

University of Louisville

## ThinkIR: The University of Louisville's Institutional Repository

---

Electronic Theses and Dissertations

---

5-2018

### Horse racing prediction using graph-based features.

Mehmet Akif Gulum  
*University of Louisville*

Follow this and additional works at: <https://ir.library.louisville.edu/etd>



Part of the [Computer Engineering Commons](#)

---

#### Recommended Citation

Gulum, Mehmet Akif, "Horse racing prediction using graph-based features." (2018). *Electronic Theses and Dissertations*. Paper 2953.  
<https://doi.org/10.18297/etd/2953>

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact [thinkir@louisville.edu](mailto:thinkir@louisville.edu).

# HORSE RACING PREDICTION USING GRAPH-BASED FEATURES

By

Mehmet Akif Gulum  
B.Eng., Erciyes University, 2014

A Thesis  
Submitted to the Faculty of the  
J. B. School of Engineering at the University of Louisville  
in Partial Fulfillment of the Requirements  
for the Degree of

Master of Science in Computer Science

Department of Computer Engineering and Computer Science  
University of Louisville  
Louisville, Kentucky

May 2018

Copyright 2018 by Mehmet Akif Gulum

All rights reserved



# HORSE RACING PREDICTION USING GRAPH-BASED FEATURES

By

Mehmet Akif Gulum  
B.Eng., Erciyes University, 2014

A Thesis Approved On

April 24<sup>th</sup>, 2018

---

Date

By the following Thesis Committee:

---

Mehmed Kantardzic, Ph.D., Thesis Director

---

Adel Elmaghraby, Ph.D.

---

James E. Lewis, Ph.D.

## ACKNOWLEDGEMENTS

I would like to convey my special regards to my distinctive advisor, Dr. Mehmed Kantardzic for his endless guidance, friendly support and exhilarated encouragements during fulfillment of this work.

I would like to thank Dr. Adel Elmaghraby and Dr. James E. Lewis for accepting to serve in my thesis committee and being a part of this special milestone.

I address, likewise, my thanks to my colleagues, and Computer Engineering and Computer Science Department for their support and friendship.

Finally, I want to convey my most heartfelt thanks to my family for their continuous support and unconditional love.

# ABSTRACT

## HORSE RACING PREDICTION USING GRAPH-BASED FEATURES

Mehmet Akif Gulum

April 24, 2018

This thesis presents an applied horse racing prediction using graph-based features on a set of horse races data. We used artificial neural network and logistic regression models to train then test to prediction without graph-based features and with graph-based features. This thesis can be explained in 4 main parts as follows:

1. Collect data from a horse racing website held from 2015 to 2017
2. Train data to using predictive models and make a prediction
3. Create a global directed graph of horses and extract graph-based features (Core Part)
4. Add graph-based features to basic features and train to using same predictive models and check improvements prediction accuracy.

Two random horses were picked that are in same races from data and tested in systems for prediction. With graph-based features, prediction of accuracy better than without graph-based features. Also We tested this system on 2016 and 2017 Kentucky Derby. Even though we did not predict top three results from 2017 Kentucky Derby, in 2016 Kentucky Derby, we predicted top four position.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
LIST OF FIGURES	vii
CHAPTER	
1 INTRODUCTION . . . . .	1
1.1 Background . . . . .	1
1.2 Motivation . . . . .	2
1.3 Scope and Objectives of the Thesis . . . . .	2
1.4 Organization of the Thesis . . . . .	2
2 LITERATURE REVIEW . . . . .	3
3 DATA PREPROCESSING . . . . .	8
3.1 Data Collection . . . . .	8
3.1.1 Download PDF files from the website with scripts . . . . .	8
3.2 Data Parsing . . . . .	11
3.3 Data Preprocessing: . . . . .	16
4 INTRODUCTION OF GRAPH THEORY . . . . .	19
4.1 Labeled Edge Graph . . . . .	23
4.2 Multiple Labeled Edge Graph . . . . .	23
4.3 Graph-Based Features . . . . .	24
5 MACHINE LEARNING MODELS . . . . .	28
5.1 Artificial Neural Networks . . . . .	28
5.1.1 Logistic Regression Model . . . . .	30
6 EXPERIMENTAL RESULTS . . . . .	33
6.1 Using Basic Horse Racing Features to Make Prediction . . . . .	38



6.2	Using Additional Graph Features to Make Prediction with ANN . . . . .	40
6.3	Using Additional Specific Graph Features to Make Prediction with ANN . . . .	41
6.4	Challenge for Kentucky Derby . . . . .	44
7	CONCLUSIONS AND FUTURE WORK . . . . .	48
7.1	Conclusions . . . . .	48
7.2	Limitations and Future Work . . . . .	49
7.2.1	Limitations . . . . .	49
7.2.2	Other Machine Learning Algorithms . . . . .	49
7.2.3	Additional Features . . . . .	49
	REFERENCES . . . . .	50
	CURRICULUM VITAE . . . . .	52

## LIST OF FIGURES

FIGURE	Page
1.1 Win, Place, Show - How To Bet On Horses. . . . .	1
2.1 The S&C Racing System . . . . .	4
2.2 The AZGreyhound System . . . . .	5
2.3 Results compared Baseline . . . . .	6
3.1 Overall Design . . . . .	8
3.2 Dates of horse Racing . . . . .	9
3.3 Example of one Track . . . . .	9
3.4 Script Code . . . . .	9
3.5 Download the PDF . . . . .	10
3.6 Example of captcha . . . . .	10
3.7 The captcha code . . . . .	10
3.8 Example of Script Results . . . . .	12
3.9 Example of Pdf File . . . . .	13
3.10 Text String . . . . .	14
3.11 Sample table . . . . .	14
3.12 Log File . . . . .	15
3.13 Wager File . . . . .	15
3.14 2017 Result File . . . . .	16
3.15 Distance and Track Type Example . . . . .	16
3.16 Example of Distance Type . . . . .	17
3.17 Track Type in Data . . . . .	18
3.18 Weather Attributes . . . . .	18
3.19 Distance Attributes . . . . .	18
4.1 Example of Graph . . . . .	19
4.2 Undirected Graph Example . . . . .	20
4.3 Directed Graph Example . . . . .	20

4.4	Adjacent Matrix . . . . .	21
4.5	Model of Horse Graph . . . . .	22
4.6	Adjacent Matrix of Horse Graph . . . . .	22
4.7	Example of Directed Edge Weighted Graph . . . . .	23
4.8	Example of Edge Connected Graph With Specific Conditions . . . . .	24
4.9	Example of Edge Connected Graph With Specific Conditions . . . . .	24
4.10	Example of Graph for HITS Algorithm . . . . .	26
4.11	Adjacent Matrix for HITS Algorithm . . . . .	26
4.12	Transpose of Adjacent Matrix for HITS Algorithm . . . . .	26
5.1	Example of Neural Network . . . . .	29
5.2	Example of Neural Network . . . . .	30
5.3	Logistic Function . . . . .	31
5.4	Schematic of Logistic Regression . . . . .	31
6.1	Horse Races Features for Some Specific Days . . . . .	34
6.2	Comparing Horse Races Results for Some Specific Days . . . . .	35
6.3	Example of Horse Race Result in 2017 . . . . .	36
6.4	Features of Two Horses in 2016 and 2017 . . . . .	37
6.5	Comparing Two Horses Results in 2016 and 2017 . . . . .	37
6.6	Compare Prediction Accuracy With ANN and The Baseline . . . . .	39
6.7	Compare Prediction Accuracy With ANN, LR and The Baseline . . . . .	39
6.8	Compare Prediction Accuracy With Additional Graph features ANN . . . . .	40
6.9	Compare Prediction Accuracy With Additional Graph features ANN and LR . . . . .	41
6.10	Compare Prediction Accuracy With Additional Specific Graph features ANN . . . . .	42
6.11	Compare Prediction Accuracy With Additional Specific Graph features ANN and LR . . . . .	43
6.12	2017 Kentucky Derby Horses . . . . .	44
6.13	Predictions of Experts . . . . .	45
6.14	Example of Model Prediction . . . . .	45
6.15	Comparing Results in Kentucky Derby 2017 . . . . .	46
6.16	Comparing Results in Kentucky Derby 2016 . . . . .	47

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

Horse racing is a sport that includes two or more horses ridden by jockeys [1]. Almost every day of the year, some race occurs all around the world. There are different types of races such as flat race, which is run over a level track at a preplanned distance; harness race, which usually pulls a two-wheeled cart for horses. In this thesis, we use flat race for the models.

While horses are sometimes raced only for sport, a majority part of horse racing is relevant with gambling components [2]. Many people in the world bet horse races and want to earn money. For example, betting on the Kentucky Derby in 2017 totaled approximately \$139.2 million [3]. There are many betting systems in horse racing but basically called win bet which is guessing the winner horse in the race. Most traditional bets are usually called straight wagers, are Win, Place Show. The win is if a person wagers money to a horse, he wins only his horse finishes first; the place is, he wins if his horse finishes the race first or the second position; the show is, he wins if his horse finishes the race first, second or third position.



Figure 1.1: Win, Place, Show - How To Bet On Horses.

## 1.2 Motivation

We choose horse racing data because horse racings contain dozens of variables such as track type, weather, jockeys to name but a few. These horse race conditions support more challenge for prediction. Also, horse racing has been a well-known subject in machine learning field, so that we would like to work horse racing data for prediction.

Horse racing is not the only sport that is researched for data mining applications. There are different sports data such as tennis, college basketball are using in especially predictive models. When we finish the model, it may be extended to other applications of data mining in different sports.

## 1.3 Scope and Objectives of the Thesis

The main objective of this thesis shows that graph-based data help better prediction of horse race results. Usually, a prediction model is based on attributes that are from horses and training based on hundreds of races and are made a prediction [4]. Besides of these set of attributes and this model, we want to extend set of features which are not directly evaluated from original data, but they are generated based on the global picture of horse races. For these attributes, we are analyzing the graph of horses. According to the graph of horses that are created with data, to establish additional relation which will be extended set of attributes and, with these extended set of attributes and the basic set of attributes, we are trying to prove prediction will be improved or not.

For prediction part, we will use predictive models for horse racing, based on two machine learning methodologies that are artificial neural network and logistic regression. One of the biggest effort was data preparation part because we don't have available data so, we need to find useful data and prepare it for using the model.

## 1.4 Organization of the Thesis

The remainder of this thesis is as follows: Chapter 2 provides a literature review of related works and applied methodologies. Chapter 3 gives data preparation and feature extraction. Chapter 4 describes applying methodologies and experimental results. Finally, chapter 5 provides conclusion and future works.

## CHAPTER 2

### LITERATURE REVIEW

Before betting a horse racing, people want to analyze some contributions such as horses and jockeys to find possible winners. There are many newspapers and websites give some tips about winning possibilities, but these possibilities are not very close accurate of predictions. Many researchers have been studying racing predictions using different several algorithms. In this review, methodologies researched and applied by different researchers to established models have been investigated.

Williams and Li [5] studied horse racing in Jamaica to the prediction of winners. They used Artificial Neural Network methodology primarily. In Jamaica, horse races in distances of 1 to 3 km. Data were collected from 143 races from January to June 2007. They used four different learning algorithms and their results were compared. They studied different variables to horse racing such as horse, jockey, past position, track distance and horses finishing time and others as input to Artificial Neural Network. They used 80% of data for training and 20% of data of testing. Four algorithms predicted correct horses with an accuracy of around 70% and they took more or less similar results.

Similar to Williams and Li [5] Davoodi and Khanteymooori [4] also used artificial neural network for prediction system. They investigated horse racing predictions for just one race track in New York using artificial neural networks. Data were collected from 100 races for analysis from January 2010. Artificial Neural Networks were used to predict the finishing time of each horse in a race. They used five supervised algorithms; Gradient Descent BP Algorithm, Gradient Descent BP with momentum Algorithm, Quasi-Newton BFGS Algorithm, Levenberg-Marquardt Algorithm and Conjugate Gradient Descent Algorithm in their study to compare performances. Eight features had been used for input: Horse weight, race type, trainer, jockey, number of horses in a race, track distance and condition and weather. In network, one neural network is used to represent one horse. Each neural network has eight inputs and one output that is horse final position time . While BP and BPM were the best at predicting winners and Levenberg-Marquardt was found to be the fastest algorithm and all algorithms used had an average accuracy of around 77%.

Pardee [6] also studied artificial neural networks, but he predicted the result of football

matches in college. A model was developed using past football seasons data and predict the winning team. He used data for training and testing from 11 games for each team which played in the seasons of 1998 and 1999. The system predicted which team will win the college football game with an accuracy of 76%. Besides if given more data about the teams previous performances, the degree of the accuracy of system could be better.

Robert P. Schumaker [7] used support vector regression to predict horse ranking for next race. He used a model name was The S&C Racing System shown figure 2.1. This system consists several major components; data module, the machine learning part, betting engine and evaluation metrics. He used the traditional betting engine that is Win, Place or Show.

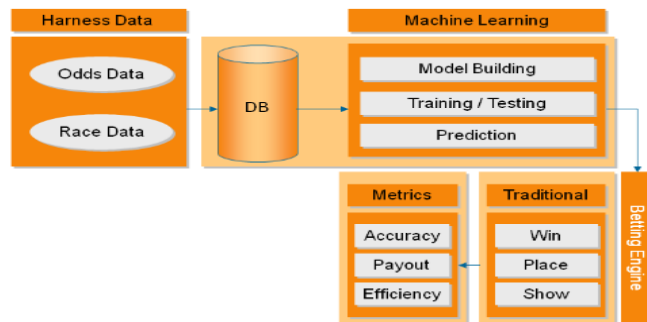


Figure 2.1: The S&C Racing System

Schumaker [7] used several features such as fastest time, win percentage, place percentage and average finishing position for the last several races. Results were tested three dimensions of evaluation that are accuracy, payout and efficiency. He was trying to find balance point for accuracy and payout when payout becomes maximum.

In addition to horse racing data, greyhound racing data are also using for prediction systems because some conditions are similar to horse racing data. Robert P. Schumaker and James W. Johnson [8] used support vector regression algorithm on more than 1900 races and 31 different dog tracks to made prediction. They used an almost similar system from Robert P. Schumaker [7], but they used different betting engine with the traditional engine in the system name was AZGreyhound System that shown figure 2.2. In Straight bets, there are exacta, trifecta and superfecta wagers. While in exacta, the bettor should predict 1st and 2nd place for dogs respectively, in trifecta is more harder by trying to predict first three places. In Superfecta, the bettor should determine best 4 dogs will pass finish line. In Box bets, the bettor should try every combination of placement.

In this paper, they tested three dimensions of evaluation; accuracy, payout and efficiency

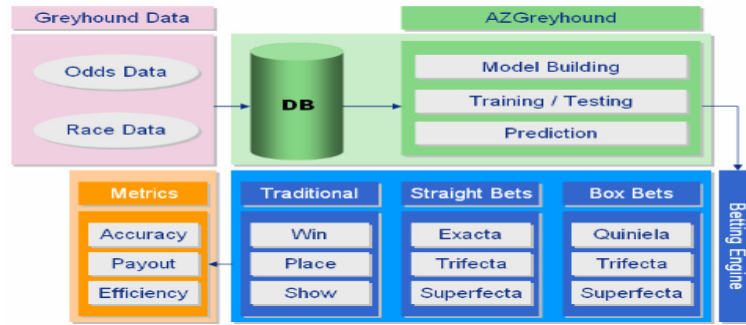


Figure 2.2: The AZGreyhound System

and they compared betting engine with .Their system predicted Wins and efficiency around for high accuracy low payout, and low accuracy high payout with different betting engine. They found that accuracy and payout are connected.

Hsinchun Chen et al. [9] also investigated prediction with greyhound racing. They tested artificial neural network algorithms (back propagation) and ID3 with 50 different variables, but they could not use all of the variables because the neural network didn't work with them. That's why they filtered more important variables and analyzed them. They used around 66% of data for training and around 33% for testing. Their system was predicted winners on 34 races and no winners were predicted for 16 races.After that, they placed bets on dogs that were predicted winners by the system in real. If a dog was predicted to finish first position, they put a \$2 wager. ID3 algorithm was %34 accuracy and \$69.20 payout, the neural network had %20 accuracy and a \$124.80 payout. In finally they found that the neural network predictions had better performance instead of human experts.

Prediction systems are not only for races, but also scientists use several sports. Baulch [10] tried to predict winners in rugby league and basketball games using artificial neural network. He focused mainly on the Backpropagation method and the Conjugate Gradient method for training. He used different type of features such as the teams previous performances, the number of wins and the average scores per game to predict the winner. His results are an accuracy that lies between %55 and %58.2 for rugby and between %49 and %59 for basketball. Finally, he showed that artificial neural networks helped a great performance in sports predictions and if research can be extended, results will be better.

