Unraveling Patterns in the Philippine Lottery System Using Machine Learning Algorithms

Gabriel Venz V. Badilles, Jewynah Mae H. Jaronay, and Renmar M. Lescano

Division of Physical Sciences and Mathematics
University of the Philippines Visayas
Miagao, Philippines
{gvbadilles,jhjaronay,rmlescano2}@up.edu.ph

ABSTRACT

This study adopted machine learning (ML) techniques to further mine out patterns in the Philippine Small Town Lottery, disproving traditional views regarding randomness in lottery. Although lotteries account for huge contributions to revenue performance of the Philippine government, attention does draw towards the complexities of STL and potential of ML in detecting hidden patterns. With this, the data used for this study was obtained from the website STL Drawn for 2017 - 2023 using Long Short-Term Memory (LSTM), and LSTM with Convolutional Neural Networks (CNN) models as a basis for analysis. The visualizations depict uniform number distribution yet showing challenges in predicting draw outcomes, revealing overfitting and randomness in STL draw results. The study reveals that there is a lack of discernible patterns in the STL draws, as LSTM and LSTM with CNN models struggle to predict lottery outcomes, which implies that STL draws follow the principles of randomness. Limitations include the constraints in the coverage of datasets and the exclusive deployment of particular ML models, thus increasingly calling for the necessity to explore wider models to broaden understanding.

KEYWORDS

Lottery Systems, Small Town Lottery (STL), Machine Learning Algorithms, Randomness, Predictive Modeling

1 INTRODUCTION

Lottery systems have traditionally been perceived as representatives of randomness, governed solely by luck and probability. Research on lottery games, including the Philippine Small Town Lottery (STL), has revealed the presence of patterns and strategic choices in players' number selection [3]. It has become a very interesting object of research that introduces a certain angle for understanding the complexity of the system. Technological advancements in the Philippines, through proper fusion of data science with machine learning (ML) techniques, proves an avenue to investigate the fundamental structure of the STL and understand them.

Lotteries have become a great source of revenue for governments in the Philippines, with the Philippine Charity Sweepstakes Office (PCSO) managing to earn more than 43 billion pesos from their games in 2021 [12]. In the Annual Report, it was stated that the positive impact of the lottery was not limited to the financial contributions extended. The lives of 55 lucky individuals were changed as they received the winning jackpot. These examples of changed lives though lottery winnings also underline the possible life changing effects and a profound reach of initiatives taken by PCSO.

In the rising field of research addressing the unpredictability of lottery outcomes, the researchers go a step forward by employing machine learning approaches to investigate historical STL draw results. Using these statistics, the researchers aim to move away from standard statistical testing and into the domain of ML pattern recognition.

This paper will evaluate various algorithms of machine learning in view of the Philippine Small Town Lottery (STL), adapting the use of datasets that will uncover trends and probable non-randomness in the lottery's results. The researchers strive to achieve two objectives with this multidimensional exploration: (a) bringing new perspectives to the discussion of lottery randomness on behalf of the scientific community, and (b) demonstrating how machine learning can be used successfully to reveal intricate underlying patterns in seemingly unpredictable events. The researchers believe that by conducting this research, it will not only improve their understanding of probability in lotto systems, but also demonstrate how a data science and statistics approach can provide us insights into the dynamics of gaming in the Philippine context and develop further research along these lines.

2 LITERATURE REVIEW

In recent years, the focus of research on lottery systems has been on their economic significance and the kind of societal implications that the lottery systems hold. According to Livingstone et al. [9], studies have explored the mechanisms of lottery systems like the Small Town Lottery (STL), studying its patterns as well as their impacts. They also examined the natural regressiveness that comes with lottery revenues, especially involving those that are sanctioned by the state [10].

