Finding a Model that Provides High Profits with Web Usage Mining: A Case Study

Şükrü Can Şayan[1], **Tahsin Çetinyokuş**[1*]

1 Gazi University, * Corresponding author, tahsinc@gazi.edu.tr

Abstract

As the methods of data collection and analysis evolve, they are developed in ways that can transform this data into intended results. For this purpose, it is necessary to draw a road map of the work to be done with planning and to reach the aimed result. In this study, a plan was made for the selected web site and the results were gathered within this plan and a result that can give an idea about the changes needed to reach the goals by using the appropriate models together.

Keywords: Web usage mining, Pre-processing, Decision trees, K-means, Feature selection.

Citation: Şayan, Ş. C., Çetinyokuş, T. (2018, October) *Finding a Model that Provides High Profits with Web Usage Mining: A Case Study*. Paper presented at the Fifth International Management Information Systems Conference.
Editor: H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey
Received: August 19, 2018, **Accepted:** October 18, 2018, **Published:** November 10, 2018

IMISC 2018 PAPER

Finding a Model that Provides High Profits with Web Usage Mining: A Case Study

Abstract

As the methods of data collection and analysis evolve, they are developed in ways that can transform this data into intended results. For this purpose, it is necessary to draw a road map of the work to be done with planning and to reach the aimed result.

In this study, a plan was made for the selected web site and the results were gathered within this plan and a result that can give an idea about the changes needed to reach the goals by using the appropriate models together.

Keywords

Web usage mining, Pre-processing, Decision Trees, K-means, Feature Selection

1. Introduction

It is possible to say that Web is the fastest growing and developing technology today. Today, the speed of setting up new secure servers and spreading new applications to the web has increased dramatically. Increase in the number of servers on the world shown in Figure 1.

INSERT FIGURE 1 HERE

This rapid development of the Web has led to the use of data mining methods to understand its structure. Web mining is classified into three types based on extracting knowledge. They are web structure mining, web content mining, and web usage mining (Sukumar, Robert, & Yuvaraj, 2016). Web use mining is basically aiming to extract meaningful information and usable results from these by using visit records.

2. Literature

There are studies in the literature about web usage mining. In these studies, advances and case studies about pre-processing processes have been abundantly processed. In a large majority of these analyzes, server-side access files held by the web server were utilized and attempts were made to analyze these data files. The troubles of this method are frequently explained in literature.

The problems frequently encountered in the literature are generally classified under the following headings.

- User identification
- Session identification
- Data cleaning and de-spidering

3. Study

Nowadays, websites are used not only for promotion and communication, but also for profit and sales. For this reason, it is important to measure how much and how much revenue this website generates as much as the number of visitors. Electronic commerce data is seen in the last 5 years in Turkey, shown in Figure 2.

INSERT FIGURE 2 HERE

3.1. Data Mining Plan

The preferred website for study is a catalog site that publishes and compares the uniform products of different companies. Firms that publish their products on the website earn prioritization at different points of the site by setting a fee per referring, and they pay money to each site in each refer. The goal of the site is to generate more revenue by providing more referring. For this reason, it will be tried to determine the factors affecting profits by analyzing the appropriate model and customers. For this purpose, the following plan has been created.

1. The data is obtained from the website.
2. Detection of dependent and independent parameters related to website.
3. Identification of high-income customers and the study of these customers
4. Selecting the appropriate ones with the selected parameter
5. Identification of the model describing independent variables that shape dependent variables

3.2. Data Collecting

An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important for Web applications due to the combination of click-stream data and other related data collected from multiple sources and across multiple channels (Liu, 2007). The recording mechanism developed in the web site was used for this data. With this recording mechanism, less preprocessing is provided from log files, and the method of income generation is included in the calculations. The data are collected according to the following conditions.

- IP
- Browser Agent
- Date
- Page
- User Status

The site is also written in each request access table. Because the logging is done at the application software layer, visual, css, etc. requests are not recorded in this system. Only the request to the

application is saved while saving the saved data. The access record module diagram is shown in Figure 3.

INSERT FIGURE 3 HERE

3.3. Pre-Processing

Web data are noisy, incomplete, inconsistent and difficult to analyze and mine. Superior quality of data gives superior quality of output; superior quality of output gives reliable information. For this reason, web data mining offers data preprocessing techniques (Sukumar et al., 2016). The obtained data has been pre-processed to be used in the modeling process. These operations are as follows.

