National College of Ireland

Using Time Series Predictive Models for Early Detection of Gambling Addiction in Problem Gamblers

MSc Research Project
Data Analytics

Ram Abhilash Vasamsetti
Student ID: x22117491

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ram Abhilash Vasamsetti
Student ID:	x22117491
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Athanasios Staikopoulos
Submission Due Date:	14/12/2023
Project Title:	Using Time Series Predictive Models for Early Detection of Gambling Addiction in Problem Gamblers
Word Count:	8,303
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using Time Series Predictive Models for Early Detection of Gambling Addiction in Problem Gamblers

Ram Abhilash Vasamsetti
x22117491

Abstract

Gambling is a game of chance and addictive in nature due to its uncertain lucrative outcomes. Gambling Addiction has become one of the most challenging aspect for gambling operators and governing bodies as it negatively impacts the user and their immediate family. Early Detection and Prevention becomes complicated due to subtle behavioural changes and patterns. Modern technology have opened doors to multiple platforms available both offline and online that allow users to engage in gambling on almost any events of change such as sports outcomes to simple lottery games.

This work focuses on early detection of gambling addiction in players who engage in online betting by analysing their bet wagering patterns and other demographic data. This paper combines K means clustering algorithm with LSTM and ARIMA/SARIMA time series forecasting models to identify various groups of players based on their demographic and wagering activity and forecast their future betting patterns. High risk players' future betting patterns are forecasted using predictive models. Identifying high risk betting early will help in providing Responsible Gambling interventions to safeguard players mental health and limit financial losses.

The work clusters the players based on their Demographic and Wagering data into 3 groups. The players with problem gambling are successfully identified by analysing the mean of features. The wagering patterns in this cluster are forecast and validated using RMSE and MAE. An Average RMSE of 1247.9 is obtained for ARIMA model and better 1239.42 RMSE for SARIMA model. LSTM received a RMSE of 1524.24 despite solving highly non stationary data points. The forecast graphs are plotted for each user showing their predicted bet wagering pattern.

The 2 model architecture has gained new insights in the domain of Gambling addiction detection and prevention where not only high risk gamblers are identified but also possible future patterns where a player could engage in high bet wagering, are predicted. The findings of this paper will enable a foundation of a method to closely monitor high risk and potential risk players and provide early interventions to prevent them from gambling addiction risks.

1 Introduction

1.1 Background

Gambling is a game of chance where a person bets on an event that is random and unpredictable. The events can be a simple as guessing the number when a dice is rolled to

betting on a horse that could win in a race. The results are either winning a reward or loosing the amount wagered. The unpredictable nature of gambling resulted in its prohibition and restrictions in various countries (Lee, 2009). The market for gambling is a 465,000 million US Dollar industry in the year 2020 with an annual CAGR of 7.7 percentage (BusinessWire, 2022) . Recent advancements in technology has made gambling available online on mobile phones and websites with a 75 billion USD industry (Bloomberg, 2023). The ever growing industry comes with a major consequences which is Gambling Addiction. The lucrative profits that the several gambling platforms claim make users hooked to the platform and this leads to addiction similar to other substance addictions such as towards alcohol. Being habitual and addictive, Gambling addiction is classified as non substance addiction in the 5th Edition of Diagnostic and Statistical Manual of Mental Disorders also known in DSM-5. DSM-5 elaborates on Gambling addiction and its harm to the person and their immediate family (American Psychiatric Association, 2013).

Having a risk to mental health and economic impact, gambling is frowned upon and governing bodies impose regulations and restrict any gambling related activities in their own countries (Lee, 2009). Despite the restrictions, the lucrative monetary gains act as major enabler for this industry. Thus, each country enforces gambling operators to follow Responsible Gambling practices which helps the players when they are adversely affected by gambling (Blaszczynski et al., 2004). They can be called as Problem Gamblers who uncontrollably engages in gambling despite large financial loss and mental health affects. This problem requires interventions and latest advancements in technology can aid to it. The section 2 describes the work done in this domain of gambling addiction detection. Unpon review of multiple works in the domain, it can be understood that Detection and Identification of Problem Gamblers is of primary focus.

This paper proposes early detection using time series prediction for this problem. The term Early Detection which will be used in this paper, is defined as predicting future wagering pattern of Players, in early, to initiate responsible intervention by the betting organizers and limit large financial losses. These include warning prompts or enforced limits on the number of risky bets.

Several interesting approaches in early detection of such Players have been identified in literature review, Suzuki et al. (2019) had utilized a concept shapelets for early detection to classifying the users into problem and casual players which needs the players to show a micro set of patterns noticeable in problem gamblers which indicates the player shows repetitive bet wagering. The work by Lajcinová et al. (2021) treats gambling addiction betting patterns as anomalies and uses anomaly detection system built using LSTM and Adversarial Autoencoder (AE). The work detects the problem gamblers as early as first weeks of joining a betting platform. Both works detect problem gamblers but this paper aims to predict when a player will engage in high risk betting than detecting if they are a problem gambling player or not. To address this, a forecast of bet patterns placed by players can be forecasted. This will provide insights on the timestamps at which high risk betting are done. This approach will help in early detecting one of gambling addiction signs which is loss chasing tendencies described in DSM-5 (American Psychiatric Association, 2013). This research niche is addressed in this paper by fitting the user Turnover, Hold and other timestamped player data into Time series models. ARIMA, SARIMA and LSTM are used for this purpose. RMSE and MAE are used to validate the model performance and the graphs to forecast the Actual vs Predicted wagering patterns are plotted.

1.2 Research Question

To what extent can LSTM and other Time Series Forecasting models be used to predict the future betting patterns of a Problem Gambler, which,in turn, can be used to detect early signs of Gambling Addiction in Players ?

1.3 Document Structure

The Introduction section briefs about the Problem background elaborating the motivation with Research Question. Section 2 elaborates the related Works and Critical Analysis leading to Research Niche. Methodology in section 3 contains the Methodology that is carried on in this paper. Details of the Models being used. Section 4 has discusses the Design Specifications of the method. Implementation in line with the methodology described in the 5th section. The 6th section discusses the evaluation metrics used, the experiments conducted to achieve the results and followed by discussion on the results obtained. The section 7 discusses the Conclusion and Future works of this paper followed by References to key papers used.

2 Related Works

Research on previous works indicate the primary focus had been the detection of problem gambling than early detection. Early Detection in terms of this paper's objective is to identify when the player might engage in heavy betting. As per DMS-5 this tendency can be described as loss chasing. Which when detected early can help in controlling gambling addiction (American Psychiatric Association, 2013). Detection of gambling addiction in players is a beneficial finding for both governing bodies and gambling operators since it helps in sustainable business and saves the companies from potential liabilities or legal issues. A heavy gambler poses risk to his immediate family and themselves and, thus, the research had been heavily biased towards detection of such players. This leaves a fewer work on research and development towards early detection and upon investigation of the latest work, one can understand the current solutions that are practiced to early detect and control players with problem gambling. This section has been divided into four subsections each covering an approach and its research towards the solution. The first section focuses on survey based detection where users are requested to fill out forms which help in screening players based on their responses. Following section looks into prediction based on voluntary self-exclusion (VSE) user data. Third subsection is crucial to this paper since it explores the research done under time series prediction for the domain. The state of the art is understood followed by the research niche that this current paper addresses.

