UPPSALA
UNIVERSITET

IT 22 014
Examensarbete 30 hp
Mars 2022

Predicting future problem gamblers using Machine Learning Algorithms

Priyadarshini Selvaraju

Institutionen för informationsteknologi
Department of Information Technology

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Predicting future problem gamblers using Machine Learning Algorithms

Priyadarshini Selvaraju

The work in this thesis attempts to identify potential gambling addicts inan online gambling website using machine learning models. Machine Learning can play a major role in predicting and identifying high-risk online players leading to self-exclusion and providing automated self-help tools to problem gamblers. It helps to predict problem gamblers based on past usage history

Machine learning aids in the training of data from users who are problem gamblers by definition. This was accomplished with the help of supervisedlearning, specifically using a classification algorithm such as Support Vector Machine and Naive Bayes. The player tracking system then creates a prediction for all active users based on their behavioral patterns. The final results would include using existing player behavior data to add predictions to improve the model and make it self-learning to detect potential gamblers.

The system then makes a prediction for all active users based on theirrecent usage history. The final result includes a system for analyzing thepotential gambling addicts compare to those non-addicts and gambling ofpotential problem gamblers who show gambling signs of gambling addiction in order to predict those in future gambling sites.

Handledare: Lijo, George
Ämnesgranskare: Andreas Hellander
Examinator: Mats Daniels
IT 22 014
Tryckt av: Reprocentralen ITC

Acknowledgements

First and foremost, I would like to thank God almighty for giving me the strength, knowledge, ability to undertake this research study and to persevere and complete it satisfactorily. I would like to thank my thesis supervisor George Lijo and the organization Kairos Logic for giving me the opportunity and for continuously steering me in the right direction during the course of this thesis. It has been an utmost pleasure working under them.

I would like to thank Andreas Hellander, my reviewer for allowing me to work on this thesis topic and proving support and regular feedbacks in the work done for this thesis.

Lastly, I must express my profound gratitude to my parents and the rest of my family and my friends in India (Chennai) for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank You

Table of Contents

LIST OF FIGURES

1. Introduction

Sustainable Interaction's (**SI**) core business is to help fighting addictions, transforming knowledge about psychology and pedagogy into innovative products for change and progress. One of the key fields is responsible gambling, which includes services and digital products such as an online training programme, a web-based self-help programme, and diagnostic tools to assist people stop gambling and improve their mental health. Gambling is very addictive in part because of today's ease of access, which allows players to wager varying amounts of money, including the entire investment. Therefore it is quite difficult to detect early problem gamblers today. With these in mind it is very important to find out problem gamblers and potential problem gamblers on the betting site so that they can be informed about the addiction and given help if needed. Gambling addiction or more formally called problem gambling or pathological gambling is an old impulse control disorder with many side effects.

Online gambling is diagnosed similarly as regular gambling problem at the casinos but it has its own problems, which are not present or can be easily diagnosed in t regular gambling [1]. Gam Test is a one-of-a-kind solution that helps both players and gambling operators to learn more about how a player is performing. The ultimate result is a set of concise, individualized recommendations that the player can use to make long-term decisions about their game.

Artificial Intelligence and Machine Learning are two types of decision-making algorithms available in today's technology. There is a possibility with machine learning to help detect problem gamblers. The data contains the input and output items in pairs [2] and supervised learning is employed. When supervised learning labels input data into two distinct classes, the classification model includes grouping data into classes. The random forest model is used in improving the model to be self-learning. It operates by constructing a multitude of decision trees and outputs a classification of the individual trees.

1.1 Main Contribution and Research Questions

This thesis attempts to answer the question of whether predictive analysis machine learning approaches can be used to identify probable gambling addicts on a high-traffic gambling site. The work done as part of this thesis includes tracking user data, building a data pipeline for storing and aggregating user data. Machine learning methods were later used to classify users as addicts and non addicts based on their usage history on the site over time. **SI**s Player tracking system (PTS) consists of two parts: The PTS gambling behavior model and the self-test Gam Test. Every individual player's gambling activity is tracked for the previous two weeks, and this two-week period is compared to the two-week period prior to measure change. From behavioral data only, it is not possible with 100 % accuracy tell if someone has a gambling problem. Thus, a crucial part of the PTS is a self-test (Gam Test) measuring over consumption and negative consequences from gambling.

Setting evaluation criteria for such an issue is problematic since, once the algorithm has classified potential gambling addicts, we have no means of knowing whether the users will become addicted in the future, which would be useful in designing a self-help tool to monitor. We can check that if predicted users turn out to be addicts but the usage can drop down or go up based on certain real life conditions.

1.2 Problem Statement

The data contains the five dimension of the self- test, over –consumption of time and money, as well as consequences for time, money and emotions. Break down of the data into self-test parts and making it dynamic based on the player tracking system and introduce motivational interventions and feedback based on risk profile and addiction probability. By training the model to predict alternative outcomes of the self-test, the analysis and classification of the player profile into risk categories will be aided (E.g. Low, medium and high risks).

1.3 Background Theory

The goal of this project is to look at data from its Player Tracking System and customers to see what gambling behaviors can forecast future self-exclusion and profile problem gamblers. Create a long-term AI tool that can assist players in changing their gambling habits.. This can be achieved with combination of SI's Self-Test data from players and player behaviors across five dimensions of problematic gambling (i.e., overconsumption of money and time, and monetary, social and

emotional negative con- sequences).

Availability of both behavior data segmented into core behaviors and a good sample of self-test data makes it good start for SIs AI journey. The outcome adds to the organization's growth, competency development, strategy, and mission of assisting society in the battle against addictions, providing psychosocial health, and upholding social responsibility. The dimensions worked on are the five dimensions of the self-test, over-consumption of time and money, as well as consequences for time, money and emotions.

Machine learning helps to make an intermediary model sub classify the self-test and player behaviors into these dimensions. We treat finding problem gamblers as a deterministic task. People who play more than a threshold amount or spend more than a threshold time can be considered problem gamblers on the site. With the meaning of problem gamblers already defined it is easy to find those from the usage data. The bigger problem is the early identification of users who might become problem gamblers later on or in other words those users who have not passed the threshold yet but still have an overall profile of gamblers.

2. LITERATURE REVIEW

Gambling is all about emotions; harm from gambling isn't about only losing money. Gambling can affect self-esteem, relationships, physical and mental health, work performance and social life. It can harm not only the person who gambles but also family, friends and workplace. Addiction typically involves initial exposure to a stimulus followed by behaviors seeking to repeat the experience. After a number of repetitions of the behavior-stimulus sequence, the addiction becomes established. The character and severity of the addiction may change over time, and it may be punctuated by attempts by the sufferer to abstain or regain control in some cases, sufferers will achieve recovery for a sustained period or even permanently.

The prevalence of gambling addiction is 2-3%. The essential feature of gambling addiction is recurrent and persistent gambling behavior. It results in disruptive personal, family or work life. Daria J et al argued that online gambling addicts have the same kind of characteristics as normal addicts [13]. The prevalence of gambling addiction has increased a great deal over the time period. Aviel Goodman et al. defined addiction as a behavior that produces pleasure and a relief from internal discomfort. It usually comes with a pattern that is characterized by constant failure to control the behavior and continuation of the behavior despite experiencing negative effects.