Another example system for prediction horse racing is using Support Vector Machine Re-



gression [11]. The author used 20 different features and all features had a value of -1,0 or 1. For example, for the feature of Horse Win%, if horse's win percentage is more than %50 "1" else "0" and with this training, he tried to predict approximative finishing line position for horses. Data was prepared from December 2016 to February 2017 for training and March 2017 for testing. He tried to compare his result with the baseline (the morning line). **The baseline** is a list of entries for a horse race with the probable betting odds as estimated by a track handicapper issued before wagering begins. In results, he tried different strategies. Generally, He picked horses that have the better winning chance from morning line. In totally around 2900 races, for morning line(the baseline), the percentage of the favorite horse was %26 and its win place show results were around %58. For his model, favorite horse winning percentage was %28 and came in 1st 2nd and 3rd 1820 times %63. He also used different type of techniques and he found that machine learning works better than baseline.

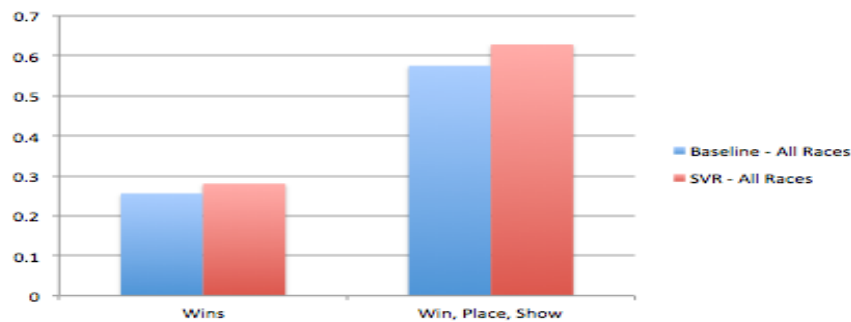


Figure 2.3: Results compared Baseline

In our thesis one of the big effort is creating a graph and extracting features from this global graph. Suvrat Bhooshan, et al. [12] investigated the application of network analysis for prediction. They used data from Major League Baseball (MLB) and the National Football League (NFL). They created directed graph as a relationship between teams for each league and it is the win-loss relationship for every two teams. They calculated the weight of edges between two teams; (the number of times team i won against j - the number of times i lost against j). For positive weighted on (i,j) means j is better than i, and negative is the exact opposite.

They used a ranking algorithm explained by Guo et al. [13], and they ordered nodes of subgraphs. That means higher ranked nodes have more incoming edges than outgoing edges. After this operation, they split nodes into two groups names were leaders and followers and they repeated these process until reached for global order. Furthermore, they applied two methods. One of them

is weight difference for a node that is;

$$d^w(i) = \sum_{(i,j) \in E} w(i,j) - \sum_{(j,i) \in E} w(j,i)$$

$w(i,j)$  is weight of  $(i,j)$ . Another algorithm is HITS algorithm that is a link analysis algorithm, rates Web pages, nodes for graphs. Hits algorithm has two different scores but, they just used authority score, not hubs score for each node as a ranking. Accurate classifier were improved with these rankings. They also used triads features similar to Jure Leskovec et al. [14]. Triad consists of three nodes and the possible connections between them. For example, if a connection between  $u$  and  $v$  nodes also can include another node that is  $w$ . Normally when we look two pairs of nodes  $(u,v)$  edges and, we can just see two different values that one of them is positive and other value is negative. In the triad, we can see four different values between  $u$  to  $v$  that are  $(++,+,-,+,-)$  from  $u$  to  $v$ . They used this triad between all pairs of nodes for taking values (scores) not directly connected between two pairs of nodes, but also undirected connections They used logistic regression algorithm for prediction In final, When they compared results, They found that accuracies NFL better than accuracies of MLB. Also when they compared expert ranking system (the baseline) they said that their system was more objective because expert systems handled recent games, not the entire season.

## CHAPTER 3

### DATA PREPROCESSING

In this chapter, we will present data preprocessing for model. Data preprocessing is a technique that is converting raw data into a clear format [15]. In the real world, data usually contain many errors or incomplete. Data preprocessing helps to resolve these issues prepare horse race data for analyzing. In this thesis, data preprocessing part design will be like in figure 3.1. Firstly we will show data collection from specific website that name is equibase [16]. After that, we will show data preparation part as Sethabhisha Naidu Nerusu et. al. [17]

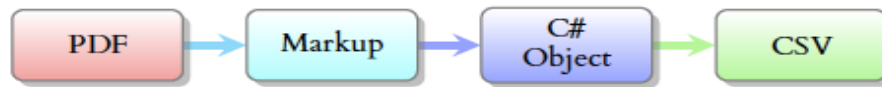


Figure 3.1: Overall Design

We will download pdfs files from website. After that we will design a program to convert the pdf to the markup text annotates text [18], and then to parse the markup text into C# objects. Finally we will convert horse race data to CSV files.

#### 3.1 Data Collection

Data collection part can be explained two main parts as follow :

1. Determine data from Equibase horse racing website held from 2015 to 2017 and downloading the daily race forms from Equibase website with semi-automated scripts.
2. Collect the pdfs downloaded put data in a tab delimited text file

##### 3.1.1 Download PDF files from the website with scripts

The goal of the scripts is to get all the race/track information from 2015 to 2017 as PDF files. In the website, race records are as pdf files and these pdf files don't only consist one race, but also include several races because website minimizes number of requests. In one day, there are several

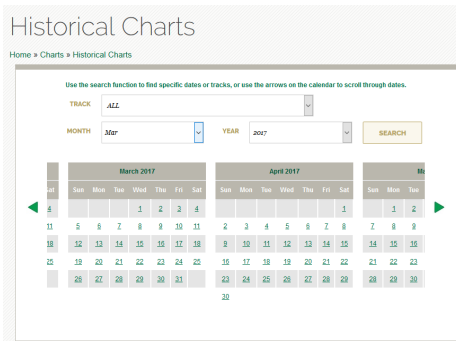


Figure 3.2: Dates of horse Racing

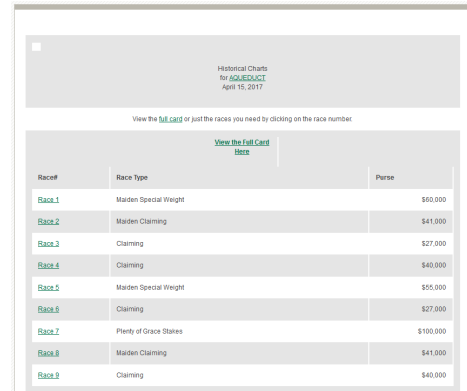


Figure 3.3: Example of one Track

tracks using for horse racing, so we will try to download each track from each day. Strategy is for each date, load the web page listing all the tracks where races happen and each track load the results for all races of the day. There are no races some days or some tracks, so we should eliminate them because system does not eliminate these empty files and it will bring them also like there are races in those days or tracks. If we also add these empty days or tracks, it is possible to receive errors during the parsing part. We will use equibase website for collecting data (<http://www.equibase.com/>). When we opened website/historical part, we can see that dates for horse racing like figure 3.2.

When we clicked one of the days, we can see tracks that there are races in that day and each track has raced as figure 3.3.

We use HTML code to frame the URLs automatically for all tracks in a day. It contains the list of all the tracks codes in the tid=xxx form. We will use them to find the tracking code. We tried to get the list of the tracks with bash code. We will use curl to get the HTML from the website, grep isolates the href lines and sed processed the href and extracts the tracking code figure 3.4, figure 3.5. We cannot call script directly because it is called by the download-pdfs script. If you call it directly, the syntax is (\$ ./get-available-tracks 2016 1 1) and output is as such as AQU, CMR, DED, FG, GG,GP.

```
#!/bin/bash
curl -s -H 'User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:43.0) Gecko/20100101
  Firefox/43.0' "http://www.equibase.com/premium/eqpVchartBuy.cfm?mo=$2&da=$3&yr=$1&trackc
  o=ALL;ALL&cl=Y" | grep "href.*lusIndex" | sed -e
  "s|.*tid=([~&]*\).*|1|g"
```

Figure 3.4: Script Code

With these scripts, we extract pdf files that contains more things. The PDF contents is returned by the web-site and saved to a pdf file. When we try to download pdf files, we have big

```
#Code extract
curl -s -A "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.2.3)
Gecko/20100401 Firefox/3.6.3" \
"http://www.equinbase.com/premium/eqbPDFChartPlus.cfm?RACE=A&BorP=P&TID=$f&C,
TRY=USA&DT=$5/$6/$1&DAY=D&STYLE=EQB">
$f/$1$5$6.pdf
```

Figure 3.5: Download the PDF

problems that is captcha. A captcha is a security for websites use to avoid robots to browse the site. It sends back a scrambled image to the user for proving user is human or not.

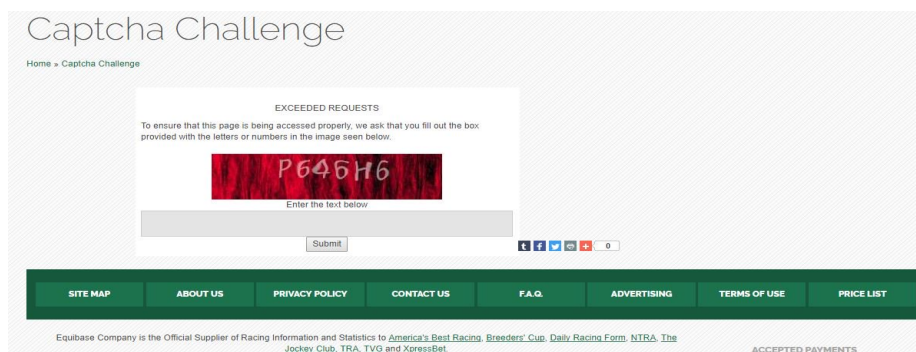


Figure 3.6: Example of captcha in Equibase website

Equibase website sends captcha every ten requests. If we don't defeat it, we cannot download data automatically. And also we should know how it works because it can be changeable every captcha. We used internet explorer debug mode and we can check every HTML code and every element on the page where we try to defeat captcha.

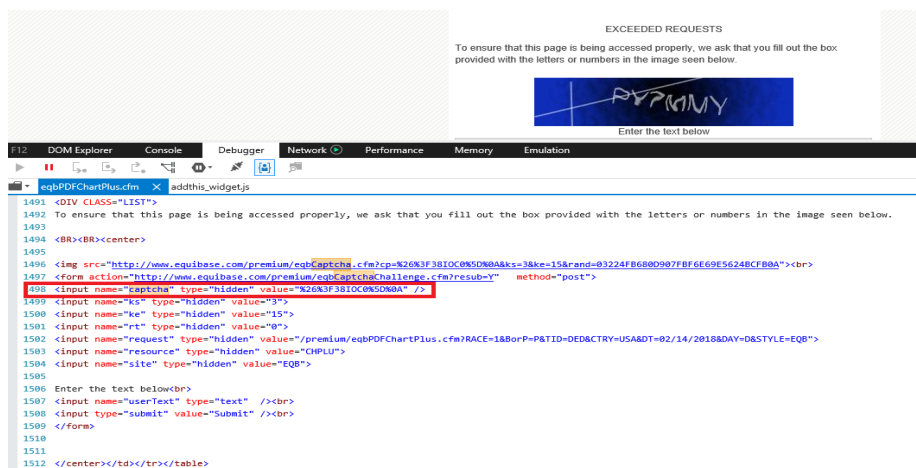


Figure 3.7: The captcha code

The web page includes a hidden captcha code, which contains a version of the captcha text. For the system that is created, we don't need to look encryption part we just have to send back the captcha text along with the captcha code. Actually, there is 2 type of captchas in the website;

protecting the tracks list and protecting the pdfs. We need to enter the captcha text and activate the Net tab than The Net tab contains a POST statement, which is the captcha send the website. There is a really important command that name is captcha-get-available-tracks is stored. We need to send the script before every get-available-tracks command because it generated a script and this script resets the allowed every 10 requests.

We have 4 different scripts captcha-get-available-tracks, get-available-tracks: called the captcha-get-available-tracks and get track list, captcha-download-pdfs send reset request for pdf and download-pdfs. When we run scripts, we should use like:

```
./generate-dates end-year end-month end-day start-year start-month start-day
```

so, until which date we want to download is beginning and from which date we want to start is final

```
$ ./generate-dates 2017 1 12 2017 1 1
```

And results are;

```
./download-pdfs 2017 1 12 17 01 10
```

```
./download-pdfs 2017 1 11 17 01 09
```

```
./download-pdfs 2017 1 10 17 01 08
```

```
./download-pdfs 2017 1 9 17 01 07
```

```
./download-pdfs 2017 1 8 17 01 06
```

```
./download-pdfs 2017 1 7 17 01 05
```

```
./download-pdfs 2017 1 6 17 01 04
```

```
./download-pdfs 2017 1 5 17 01 03
```

```
./download-pdfs 2017 1 4 17 01 02
```

This scripts will run for everyday and generates results. For example in above we can see that download races from 2017/1/12 to 2017/1/1. In final we took following results (figure 3.8) that displays date and track name (AQU is Aqueduct);

Scripts didn't always work sometimes we got some errors and when we took this error we regenerated manually. Also rarely, some pdfs files downloaded as HTML file so, sometimes we need to download them manually.

### 3.2 Data Parsing

A PDF Parser is a software which you can use to extract data from PDF documents. There are many different document types you can extract data PDF files such as text paragraphs, single

```
oadFile/codes $ ./generate-dates 2017 12 31 2015 1 1 | bash
2017-12-31-AQU
2017-12-31-CMR
2017-12-31-FG
2017-12-31-GG
2017-12-31-GP
2017-12-31-LRL
2017-12-31-PRX
2017-12-31-SA
2017-12-31-SUN
2017-12-31-TAM
```

Figure 3.8: Example of Script Results

data fields, tables and lists and images. If a pdf file contained image, we need to use optical character recognition to extract these images from pdf files. Horse race data doesn't include images figure 3.9 so, we don't need to use OCR system.

We need to extract horse race data from pdf files automatically like project [17] We don't need to write hundreds line of codes. We will use the iTextSharp framework and C#. Important part for parsing is a strategy to reliably and correctly parse the data because parser give us which texts are in pdf file but, we should find exact location of texts that we want to extract. When we process this PDF document with iTextSharp with we end up with the following text string in figure 3.10.

In this figure, the first line includes track, date, and race number. Other lines contain the race and horse types, layout becomes more variable so, text position becomes less reliable, except when we can find a beacon to fix on and start counting rows again. We will use text label technique that allows us reliable and more data. It looks like a key that helps find data quickly.

In the basic parsing, some data are created in a superscript font and there are three tables in the PDF page that can have a variable number of rows and columns. It is not enough to identify all parts of a document with raw text so, we need to find a mechanism that is reliable and we need to find way to provide the parser with hints. We used two kinds of markup tags to help to parse the PDF.

**Superscripts:** Some parts of text are surrounded by superscript text. For example, text like <sup>4</sup>AQU<sup>5</sup> output text is < 4 > AQU < 5 > so, superscript text is really easy to find.