2.1 Problem Gambling Prediction using Surveys

Problem Gambling Severity Index (PGSI) is a standard set of questionnaire designed to screen players having gambling addiction and is highly reliable to determine the user's risk taking tendencies. (Jim Orford and Erens, 2010). This index have been extensively used by many researchers to predict problem gamblers by analysing the players betting wagers and validating against their PGSI score. The paper by Howe and et al. (2019) indicates the essential predictors that can be derived from PGSI such as affects on family, mental health and frequency of gambling. PGSI has also effectively been used to

understand the risks of un-intervened problem gambling which highlights the need for early intervention (Currie et al., 2021). The works of Seo et al. (2020) and Murch et al. (2023) utilize these metrics and perform Random Forest model training using players demographic and wagering data to predict the PGSI values of users who have self filled the survey. Seo et al. (2020) uses the PGSI survey results derived from Korea center for controlling gambling addiction in youth, South Korea. Bet outcome in terms of profit and loss, turnover have been considered as major features for predicting the index. PGSI being a survey has also given the researchers information about their personal relationships. This plays a crucial role as DSM-5 identifies gambling addiction to cause issues in personal relationship (American Psychiatric Association, 2013). Murch et al. (2023) has derived similar features from the data but has requested up to 12 months of player user data whereas Seo et al. (2020) used 3 months of usage data. Both the works propose Machine Learning classification models such as SVM and Random Forest for the prediction. Despite having different feature selections in both the papers due the nature of data and the regional factors, the papers achieve similar accuracy. Seo et al. (2020) have achieved 71.8 using Random Forest and 0.507 with SVM and Murch et al. (2023) have gained a slight better accuracy of 75.93 for PGSI 5+ users and 73.4 for PGSI 8+ users. It is worth noting that the F1 score has a marginal degrade for 5+ and 8+ PGSI and a larger degrade from 46 percent to 29 percent in Precision for the same.

The papers that were discussed above have used PGSI to validate the prediction applied on the player usage data. The PGSI has few disadvantages as it can be biased and dependent on the players truthfulness while answering the survey. The paper Jim Orford and Erens (2010) also indicates the gender based bias in the index where the female compulsive betting is underestimated. It is also worth considering the stereotypes that are associated with gambling which causes many players to not admit and honestly answer the survey Hing et al. (2015). Thus, considering all the facts, this methodology doesn't utilize PGSI responses for prediction. The next subsection will focus on Self exclusion and limit setting data to predict problem gambling.

2.2 Predicting Limit Setting and VSE to Detect Problem Gambling

Limit-setting and voluntary self-exclusion (VSE) is a service provided by the both offline and online gambling operators to let the players self exclude themselves from using the gambling platform. Once initiated the players cannot access the the platform for a predefined period of time. Walker et al. (2015) suggests the usage of Limit-setting where a limit is placed on the amount gained or lost per day which facilitates gambling in a responsible way. The works of Auer and Griffiths in this domain provide valuable information for this literature review. Their multiple papers include Limit-Setting prediction (Auer and Griffiths, 2022a), self-reported problem gambling prediction (Auer and Griffiths, 2022b) and Predicting high-risk gambling players based on their first few weeks of activity Auer and Griffiths (2023). The initial paper demonstrates the possibility to predict the limit setting tendency of a player in near future based on the 33 features derived from the player usage data. The features included demographic information from age, gender and bet related features such as bet limit, daily Turnover etc. Users who exceed 80 percent of their monthly limit threshold and users who changed their limit to higher values have been considered as a PG. The change in limit shows the person is engaging in compulsive betting. The work also concludes that the users were, significantly, able to limit compulsive betting while using VSE tools. The work by Auer and Griffiths

(2022a) uses Random Forest algorithm for the prediction and have received an average of 73 percent accuracy. Random Forest Tree and logistic regression model was combined to achieve the best solution with 75 percent accuracy. The work of Auer and Griffiths (2022a) facilitates early detection and sheds light on advantages of early prediction of limit setting behaviour.

The later paper by (Auer and Griffiths, 2023) uses the first week of player data after registration and validates against the high risk classification of 90 day player data. The independent variables derived from the first week of player data, which include, amount wagered, amount lost or won (ranging from positive value for profit and negative for loss), Gambling account depletion (sessions that end up having less balance) and Frequentdepositing (Deposits made in a session in average), are used. The work had achieved 73 percent accuracy using Random Forest and 67 percent using Gradient Boost Machine Learning model.

Both the solutions from Auer and Griffith, and (Percy et al., 2016) are successful in detecting problem gamblers with 70 percent average classification accuracy. The literature review done so far shows that for the reviewed methodologies to work, the users need to either use VSE tools or engage in limit betting. This paper aims to investigate methods to predict problem gambling that do not require users to use any VSE tools. This will help in identifying and helping many users that do not realize being addicted to gambling. This research niche identified, requires further investigation towards deep learning models. Deep learning models enable forecasting and better analyse of time series data when compared to Machine Learning model (Lim and Zohren, 2021). The next subsection explores the research done in time series prediction in the gambling addiction domain using Deep Learning.

2.3 Using Time Series models and Deep Learning Techniques to detect Problem Gambling

The core idea of this paper is to solve the compulsive gambling prediction problem using Time series models. The paper by Akhter (2017) shows a systematic approach to address the time series data. The paper implements a Detection system that monitors the player usage data for a fixed period of time (30 days). The features used are aligned with the current research where Profit/ Loss, number of limits exceeding, Wager amounts, Turnover, Loss Chasing are used with SVM and other Supervised Learning models. The work highlights the need for clustering in-order to identify clusters of users that are ranging from low likely to have gambling addiction to problem gamblers. Usage of Classification models have been long observed in this problem where Random Forest have been used to detect two classes (Auer and Griffiths, 2022a) (Auer and Griffiths, 2022b). But Further research in Time series clustering indicate the usage of multi clusters to segregate players based on player usage data. The work of Peres et al. (2021) divides the user data into 4 clusters using K means Clustering and this work gives the current research a direction in its methodology. Akther's work has shows existence of bias due to oversampling and it is expected since problem gambling is considered as an anomaly with a percentage of less than 5 percent of the total players in an average Casio setup (Lajcinová et al., 2021). The work uses SMOTE method to mitigate the oversampling, but Peres et al. (2021) doesn't specify the need for balancing for clustered data. Thus, the methodology will compare the impact of oversampling to ensure better accuracy. Akther's work has a drawback of large computation requirement and a constant monitoring system to handle

real time data. The work of Peres et al. (2021) is mitigating the issue with clustering but the research question of early detection is still not answered. Extending the research to future predict the classes identified will bring new insights about players that are showing early signs of gambling addiction.

Another approach by Suzuki et al. (2019) for solving time series data in this domain includes the concept of shapelets. The micro timestamps are extracted from Problem gamblers. The timestamps are captured against turnover, profit loss per session and other parameters over variable time frames. The variable time frames are addressed using Dynamic Time Warping to correlate shapelets against Identified problem gamblers entire time series data. This method received an accuracy of 79 percentage and proves to be a novel approach for this domain. The paper's work is very relevant to this research in terms of objectives but has a drawback for have a requirement that the novice players to express similar betting pattern of that of a established problem gamblers which is causes an issue due to a very less availability of problem gamblers data (¡ 5 % of all cases). The model doesn't identify users with low money wagering but repetitive patterns. The model will also cause higher false positive predictions as number of PG shapelets increase. The current research objective plans to early detect and predict while staying scalable and support in detecting high risk players and low wagering compulsive betting players as well.