Problem gambling is an urge to continuously gamble despite harmful negative consequences, the prevalence of problem gambling has been evaluated at 7.8% among university students which is

considerably high than the roughly 5% rate found among the general population (Blinn, Pike,Worthy, Jonkman, 2006). Students facing problem gambling illustrate many signs including isolating behavior, lowered academic performance, poor impulse control and displaying extreme overconfidence, and participating in other high risk.

3. GAM-TEST

Gam-Test is an online test of gambling behavior which provides information that can be used to give players individualized feedback and recommendations for action. Gam-Test consists of five dimensions which all affects the communication within the SI PTS system. The dimensions are overconsumption of money, time, and monetary, social and emotional negative consequences. It is based on 15 self-reported statements. The aim is to find early signs of gambling problems such as over consumption of time and money.

OC time	GT2	Sometimes I forget the time when I'm gambling
	GT1	Sometimes I gamble for longer than I intend
	GT4	I devote time to my gambling when I really should be doing something else
OC money	GT5	Sometimes I gamble more money than I intend
	GT6	I sometimes try to gamble back money that I have lost
NC money	GT8	I sometimes borrow money to enable me to gamble
	GT7	I sometimes [illegible] with money that really [illegible] been used for something else
	GT12	Sometimes my gambling has left me short of money
NC social	GT10	People close to me think that I gamble too much
	GT3	Other people say that I spend too much time gambling
NC emotions	GT14	Sometimes I feel bad when I think about my gambling.
	GT15	My gambling sometimes makes me irritated
	GT11	Sometimes I feel bad when I think of how much I have lost gambling
	GT13	I feel restless if I do not have the opportunity to gamble
	GT9	I do not want to tell other people about how much time and money I spend on my gambling

Figure 1. Self-assessment test (Gam-test) showing different aspects on dimension money, emotions, time and social

The tracked aggregated and filtered data has 1263 records of monthly user activity and an independent variable event_grouped defining if the user was addicted in that month. Fig 7.1 shows the list of features used for classification task, data types and the number of records.

```
In [21]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1263 entries, 0 to 1262
Data columns (total 26 columns):
amount_won_avg                      1263 non-null float64
session_start__$date                1263 non-null object
amount_bet                          1263 non-null float64
amount_won                          1263 non-null float64
session_end__$date                  1263 non-null object
amount_netto                        1263 non-null float64
amount_won_avg                      1263 non-null float64
amount_netto                        1263 non-null float64
amount_bet                          1263 non-null float64
amount_won                          1263 non-null float64
session_start__$date                1263 non-null object
session_end__$date                  1263 non-null object
session_avg__amount_bet_avg         1263 non-null float64
session_avg__amount_won_avg         1263 non-null float64
session_avg__amount_netto_avg       1263 non-null float64
type                                1263 non-null object
amount                              1263 non-null float64
type                                1263 non-null object
amount                              1263 non-null float64
date_of_aggregation__$date          1263 non-null object
modelVersion                        1263 non-null float64
variables                           1263 non-null object
created_at__$date                   1263 non-null object
result__|                           1263 non-null object
result__|__score                    1263 non-null float64
result__|__group                    1263 non-null float64
dtypes: float64(16), object(10)
memory usage: 256.6+ KB
```

Figure 1. List of features. Data type and number of records used while feature scaling

4. METHOD

The work in this thesis is done through the following steps as illustrated in the figure 1.

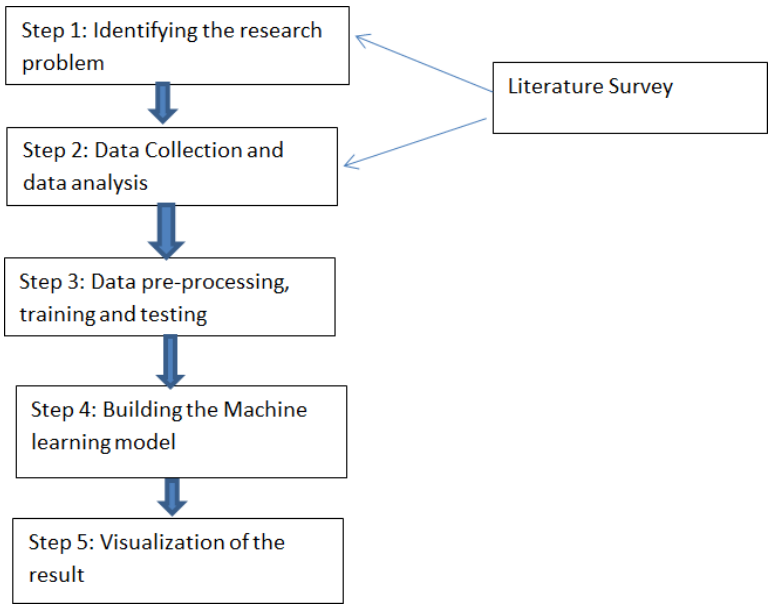

Figure 2. Flowchart representing the process flow of the data while analyzing and the work done in sequential format

The scope and objectives of the research is identified and defined. The Second step will be the necessary data extracted and collected, meanwhile researching on various literature articles to get

some knowledge on how to carry the whole process, on how to build the model and the techniques used in the project is studied.

In the third setup, data pre-processing and analyzing the data is done with tools in order to prepare the data set for further process of how to handle it. For a high-traffic site, tracking, cleaning, analyzing, and storing data must be both efficient and scalable. A custom log analyzer and an indexed relational database were used to overcome the scalability and efficiency issues. The algorithms Naive Bayes and Support Vector Machines (SVM) algorithms for prediction but found linear SVM to have the highest accuracy, precision and recall from the usage data. The classifier is used to predict all the users based on their usage history for last month. The data for model creation is collected and summarized monthly for model creation and the predictor predicts every day. A brief explanation of the work completed, results from data analysis and predictive analytic system, and recommendations for future work to solve this or similar problems are included in the result section.

5. TOOLS USED

5.1 Python pandas

A Python panda is a powerful library for data processing. It can take (Comma Separated Values) CSV files as input and convert them into a data frame, a table-like data type for easy understanding and easy handling of data. Python is known to be flexible and user-friendly.

Due to its ease of use and its extensive documentation regarding machine learning, python was chosen. This library is generally used for data cleaning, data engineering and JSON (Java Script Object Notation) to CSV conversion required for this project.

5.2 Git Version Control

Git is a free and open source distributed version control system which handles everything from small to very large projects [3]. A version control system is software which helps a software team to manage the source code over time. It tracks the commits, in case some code breaks in the future; it can always be tracked back. Cloned the repositories from the GitHub repository for the customer data Player tracking system to a local folder.

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. Building the

container is done by navigating to the local folder for the container and running. Container start-up sequences:

- Transform container (aggregates incoming data transaction into the required data format used by the analysis container, runs a task every 15 minutes)
- Analysis container runs and save the data analysis of gaming transactions, runs a task every day at 6:00pm.

6. MONGO DB

Mongo DB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. Mongo DB works on concept of collection and document. Database is a physical container for collections. Each database gets its own set of files on the file system [4]. A single Mongo DB server typically has multiple databases. Collection is a group of Mongo DB documents. It is the equivalent of an RDBMS table.

