

**UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION
INSTITUTE OF SCIENCE AND TECHNOLOGY**

**AIRLINE CUSTOMER DATA ANALYTICS
INTEGRATED WITH SOCIAL NETWORK INFORMATION**



**MASTER THESIS
Ahmet Birol avdar**

Engineering Management Department

Master of Science in Engineering Management Program

JANUARY 2017

**UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION
INSTITUTE OF SCIENCE AND TECHNOLOGY**

**AIRLINE CUSTOMER DATA ANALYTICS
INTEGRATED WITH SOCIAL NETWORK INFORMATION**

MASTER THESIS

Ahmet Birol avdar

1403670028

Engineering Management Department

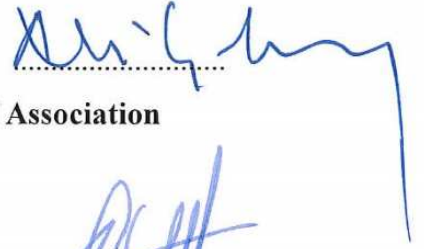
Master of Science in Engineering Management Program

Supervisor: Prof. Dr. Ali oban

Ahmet Birol ÇAVDAR, having student number 1403670028 and enrolled in the Engineering Management Master Program at the Institute of Science and Technology at the University of Turkish Aeronautical Association, after meeting all of the required conditions contained in the related regulations, has successfully accomplished, in front of the jury, the presentation of his thesis prepared with title of "Airline Customer Data Analytics Integrated with Social Network Information".

Supervisor:

Prof. Dr. Ali ÇOBAN



University of Turkish Aeronautical Association

Co-advisor:

Assist. Prof. Dr.

Nilgün FERHATOSMANOĞLU



University of Turkish Aeronautical Association

Jury Members:

Assist. Prof. Dr.

Hasan Umut AKIN



University of Turkish Aeronautical Association

Assist. Prof. Dr.

Gülesin Sena DAŞ

Kırıkkale University



Thesis Defense Date: 6 January 2017

**INSTITUTE OF SCIENCE AND TECHNOLOGY OF
THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION**

I hereby declare that all the information in this study I presented as my Master's Thesis, called "Airline Customer Data Analytics Integrated with Social Network Information" has been presented in accordance with the academic rules and ethical conduct. I also declare and certify on my honor that I have fully cited and referenced all the sources I made use of in this present study.

06.01.2017

Ahmet Birol ÇAVDAR



PREFACE

The thesis, *Airline Customer Data Analytics Integrated with Social Network Information* aims to analyze airline customer data by combining customers' flight data with their social network factors. It has been written to fulfill the graduation requirements of the Engineering Management graduate program at the University of Turkish Aeronautical Association. I was engaged in researching and writing this thesis starting from October 2015.

My research question was formulated together with my co-advisor, Assist. Prof. Dr. Nilgün Ferhatosmanođlu. In this research, it was challenging to develop an effective method to integrate the social network information with the traditional data analytical methods. Fortunately, I was able to overcome the problems with the help of Dr. Ferhatosmanođlu.

I would like to start my acknowledgements statements with thanking my co-advisor for her excellent guidance, broad vision and incredibly useful advice during this process. I also wish to thank my supervisor, Prof. Dr. Ali oban, for his support to complete my thesis. I am grateful to řafak Tulumođlu because of his contribution to the maturation of my thesis, to Bařar Kasım for his motivation support and advice. My mother-in-law, brother-in-law and sister-in-law: Thanks for your hospitality and great working environment you've provided. My wife, Betül avdar, deserves the greatest thanks because of her extra dedication in the care of our children and her motivation support.

Since this thesis is an output of the studies conducted related to the career project (112M950) of Dr. Ferhatosmanođlu, here I would like to thank The Scientific and Technological Research Council of Turkey too, because of their financial and academic support.

I dedicate this thesis to my wife and my children, for times I am not able to bring back.

January, 2017

Ahmet Birol AVDAR

CONTENTS

PREFACE	iii
CONTENTS	iv
TABLES	v
FIGURES	vi
ABBREVIATIONS	vii
ABSTRACT	viii
ÖZET	ix
CHAPTER ONE	10
1. INTRODUCTION	10
CHAPTER TWO	12
2. LITERATURE REVIEW.....	12
CHAPTER THREE.....	16
3. MATERIALS AND METHODS.....	16
3.1. Flight Data.....	16
3.2. Social Network Data	18
3.3. Modeling Supported by Network Information.....	25
3.3.1. Specification of Model Selection Criteria.....	27
3.3.2. Construction of Base Model Candidates.....	28
3.3.3. Base Model Selection	30
3.3.4. Integration of Social Network Information.....	37
3.3.5. Construction of the Proposed Model	39
3.3.6. Comparison of the Base and the Proposed Models	42
CHAPTER FOUR.....	47
4. RESULTS AND DISCUSSION	47
4.1. Airline Customer Data Analytics	49
4.1.1. Customer Lifetime Value Analysis.....	49
4.1.2. Membership Category Analysis.....	51
4.1.3. Profitability Analysis	53
4.1.4. Churn Analysis.....	54
4.2. Time Series Forecasting Models	56
CHAPTER FIVE.....	57
5. CONCLUSION	57
REFERENCES	59
APPENDICES	63
Appendix A - Descriptive Statistics of the Flight Data.....	64
Appendix B - 500 Most Valuable Customers Rankings of the Models.....	65
Appendix C - R Source Code	72
RESUME	86

TABLES

Table 3.1: Example flight data gathered from Mile Program web site.	16
Table 3.2: Demographic and behavioral attributes of converted data.	17
Table 3.3: The social network unique identifiers of the egos.	19
Table 3.4: Degree centrality values of the example network.	21
Table 3.5: Closeness centrality values of the example network.	22
Table 3.6: Betweenness centrality values of the example network.	23
Table 3.7: Betweenness centrality calculation details of the example node.	23
Table 3.8: PageRank values of the example network.	23
Table 3.9: Hub score values of the example network.	25
Table 3.10: Arguments passed to 'regsubsets' to find the base model.	29
Table 3.11: Significances of the factors of the initial model.	31
Table 3.12: Explanations about the significance measures.	31
Table 3.13: Values of the variables used in synthetic data size calculation.	32
Table 3.14: Significances of the factors of the model using extended data.	36
Table 3.15: Significances of the factors of the base model.	37
Table 3.16: Variables of social network data.	39
Table 3.17: Arguments passed to 'regsubsets' to find the proposed model.	40
Table 3.18: Significance levels of the factors of the proposed model.	41
Table 3.19: Comparison of the determination coefficients of the two models.	42
Table 3.20: Comparison of the confidence levels of the two models' factors.	43
Table 3.21: Top 60 of the models' customer value rankings.	45

FIGURES

Figure 3.1: Graph of the Facebook data.	19
Figure 3.2: Graph of the example network used in social score explanations.	20
Figure 3.3: Degree centrality graph of the social network data.....	21
Figure 3.4: Flowchart of the proposed methodology.	26
Figure 3.5: Best Subsets Regression plot for discovery of the base model.....	29
Figure 3.6: Monthly flight frequencies.....	34
Figure 3.7: Monthly total distances.	34
Figure 3.8: Monthly total sales.	35
Figure 3.9: Best Subsets Regression plot using extended dataset.	36
Figure 3.10: Best Subsets Regression plot of the model with social scores.....	40
Figure 3.11: Customer Values of the Two Models.....	44
Figure 4.1: Customer lifetime value analysis plot of the whole network.	50
Figure 4.2: Customer lifetime value analysis plot of the filtered network.	50
Figure 4.3: Membership category plot of the whole network.	52
Figure 4.4: Membership category plot of the filtered network.....	52
Figure 4.5: Profitability analysis plot of the whole network.	53
Figure 4.6: Profitability analysis plot of the filtered network.	54
Figure 4.7: Churn analysis plot of the whole network.....	55
Figure 4.8: Churn analysis plot of the filtered network.....	55
Figure 4.9: The flowchart of the proposed time series forecasting method.	56

ABBREVIATIONS

ARIMA	: Auto-Regressive Integrated Moving Average
CLV	: Customer Lifetime Value
CRM	: Customer Relationship Management
DM	: Data Mining
SNAP	: Stanford Network Analysis Project

ABSTRACT

AIRLINE CUSTOMER DATA ANALYTICS INTEGRATED WITH SOCIAL NETWORK INFORMATION

ÇAVDAR, Ahmet Birol

Master, Department of Engineering Management

Supervisor: Prof. Dr. Ali ÇOBAN

January-2017, 86 pages

Customer Relationship Management (CRM) has become essential in the business world where competition exhibits a steady increase. Business performance can be significantly improved through analytical applications such as estimation of customer lifetime value (CLV), profitability computation, profiling, classification, customer retention and churn analyses. In recent years, besides traditional data, social network behaviors of users and the interactions among them can be obtained. Although the need for social network data mining activities has been increasing, research on their integration with existing data analytical models is still limited.

The purpose of this study is to develop a model for estimating customer value in airline industry that utilizes customers' flight information as well as their social network information. For this purpose, firstly, a regression model for airline customers to estimate their CLVs is adopted. After that, a method for enhancing this base model with customers' social network information is proposed and the performances of the both models are compared. Lastly, some airline customer analysis cases using the proposed first model for CRM applications are exemplified.

Keywords: Social network analysis, regression and classification models, time series, forecasting methods, customer profiling, data analytics, customer relationship management.

ÖZET

SOSYAL AĞ BİLGİSİ İLE BÜTÜNLEŞTİRİLMİŞ HAVAYOLU MÜŞTERİSİ VERİ ANALİTİĞİ

ÇAVDAR, Ahmet Birol

Yüksek Lisans, Mühendislik Yönetimi Anabilim Dalı

Tez Danışmanı: Prof. Dr. Ali ÇOBAN

Ocak-2017, 86 sayfa

Rekabetin sürekli artış göstermesiyle iş dünyasında müşteri ilişkileri yönetiminin önemi artmaktadır. Müşteri yaşamboyu değeri, müşteri kârlılık hesaplama, profilleme ve sınıflandırma, müşteri ayrılma analizleri, müşteri geri kazanımı gibi analitik uygulamalarla sektör performansı önemli düzeylerde arttırılabilmektedir. Son yıllarda geleneksel verilerin yanı sıra, kullanıcıların başka kullanıcılarla etkileşimlerini gösteren sosyal ağ hareketleri de elde edilebilmektedir. Sosyal ağ veri madenciliğine ihtiyaç artış göstermekle birlikte, bu konunun mevcut veri analitik modelleri ile bütünleştirilmesi alanındaki araştırmalar henüz kısıtlıdır.

Bu çalışmanın amacı, havayolu sektöründe müşteri yaşamboyu değerini tahmin etmek için, müşterinin uçuş bilgilerinin yanı sıra sosyal ağ bilgilerini de kullanan bir model geliştirmektir. Bu amaçla, öncelikle havayolu müşterilerinin müşteri yaşam boyu değerlerinin tahmini için bir regresyon modeli seçilir. Ardından, bu temel modelin müşterilerin sosyal ağ bilgileriyle geliştirilmesi için bir model önerilip, bu iki modelin performanslarını karşılaştırılır. Son olarak, müşteri ilişkileri yönetimi uygulamaları için, önerilen model kullanılarak birkaç havayolu müşteri analizi vakası örneklendirilir.

Anahtar Sözcükler: Sosyal ağ analizi, regresyon ve sınıflandırma modelleri, zaman serileri, öngörü yöntemleri, müşteri profili belirleme, veri analitiği, müşteri ilişkileri yönetimi.

CHAPTER ONE

1. INTRODUCTION

The increase in the amount of collected data and applications in the organizations in parallel with the developments in information technology has also increased the need for the data mining (DM) methods. Discovery of the trends and extraction of meaningful information from the collected data and then arranging the customer relations based on the results obtained require holistic approaches. In the air transport sector, customer relationships are carried out mostly by focusing on travels. Even the customer relationship management (CRM) units of the leading airline companies are still young and not yet benefiting fully from the data mining activities to better recognize the customers and provide them with the services.

In the air transport sector, the most successful CRM practices are the *Mile Programs*. It is known that the customers who are the members of the mile program are usually ahead in terms of loyalty and profitability. Although mile program data contain rich amount of customer information, each customer is mostly assessed individually. However, the position, influence and features of the customers in the social network topology are important in many other sectors as well as in the air transport sector and should be used as a factor in the data analytical models. For example, a customer's direct expenditures as well as his/her positive or negative influence potential on other customers can provide information about his/her customer value for the company. In addition, social network neighborhoods and topological similarities between people might give important hints about the tendencies of the customers.

With the rapid spread of social networks in recent years, it is clear that information gathered from the social networks plays an important role in customer analysis and corporate decision-making processes. Such information is usually not

considered in enough depth in existing systems. Network information can be used to improve accuracy of time series forecasting results. Although the statistical forecasting methods have been widely studied in the literature, an approach that uses flight information accompanied with the social network information has not attracted enough attention in the air transportation sector.

In this work, social network information is integrated into a multiple regression model to improve estimation of the customer lifetime value (CLV) of airline customers. The proposed method involves methods for both customer value estimation and social network analyses.

Rather than examining an actual large scale data and making conclusions from its content, this study focuses on developing a high level methodology that integrates social network information into the traditional data analytical techniques and showing that the social factors may improve the accuracy of the models.

CHAPTER TWO

2. LITERATURE REVIEW

Companies in general, the ones in the service sector in particular, need to build long-term and strong relationships with their customers to ensure profitability and sustainable growth. As an indication of the importance given to the customer by the companies, recently, marketing activities and performance evaluations are increasingly carried out using the customer relationships rather than products (Jain & Singh, 2002). The inclusion of a separate section for managing customer relationships in companies is another indication of the fact that the customer relationships is treated as a serious factor that influences the organizational structure of a company. Within this context, customer segmentation, customer retention, and customer lifetime value (CLV) are the key issues that have been studied by the research communities of marketing, customer strategy and data analytics.

Companies need to retain their existing customers as well as to acquire new customers in order to be profitable (Anderson & Mittal, 2000). In the context of customer retention, Woo and Fock (2004) state that the axiom "the customer is always right" is no longer valid and emphasize the categorization of the customers with respect to their profitability and loyalty attributes. They categorize the customers as "right", "at-risk right customers" and "wrong" customers and propose a discriminant analysis to identify them. There are also other studies in the literature emphasizing that retaining every customer is not profitable ((Anderson & Mittal, 2000); (Niraj, et al., 2001); (Reinartz & Kumar, 2002)). Therefore, companies need to determine which customers are important to them and classify them according to their importance.

