УДК 004.4

Amanuel Mehretab Zeikos,
Student,
Department "Big Data Analytics and Video Analysis Methods",
Engineering School of Information Technologies,
Telecommunications and Control Systems,
Ural Federal University named after the first President of Russia B.N. Yeltsin
Yekaterinburg, Russian Federation

Jorge Martinez,
Student,
Department "Big Data Analytics and Video Analysis Methods",
Engineering School of Information Technologies,
Telecommunications and Control Systems,
Ural Federal University named after the first President of Russia B.N. Yeltsin
Yekaterinburg, Russian Federation

Al-Lami Mustafa Ali,
Student,
Department "Big Data Analytics and Video Analysis Methods",
Engineering School of Information Technologies,
Telecommunications and Control Systems,
Ural Federal University named after the first President of Russia B.N. Yeltsin
Yekaterinburg, Russian Federation

DATA-DRIVEN APPROACHES TO DETECTING SPORTS BETTING IRREGULARITIES

Abstract:
Today, people can place their bets from the commodity of their couches by using their cell phones. This means a person can technically place a bet wherever they are and whenever they want. When you have access to a betting website you can bet on any sport event you can imagine every day and all at times of the day. According to the blog “A Football report” the sports betting industry is worth 3 trillion dollars. For this project, we are going to focus on the most popular sport in the world, Football.

Keywords:
Anomaly Detection, Betting, Sports, Data-Drivenю

Introduction

Betting on sporting events is extremely popular world-wide. Sports betting is one of the most profitable businesses. Technology has made this business much bigger. Today, people can place their bets from the commodity of their couches by using their cell phones. This means a person can technically place a bet wherever they are and whenever they want. When you have access to a betting website you can bet on any sport event you can imagine every day and all at times of the day. According to the blog “A Football report” the sports betting industry is worth 3 trillion dollars. These number includes both illegal and legal betting. For this project, we are going to focus on the most popular sport in the world, Football, which is the most played and most watched sport on the world. For these reasons, it is no surprise that it is also one of the most popular sports for betting.

Literature Review

In the area of betting and gambling in general, there are tendencies of individuals and bookmakers wanting to cheat the system and as such it is important to identify these irregularities.

Anomaly detection is the process of identifying unexpected items or events in datasets, which differ from the norm [1]. It has been applied in data types such as textual data, images, computer network data, web clicks, census data, crime data, e-commerce transactions, banking data, energy and power transmission data, and many more [2]. According to the researchers, anomaly detection's importance is as a result of the fact that anomalies that exist in data translate to significant (mostly critical) actionable information in a wide variety of application domains. Hence, our research aims at proposing techniques for irregularity detection within sports betting to provide bookmakers and stakeholders with the necessary tools to prevent financial loss.

Methodology

Bets on football matches can be quite simple. A bet for a match result has only 3 options, home win, tie (draw) or away win. Many sites offer other options as betting for how many corner quicks or yellow cards will be during a game. However, we will only focus on the match result aspect. A dataset provided by Indatabet (*the Soccer_All-in-One dataset*) containing betting odds from English Premier League (EPL) with 380 observations and 118 variables (which we narrowed down to 72 variables from BET 365 necessary for this study). This includes matches; scores; Draw No Bet (DnB); Double Chance (DC); Home Win, Draw and Away Win (1X2) odd lines (Odds block and Double Chance Odds block) with opening and closing odds for full time, first half and second half for games between September 2018 to May 2019 [3].

The EPL was sampled out as our choice league due to it being the top tier. According to the football database website, The English Premier League has 2 teams in the top 5 of the world club ranking. Moreover, English football league is the highest revenue generating league in European football grossing approximately 2.1 billion euros as of 2019 and is the most watched football league in the [4]. From records, the EPL is reported to have had a peak audience of 3.24 million people in the United Kingdom alone in 2017/2018 [5]. Furthermore, the 2018- 2019 season had a global audience of 3.2 billion people. Thus, indicating the superiority of the league and its relevance with regards to understanding betting anomalies.

The analytics software used was R (via the RStudio IDE) as well as the anomalize [6]. Package to perform the anomaly detection tasks. The anomalize package allows the user to decompose time series, detect anomalies in the dataset and create bands separating the non-anomalous data from the anomalous spikes [7]. The anomalize library/package is used on datasets which are typically characterized by inherent or erratic seasonality trends [8]. The Anomalize library uses the Seasonal Hybrid ESD (S-H-ESD) algorithm to compensate for these trends.