Several studies have found that analyzing non-random structures in datasets using machine learning algorithms such as the Random Forest (RF), Support Vector Machines (SVM), and Neural Networks can produce significant results. A study by Dewi and Chen [6] presents the result by combining a model of RF, SVM and tuned SVM regression, producing an improved performance for regression analysis. The paper further depicted that this integration consistently reduced the Root Mean Square Error (RMSE) and enhanced correlation coefficient (r) values as demonstrated across various datasets. Highlighting the feature selection by Random Forest, this study demonstrated how it is necessary to select key features, hence, better model performance. Besides, it was found

that SVM parameter tuning was important for better results to be attainable, including remarkable decreases in RMSE. Thus, the results indicate that not only the proposed algorithm is proven to be feasible and accurate but it also has potential for expanding applications which implies future researches in exploring different models, kernels, methods, datasets, and influencing factors in prediction models. However, the difficulty in using machine learning algorithms such as RF, SVM, and tuned SVM regression to predict lottery numbers is due to unstandardized data as a result of randomness during lotto draws. While these algorithms are excellent tools to recognize patterns, lottery systems have been designed in such a way that the outcomes of a lottery appear to be unpredictable [14].

Other researches have used time series forecasting models in predicting outcomes. For instance, Albeladi et al. [1] utilized Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA) models to forecast temporal dependencies or long-term dependencies using historical data. These models also have a great potential in predicting financial and stock market data as demonstrated in a study by [2], which specifically emphasized ARIMA as having significant results in the short term prediction.

The changing landscape in the application of machine learning to lottery analysis involves evolving methodological choices. Researchers continuously have to change and revise the approaches to reflect the intricate patterns in these systems [8]. This clearly shows the passion that researchers possess for advancing knowledge in understanding puzzles surrounding data from lotteries as evidenced by experimenting with various machine learning algorithms [5]. This basically confirms that, indeed, there is a search for precision and insight in understanding the diverse aspects of lottery systems as the field progresses and with its lenses through machine learning.

The integration of machine learning into the lottery system has given insights to the researchers on how those systems work, especially in determining patterns, anomalies, and their predictive capabilities. Indeed, such predictive models have shown mixed results in terms of forecasting future outcomes as they tend to challenge the notion of pure randomness in such lottery draws [4]. These findings can be believed to add complexity in an emerging discourse which questions the ability of machine learning algorithms in navigating patterns present in the datasets of lotteries.

As much as machine learning offers a plausible way to the analysis of lotteries, there exists several limitations and challenges in its works. Problems like data quality, overfitting, and changing aspects of the lottery systems may limit application of the machine learning algorithms [13]. Indeed, the field is not without challenges and continuing research endeavors are necessary in further refining of methodologies, overcoming of existing limitations, and understanding of a complex relationship between machine learning and lottery systems. As the lottery systems develop, so will be the methodologies of analyzing it, all with the explorative journey of unraveling its patterns that govern probability.

3 METHODOLOGY

3.1 Data Collection

The data gathered for the project was obtained from the STL Drawn website of Negros Occidental, spanning from January of 2017 until November of 2023 [7]. The dataset comprises the date, which is indicated by day, month, and year, and the three draws for each day. Each draw consists of a distinct pair of numbers ranging from 1 to 38. This means that the two numbers in a pair are different. However, the same number can appear in different pairs within the other draws of the same day [11]. This dataset structure provides a comprehensive overview of the STL outcomes over six years, providing a solid basis for further analysis.

3.2 Model Training

The LSTM model is known for how well it can analyze time series data because it can learn and remember how things depend on each other over time. Because of this, it is very good at predicting lottery results and randomness in it. Since lottery sequences are dependent events, the can find some pattern that will explain the sequence. The model is trained by looking at past lottery results and figuring out how draws happen at different times and how they are related.

When we combine LSTM and CNN, we can use CNN's ability to recognize patterns in space along with LSTM's ability to handle data over time. Using this mode, we are not relying on the time based draws but trying to infer spatial features. This mixed model works great for finding complicated trends in lottery draws that may be important in both space and time. First, CNN layers go through the raw data to find features that are important in the string of lottery numbers. Then, these features are sent to LSTM layers, which look at the time aspects of the data to guess what draws will happen in the future.

4 RESULTS

The team utilized various libraries such as Matplotlib and Keras in order to visualize and display the results of the study. Matplotlib was used in order to visualize and determine if the dataset are randomly distributed or not during exploratory analysis and result analysis after data training and testing. Keras library was used in testing and training the dataset using LSTM and LSTM with CNN.