The work of Lajcinová et al. (2021) has an interesting approach towards solving Time Stamp data where gambling addiction players are considered as Anomaly and a transformer based auto-encoder is designed to detect players showing early signs of gambling addiction. A completely unsupervised learning method. The paper uses deep learning methods such as LSTM, CNN and Transformer based Autoenoder. The Autoender score is projected for each time series data to detect players that are showing unusual betting patterns. The work signifies the effectiveness of Deep learning models such as LSTM and CNN. Lajcinová et al. (2021) have used Deep learning models to predict the anomaly score. The players are classified based on the anomaly score but their future betting patters are not known. The current paper understands the methods that can be utilized to detect a problem gambler but aims to extend it further and forecast their future betting pattern to early detect problem gambling. Thus, the work of Lajcinová et al. (2021) aligns greatly with the current paper and deep learning models are used. Keeping in mind the research question of forecasting problem gambling patterns, the anomaly detection method is not utilized.

2.4 Research Niche

After studying the papers and understanding the work of researchers in the same domain and direction of gambling addiction prediction, the methodology is designed. The paper's objective of early detection is aimed to be achieved by forecasting the player's future betting pattern based on their historical performance. The current paper focuses on unsupervised learning models to classify clusters as work done by Peres et al. (2021) and Suzuki et al. (2019). The solution is later extended by utilizing ARIMA/SARIMA for stationary data points and Deep Learning methods as suggested by Lajcinová et al. (2021) for non stationary data points. The clusters with moderate to problem gamblers is considered for the forecasting models. This is to reduce the computation efforts and scale solution to larger data sets. The user's actual vs predicted wagering pattern is plotted. The final results are validated using RMSE and MAE and reported in results section of

this paper.

3 Methodology

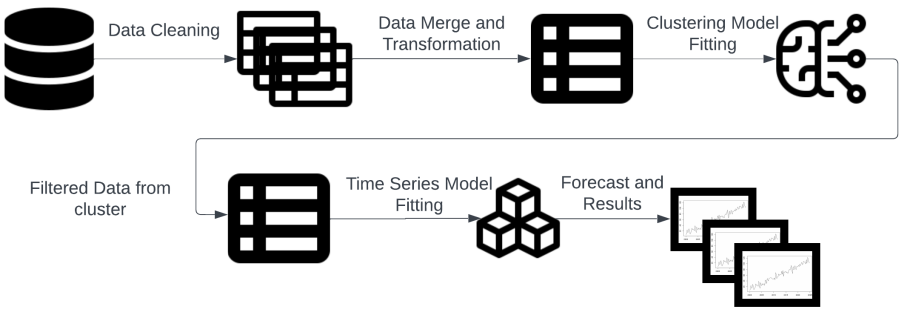

Figure 1: Methodology workflow

3.1 Sourcing of Data

The dataset has been derived from online betting operator Bwin Interactive Entertainment, from Vienna, Austria. This anonymized dataset containing the player usage data is made available by The Transparency Project. A publicly available website by Harvard Medical School to encourage research and development towards Addiction Prevention and Control Division on Addiction (2021). The data collection of Gray et al. (2012) have been used for this research. The entire data is divided into three datasets. The first dataset contains demographic data such as [Player Gender], [Preferred Language], [Joining Date] and [Place of Origin]. The second dataset is the player usage data holding the Turnover, Hold, NumberofBets information. This timestamped data is later used for forecasting. The third dataset contains the Responsible gambling interventions and their logs with important features such as [RG] which states if the player has undergone any initiatives by the operator. [InterventionType_first] indicates the type of intervention initiated.

3.2 Data Preprocessing

The datasets undergo cleaning as first step in the methodology flow as in Fig. 1 and checked for null values or missing values. The three datasets are merged over common key. Label Encoding is performed for class columns. The new merged dataset is checked for duplicate values, null and empty values. The zero value cells are unchanged due to the nature of data. 0 values indicate the user not engaging in gambling on the given date and thus this values are not changed. After performing feature selection using Correlation matrix, The dataset is fitted with K means clustering algorithm as its first stage of model fitting. Post model fitting the dataset is further filtered using the UserId's of the Moderate Problem Gambler's cluster identified by K- means algorithm. The time series data of all the users is adjusted between the window of 2002 and 2010 for better model performance. Each user's timestamp data indicating Turnover, Hold, Number of

Bets over the time of 2002 till 2010 is checked for stationarity as it is an important factor for time series prediction as discussed by Puranik et al. (2023).

3.3 Data Modelling

The first model classifies the merged dataset into clusters using K - means clustering. Peres et al. (2021)'s work shows the 4 major ways problem gamblers can be classified from casual players to problem gamblers. The model is first fitting with K - means clustering as the first model shown in the Fig. 1 and tested with K value range from 2 to 10 and the best k value is identified using elbow method. The clusters are labeled by exploring the mean Turnover, mean Hold and other Features. The cluster with moderate Problem gambling users are identified. Puranik et al. (2023)'s work indicates stationarity as a requirement in time series prediction. Thus, the model uses stationarity to reduce its computation time by fitted stationary data points with ARIMA and SARIMA models. The non stationary data points are much complex to be predicted with statistical models. Thus, LSTM (Long Short Term Memory) is used as suggested by the work of Cao et al. (2019).

3.4 Evaluation Metrics

The performance of K - means clustering algorithm is validated using Silhouette score and Elbow method for best K value. The results of Test and Predicted forecast of Problem Gambling users are visualized using Mathplot graphs and are are validated against RMSE and MAE and can be found at the Section 6 of Evaluation.

This method has been implemented to forecast the wagering pattern of a player which can be used to early detect high risk gambling behaviour and limit its effect. This problem of gambling detection has been addressed as a classification problem in the Machine learning domain (Peres et al., 2021), (Akhter, 2017) and (Suzuki et al., 2019). This solution will open doors to new approach to early detection and predicting the problem gambling behaviour. One of the research objectives is to pivot the research focus from detecting problem gambling to develop methods to detect problem gambling and high risk behaviour in advance. It is to be understood that without proper intervention almost 90 percent of Problem Gamblers tend to relapse (Schreier, 2022)

4 Design Specification

This paper follows a three stage model with the first stage as Data Processing, followed by Model Fitting and Lastly Visualization as shown in the Fig. 2 The Data Processing stage uses the three datasets in .csv format and undergoes data transformation detailed in section 5 of Implementation. The formatted data goes through two models in Model fitting stage. The Data transformation is optimized based on K means clustering results. The first model is K- means Clustering algorithm implemented by using KMeans library from SKlearn. The filtered data goes through ARIMA, SARIMAX (Seasonal ARIMA with exogenous Factor) and LSTM. The final stage of visualization plots the forecasted wagering pattern(time series forecast) of the users which are identified by the k means clustering model. Both models are configured based on visualization output (number of clusters, LSTM units and hyper parameter tuning). The data preparation is also iterated based on the visualization output. The features are selected based on visualization output

of the clusters. This design with iterative approach makes sure the methodology has scope for improvement at every stage of process. The entire implementation is carried out in Jupyter Notebook using Python programming language and Visualizations are plotted within jupyter Notebook using matplotlib which is a python library used for visualizing data.

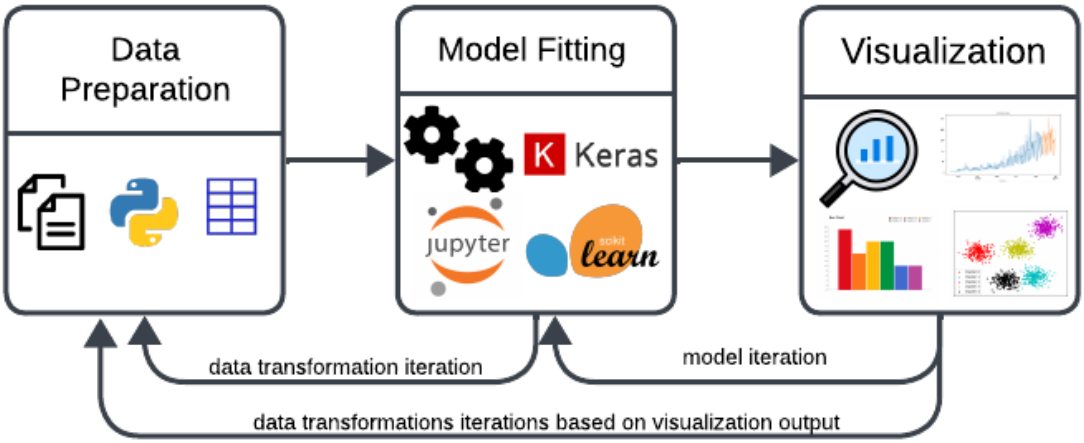

Figure 2: Design Specification

5 Implementation

The Fig. 3 shows the implementation flow which will be elaborated in the following subsections.