In the classification of customers according to their importance, "customer lifetime value" emerges as a commonly used feature. Berger and Nasr (1998) define

CLV by using “profitable customer” definition of Kotler and Armstrong (1996) as the acceptable amount of excess between a customer’s revenue and company costs of attracting, selling, and servicing that customer over time. Another commonly used definition of CLV is "the present value of all future profits generated from a customer" (Gupta & Lehmann, 2003). The other CLV definitions in the literature and the mathematical models on this subject such as the basic CLV model, the models that consider the customer retention, the ones which reflect the fluctuations of the sales and costs, etc. are explained in detail in the work of Ferrentino et al. (2016) who formulated the basic mathematical form of CLV in terms of customer revenues, company costs and discounting factor as in Equation 2.1:

$$CLV = \sum_{i=1}^n \frac{(V_i - C_i)}{(1 + d)^i} \quad \text{(Equation 2.1)}$$

In the field of air transport, Tirenni et al. (2007) introduce a decision-tree-based CLV model that segments customers by estimating their future CLVs. More recently, Ekinci et al. (2014) propose a model that uses the Markov decision process to quantify the CLV and exemplify an application of this model in the banking sector. They also suggest a regression-based method of predicting the future state of the attributes in the Markov decision process.

The studies in social network literature can be grouped in two categories:

- i. Structural analysis of social networks and
- ii. Data mining studies on social networks.

The research in the first group, which has also been the subject of sociological studies, has started with the experiment of Milgram (1967) that follows chain of mail letters manually. Similarly, Granovetter (1973) is one of the first studies examining the indirect structure of the networks. Watts and Strogatz (1998) observe the small world, grouping and short path features in many social network applications. It is also observed in social networks that, people affect each other’s behaviors and adjacent people act similarly (Trusov, et al., 2009). As a result of this, in addition to the traditional media, social networks have emerged as a new media for advertising.

The methods developed in the field of social network analysis models are discussed in detail in the books: (Wasserman & Faust, 1994), (Watts D. J., 2004) and (Carrington, et al., 2005).

With the advancement of communication technologies, the social network data generated by electronic chat environments like electronic mail, Skype, Google Talk or MSN Messenger have increased significantly (Lescovec & Horvitz, 2008). Web 2.0 technologies such as Facebook, Twitter, Myspace and Orkut have further diversified the social network data structures. As a result, social media data mining works have gained momentum (Kleinberg, 2007).

In fact, social networks are often reminiscent of social network applications. However, companies are able to create their own social networks from their trading activity logs and customer databases. As an example, one can determine the relationships between customers using outside public information such as Facebook friendships. With the help of this social network information, business intelligence products can be developed. For example, if a person's many friends have begun to prefer another company, the possibility of choosing the same preference will increase for that person (Dasgupta, et al., 2008). Such analyses for social networking have an important role in business intelligence and customer relationship management applications.

The structure of a social network is generally defined as a graph. There exist studies in the literature that try to find patterns on a graph using theoretical graph algorithms. A fundamental example of these studies are network influence measures that are generally used by the search engines such as PageRank (Brin & Page, 2012) and Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) (Kleinberg, 1999). In their original forms, these algorithms have been developed to model the relationships between the web pages instead of the social network nodes. However, different versions of them for social network analysis studies have been also developed. There are also many studies showing that people are affected by each other or similar people do similar behaviors (Trusov, et al., 2009). There are studies that formally define the *influence* concept (Tang, et al., 2009). Network mining on

the graph lies on the basis of these studies. Another area of social networking has found a wide application area is security applications. Significant research was conducted on monitoring the terrorist activities (Krebs, 2002) and epidemic diseases (Eubank, et al., 2004).

Approaches in the social network literature summarized above have mostly focused on the network mining. There is less work on integrating the social network information with data modeling tasks such as forecasting. Likewise, modeling studies have not sufficiently integrated the information obtained from social networks. In particular, a comprehensive study on these subjects is not available for the airline industry. In this study, such integration for airline customer value estimation is addressed. There have been some efforts on customer value estimation for airline industry (Tirenni, et al., 2007). However, these methods have not considered social network information.

Data mining for CRM has been successfully used in the telecommunications and financial sectors. For example, support vector machine (SVM) data classification was used in customer credit scoring system (Chen, et al., 2009). Regression was used in customer loyalty (Lariviere & Van den Poel, 2005). Decision tree, naive Bayes, and k-NN were used in marketing (Jiang & Tuzhilin, 2006). These approaches are not integrated with the social network information either. However, applications in the field of credit scoring, churn analyses and marketing can be extended with the customers' social network information. This study is mainly focused on customer value determination and scoring applications. All the methods can be applied in other CRM applications, as well.

CHAPTER THREE

3. MATERIALS AND METHODS

In this study, it is proposed to enrich the traditional models to determine the customer value in aviation industry by integrating customer's social position and relationships. Throughout this work, *Microsoft Excel* (Spreadsheet Software Programs | Excel Free Trial) is used for editing data and converting data files to other formats. *RStudio* software (RStudio, Inc.), which is a development tool for *R* statistical computing environment (The R Foundation), is used for all kinds of the programming needs. The R source code written in this study is given in Appendix C.

3.1. Flight Data

In this study, firstly a customer lifetime value determination method is adapted to the aviation industry. For this purpose, an anonymized dataset that includes flight information about the members of an airline's frequent flyer program is used. The dataset consists of *Date*, *Flight*, *Class*, *Activity*, *Description*, *Bonus Miles* and *Status Miles* attributes as seen in Table 3.1. The *Distance* attribute, measured in miles, is added to this dataset by using source and destination airport information as inputs for the flight distance calculator service provided by the travelmath web site (Flight Calculator).

Table 3.1: Example flight data gathered from Mile Program web site.

Date	Flight	Class	Activity	Description	Bonus Miles	Status Miles	Distance
13.12.2015	X8	F	SAW – ANK	SABIHA GOKCEN-ANKARA	0	150	190
11.12.2015	X9	F	ANK – SAW	ANKARA-SABIHA GOKCEN	0	150	190

The attributes of the dataset that involve Mile Program members are converted to the format given in Table 3.2. All but the one of the attributes (namely *MonthlyRecency*) are selected from demographical and behavioral features used in the study of Tirenni, et al. (2007).

Table 3.2: Demographic and behavioral attributes of converted data.

Attribute	Description
Id	Customer unique identifier
Value_New (Dependent variable)	Sales generated from Jan 2015 to Dec 2015
Value	Sales generated in Jan 2015
Value_3	Sales generated from Oct 2014 to Dec 2014
Value_6	Sales generated from Jul 2014 to Dec 2014
Value_12	Sales generated from Jan 2014 to Dec 2014
Freq	Number of trips in Jan 2015
Freq_3	Number of trips from Oct 2014 to Dec 2014
Freq_6	Number of trips from Jul 2014 to Dec 2014
Freq_12	Number of trips from Jan 2014 to Dec 2014
Av.Tran.Size	Average amount of money spent in each transaction
Av.Tran.Size_3	Average amount of money spent in one transaction between Oct 2014 and Dec 2014
Av.Tran.Size_6	Average amount of money spent in one transaction between Jul 2014 and Dec 2014
Av.Tran.Size_12	Average amount of money spent in one transaction between Jan 2014 and Dec 2014
Miles	Number of miles flown in Jan 2015
Miles_3	Number of miles flown between Oct 2014 and Dec 2014
Miles_6	Number of miles flown between Jul 2014 and Dec 2014
Miles_12	Number of miles flown between Jan 2014 and Dec 2014
Longevity	Number of days since first transaction
Recency	Number of days since last transaction
Age	Age of the customer
MonthlyRecency	How many months before the customer has flown between Jan 2014 and January 2015

The *MonthlyRecency* attribute indicates how many months before the last flight of the customer, starting from the end of the modeling time interval (the closed time interval of 01.2014 – 01.2015). The calculation starts with 1 for January of the year 2015 and this value is increased by 1 for each month that the person has no flight to the backwards in the time axis. According to the rule, the value of *MonthlyRecency* attribute will be 2 for a person, whose last flight in November, 2014. As another example, the value of this attribute will be 5 for a person whose last flight in August, 2014.

Since revenue derived from the customers is private and is directly used in calculation of CLV, it should somehow be represented in the models. For this purpose, *Status Miles* attribute of the collected flight data is used. Checking gathered data and written documentation, it is confirmed that the higher the ticket prices, the higher the *Status Miles*. So, values of *Status Miles* have been used as the values of *Value_New* variable in Table 3.2. The descriptive statistics of the variables of the flight data are given in Appendix A.

3.2. Social Network Data

An anonymous social network dataset is used for illustrating the social network part of the study. This dataset is obtained from the survey participants of a Facebook application implemented by Stanford University researchers. It is publicly available from the web site of the Stanford Network Analysis Project (SNAP) (Lescovec, SNAP: Network Datasets: Social Circles).

The Facebook dataset is anonymized by replacing the users' Facebook unique identifiers with new identity values. In addition, the interpretation of the node features (profiles) are obscured by changing the values with anonymized data. This social network consists of 10 people that are analyzed (egos) and their friends (alters) (Lescovec, SNAP: Network Datasets: Social Circles). The social network unique identifiers of the egos are given in Table 3.3.

Table 3.3: The social network unique identifiers of the egos.

0	107	348	414	686	698	1684	1912	3437	3980
---	-----	-----	-----	-----	-----	------	------	------	------

A graph data structure is generated to represent this social network, using the *igraph* package of *R* software. The nodes of the graph represent the customers; the edges represent the friendship relationships between the customers. This graph, which has 4039 nodes and 88234 edges, is visualized in Figure 3.1.

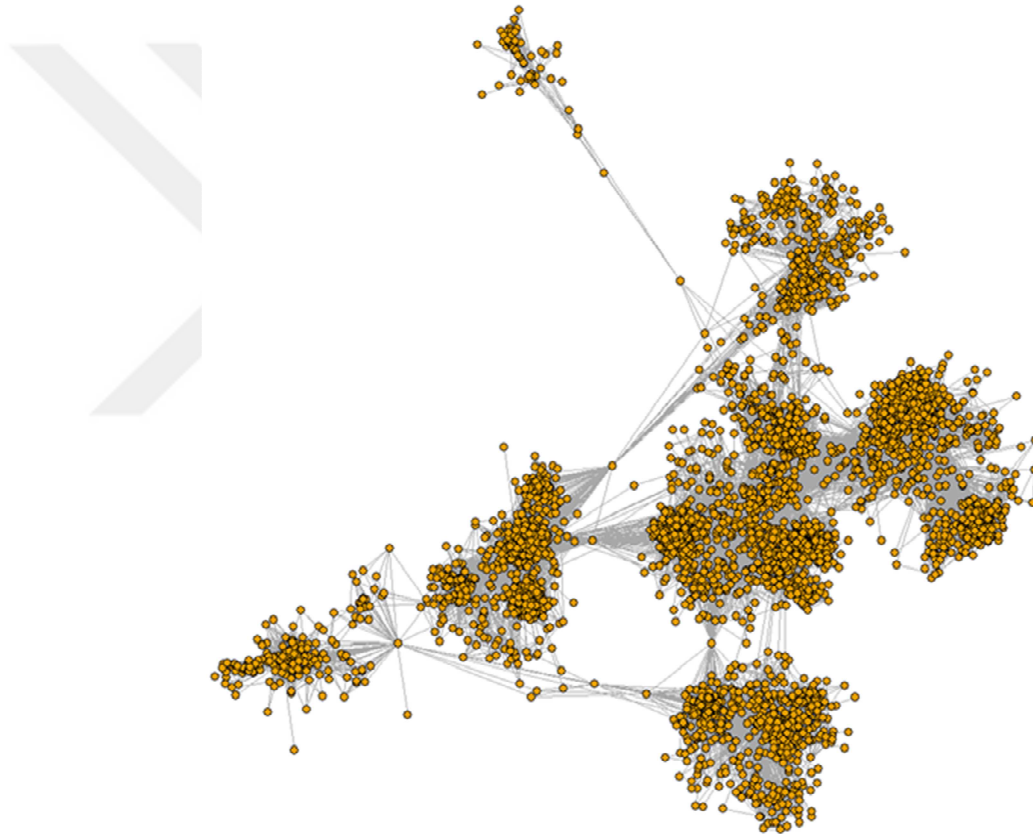


Figure 3.1: Graph of the Facebook data.

The graph data structure is formed using *graph.data.frame* function of the *igraph* package. Then, five structural social network scores, namely degree centrality, closeness centrality, betweenness centrality, PageRank and hub score are calculated in order to be used in integration of social dimensions to traditional data analytics methods. Following is a brief explanation of these measures together with the illustrations.

An example network is given in Figure 3.2 in order to make clear the social score definitions used throughout this study. This is a very simple undirected network containing 6 nodes and 7 edges.

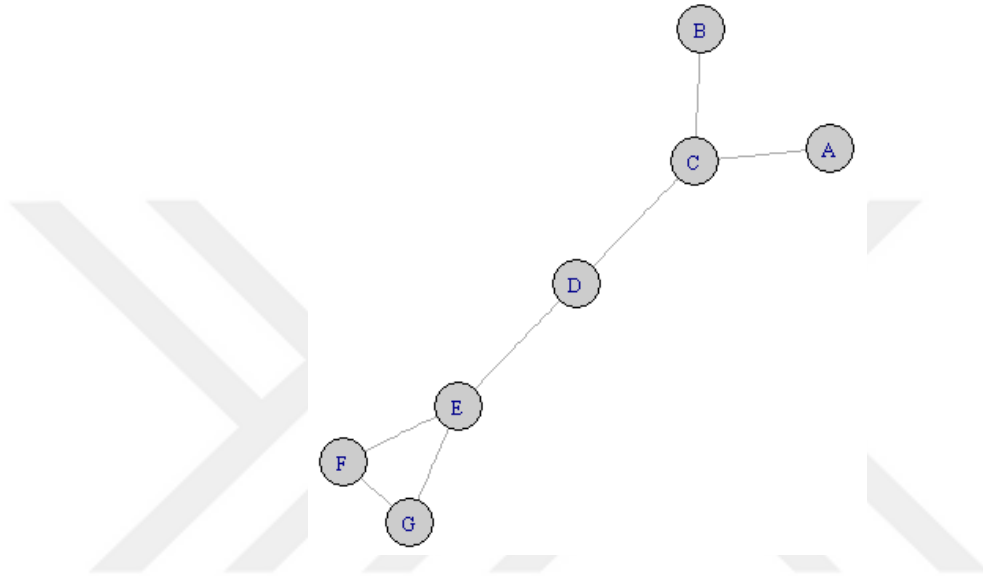


Figure 3.2: Graph of the example network used in social score explanations.