7. NATURE OF THEDATA

This is an empirical study in the sense that is based on real-world data. As consequences, any conclusions drawn solely concern the performance of the statistical methods employed. The dataset is anonymous and it contains the player behavior tracking on the basis of weekly data. The original data set provided consists of several rows and columns. The data is generally in JSON (JavaScript Object Notation Format) files which contains one JSON document per player per variable, so 10x10 matrix containing 100 cells (10 players and 10 variables).

The three main general category on gambling were Horse, Casino and Sport betting in which the Variables contains the player's behavior which is based on the dimensions; Speed, Change, Money and Time. It has every individual's activities on gambling basically the tracking of all activities during the betting. The result of data aggregation is placed into the mongo collection.

Data is exported from mongo database in JSON format. The analysis container uses two schedules for running transform and analysis. Each container will perform a task based on where the data is generated. Data aggregation generates a matrix of data, where each row contains values for a single player, and each column contains one variable [5].

7.1 Data Collection

Data collection and data creation is the initial step when you want to build the statistical model using machine learning and AI. The data set is commonly divided into three subsets: a training set, a validation set and a test set. The training set is used to train the statistical model, the validation set is used to determine how well the model has been trained, and the test set is used to assess the model's performance. Data is stored in databases. Mongo DB exports the data from the data hub: It exports PTS data and information from Mongo DB to the API's web front end (Application Programming Interface, which is a software intermediary that allows two applications to talk to each other).PTS results, self-assessment tests, tracking web traffic, and error management are all handled by the container.

7.2 Features and labeling

Featuring is a way an object is presented in machine learning and pattern recognition. Feature vectors are n-dimensional vectors where each vector represents an object. A numeric representation of the features will amplify statistical analysis. The features and labels among supervised learning, the 18 features are descriptive attributes and the label is what the prediction or forecast. The output which gives after training the model will be label.

8. IMPLEMENTATION

8.1 Data pre-processing

The performance of supervised learning algorithms can greatly be affected by data pre-processing and proper data pre-processing can lead to improved generalization. Pre-processing may include feature selection, normalization of numerical features and encoding of categorical features. The first step was cleaning (identifying and correcting errors in the dataset that may negatively impact a predictive model). The player tracking system data which is the customer data is exported from mongo database in JSON format as multiple files.

The data was initially noisy thus cleaning was required. The project cleaning and converting JSON to CSV was completed using python programs. Python was chosen because it has features that make JSON to CSV conversion very simple and efficient, as well as the ‘pandas’ which makes

cleaning a breeze.The relevant features and labels were imported into the data frame, and the features were used to train and test the system. The pre-processed dataset will include the same number of samples as theoriginal dataset because the goal of the study was to forecast whether the player will result in a danger level or not. The PTS- model is built upon a number of variables; each standardized scored 0-10, evolving from gambling data from the operator. Each dimension has its own variables for thevariables money, time, change, and speed, which include summed and cut-off scores that determine the risk level. The data set was divided into two parts: training and testing. The split was done in a tiered manner, which preserved much of the previous player behavior. As a test dataset, 20% of the samples' pre-processed dataset was set aside.

As the dataset is imbalanced a resampling algorithm SMOTE (SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the over fitting problem posed by random oversampling), is implemented in order to try to improve the performance of the predictor. This implemented of SMOTE is intended for one-dimensional samples, firstly it was reshaped into one-dimensional vectors when applied. Each sample was then reshaped back to original sample shape. The eighty per cent of the dataset reserved for training was further split into training and a validation set. Twenty per cent of this subset was reserved and the validation set are used for tuning the hyper parameters.

8.2 Responsible Gambling section

The player tracking system has responsible gambling activities where the users can limit the interactions on the system. The ways the user can limit their usage on the system are:

- Self-Exclusion: Removes the user permanently from the system. A hard indicator that the player is worried about his usage in the PTS.
- Time out: Removes the player from the system for specific period of time. Hard indicator if it is more than a couple of months.
- Deposit Limit: Setting a deposit limit will not let users deposit more money than they have selected. It is also an indicator because it can be for precautionary reasons.

8.3 Log Analysis and Aggregating Data

The customer data is tracked and stored in files, and the data is being processed and analyzed, it is aggregated using pandas and Numpy libraries. A python log analyzer which runs everyday data and

monthly data to analyses the data pattern.

8.4 Tables for Daily Analysis

- Acc_user: User basic details (login ID, gender, language, DOB, currency and country etc.)
- Acc_user_action: There are several actions a player can perform on the player tracking system (e.g.: setting deposit limit, logging in, system usage and reality check)
- Acc_user_deposit: Each record is the history of deposit and withdrawal for each player as a transaction.

8.5 Tables for Monthly Analysis

As mentioned in previous section, a python script which runs every month, aggregates the data for previous month and store it in the separate monthly tables, some of them are:

- Acc_montly_action: Count and aggregation of each player's action for previous month.
- Acc_monthly_bet: Aggregated summary of each player's previous months betting. Columns include how much money was stake, sum of win/loss, count of bets etc.
- Acc_monthly_timespent: Time user spent on the system in the previous month.
- Acc_monthly_deposit: Aggregation of withdrawal and deposit of previous month with deposit and withdrawal frequencies.

9. MACHINE LEARNING

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience T." Thomas M. Mitchell, 2007. In laymen's terms Machine learning is the ability of a computer to learn from experiences (data) and generalize the solution over all kinds of datasets. Machine learning applications are useful for non-deterministic problems where the problems are either too complex or too big to be broken down in to step by step instructions [6]. Some machine learning applications are spam detection, predicting stock values, finding patterns or hidden structures in a structured data set etc. The method of learning from previous examples or experiences is a sub field of Artificial Intelligence called supervised learning. The two main applications of supervised machine learning are classification and regression. Classification is used to predict which class a record belongs to whereas regression is used to predict a real numerical

value. Machine learning can also be used to reveal a hidden class structure in unstructured data, or it can be used to find dependencies in a structured data to make predictions.

10. PREDICTIVE ANALYSIS

Predictive analytics is the act of predicting future events and behaviors present in previously unseen data, using a model built from similar past data. It has applications in wide variety of domains such as finance, healthcare, academics and gambling industry etc. The application and methods of predictive analysis in all these fields is similar. A machine learning algorithm finds the relation between different properties of the data to build a model. The resulting model is able to predict one of the properties of future unseen data. Creating a prediction model from previously known dataset is called training. The data used for training a model is called training data or training set.

Once the model is created it is tested against a dataset which is not part of training set to test the efficiency and effectiveness of the model. The data used for testing the model is called test data or test set. The model is then fine-tuned and reiterated multiple times to find the best accuracy score for the test set. Once the model is efficient enough for test set, it can be said that it is generalized for any unseen data and can be used on production [7]. The reason for splitting the data in two different sets (training and testing) is to avoid over fitting. If we use all the data for training then the model will be highly efficient for training set but will perform poorly on unseen data.

We have used predictive analysis in learn the relationship of different properties of a betting addict, and then used the model to predict for all users if they are potential betting addicts or not. The implementation and system overview part tells more about the methodology and implementation.

11. Classification and Selected Methods

There are hundreds of different classification algorithms which learn from a dataset and predict whether a record belongs to a class or not. Each algorithm has the same task i.e. predicting a dependent variable. They are based on different mathematical methods with their own weaknesses and strengths. We have used two different algorithms, Support vector machines with linear kernel and Naive Bayes classifier for the use case and found SVM to be performing better.