- Data Cleaning and de-spidering: The registration module does not contain unnecessary entries only because it registers application requests. Data cleaning also entails the removal of references due to crawler navigations. It is not uncommon for a typical log file to contain a significant (sometimes as high as 50%) percentage of references resulting from search engine or other crawlers (or spiders) (Liu, 2007). The logging system recognizes webbots and saves their requests in the BrowserAgent field with the "Bot:[Bot Name]" template.
- User Idenfication: Users are identified by IP address and browser name However; the page control is not considered because the companies can change the link structure by changing their bids. The resulting IP and header data are compressed by hashing with md5. The algorithm takes as input a message of arbitrary length and produces as output a 128-bit "fingerprint" or "message digest" of the input (Rivest, 1992).
- Session İdentification: After the users are identified, the user actions are divided into 30 minute sessions. The simplest method of achieving this is through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout, and established a timeout of 25.5 minutes based on empirical data (Cooley, Mobasher, & Srivastava, 1999).

3.4. Evaluation Parameters

After preprocessing of records, 6 different parameters were selected for modeling operations:

- .user: Logged-on user status
- pageCount: Number of pages visited
- firmCount: Company profile page shown
- packageCount: Number of forwarding
- totalTime: Time spent on site
- blogCount: The number of blogs read

The evaluation parameters are obtained by the following coding.

While(Not End Of Table)

Calculate user key

Test for userStatus and set it 1 if has auth

If(Request time is not 1800 second later than previous request)

Add Time diff to totalTime

Increase pageCount for user

IF(url contains "blog")

Increase blogCount for user

IF(url contains "firma")

Increase firmCount for user

IF(url contains "yonlendir")

Increase packageCount for user

Examples for the data obtained after this calculation are shown in Table 1.

INSERT TABLE 1 HERE

3.5. Establishing the model

The data was analyzed and modeled using the IBM SPSS Modeller application. The data were first subjected to anomaly detection. The extreme points of the data were sifted and worked with the remaining data. After the elimination process, the users are clustered. The K-means algorithm is used for this grouping. For this grouping, high profitable clusters were obtained in all the data. CRT decision tree model was applied for finding high profit users.

3.5.1. Detecting Abnormal Records

Anomaly modeling which is included in Modeller was used for the detection of abnormal records. The layout view for this model is shown in Figure 4.

INSERT FIGURE 4 HERE

With the model installed, 137 rows were found as abnormal and extracted from the data. The distribution of the anomaly priority for these records is given in Table 2.

INSERT TABLE 2 HERE

3.5.2. K-means Clustering

Clustering of user records (sessions or transactions) is one of the most commonly used analysis tasks in Web usage mining and Web analytics. Clustering of users tends to establish groups of users exhibiting similar browsing patterns (Liu, 2007). It is possible to find groups that provide income by dividing this into different groups. Primarily, the clusters were clustered 4 times with 3, 4, 5, and 6 clusters, and the best clusters in these clusters were found. For clustering The K-means model in the Modeller application is used. The K-means algorithm is the best known partitional clustering

algorithm. It is perhaps also the most widely used among all clustering algorithms due to its simplicity and efficiency (Liu, 2007). The best clusters were selected as clusters with the highest average target value of packageCount. Selected clusters are shown in Table 3.

INSERT TABLE 3 HERE

3.5.3. Feature Selection

Feature selection process refers to choose a subset of attributes from the set of original attributes. The purpose of the feature selection is to identify the significant features, eliminate the irrelevant of dispensable features to the learning task, and build a good learning model (Inbarani, Thangavel, & Pethalakshmi, 2007). Each of the clusters determined by clustering is subjected to the Feature Selection Model located in Modeller, so that parameter groups that can best represent the cluster have been identified. Important and Marginal metrics were selected in the selection of these detected metrics. Selected metrics are shown in Table 4.

INSERT TABLE 4 HERE

3.5.4. Decision Trees Creation

The best clusters obtained were used to generate decision trees according to the best metrics obtained for these clusters. Decision trees are a very effective method of supervised learning. It aims is the partition of a dataset into groups as homogeneous as possible in terms of the variable to be predicted (Hssina, Merbouha, Ezzikouri, & Erritali, 2014). Among these methods, CHAID, CART, ID3 and C4.5 algorithms are frequently used (Seyrek & Ata, 2010). The CRT algorithm, which is included in the Modeller, is used to construct these decision trees. The statistical information obtained as a result of these models is shown in Table 5.