4.1 Exploratory Data Analysis

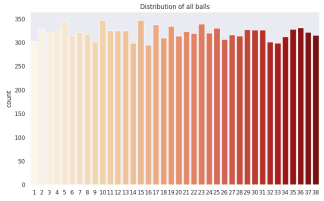

Figure 1: Distribution of all balls

In analyzing the data, the team first visualized the overall distribution of all balls in the dataset. Figure 1: Distribution of all balls displays the graph having a uniform distribution with the lottery numbers being picked between 250 and 350 times.

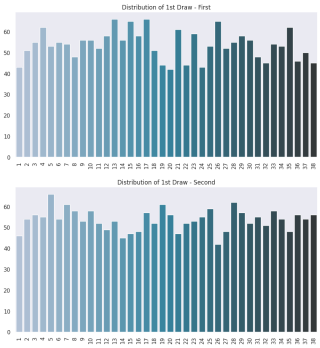

Figure 2: Distribution of 1st draw

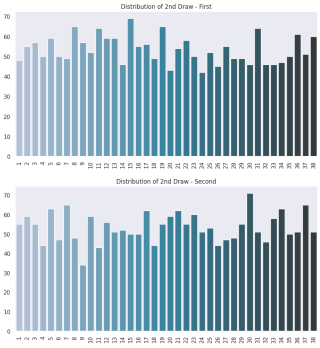

Figure 3: Distribution of 2nd draw

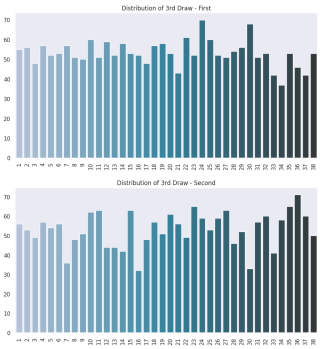

Figure 4: Distribution of 3rd draw

After getting the overall distribution of all balls, we then examined the distribution of balls for every draw. Figures 2, 3, and 4 display the distribution of each draw, and it can be observed that the graph shows variability in the occurrences of each number drawn.

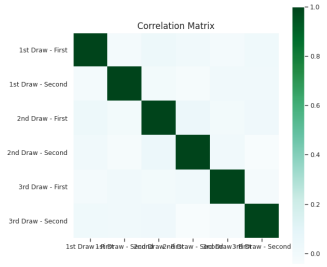

Figure 5: Correlation Matrix

To gain more insights into the dataset, a correlation matrix was used in order to identify patterns or trends in the data. Figure 5 displays that the correlation matrix had a perfect positive correlation in the same draw and that there is no correlation between different draws.

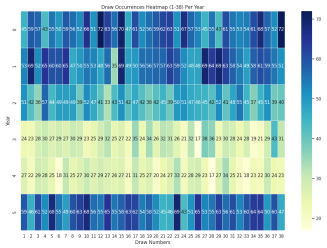

Figure 6: Distribution heatmap

A heatmap was generated along with the correlation matrix to identify patterns as well as analyze the large dataset in a digestible and simpler format. Figure 7 shows the heatmap of all draws for all years, and it can be observed that the results are varied. Year 0 and 1 had a consistently high score with a noticeable decrease in year 2 and a further decrease in scores in years 3 and 4, and lastly, a return of an increasing score in year 5.

4.2 STL Lottery LSTM Architecture

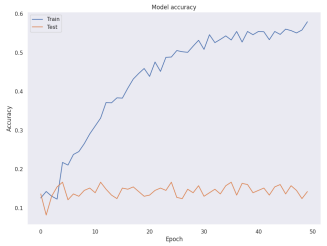

Figure 7: LSTM Model Accuracy

In evaluating the accuracy of the model, the team used plotted graphs in order to have a visual representation of the results. Figure 7: LSTM Model Accuracy displays the graphical presentation of the accuracy of the model, and it can be seen that the accuracy of

the training set (blue line) is higher than that of the validation set (orange line).