12. SUPPORT VECTOR MACHINE

In machine learning, support vector machine (SVMs, also known as support-vector networks) are supervised learning models associated with learning algorithms that analyze data for classification and regression analysis. SVMs have been used in many pattern recognition and regression estimation problems and have been applied to the problems of dependency estimation, forecasting and constructing intelligent machines [8]. In Multi-Layer Perceptron (MLP) classifiers, the weights are updated during the training phase for which the total sum of errors among the network outputs and desired outputs is maximized.

The performance of the network strongly degrades for small data sizes, as the decision boundaries between classes acquired by training are indirect to resolute and the generalization ability is dependent on the training approach. In contrast to this, in SVM the decision boundaries are directly determined from the training data set for which the separating margins of the boundaries can be maximized in feature space. A SVM is a maximum fringe hyper plane that lies in some space and classifies the data separated by non-linear boundaries which can be constructed by locating a set of hyper planes that separate two or more classes of data points.

After construction of the hyper planes, the SVM discovers the boundaries between the input classes and the input elements defining the boundaries (support vectors). From a set of given training samples labeled either positive or negative, a maximum margin hyper plane splits the positive or negative training sample, as a result the distance between the margin and the hyper plane is maximized. If there exist no hyper planes that can split the positive or negative samples, a SVM selects a hyper plane that splits the sample as austerely as possible, while still maximizing the distance to the nearest austerely split examples.

Mathematically it can be defined as the following:

We are given a training set of n data points.

$$(x_1/y_1),\ldots\ldots\ldots\ldots(x_n y_n)$$

Where y_i is either 1 or -1 indicating the class in which the data element x_i belongs, SVM tries to find the 'maximum margin hyper plane' that divides the group of points for which $y_i = 1$ from the group of x_i for which $y_i = -1$ so that the distance between the hyper plane and the nearest point from either group is maximized. A hyper plane can be written as the set of point's x satisfying

Where w is the normal vector to the hyper plane

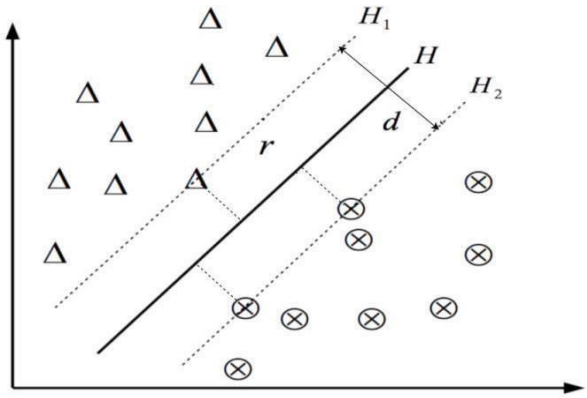

Figure 3. Representation of linearly separable samples of two classes indicated with a hyper plane

Figure 2.1 depicts a linearly separable hyper plane in which two groups of data points are shown. by '⊗' and 'Δ'. There may be possibility of an infinite no. of hyper planes but in the described figure, only one hyper plane represented by solid line optimally separates the sample points and is situated in between the maximal margins.

13. Naive Bayes Classifier

Naive Bayes classification is a machine learning algorithm which relies on Bayes' Theorem:

$$P(c/x)=\frac{P(\frac{X}{C})P(C)}{P(X)}$$

Where A and B are two different events, P (A) and P (B) are the probability of A and B occurring, respectively. P (A—B) is the probability of A occurring given that B has already occurred [9]. The Naive Bayes Classifier makes strong assumptions about how the data is generated, and posit a probabilistic model that embodies these assumptions; then they use a collection of labeled training examples to estimate the parameters of the generative model. Classification on new examples is performed with Bayes' rule by selecting the class that is most likely to have generated the example. It assumes that all attributes of the examples are independent of each other given the context of the class. This is the so-called "naive Bayes assumption."

14. FEATURE ENGINEERING

In machine learning, feature engineering is the process of selecting and creating features

(Variables) in a data set to improve machine learning results [10]. Feature engineering when done correctly results in improved model accuracy on unseen data, the common methods of feature Engineering are:

Feature selection

- Feature selection is the process of selecting important features and removing redundant or useless features which have no correlation to the dependent variable.
- Creating a model to test the correlation of the variable with the dependent variable.

Feature Creation

- Feature creation includes modifying the variables and creating new ones by combining multiple different variables.
- A mixture of aggregating or combining features to create new features, and decomposing or splitting features to create new features.

Feature Extraction

- Feature extraction is the process of automatically reducing the number of features (dimensions) of a data set which can be modeled.
- Some data sets are having too many features which results in a very complex model if all the features are used as it is.

Feature Scaling.

- If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance

- The range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization

14.1 Imbalanced classes in predictive analysis

A dataset with majority of data belonging to one class, modeling a predictive classifier will always be sensitive towards majority class. If this issue is not taken care of, the classifier will be biased

and will predict the majority class in most cases. There are various methods to tackle this problem; we have used oversampling (**techniques used to adjust the class distribution of a data set**) for this thesis.

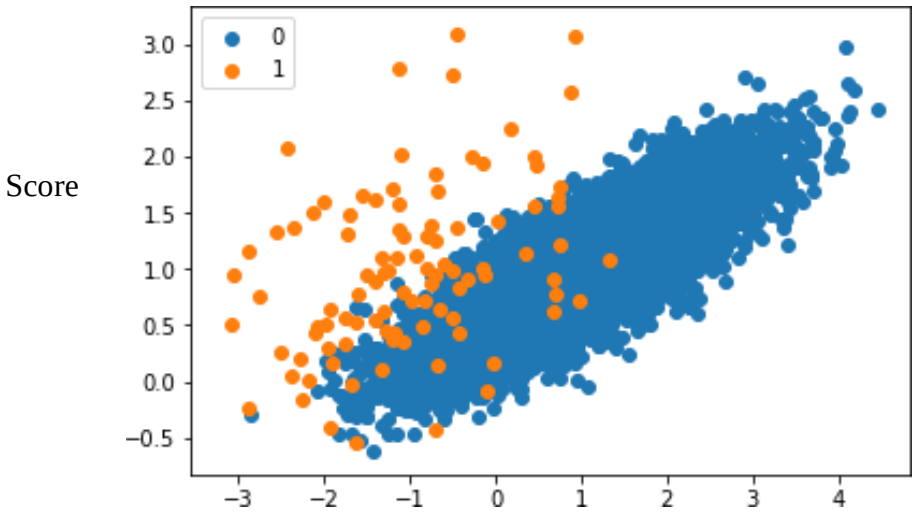

Figure 4. Scatter plot by class labels comparing addicted and non-addicted players (class distribution of the data)

A scatter matrix plots all the numeric variables in a dataset against each other. In python the data visualization technique can be carried out with many libraries but while using Pandas to load the data we use scatter matrix to visualize the dataset.

14.2 SYSTEM OVERVIEW AND METHODOLOGY

This chapter is a brief overview of the overall system and how different modules are communicating.

14.3 Data and Domain Information

If a user satisfies any of the following criteria, they are considered a gambling addict which will be helpful in developing a self-help tool:

- Users who spend more than 10k GBP per month on the gambling site.
- Users who spend more than 8 hours on average on the gambling site.
- Users who have self-excluded themselves on the website.
- Users who deposit money more often are more likely to be gambling addicts.