5.1 Data Gathering

The 3 datasets used in the implementation are derived in .CSV format. The datasets are made available by the Transparency Project which is an Harvard Medical University webside dedicated to progress research towards substance and non substance addiction (Division on Addiction, 2021). Datasets have been imported into Jupyter Notebook and loaded into Python Pandas Dataframe. The datasets have key features such as UserId, Timestamped Turnover, Hold, Number of Bets for each UserID, Demographic Data of User such as Language, Gender and Age. The dataset also contains Responsible Gambling data where UserID's are logged against a Player when the they have undergone Responsible Gambling interventions. The table shows the Features and their definitions available across the three datasets.

5.2 Data Transformation

The three datasets are loaded into three dataframes as mentioned in table above. The Transformations begin by filling the empty and null date columns with a standard date of

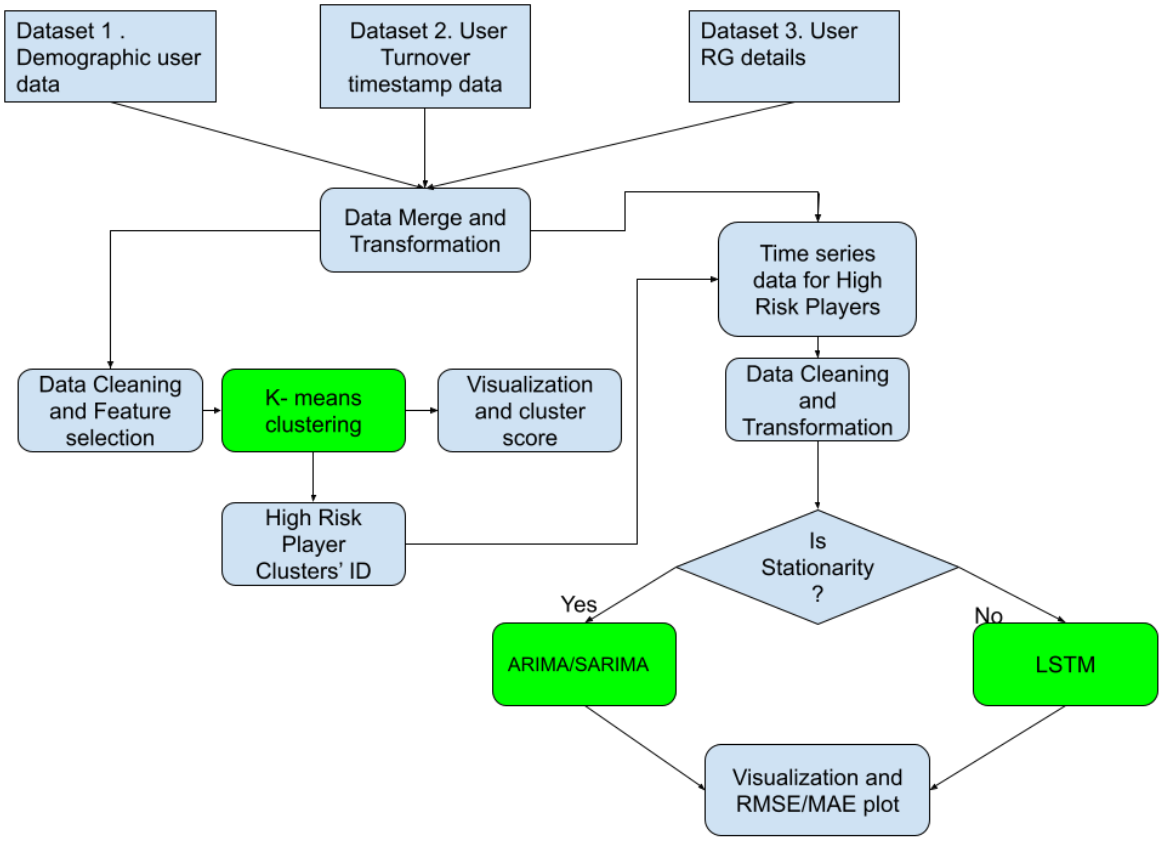

Figure 3: Implementation Flow

"01/01/1900". 5 date columns are filled over the three dataframes daily_agg_df, rg_det_df and demog_df (Date,Registration_date, First_Deposit_Date, RGFirst_Date , RGLast_Date). Label encoding is performed for 3 columns (CountryName, LanguageName, Gender). Year of birth has been converted to age at 2010 (Age_until_2010) since the prediction is done with data till 2010. Renaming of columns are done and finally all the three datasets are merged over UserId as key. Post merging the dataset, the merged_df has 981803 rows and 18 columns. All date columns except Aggregate_date are dropped. The timestamps of users who have played casino online has been filtered using Product_type and stored in filtered_df. The dataset indicate 331828 entries for online casio out of which 294375 entries have RG column as 1 and remaining 37453 entries have RG as 0. RG is a primary factor that enables us to identify users who have been identified by casino operator as people who have either taken help of the company to gamble responsibly or have undergone interventions from the company to limit their gambling due to Problem Gambling. Duplicates and null values are checked and removed for the filtered_df and Heat map in Fig. 4 is used to identify high correlation. Based on high correlation value LanguageName is removed from features. The dataframe is pivoted to hold UserID as column along with each of the Turnover, Hold , Number of Bets and other feature values in a series of date columns from 2002-11-12 till 2010-11-10. The data is stored in a 3 dimensional array named user_data_3d. 3161 users have been identified in total.

After fitting the transformed data with K- means clustering algorithm. The moderate PG cluster is considered for further model input. The cluster with label [Moderate

Table 1: Feature Description

Variable	**Type of Data**	**Description of the Variable**
UserID	Number	User ID that has been given to each user anonymizing them from identification.
RG	Number	Involved in any of the Responsible Gambling Interventions either by themselves or by Operator. (0 = yes, 1 = no)
Country	String	Players contry of origin
Language	String	Language preferred by the user
Turnover	Number	Total Betting amount wagered in a day in EUR.
Hold	Number	Total amount lost in a day in EUR.
NumberofBets	Number	Total number of bets placed in a day ini EUR.
Interventiontype first	Number	Each number indicates different type of intervention first experienced by the user.

Problem Gambling] contain 39 users. Stationarity check is used to divide the 39 users into two datasets arimausers and lstmusers. Augmented Dickey–Fuller (ADF) test is used to perform Hypothesis testing and The datapoints that have p value less than 0.05 for all of the 5 columns (Turnover, Hold , NumberofBets ,Age_until_2010, Interventiontype_first) are classified as stationary data points. The p value 0.05 for any of the column indicates presense of non stationarity and thus they are stored under lstmusers dataset. In total 29 users have been identified as stationary users and ARIMA/SARIMA is performed on them to forecast and 9 users have been identified as non stationary which are forecast using LSTM.