**Tables:** When we look tables that we want to parse, we can see that we should make clear

**FAR HILLS - October 21, 2017 - Race 1**  
**STAKES Gladstone Hurdle S. - Thoroughbred**  
**TO BE RUN OVER NATIONAL FENCES FOR THREE YEAR OLDS.** 148 lbs. Winners, other than claiming, 4 lbs. extra for each race won. Owner of the winner to receive The Gladstone, a perpetual challenge bowl presented in memory of Grafton H. Pyme by Mrs. John H. Ewing to be held by the Association. The owner of the winner to receive a piece of plate. A piece of plate to the rider of the winner.  
 Two And One Eighth Miles On The Hurdle **Track Record:** (Kisser N Run - 3:55.10 - October 19, 2013)  
**Purse:** \$50,000 Guaranteed  
**Available Money:** \$50,000  
**Value of Race:** \$50,000 1st \$30,000, 2nd \$9,000, 3rd \$5,000, 4th \$2,500, 5th \$2,000, 6th \$1,500  
**Weather:** Clear **Track:** Good  
**Off at:** 1:00 **Start:** Good for all

Last Raced	Pgm	Horse Name (Jockey)	Wgt	M/E	PP	1/2	1m	11/2	13/4	Str	Fin	Comments
23Sep17 'SHW'	2	DQ-Snuggling (Galligan, Gerard)	148	L	2	5 <sup>1/2</sup>	5 <sup>1/2</sup>	5 <sup>1/2</sup>	4 <sup>1/2</sup>	1/2	1 <sup>1/2</sup>	impeded foe, hard drive
23Sep17 'SHW'	6	Menacing Dennis (Watts, Mark)	148	L	6	4 <sup>1/2</sup>	2 <sup>1/2</sup>	2 <sup>1/2</sup>	1 <sup>1/2</sup>	2 <sup>1/2</sup>	2 <sup>1/2</sup>	led, forced wide
23Sep17 'SHW'	3	Down Royal (Dalton, Bernard)	140	L	3	3 <sup>1/2</sup>	3 <sup>1/2</sup>	4 <sup>1/2</sup>	2 <sup>1/2</sup>	5 <sup>2</sup>	3 <sup>1/4</sup>	took up, went wide
27Jun17 'LE'	9	Aishibaa (IRE) (Townend, Paul)	148	L	9	8 <sup>1/2</sup>	8 <sup>2</sup>	7 <sup>1</sup>	8 <sup>24</sup>	6 <sup>5</sup>	4 <sup>1/2</sup>	took up, went widest
23Sep17 'SHW'	7	First Friday (Doyle, Jack)	152	b	7	4 <sup>1/2</sup>	4 <sup>1</sup>	6 <sup>2</sup>	5 <sup>1/2</sup>	3 <sup>1/2</sup>	5 <sup>1/2</sup>	lacked late response,
23Sep17 'SHW'	1	Flash Jackson (Geraghty, Ross)	148	L	1	1 <sup>3</sup>	1 <sup>1/2</sup>	1 <sup>2</sup>	3 <sup>1/2</sup>	4 <sup>1/2</sup>	6 <sup>5</sup>	battled, gave way,
15Apr17 'TRY'	8	Burning Approval (Nagle, Darren)	148	L	8	9	9	8 <sup>1</sup>	7 <sup>2</sup>	7 <sup>10</sup>	7 <sup>13</sup>	took up, went wide
16Aug17 'SAR'	4	Britain (Mitohel, Michael)	148	L	4	7 <sup>2</sup>	6 <sup>1/2</sup>	3 <sup>1</sup>	6 <sup>1</sup>	8 <sup>23</sup>	8 <sup>29</sup>	gave way,
30Jul17 'LRL'	5	What Gives (McDermott, Sean)	148	L	5	6 <sup>2</sup>	7 <sup>2</sup>	9	9	9	9	gave way,

**Final Time:** 4:05.70  
**Run-Up:** 0 feet

**Winner:** Menacing Dennis, Chestnut Gelding, by Greatness out of Soldier Girl, by Lost Soldier. Foaled Mar 26, 2014 in Florida.  
**Breeder:** Tom McCrocklin & Frank Mermenstein. **Winning Owner:** Bonnie Rye Stable

**Disqualification(s):** # 2 Snuggling from 1 to 3

Past Performance Running Line Preview						
Pgm	Horse Name	1/2	1m	11/2	13/4	Fin
2	Snuggling	5 <sup>1/2</sup>	5 <sup>3</sup>	5 <sup>4</sup>	4 <sup>1</sup>	1 <sup>1/2</sup>
6	Menacing Dennis	2 <sup>1/2</sup>	2 <sup>1/2</sup>	2 <sup>3</sup>	1 <sup>1/2</sup>	2 <sup>1/2</sup>
3	Down Royal	3 <sup>1/2</sup>	3 <sup>2</sup>	4 <sup>4</sup>	2 <sup>1/2</sup>	5 <sup>2</sup>
9	Aishibaa (IRE)	8 <sup>1/2</sup>	8 <sup>7</sup>	7 <sup>7/2</sup>	8 <sup>5</sup>	6 <sup>5</sup>
7	First Friday	4 <sup>1/2</sup>	4 <sup>2</sup>	6 <sup>5</sup>	5 <sup>3</sup>	3 <sup>3</sup>
1	Flash Jackson	1 <sup>3</sup>	1 <sup>1/2</sup>	1 <sup>2</sup>	3 <sup>1/2</sup>	4 <sup>2</sup>
8	Burning Approval	9 <sup>19</sup>	9 <sup>9</sup>	8 <sup>8</sup>	7 <sup>4</sup>	7 <sup>10</sup>
4	Britain	7 <sup>15</sup>	6 <sup>5</sup>	3 <sup>2</sup>	6 <sup>3</sup>	8 <sup>10</sup>
5	What Gives	6 <sup>13</sup>	7 <sup>5</sup>	9 <sup>10</sup>	9 <sup>30</sup>	9 <sup>33</sup>

**Trainers:** 2 - Sheppard, Jonathan; 6 - Gomena, Julie; 3 - Dalton, Kate; 9 - Young, Leslie; 7 - Fout, Paul; 1 - Wyatt, Todd; 8 - Murphy, Brian; 4 - Hendriks, Richard; 5 - Fisher, Jack

**Owners:** 2 - Hudson River Farms; 6 - Bonnie Rye Stable; 3 - Riverdee Stable; 9 - Ballybristol Farm LLC; 7 - Beverly R. Steinman; 1 - Ann Jackson; 8 - Murphy, Brian E. and Murty, Tara; 4 - Eve Ledyard; 5 - Andrea S. Knight

**Footnotes**  
 SNUGGLING was washy in the paddock and the race, rated into the second circuit, rallied between rivals on the final turn, impeded DOWN ROYAL near the quarter pole, continued on willingly to duel up the hill to the final hurdle, crossed the paths of FLASH JACKSON and MENACING DENNIS before the hurdle and duelled with MENACING DENNIS over the hurdle and to the finish under strong urging, MENACING DENNIS pressed the pace into the second circuit, gained a short lead before the second-last hurdle, was forced wide by the winner near the final hurdle, and battled to the end, DOWN ROYAL saved ground pressing the pace, rallied on the final turn, took up sharply near the quarter pole, went wide approaching the hill, finished willingly and the filly may have been best, ALSHIBAA (IRE) was outgun to the final turn, rallied on the inside, took up behind DOWN ROYAL, went wide and finished very willing, FIRST FRIDAY pressed the pace into the final turn and lacked the needed late response, FLASH JACKSON set the pace to the second-last hurdle, battled for the lead while wide on the final turn and gave way when SNUGGLING out across him closing in on the final hurdle, BURNING APPROVAL was outgun to the final turn, rallied on the rail, took up sharply and went wide passing the quarter pole, BRITAIN raced in the middle of the pack to the second circuit and gave way, WHAT GIVES raced in the middle of the pack to the second circuit and gave way, AFTER THE RIDERS OF MENACING DENNIS AND DOWN ROYAL CLAIMED FOUL AGAINST THE RIDER OF SNUGGLING, THE STEWARDS DISQUALIFIED SNUGGLING AND PLACED HIM THIRD.


 Denotes a Keeneland Sales Graduate

Figure 3.9: Example of Pdf File

such as table's start and end etc.

When the text is completely marked up, we are ready to convert the pdf to the text and to parse the markup text into C# objects.

The data goes to PDF file than markup language defined earlier in the document than C# representation of the race objects. This is an in-memory data model which will be transformed into the final format. In final, we are trying to convert data into a CSV file through a CSV renderer figure 3.11.

The markup generation is class HorizontalTextExtractionStrategy, derived from the iTextSharp class LocationTextExtractionStrategy. The main function of this class is, it is invoked by the iTextSharp parser for every single word of the document. The parameters of this function contain the word,data position and format. The function can determine if a new line starts.

We need to take a lot of data information and should modify each word on the fly we will use a **filter chain**. Each word makes its way through the filter in filter chain and the output of the



```

AQUEDUCT - January 11, 1994 - Race 1
CLAIMING - Thoroughbred
Claiming Price: $12,500 - $0
One And One Eighth Miles On The Inner track Track Record: (Conveyor - 1:47.33 - March 6, 1993)
Purse: $11,000
Available Money: $11,000
Value of Race: $11,000 1st $6,600, 2nd $2,420, 3rd $1,320, 4th $660
Weather: Cloudy Track: Fast
Off at: 12:00 Start: Good for all
Last Raced Pgm Horse Name (Jockey) Wgt M/E PP Start 1/4 1/2 3/4 Str Fin Odds Comments
4 5 1 1 Head 2 1/2 1
1Jan94 AQU 9 Hudlam's Sidekick (Rodriguez, Rudy) 112 b 8 2 5 5 2 1 3.10 driving
9 6 1/2 1 1 1 3/4
31Dec93 AQU 11 Star Kalibur (Mojica, Jr., Rafael) 113 c 9 6 6 4 2 1 2 18.70 dueled, weakened
9 2 1/2 3
1Dec93 AQU 7 Private Access (Chavez, Jorge) 117 b 6 11 12 12 12 5 3 2.00* closed fast
1 5 1 1/2 1 1/2 1 Neck
8Dec93 AQU 12 Crafty Investment (Leon, Filiberto) 108 fc 10 8 8 6 6 3 4 19.60 rallied, 3 wide
4 7 Head 1/2 Head Head 1/2
2Jan94 AQU 5 Electret (Mickens, Tony) 106 - - 4 7 4 7 7 6 5 39.10 rallied, inside
9 1 1 1/2 1/2 1 1/2 2 1/2
31Dec93 AQU 1a Stack Um Up (Luzzi, Michael) 117 b 12 9 10 9 8 4 6 4.60 wide thru-out
9 4 1/2 1/2 1 Head 1 1/4
2Jan94 AQU 8 Get the Bags (Sweeney, Katy) 113 f 7 10 9 8 9 8 7 34.40 3 wide thru-out
1 8 1 4 1/2 1 2 3/4
22Dec93 AQU 13 Diamond Skater (Luttrell, Michelle) 112 fc 11 12 11 11 10 9 8 17.40 no late rally
4 6 1 2 2 5 3 1/2
2Jan94 AQU 6 Joe's Jackie (Alvarado, Frank) 117 bf 5 5 7 10 11 11 9 50.40 failed to menace
9 10 1/2 1 Head 1 1
29Dec93 AQU 2 Return to Newport (Santagata, Nick) 117 - - 1 1 1 1 1 7 10 12.10 dueled, gave way
1 6 Head 1 Head 1/2 23
11Dec93 AQU 4 Presence (Vega, Harry) 113 f 3 4 3 3 3 10 11 10.40 gave way stretch
6 6 1/2 Head 1/2
2Dec93 PHA 3 North Warning (Velazquez, John) 113 - - 2 3 2 2 4 12 12 13.30 faded
Fractional Times: 23.79 48.55 1:13.74 1:40.42 Final Time: 1:53.80
Split Times: (24:76) (25:19) (26:68) (13:38)
Run-Up: 0 feet
Winner: Hudlam's Sidekick, Dark Bay or Brown Gelding, by Kris S. out of Dream Buster, by Spellcaster.
  -> Foaled Apr 18, 1989 in Florida.
Breeder: Meadowbrook Farms, Inc. Winning Owner: Steve Simon
Claiming Prices: 9 - Hudlam's Sidekick: $12,500; 11 - Star Kalibur: $10,500; 7 - Private Access:
  -> $12,500; 12 - Crafty Investment:
$10,500; 5 - Electret: $10,500; 1a - Stack Um Up: $11,500; 8 - Get the Bags: $10,500; 13 - Diamond
  -> Skater: $12,500; 6

```

Figure 3.10: Text String

Pgm	Horse Name	Start	1/4	1/2	3/4	Str	Fin
1	Dublinoymoney	9	3 <sup>1/2</sup>	4 <sup>1</sup>	4 <sup>2</sup>	4 <sup>8</sup>	4 <sup>10 1/2</sup>
2	Odyssey	8	1 <sup>1/2</sup>	2 <sup>1</sup>	2 <sup>2</sup>	2 <sup>8</sup>	2 <sup>10 1/2</sup>
6	Nseventeen	3	7 <sup>2 3/4</sup>	6 <sup>4 1/2</sup>	4 <sup>5 1/2</sup>	4 <sup>8 1/2</sup>	3 <sup>10 1/2</sup>
3	My Boy Lane	1	4 <sup>3/4</sup>	4 <sup>3</sup>	3 <sup>5</sup>	3 <sup>8</sup>	4 <sup>12 1/4</sup>
8	Reydell	5	5 <sup>1 1/4</sup>	5 <sup>3 1/2</sup>	5 <sup>8</sup>	5 <sup>11 1/2</sup>	5 <sup>16 1/2</sup>
4	Chill Phil	7	6 <sup>2 1/4</sup>	7 <sup>5 1/2</sup>	6 <sup>13</sup>	6 <sup>21</sup>	6 <sup>24 3/4</sup>
5	Negu	4	8 <sup>4 1/4</sup>	8 <sup>8 1/2</sup>	8 <sup>14</sup>	7 <sup>22</sup>	7 <sup>25 1/2</sup>
9	Awesome Mil	6	9 <sup>5 3/4</sup>	9 <sup>10 1/2</sup>	9 <sup>15 1/2</sup>	8 <sup>22 1/2</sup>	8 <sup>26 1/2</sup>
7	Teenertown Prince	2	2 <sup>1/2</sup>	3 <sup>2</sup>	7 <sup>14</sup>	9 <sup>25</sup>	9 <sup>29 3/4</sup>

Figure 3.11: Sample table

first filter is fed into the second filter. Each filter has opportunities the following:

1. Change the word that transmits like parameter before you passed the next filter
2. Bypass subsequent filters and ignore the word. After that, word that is ignored will not be in the markup text. Every word is passed into the filter chain.

Every filters take one word so, text passed through a pre-processor chain, in fact, to make calculation again or pattern recognition and after that, it goes preprocessing. Similarly, the markup text is passed through a post-processing chain to prepare the final result. Some filters connect with

some other filter status and all filters communicate with each other via publication and subscription. Usually one filter in the chain has already finished its job than other filter will be active. If we do not do this, a filter can be confused if a horse name contains the word. That is one of the errors we faced. The markup text is converted to C# objects, and These objects are defined in the Model subdirectory. While the C# object hierarchy cannot render CSV file, the C# objects are de-normalized will be written as a CSV file in different objects.

The program interface is easy: EquibaseParser.exe input1 input2 input3 ...inputn and output has three files are created as a result.

```

log-20180305-161902.txt - Notepad
File Edit Format View Help
.\20170102 (2).pdf starting, 8 page(s) to process
.\20170102 (2).pdf finished, 8 pages processed, 8 races OK, 0 failed.
.\20170102 (3).pdf starting, 9 page(s) to process
Error in .\20170102 (3).pdf page 1 (Input string was not in a correct format.)
.\20170102 (3).pdf finished, 9 pages processed, 8 races OK, 1 failed.
.\20170102 (4).pdf starting, 11 page(s) to process
Error in .\20170102 (4).pdf page 3 (Input string was not in a correct format.)
.\20170102 (4).pdf finished, 11 pages processed, 8 races OK, 1 failed.
.\20170102 (5).pdf starting, 9 page(s) to process
.\20170102 (5).pdf finished, 9 pages processed, 7 races OK, 0 failed.
.\20170102 (7).pdf starting, 11 page(s) to process
Error in .\20170102 (7).pdf page 2 (Input string was not in a correct format.)
Error in .\20170102 (7).pdf page 4 (Input string was not in a correct format.)

```

Figure 3.12: Log File

**Log File** has the log of all the races processed. If any file stopped in error we can see in log file and we can fix error that occurred while running

```

wagers-20180305-162154.csv - Notepad
File Edit Format View Help

```

TrackName	Date	Number	WagerType	WinningNumbers	Payoff	Pool	Carryover
MAHONING VALLEY RACE COURSE	1/2/2017	1	\$2.00 Exacta	2-1	57.6	35524	
MAHONING VALLEY RACE COURSE	1/2/2017	1	\$0.50 Trifecta	2-1-4	61.2	24648	
MAHONING VALLEY RACE COURSE	1/2/2017	1	\$0.10 Superfecta	2-1-4-5	22.65	17013	
MAHONING VALLEY RACE COURSE	1/2/2017	2	\$2.00 Daily Double	2-4	149.2	12958	
MAHONING VALLEY RACE COURSE	1/2/2017	2	\$2.00 Exacta	4-6	184.6	51046	
MAHONING VALLEY RACE COURSE	1/2/2017	2	\$0.50 Trifecta	4-6-1	660.65	34923	
MAHONING VALLEY RACE COURSE	1/2/2017	2	\$0.10 Superfecta	4-6-1-8	716.95	24999	
MAHONING VALLEY RACE COURSE	1/2/2017	3	\$2.00 Exacta	5-6	47.4	80566	
MAHONING VALLEY RACE COURSE	1/2/2017	3	\$0.50 Trifecta	5-6-10	40.2	53128	
MAHONING VALLEY RACE COURSE	1/2/2017	3	\$0.10 Superfecta	5-6-10-8	56.98	36242	
MAHONING VALLEY RACE COURSE	1/2/2017	4	\$2.00 Exacta	8-6	165	57670	
MAHONING VALLEY RACE COURSE	1/2/2017	4	\$0.50 Trifecta	8-6-4	299.95	34275	
MAHONING VALLEY RACE COURSE	1/2/2017	4	\$0.10 Superfecta	8-6-4-9	153.52	21322	
MAHONING VALLEY RACE COURSE	1/2/2017	4	\$0.50 Pick 3	4-5-8 (3 correct)	344.35	8123	
MAHONING VALLEY RACE COURSE	1/2/2017	5	\$2.00 Exacta	3-5	20.4	52055	
MAHONING VALLEY RACE COURSE	1/2/2017	5	\$0.50 Trifecta	3-5-8	28.05	32030	

Figure 3.13: Wager File

**Wager File** includes especially standard wager types other attributes such as track name, date and payoff for horse races (figure 3.13).