- Users who play more than 1000 bets per month are gambling addicts.
- Users who log in more than 300 times per month are gambling addicts.

Name	Explanation
Basic Player detail	Currency, Country, login ID, age
Deposit/Withdrawal History	Each deposit/withdrawal the player made during the specific time
Gambling History	The detailed gambling history of every individual
Responsible Gambling	Restrictions (Reality check, self-exclusion, time out)
Usage Summary History	How much money user is spending per each game

Table 1: Data Tracking- tracked for the last two weeks for every individual player

14.4 DATA INFORMATION

Mongo DB utilized to capture raw data, which was then used as features in the machine learning application. The proxy server aided in the ftp access of the files (File Transfer Protocol).

The data was collected on the proxy server and then pre-processed (used to transform the raw data in a useful and efficient format).

- The data had a separate sports section. It was not possible for me to track any sports section specific betting data. The data collected is from all the rest of the sections of the website.
- User's device (mobile or desktop site), IP, number of times logged in/out, timestamp etc. is also tracked for each user.

Python log analyzer (which analyses the CSV files which will store it in the database to use simple code whenever we need it) runs every day and reads the previous day log, cleans the data and stores it. The relational database has a schema tailored for collecting data to avoid duplication of records. There is another python based script that runs every month and stores the aggregated monthly usage data in the respective database tables.

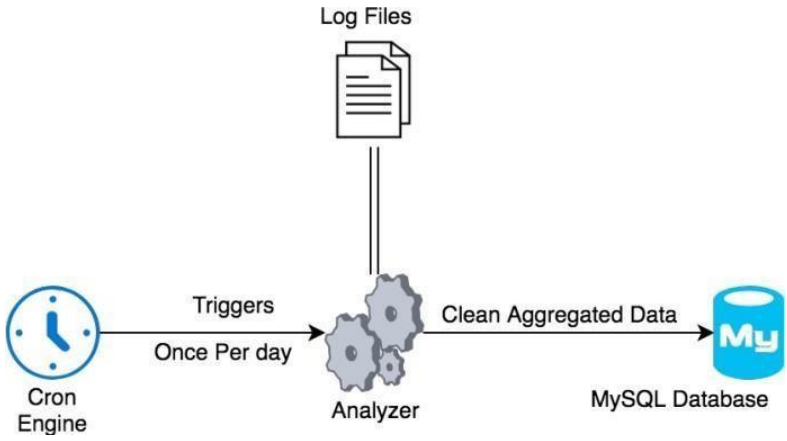

Figure 5. Cron Engine pre-processing data and storing in the Database where every day analysis is done

Data Normalization

The data set features are rescaled so that they have the properties of a standard normal distribution with 0 mean (μ) and 1 standard deviation (σ). Data normalization resulted in features which had different unit but were all in similar scale. Pearson's Correlation method is used to find the correlation between the features. The yellow box in figure 5 shows that the two features are highly correlated, while the black section shows no correlation. It can be concluded that from figure 5 used to solve the machine learning problem are not correlated and hence can all be used together [11].

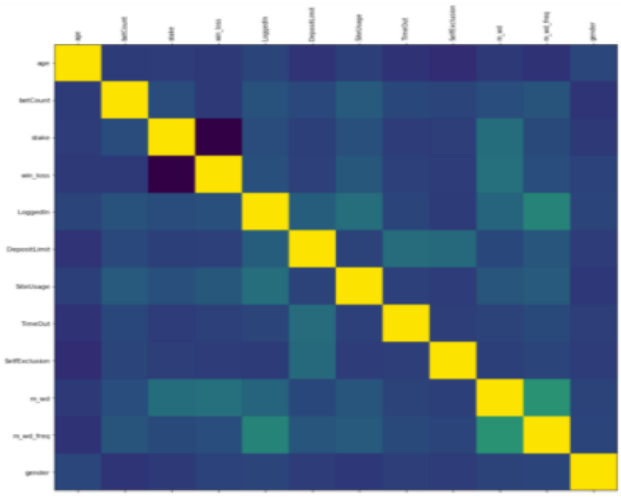

Figure 6. Pearson's Correlation between two variables having 0.5 which is moderately correlated

15. MACHINE LEARNING MODELLING

The data of gambling addicts and non-addict users were used to create a machine learning model

of betting addiction. Firstly scaling the features, oversampled the minority class, used grid search to find the best hyper parameters for the algorithm and then trained the model for potential betting addicts. Once the model is trained, it can predict users based on their behavior history for potential betting addiction, and then used all the non-betting addicts to predict if they are potential betting addicts based on their last 30 days usage history. This whole process is automated and repeated every day through Python scripts (A Python script **a file containing code written in Python)** [12].

The file containing python script has the extension '.py '). The supervised learning algorithms used in this thesis generalize the usage data to form a machine learning model for prediction. For this reason we need to track user data. Due to daily traffic reasons, ended up setting a custom tracking module using Frosmo solutions.

Frosmo solutions (Frosmo Script is used for data tracking module which will help to build into the model). The data tracking module then was able to subscribe to user actions on the page and send the data to a dedicated server.

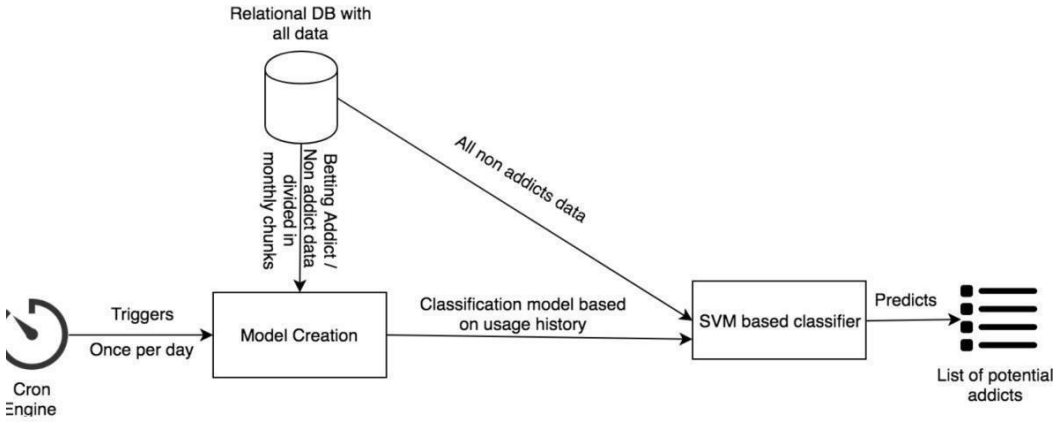

Figure 7. Cron Engine predicting classification of potential betting, the process of model creation and prediction which stores in Database

Basic User Details

We need basic user details like currency, country, age and login ID for analyzing data set. The methods we get the user's transactions history do not return the currency of the transactions as it can be in GBP, EUR or USD. In order to normalize the amount in one currency unit we need to track the currency unit for each user. Tracking the age of the user (Date of Birth) since it could be one of the factors for user's addiction. Tracking the login ID, we need to identify the users later on

while sending the list of potential betting addicts back to the game. Tracking the country because it it could be helpful in seeing the geographic map of betting addicts.

