultimately leading to better decision-making. Because of a lack of data, we were unable to completely model the seasonality pattern.

2.3 Method 3: Time series analysis using LSTM

Modeling complex sequences and time-series data has shown to be a notable strength of the Long Short-Term Memory (LSTM) class of recurrent neural networks. With the use of memory cells, input gates, output gates, and forget gates, LSTM models may effectively capture long-term dependencies in the data.

We initially preprocess the data by putting it into a time-series format in order to employ LSTM for gym attendance prediction. The data is then divided into training and testing sets, and it is normalized to make sure that all of the features have a comparable scale.

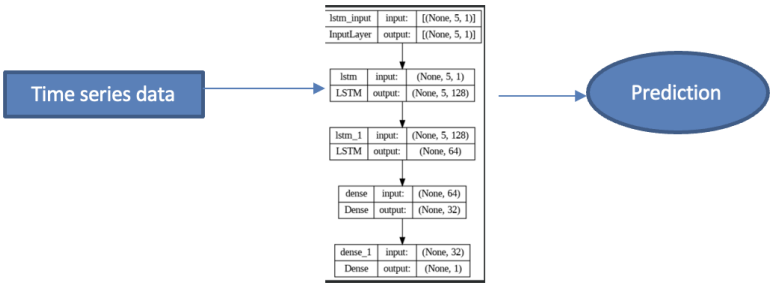

Fig 2: Flow of LSTM to use prediction

Next, we create a deep LSTM model in Python using the Keras framework. The model has several layers, including dense layers for output prediction and LSTM layers with dropout regularization to avoid overfitting. The dense layers deliver the ultimate output prediction, while the LSTM layers record the temporal dependencies of the input data.

We employ the Adam optimizer with a mean squared error loss function to train the LSTM model. In order to avoid overfitting and preserve the model that performs the best based on validation loss, we additionally use early stopping.We assess the LSTM model's performance on the testing set once it has been trained. Metrics like mean absolute error (MAE) and root mean square error (RMSE) are used to gauge the accuracy

3. Experimental Results

3.1 Dataset

We gathered actual data from the Marino recreation center in order to train and evaluate the prediction models for gym attendance. The dataset includes data on the number of patrons using the gym at various times during the day and on various gym floors. A combination of manual headcounts and an automated tracking system put in place at the gym's entrance were used to gather the data. The tracking system counts the number of persons entering and leaving the gym at any given time by using infrared sensors to track their entry and leave.

The dataset includes attendance information for the calendar year 2023, which runs from January 1 to March 23. The data comprises information on the number of persons present on each floor of the gym and is collected hourly between 5:00 AM and 11:00 PM. The preprocessing of this real world data is done separately for each model which is described as follows

3.1.1 Preprocessing of data for ARIMA, SARIMAX

The Dataset was our biggest challenge. To tackle this, we interpolated data for every 5 minute interval from the given hourly dataset for twenty days.

After interpolating the data it looks like:

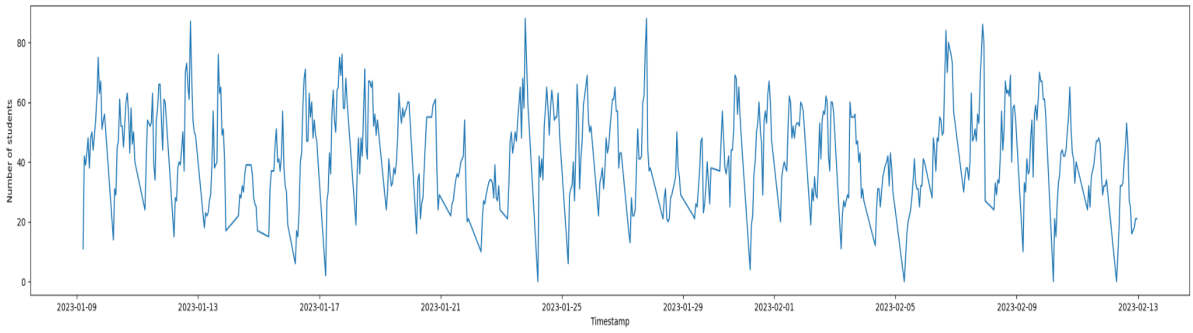

Fig 3: A plot of the time series data after interpolating it

The average number of students distributed over the day from our data looks like:

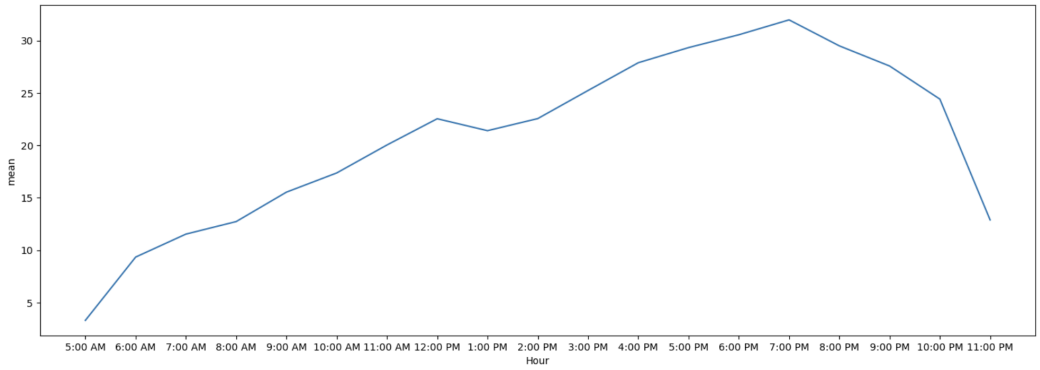

Fig 4: Average distribution based on hours of the day

3.1.2 Preprocessing of data for LSTM

Before we could feed the data to the LSTM model we had to do a far bit of preprocessing. The datatypes of the columns were not suitable so we had to convert it to the appropriate data types. Next we grouped the data by date to get the total number of huskies per hour in all floors to make it easier to feed the data to the network.

We had to later convert this data into a time series format where the first data 5 days would be taken as an input to predict the 6th day and then it would be slid by one where the 6th day becomes the input to predict the 7th day. The model learns the complex time series pattern this way Next, we normalized the data to ensure that all features have a similar scale. We used the MinMaxScaler from the scikit-learn library to scale the data to a range between 0 and 1

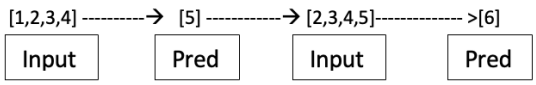

Fig 5: Flow of input through predictions, for time series data

3.2 Training and prediction accuracy for ARIMA

The Augmented Dickey-Fuller (ADF) test is a statistical test commonly used in econometrics and time-series analysis to determine whether a given time series is stationary or non-stationary. Stationarity is a key assumption in many time-series models and refers to the idea that the statistical properties of the series, such as mean and variance, do not change over time. The ADF test works by estimating the degree of dependence between the current observation and its lagged values, and then comparing this to a null hypothesis of non-stationarity. If the test statistic is less than the critical value at a given level of significance, the null hypothesis of non-stationarity is rejected, indicating that the time series is stationary. The following plot shows that our dataset is stationery especially for values of lag between 0 and 30.

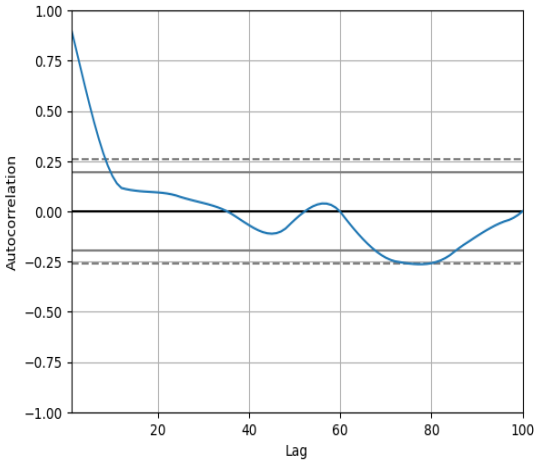

Fig 6 : lag with respect to Auto correlation

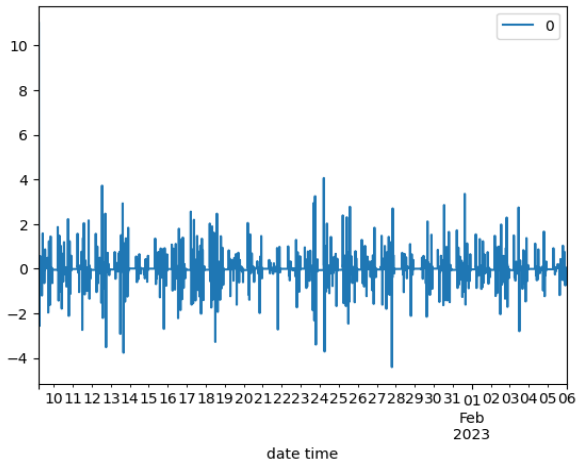

Fig 7 Covariance of the data

Finally, trying a few variables for the parameters and understanding the model's behavior based on the data using tabulated results such as:

SARIMAX Results

Dep. Variable:	count	No. Observations:	8007
Model:	ARIMA(5, 1, 0)	Log Likelihood	-1668.279
Date:	Mon, 24 Apr 2023	AIC	3348.557
Time:	11:58:52	BIC	3390.485
Sample:	01-09-2023	HQIC	3362.908
	- 02-06-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9480	0.039	24.268	0.000	0.871	1.025
ar.L2	-2.221e-06	0.057	-3.9e-05	1.000	-0.111	0.111
ar.L3	-3.959e-06	0.057	-6.96e-05	1.000	-0.111	0.111
ar.L4	1.389e-05	0.057	0.000	1.000	-0.111	0.111
ar.L5	-0.0528	0.040	-1.331	0.183	-0.130	0.025
sigma2	0.0888	0.000	214.046	0.000	0.088	0.090

Ljung-Box (L1) (Q):	0.13	Jarque-Bera (JB):	1090523.01
Prob(Q):	0.72	Prob(JB):	0.00
Heteroskedasticity (H):	0.67	Skew:	-0.34
Prob(H) (two-sided):	0.00	Kurtosis:	60.17

Fig 8: Different parameters of evaluation for ARIMA(5,1,0)

We were able to train a model but our results were subpar and either underfitting or overfitting the data:

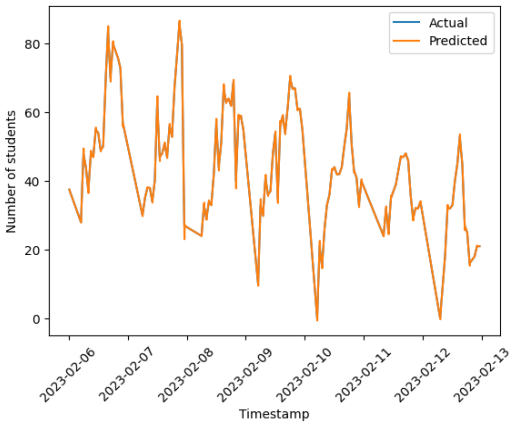

Fig 9: Overfit model

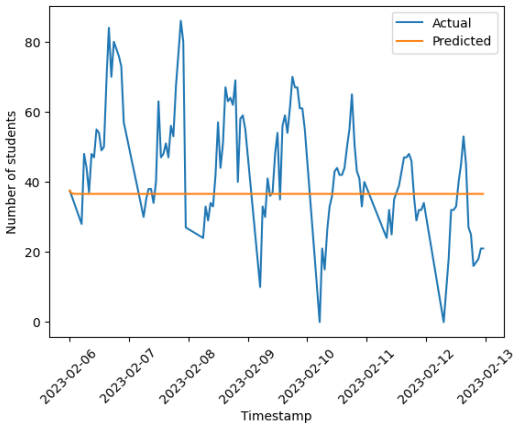

Fig 10:Underfit model

3.3 Training and prediction accuracy for SARIMAX

Upon studying the data, we realized we can take advantage of the two peaks in the average number of students per day at any given hour in the day shown in fig 2.

Using the parameters (5,1,0) for p,d,f of the order and a seasonal order of (1,0,0,7) and a cumulative model, we found some reasonable results as indicated in the graph below:

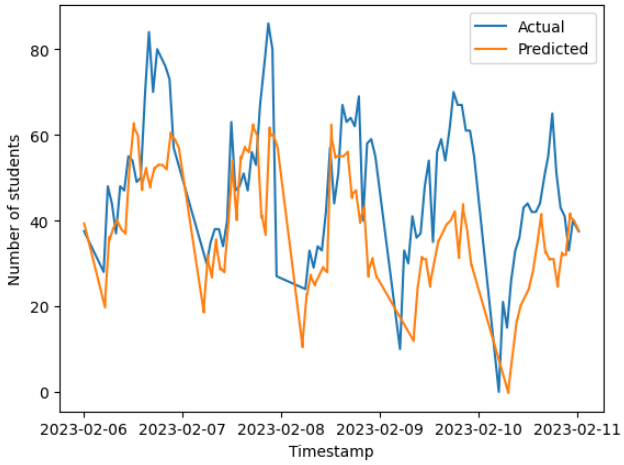

FIg 11: SARIMAX result

The RMSE dropped from 45.916 to 17.25

3.4 Training and prediction accuracy for LSTM

The LSTM model was trained using the preprocessed gym attendance data, with the training set consisting of 80% of the data and the remaining 20% used for testing. We used the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as performance metrics to evaluate the accuracy of the LSTM model