Thus, for our study, we employed the data, tools and techniques indicated above to identify betting irregularities on EPL data for the 2018/2019 season.

On Figure number one we can see the process that was followed in order to perform the study. In short we can say that we obtained the data set online. The data was in a CSV file. We selected the information we needed from the data set and loaded it into the R data frame. Then we applied the anomaly algorithm (*time_decompose(), anomalize(),* and *time_recompose()*.) to detect the anomalies in the data. After that, we visualized the anomalies using the *plot_anomalies()* function and got our chart. We had to do these 3 times filtering the data with different bet odds.

Results and Discussion

In order to understand the plots, here are the definitions of certain components of the "anomalize" package [9]:

- "observed": The observed values (actuals).
- "season": The seasonal or cyclic trend. The default for daily data is a weekly seasonality.
- "trend": This is the long term trend. The default is a Loess smoother using spans of 3-months for daily data.

– “remainder”: This is what we want to analyze for outliers. It is simply the observed minus both the season and trend.

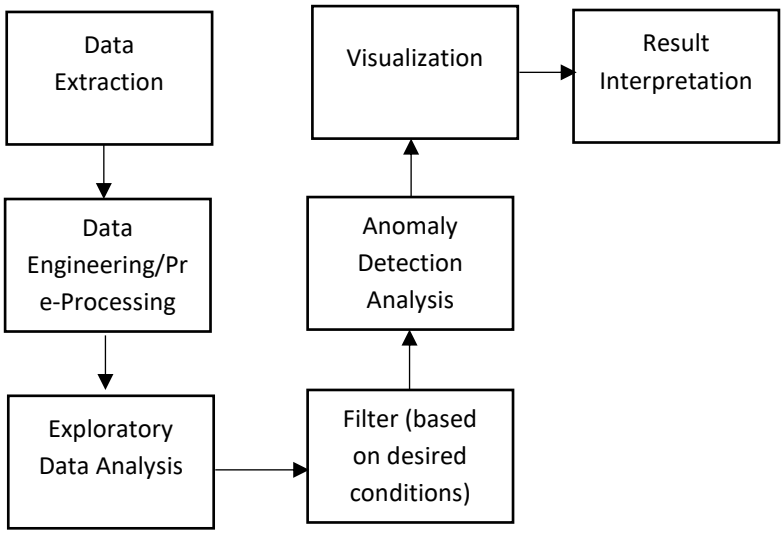

Figure 1 – Research Methodology

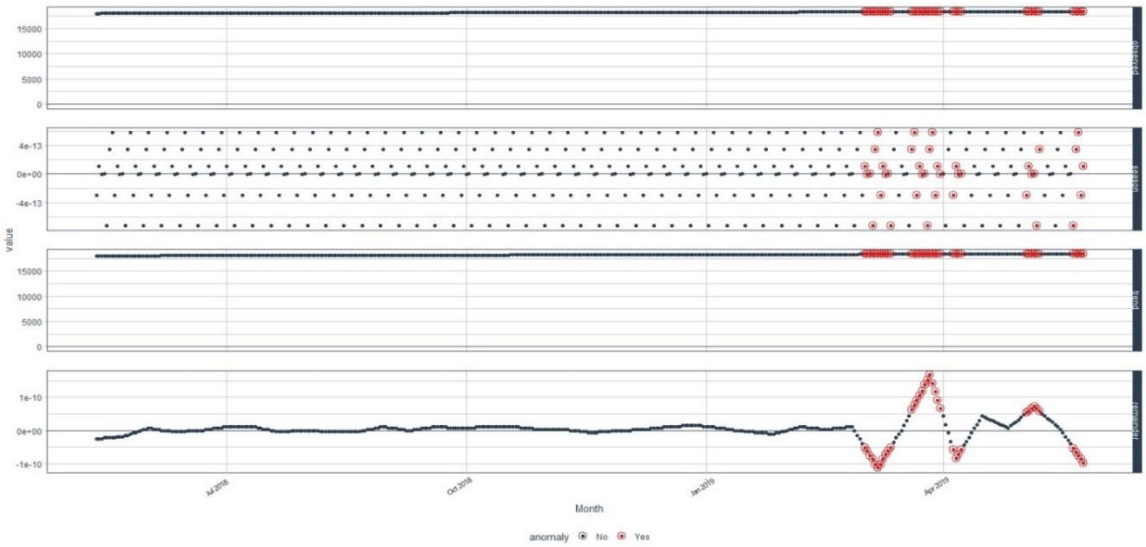

Figure 2 – Anomaly Detection Result for Overall Dataset

Conclusion

We analyzed the data from 380 matches of the 2018-2019 EPL. We have 3 charts that represent the full data set, anomalies found on the double chance odds for away wins and anomalies found on the DNB odds for away wins. There were anomalies on all charts. We can conclude that the anomalies mostly happened at the end of the season. On Figure 2 representing the whole data we can see that the number of anomalies increase on the last 3 months of the season. One of the reasons could be that towards the end of the season is when the league is decided. Therefore, the number of audience and the number of bets increase. With more people betting there is an incentive for bookmakers to try to increase their profits as well. This could be one of the reasons why the anomalies show on such dates. Though in order to make such conclusion other studies would need to be made. We can’t assume that the anomalies actual come from an illegal activity. In order to figure this out more studies need to be made. Perhaps our study could be used to make audits. From these audits a pattern would need to be found as to the teams involved or perhaps the bookmakers in charge of the odds. This would have to be done by analyzing the data of the sport bet websites. With the number of legal online casinos it would not be an easy task. However, the fact that everything is done online also means the data can be obtained fast.