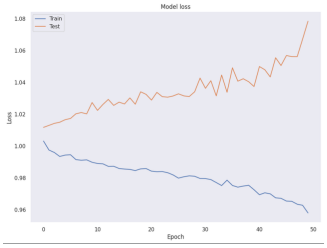

Figure 8: LSTM Model Loss

On the other hand, Figure 8: LSTM Model Loss shows the model loss on the testing and validation sets, and there is an inverse result wherein the training set had a higher loss than the validation set.

4.3 STL Lottery LSTM with CNN Architecture

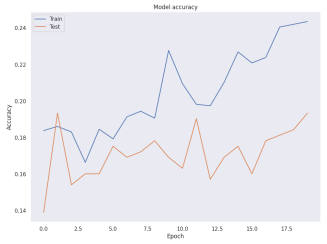

Figure 9: LSTM with CNN Model Accuracy

Figure 9: LSTM with CNN Model Accuracy shows how well the LSTM+CNN model worked on both the training set (Train) and the testing set (Test). It is noticeable that the accuracy achieved during training consistently surpasses that of testing. Training precision starts just above 0.18 and goes up steadily until it reaches about 0.24. The testing accuracy, on the other hand, starts out a little lower and, despite some changes, stays on an overall upward path until it ends around 0.19.

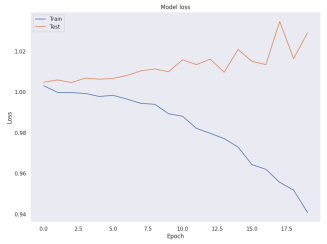

Figure 10: LSTM with CNN Model Accuracy

The training and testing losses have an inverse relationship, as shown in Figure 10. The training loss goes down over time, starting just above 1.00 and ending just above 0.94. On the other hand, the testing loss seems to start at the same level without a downward trend.

5 DISCUSSIONS

In analyzing the exploratory analysis, the overall distribution of all balls in figure 1 is uniformly distributed which inferred that the each ball has an equal probability of being drawn. This result is important in the study as it is a fundamental principle of fairness in random selection processing.

It was observed that the distribution of draws does not follow a uniform distribution, even though it should be based on Figures 2, 3, and 4. The graph shows the distribution of ordered data, which is irrelevant in this study. Due to this reason, the team focused more on the distribution of all balls, as the order of the balls does not affect the result.

The varying results in the distribution heatmap in figure 6 indicated that there may be various factors at play that resulted in the rise or fall on the number of results.

Looking at both the accuracy on figure 7, the training set had a higher accuracy than the testing set, while on figure 8, the testing set had a higher loss than the training set. By analyzing the graph and the result, it was apparent that the model was overfitted and was not able to train the data properly, indicating that there was no apparent pattern that could be found in the drawing of lottery balls.

Furthermore, the results of the STL lottery draws using LSTM+CNN models, shown in Figures 9 and 10, show that lottery results are naturally hard to predict. The fact that the model could find trends in the training set but not in new data shows how random lottery draws are.

This study was conducted in order to determine if there is a prominent pattern in the drawing of STL and after running the dataset on the model chosen, it revealed an overfitting of the data, indicating that the result is non-conclusive and no pattern can be observed.

6 CONCLUSIONS

The purpose of this study was to conduct a thorough investigation of the Philippine Small Town Lottery (STL) system using machine learning models. Our main goal was to look deeper into the STL's lottery results to see if there were any underlying trends that would go against the common belief that these things happen by chance. Long Short-Term Memory (LSTM) and LSTM mixed with Convolutional Neural Networks (CNN) were two of the machine learning models used in the study. The use of these machine learning models revealed that STL drawings do not exhibit obvious, predictable patterns that can be consistently identified using the machine learning algorithms used. The fact that our models showed overfitting made it even more clear that the STL results mostly follow the rules of chance and randomness that are common in lottery systems.