5.3 Model Fitting

5.3.1 Using K means Clustering as first model to Group the Players

K means Clustering Machine Learning Algorithm divides the datapoints into K number of distinct, non overlapping clusters. The model can utilize temporal data and divide them into clusters based on the trends and patterns in the data over time Ikotun et al. (2023). The model is implemented using KMeans Library from Sklearn. 8 features have been selected from feature selection ('Turnover', 'Hold', 'NumberofBets','Age_until_2010','Interventiontype_first', 'CountryName','Gender','Event_type_first') and pivoted to add users and timestamp as columns with features as layers . This data is stored in dataframe user_data_3d. The dataframe is checked for null values and cleaning to make sure the pivoting was successful. The K value which determines the number of clusters is set to 4 initially with suggestions from (Peres et al., 2021) and (Suzuki et al., 2019). But the elbow method indicates 2 to 3 as best K value and thus , 3 has been taken as K value for the Cluster. The model is fit and it was observed that the clusters cannot be visualized on a two dimension plane as non overlapping clusters. In order to visualize the clusters effectively. Principal Component Analysis (PCA) has been implemented. The Fig. 6 shows the final clusters with PCA as x and y axis. The Silhouette score of 0.90 (range of -1 to +1) indicate good K value and cluster separation. The

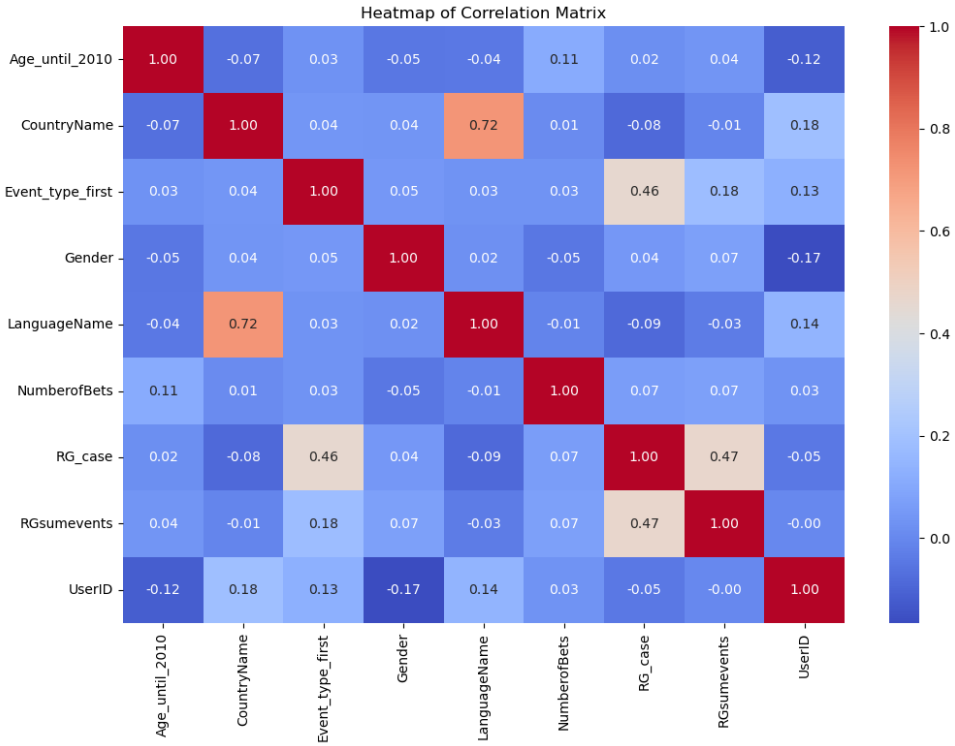

Figure 4: Heat Map for Feature Selection

output clusters are unlabeled and thus its crucial to understand what the cluster mean. The mean of all the features used are analysed in Fig. 5 and it can be observed that high mean Turnover, Hold and NumberofBets are clustered. The labels are assigned to the clusters with low mean as [Early Players] and later as [Moderate Problem Gambler] and one user has been identified as [Problem Gambler]. It is interesting to know that 90 percent of users in cluster [Moderate Problem Gambler] had undergone Responsible Gambling initiates at least one time between 2002 till 2010. This is realized from the column [RG] which was not included in the clustering model. This indicates the model is able to categorize Problem gamblers with possible problem gambling and we can further analyse their wagering pattern to forecast their future betting patterns. Moderate Problem Gambling cluster contains 38 users and PG cluster contains a single user and added to the Moderate Problem Gambling cluster totalling 39 users which will be further analyzed.

5.3.2 Using ARIMA and SARIMA on UsersId with Stationary columns

Puranik et al. (2023) work calls for the need for early detection and importance of stationarity for better performance of Time series forecasting models. The Moderate Problem Gambler cluster data is tested to check their stationarity using Augmented Dickey-Fuller (ADF) Test (Mushtaq, 2011). Hypothesis testing is performed over 5 features (Turnover, Hold, NumberofBets, Age_until_2010, Interventiontype_first) of each user in Moderate Problem Gambler cluster. ADF value than 0.05 reject the null hypothesis and are labelled as stationary. The rest are marked as non stationary. At total 30 users out of 40 are identified as stationary and 10 as non stationaary. The 29 user timstamp data are fit are looped through Time series Forecasting model. Test Train split is done for each

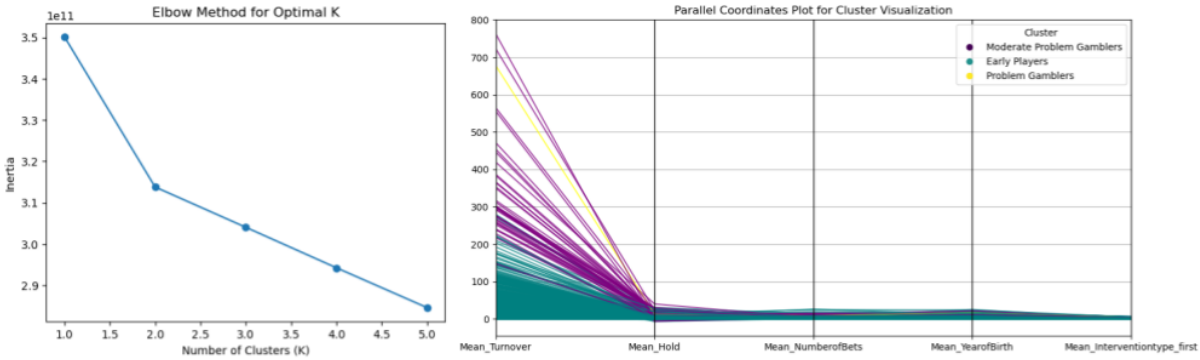

Figure 5: Elbow method for Optimal K (left) and Cluster Labeling Analysis (right)

datapoint and undergone Standardization using min-max scaler. The data is fit into two statistical models : ARIMA and SARIMAX. ARIMA model (Autoregressive Integrated Moving Average) is implemented from statsmodels library with three parameters p(order of Autoregression as 1), d (Degree of Differentiation as 1) and q (order of moving average as 1). The SARIMA (ARIMA with seasonality) model is implemented by SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Factors) method from statsmodel library. This is an advanced version of ARIMA with seasonality and Exogenous factors. Exogenous features cannot be influenced by other features and such features can be fitted with SARIMAX. SARIMAX has the same parameters as ARIMA with a seasonality factor (Arunraj et al., 2016). The model is given 14 which indicates a pattern repetition of 14 days. This value was set keeping in mind the bi-weekly repetition observed in user wagering pattern. The forecasting betting pattern for each user is plotted in a loop. The forecast can be observed in Fig. 7 for a single user with UserID 1175809. The metrics used for the results are RMSE (Root Mean Squared Error) and MAE (Moving Average Error) plotted for each user. The average RMSE and MAE is calculated for 10 users and are discussed in Evaluation. The ARIMA RMSE of user 1175809 is observed to be 784.768 and MAE of 331.96. The SARIMA results of the same are 777.97 and 336.02. This indicates SARIMA model is better in comparison to ARIMA. SARIMA graph in Fig. 7 (top) shows the forecast was able to identify the timestamps at which Problem Gambling or Risky betting is observed. The figure also shows the datapoints at which high variance in gambling has occurred. This is achieved by finding the absolute difference in Turnover column over a threshold of 100.