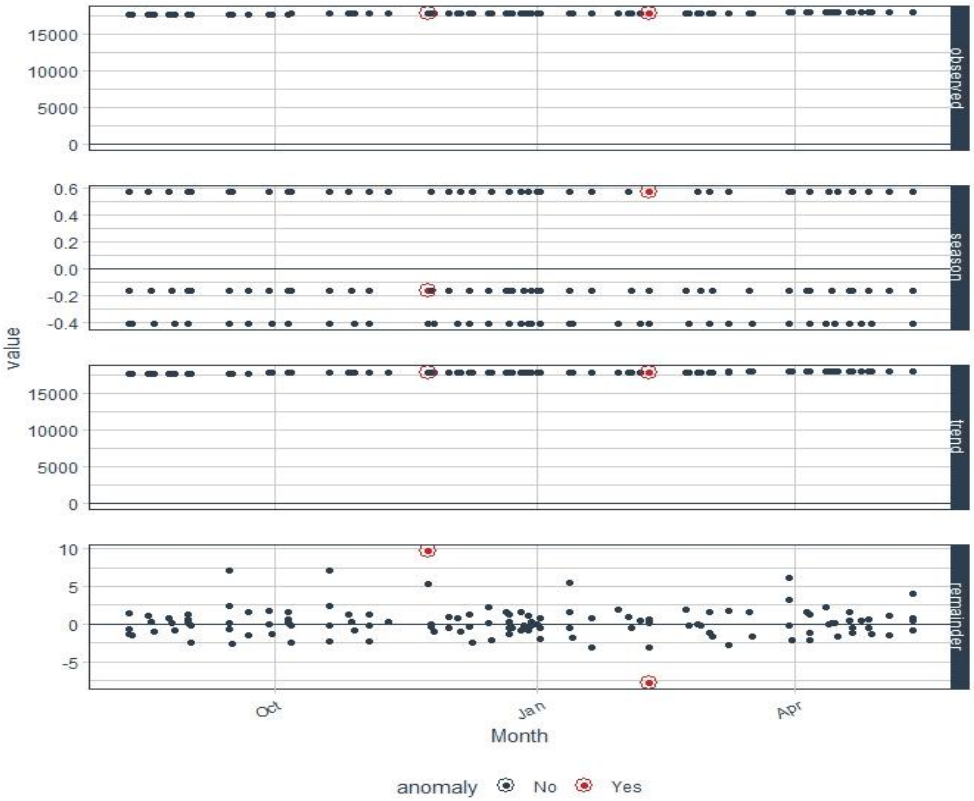

Figure 3 - Anomaly Plot of Double Chance Odds for Away Wins

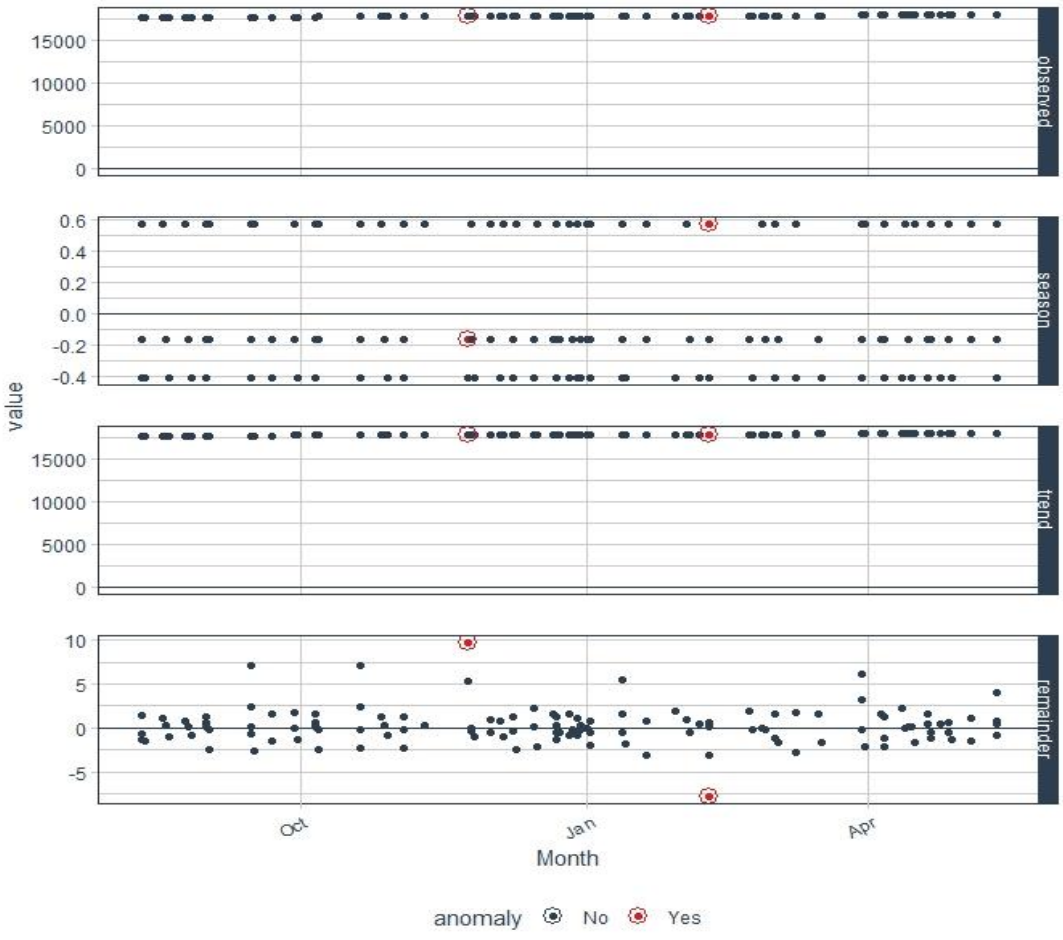

Figure 4 - Anomaly Plot of DNB Odds for Away Wins

References:

1. Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS One, 11(4), e0152173.
2. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1–58.
3. Indatabet. (n.d.). Soccer Historical Data of Odds & Results. Retrieved 13 January 2021, from https://www.indatabet.com/soccer.html
4. Lange, D. (2020). Premier league revenue streams England 2014/15-2020/21 Statistic. Statista. https://www.statista.com/statistics/874020/revenue-of-premier-league-football-clubs-by-stream/
5. EPL. (n.d.). Entertaining audiences. Retrieved 13 January 2021, from http://www.premierleague.com/news/686489
6. Dancho, M, & Vaughan, D. (2018). Anomalize: Tidy anomaly detection. R Package Version 0.1, 1.
7. Ryan, P. M., & Ryan, C. A. (2019). Mining Google Trends Data for Health Information: The Case of the Irish "CervicalCheck" Screening Programme Revelations. Cureus, 11(8). https://doi.org/10.7759/cureus.5513
8. North, S., Piwek, L., & Joinson, A. (2020). Battle for Britain: Analyzing Events as Drivers of Political Tribalism in Twitter Discussions of Brexit. Policy & Internet.
9. Dancho, Matt, & Vaughan, D. (2020). Anomalize Quick Start Guide. https://business-science.github.io/anomalize/articles/anomalize_quick_start_guide.html