INSERT TABLE 5 HERE

The model gives the best correlation value and the lowest error value with 4 clusters. 4 The tree created by the cluster is shown in Figure 5.

INSERT FIGURE 5 HERE

According to the result nodes, designing the websites to provide the most accurate combination will be very helpful in increasing the revenue from the site. Assuming that each forwarding produces equivalent profits, revenue estimates based on the result nodes will be as given in Table 6. The highest revenue comes from Node 31.

INSERT TABLE 6 HERE

The model's decision text is as follows.

```
firmCount <= 188.500
        totalTime <= 26255771.500
                firmCount <= 28.500 => 1.667
                firmCount > 28.500 => 11.805
        totalTime > 26255771.500 => 33.111
firmCount > 188.500
        pageCount <= 3047.500
                firmCount <= 429.500
                        blogCount <= 6.500 => 133.5
                        blogCount > 6.500
                                firmCount <= 296.500 => 32.0
                                firmCount > 296.500 => 97.0
                firmCount > 429.500
                        totalTime <= 21,853,204 => 281.25
                        totalTime > 21,853,204 => 151.75
        pageCount > 3047.500
                totalTime <= 25,485,426
                        blogCount <= 51
                                firmCount <= 568 => 33.279
                                firmCount > 568 => 69.75
                        blogCount > 51 => 110.333
                totalTime > 25,485,426 => 77.783
```

4. Result and Discussion

Today, web mining is a work and decision support system that can be adapted to a large part of the web and beneficial results can be obtained. With these outputs, it is possible to reconstruct the web sites according to their purposes and to orient the users in the desired direction.

In this study a module was developed on the website for the collect data and reduce the necessary preprocessing processes for the data mining. It is possible to perform the current work via server access records. It has been tried to show how much data cleaning processes can be shortened with additional modules.

In addition, there are many methods that can be used in this area, and this study tried to explain how these methods can be combined. The data and experience obtained from this study can be used to

rebuild the website. In addition, a similar operation may be tried again by changing the selected independent parameters.

References

BDDK. (2018). *CARDED PAYMENT PROCEDURES WITHIN INTERNET*. Retrieved from: https://bkm.com.tr/internetten-yapilan-kartli-odeme-islemleri/

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems, 1*(1), 5-32.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications, 4*(2).

Inbarani, H. H., Thangavel, K., & Pethalakshmi, A. (2007, 13-15 Dec. 2007). *Rough Set Based Feature Selection for Web Usage Mining.* Paper presented at the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007).

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*: Springer Science & Business Media.

Rivest, R. (1992). *The MD5 message-digest algorithm* (2070-1721). Retrieved from

Seyrek, İ. H., & Ata, H. A. (2010). Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü. *Journal of BRSA Banking & Financial Markets, 4*(2).

Sukumar, P., Robert, L., & Yuvaraj, S. (2016). *Review on modern Data Preprocessing techniques in Web usage mining (WUM).* Paper presented at the Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on.

WorldBank. (2018). *Secure Internet Servers*. Retrieved from: https://data.worldbank.org/indicator/IT.NET.SECR.P6?end=2017&locations=1W-EU-US&start=2010&type=shaded&view=chart