6.1 Limitations

One of the study's limitations is the dataset, which includes STL draw results from January 2017 to November 2023. Although the dataset offers an adequate amount of information, it is not comprehensive enough to encompass all enduring patterns and trends, especially in a system that is inherently randomized, like a lottery. Making the information cover a longer period of time gives a more complete picture of how the lottery system works. Another is the

limited set of machine learning models that were used to analyze the Small Town Lottery (STL) data. The models using only LSTM and LSTM with CNN are both effective on their own. However, the omission of alternative models that have the potential to be effective has constrained the scope of the analysis. Also, the study mostly looked at models that are good at handling time series data, so the models that are good at classifying or clustering jobs have been overlooked. Adding a wider range of machine learning models will help understand the dataset better and find patterns that the present models missed.

REFERENCES

[1] Zafar-B. Mueen A. Albeladi, K. 2023. Time Series Forecasting using LSTM and ARIMA. *International Journal of Advanced Computer Science and Applications* 14 (2023). Issue 1. https://doi.org/10.14569/IJACSA.2023.0140133

[2] Alwadi S. Almasarweh, M. 2018. ARIMA Model in Predicting Banking Stock Market Data. *Modern Applied Science* 12, 11 (2018), 309. https://doi.org/10.5539/mas.v12n11p309

[3] Zoltayné Paprika Z. Becser, N. 2016. Patterns in the lottery game. *Forum Scientiae Oeconomia* 4, 1 (2016), 55–70. https://doi.org/10.14569/IJACSA.2021.0120480

[4] Cook P. J. Clotfelter, C. T. 1991. Lotteries in the real world. *Journal of Risk and Uncertainty* 4 (1991), 227–232. https://doi.org/10.1007/BF00114154

[5] Rachel Courtland. 2018. *Bias detectives: the researchers striving to make algorithms fair.* Retrieved January 13, 2024 from https://www.nature.com/articles/d41586-018-05469-3

[6] Chen R. C. Dewi, C. 2019. Random forest and support vector machine on features selection for regression analysis. *International Journal of Innovative Computing, Information and Control* 15, 6 (2019), 2027–2037.

[7] STL Drawn. [n.d.]. *STL Negros Results.* Retrieved December 1, 2023 from https://stldrawn.blogspot.com/search/label/STL%20Negros%20Results

[8] Gray H. M. Bosworth L. Shaffer H. J. LaPlante, D. A. 2010. Thirty Years of Lottery Public Health Research: Methodological Strategies and Trends. *Journal of Gambling Studies* 26 (2010), 301–329. https://doi.org/10.1007/s10899-010-9185-1

[9] Adams P. Cassidy R. Markham F. Reith G. Rintoul A. Schüll N. Woolley R. Young M. Livingstone, C. 2018. On gambling research, social science and the consequences of commercial gambling. *International Gambling Studies* 18 (2018), 56–68. Issue 1. https://doi.org/10.1080/14459795.2017.1377748

[10] Hansen A. Sprott D. E. Miyazaki, A. D. 1998. A Longitudinal Analysis of Income-Based Tax Regressivity of State-Sponsored Lotteries. *Journal of Public Policy Marketing* 17 (1998), 161–172. Issue 2. https://doi.org/10.1177/074391569801700202

[11] Lotto PCSO. [n.d.]. *STL Pares.* Retrieved December 27, 2023 from https://www.lottopcso.com/stl-pares-result-january-9-2020/

[12] Philippine Charity Sweepstakes Office (PCSO). 2021. PCSO 2021 Annual Report. (2021), 17–21. https://www.pcso.gov.ph/pcsofiles/CGS/Annual%20Report%202021%20pp17-21.pdf

[13] Azar A. T. Elgendy M. S. Fouad K. M. Salam, M. A. 2021. The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. *International Journal of Advanced Computer Science and Applications* 12, 4 (2021), 641–655. https://doi.org/10.14569/IJACSA.2021.0120480

[14] Galbo-Jørgensen C. B. Tyran J. R. Suetens, S. 2016. Predicting Lotto Numbers: A Natural Experiment on the Gambler's Fallacy and the Hot-Hand Fallacy. *Journal of the European Economic Association* 14 (2016), 584–607. Issue 3. https://doi.org/10.1111/jeea.12147