In graph theoretical terminology, the degree centrality, $d(i)$ of a node i , can be defined as the number of edges adjacent to node (i). In Equation 3.1, the definition of it is given in mathematical terms:

$$d(i) = \sum_j m_{ij} \quad \text{(Equation 3.1)}$$

where $m_{ij} = 1$ if there is an edge between nodes i and j , and $m_{ij} = 0$ if there is no edge between them. The standardized degree centrality, $\mathbf{d}_s(\mathbf{i})$, is obtained by dividing $\mathbf{d}(\mathbf{i})$ by $(N - 1)$, where N is the number of nodes in the graph (Otte & Rousseau, 2002) (Freeman, 1978).

Degree centrality graph of the example network is given in Figure 3.3. The sizes of the nodes in this graph increases in proportion to their degree centrality value. In order to emphasize how degree centrality is calculated, the edges to its direct

neighbors of the node D are drawn wider and in orange color. The degree centrality values of all the nodes in the example network are given in Table 3.4.

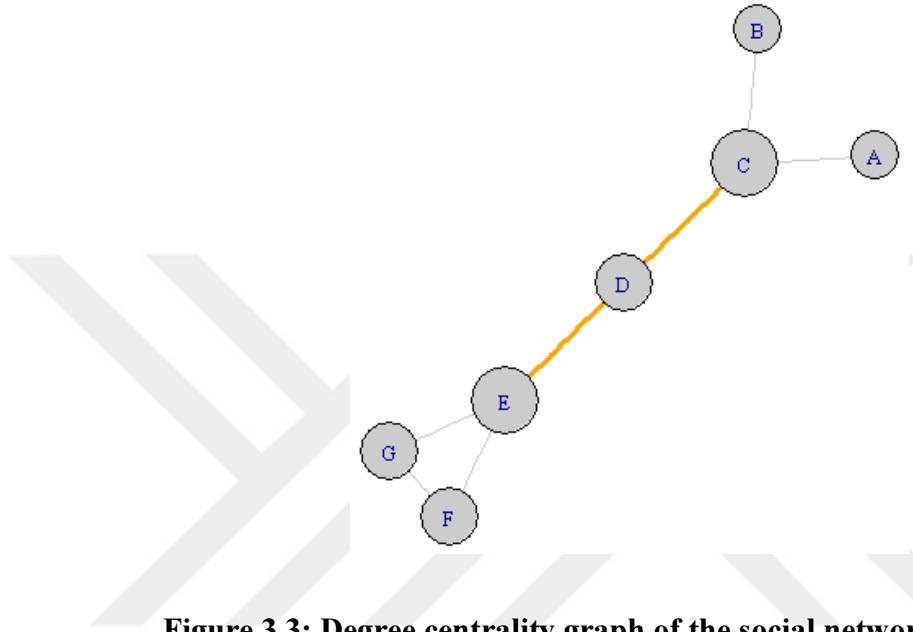


Figure 3.3: Degree centrality graph of the social network data.

Table 3.4: Degree centrality values of the example network.

	A	B	C	D	E	F	G
$d(i)$	1	1	3	2	3	2	2
$d_s(i)$	0.17	0.17	0.5	0.33	0.50	0.33	0.33

The second measure, closeness centrality, is defined as the total distance of a node in the graph from all other nodes (Otte and Rousseau, 2002). This centrality measure can be defined as a mathematical formula as given in Equation 3.2:

$$c(i) = 1/\sum_j d_{ij} \quad \text{(Equation 3.2)}$$

where $c(i)$ is the closeness centrality of node i , d_{ij} is the number of edges in a shortest path from node i to node j (Otte and Rousseau, 2002) (Freeman, 1978). This measure needs to be standardized by multiplying it 1 minus the number of nodes in the network ($N - 1$) as in Equation 3.3 so that it can be used directly as other measures.

$$c_s(i) = (N - 1) * c(i) \quad \text{(Equation 3.3)}$$

The closeness centrality values of all the nodes in the example network are given in Table 3.5. The unstandardized closeness centrality value of the node D is calculated as 0.10 whose calculation details are given in Equation 3.4. When this value is multiplied by $(N - 1)$, the standardized closeness centrality value is obtained which is 0.60 in this case.

$$c(D) = 1/(2 + 2 + 1 + 1 + 2 + 2) = 0.10 \quad \text{(Equation 3.4)}$$

Table 3.5: Closeness centrality values of the example network.

	A	B	C	D	E	F	G
$c(i)$	0.06	0.06	0.09	0.10	0.09	0.07	0.07
$c_s(i)$	0.38	0.38	0.55	0.60	0.55	0.40	0.40

The third one, betweenness centrality, measures the number of the node's existence in the shortest path between another two nodes in the network. Betweenness centrality, $b(i)$, of node i is defined in mathematical terms as in Equation 3.5:

$$b(i) = \sum_{j,k} \frac{g_{jik}}{g_{jk}} \quad \text{(Equation 3.5)}$$

The standardized form of betweenness centrality $b_s(i)$ is obtained by dividing $b(i)$ through the number of pairs of vertices except from i , which is $(N - 1)(N - 2)$ for directed graphs and $(N - 1)(N - 2)/2$ for undirected graphs (Otte and Rousseau, 2002) (Freeman, 1978).

The betweenness centrality values of all the nodes in the example network are given in Table 3.6. The unstandardized value of the node D is calculated as 9 with the help of Table 3.7 which consists of g_{jk} , g_{jik} and g_{jik}/g_{jk} values calculated for each node in the network except from the node D itself. Since the example network is an undirected one, the standardized value, which is 0.60 in this case, is obtained by dividing this value by $(N - 1)(N - 2)/2$.

Table 3.6: Betweenness centrality values of the example network.

	A	B	C	D	E	F	G
$b(i)$	0	0	9	9	8	0	0
$b_s(i)$	0.00	0.00	0.60	0.60	0.53	0.00	0.00

Table 3.7: Betweenness centrality calculation details of the example node.

	g_{jk}	g_{jik}	g_{jik}/g_{jk}
A	2	3	1.5
B	2	3	1.5
C	2	3	1.5
E	2	3	1.5
F	2	3	1.5
G	2	3	1.5
Total			9

The fourth measure, PageRank, is an algorithm developed mainly for grading the web sites in the search results of Google Search. It is mainly used for measuring the importance of the web sites; however, it is also used in social network analysis to measure the importance of the nodes.

PageRank can be defined shortly as an iterative “voting” of all other pages in the web about the importance of a page. A link to a page is accepted as a supporting vote, in other words, no link means no vote.

Brin and Page (2012) define the PageRank of web page A as in Equation 3.6:

$$PR(A) = (1 - d) + d(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)) \quad \text{(Equation 3.6)}$$

The PageRank values of all the nodes in the example network are given in Table 3.8.

Table 3.8: PageRank values of the example network.

	A	B	C	D	E	F	G
$PR(i)$	0.09	0.09	0.23	0.14	0.19	0.13	0.13

Since calculation of the PageRank values requires solving as many equations as the number of nodes in the network, for the sake of simplicity and clarity, only the PageRank equations of the example network are given in equations between Equation 3.7 and Equation 3.13. There are seven equations here and seven unknowns, namely PageRank values of each node in the network. The PageRank values in Table 3.8 should be obtained after solving those linear equations. In these equations, the value of dumping factor, d , is used as 0.85, which is the recommended default in the article.

$$PR(A) = 0.15 + 0.85 (PR(C)/3) \quad \text{(Equation 3.7)}$$

$$PR(B) = 0.15 + 0.85 (PR(C)/3) \quad \text{(Equation 3.8)}$$

$$PR(C) = 0.15 + 0.85 (PR(A) + PR(B) + PR(D)/2) \quad \text{(Equation 3.9)}$$

$$PR(D) = 0.15 + 0.85 (PR(C)/3 + PR(E)/3) \quad \text{(Equation 3.10)}$$

$$PR(E) = 0.15 + 0.85 (PR(F)/2 + PR(G)/2 + PR(D)/2) \quad \text{(Equation 3.11)}$$

$$PR(F) = 0.15 + 0.85 (PR(G)/2 + PR(E)/3) \quad \text{(Equation 3.12)}$$

$$PR(G) = 0.15 + 0.85 (PR(F)/2 + PR(E)/3) \quad \text{(Equation 3.13)}$$

The last social network measure used in this study, Hyperlink-Induced Topic Search, is an algorithm that analyzes the links in order to rate the web pages. In this method, a page pointing to many other pages is represented as a good hub and a page which is linked by many different hubs is represented as a good authority. The numerical weights of the pages are calculated using the relationships between hubs and authorities by means of an iterative algorithm (Kleinberg, 1999).

Since calculation of the hub score values requires execution of an iterative algorithm, only the results obtained by executing the *hub.score* function of *R*, are given in Table 3.9 for this network measure.

Table 3.9: Hub score values of the example network.

	A	B	C	D	E	F	G
<i>HS(i)</i>	0.21	0.21	0.48	0.66	1.0	0.80	0.80

3.3. Modeling Supported by Network Information

In order to clearly define the proposed methodology that aims to integrate social network information into traditional data analytical techniques, a flowchart is presented in Figure 3.4 showing the steps of the process. The details of those steps are given in the subsections of this section after the techniques used during the study as well as changes made on the data are briefly explained.

In this study, multiple linear regression technique is used as traditional data analytical method for modeling CLV. Multiple linear regression is an approach to model the relationship between the numerical dependent variable and two or more explanatory variables (or independent variables or factors). The general form of multiple linear regression is given in Equation 3.14:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} + e_i \quad \text{(Equation 3.14)}$$

where y_i represents the dependent variable, $x_{1,i} \dots x_{k,i}$ represent k independent variables, β_0, \dots, β_k represent the regression coefficients and e_i represents the error term (Hyndman & Athanasopoulos, 2014). Regression models are constructed using *lm* function from *stats* package of *R* software.

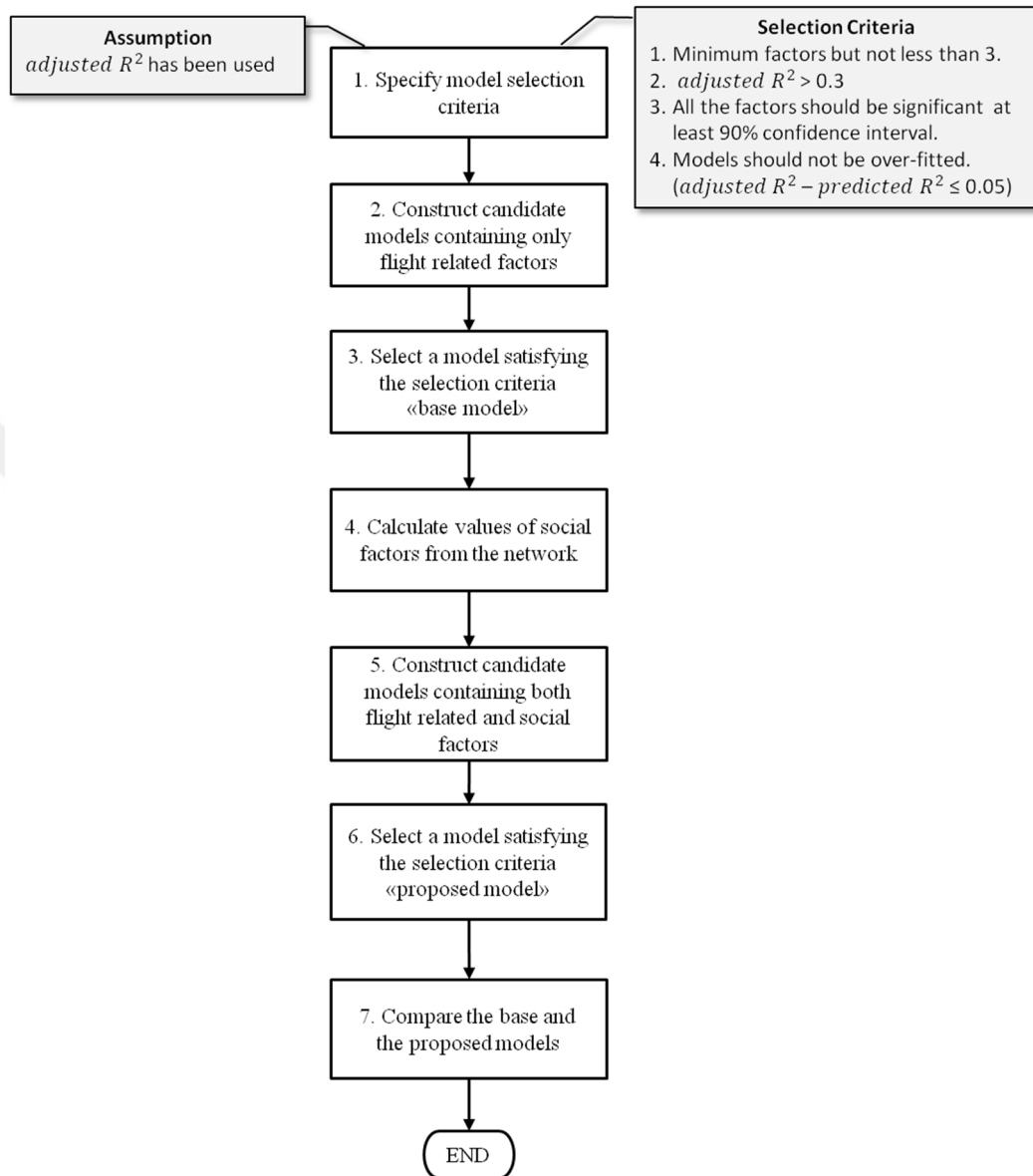


Figure 3.4: Flowchart of the proposed methodology.

Although the dataset contains 20 independent variables, only 16 of them have been used in the models. The variables, *Longevity*, *Recency*, *Age* and *MonthlyRecency*, have been omitted because the variabilities of these factors are small for the sample used.