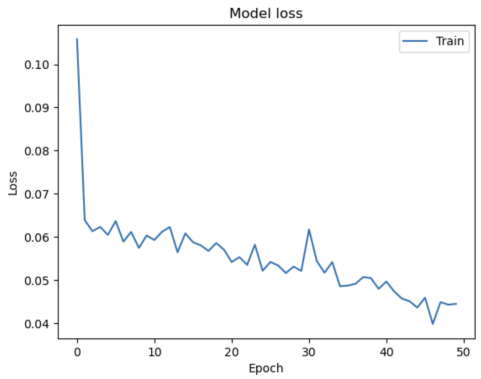

Fig 12: Training accuracy

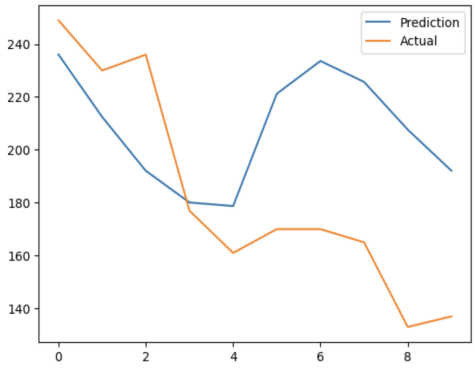

Fig 13:Testing accuracy

3.5 Creation of User Interface (UI) for Predictive Model

We created a graphical user interface (GUI) for the gym attendance prediction system using the Python Tkinter module to make it easier for users to use and more accessible. Users can communicate with the system through the GUI by choosing the time slot for which they want to view the attendance projection.The GUI shows a line graph with the l attendance data for the same time period.

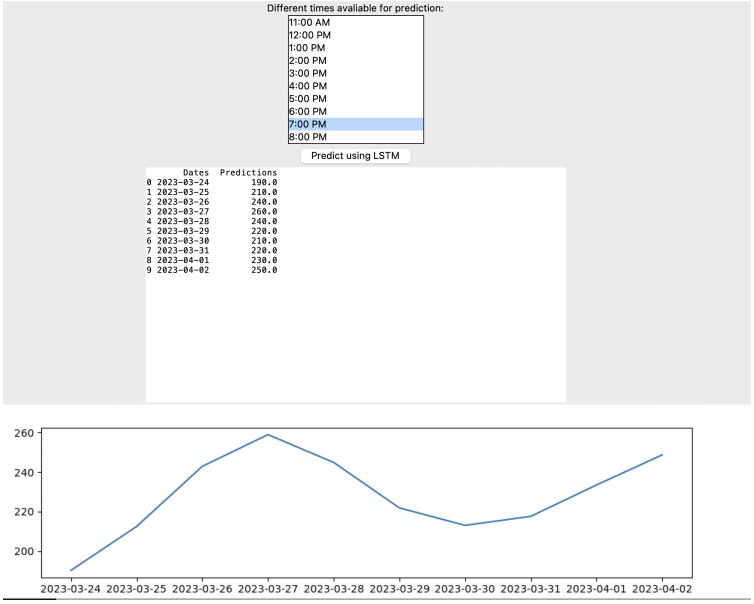

Fig 14: GUI for students

4. Conclusion and future work

In this project, we created a system for predicting gym attendance using machine learning methods including ARIMA, SARIMA, and LSTM models. We gathered actual attendance data from a gym and preprocessed it so that the models could be trained and tested. For the gym attendance prediction system, we also created a user interface that enables users to quickly choose the time period for which they wish to examine attendance forecasts and to display the projected and actual attendance data in a graphical style.

We found this approach and analysis very interesting however, it should be noted that due to limited access to real-world data, the accuracy of the developed model may be limited.We requested for more data but we were not given the access to the required and ML models require large amounts of data to be highly accurate.

Another area of future work could also include integrating the created GUI with the Marino website so that gym patrons can view real-time attendance data and plan their visits accordingly.

5. Participants Contributions

Harishraj Udaya Bhaskar

- Contributed to the collection and preprocessing of data,
- Trained and developed the LSTM network
- Developed a GUI for better accessibility to the real world

Siddhant Ashay Shah-

- Contributed to the collection and preprocessing of data,
- Trained and developed the ARIMA model
- Trained and developed the SARIMA model

6. References

[1] Hochreiter, Sepp; Schmidhuber, Jürgen (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276.

[2] Title: Autoregressive integrated moving average. (2022, April 14). In Wikipedia, The Free Encyclopedia. Retrieved 20:15, April 27, 2023, from https://en.wikipedia.org/w/index.php?title=Autoregressive_integrated_moving_average&oldid=1110582437

[3] "SARIMA." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 17 September 2021. Web. 27 April 2023. https://en.wikipedia.org/wiki/Seasonal_autoregressive_integrated_moving_average.

[4]Python Software Foundation. (n.d.). tkinter – Python interface to Tcl/Tk. Retrieved April 27, 2023, from https://docs.python.org/3/library/tkinter.html