Figures and Tables

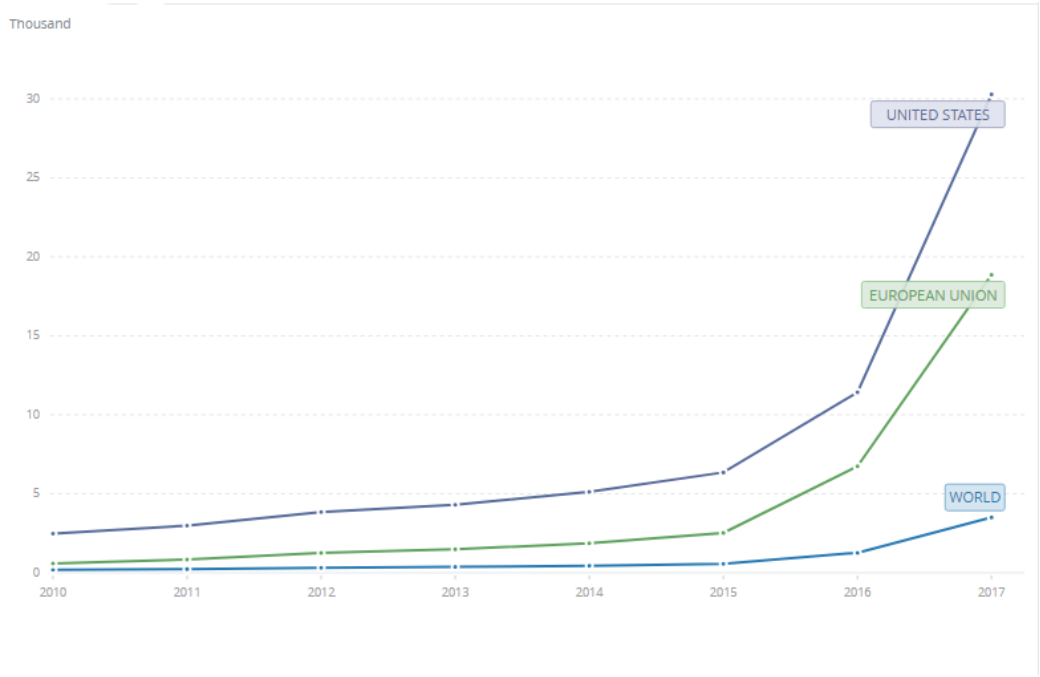

Figure 1. Secure Internet Server per million people(WorldBank, 2018)

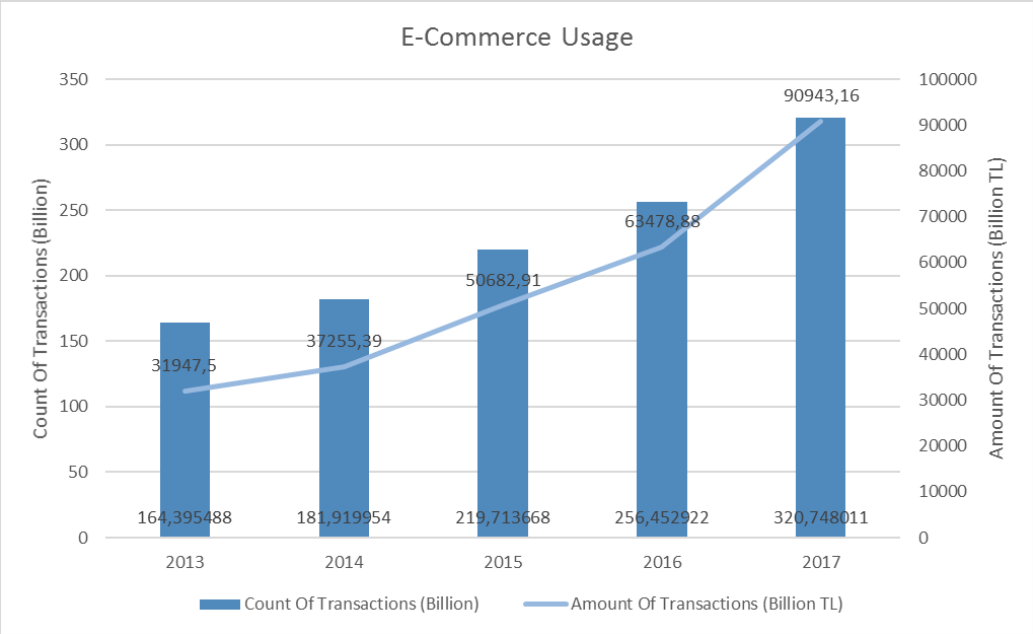

Figure 2. E-commerce credit card usage (BDDK, 2018)