One of the challenges in modeling using multiple linear regression is to select the factors that result in more accurate models or in other words models that have

better determination coefficients. This becomes more problematic when the number of factors is increased. Another challenge of regression modeling is that obtaining as high accuracy as possible with minimum number of factors.

The technique known as *Best Subset Regression* (Makridakis, et al., 1998) has been developed to overcome those challenges. Given that, the dependent variable, candidate factors, number of best models to be returned, the maximum number of factors in the resulting models and together with some other optional parameters, it searches for the best models having different number of independent variables. For this purpose, the *regsubsets* function from the *leaps* package of *R* software is used throughout this study.

3.3.1. Specification of Model Selection Criteria

The first step of the process in Figure 3.4 is to determine the model selection criteria, which guides the modeler to choose between alternative models. At this stage, the model selection criteria that is not very challenging but still suitable for making statistical inferences have been chosen in accordance with the purpose of the study.

The *adjusted R^2* is used for comparing the accuracy of the models. Since it is not sensitive to the number of factors in the models, it is preferred over R^2 (or *multiple R^2*). Actually, this is an assumption rather than a selection criterion.

Because of the need to work with data in a limited size, the number of independent variables of the models to be developed is an important model selection criterion. The models should contain as few independent variables as possible but not less than three. Such a minimum number constraint is added in order to let the more important social factors replace the less important flight related factors.

A very modest minimum *adjusted R^2* value of 0.3 is set as an acceptance criterion for models. Models having *adjusted R^2* value of less than 0.3 is rejected.

All the independent variables of the model to be selected must be significant at least in the 90% confidence interval. This criterion is checked using the probability values of the regression coefficients in the model summary.

The models should not be over-fitted. Over-fitted models refer to those that seem to explain the relationship between the dependent variable and the independent variables for the dataset used during model creation, but that fail to make valid predictions for new observations. The coefficient of determination used for understanding whether the models are over-fitted is called the *predicted R^2* . Significant range of values of this coefficient is between 0 and 100%. The greater the value of this coefficient, the ability to predict new observations of the model increases. If the difference between *adjusted R^2* and *predicted R^2* values of a model is greater than 0.05, the model would be accepted as over-fitted and be rejected.

In real life scenarios where all the flight transactions of the airline companies are available, it is expected that the minimum number of independent variables and minimum *adjusted R^2* value constraints would be much higher. In addition, models also would not be over-fitted because the data size would be much higher than this case.

3.3.2. Construction of Base Model Candidates

As the second step of the process, the base model candidates containing only flight related factors have been constructed. For this purpose, the *regsubsets* function of *R*, an implementation of Best Subset Regression technique, has been used. The arguments passed to the *regsubsets* function call are given in Table 3.10.

By using this function with the arguments in Table 3.10, only the best model of each subset is selected among the model subsets containing at most 10 independent variables. The plot of the *regsubsets* function's output is given in Figure 3.5.

Table 3.10: Arguments passed to 'regsubsets' to find the base model.

Argument	Value	Description
x	Value_New ~ Value + Value_3 + Value_6 + Value_12 + Freq + Freq_3 + Freq_6 + Freq_12 + Av.Tran.Size + Av.Tran.Size_3 + Av.Tran.Size_6 + Av.Tran.Size_12 + Miles + Miles_3 + Miles_6 + Miles_12	This argument takes the model in different formats: design matrix, model formula for full model, biglm object. The model formula has been passed in this case.
data	clv_data	Actual flight data of 15 people containing the attributes given in Table 3.2.
nbest	1	When 1 is used, the function returns only one best model in each of the n-variable subsets.
nvmax	10	Maximum size of the subsets to examine. This means that the resulting models contain at most 10 independent variables.

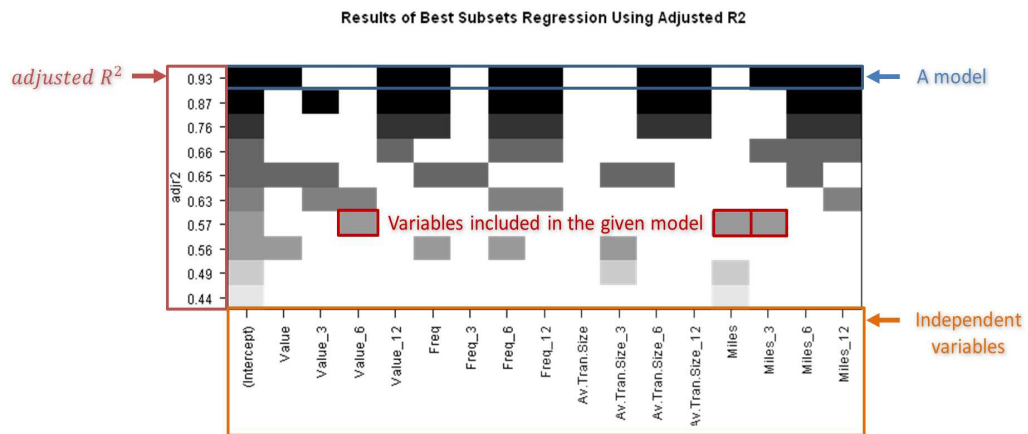


Figure 3.5: Best Subsets Regression plot for discovery of the base model.

The output should be interpreted as follows:

Each row of the matrix represents a model; the predictors included in the given model are symbolized as shaded rectangles in the columns. The values of the determination coefficient used (such as, *adjusted R²*, BIC, etc.) are on the left margin; note that this axis is ordered but not quantitative. The level of the darkness

shows the ordering of the determination coefficient's values: the darker is better (Taylor, 2016).

3.3.3. Base Model Selection

The third step of the process requires selection of a model that satisfies the model selection criteria specified in the first step. As a summary, the model should contain at least three independent variables all of which are significant at least 90% confidence interval, *adjusted R²* value of the model should be greater than 0.3 and the model should not be over-fitted.

Among the candidate models in Figure 3.5, all models starting from the third row onwards meet the criteria related to the number of independent variables and the *adjusted R²* value. So, the model with three variables (*Value_6*, *Miles* and *Miles_3*) which is located on the fourth row of the matrix has been selected. Note that, compared to the model in the third row, this model contains less variables but has a higher *adjusted R²* value.

In order to check that the criterion related to significance of the independent variables is met, the regression coefficients together with the probability values that determine the significance of those must be calculated. For this purpose, the *lm* function from the *stats* package of *R* software has been used by passing the flight data of 15 people and the model formula which is constructed with respect to fourth row of the matrix in Figure 3.5 (*Value_New ~ Value_6 + Miles + Miles_3*) as arguments. The resulting regression formula of this initial model has been constructed as shown in Equation 3.15. The *adjusted R²* and *predicted R²* values of the initial model are 0.57 and 0.36 respectively. The measures used to interpret the significance of the factors are given in Table 3.11. The explanations about the measures themselves are given in Table 3.12.

$$Value_New_i = 3188.7250 - 0.6557(Value_6_i) + 4.5634(Miles_i) + 1.1508(Miles_3_i) \quad \text{(Equation 3.15)}$$

Table 3.11: Significances of the factors of the initial model.

Factor	Std. Error	t value	Pr(> t)	Significance Code
Value_6	0.3441	-1.906	0.08313	.
Miles	1.1874	3.843	0.00273	**
Miles_3	0.5019	2.293	0.04255	*

As the last model selection criterion, it is checked whether the initial model in Equation 3.15 is over-fitted. Remember from Section 3.3.1 that, according to the model selection criteria specified, if the difference between the values of *adjusted R²* and *predicted R²* is greater than 0.05, it is considered that the model is over-fitted. This difference is 0.21 for the initial model, so it is concluded that the initial model is over-fitted.

It is known that there is a strong relationship between the data size and the value of *predicted R²*. So, it is decided to generate flight data synthetically.

Table 3.12: Explanations about the significance measures.

Measure	Explanation
Std. Error	"The standard deviation of the sampling distribution of the estimate of the coefficient under the standard regression assumptions." (Stack Exchange, n.d.)
t value	"Value of the t-statistic for testing whether the corresponding regression coefficient is different from 0." (Stack Exchange, n.d.)
Pr(> t)	"p-value for the hypothesis test for which the t value is the test statistic." (Stack Exchange, n.d.)
Significance Code	Categorizations of p-values around typical confidence intervals. (.): $p < 0.1$ (*): $p < 0.05$ (**): $p < 0.01$ (***): $p < 0.001$

After deciding that synthetic data generation is necessary, the question of how much data to generate is arisen. This question is addressed with the help of Equation 3.16 where

n : Necessary sample size

z_c : Z-score of the expected confidence level

s : Expected standard of deviation of the sample

E_m : Expected margin of error of the sample (Smith, n.d.)

$$n = \frac{(z_c)^2 * s * (1 - s)}{(E_m)^2} \quad \text{(Equation 3.16)}$$

In order to calculate the necessary synthetic data size, typical values given in Table 3.13 of the relevant parameters are used. When the parameter values are substituted in Equation 3.16, the expected value of synthetic data size is calculated as 68 as shown in Equation 3.17.

Table 3.13: Values of the variables used in synthetic data size calculation.

Variable	Value	Explanation
c	0.90	Confidence level which can be defined as probability of actual mean falls within the confidence interval.
z_c	1.645	Z-score of the 90% confidence level
s	0.5	Standard of deviation
E_m	$\pm 10\%$	Margin of error (confidence interval) of the sample

$$n = \frac{(1.645)^2 * 0.5 * (1 - 0.5)}{(0.1)^2} = 67.6 \quad \text{(Equation 3.17)}$$

Although 68 samples are enough for %90 confidence level, it is decided to generate synthetic data for 200 customers.

As far as the independent variables in the initial regression model are concerned, it is evident that data distribution characteristics of the following variables are needed:

1. Monthly flight frequencies
2. Monthly total distances
3. Monthly total sales (monetary values)

Within this context, three charts showing the characteristics of the flight data is given in Figure 3.6, Figure 3.7 and Figure 3.8 respectively. This data have been used to discover the distribution characteristics of those three variables. When working with the data, it is also noticed that the data have clusters in terms of similar flight habits of the customers. In fact, this is not so surprising because the Mile Programs of the airline companies also have different membership categories. It is observed that there are three categories in the dataset which are named as 1, 2 and 3 for the sake of anonymity.

For the first variable, the random frequencies have been generated for each category with respect to Poisson distribution using each category's average flight frequency as λ parameter. The number of randomly generated frequencies in each category (n) has calculated using the formula in Equation 3.18 where n_c represents the number of people in relevant category, n_f represents the number of people in flight data sample, and n_s represents the number of nodes in social network data sample.

$$n = \frac{n_c}{n_f} (n_s) \quad \text{(Equation 3.18)}$$

For the second variable, the random distances have been generated for each category with respect to normal distribution using each category's average of total distance flown in respective month as the mean and each category's standard deviation of total distance flown in respective month as the standard deviation. The number of randomly generated distances in each category (n) is calculated using the formula in Equation 3.18.

For the third variable, the random sales have been generated for each category with respect to normal distribution using each category's average of total sales in respective month as the mean and each category's standard deviation of total sales in respective month as the standard deviation. The number of randomly generated sales in each category (n) is calculated using the formula in Equation 3.18.

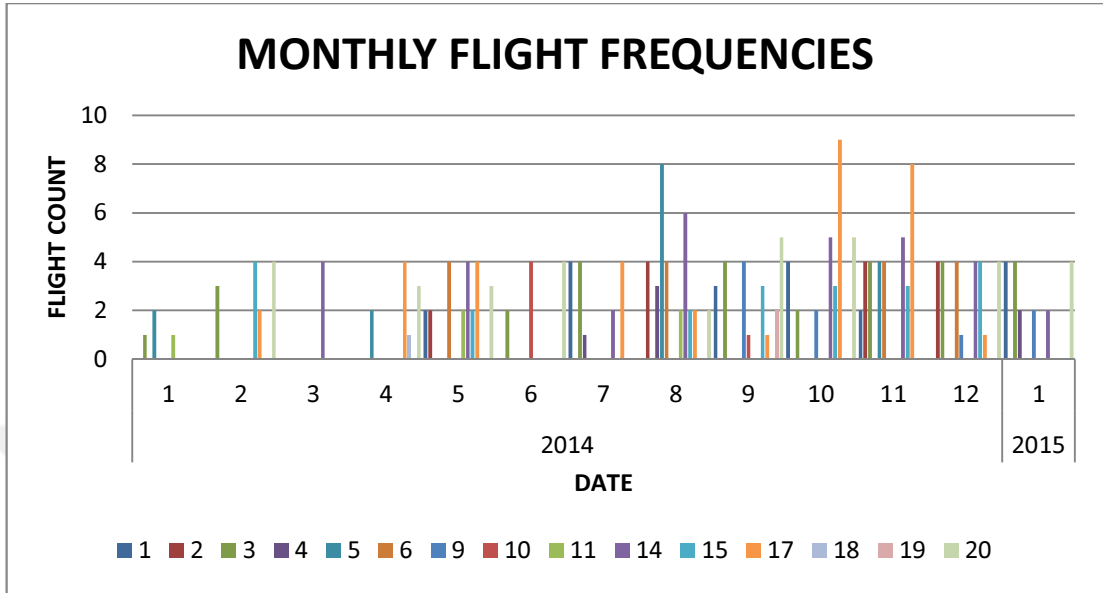


Figure 3.6: Monthly flight frequencies.