	A	B	C	D	E	F	G	H	I	J	K	L
1	TrackName	Date	Number	Type	HorseType	ClaimingP	Distance	Purse	Available	ValueOfR	Weather	Track
2	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
3	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
4	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
5	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
6	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
7	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
8	MAHONIN	1/2/2017	1	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
9	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
10	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
11	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
12	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
13	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
14	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
15	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
16	MAHONIN	1/2/2017	2	STARTER /	Thoroughbred		Six Furlon	15000	16000	\$15,000 1s	Foggy	Muddy
17	MAHONIN	1/2/2017	3	CLAIMING	Thorough	\$4,000	One Mile	8100	9100	\$8,100 1st	Showery	Muddy
18	MAHONIN	1/2/2017	3	CLAIMING	Thorough	\$4,000	One Mile	8100	9100	\$8,100 1st	Showery	Muddy
19	MAHONIN	1/2/2017	3	CLAIMING	Thorough	\$4,000	One Mile	8100	9100	\$8,100 1st	Showery	Muddy
20	MAHONIN	1/2/2017	3	CLAIMING	Thorough	\$4,000	One Mile	8100	9100	\$8,100 1st	Showery	Muddy
21	MAHONIN	1/2/2017	3	CLAIMING	Thorough	\$4,000	One Mile	8100	9100	\$8,100 1st	Showery	Muddy

Figure 3.14: 2017 Result File

**Result File** is actual data and attributes such as weather, track name, date, distance and track type that we use in model shown in figure 3.14.

### 3.3 Data Preprocessing:

In horse racing distance of track and track type are important attributes that are affect horse racing results. In the data, these attributes are together like one attribute as shown in figure 3.15.

One Mile And Seventy Yards On The Inner track  
Six Furlongs On The Dirt

Figure 3.15: Distance and Track Type Example

We should split distance and type of track into 2 attributes. For example One Mile and Seventy Yards as a distance and Dirt as a track type. We should split 2 different attributes because there is no connection with these attributes for all of races.

	A	B
1	Seven Furlongs	0.87
2	Five And One Half Furlongs	0.68
3	Six Furlongs	0.75
4	One Mile	1.00
5	Five Furlongs	0.62
6	Six And One Half Furlongs	0.81
7	One And One Sixteenth Miles	1.06
8	One Mile And Seventy Yards	1.03
9	One And One Fourth Miles	1.25
10	One And One Eighth Miles	1.13
11	One And Three Sixteenth Miles	1.19
12	Seven And One Half Furlongs	0.93
13	Four And One Half Furlongs	0.56
14	Four Furlongs	0.50
15	One Mile And Forty Yards	1.02
16	About One And One Sixteenth Miles	1.06

Figure 3.16: Example of Distance Type

On the other hand, distance section is in words and in different units like yards, miles, and furlongs. So, we should convert all type of distance to one type as a mile. Also, we should converted words to numeric values manually like figure 3.16.

When we started to search horse races, we saw that there are a lot of conditions and attributes affect results but, some of them are important because of they affect directly results. Like most of the research and paper, Elnaz Davoodi et al [4], they used as features horse weight, type of race, horse's trainer, horse jockey, number of horses in the race, race distance, track condition and weather. When we analyzed other research and races, we can see that jockey is more effective than trainer in races. Also horse race data just include same type of races. track type weather, horse weight, number of horses in one race are important conditions and attributes in horses races. With all of these researches and horse races results we decided to use 8 different features for horse racing prediction: track type, weather , jockey , horse's past final position, horse weight, number of horses in the race, track condition and distance and also for comparing we will use probable betting odds (the baseline) attributes. Before input data for training and testing we should change some of them nominal to numeric. Horse' weight, number of horse, past position and distance are numeric values. For track type, weather and track condition, we encode them into numerical values for using as input data. For horse jockey values is calculated by dividing the number of wins by the number of rides in horse racing data. Some of the attributes in data are below figures.

Name: TrackType		Type: Nominal	
Missing: 1948 (2%)		Distinct: 6	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	Dirt	74473	74473.0
2	Turf	13322	13322.0
3	AllWeather	5643	5643.0
4	Inner	2033	2033.0
5	Inner turf	913	913.0
6	Downhill turf	18	18.0

Figure 3.17: Track Type in Data

In figure 3.17, we can see that track types such as dirt, turf and inner. Dirt type is more than other type in horse races.

Name: Weather		Type: Nominal	
Missing: 2442 (2%)		Distinct: 7	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	Clear	36338	36338.0
2	Cloudy	31929	31929.0
3	Rainy	1867	1867.0
4	Showery	2621	2621.0
5	Hazy	593	593.0
6	Snowing	378	378.0
7	Foggy	195	195.0

Figure 3.18: Weather Attributes

in figure 3.18 we can see weather conditions in horse races. In data, most of the races are organized in clear and cloudy weather

Name: Distance		Type: Numeric	
Missing: 1922 (2%)		Distinct: 41	
		Unique: 0 (0%)	
Statistic	Value		
Minimum	0.056		
Maximum	2.25		
Mean	0.761		
StdDev	0.26		

Figure 3.19: Distance Attributes

In figure 3.19, we can see that distances in data. Minimum distance in one race is 0.056 mile and maximum distance in one race is 2.25.

## CHAPTER 4

### INTRODUCTION OF GRAPH THEORY

The graph is basically a network that defines the relationship between points [19]. Each object in a graph is called a node. Objects in a graph are described by nodes, and relations between these objects are described by graph edges [20].

If we want to understand system that is really complicated, we need to know how its elements interact with each other. In other words, we need a map of its global picture.

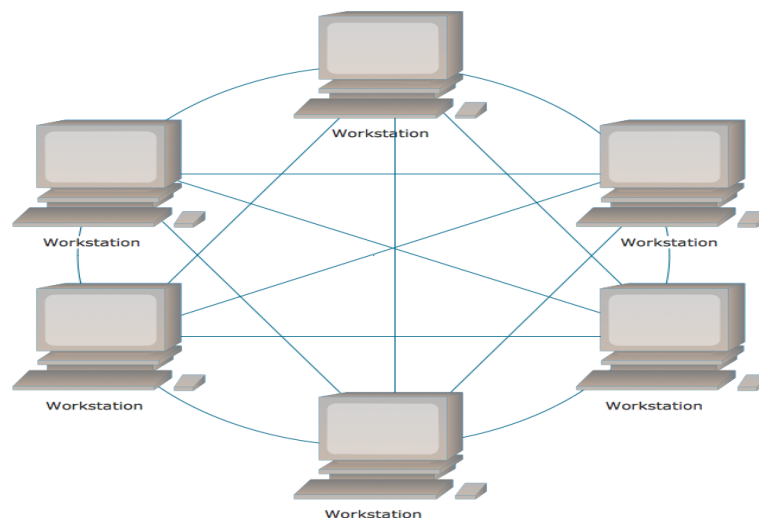


Figure 4.1: Example of Graph

In figure 4.1, we can see a network of graph. In this network, components (computers) often called nodes or vertices and the interactions between them, called links or edges. With this graph network, we can see whole system of computer network and connections like a big picture.

It is possible to divide graphs as directed graphs and non-directed graphs, depending on whether the edge types [21]. If the set of edges are unordered, this graph called undirected graph and if a graph consists of vertices and ordered pairs of edges, this graph called directed graph [22].

### Undirected Graph

Undirected graph is a graph that has relations between pairs of nodes, so that each edge has no directional character [23].

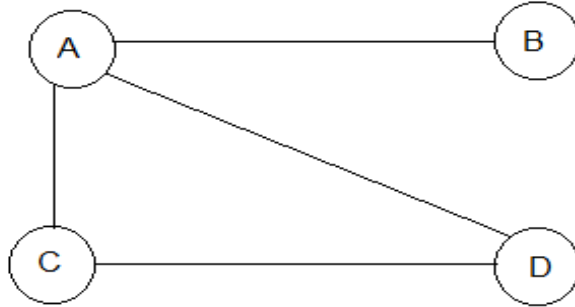


Figure 4.2: Undirected Graph Example

For example, in figure 4.2 a undirected graph shows consisting of 4 nodes and 4 edges. This graph can be expressed as  $G = ([A,B,C,D], [(A,B),(A,C),(C,D),(A,D)])$ . Therefore, graphs are written in the form of  $G = (V, E)$ , ie, nodes and edges.

### Directed Graph

Unlike undirected graph, directed graph set of nodes are connected pairwise by directed edges [24] so, we can say that each edges like one way streets.

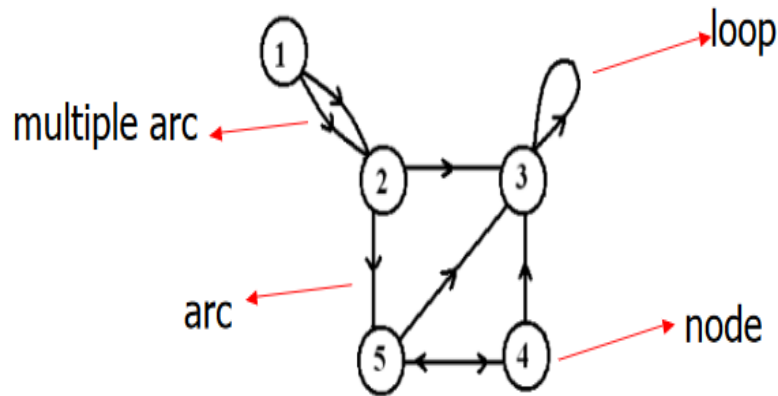


Figure 4.3: Directed Graph Example

In figure 4.3, There is a directed graph. Arc is directed edge way from a node to another node. Sometimes some nodes has edges for themselves (loop edge). if one node has two edges to

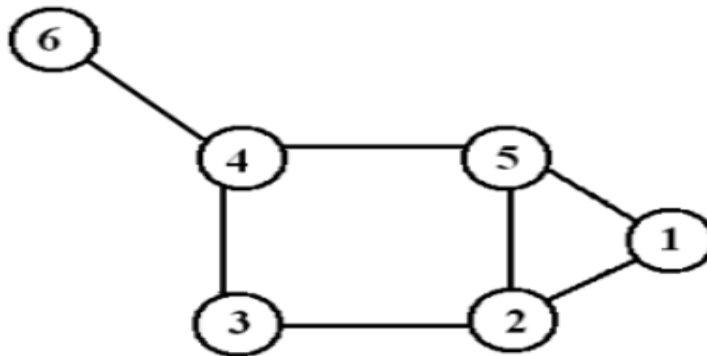
other node it called multiple arc or multiple edges.

### Labeled Graph

In graph, nodes have some degrees. Degree is number of edges incident on a node and two type of degree a node can has [25]. In-degree is number of edges entering and out-degree is number of edges leaving. For example graph that is above, node three (Node3) has 1 out-degree and 4 in-degree and total degree is 5.

### Adjacency Matrix

Graphs also represent as a matrix that is called adjacent in graph theory and computer science. Adjacent matrix show pairs of vertices connect in a finite graph or not [26].



	1	2	3	4	5	6
1	0	1	0	0	1	0
2	1	0	1	0	1	0
3	0	1	0	1	0	0
4	0	0	1	0	1	1
5	1	1	0	1	0	0
6	0	0	0	1	0	0

Figure 4.4: Adjacent Matrix

In graph and matrix in figure 4.4, we can see that connections represent 0 and 1. For example node 6 has one connection with node 4 so, other numbers are 0 and with node 4 value is 1 in diagonal. Also node 5 has 3 connections and it has three 1 value that are with node 1, node 2



and node 4. This graph is also example of non-labeled graph

We can used graphs in data organization, computational devices as well as many types of processes in physical, biological, social fields and especially network systems as a model [27]. We can see graphs in everywhere in the world. Even if friendship with other people is a graph network. [28]

In this thesis, we want to create a global picture of horse races with directed graph and extract some additional features that generated connection horses from a directed graph as well as basic features that are directly come from horse races data. Important part that, we want to create two type of graph for new features. One of them includes only basic edge labeled connections between horses and other describes graph contains edge labeled connections that include some conditions. We will use similar approaches as Suvrat Bhooshan et al. [12] for this models.

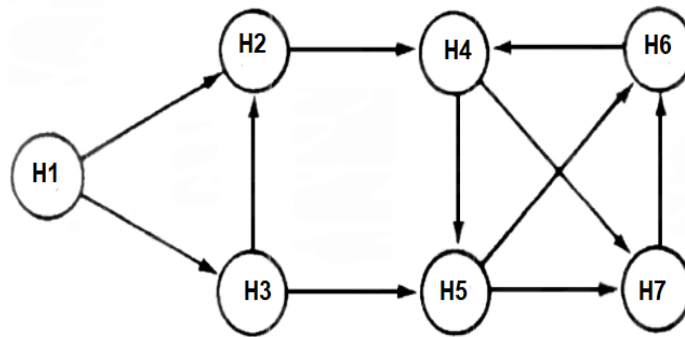


Figure 4.5: Model of Horse Graph

	1	2	3	4	5	6	7
1	0	1	1	0	0	0	0
2	0	0	0	1	0	0	0
3	0	1	0	0	1	0	0
4	0	0	0	0	1	0	1
5	0	0	0	0	0	1	1
6	0	0	0	1	0	0	0
7	0	0	0	0	0	1	0

Figure 4.6: Adjacent Matrix of Horse Graph

For example in graph and adjacent matrix in figure 4.5 and 4.6, if we look H4, it has 2 out-degree and 2 in-degree. Out-degree represents win edges and in-degree represents lose edges. In this thesis, two kind of graphs are created: labeled edge graph and multiple labeled edge graph.

#### 4.1 Labeled Edge Graph

For basic edge labeled graph model, we want to use the win-loss spread [29] of horses for all races in the data. Each horse has a relationship with other horses as weighted in edges. For example if a horse  $i$  beat another horse  $j$ , we will put an edge from horse  $i$  to horse  $j$ , or in a race, if a horse loses to another horse, we will put an edge from winner horse to other horse. Then, we will calculate win-loss spread between horse  $i$  and horse  $j$  as a weighted edge  $(i,j)$ . For instance, if a horse  $i$  beat 3 times other horse  $j$  we will put 3 edges from horse  $i$  to horse  $j$  and horse  $j$  beat 2 times horse  $i$  we will put 2 directed edges from  $j$  to  $i$ . Then we will calculate these edges  $(3 - 2 = 1)$  and we can say edge weight is 1 from  $i$  to  $j$ . One of example is in figure 4.7

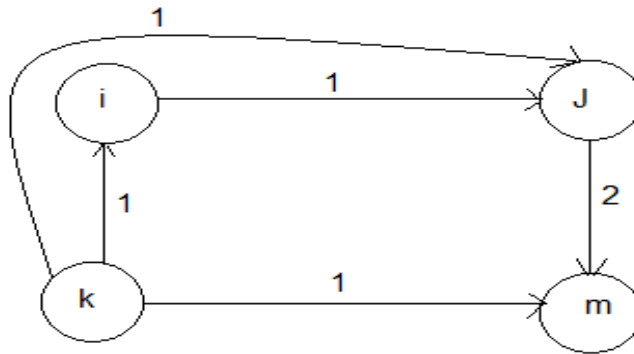


Figure 4.7: Example of Directed Edge Weighted Graph

#### 4.2 Multiple Labeled Edge Graph

Other type of graph is same the system as basic edge labeled graph but, we added some conditions on edges such as weather, track condition or track type. For example, when we want to predict a horse racing in some specific conditions such as cloudy weather and muddy track, we can directly extract features from graph that contains these conditions in edges and we can find more specific results. For example we take 2 horses and we want to know which will beat other horse in sunny weather and Dirt track. After creating the graph with weighted edges and these two conditions on edges, we will eliminate other edges that include different conditions and we will extract features from this edge labeled graph with specific conditions. In figure 4.2 and 4.3 show that an example of specific edge graph. Different colors include same type of conditions and we just look one type of color that is same conditions we have determined.

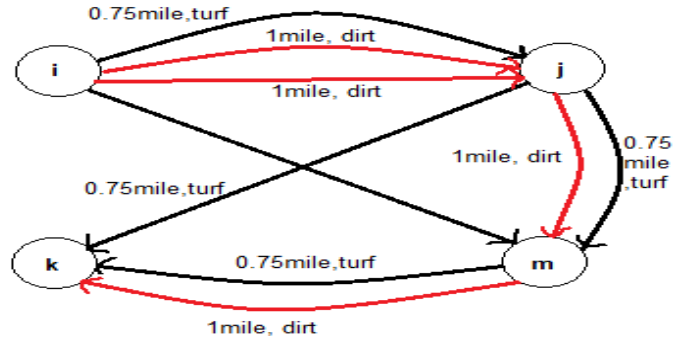


Figure 4.8: Example of Edge Connected Graph With Specific Conditions

For Example in figure 4.8, we can see that some edges have two conditions that distance 1 mile and dirt track type and other edges contain different conditions that are distance 0.75 and turf track type.

With these conditions, we eliminated other connections like figure 4.9 and we created the graph specifically with these conditions. These conditions are not only distance and track type, we can decide when we picked two pairs of horses and we can change with different mutual conditions.