Deposit/ Withdrawal History

Each transaction record and gaming history data record had a unique identifier as the timestamp of when the transaction was made. Even if the data is sent twice in the case of a user using multiple devices duplicates are removed in the data analysis step.

15.1 Database views

Since the data is stored in the database with their specific tables, we needed to create several database views to connect different tables and get specific data from there for data analysis and machine learning algorithms. Following is the list of some most important database views and their purposes.

- **ml user monthly**: It connects different monthly aggregated tables and creates generalized view for each user for each month.
- **ml addicted monthly**: Augments the ml user monthly view with is Addicted column based on the addicted definition. This serves as the training data for machine learning algorithm.
- **ml addicted**: This is also the last month's summary for each user. It is different from the above view because it shows the history for last 30 days rather than from monthly aggregated data. After the model is trained it classifies the user every day based on their 30 days usage from this view.
- From gambler's point of views, it would reduce damage enabling the system to prevent gamblers from reaching a critical stage.
- Time between bet should be limited since the data on the dimension time has a lot more usage than any other dimension. Time limitation need to be set by the platform which could provide a guide to calculate disposable income.

The dimension money was useful for determining the general risk of the population, additionally having individual player identification codes allow to subset the data to observe and complete gambling history of each individual gambler.

15.2 Data Analysis and Visualization

Analyzing and visualizing the data will help in feature engineering of the actual machine learning implementation

TRENDS VALIDATION

ATTRIBUTE	NOT ADDICTED	ADDICTED
Average usage per day	69 minutes	143 minutes
Average time logged in	15.82	25.39
Average time deposited in a month	0.06	2.08
Average time of withdrawal	0.71	1.70
Average withdrawal amount	191 GBP	633 GBP
Average user age	42.3 years	36.8 years

Some of the trends in above section are pretty self-explanatory. For example on average addicts use twice as much time on the site as compared to non addicts, they have a higher withdrawal frequency and much higher deposit frequency. They are also more likely to be concerned about their gambling problems and have used the responsible gambling section of the site (timeout, self-exclusion and deposit limit). They are likely to spend much more money during the game.

15.3 USER DATA TRACKING

TRACKING MECHANISM

The data tracking mechanism needed to be failsafe and scalable because there was a high traffic in the data. The data is sent from the user's browser to a server through API calls (submitting a request to retrieve the data from external server).

15.4 HIDDEN TRENDS

Some of the hidden patterns during the visualization are bet count, win/loss ratio and the method of time tracking. The bet count is the number of time the user gambles. By the definition of addicted users anyone who plays more than 1000 times per month is considered a betting addict. The first intuition about win/loss ratio was betting addicts who cannot stay away from gambling must be losing a lot of money, the data proves it wrong.

The addicts on average win 50GBP more than non addicts with 90th percentile winning three times as much as non addicts. However, the bet count among addicts are reaching much higher with an average of 1121 and 90th reaching over 2,000+ games count per month.

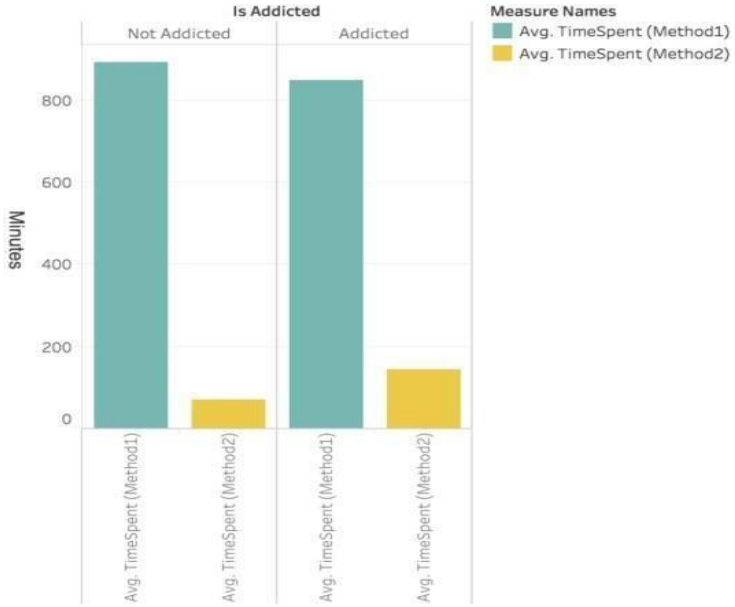

Figure 8. Comparison between Svm and Naive Bayes algorithm, where Svm showing the accurate average of time spent (addicted players)

As discussed in the data tracking section, we have two methods of tracking user time spent during the game. One is to spend time on each game after the user leaves the game and other user is to send a flag of time spent every 15 minutes. Figure shows that the two methods are not reporting relevant similar results. When compared to the second approach, the first method has about the same average value for both addicts and non-addicts and has a significantly higher average. The second technique is more intelligently tracked; after 15 minutes of active use, it merely reports time spent. As a result, we will just look at the second method, which is model generation. [13].

15.5 DATA ANALYSIS

Data visualization and analysis resulted in validation of some already known trends and identifying hidden trends among addicts and non addicts. These findings can be found as:

TREND VALIDATION

- Addicted users spend twice as much as time compared to non-addicts

- Addicted users on an average log 1.6 times more than non-addicts
- Most of them are aware of their problems and they try to control their gambling habits through various responsible gambling means.
- Addicted users deposit money more frequently and larger sums of money on average when compared to non-addicts
- Addicted users on average wins 50GBP more than non-addicts with 90th percentile winning three times as much as non addicts.
- Users with age group between 23-32 are more likely to be addicts than my other group.
- The mean age for gambling addicts is 37 years as compared to 42 for non addicts.

16. AUTOMATION

The whole machine learning modeling and prediction method was needed to be performed on daily basis [15]. The monthly record for each user was used for model creation but the prediction was done on daily basis based on last 30 day's history of each user. A daily Python script triggered the learning and prediction process. After the prediction was made, the results will be stored in a log file.

17. RESULTS

The work done as a part of the thesis resulted in a complete end to end system from tracking the user data to predicting potential betting addicts. Following figure is the architecture of the complete system.

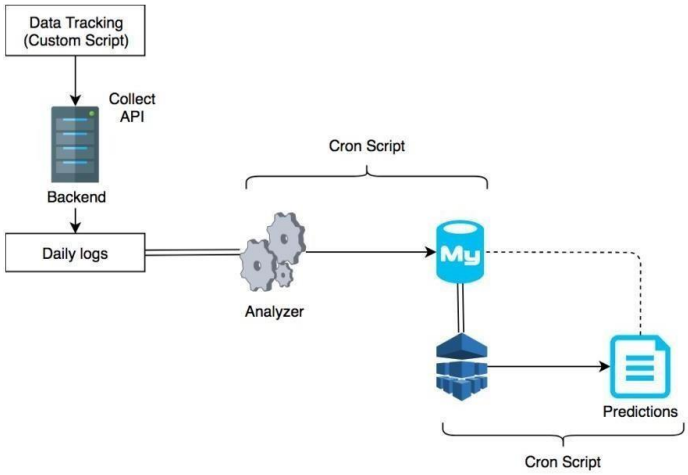

Figure 10. End to End architecture of the database having predictions stored in python script

This chapter presents the key findings from the system through data analysis and machine learning predictions. Furthermore, the recommendations for how the system can be improved further for

identifying gambling addicts with more precision.