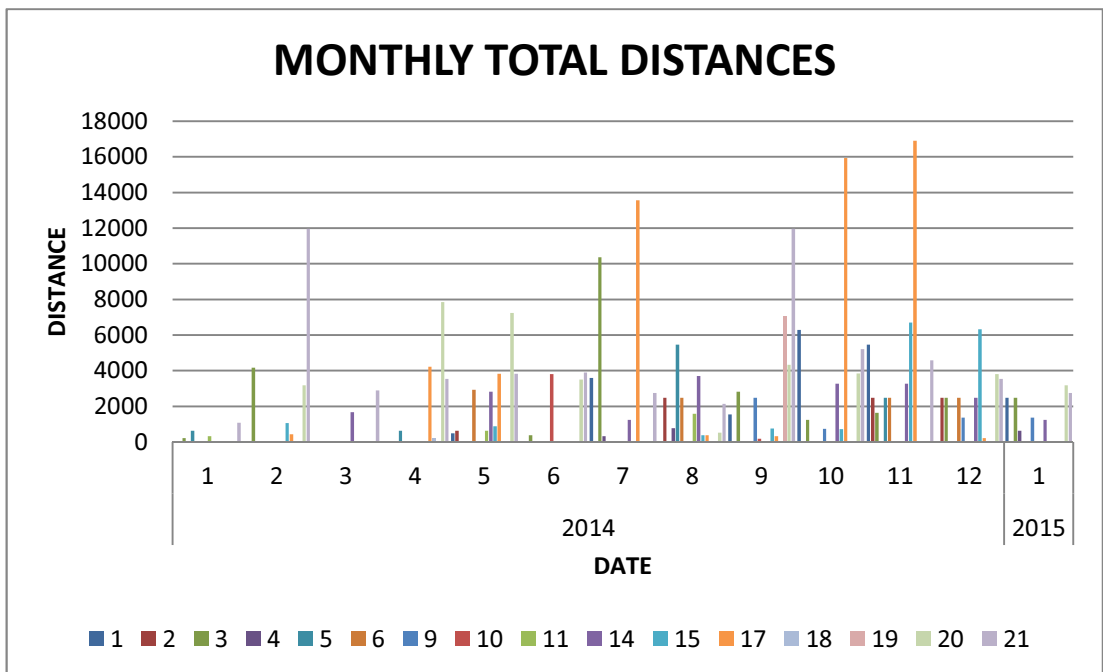


Figure 3.7: Monthly total distances.

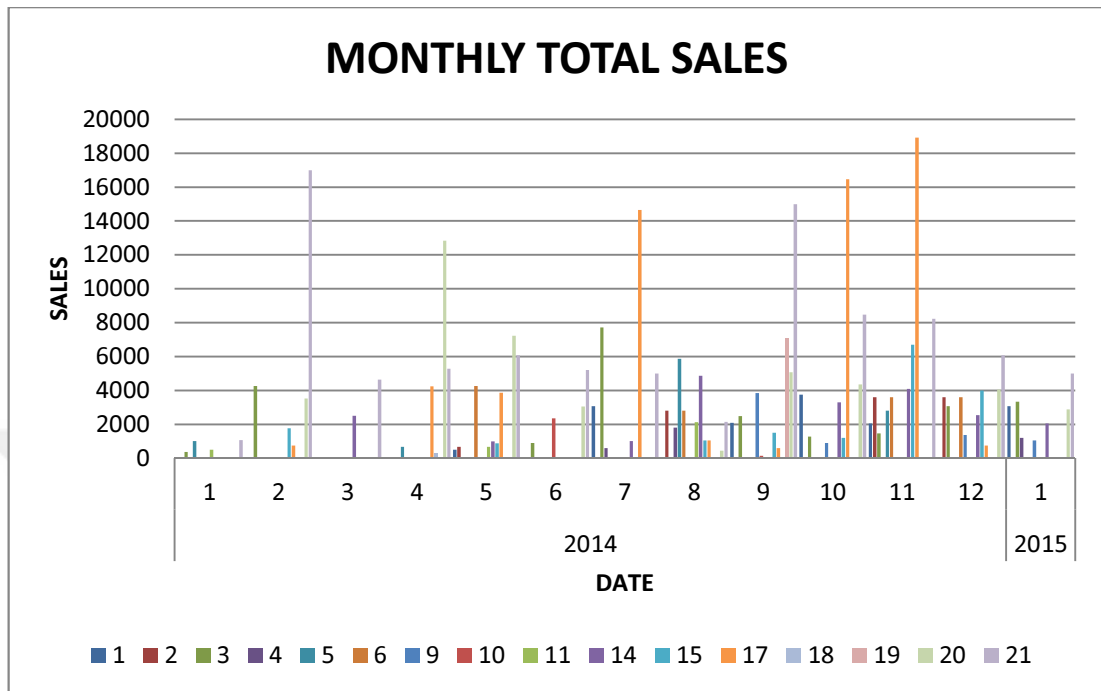


Figure 3.8: Monthly total sales.

Random data have also been generated according to normal distribution for the dependent variable, *Value_New*, which is significant on annual basis. The mean and standard deviation parameters necessary for random value generation are obtained from respective category's average and standard deviation values which are calculated from the actual flight data for each category.

Lastly, all flight related data in Table 3.2 are obtained synthetically by taking the cumulative sums of generated monthly data for the respective months.

After synthetic flight data generation is completed, the regression coefficients of the initial model have been re-calculated using the dataset extended synthetically. The resulting regression formula of this model has been constructed as shown in Equation 3.19. The *adjusted R²* and *predicted R²* values of this model are 0.33 and 0.27 respectively. The measures used to interpret the significance of the factors are given in Table 3.14.

$$\begin{aligned}
 ext(Value_New_i) = & -2703687.4 + 599.4(Value_6_i) + 6028.8(Miles_i) - \\
 & 377.8(Miles_3_i)
 \end{aligned}
 \tag{Equation 3.19}$$

Table 3.14: Significances of the factors of the model using extended data.

Factor	Std. Error	t value	Pr(> t)	Significance Code
Value_6	630.7	0.950	0.3430	
Miles	1350.9	4.463	1.32e-05	***
Miles_3	973.4	-0.388	0.6983	

Although the model in Equation 3.19 is not over-fitted any more (the difference of the determination coefficients is now 0.05), it still needs to be rejected because of the third selection criterion. With a quick examination of Table 3.14, it can be noticed that the p-values of Value_6 and Miles_3 factors are high. This means that, those variables are not significant anymore and violate the third selection criterion. Because of this reason, base model candidates need to be searched again using the extended dataset. For this purpose, the *regsubsets* function is called again by changing only the *data* argument. The Best Subset Regression plot that is generated after this function call is shown in Figure 3.9.

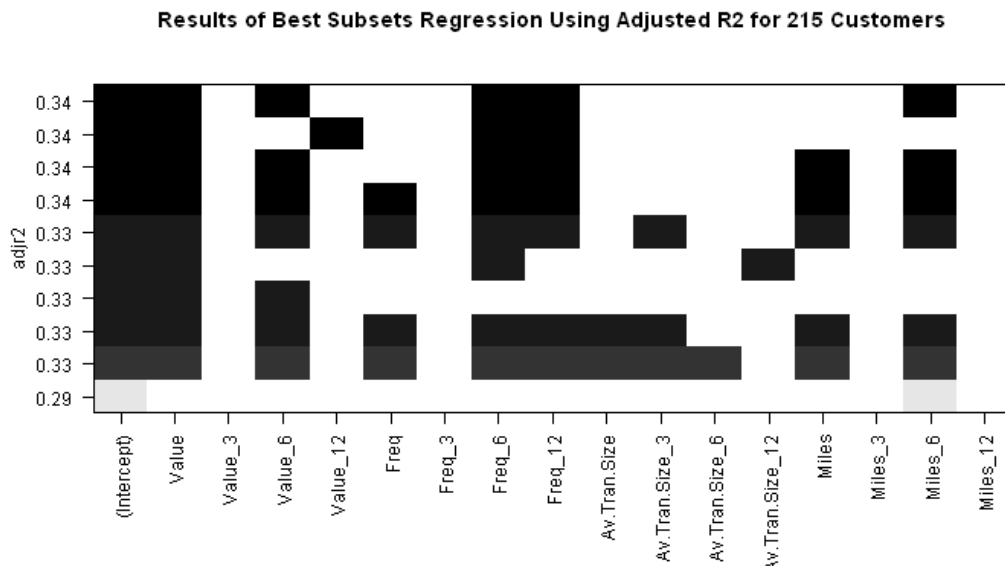


Figure 3.9: Best Subsets Regression plot using extended dataset.

At first glance, it can be seen that the model on the second row from the top is a strong candidate to meet the selection criteria. Because, this model has the highest *adjusted R²* value and contains minimum number of independent variables among the subsets having the same *adjusted R²* value. The exact regression coefficients together with their significance values are obtained by calling *lm* function. As a first argument of this function, the model formula ($Value_New \sim Value + Value_12 + Freq_6 + Freq_12$) should be passed, and as the *data* argument of this function, the extended dataset should be passed.

The resulting regression formula of this base model candidate has been constructed as shown in Equation 3.20. The *adjusted R²* and *predicted R²* values of the model are 0.34 and 0.29 respectively. All of the factors are significant with respect to the measures given in Table 3.15. The model is not over-fitted since the difference between its determination coefficients is less than or equal to 0.05. So, all of the model selection criteria are met and the 4-variable model in Equation 3.20 is accepted as the base model containing only flight related factors.

$$base(Value_New_i) = -2123484.1 + 4358.0(Value_i) + 331.1(Value_12_i) + 780760.3(Freq_6_i) - 602153.0(Freq_12_i) \quad \text{(Equation 3.20)}$$

Table 3.15: Significances of the factors of the base model.

Factor	Std. Error	t value	Pr(> t)	Significance Code
Value	1116.4	3.903	0.000128	***
Value_12	118.3	2.798	0.005621	**
Freq_6	345450.8	2.260	0.024840	*
Freq_12	306666.5	-1.964	0.050903	.

3.3.4. Integration of Social Network Information

Since the actual social network of the customers is not available in this study, a publicly available anonymous social network is used in the fourth step of the process. As a direct consequence of this situation, airline customers in the dataset need to be matched to the nodes of the social network in some way. To achieve this, customer

identifiers (*Id*) in the flight data need to be matched with the social network identifiers of the social network (*SN_Id*). In addition, flight and social network data that belong to matched people need to be consolidated. In this context, first, 10 of 15 real people are directly matched with the 10 egos of the social network. Next, the remaining 5 real customers are matched with 5 random nodes in the social network. Then, a dataset that contains the values of the variables in Table 3.16 for real Mile Program members is obtained. After that, synthetically generated flight data of 200 people are randomly matched to 200 nodes of the social network, that haven't matched before. The customer values (*Value_New*) of these 200 people are also updated according to the formula in Equation 3.21. Such an action is performed in order to model the impact of social network on the customer value. The real and synthetic data are consolidated by rows and the expanded dataset of 215 Mile Program members containing the attributes in Table 3.2 and Table 3.16 is obtained.

$$Updated_Value_New_i = (SN_PageRank_i * 1000 * Value_New_i) + Value_New_i \quad \text{(Equation 3.21)}$$

Next, the social network scores except from *SN_PageRank* are tried to be added as new independent variables to the base model via the expanded dataset. However, a model that satisfies the model selection criteria cannot be obtained. It is considered that this unexpected situation may result from the fact that the values of *SN_PageRank* are not differed enough between the nodes of the graph. In this context, rather than selecting the 200 people randomly, it is decided to select them according to variances of their social scores by following the procedure below:

- 1 Find the social score that has the highest variance.
- 2 Sort all the nodes according to the social score found in Step 1.
- 3 Select first 100 and last 100 nodes from the sorted list obtained in Step 2.
- 4 Update *Value_New* values of selected 200 people using the social score found in Step 1, after the flight data have been assigned.

Table 3.16: Variables of social network data.

Variable	Description
SN_Id	Unique Identifier of the social network node (Customer Id)
SN_Btw_Cntr	Betweenness centrality score
SN_Cls_Cntr	Closeness centrality score
SN_Degree	Degree centrality score
SN_HubScore	Hub score
SN_PageRank	PageRank score

When the above procedure is applied on the social network data, the highest variance is obtained by using hub score ($SN_HubScore$). Then, the unique identifiers of all the nodes but the 15, which match the real flight data, are sorted according to hub score in increasing order. Next, the first and the last 100 identifiers from this sorted list are selected as the unique identifiers of the people accepted as Mile Program members. The flight data for those 200 people are generated in the way explained before and combined dataset containing the values of the variables in Table 3.2 and Table 3.16 are constructed for them. After that, the values of $Value_New$ variables of those people are updated by using $SN_HubScore_i$ instead of $SN_PageRank_i$ in Equation 3.21.

3.3.5. Construction of the Proposed Model

When the fifth step of the process is reached, the social network scores obtained from real or synthetic network have been calculated and consolidated with the flight related data already. At this stage of the process, it is questioned whether a better model can be achieved if the base model is combined with social factors by using this consolidated dataset, which is called "expanded dataset". For this purpose, the *regsubsets* function is called again by passing the arguments given in Table 3.17. The plot of the *regsubsets* function's output is given in Figure 3.10.

Table 3.17: Arguments passed to 'regsubsets' to find the proposed model.

Argument	Value	Description
x	Value_New ~ Value + Value_12 + Freq_6 + Freq_12 + SN_Btw_Cntr + SN_Cls_Cntr + SN_Degree + SN_PageRank	The model formula which consists of the independent variables of the base model and the social network related ones.
data	combined_data_ms_ego	Expanded dataset of 215 people.
nbest	1	When 1 is used, the function returns only one best model in each of the n-variable subsets.
nvmax	8	Maximum size of the subsets to examine.

As the base model consists of four independent variables, in order to guarantee that comparison baseline has kept, a model with the same number of variables must be selected among the model candidates. So that, the model which is located in the fourth row from the bottom of the matrix shown in Figure 3.10 has been selected as the proposed model candidate.

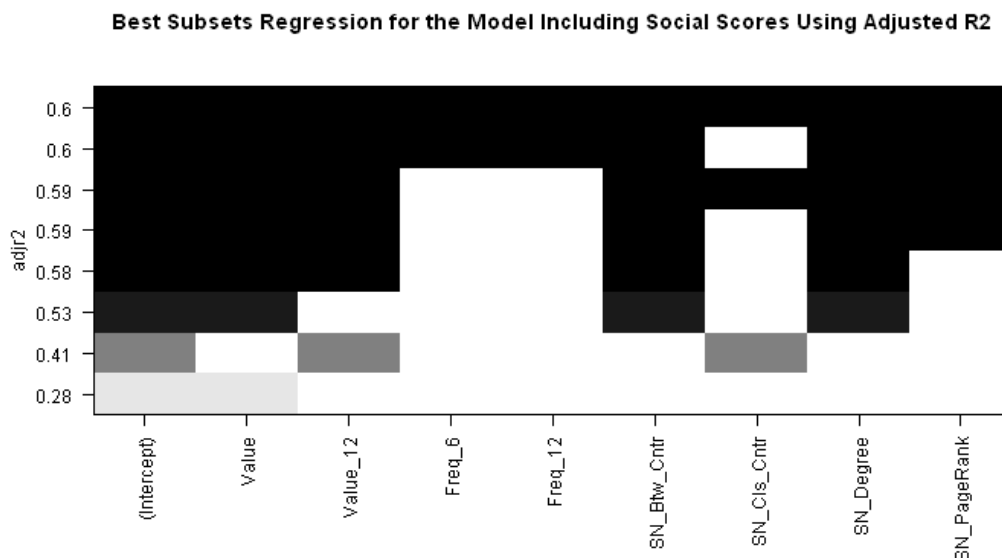


Figure 3.10: Best Subsets Regression plot of the model with social scores.