5.3.3 Using LSTM on UserId's with Non Stationary columns

Long Short-Term Memory (LSTM) is a RNN (Recurring Neural Network) model widely used for Time Series data forecasting (Cao et al., 2019). LSTM can capture long term dependencies of time series data and thus can be used for non stationary datapoints (Preeti et al., 2019). 10 users are identified as non stationary. Turnover is used as data series input, flattened and split into test train data. The testing and training data are converted into windows of 1 week each. The model uses previous window of training data and forecast the next window. This enables the model to learn from historical data and reduce the loss error. The test and training data are scaled and hyper-parameter tuning is performed. The parameters were tested for units_values = [64, 128, 256] and learning_rate_values = [0.01, 0.001, 0.0001]. Hyper-parameter turning has returned the

values 64 and 0.001 as best parameters. The LSTM model is trained using the training data with the above mentioned units and learning rate. The model is further optimized using Adam optimizer. The model is looped over all the 10 user's data and the results are plot in sequence. The LSTM model is evaluated using RMSE and the plots for each user indicate good overlap between the forecast and test data. The Fig. 8 shows the results for one of the user 1776178 where the forecast can be observed. The timestamps where the model has predicted the occurrence of high risk gambling is logged. The RMSE 699.691 and MAE of 226.756 has been observed for User 1776178 and The overall Average RMSE of 1524.246 and MAE of 671.775 has been observed for LSTM model.

6 Evaluation

This section elaborates the Key metrics used in evaluating the performance of the models used. Clustering Model is validated using Silhouette score. RMSE and MAE is used for time series models used after clustering the dataset.

6.1 Silhouette Score

Silhouette score is metric used to assess the performance of clustering models such as K - means clustering algorithm (Shahapure and Nicholas, 2020). This is achieved by measuring how isolated the datapoints are present in a cluster when compared to other cluster. High scores can be observed in clusters where there is no overlap. The score is ranged from -1 to 1 where higher value indicates best separation of clusters. The score is attained for each i data point and the average of the score of each datapoint gives the Average silhouette score.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where a(i) The distance between i data point to other data points in same cluster. b(i): The distance from the i data point to data points in a different cluster

6.2 RMSE score

Root Mean Squared Error or RMSE is an metric used for evaluating the models performance in regression data. RMSE are useful metrics in time series data as the model is evaluated based on its capability to forecast values as close as possible to real values (Yorucu, 2003). This paper uses RMSE at the second part of implementation where Time series models are implemented. Both ARIMA/SARIMA and LSTM are evaluated based on their Root mean Squared Error value. The range of RMSE lies between 0 till infinity and the closer the value to zero the better.

$$RMSE = \sqrt{(\sum (y\_pred - y\_true)^2)/N)}$$

Where y_pred is the predicted values and y_true is the validation datapoints.

6.3 MAE score

Mean Absolute Error (MAE) is used for evaluating the accuracy of regression values. In time series it can be used to understand the models efficiency in predicting the true values (Yorucu, 2003). It is achieved by calculating the average absolute difference between actual and predicted values. MAE is calculated with the formula below.

$$MAE = \frac{1}{N} \sum |y_{\text{pred}} - y_{\text{true}}|$$

Where pred is the predicted values and true is the true values.

6.4 Experiment 1 - Evaluation of K means Clustering

Clustering using K means Clustering for Identifying Problem Gambling

The K - means clustering algorithm is fitted using player's pivoted data with optimal K value as 3. The silhouette score of 0.9098 is achieved. The cluster output is visualized in the Fig. 6. The users are divided into three clusters and are labelled as Early Players, Moderate Problem Gamblers and Problem Gamblers. The labelling is validated by Analysing mean of Turnover, Hold, Number of Bets, age and Intervention type, in the Fig. 5

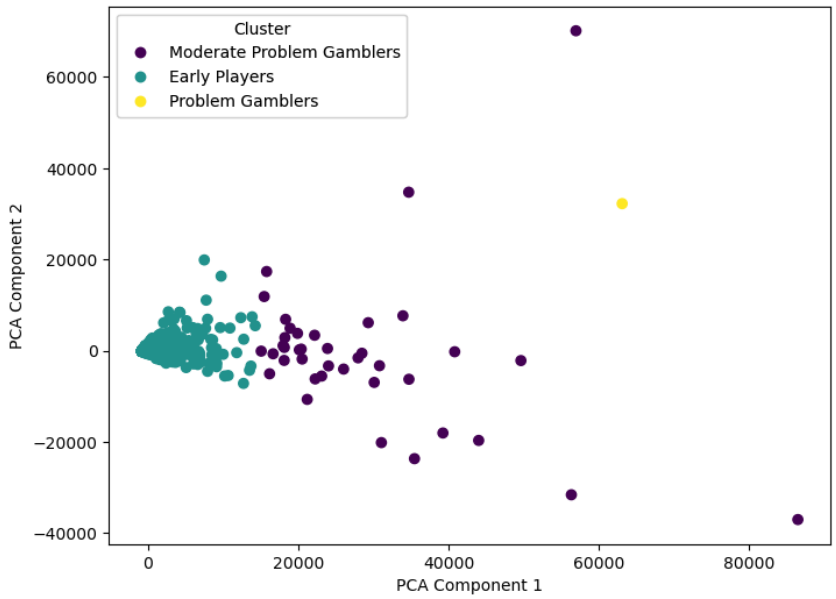

Figure 6: K - Means Clustering Graph

6.5 Experiment 2 - Evaluation of ARIMA on Forecasting

The 29 users are fitted with ARIMA model with order parameters(1,1,1) and their RMSE, MAE values are reported in the table below. The mean of RMSE and MAE are calculated for 10 users to calculate the average error. We can notice in the Table 2, the ARIMA RMSE and MAE is varying for each user. This is due to each users relative wagering amount while playing. Thus, Average RMSE and MAE is calculated. The ARIMA model has an average RMSE of 1247.96 and average MAE of 877.19