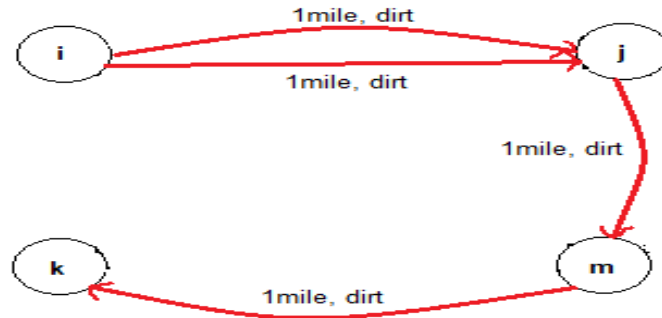


Figure 4.9: Example of Edge Connected Graph With Specific Conditions

### 4.3 Graph-Based Features

When we create a system, we need to extract some features or values from the system, and we can use them in different scientific fields such as machine learning and statistics. The graph is a big global system. That is why we need to extract some features to use approaches. Feature extraction covers narrowing resources to describe a large set of data [30]. When we are performing analysis of a system or graph, important problem is number of variables. Analyzing large data or system requires a lot of memories. That is why we need to construct combinations of the variables. Important part of the feature extraction is determine which features you use [31].

After creating the graph, we need to find which features we can use and how to extract features from the graph. In Guo et al. [13] investigated an algorithm that ordered the nodes by the degree difference. They used how many edges coming from other nodes to node  $i$  minus how many edges going from nodes  $i$   $d^\Delta = d_{in} - d_{out}$  and if a node has higher ranked it should have more incoming edges than outgoing edges. After this ranked score, they used a parameter  $\alpha$  and separated into two class of nodes that are leaders  $V_l$  and followers  $V_f$ . Than they repeated same processes and until global achieved.

In this thesis, we will use similar ranking score but we also need to give averages because some horses have more than races than others so, if we give directly incoming edges minus outgoing edges it can be a problem. For example, a horse has 3 races and 3 times win. Other horses has 9 races and 6 times win 3 times lose. If we directly calculate  $d^\Delta = d_{in} - d_{out}$  it will be not equal system. That is why we made  $d^\Delta = \frac{d_{in}-d_{out}}{n}$ ,  $n$  is number races that attend a race.

Suvrat Bhooshan et al. [12] also investigated for a node  $i$  edge weight difference as

$$d^w(i) = \sum_{(i,j) \in E} w(i,j) - \sum_{(j,i) \in E} w(j,i)$$

We will also use the same approach as a graph feature in the graph.

There is another algorithm that can be useful and improve nodes ranking score is HITS algorithm. The HITS algorithm that is Hyperlink-Induced Topic Search (HITS) is an analyze algorithm for links of networks that rates Web pages, nodes in graphs etc. [32]. The HITS algorithm computes two numbers for a node. Authorities calculate the node value based on the incoming edges. Hubs calculates the node value based on outgoing edges. For example there is a graph and three node that are X, Y and Z figure 4.10.

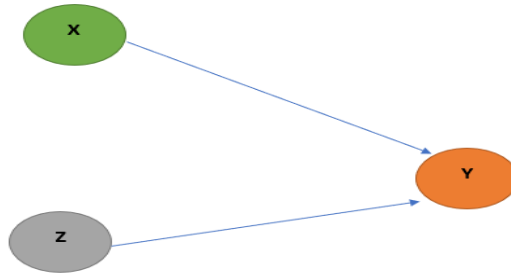


Figure 4.10: Example of Graph for HITS Algorithm

The adjacency matrix of the graph is in figure 4.11;

	<b>1</b>	<b>2</b>	<b>3</b>
<b>1</b>	0	0	1
<b>2</b>	0	0	1
<b>3</b>	0	0	0

Figure 4.11: Adjacent Matrix for HITS Algorithm

After compute adjacent matrix we will compute transpose of matrix in figure 4.12;

	<b>1</b>	<b>2</b>	<b>3</b>
<b>1</b>	0	0	0
<b>2</b>	0	0	0
<b>3</b>	1	1	0

Figure 4.12: Transpose of Adjacent Matrix for HITS Algorithm

If hub weight vector is accept in the beginning like:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

We compute the authority weight, matrix transpose \* hub weight vector;

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

And after that we update the hub weight matrix \* authority weight ;

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$$

When we look these two matrix we reached that node X and Z is a hub since it has 2 hub score and node Y is the most authoritative since it has 2 authority score. In Suvrat Bhooshan et al. [12], they applied HITS algorithm just only hubs score for each node as a ranking. We will use also similar algorithm for the graph to improve ranking score's quality.

When we extracted graph features, we faced a problem that is about undirected edge scores. In the graph model, we just have directed edge features. Similar to Jure et al. [14] we also will use triad features for the graph. For example, the features between node u and node v would also be calculated with another node w that is between (u,v).

$$(sign(u, w), sign(w, v)) \forall_{w \in V: (u,w) \in E, (w,v) \in E}$$

For example, in two nodes (u,v) if we look directed edges, we just can see two different values that are win or lose from u to v but, when we look undirected connection between u to v, we will add also w node. Moreover, with this connection, we have four different values from u to v that are u beat w w beat v, u beat w w lose v, u lose w w beat v and u lose w w lose v. We will use this triad methods for all pairs of horses and took values for undirected edge scores.

In conclusion, we will use edge weight score , node score and triad feature as graph features in the graph and also we will use HITS algorithm for improving quality of features. We will use these features because we wanted to use directly relationship with horses. Basic features are coming from directly horse racing data so, there is no relationship between horses in horse race data but, in the graph, features are extracting from big picture of horses and there are connections between each horses at least in one or two races. That is why we are using graph features and we will use these features to check improvements of prediction.

## CHAPTER 5

### MACHINE LEARNING MODELS

Machine learning is one of the important areas of Artificial Intelligence that researches algorithms [33]. There are many machine learning methods that can be used for prediction but, in this thesis, we will use two machine learning models for prediction: artificial neural network and logistic regression models. There are many researchers that have published models are using logistic regression. For example, Clarke and Dyte [34] applied a logistic regression model to the difference in the rating points of the two players for predicting the results. Logistic regression uses to predict an outcome feature that is categorical form predictor features that are continuous and categorical. Also a logistic regression model took at seconds in maximum to train. Also we chose the artificial neural network because ANN model development is highly experimental, and the selection of the hyper-parameters of the model usually requires a tests and error approach. Moreover, most of the researchers used the artificial neural network and logistic regression methods for prediction. Also artificial neural networks can discover complex relationships between the various features in races or matches. The ANN can give more challenge instead of logistic regression because when we search it, it takes more time and also the model configuration is more difficult than logistic regression, that is why we chose these two models and will try to compare results.

#### 5.1 Artificial Neural Networks

An artificial neuron network (ANN) is a system based on the interconnected neurons from biological neural networks [35]. Each neuron or it is called nodes, calculates a value from its inputs. Artificial neural networks are typically designed with several layers and each neuron that is in non-input layer are connected all other neurons. Each connection in the network is a weight and neurons uses its inputs and their weights to calculate an output value. In figure 5.1, we can see that four input data come for neural network and processed.

We should train input data before analyzing or testing [35] because neural networks are trained presenting samples to the network again and again. The general architecture of the network

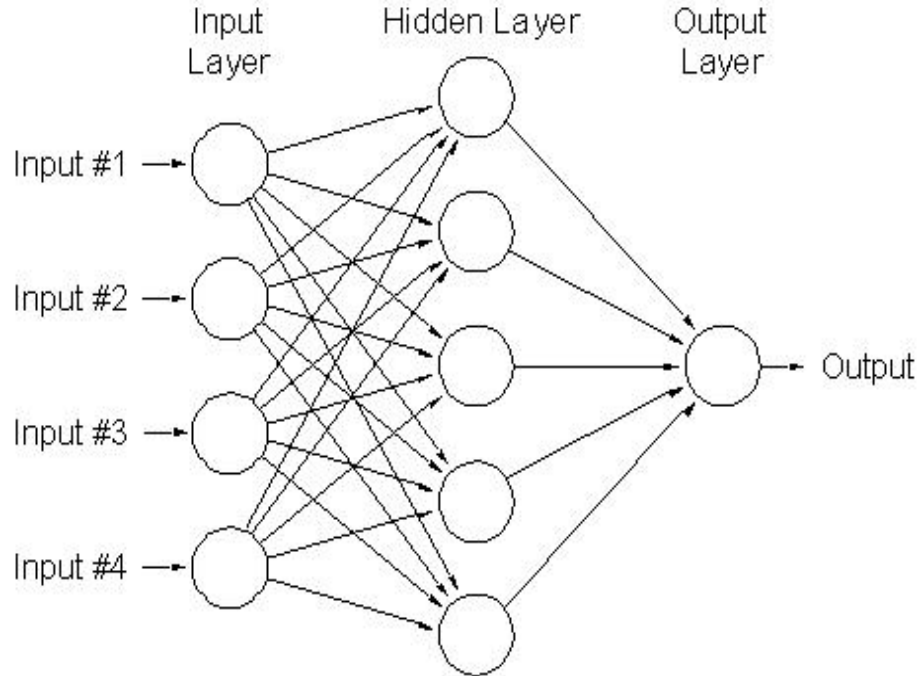


Figure 5.1: Example of Neural Network

is that of a multilayer perceptron (MLP), a feedforward network trained by backpropagation. Each sample contains input data that we put for set conditions and output data that resulting prediction or answer. The network learns each sample and computes the output using input that we provided. If there is an imbalance between network output and actual output, the network corrects itself via exchange its internal connections. This loop process continues until network reached accuracy that we specified. After neural network is trained, it starts to check similarities between new input data and give the result a predicted output data.

In this thesis, following system for artificial neural network is used in figure 5.2: In 2015-2016 years data, we will take two horses with 8 different features in a race and we will put network system for training and we will repeat it until it becomes global achieved. After that, we will pick another two horses and we will train network repeatedly. For the end, we will have trained each horse with all other horses in the network.



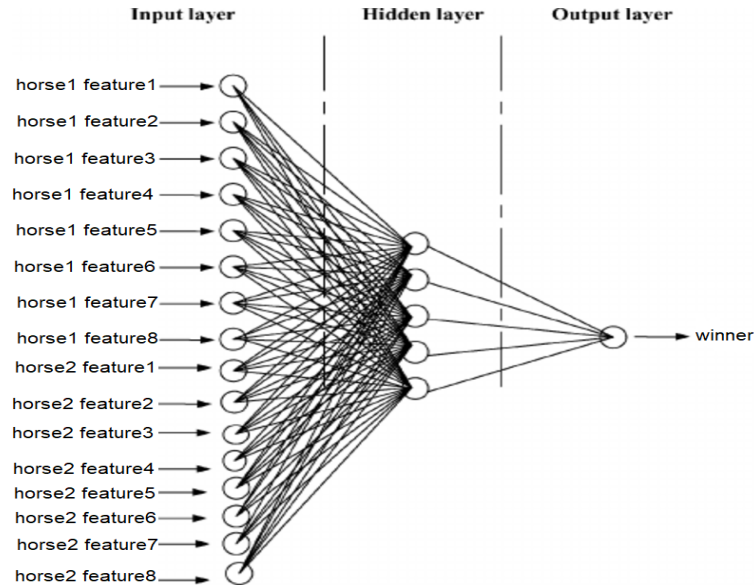


Figure 5.2: Example of Neural Network System

After training part, we will use 2017 data for testing. We will pick 2 horses in same race and we will put these 2 horses and their features in trained network and make a prediction of accuracy compared real race in 2017.

### 5.1.1 Logistic Regression Model

Logistic regression is a classification algorithm which uses the calculated logits (score ) to predict the target class [36]. It transforms its output to return a prediction value which can then be mapped to two or more discrete classes [36]. Generally, when we create a machine learning program, we are trying to arrive a prediction for future inputs based on the experience with using the past inputs and outputs. The logistic function  $\sigma(t)$  is defined as:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

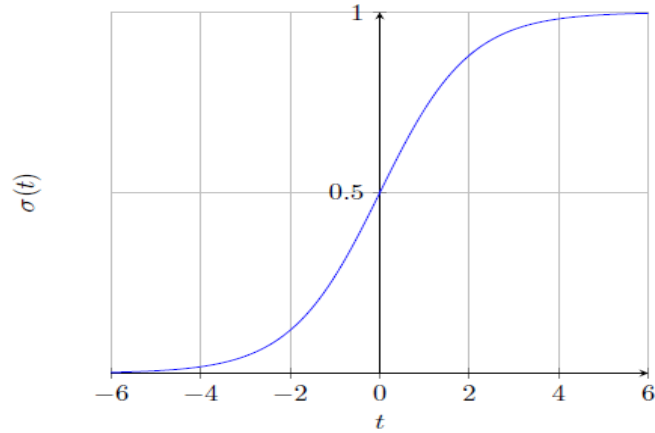


Figure 5.3: Logistic Function

A logistic regression model for race prediction consists of a vector of  $n$  race features [36]  $x = (x_1, x_2, x_3, \dots, x_n)$  and a vector of  $n + 1$  real-valued model parameters  $\beta = (\beta_0, \beta_1, \dots, \beta_n)$  so for prediction, they design a point in  $n$ -dimensional feature space to a real number then they match  $z$  to a value in the admissible range of probability (0 to 1) using the logistic function.

$$z = (\beta_0, \beta_1 x_1, \beta_2 x_2, \dots, \beta_n x_n)$$

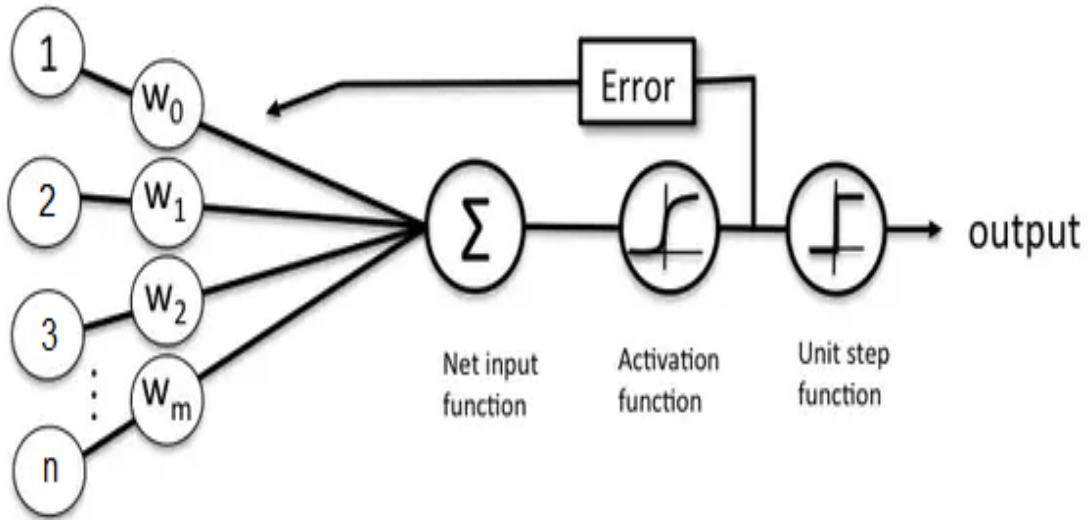


Figure 5.4: Schematic of Logistic Regression

Model occurs of optimizing the parameters  $\beta$  the model gives the best results of outcomes for the training data [36] . In logistic regression, we will use stochastic gradient descent method that is simple but very efficient approach to discriminative learning of linear classifiers and fit for large

datasets, updates parameters in every iteration to minimize the error of a model on training data.

This optimization algorithm works like in [37];

-Each training sample is shown to the model one at a time than the model makes a prediction for a training sample

-Model is updated with errors

- Errors are considered for reducing next prediction.

## CHAPTER 6

### EXPERIMENTAL RESULTS

In experimental results, our goal is to extend basic features with graph-based features, train and test with artificial neural network and logistic regression and compare their results of accuracy between each other. We will compare artificial neural network and logistic regression with basic features, the artificial neural network and logistic regression with additional edge labeled graph-based features, the artificial neural network and logistic regression with additional multiple edge labeled graph-based features, the baseline that is probable betting odds as estimated by a bookmaker and also with previous work that is Elnaz Davoodi et. al [4]. Our hypothesis is to try to obtain better results with additional graph-based based features. In experimental part, firstly we will look some specific dates and compare the results, after that we will compare results of accuracy for each methods and previous works and finally we will try our model in real race that is Kentucky Derby.

When we prepared our model, we need to compare some works and show our improvements. In every race, the bookers or betting operators displayed some values that which horses are most favorite and which horses are less favorite. We tried to compare morning line (baseline) that is in our data, and compare horse's winning probabilities with our model. Also we will compare our results with previous works.

In horse racing, it is difficult to find 2 horses in same race in same year. It was one of the big effort to find 2 horses in same race and these horses should be trained from 2015-2016 years data. Firstly, we used 8 basic features that are directly from horse racing data and implemented artificial neural network and logistic regression. After trained our 2015-2016 years data, we selected in 2017 data with each time 2 horses randomly in same race, compared accuracy results with horse's winning probabilities.

Our important target is, using additional graph features with basic features, make prediction and we want to improve our accuracy of prediction. While our basic features are coming directly from horse racing data, our graph features come a big global picture of horse racing graph. Before we look accuracy of prediction, we will look some specific dates and races and we will show results.