The resulting regression formula of this proposed model candidate has been constructed as shown in Equation 3.22. The *adjusted R²* and *predicted R²* values of the model are 0.58 and 0.56 respectively. All of the factors are significant with respect to the measures given in Table 3.18. The model is not over-fitted since the difference between its determination coefficients is 0.02. So, all of the model selection criteria are met for this model candidate also and the 4-variable model in Equation 3.22 is accepted as the proposed model containing both flight related and social factors.

$$\begin{aligned}
 \text{proposed}(\text{Value_New}_i) = & -7.863 \times 10^6 + 4.456 \times 10^3(\text{Value}_i) \\
 & + 2.514 \times 10^2(\text{Value_12}_i) \\
 & - 1.475 \times 10^8(\text{SN_Btw_Cntr}_i) \\
 & + 1.321 \times 10^8(\text{SN_Degree}_i)
 \end{aligned}
 \tag{Equation 3.22}$$

Table 3.18: Significance levels of the factors of the proposed model.

Factor	Std. Error	t value	Pr(> t)	Significance Code
Value	8.100e+02	5.501	1.09e-07	***
Value_12	4.928e+01	5.102	7.51e-07	***
SN_Btw_Cntr	1.596e+07	-9.241	< 2e-16	***
SN_Degree	1.160e+07	11.386	< 2e-16	***

After a model containing social scores is obtained, synthetic flight data are assigned to the rest of the people in the social network (3824 nodes) by applying the same method used in Section 3.3.3. Of course, the selection step is skipped since the entire network would be covered. In addition, just before the assignment process, the standard deviations of monthly flight distances and monthly sales are multiplied by 4 in order to address the differences in flight habits of the people outside of the Mile Program. Then, the *Value_New* values of those 3824 people are replaced with the estimated values from the model in Equation 3.22. In the end, the dataset of 4039 people which include both flight and social network related attributes is obtained.

The mile program category (*Category*), the customer profitability (*Profitability*) and the customer value display index (*ValueNewDisplayIndex*) attributes together with their values are added to this dataset in order to be used in customer analyses. Lastly, the graph data structure is constructed using *graph.data.frame* function from *igraph* package of *R* software by passing *SN_Id*, *Value_New*, *MonthlyRecency*, *Category*, *Profitability* and *ValueNewDisplayIndex* variables as *vertices* parameter and the connection information between the vertices (edges) from the Facebook data as *d* parameter.

3.3.6. Comparison of the Base and the Proposed Models

In the last step of the process, the base and the proposed models are compared and the results are discussed.

One of the methods used in comparison of statistical regression models is to compare the *adjusted R²* determination coefficient. The significant value range of this coefficient is between 0 and 1 and the greater the value of this coefficient, the prediction ability of the model increases.

In addition, checking the significance levels of the models' independent variables should also be considered when comparing the reliability of the models. The *p-values* of the independent variables calculated during the construction of the models should be used. Within the context of modeling with regression, the *p-value* of an independent variable represents the probability of accepting the null hypothesis that the true coefficient of the variable is zero. So, the smaller the *p-value*, the more significant the variable is.

The comparison of the base model and the proposed model in accordance with the explanations above is summarized in Table 3.19 and Table 3.20.

Table 3.19: Comparison of the determination coefficients of the two models.

	<i>adjusted R²</i>	<i>predicted R²</i>
Base Model	0.34	0.29
Proposed Model	0.58	0.56

Table 3.20: Comparison of the confidence levels of the two models' factors.

	Base Model	Proposed Model
Value	0.000128	1.09e-07
Value_12	0.005621	7.51e-07
Freq_6	0.024840	-
Freq_12	0.050903	-
SN_Btw_Cntr	-	< 2e-16
SN_Degree	-	< 2e-16

By examining Table 3.19, it can be concluded that with the addition of the social factors;

- The accuracy of the model has been increased by 24%.
- The prediction ability of the model has been also increased by 27%.

It can be concluded from Table 3.20 that addition of social factors to the base model makes its independent variables more significant also. It should be noticed that, the p-values of *Value* and *Value_12* in the proposed model have been decreased significantly compared to the base model. The p-values of the social factors are also very low, showing that they are strongly significant.

In order to demonstrate the impact of social scores on customer value quantitatively, the two models proposed in the study are compared in terms of customer values they generated. For this purpose, all the customers in the social network are sorted according to their *Value_New* values generated by the two models separately. Then, 500 most valuable customers from both models are picked up for comparison. It is observed that, 457 of 500 customers in the base model's ranking do not take place in the new ranking by the proposed model. In other words, when the social factors are added to the regression model, 91 percent of the 500 most valuable customers list is changed. The whole list of the 500 most valuable customer rankings is given in Appendix B.

The unique identifiers together with the generated customer values of the most valuable 60 customers' rankings for both models are given in Table 3.21. The

customers that are taken place in the 500 most valuable customer rankings of the proposed model but not in the base model's rankings are shaded in this table and in Appendix B. For example, the most valuable customer of the proposed model (whose ID is 656) does not take place in the base model's top 500 rankings. In this regard, only 3 of 60 customers in Table 3.21 are also in the rankings of the base model.

In order to see the correlation between the customer values generated by the two models, the scatter plot in Figure 3.11 is drawn. This chart is made up of customer values generated by the two models for the 100 customers with ID index 1-100. The customer values are sorted based on the proposed model.

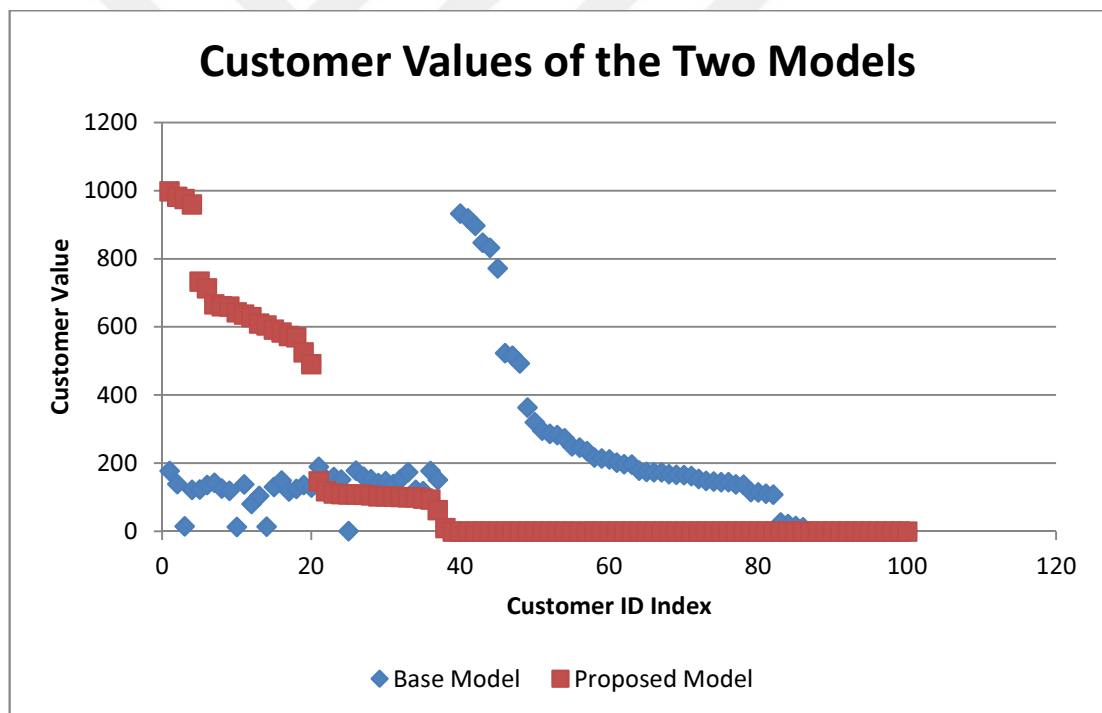


Figure 3.11: Customer Values of the Two Models.

It can be clearly seen from the figure that the two models may generate very different customer values for the same customer. This can cause employees of the CRM departments to make incorrect decisions.

Table 3.21: Top 60 of the models' customer value rankings.

No	Base ID	Base Value*	Prop. ID	Prop. Value*
1	920	998,03	656	999,19
2	694	996,56	2554	998,57
3	1284	995,93	68	998,32
4	678	995,24	1560	996,95
5	207	995,06	1685	996,52
6	2152	993,47	3398	996,18
7	235	992,65	2994	996,13
8	2881	992,13	3030	995,99
9	345	991,52	112	995,91
10	2667	991,23	629	995,62
11	1408	990,80	641	995,10
12	3889	990,79	2301	995,03
13	2782	990,73	3539	994,62
14	1456	989,79	2391	994,01
15	894	989,75	3932	993,98
16	1849	988,83	2171	991,89
17	784	988,64	478	991,84
18	1047	986,75	2805	991,77
19	1379	986,07	4017	991,57
20	3105	985,84	3419	988,91
21	3874	985,74	1865	988,31
22	3004	984,49	2653	988,27
23	1275	984,33	1730	988,09
24	3272	980,02	1701	987,93
25	1072	979,49	880	987,82
26	3150	979,03	2541	987,50
27	2669	978,84	3843	987,48
28	3703	975,60	3002	987,45
29	340	973,73	2904	986,41
30	2153	972,75	587	985,81
31	2026	971,73	3970	985,57
32	2474	971,70	1461	985,40
33	1768	971,61	2182	985,18
34	2630	970,76	91	982,68
35	2438	970,33	1087	982,54
36	1194	970,00	4019	982,29
37	2186	968,68	871	982,18
38	1391	968,64	3513	981,82
39	2530	967,73	2729	981,21
40	3467	967,14	3185	981,14
41	2083	965,34	431	980,24
42	1062	964,69	1273	979,87
43	1271	963,76	2499	979,29
44	1567	963,61	3108	978,98
45	2972	963,31	3270	978,93
46	2195	961,06	3864	978,87
47	2347	958,27	2763	978,55
48	3832	956,75	2824	978,45
49	1494	954,05	3716	978,07
50	1189	953,91	1718	977,80
51	1786	953,87	673	976,78
52	3403	953,31	3771	976,60
53	2683	951,98	8	975,98
54	146	951,46	3019	974,80
55	1179	950,99	593	974,63
56	3758	950,75	2767	973,25
57	3583	950,40	3953	973,16
58	2457	947,91	1181	973,03
59	2066	947,05	202	972,48
60	2637	946,46	2630	970,76

(*) The actual values calculated are divided by 10^7 .

To explain this situation with an example, assume that for decision making, CRM-1 uses the base model and CRM-2 uses the proposed model. With respect to the graph given in Figure 3.11, the customer values generated by the base model and the proposed model for the customer whose ID index is 2 (Customer-2) are 138.7 and 982.7 respectively. Also assume that Customer-2 who is socially very prominent, wants to leave from the Mile Program of the airline company. Since his customer value with respect to the base model is relatively low and CRM-1 has no other measure except from its flight related CLV, it would be easier for CRM-1 to let him leave without having taken any action in the context of customer retention. However, from the CRM-2's perspective, the situation is more difficult because Customer-2 is a valuable customer. In addition, by evaluating his customer value on the social network graph similar to customer lifetime value analysis plot given in Figure 4.2, CRM-2 is able to see that he is very prominent in the social network. CRM-2 is able to analyze his status by means of the decision support facilities as discussed in Section 4.1 which are not available in traditional data analytical techniques. As a result, CRM-2 can make more comprehensive decisions.

CHAPTER FOUR

4. RESULTS AND DISCUSSION

In this study, a methodology has been proposed explaining how the social network information can be integrated into the data analytical models. It is questioned whether the traditional data analytical models could be improved with the information obtained from social networks. Modeling customer lifetime value of the airline customers is chosen as the example case.

Firstly, the steps of the proposed methodology for integrating social network information into the data analytical methods have been presented (See, Figure 3.4). Next, those steps have been applied to the sample case and findings have been evaluated (See, Section 3.3). In order to compare the effects of social factors in the model, an initial model using only flight data has been created first. This model calculates customer value with respect to regression formula given in Equation 3.15. Observing that this model has over-fitting problem resulting from lack of enough data, synthetic flight data have been generated to increase the sample size. When the larger flight data are used, some of the factors in the initial model have become insignificant and the model has been rejected. Model candidates have been generated again using expanded dataset. The model whose regression formula is given in Equation 3.20 has been accepted as the base model. The *adjusted R²* and the *predicted R²* values of the base model are 0.34 and 0.29 respectively.

For social network analysis, the social scores; degree centrality, closeness centrality, betweenness centrality, page rank and hub score are calculated first. Then, flight data are assigned to the people in the social network as described in Section 3.3.3. After the social scores are added as new independent variables to the base regression model, a better model in terms of the value of *adjusted R²* among the expanded independent variable set are searched using best subset regression

technique. In order to preserve comparison baseline, particular attention is paid to the model obtained at the end of selection process consisting of four independent variables as in the base model. In the end, the model with the regression formula given in Equation 3.22 has been reached. It should be noted that, in the new model two of the independent variables are social scores. *adjusted R²* and *predicted R²* values of this new model are 0.58 and 0.56 respectively. This means a 24 percent increase in the *adjusted R²* value and 27 percent increase in the *predicted R²* value compared to the base model.

The 500 most valuable customer rankings of both models; the base model and the proposed model are also examined in terms of the customer values they generated. It is observed is that the models calculate very different customer values. There are only 43 people that take place in both rankings. However, all of them are in different order. It can be concluded from here that with the addition of the social factors the models become radically different.

If above results are combined with the research findings of (Trusov, et al., 2009) that people affect each other in social networks or in other words neighbors in the network do similar actions, it can be concluded that using the proposed model for determination of customer value in airline industry may produce more comprehensive results especially for people who fly rarely but have high social impact.

One of the major achievements obtained by adding social factors to the traditional data analytical models is the ability to analyze the customers together with their social networks and visualize the results of the analyses directly in the network. In the subsections of Section 4.1, customer analyses that can be made using social network supported models are exemplified.