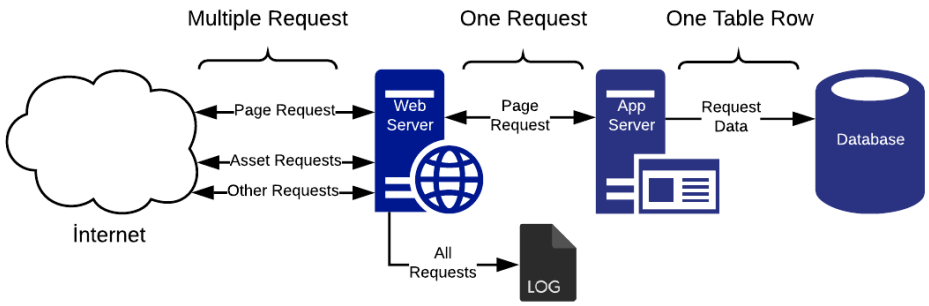

Figure 3. Access Record Module Process

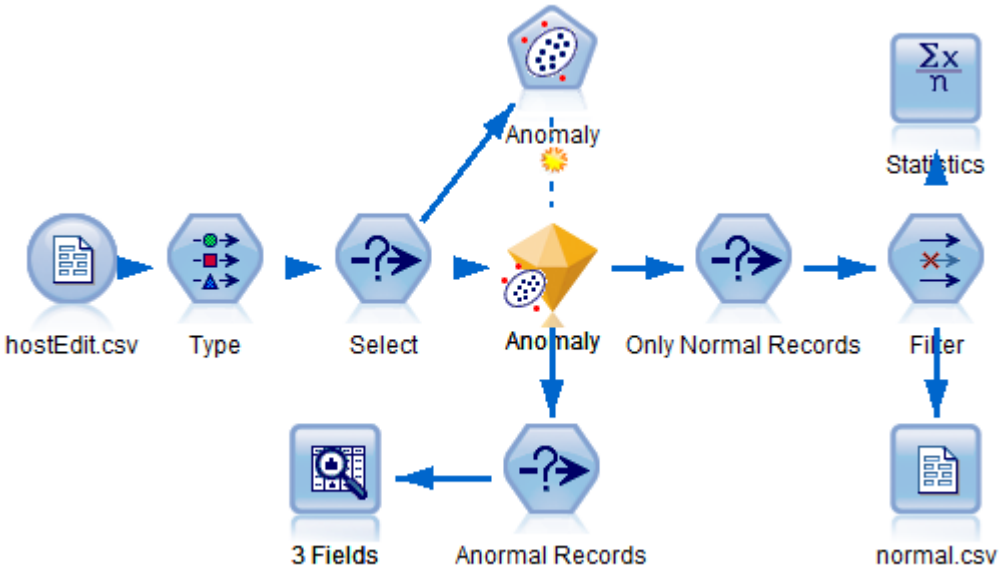

Figure 4. Anomaly Model Setup

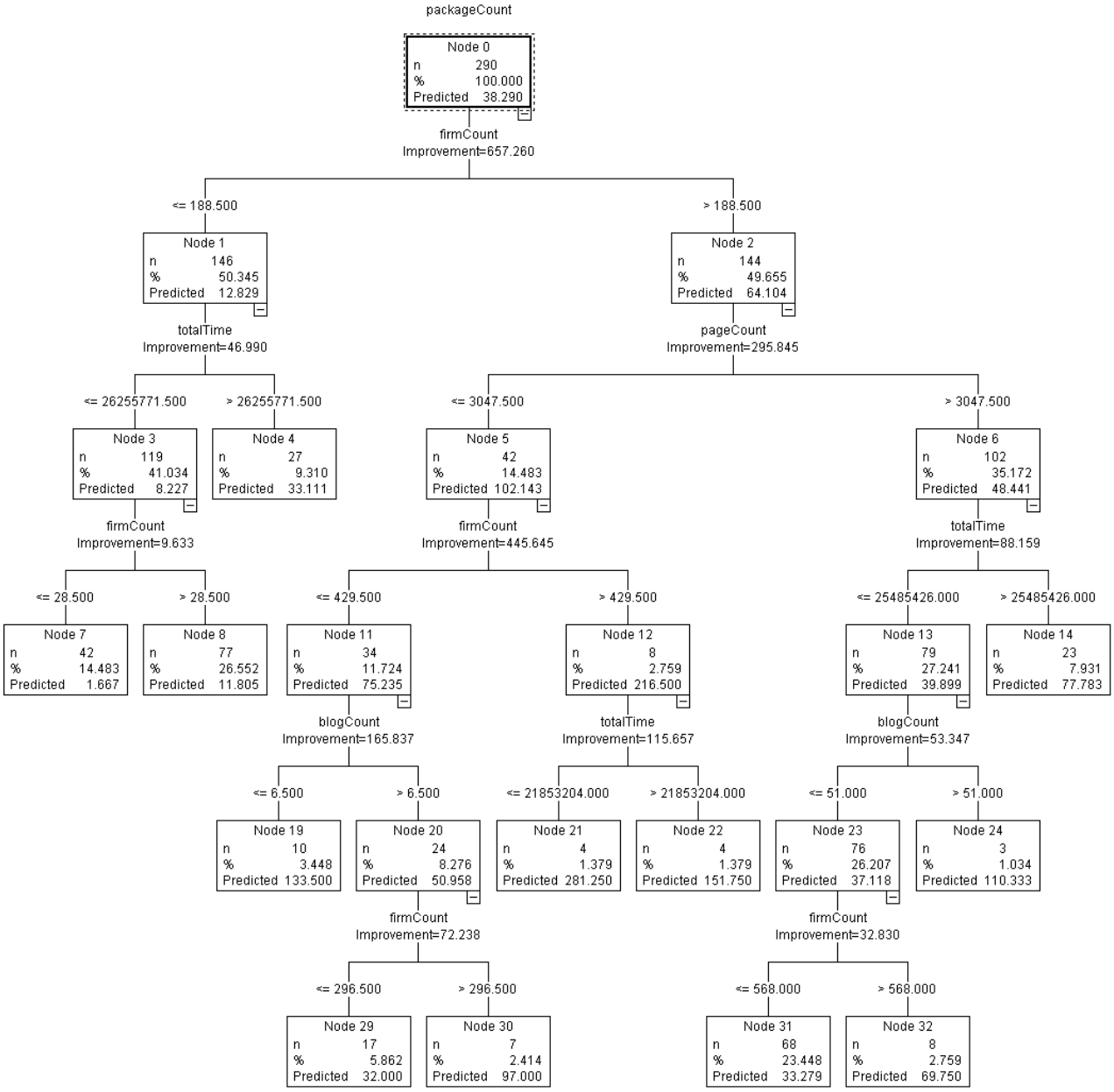

Figure 5:CR Tree Diagram for 4 Cluster Data

Table 1.

Example data table after pre-processing

id	client	user	pageCount	firmCount	packageCount	totalTime	blogCount
3	179a0585f5dc63ce0eba4fcbe7ca08f5	0	2017	157	3	27554703	11
25	f5dbfe27801f503f0630684f006207b1	0	37	0	3	1977	0
37	361169b1d705f6b65415c18bd8c7cc15	0	2377	199	5	6721861	12
54	35b802948ed876a1287fc71dbcad905c	0	420	39	53	7164212	20
76	8b36bdb44d5e5128a21d4fdfb3745d74	0	1607	0	4	2170578	0

Table 2.

Anomaly Repeats

	1st priority	2nd priority	3rd priority	Total
firmCount	7	44	75	126
pageCount	59	44	20	123
totalTime	50	44	23	117
blogCount	9	1	11	21
packageCount	8	4	8	20
user	4	0	0	4

Table 3.

Best clusters for each clustering

	3 Cluster	**4 Cluster**	**5 Cluster**	**6 Cluster**
Best Cluster	cluster-1	cluster-1	cluster-1	cluster-1