Track Name / Date	Horse Name	Track Type	Weather	Jockey	Past final position	Number of horses in the race	Horse weight	Track condition	Distance
MAHONING VALLEY RACE COURSE 3/7/2017	Majestic Contender	Dirt	Showery	Bryan Metz	3	8	126	Sloppy	1
	Bobbys Approval	Dirt	Showery	Bobby R. Rankin	4	8	120	Sloppy	1
PARX RACING 3/7/2017	Majestic Woody	Dirt	Cloudy	Sanchez, Mychel	7	10	121	Sloppy	1
	Old Sport	Dirt	Cloudy	Rivera, Edwin	6	10	121	Sloppy	1
FAIRMOUNT PARK 6/6/2017	Seeyalaterbye	Dirt	Clear	James, Michael	4	5	124	Fast	0.75
	Lookin for Luv	Dirt	Clear	Molina, Jr., Juan	5	5	121	Fast	0.75
ARAPAHOE PARK 6/11/2017	Twisted Regan	Dirt	Clear	Ziegler, Michael	4	6	124	Fast	1
	Sandhill Lady	Dirt	Clear	Barton, Jake	6	6	121	Fast	1
ARAPAHOE PARK 6/18/2017	Furriddle	Turf	Cloudy	Williams, Carl	4	7	124	Fast	0.2
	Blazin Jim	Turf	Cloudy	Guerra, Vince	3	7	124	Fast	0.2
ALBUQUERQUE 9/8/2017	Allied Kid	Dirt	Clear	Escobar, Victor	2	6	123	Fast	0.87
	Stylish Copy	Dirt	Clear	Garcia, Enrique	6	6	123	Fast	0.87

Figure 6.1: Horse Races Features for Some Specific Days

We can see in the figure 6.1; we picked randomly two pairs of horses in the same race in 2017 data in different dates. For example first two horses raced in Mahoning Valley Race Course in March Seventh 2017 and in dirt track, showery weather, with 8 horses, and 1-mile sloppy track. First horse that name is Majestic Contender 's weight is 126lb, jockey name is Bryan Metz, final position in previous race is 3, second horse that name is Bobbys Approval's weight is 120lb, jockey name is Bobby R. Rankin and final position in previous race is 4. We compared these horses because they were in same races in 2017 data and also 2015 2016 data. When we wanted to test these two pairs of horses, we compared their baseline values that are probable betting odds with our methods.

Track Name / Date	Horse Name	The Baseline (Odds) *Favorite in the baseline	The Real Race Final Position	The ANN with Basic Features	Logistic Regression with Basic Features	The ANN with Graph Features	Logistic Regression with Graph Features	The ANN with Specific Graph Features	Logistic Regression with Specific Graph Features
MAHONING VALLEY RACE COURSE 3/7/2017	Majestic Contender	2.3	3	2	2	1	1	No connection with specific condition	No connection with specific condition
	Bobbys Approval	2.1*	5	1	1	2	2	No connection	No connection
PARX RACING 3/7/2017	Majestic Woody	15	5	2	2	2	1	1	1
	Old Sport	3.7*	7	1	1	1	2	2	2
FAIRMOUNT PARK 6/6/2017	Seeyalaterbye	15.2	4	2	2	1	2	1	1
	Lookin for Luv	9.1*	5	1	1	2	1	2	2
ARAPAHOE PARK 6/11/2017	Twisted Regan	5.2	3	2	2	1	2	1	2
	Sandhill Lady	2.3*	4	1	1	2	1	2	1
BELTERRA PARK 6/18/2017	Furriddle	6.5	2	1	2	1	1	No connection	No connection
	Blazin Jim	4.3*	4	2	1	2	2	No connection	No connection
ALBUQUERQUE 9/8/2017	Allied Kid	8.4*	2	1	2	1	1	1	1
	Stylish Copy	12.5	6	2	1	2	2	2	2

Figure 6.2: Comparing Horse Races Results for Some Specific Days

In figure 6.2, we can see results of the baseline, actual races, the ANN with basic features, logistic regression with basic features, the ANN with graph features, logistic regression with graph features, the ANN with specific graph features and logistic regression with specific graph features. Specific graph features mean for instance, when we looked first two horses previous races and their 2017 races, we saw that they raced same distance (1 mile) and same track type (dirt) so, we added these two conditions in graph edges. We eliminated other edges that are include different conditions in edges about track type and distance. After that we extracted features from these specific graph and. In the baseline, if a horse has lower rate, this mean have a higher winning rate. For example if a horse has 1 odds and other horses has 2 odds, it means first horse winning probability is higher

than second horses. If we look first two horses we can see that first horse (Majestic Contender) wager odds is 2.3 but, second horse (Bobbys Approval) wager odds is 2.1 so, first horse is favorite in this race in the baseline. When we tested these horses, Although with basic features, The ANN and Logistic Regression predicted Bobbys Approval beat Majestic Contender, with graph features, it is opposite so Majestic Contender beat Bobbys Approval. In real race Majestic Contender finished third position and Bobbys Approval finished fifth position. We can say that we predicted correct with graph based features but, The ANN and logistic Regression did not predict correct when we used only basic features.

Before testing many times to our methods, we wanted to look more deeply racing results and we wanted to check more closely at the effects of attributes. That is why we picked two pairs of horses that raced both 2016 and 2017 in same races.

**PARX RACING - December 5, 2016 - Race 1**

MAIDEN CLAIMING - Thoroughbred

(PLUS UP TO 40% PABF) FOR MAIDENS, THREE YEARS OLD AND UPWARD. Three Year Olds, 123 lbs.; Older, 124 lbs.

Claiming Price \$20,000. **Claiming Price:** \$20,000

Six Furlongs On The Dirt **Track Record:** (Iredell - 1:07.21 - December 30, 2017)

**Purse:** \$27,000

**Plus:** \$3,348 PABF - Pennsylvania Bonus Program

**Available Money:** \$30,598

**Value of Race:** \$30,598 1st \$16,200, 2nd \$7,560, 3rd \$4,158, 4th \$1,620, 5th \$810, 6th \$250

**Weather:** Cloudy **Track:** Fast

**Off at:** 12:31 **Start:** Good for all

**EQUIBASE**

 [Video Race Replay](#)

Last Raced	Pgm	Horse Name (Jockey)	Wgt	M/E	PP	Start	1/4	1/2	Str	Fin	Odds
10Nov16 <sup>9</sup> AQU <sup>2</sup>	8	Mr. Neetie (Cintron, Alex)	124	L bf	6	1	1 <sup>2</sup>	1 <sup>4</sup>	1 <sup>10</sup>	1 <sup>16 1/2</sup>	0.20*
18Oct16 <sup>4</sup> MED <sup>4</sup>	3	<b>Spanish Villa</b> (Ocasio, Luis)	118 <sup>5</sup>	L	3	2	3 <sup>Head</sup>	3 <sup>2</sup>	3 <sup>2 1/2</sup>	2 <sup>1</sup>	14.80
24Oct16 <sup>2</sup> PRX <sup>5</sup>	5	<b>Risky Persuasion</b> (Lopez, Charles)	123	L b	5	4	5 <sup>9</sup>	5 <sup>8</sup>	4 <sup>4</sup>	3 <sup>3/4</sup>	15.00
18Nov16 <sup>5</sup> LRL <sup>7</sup>	2	Hennessy Fire (Bisono, John)	123	L	2	5	4 <sup>1</sup>	2 <sup>1 1/2</sup>	2 <sup>1/2</sup>	4 <sup>9</sup>	3.90
28Sep16 <sup>3</sup> PRX <sup>9</sup>	1	Shur Fox (Castillo, David)	123	L	1	6	6	6	6	5 <sup>3</sup>	59.00
13Nov16 <sup>3</sup> PRX <sup>7</sup>	4	Just Junior (Acosta, J.)	124	L bf	4	3	2 <sup>1/2</sup>	4 <sup>1/2</sup>	5 <sup>4</sup>	6	16.70

Figure 6.3: Example of Horse Race Result in 2017

We picked two horses that are Spanish Villa and Risky Persuasion. In figure 6.3, we can see that Spanish Villa (Final position is 2) beat Risky Persuasion (Final Position is 3).

Track Name / Date	Horse Name	Track Type	Weather	Jockey	Past final position	Number of horses in the race	Horse weight	Track condition	Distance
PARX RACING 11/5/2016	Spanish Villa	Dirt	Cloudy	Ocasio, Luis	2	6	118.5	Fast	0.75
	Risky Persuasion	Dirt	Cloudy	Lopez, Charles	3	6	123	Fast	0.75
PARX RACING 1/2/2017	Spanish Villa	Dirt	Showery	Perez, Jorge	1	9	117.5	Muddy	0.75
	Risky Persuasion	Dirt	Showery	Lopez, Charles	2	9	124	Muddy	0.75

Figure 6.4: Features of Two Horses in 2016 and 2017

These two horses also raced in 2017 in same race and their attributes are in the figure 6.4. We can see that weathers are different (cloudy, showery), and number of horses in races are 6 and 9 respectively. Track condition is also different because in 2017, weather was showery so, track condition was muddy in 2017. Risky Persuasion has same jockey in these races but, Spanish Villa has different jockeys. When we tested these two horses, we can see results in figure 6.5.

Track Name / Date	Horse Name	The Baseline (Odds)	The Real Race Final Position	The ANN with Basic Features	Logistic Regression with Basic Features	The ANN with Graph Features	Logistic Regression with Graph Features	The ANN with Specific Graph Features	Logistic Regression with Specific Graph Features
PARX RACING 11/5/2016	Spanish Villa	14.8	1	2	2	1	1	1	2
	Risky Persuasion	15	2	1	1	2	2	2	1
PARX RACING 1/2/2017	Spanish Villa	33.3	2	2	2	2	2	2	2
	Risky Persuasion	5.5	1	1	1	1	1	1	1

Figure 6.5: Comparing Two Horses Results in 2016 and 2017

As we can see in figure 6.5, while Spanish Villa was favorite in the baseline in 2016 (Spanish Villa is 14.8 and Risky Persuasion is 15 ), in 2017, Risky Persuasion was favorite compared to Spanish Villa (Spanish Villa is 33.3 and Risky Persuasion is 5.5). In the real race in 2016 bettors predicted and Spanish Villa beat Risky Persuasion even if it had really close wager odds. On the other hand in 2017 race bettors also predicted this two horses in the baseline. Also we tested 2017 race and when we compared the results, we can see that we also predicted which horse beat other



horse. There are some difference between 2016 and 2017 races but, big differences are weather, track condition, and Spanish Villa's jockey. When we checked Risky Persuasion we can see that this horse is really good when weather is showery and condition is muddy. On the other hand Spanish Villa is really bad in muddy condition. These are the biggest facts so, bettors gave wager odds these two horses really different (Spanish Villa is 33.3 and Risky Persuasion is 5.5). In our methods, we also looked these ingredients and we predicted that Risky Persuasion beat Spanish Villa. In conclusion, we can clearly say that Risky Persuasion is really favorite in showery weather and muddy track.

We also tested these two horses with their 2016 race conditions that are cloudy weather and 6 horses. Track condition and weather is different that is why Risky Persuasion is not really favorite in this race even if in the baseline (15 and 14.8). When we tested our methods, The ANN and Logistic Regression with only basic features predicted Risky Persuasion beat Spanish Villa but, with graph based features, Spanish Villa beat Risky Persuasion. In real race Spanish Villa beat Risky Persuasion with very little difference. So, we predicted correct with graph based features but, with specific graph based features in logistic regression predicted wrong in first time after that we tested again and it predicted correct.

We tested our models 512 times with different two pairs of horses in same race in 2017 data and we will discuss accuracy of prediction results obtained from horse racing data with methodologies.

## **6.1 Using Basic Horse Racing Features to Make Prediction**

Firstly we compared the baseline (probable betting odds), previous work that is Elnaz Davoodi et. al [4] and artificial neural network with basic horse racing features and we looked how many times we will predict winner in our methods.

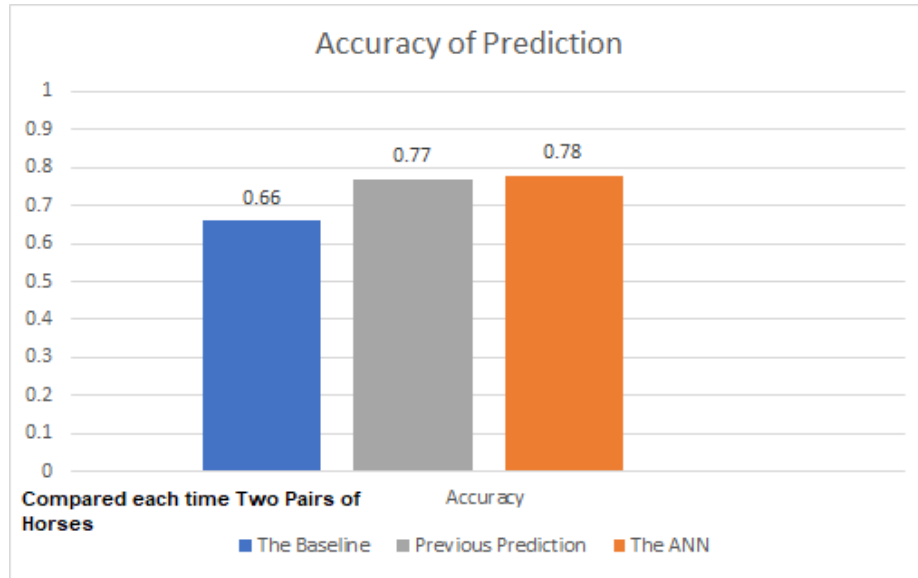


Figure 6.6: Compare Prediction Accuracy With ANN and The Baseline

In figure 6.6, we can see that, while the baseline predicted 338 times and accuracy is 66%, artificial neural network predicted 401 times and accuracy is 78%. Also in previous work that is Elnaz Davoodi et. al. [4] predicted 77% We can say that our artificial neural network system is better than the baseline (probable betting odds) a little better than previous horse racing prediction work.

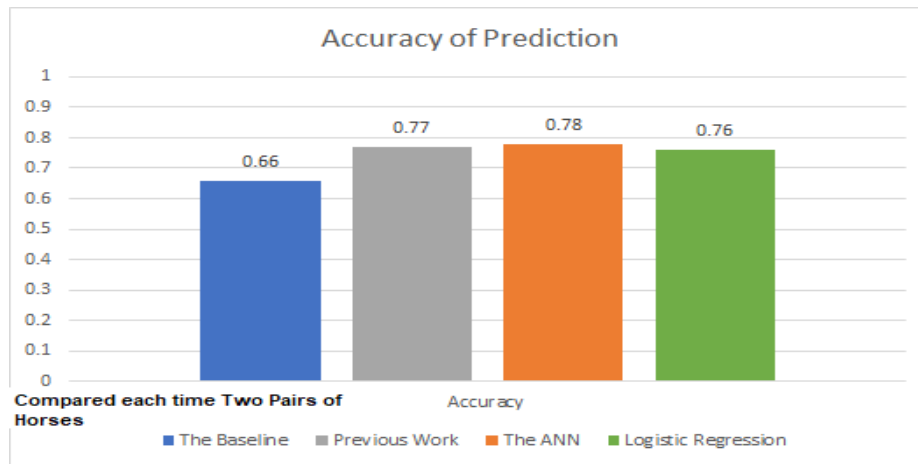


Figure 6.7: Compare Prediction Accuracy With ANN and LR and The Baseline

When we tested with logistic regression model we can say that, logistic regression is better than the baseline for prediction of winner with two horses in same race but, just a little bit less accuracy if we compare the ANN. As can we see in figure 6.7, logistic regression predicted 390 times and accuracy is 76% that is more than the baseline but a little less than the ANN. Also logistic

regression model has almost similar accuracy value with previous work.

## 6.2 Using Additional Graph Features to Make Prediction with ANN

In this section, we added different graph features such as HITS, edge weight difference and nodes degree difference in our approaches with basic features and we will compare the results.