Flight patterns of airline customers as well as their relationships in the social network change over time. In today's fiercely competitive air transportation sector dynamics, determining customer lifetime value considering these changes is a critical issue that will provide a competitive advantage for airline companies. In Section 4.2, a method is proposed for identifying customer value that considers the changes in

customers' relationships in the social network as well as their flight patterns during specified time period. Since, this method requires use of the historical flight and social network data which is not available in this study, the topic is remained as a future work.

4.1. Airline Customer Data Analytics

4.1.1. Customer Lifetime Value Analysis

This analysis aims to categorize the customers with respect to their CLV and visualize them in the social network graph. In the example here, this aim is realized by using *Value_New* and *ValueNewDisplayIndex* variables as vertex attributes in the social network graph. As recalled from previous sections, the CLVs of the customers are represented with the *Value_New* variable. In order to represent the CLV categories, it is also derived *ValueNewDisplayIndex* variable from *Value_New* according to the specific ranges of the customer values. Then, *ValueNewDisplayIndex* is used to specify the diameter and the color of the nodes while plotting the social network graph in Figure 4.1.

In reporting applications, after seeing the big picture drilling down into the details through filtering is a commonly used approach to achieve better focused results. To illustrate how this can be done in practice, the plot of the filtered graph which contain only the customers whose CLVs were greater than or equal to 10 million are given in Figure 4.2.

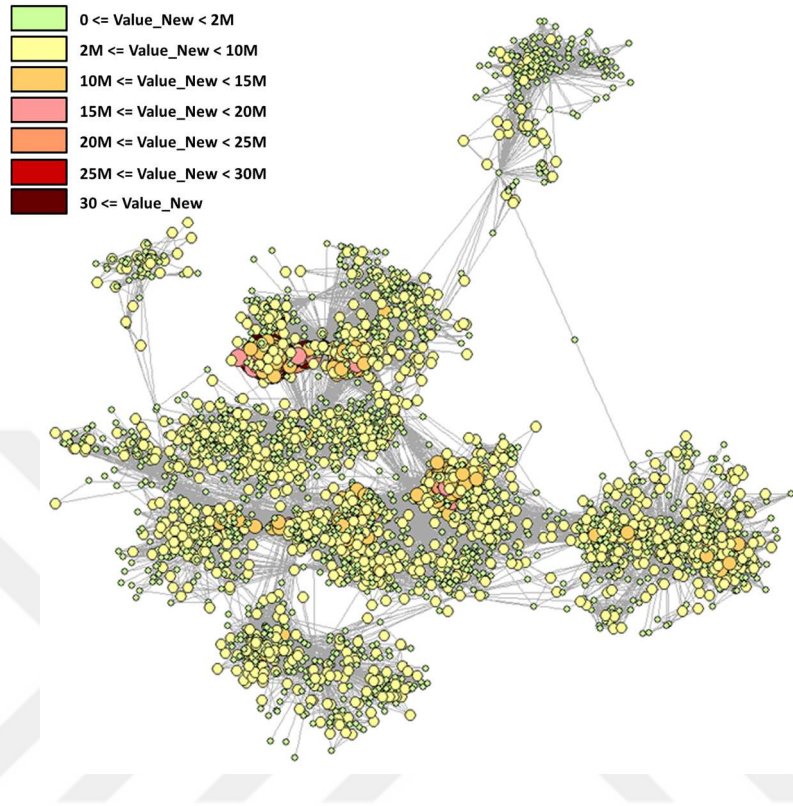


Figure 4.1: Customer lifetime value analysis plot of the whole network.

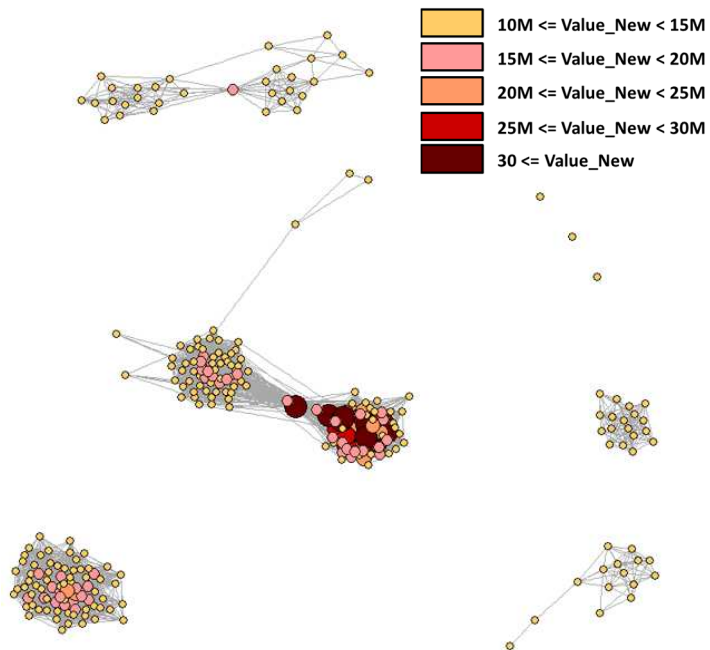


Figure 4.2: Customer lifetime value analysis plot of the filtered network.

Note in Figure 4.2 that, the customers are actually clustered among themselves according to their CLVs. For example, the clusters on the center right and bottom right are comprised of the people whose customer values are only between 10 and 15 million. The CLVs of all of the friends of these people in this network also contain no people from other clusters. Similarly, the cluster containing the most valuable customers takes place in the center of the plot. Although this cluster is not a completely homogeneous one, it is very striking that there is no customer outside of this cluster whose CLV is more than 25 million.

4.1.2. Membership Category Analysis

A classification can be made according to customers' membership categories of the Mile Program, such as *Basic*, *Intermediate* and *Advanced*. For this analysis, if the annual total distances flown by the people outside the Mile Program are within the range required by a specific mile category, it is assumed that those people belong to that category. Within this context, the membership category analysis plots in Figure 4.3 and Figure 4.4 are drawn. Both plots are colorized according to membership categories. The sizes of the nodes are adjusted with respect to CLVs. Figure 4.4 shows the membership categories of the customers whose CLVs are greater than 10 million.

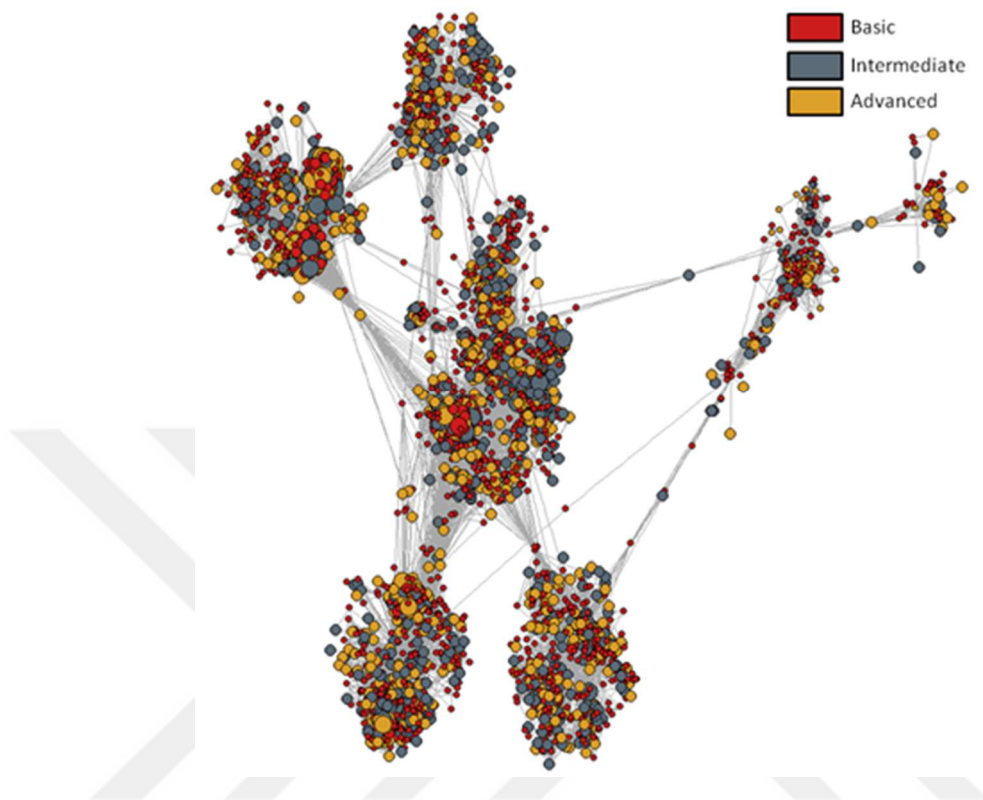


Figure 4.3: Membership category plot of the whole network.

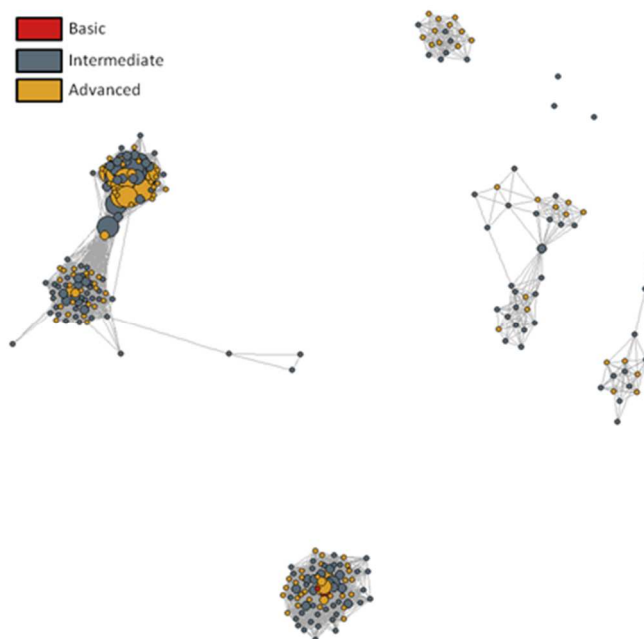


Figure 4.4: Membership category plot of the filtered network.

4.1.3. Profitability Analysis

Since the ticket prices in flight transactions are not available, in profitability analysis, the *Profitability* value which is obtained by dividing total customer value in year 2014 (*Value_12*) by total distance flown in 2014 (*Miles_12*) is used. The profitability plot for the network is shown in Figure 4.5 that is colorized with respect to three profitability levels. Figure 4.6 is the plot of the customers whose profitability values greater than 1.0.

Determining the customer profitability could be seen as merely a simple database query especially for airline companies which have the detailed knowledge about ticket fares and flight transactions. However, less profitable customers in this query may actually turn out to be more valuable when the analytics method proposed in this study is used for customer value estimation. To substantiate this claim, a method can be applied which combines the customer lifetime value and profitability analyses and then compares the results side by side.

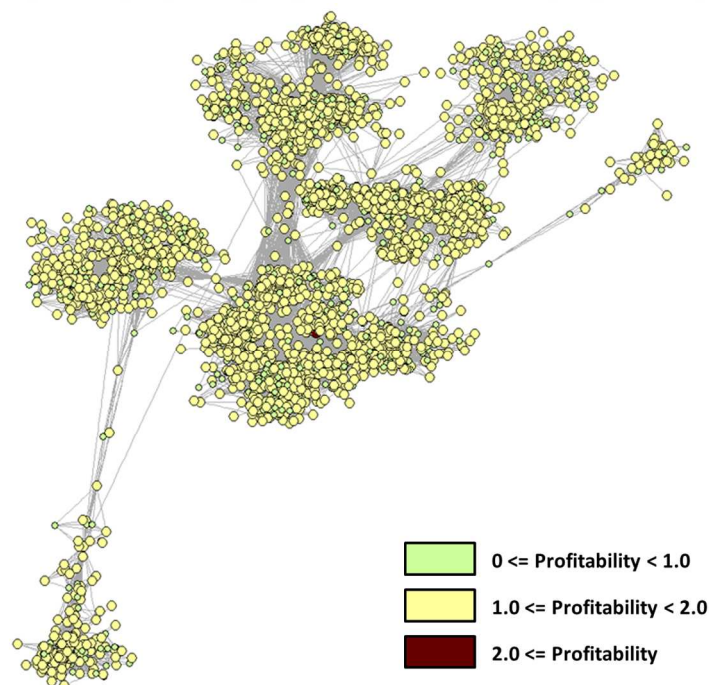


Figure 4.5: Profitability analysis plot of the whole network.

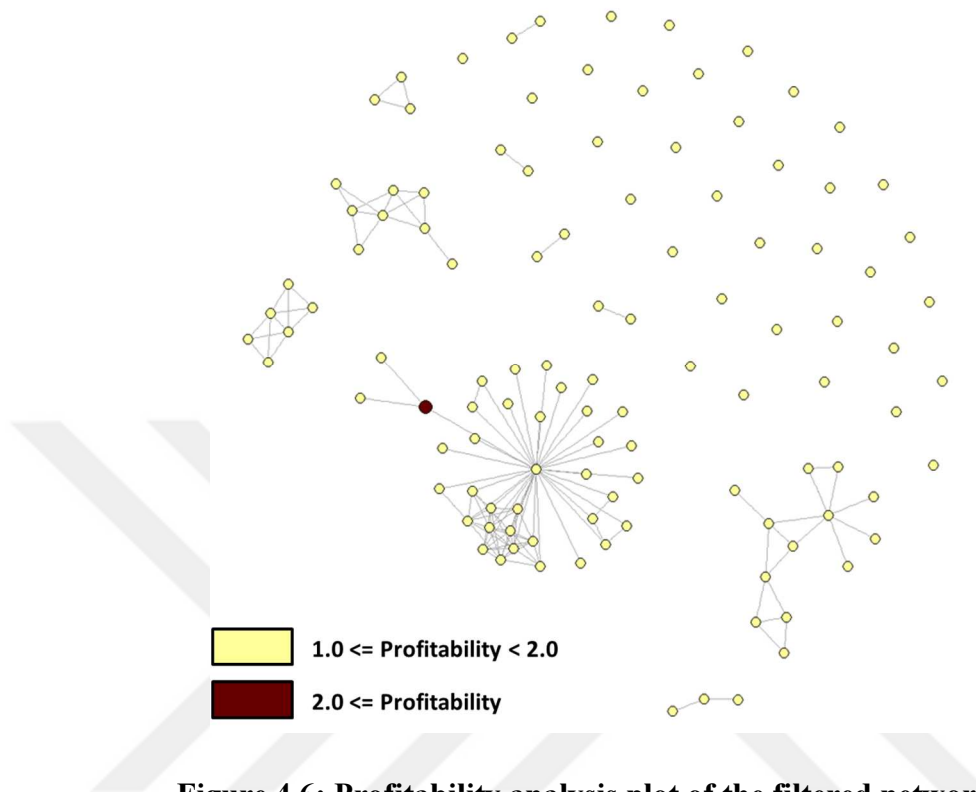


Figure 4.6: Profitability analysis plot of the filtered network.