18. MACHINE LEARNINGPREDICTIONS

The system predicted that 14.5% of all the users were potential addicts which weren't identified from gambling addiction. The system predicted that most users, who are predicted as gambling addicts are predicted for one month only, this indicates that not all potential gambling addicts end up being gambling addicts. The system predicted that only 4.6% of potential gambling addicts continued to be potential gambling addicts for more than 2 months in continuation. From these results it can be deducted that the system can be used as an indicator to find potential gambling addicts. However, the trend shows that out of those potential gambling addicts only 4.6%of the users continue to show addiction behavior for more than 2 months in continuation.

18.1 OVERSAMPLING

The classification training data had an imbalance class distribution. More than 80% of the users were not addicted and the remaining was addicted [14]. Creating a predictive model from such a data yields a highly biased classifier which predicts the majority class most number of times. For this reason the minority data was oversampled to match the number of majority class elements. Python's imbalanced-learn class was used to oversample the minority class using SMOTE method. It resulted in a dataset with 27955 records for each class.

18.2 The ROC Curve

An ROC curve (receiver operating curve) is a graph showing the performance of a classification model at all classification thresholds. The curve plots two parameters:

- True Positive Rate
- False Positive Rate

An ROC curve plots different classification thresholds, lowering the classification thresholds classifies more items as positive, thus increasing both false positive and trues positive.

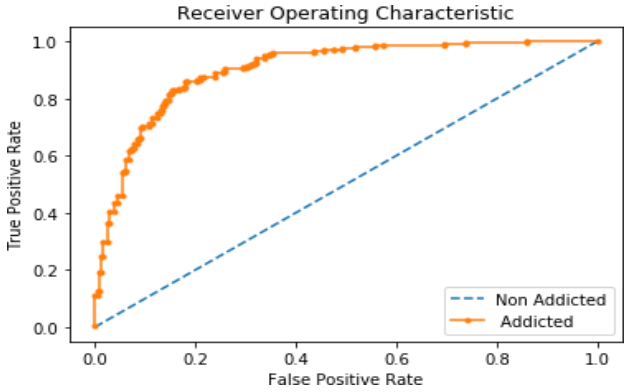

Figure 11. ROC curve showing the accuracy of the performance of classification model generated

19. CONFUSION MATRIX

Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives a holistic view of how well the classification model is performing and what kind of errors it is making.

- The target variable has two values: **Positive** or **Negative**
- The **columns** represent the **actual values** of the target variable
- The **rows** represent the **predicted values** of the target variable

The confusion matrix provides much more granular way to evaluate the results of the classification algorithm than by just giving the accuracy value. It divides the result into two categories which join together with the matrix: the predicted labels and the actual labels of the data points which are the percentile value of each player and the actual value.

Recall- $$\frac{TP}{TP + FN}$$

Precision $$\frac{TP}{TP + FP}$$

F-Measure $$\frac{2 * Recall * Precision}{Recall + Precision}$$

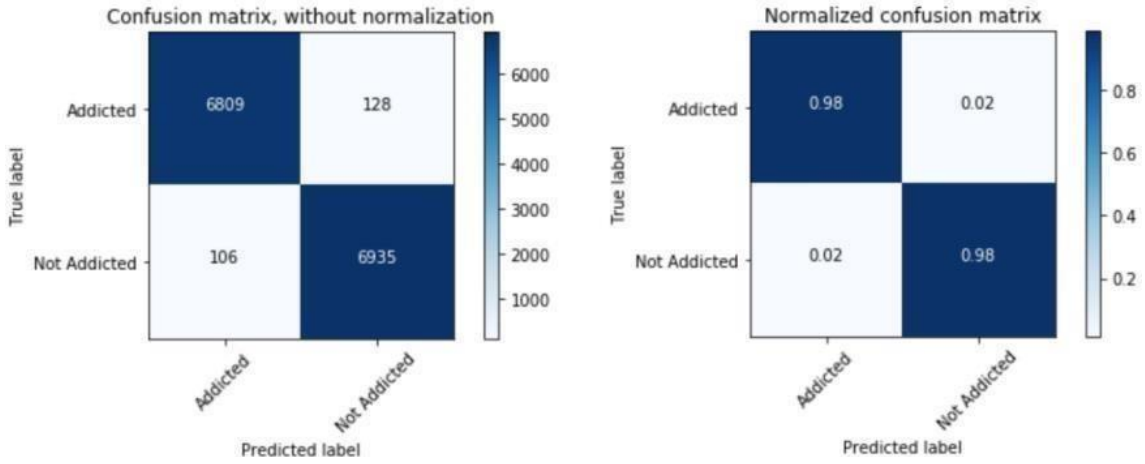

Figure 12. Confusion matrixes predicting the result with summarized count values broken down by each class (correct and incorrect predictions)

	Precision	**recall**	**F1-score**	**support**
0	1.00	0.93	0.96	14
1	0.86	1.00	0.92	6
Triggered	0.95	0.95	0.95	20
Gam-test	0.93	0.96	0.94	20
Result	0.93	0.93	0.93	20

The precision value lies between 0 and, out of all positive predicted at what percentage is truly positive will be a precision. Out of all positive, what percentage is predicted positive Recall should be high when compared with precision. F1-score is the mean of recall and precision, it takes both false positive and false negative into account.

The figure shows the confusion matrix with and without normalization by class support size (number of elements in each class). The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

20. DISCUSSION

This chapter will include why certain decisions were made and what could have been done differently if the thesis were to be conducted again.
When comparing the Machine Learning Algorithms, Support Vector Machine an Naïve Bayes Theorem: SVM works more accurately since there is an hyper plane which divides the two classes and maintains the accuracy of 93% whereas in Naive Bayes Theorem there were imperfection in the algorithm when the data grows, the implementation was quite slow when compared to SVM.

Problem gamblers on average refute or minimize the problem. They also go great lengths to bury their gambling habits. Internet (or online) gambling was the most commonly researched emerging technology/trend. While greater rates of problem gambling were associated with this form of gambling, recent research suggests gambling via the Internet is not inherently problematic but rather appears to affect different gamblers in different ways.