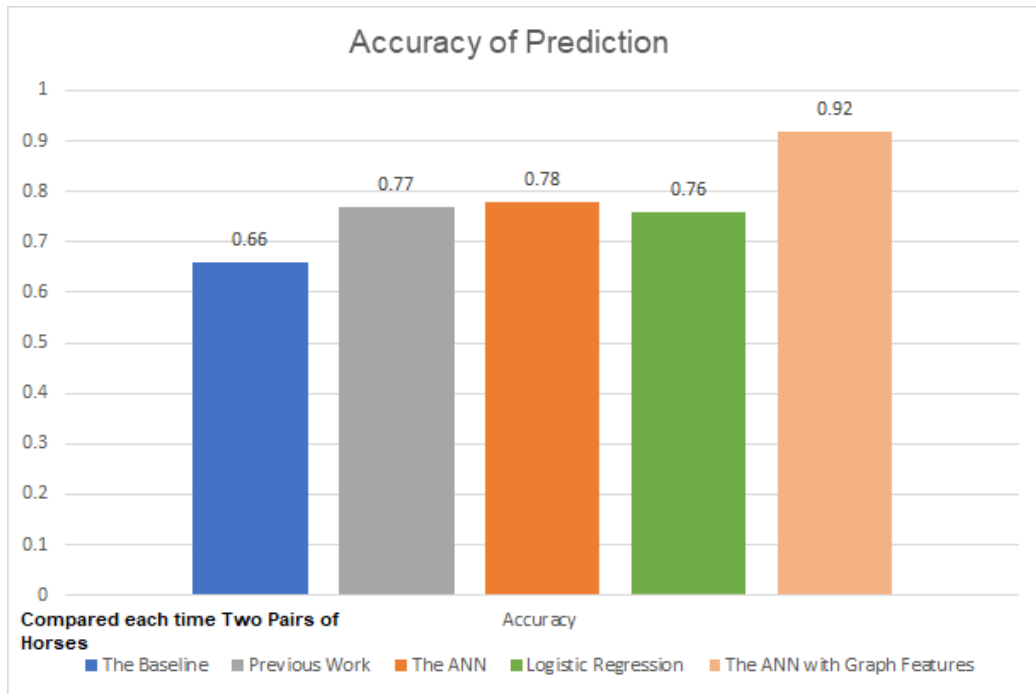


Figure 6.8: Compare Prediction Accuracy With Additional Graph features ANN

When we used additional features from global graph in the artificial neural network, we saw big improvement than we used only basic features. The ANN predicted with graph features more than using only basic features(472 times) and accuracy is 92% in figure 6.8.

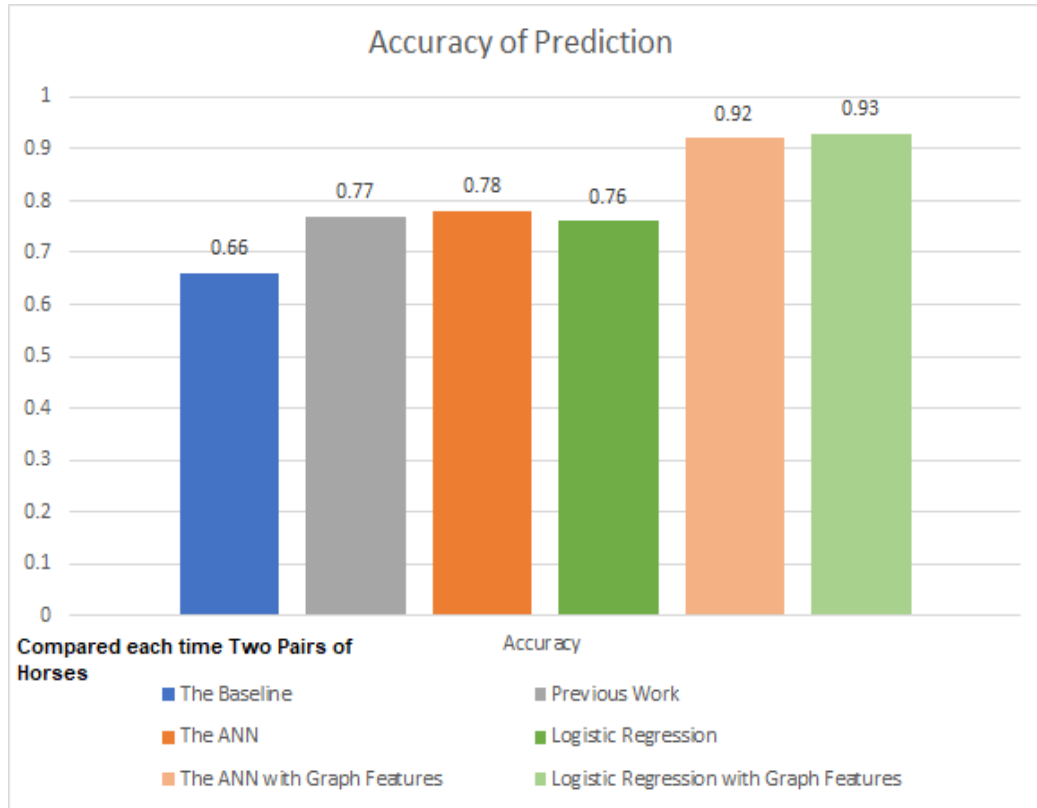


Figure 6.9: Compare Prediction Accuracy With Additional Graph features ANN and LR

If we applied also in logistic regression methodology in figure 6.9, we can say that The ANN and logistic regression are almost similar prediction, but they have really big improvement according to previous results that we applied with basic features. Figure 6.4 shows that, logistic regression predicted 476 times and accuracy is 93%. So we can say that, if we add graph features that are relationship between horses, accuracy of prediction will be better than previous results.

### 6.3 Using Additional Specific Graph Features to Make Prediction with ANN

We created graph more specifically. It means we added some conditions to edges in our graph and we eliminated other edges that include different conditions so, we extracted features from directly this graph. We picked 2 conditions such as weather and distance (it depends on the races). In our 2015-2016 and 2017 data, we just found 202 races in same two conditions and 2 horses in same races . For this reason, we just put 202 times two pairs of horses in our approaches and compare the results.

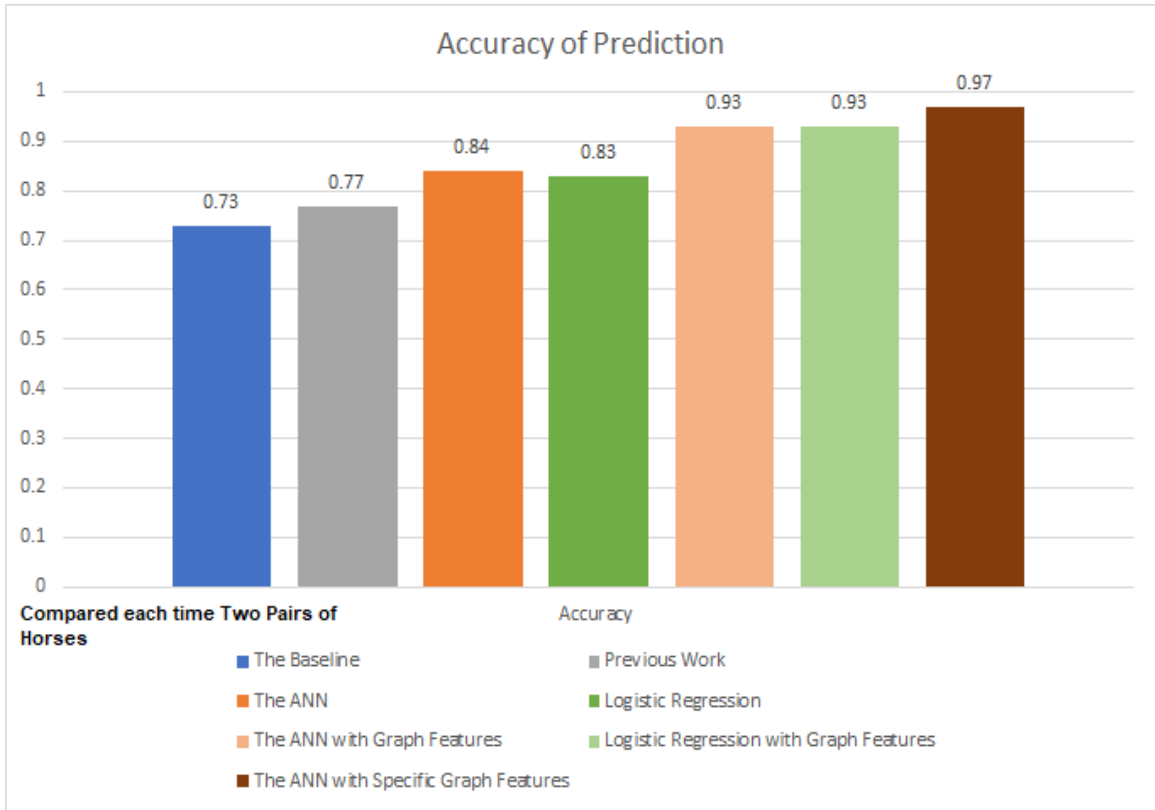


Figure 6.10: Compare Prediction Accuracy With Additional Specific Graph features ANN

When we compared the results we saw that, the baseline predicted winner 148 times and accuracy is %73. In basic features, the artificial neural network and logistic regression predicted almost similar time that 170 times the ANN and 167 times logistic regression and accuracy is 84% and 83% respectively. In addition, with additional graph features, the ANN predicted 188 times and logistic regression predicted 186 times and accuracy are same %93. When we add specific graph features in our approaches with basic features, we saw that the ANN predicted 196 times and just only 6 times it didn't predict and accuracy is 97% in figure 6.10

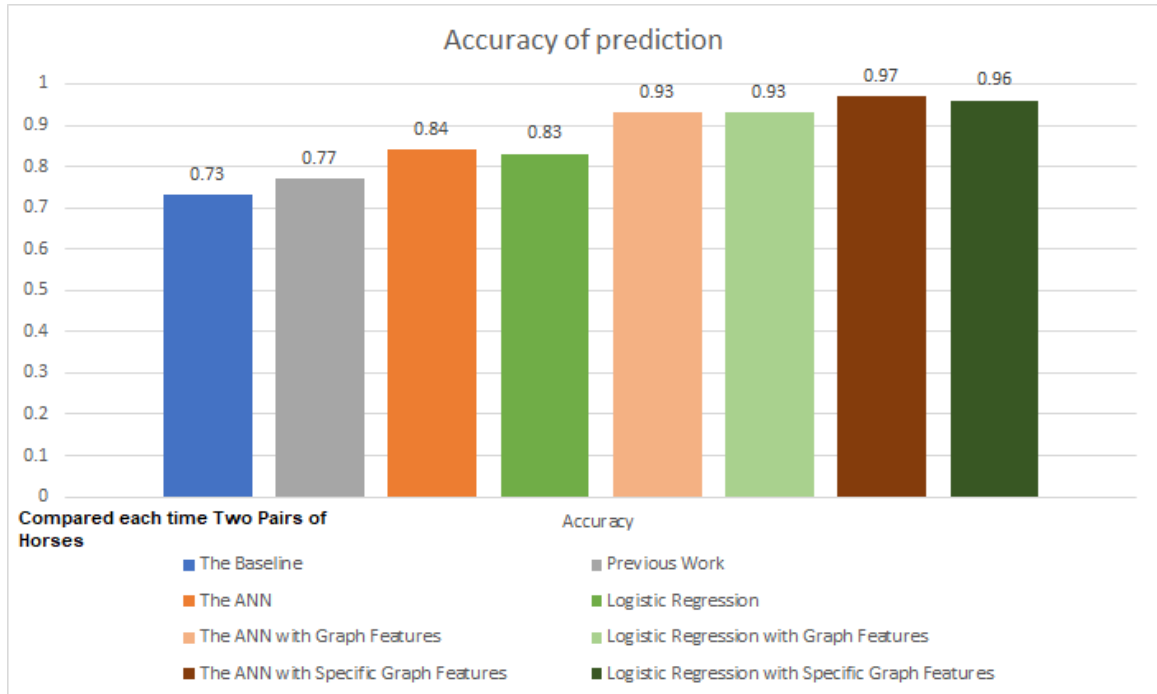


Figure 6.11: Compare Prediction Accuracy With Additional Specific Graph features ANN and LR

As can we see in figure 6.11, when we tested logistic regression for these specific features and basic features, logistic regression predicted winner 195 times and 7 times prediction was wrong. Accuracy is 96%. **These values are really good and high but, we compared each time two pairs of horses not whole races that is why it can be high.**

As a conclusion, we showed that, using additional graph features that are relationship between horses are made better accuracy than using only basic features that are extract directly horse racing data. Moreover, when we created our graph with specific edge weight conditions, extracted features from this graph and apply our approaches we had taken better results than normal using additional graph features. In previous works that are Elnaz Davoodi et. al. [4], we can see that accuracy of winner prediction is 77% and and in prediction horse racing results [11] win place show accuracy of prediction is 83%. These values are almost similar our prediction with basic features, but with graph based features our prediction accuracy values are better than previous works and previous results.

## 6.4 Challenge for Kentucky Derby

The Kentucky Derby is a horse race that is organized annually (since 1875) in state of Kentucky, United States. [38] The race is held first weekend of May at a distance of 1 and 1/2 miles at Churchill Downs in city of Louisville. Most of the people called this race "The Run for the Roses" [39] because wreath of roses is draped over the winner. Also the race known as "The Fastest Two Minutes in Sports" [40] for its approximate duration.

We wanted to test our methods in Kentucky Derby and we compared to result with Kentucky Derby 2017. In figure 6.12 shows that horses and their odds in the morning line and post position in Kentucky Derby 2017.

2017 Kentucky Derby Odds		
HORSE	POST	ODDS
Always Dreaming	5	9-2
Irish War Cry	17	9-2
McCraken	15	6-1
Classic Empire	14	7-1
Gunnevera	10	10-1
Patch	20	12-1
Hence	8	16-1
Thunder Snow	2	16-1
Gormley	18	21-1
Girvin	7	22-1
Tapwrit	16	26-1
Practical Joke	19	30-1
Lookin at Lee	1	31-1
Sonneteer	12	36-1
Battle of Midway	11	37-1
Fast and Accurate	3	37-1
Irap	9	38-1
J Boys Echo	13	46-1
State of Honor	6	51-1
Untrapped	4	57-1

Figure 6.12: 2017 Kentucky Derby Horses

In figure 6.12 we can see that Always Dreaming, Irish War Cry, McCraken and Classic Empire are the most favorite horses in this race then other horses follow these horses. Kentucky Derby is really famous horse race in United State. That is why many experts and bettors try to predict result of races. In figure 6.13 we can see some predictions of experts for The Kentucky Derby 2017. They tried to predict first 3 horses.

In model that are created, we wanted to compare each time two pair of horses we tried to find a ranking. Each time we compared two horses in our methods and find a winner and we found a ranking.

EXPERT	1st	2nd	3rd
Joe Kristufek	Always Dreaming	McCraken	Classic Empire
John Asher	McCraken	Classic Empire	Always Dreaming
Jill Byrne	McCraken	Practical Joke	Irish War Cry
Ed DeRosa	Classic Empire	Irish War Cry	Hence
James Scully	Always Dreaming	Practical Joke	McCraken
Kellie Reilly	McCraken	Irish War Cry	Classic Empire

Figure 6.13: Predictions of Experts

In our model we wanted to compare each time two pair of horses we tried to find a ranking. Each time we compared two horses in our methods and find a winner and we found a ranking. Example of how we compared horses is in figure 6.14 that are in below;

Horse Name	Lookin at Lee	Thunder Snow	Fast and Accurate	Untrapped	Always Dreaming	State of Honor	Girvin	Hence	Irap	Gunnevera	Battle of Midway	Sonneteer	J Boys Echo	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Lookin at Lee	X	Thunder Snow	Lookin at Lee	Lookin at Lee	Always Dreaming	Lookin at Lee	Lookin at Lee	Hence	Lookin at Lee	Gunnevera	Lookin at Lee	Lookin at Lee	Lookin at Lee	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Thunder Snow	Thunder Snow	X	Thunder Snow	Thunder Snow	Always Dreaming	Thunder Snow	Thunder Snow	Thunder Snow	Thunder Snow	Thunder Snow	Thunder Snow	Thunder Snow	Thunder Snow	Classic Empire	McCraken	Thunder Snow	Irish War Cry	Thunder Snow	Thunder Snow	Thunder Snow
Fast and Accurate	Lookin at Lee	Thunder Snow	X	Fast and Accurate	Always Dreaming	Fast and Accurate	Girvin	Hence	Fast and Accurate	Gunnevera	Battle of Midway	Sonneteer	J Boys Echo	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Untrapped	Lookin at Lee	Thunder Snow	Fast and Accurate	X	Always Dreaming	Untrapped	Girvin	Hence	Irap	Gunnevera	Battle of Midway	Sonneteer	J Boys Echo	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	X	Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	Always Dreaming	Classic Empire	McCraken	Always Dreaming	Irish War Cry	Always Dreaming	Always Dreaming	Always Dreaming
State of Honor	Lookin at Lee	Thunder Snow	Fast and Accurate	Untrapped	Always Dreaming	X	Girvin	Hence	Irap	Gunnevera	Battle of Midway	Sonneteer	J Boys Echo	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Girvin	Lookin at Lee	Thunder Snow	Girvin	Girvin	Always Dreaming	Girvin	X	Hence	Girvin	Gunnevera	Girvin	Girvin	Girvin	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Hence	Hence	Thunder Snow	Hence	Hence	Always Dreaming	Hence	Hence	X	Hence	Gunnevera	Hence	Hence	Hence	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Hence	Patch
Irap	Lookin at Lee	Thunder Snow	Fast and Accurate	Irap	Always Dreaming	Irap	Girvin	Hence	X	Gunnevera	Battle of Midway	Sonneteer	J Boys Echo	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Gunnevera	Gunnevera	Thunder Snow	Gunnevera	Gunnevera	Always Dreaming	Gunnevera	Gunnevera	Gunnevera	Gunnevera	X	Gunnevera	Gunnevera	Gunnevera	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Gunnevera	Patch
Battle of Midway	Lookin at Lee	Thunder Snow	Battle of Midway	Battle of Midway	Always Dreaming	Battle of Midway	Girvin	Hence	Battle of Midway	Gunnevera	X	Battle of Midway	Battle of Midway	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Sonneteer	Lookin at Lee	Thunder Snow	Sonneteer	Sonneteer	Always Dreaming	Sonneteer	Girvin	Hence	Sonneteer	Gunnevera	Battle of Midway	X	Sonneteer	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
J Boys Echo	Lookin at Lee	Thunder Snow	J Boys Echo	J Boys Echo	Always Dreaming	J Boys Echo	Girvin	Hence	J Boys Echo	Gunnevera	Battle of Midway	Sonneteer	X	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	Practical Joke	Patch
Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire	X	McCraken	Classic Empire	Classic Empire	Classic Empire	Classic Empire	Classic Empire
McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	McCraken	X	McCraken	McCraken	McCraken	McCraken	McCraken
Tapwrit	Tapwrit	Thunder Snow	Tapwrit	Tapwrit	Always Dreaming	Tapwrit	Tapwrit	Tapwrit	Tapwrit	Tapwrit	Tapwrit	Tapwrit	Tapwrit	Classic Empire	McCraken	X	Irish War Cry	Gormley	Practical Joke	Patch
Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Irish War Cry	Classic Empire	McCraken	Irish War Cry	X	Irish War Cry	Irish War Cry	Irish War Cry
Gormley	Gormley	Thunder Snow	Gormley	Gormley	Always Dreaming	Gormley	Gormley	Gormley	Gormley	Gormley	Gormley	Gormley	Gormley	Classic Empire	McCraken	Gormley	Irish War Cry	X	Gormley	Patch
Practical Joke	Gormley	Thunder Snow	Practical Joke	Practical Joke	Always Dreaming	Practical Joke	Practical Joke	Hence	Practical Joke	Gunnevera	Practical Joke	Practical Joke	Practical Joke	Classic Empire	McCraken	Tapwrit	Irish War Cry	Gormley	X	Patch
Patch	Patch	Thunder Snow	Patch	Patch	Always Dreaming	Patch	Patch	Patch	Patch	Patch	Patch	Patch	Patch	Classic Empire	McCraken	Patch	Irish War Cry	Patch	Patch	X