PackageCount Average	37.5	28.29	45.73	48.44
Size	2.8%	2.1%	1.1%	1%

Table 4.

Feature Selection Metrics

	3 Cluster	**4 Cluster**	**5 Cluster**	**6 Cluster**
Important	firmCount blogCount pageCount totalTime	firmCount blogCount pageCount totalTime	firmCount	firmCount
Marginal			blogCount	blogCount

Table 5.

CRT Model Results

	Clusters			
Decision Tree Depth	3 Cluster	4 Cluster	5 Cluster	6 Cluster
Minimum Error	4	5	5	5
Maximum Error	-102,8	-132,5	-107,8	-200,7
Mean Absolute Error	307,23	196,21	176,20	167,60
Standard Deviation	30,132	25,750	30,460	30,803
Linear Correlation	49,604	41,599	45,403	45,263
Occurrences	0,552	0,731	0,685	0,722
	377	290	153	130

Table 6.

CRT Model Profit

Node	packageCount	Count	Percent	Profit
7	1,66	42	14,49%	69,72
8	11,805	77	26,55%	908,99
4	33	27	9,31%	894,00
19	133,5	10	3,45%	1335,00
29	32	17	5,86%	544,00
30	97	7	2,41%	679,00
21	281,25	4	1,38%	1125,00
22	151,74	4	1,38%	606,96
31	33,279	68	23,49%	2262,97
32	69,75	8	2,76%	558,00
24	110,333	3	1,03%	331,00
14	77,783	23	7,90%	1789,01