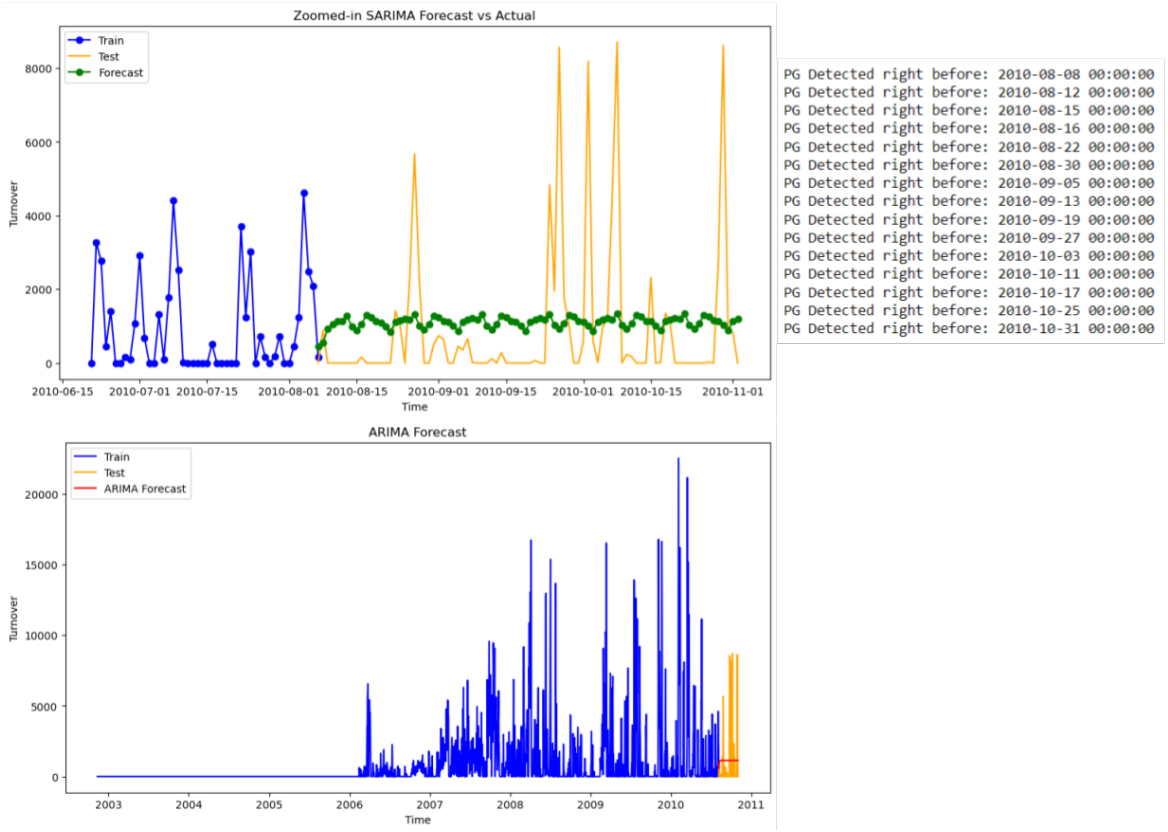

Figure 7: SARIMA (top) and ARIMA (bottom) forecast for UserID 1457496.

6.6 Experiment 3 - Evaluation of SARIMA on Forecasting

The same 29 users are fitted with SARIMAX model which is a SARIMA model with exogenous factors. 97 % for training, 3 % for testing is being used with same order parameters as ARIMA model but with additional seasonal order of (1, 1, 1, 14) The last additional parameter is the seasonality which has been set bi-weekly. The results are compiled in the Table. 2 using RMSE and MAE. The SARIMA model has slightly better average RMSE of 1239.42 but higher MAE of 1039.80 when compared to ARIMA model.

Table 2: ARIMA and SARIMA Results for 11 Stationary Users

UserID	ARIMA RMSE	ARIMA MAE	SARIMA RMSE	SARIMA MAE
868583	22.402	22.396	102.576	79.833
1175809	784.768	331.960	777.976	336.026
1411743	4616.157	2204.101	4613.733	2212.505
1457496	1978.891	1324.978	1977.662	1313.908
1486136	3753.673	3049.424	3891.255	3338.057
1662632	0.347	0.3459	65.029	53.443
1679490	978.136	636.149	985.506	612.500
1790848	0.001	0.001	55.193	41.993
1921204	573.983	453.599	564.291	437.393
2070894	1318.902	939.989	1338.566	975.863
4754125	1771.224	946.181	1841.963	1036.222
AVG :	**1247.969**	**877.195**	**1239.426**	**1039.804**

6.7 Experiment 4 - Evaluation of LSTM on Forecasting

The LSTM model is fitted with 9 users Turnover data which are identified as non stationary. The LSTM model is fitted with 64 units and 0.001 learning rate. The results are forecasting in plot as in Fig. 8 and the Table 3 indicates the RMSE and MAE of each user and their Average Accuracy. Despite having Highter Average RMSE of 1524.246, the LSTM model has lowest error of 671.775 and the Fig. 8 shows LSTM model's forecast with good overlap.

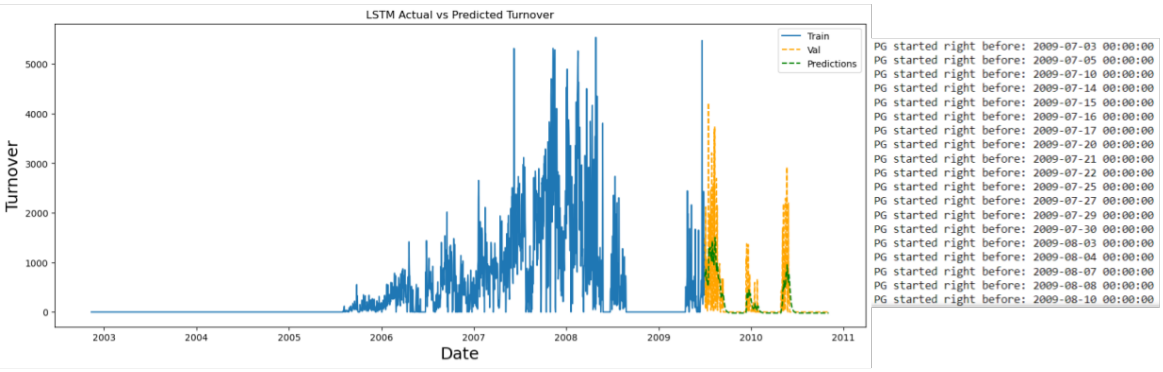

Figure 8: LSTM Forecast for the UserID 1776178

Table 3: LSTM Results for 10 Non Stationary Users

UserID	**LSTM RMSE**	**LSTM MAE**
1776178	699.691	226.756
3852889	1465.642	844.466
3904422	1438.356	607.398
4371320	1904.047	938.596
4412550	817.908	435.184
4495603	452.762	116.345
5308271	2993.712	1485.720
6158120	2064.926	878.535
6239380	322.380	86.723
6985339	3639.499	1516.271
AVG:	**1524.246**	**671.775**

6.8 Discussion

The silhouette score of 0.91 indicates good clustering of the datapoints using Principal Component Analysis (PCA). The Elbow method indicates the ideal K value of 3. Literature review suggest the ideal cluster size of 4, but by analysing the mean of Turnover and Hold of the clusters, It is observed that most of the data points are justified within 3 clusters ranging from low mean cluster, medium mean cluster and large mean cluster. The model doesn't make meaningful clusters beyond k value of 3 and drop in silhouette score can be noticed for higher K value.

The second part of implementation uses Time Series forecasting models. Based on Stationaraity, the solution has significantly reduced the computation efforts by utilizing statistical model such as ARIMA/SARIMA when the user data points are stationary. The results of each RMSE and MAE of all users is plotted in the Tables 2 3. It can be observed that in terms of Average RMSE, SARIMA [1239.426] has performed better than ARIMA [1247.969]. But Average MAE is lowest for ARIMA. But it is worth noting that ARIMA model is unable to capture the trends in the datapoints as in Figure. 7. Thus, SARIMA model has better forecasting than ARIMA model. On the other hand LSTM has performed better than ARIMA and SARIMA model in terms of capturing the sharp increases in turnovers as can be seen in the Fig. 8. But has an Average RMSE of 1524.246. This is larger than both SARIMA and ARIMA model but considering LSTM has been implemented on UserID with non stationary datapoints which are far complex to forecasting and differ from ARIMA/SARIMA UserID's, the comparison is not applicable . In terms of Average MAE, the LSTM model has the least error of 671.775, despite having non stationary data to forecast. Overall the models have captured significant part of the future betting patterns, despite have larger error rates as seen in their plots 7 and 8.