Figure 6.14: Example of Model Prediction

As we can see in figure 6.14, we compared each time two horses and we found a winner. We tested methods for each horses and compared all of the horses in the race and we found rankings for each methods in table that are in below;



Position	Real Results	The Baseline	The ANN	LR	The ANN with Graph features	LR with Graph features
1	Always Dreaming	Always Dreaming	McCraken	McCraken	McCraken	McCraken
2	Lookin At Lee	Irish War Cry	Classic Empire	Classic Empire	Classic Empire	Classic Empire
3	Battle of Midway	McCraken	Irish War Cry	Irish War Cry	Always Dreaming	Always Dreaming
4	Classic Empire	Classic Empire	Always Dreaming	Always Dreaming	Irish War Cry	Irish War Cry
5	Practical Joke	Gunnevera	Thunder Snow	Patch	Tapwrit	Tapwrit
6	Tapwrit	Patch	Patch	Gunnevera	Gunnevera	Patch
7	Gunnevera	Hence	Gormley	Thunder Snow	Gormley	Thunder Snow
8	McCraken	Thunder Snow	Tapwrit	Gormley	Thunder Snow	Gunnevera
9	Gormley	Gormley	Gunnevera	Tapwrit	Practical Joke	Gormley
10	Irish War Cry	Girvin	Hence	Hence	Hence	Hence

Thunder Snow (DNF) No Connection for Specific graph-based features

Figure 6.15: Comparing Results in Kentucky Derby 2017

As we can see in the figure 6.15, Although McCraken finished 8th position in real result, it finished 1st position in our approaches every time. Also Classical Empire was big disappointment in real race. Even though It finished 4th position in real race, it has never lost 2nd position in our methods. **Lookin at Lee made and Battle of Midway** made big surprise in real race even if they didn't get into the first ten in our results. As a conclusion, although we didn't predicted real results but, at least we predicted favorite horses in our methods because when we looked McCraken we can see that it was the most favorite horse in this race because it was very familiar and successful at Churchill Downs when we checked previous races, and also McCraken's jockey whose name is Brian Hernandez, Jr. was really successful in previous races. These contributions and with some other attributes, our methods predicted McCraken as winner. When we compared with the baseline we can see that we almost predicted first 5th position like the baseline with graph based features and other positions are depends on our methods. **if we look results table, Always Dreaming is first, Irish War Cry is second, McCraken is third and Classic Empire is forth in the baseline. In The ANN with graph-based features, we did not predict Always Dreaming but, we predicted Classic Empire beat Irish War Cry, instead of the baseline Tapwrit is in the first six position, Gunnevera and Practical Joke are first ten position.** 2017 Kentucky Derby had one of the big surprise results. Many bettors and experts did not predict results. We did not predict really healthy because our methods tested each time two pairs of horses but, our results are sometimes correct when we compared with the baseline and sometimes correct

when we compared with real results.

When we tested 2017 Kentucky Derby and we did not good healthy results we applied models in 2016 Kentucky Derby Race horses. We applied graph-based features with basic features and we tested our methods. For training part, we only used 2015 horse race data.

Position	The Baseline	Real Results	The ANN with Graph features	LR with Graph features
1	Nyquist	Nyquist	Nyquist	Nyquist
2	Exaggerator	Exaggerator	Exaggerator	Exaggerator
3	Mohaymen	Gun Runner	Gun Runner	Gun Runner
4	Gun Runner	Mohaymen	Mohaymen	Mohaymen
5	Brody's Cause	Suddenbreakingnews	Destin	Destin
6	Destin	Destin	Suddenbreakingnews	Brody's Cause
7	Mo Tom	Brody's Cause	Brody's Cause	Suddenbreakingnews
8	Suddenbreakingnews	Mo Tom	Mo Tom	Mo Tom
9	Mor Spirit	Lani	Mor Spirit	Creator
10	Outwork	Mor Spirit	Creator	Mor Spirit

Figure 6.16: Comparing Results in Kentucky Derby 2016

As we can see in the figure 6.16, our results are better than 2017 Kentucky Derby. We predicted top four positions correctly when we compared with real results. Unfortunately we did not predict Suddenbreakingnews's position. In The ANN with graph-based features, it is in 6th position and logistic regression with graph-based features, it is in 7th position. Although Brody's Cause beat Destin in the baseline, in our results and real results, Destin beat every time Brody's Cause. In real results Lani made big surprise and finished 9th position. In our results, it did not finish top ten. Even if some difference with real results, we predicted almost correct with our methods.

## CHAPTER 7

### CONCLUSIONS AND FUTURE WORK

#### 7.1 Conclusions

In this thesis, we presented horse racing prediction using additional graph based features and compared the results. We used horse racing data from 2015 to 2016 for training and 2017 for testing and we used two different methodologies that are artificial neural network and logistic regression. We trained methods from 2015-2016 data and tested in same races from 2017 data.

Besides applying of machine learning methods, it was hard to prepare horse racing data. We collected, parsed, preprocessed horses racing data from website and we made it ready for training and testing. We tried to take each time 2 pairs of horses in same races and compared their winning probabilities. In horse racing, it is really hard to find similar horses in different years in same races. We could not take results with more data in but, our results were satisfactory.

Firstly, we applied methodologies for basic features that are from horse racing data and compared the results with the morning line (baseline). We reached that machine learning approaches gave better accuracy than the betting operator's odds. While machine learning approaches predicted around with 77%, the baseline predicted 66%.

After finished prediction with basic features, we created a graph includes horses and their relationship in edges as win-loss spread. We extract features from this global horse graph and we added our system with basic features for prediction. Although with only basic features, the artificial neural network and logistic regression predicted winner approximately 77% , with additional graph features, these accuracies went through the roof and reached nearly 93%.

When we tried to graph features in our approaches we saw that accuracy of prediction is better. That is why we used more strategy that is using specific graph features. We used some conditions in edges of graph and we extract features from graph with these informations. After training, we tested and results were better than previous results. In both of our approaches, we almost predicted which horse will win in two pairs of horses. Accuracies are around 96% percent for both artificial neural network and logistic regression methods.

Finally, we tested our methods in 2017 Kentucky Derby and we compared our results with real race results and the baseline. Although we did not predict correctly when we compared with real race, we predicted similar correct when we compared with the baseline. After that we applied our methods in 2016 Kentucky Derby. When we checked the results, we predicted almost similar with real results and our results were better than the baseline.

In a nutshell, we demonstrated that, machine learning approaches give more successful results than bookmaker's probable odds. Although betting operators have to investigate and determine the probabilities, machine learning approaches gave better probabilities.

## **7.2 Limitations and Future Work**

### **7.2.1 Limitations**

In this thesis , we had some problems with data set such as not enough past performance of horses, and not much comparison. Some horses didn't have enough past performances as much as we need. For example, some horses had only one or two past races and these informations were sometimes. Also for training we used 2015-2016 data and for testing we used 2017 horse racing data and the most of time it is hard to find horses 2017 and 2015-2016 data at the same time.

### **7.2.2 Other Machine Learning Algorithms**

We focused our efforts on two machine learning algorithms: logistic regression and artificial neural network. If we use other approaches such as support vector machine, we may take better results. Essentially, support vector machine sometimes have greater accuracy than artificial neural network. It is important to note that while support vector machine needs calibration step for prediction, logistic regression and artificial neural network do not need it. Also, Bayesian networks may use for prediction for horse racing prediction

### **7.2.3 Additional Features**

We selected features based on horse races results and previous works, but we can also use some methods for selecting features. We used 8 basic features graph features. Additional features also may be used such as value of races, medication that uses on horse, Comments on each horse in race and available money. All of our features represent quality of horses and condition of race. Hence, we can also use feature of jockey like weight.

## REFERENCES

- [1] "<http://www.britannica.com/ebchecked/topic/272329/horse-racing>," .
- [2] Ben Nighthorse Campbell, *National Gambling Impact Study Commission Final Report: Congressional Hearing*, DIANE Publishing, 2001.
- [3] "<https://www.betfirm.com/how-much-is-bet-on-the-kentucky-derby/>," .
- [4] Ali Reza Khanteymoori Elnaz Davoodi, "Horse racing prediction using artificial neural networks. NN'10/EC'10/FS'10 Proceedings of the 11th WSEAS international conference on neural networks and 11th WSEAS international conference on evolutionary computing and 11th WSEAS international conference on fuzzy systems pages 155-160 . Iasi, Romania June, 2010," .
- [5] Janett Williams and Yan Li, "A case study using neural networks algorithms:horse racing predictions in Jamaica. In International Conference on Artificial Intelligence, Las Vegas, NV, 2008. ," .
- [6] Michael Pardee, "An artificial neural network approach to college football prediction and ranking," Technical Paper. Madison, WI: University of Wisconsin, Electrical and Computer Engineering Department, 1999.
- [7] Robert P. Schumaker, "Using SVM Regression to predict harness races: A one year study of northfield park," Computer and Information Science's Department Cleveland State University, Cleveland , Ohio 44115, USA 2011.
- [8] "Schumaker, Robert P. and Johnson, James W. " An Investigation of SVM Regression to Predict Longshot Greyhound Races," Communications of the IIMA: Vol. 8 : Iss. 2 , Article 7. available at: <http://scholarworks.lib.csusb.edu/ciima/vol8/iss2/7>," 2008.
- [9] Linlin She Siunie Sutjahjo Chris Sommer Daryl Neely Hsinchun Chen, Peter Buntin Rinde, "Expert prediction, symbolic learning, and neural networks: An experiment on greyhound racing, IEEE Intelligent Systems, vol. 9, no. 6, pp. 21-27," Journal, Dec 1994.
- [10] Michael Baulch, "Using machine learning to predict the results of sporting matches Baulch, M. "Using Machine Learning to Predict the Results of Sporting Matches". Department of Computer Science and Electrical Engineering, University of Queensland, 2001. ," .
- [11] "<https://github.com/dominicplouffe/horseracingprediction>," .
- [12] Wayne Lu Suvrat Bhooshan, Josh King, "Edge direction prediction of sporting tournaments graphs, Stanford University, Final Project," 2016.
- [13] Zimo Yang Guo, Fangjian and Tao Zhou, "Predicting link directions via a recursive subgraph-based ranking Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, PR China, 2013.," .
- [14] Daniel Hunttenlocher Jon Kleinberg Leskovec, Jure, "Predicting positive and negative links in online social network. Proceedings of the 19th international conference on World wide web, ACM," Raleigh, North Carolina, USA April 26 - 30, 2010.
- [15] "<https://www.techopedia.com/definition/14650/data-preprocessing>," .

- [16] "<https://www.equibase.com/premium/eqbracechartcalendar.cfm?sap=tn>," .
- [17] Sethabhisha Naidu Nerusu, "Master project: Parsing and data mining final project on horse races data set," University of Louisville, 2016.
- [18] "<https://www.lifewire.com/what-are-markup-languages-3468655>," .
- [19] "<https://www.britannica.com/topic/graph-theory>," .
- [20] "<http://bilgisayarkavramlari.sadievrenseker.com/category/graf-teorisi/page/2/>," .
- [21] "<http://www.cs.yale.edu/homes/aspnes/pinewiki/graphtheory.html>," .
- [22] David Goodwin, "An Introduction to Graph Theory CIS008-2 Logic and Foundations of Mathematics University of Bedfordshire Friday 17th February 2012," .
- [23] "<http://mathworld.wolfram.com/undirectedgraph.html>," .
- [24] Robert Sedgewick and Kevin Wayne, "Textbook: Algorithms, 4th Edition Mar 19, 2011," .
- [25] Reinhard Diestel, "Graph Theory, (3rd ed.), Electronic Edition 2000, Springer-Verlag New York 1997, 2000," .
- [26] "<http://ceadserv1.nku.edu/longa//classes/mat385/resources/docs/matrix.html>," .
- [27] Martin Grandjean, "A social network analysis of Twitter: Mapping the digital humanities community, Research Article 15 April 2016," .
- [28] George B. Mertzios and Walter Unger, "The Friendship Problem on Graphs, Department of Computer Science RWTH Aachen University 52056 Aachen, Germany, ROGICS'08 12-17 May 2008, Mahdia, Tunisia Relations, Orders and Graphs:Interaction with Computer Science," .
- [29] Roger C. Vergin and John J. Sosik, "No Place Like Home: An Examination of the Home Field Advantage in Gambling Strategies in NFL Football, Journal of Economics and Business 1999; 51:2131 1999 Elsevier Science Inc., New York, New York," .
- [30] "<https://reality.ai/it-is-all-about-the-features/>," .
- [31] Ethem ALPAYDIN, "Introduction to Machine Learning, 3rd Edition, The MIT Press, September 2014," .
- [32] Tetsuo Hayamizu Toru Ishida Saeko Nomura, Satoshi Oyama, "Analysis and Improvement of HITS Algorithm for Detecting Web Communities, Conference Paper in Systems and Computers in Japan, February 2002," .
- [33] Michal Sipko, "Master thesis: Machine learning for the prediction of professional tennis matches," Imperial College London, 2015.
- [34] S. R. Clarke and D. Dyte., "Using official ratings to simulate major tennis tournaments. International Transactions in Operational Research, 7(6):585594, 2000," .
- [35] "[http://www.marcom-china.com/english/mva\\_neural.html](http://www.marcom-china.com/english/mva_neural.html)," .
- [36] "<http://ml-cheatsheet.readthedocs.io/en/latest/logistic-regression.html>," .
- [37] "<https://machinelearningmastery.com/implement-logistic-regression-stochastic-gradient-descent-scratch-python/>," .
- [38] "Tenth race churchill may 1, 2004, daily racing forum," 2004.
- [39] Katherine Mooney., "The Kentucky Derby: How the Run for the Roses Became America's Premier Sporting Event. By James C. Nicholson. (Lexington: University Press of Kentucky, 2012.)," .
- [40] "[http://gotolouisville.s3.amazonaws.com/cms/4239/fall\\_fun](http://gotolouisville.s3.amazonaws.com/cms/4239/fall_fun)," .

## CURRICULUM VITAE

**NAME:** Mehmet Akif Gulum

**ADDRESS:** Computer Engineering & Computer Science Department  
Speed School of Engineering  
University of Louisville  
Louisville, KY 40292

**EDUCATION:**

M.Sc., Computer Science & Engineering

May 2018

**University of Louisville, Louisville, Kentucky**

B.Eng., Computer Engineering

September 2014

**Erciyes University, Kayseri, Turkey**

**PROJECTS AND INTERNSHIPS**

1. Master Thesis: Horse Racing Prediction Using Graph Based Features April 2018
2. Bachelor Degree Final Project: Secret Tracking Systems for Mobile Systems in Android
3. Internship: Inonu University Computer Engineering Department Malatya/Turkey 2013
4. Internship: Novalab Technology Istanbul/Turkey. 2012