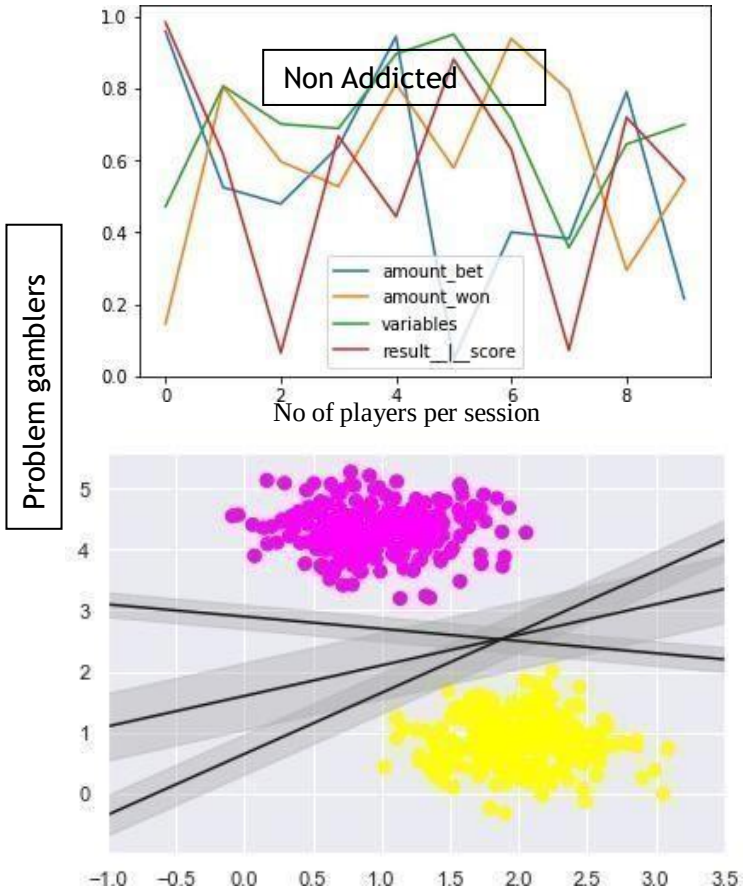

Figure 9. Classes separated by hyper plane using Svm algorithm, result showing the predicted classes

The data set is linearly separable, since the two classes (variable and analysis are classified into single straight line, the data is linear. The region that the closest point defines around the decision boundary is the margin. That is why the decision boundary of a support vector machine model is known as the maximum margin classifier.

20.1 LIMITATION OF THE SYSTEM

The data tracked did not include the sports section of the website because of technical limitations. It resulted in the limitation of finding potential betting addicts who spend a lot of time and money on sports section of the website.

The time tracking mechanism for each user sends a flag every 15 minutes of usage from the site, the time before the next time flag is sent is not tracked resulting in an inaccurate picture of time spent tracking.

20.2 FUTURE RECOMMENDATION

Based on the work done for this thesis, the results can be improved further on choosing better tools and platforms and improvising the process. Recommendations for improving the performance of the feature are as follows

- The system can be built on cloud like Amazon Web Services or Microsoft Azure it can use a lot more powerful off the shelf components for building the data pipeline.
- Using relational database was a major limitation in fetching customized data from the database.
- The data tracked from the server did not include usage on the sports section of the site because of technical limitations. To get a full picture of addiction behavior the data from sports section needs to be included.
- If a user spends average amount of time and money gambling throughout the week but gambles heavily on Friday evenings or on the weekends, it shows an anomaly and a problem gambling behavior. This can be further analyzed for correlation with gambling addiction.
- If a user uses multiple devices and gambles from multiple locations (home, office or on vacations), it can be analyzed for correlation with gambling addiction.

Developing an artificial intelligence Chabot which provides emotional support for those who are

addicted to the gambling world with mood tracking, finding optimism, easy to understand and friendly chats so that it could be useful in controlling humans who are fighting addictions. Chabot's and intelligent search bars are two of the fast-evolving channels within the online self-service portal that constitute the self-service portal. Few recommendations will help to build the system using artificial intelligence which includes:

Virtual customer assistants

- Intelligent search bars
- Optimizing user experience
- Machine learning to interpret a large volume of past requests and correlate themto past solutions.
- Engage and empower users

21. CONCLUSION

The work done in this thesis tries to find potential gambling addicts in order to develop a self- help system based on their usage history. The labeled data was used to train the models to find potential gambling addicts based on their usage statistics. The data was tracked, cleaned, aggregated and analyzed to find patterns in usage behavior. The cleaned and aggregated data was then used for training the machine learning models with two different algorithms and found Support Vector Machine to be more preforming than Naïve Bayes. The whole system was then automated to perform daily leaning as a self-help system to predict automatically through Python scripts. The post prediction evaluation of the model found that only 4.6% of the users turn out to be gambling addicts for more than two months after being identified as potential betting addicts by the system.

Humans, after all, are built on inputs and outputs that may be converted into output patterns that the system can understand. As a result, the behavioral patterns that problem gamblers exhibit, albeit numerous, can be used to identify them. This led to a belief that, if analyzing problem gamblers conducted right, machine learning could be used to create a solution in the area of problem gambling in future and also in identifying the risk. The majority of those who took part in the study, which was expressly targeted at problem gamblers, seemed to think that a system designed to aid them would be acceptable.

2. REFERENCES

[1] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[2] Sally Monaghan. "Responsible gambling strategies for Internet gambling: The theoretical and empirical base of using pop-up messages to encourage self-awareness". In: Computers in Human Behaviour 25.1 (Jan. 2009), pp. 202–207. ISSN: 0747-5632. DOI: 10.1016/J.CHB.2008.08.008. URL: https://www.sciencedirect.com/ science/article/pii/S0747563208001659

[3] Auer, M., and Griffiths, M. D. (2015). The use of personalized behavioural feedback for online gamblers: an empirical study. Front. Psychol. 6:1406. doi: 10.3389/ fpsyg.2015.01406

[4] "What Is Git". In: Tutorials Point. url: https://git-scm.com

[5] Anjali Chauhan, 2019, A Review on Various Aspects of MongoDb Databases, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 05 (May 2019),

[6] Guo Haixiang et al."Learning from class-imbalanced data: Review of methods and applications". In: Expert Systems with Applications 73 (Dec. 2016). doi: 10.1016/j.eswa.2016.12.035

[7] Mitchell, T. M. Machine Learning, 1 ed. McGraw-Hill, Inc., New York, NY, USA, 1997

[8] Bhumika Gupta, Pauri Uttarakhand, and India Aditya Rawat. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. Tech. rep. 8. 2017, pp. 975– 8887. URL: https://pdfs.semanticscholar.org/fd39/e1fa85e5b3fd2b0d000230f6f8bc9dc694ae. Pdf.

[9] Alex Galakatos, Andrew Crotty, and Tim Kraska. Distributed machine learning. 2018.

[10 Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv:1909.06335, 2019.

[11] Feature Engineering Vipin Kumar

https://www.researchgate.net/publication/272799572_Feature_Selection_A_literature_Review

[12] Zhang, L., Kaschek, R. and Kinshuk. (2005): Developing a knowledge management support system for database normalization. Accessed May14, 2008. (http://ieeexplore.ieee.or/apl/freeabs_all.jsp?tp=&arnumber=1508695&isnumber=32...
)

[13] Ambikesh Jayal, Stasha Lauria, Allan Tucker & Stephen Swift (2011) Python for Teaching Introductory Programming: A Quantitative Evaluation, Innovation in Teaching and Learning in Information and Computer Sciences, 10:1, 86-90, DOI: 10.11120/ital.2011.10010086

[14] Wang, Y.L., Tainyi, L.U.O.R., Luarn, P. and Lu, H.P., 2015. Contribution and Trend to Quality Research--a literature review of SERVQUAL model from 1998 to 2013. *Informatica Economica*, *19*(1).

[15] Hong, Juwon, Hyuna Kang, and Taehoon Hong. "Oversampling-based prediction of environmental complaints related to construction projects with imbalanced empirical-data learning." *Renewable and Sustainable Energy Reviews* 134 (2020): 110402.

[16] Martens, D., and F. Provost, “Explaining Data-Driven Document Classification”, MIS Quarterly 38(1), 2014, pp. 73–99..