LSTM model has run 1934 steps over average of 300 seconds (5 Minutes) per user where SARIMA and ARIMA have an average of 15 seconds per user. The methodology has saved significant amount of time by processing stationary data points using statistical time series models without having much increase in the average error rate and as a whole solved the research objective of using Time series forecasting models to forecast Problem Gambling patterns in Players.

7 Conclusion and Future Work

The results from Fig. 7 8 answer the research question by forecasting the future bet wagering pattern exhibited by Problem Gamblers. By predicting in early when the user is going to place large and risky bets, the gambling operators can perform Responsible gambling interventions to avoid loss chasing. A tendency where a person keeps betting large and risky bets to recover all their losses. This tendency is known to exacerbate gambling addiction (American Psychiatric Association, 2013). Table 2 3 indicate the indivual and average RMSE and MAE for each models used. The division of models based on stationartiy has helped reduce the time taken to compute forecast for each user.

The work is able to answer the research question but has its own limitations. The work indicates the possibility of using time series forecasting as a method to early detect problem gambling. Further research using complex time series deep learning models such as Transformers can other RNN models can help further identify the Problem Gambling patterns. Since forecasting is working for each user. A dedicated interface application can be developed which can be used by gambling operators and governing bodies to monitor players and make sure they are gambling responsibly.

References

Akhter, S. A. (2017). Using machine learning to predict potential online gambling addicts, *ResearchGate* .

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, 5th edn, American Psychiatric Association, Arlington, VA.

Arunraj, N., Ahrens, D. and Fernandes, M. (2016). Application of sarimax model to forecast daily sales in food retail industry, *International Journal of Operations Research and Information Systems* **7**: 1–21.

Auer, M. and Griffiths, M. (2022a). Predicting limit-setting behavior of gamblers using machine learning algorithms: A real-world study of norwegian gamblers using account data, *International Journal of Mental Health and Addiction* **20**: 1–18.

Auer, M. and Griffiths, M. D. (2022b). Using artificial intelligence algorithms to predict self-reported problem gambling with account-based player data in an online casino setting, *Journal of Gambling Studies* .
URL: *https://doi.org/10.1007/s10899-022-10139-1*

Auer, M. and Griffiths, M. D. (2023). Predicting high-risk gambling based on the first seven days of gambling activity after registration using account-based tracking data, *International Journal of Mental Health and Addiction* .
URL: *https://doi.org/10.1007/s11469-023-01056-4*

Blaszczynski, A., Ladouceur, R. and Shaffer, H. J. (2004). A science-based framework for responsible gambling: The reno model, *Journal of Gambling Studies* **20**(3): 301–317.

Bloomberg (2023). Global online gambling market set to reach usd 75.15 billion by 2031, with a sustainable cagr of 12.5url = https://www.bloomberg.com/press-releases/2023-02-02/global-online-gambling-market-set-to-reach-usd-75-15-billion-by-2031-with-a-sustainable-cagr-of-12-5-growth-market-reports.

BusinessWire (2022). Global gambling market opportunities and strategies report 2022: Market is expected to grow from $674,703.9 million in 2025 to $895,720.3 million in 2030. url = https://www.businesswire.com/news/home/20220316005705/en/Global-Gambling-Market-Opportunities-and-Strategies-Report-2022-Market-is-Expected-to-Grow-from-674703.9-Million-in-2025-to-895720.3-Million-in-2030—ResearchAndMarkets.com.

Cao, J., Li, Z. and Li, J. (2019). Financial time series forecasting model based on ceemdan and lstm, *Physica A: Statistical Mechanics and its Applications* **519**: 127–139.

Currie, S. R., Hodgins, D. C., Williams, R. J. and Fiest, K. (2021). Predicting future harm from gambling over a five-year period in a general population sample: a survival analysis, *BMC Psychiatry* **21**(1): 15.
URL: *https://doi.org/10.1186/s12888-020-03016-x*

Division on Addiction, Cambridge Health Alliance, a. H. M. S. t. h. (2021). Behavioral characteristics of internet gamblers who trigger corporate responsible gambling interventions, Medford, MA: Division on Addiction, The Transparency Project [database distributor].

Gray, H. M., LaPlante, D. A. and Shaffer, H. J. (2012). Behavioral characteristics of Internet gamblers who trigger corporate responsible gambling interventions., *Psychology of Addictive Behaviors* **26**(3): 527–535.

Hing, N., Russell, A., Gainsbury, S. and Nuske, E. (2015). The public stigma of problem gambling: its nature and relative intensity compared to other health conditions, *Journal of Gambling Studies* **32**(3): 847–864.

Howe, P. D. L. and et al. (2019). Predictors of gambling and problem gambling in victoria, australia, *PLOS ONE* **14**(1): e0209277.
URL: *https://doi.org/10.1371/journal.pone.0209277*

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B. and Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Information Sciences* **622**: 178–210.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025522014633*

Jim Orford, Heather Wardle, M. G. K. S. and Erens, B. (2010). Pgsi and dsm-iv in the 2007 british gambling prevalence survey: reliability, item response, factor structure and inter-scale agreement, *International Gambling Studies* **10**(1): 31–44.
URL: *https://doi.org/10.1080/14459790903567132*

Lajcinová, B., Gall, M. and Pitoˇnák, M. (2021). Anomaly detection in time series data: Gambling prevention using deep learning.

Lee, P. (2009). An analysis of gambling expenditure across countries, *UNLV Gambling Research Review Journal* **13**(1).

Lim, B. and Zohren, S. (2021). Time-series forecasting with deep learning: a survey, *Phil. Trans. R. Soc. A* **379**(2200): 20200209.

Murch, W. S., Kairouz, S., Dauphinais, S., Picard, E., Costes, J.-M. and French, M. (2023). Using machine learning to retrospectively predict self-reported gambling problems in quebec, *Addiction* **118**(8): 1569–1578.

Mushtaq, R. (2011). Augmented dickey fuller test, *SSRN Electronic Journal* .

Percy, C., França, M., Dragičević, S. and d'Avila Garcez, A. (2016). Predicting online gambling self-exclusion: an analysis of the performance of supervised machine learning models, *International Gambling Studies* **16**(2): 193–210.

Peres, F. A., Fallacara, E., Manzoni, L., Castelli, M., Popovič, A., Rodrigues, M. and Estevens, P. (2021). Time series clustering of online gambling activities for addicted users' detection, *Applied sciences* **11**(5): 2397.

Preeti, Bala, R. and Singh, R. P. (2019). Financial and non-stationary time series forecasting using lstm recurrent neural network for short and long horizon, *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–7.

Puranik, P., Taghva, K. and Ghaharian, K. (2023). Descriptive analysis of gambling data for data mining of behavioral patterns, *International Conference on Interactive Collaborative Robotics*, Springer Nature Switzerland, Cham, pp. 40–51.

Schreier, A. J. (2022). Relapse: Substance use vs. gambling — wisconsin department of health ...

Seo, W., Kim, N., Lee, S. K. and Park, S. M. (2020). Machine learning-based analysis of adolescent gambling factors, *Journal of Behavioral Addictions* **9**(3): 734–743.

Shahapure, K. R. and Nicholas, C. (2020). Cluster quality analysis using silhouette score, *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748.

Suzuki, H., Nakamura, R., Inagaki, A., Watanabe, I. and Takagi, T. (2019). Early detection of problem gambling based on behavioral changes using shapelets, *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 367–372.

Walker, D. M., Litvin, S. W., Sobel, R. S. and St-Pierre, R. A. (2015). Setting win limits: An alternative approach to "responsible gambling"?, *Journal of Gambling Studies* **31**: 965–986.

Yorucu, V. (2003). The analysis of forecasting performance by using time series data for two mediterranean islands, *Review of Social, Economic and Business Studies* **2**.