4.1.4. Churn Analysis

Another important issue for airline companies is to retain customers. In this context, CRM departments of airline companies need to specify the customers who are likely to leave and take actions to avoid losing them. Here, it is provided that how this analysis could be done using monthly recent flights (*MonthlyRecency*) as seen in the churn analysis plot, in Figure 4.7. In this plot, the customers who have no flights during last 6 months are shown as larger red circles. In Figure 4.8, the customers who have no flights during last 6 months are shown together with their unique identifiers in the node circles.

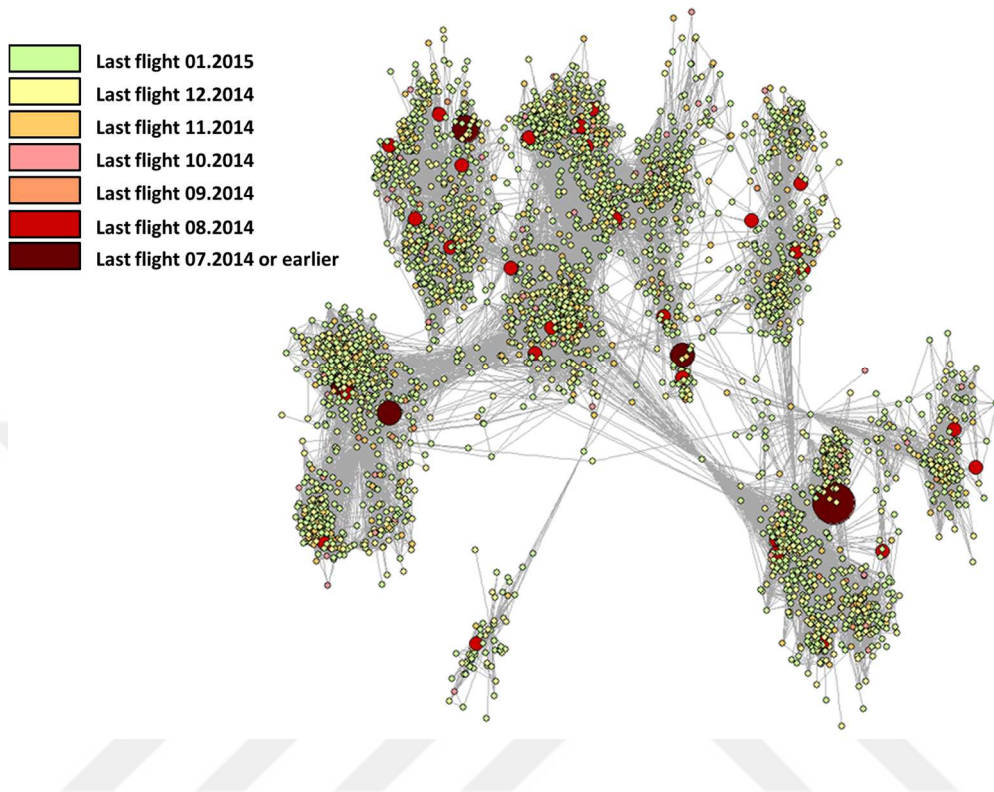


Figure 4.7: Churn analysis plot of the whole network.

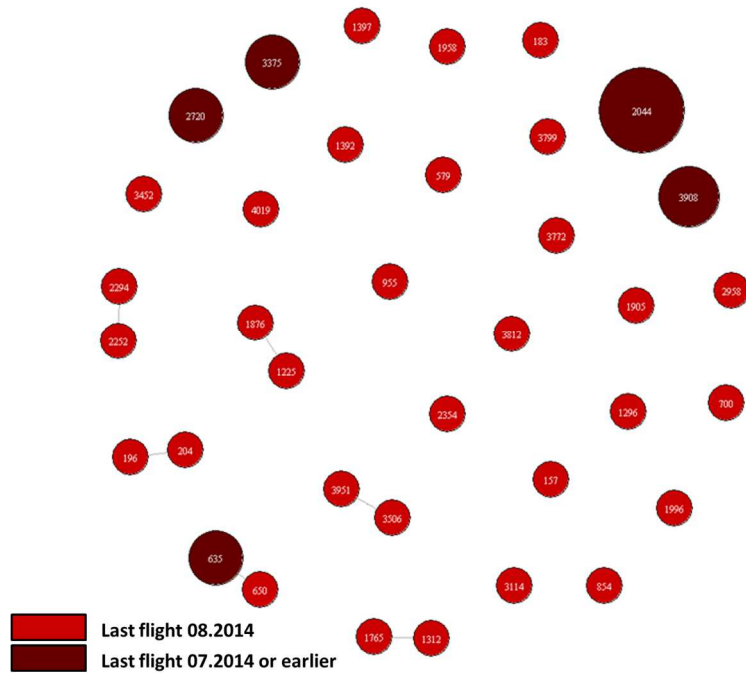


Figure 4.8: Churn analysis plot of the filtered network.

4.2. Time Series Forecasting Models

Social network information can potentially be integrated into the time series forecasting models. In order to achieve this goal, both domain related data and the social factors can be represented as time series data structure first. It is also important to be able to take the snapshots of the social network and gather the social factors concerning to that time. For forecasting part, another important topic is to model the network growth. The flowchart shown in Figure 4.9 includes the steps of the proposed method for airline customer value determination.

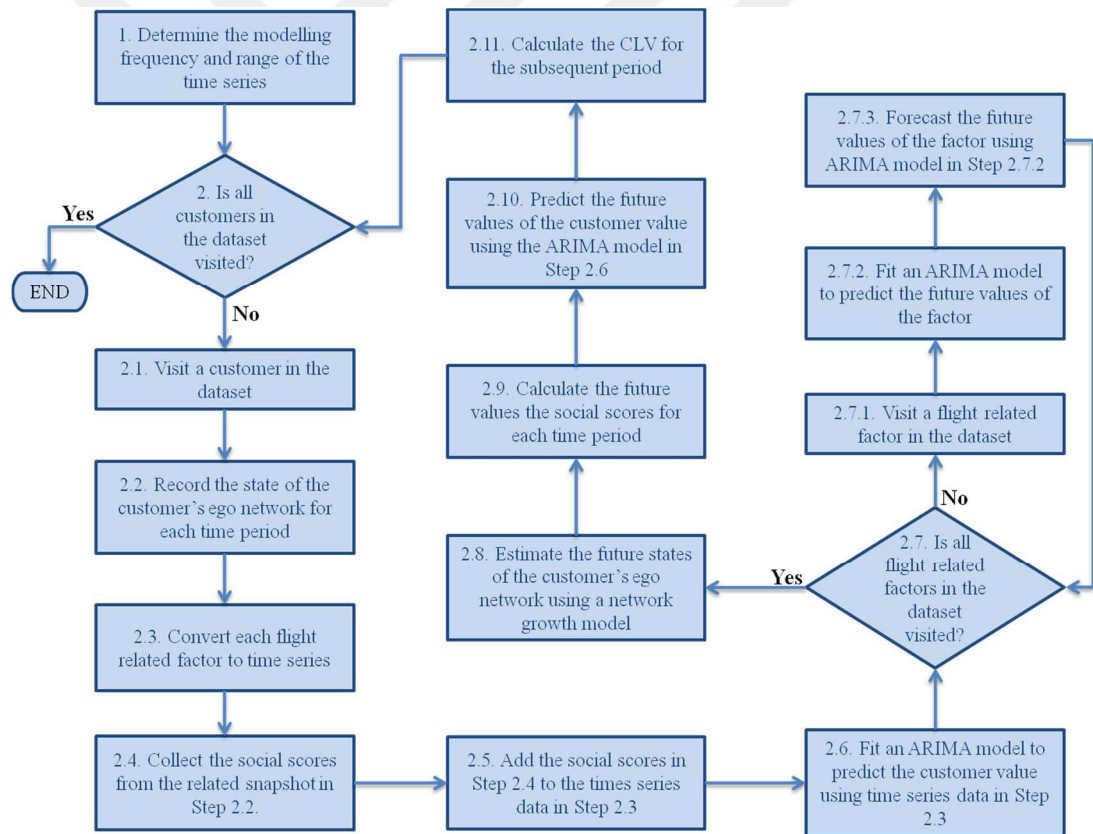


Figure 4.9: The flowchart of the proposed time series forecasting method.

CHAPTER FIVE

5. CONCLUSION

In this thesis, a methodology that combines traditional data analytics methods such as multiple regression with social network analysis methods is proposed for the airline industry. In particular, the customers' flight data and their social network data are combined as independent variables in a multiple regression formula for determining the customer value.

As mentioned in Section 3.3.6, with the addition of social network information airline customer lifetime value model has been improved in terms of accuracy and prediction ability. This method can also be adapted for other models from different domains.

Since both flight and social network data are private and commercial, the study has been conducted using limited sized, anonymous and sometimes synthetic data. Although, this is a major limitation, it should be noted that the main purpose of the study is to develop a methodology for integration of social factors into the data analytical models. The methods presented here can be used by the CRM departments of the airline companies on their real data repositories.

Actually the knowledge to be gained from social networks is not limited to only the static relationships between the nodes. One can add enriched information about the social circles of people, such as kinship, working in the same company, having graduated from the same school, residing in the same city, to the social network graph in the form of edge attributes. By adding this social knowledge to the model, it would be possible to reach a customer value determination model in which different social dimensions are taken into account. Since such models require knowing the actual relationship of the customers which cannot be possible in anonymous networks, this topic is left as a future study.

In the field of air transportation which has tough competition, it is critical to follow temporal variations of the airline customers' social states and to predict the reflection of these variations on the customers' airline preferences. In fact, this is not only a requirement of the air transportation sector but also all other sectors whose central focus are the customers such as those in services domain. In Section 4.2, a method is proposed that estimates the customer value by using time series forecasting. Time series forecasting models integrated with the social networks require historical snapshots of the network during the modeling time interval. Since the network used in this study is an anonymous one, such an opportunity is not available for this study. Because of this reason, integration of social network information into time series models is left as a future study.

As far as the types and diversity of the analyses are concerned, since the majority of data under study is synthetic, the illustrated analyses in Section 4 are relatively limited compared to the case where a large social network of the customers could be obtained. Although real flight data of the members of a mile program are used, such as their date of flight, departed and destination airport and total earned status miles, the variety and the size of the flight data were limited and the social network of these customers was not available. Therefore, a publicly available social network data are utilized and integrated with the data summarized above. If the size of real flight data were big enough to make more consistent statistical inferences, it would be possible to model the customer lifetime value better and to make more comprehensive analyses. For example, as a future work flight data can be enriched by adding factors related to real ticket price, flight class (such as, economy, business, first class, etc.) and passenger name record (PNR) number to the model. Similarly, social network factors can be expanded by using detailed semantic edge attributes such as being a member of same family, working in the same company, living in the same city, etc.

An article about the studies explained in this thesis submitted to the Journal of Air Transport Management is in the second round of revision. (Çavdar, Ferhatosmanoğlu, & Tulumoğlu, 2017).

REFERENCES

- Anderson, E. W., & Mittal, V. (2000). Strengthening The Satisfaction-Profit Chain. *Journal of Service Research*, 3(2), 107-120.
- Berger, P. D., & Nasr, N. I. (1998). Customer Lifetime Value: Marketing Models And Applications. *Journal Of Interactive Marketing*, 12(1), 17-30.
- Brin, S., & Page, L. (2012). Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 56(18), 3825-3833.
- Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and Methods in Social Network Analysis*. Cambridge University Press.
- Chen, W., Ma, C., & Ma, L. (2009). Mining the Customer Credit Using Hybrid Support Vector Machine Technique. *Expert Systems with Applications*, 36(4), 7611-7616.
- Çavdar, A. B., Ferhatosmanoğlu, N., & Tulumoğlu, Ş. (2017). Airline Customer Lifetime Value Estimation Using Data Analytics Supported by Social Network Information. *Journal of Air Transport Management (2nd round review)*.
- Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., & Nanavati, A. A. (2008). Social Ties and Their Relevance to Churn in Mobile Telecom Networks. *Proceedings of the 11th International Conference on Extending Database Technology* (pp. 668-677). ACM.
- Ekinci, Y., Ülengin, F., Uray, N., & Ülengin, B. (2014). Analysis Of Customer Lifetime Value And Marketing Expenditure Decisions Through A Markovian-Based Model. *European Journal Of Operational Research*, 237(1), 278-288.
- Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling Disease Outbreaks in Realistic Urban Social Networks. *Nature*, 429(6988), 180-184.

- Ferrentino, R., Cuomo, M. T., & Boniello, C. (2016). On the customer lifetime value: a mathematical perspective. *Computational Management Science*, 13(4), 521-539.
- Flight Calculator*. (n.d.). Retrieved 10 17, 2016, from travelmath Web Site: <http://www.travelmath.com/flights/>
- Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1(3), 215-239.
- Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 1360-1380.
- Gupta, S., & Lehmann, D. R. (2003). Customers As Assets. *Journal Of Interactive Marketing*, 17(1), 9-24.
- Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: Principles and Practice*. OTexts. Retrieved 10 20, 2016, from <https://www.otexts.org/book/fpp>
- Jain, D., & Singh, S. S. (2002). Customer Lifetime Value Research In Marketing: A Review And Future Directions. *Journal Of Interactive Marketing*, 16(2), 34-46.
- Jiang, T., & Tuzhilin, A. (2006). Segmenting Customers from Population to Individuals: Does 1-to-1 Keep Your Customers Forever? *IEEE Transactions on Knowledge and Data Engineering*, 18(10), 1297-1311.
- Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5), 604-632.
- Kleinberg, J. M. (2007). Challenges in Mining Social Network Data: Processes, Privacy, and Paradoxes. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 4-5). ACM.
- Kotler, P., & Armstrong, G. (1996). *Principles Of Marketing* (7th ed.). Englewood Cliffs: Prentice Hall.
- Krebs, V. E. (2002). Mapping Networks of Terrorist Cells. *Connections*, 24(3), 43-52.

- Lariviere, B., & Van den Poel, D. (2005). Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques. *Expert Systems with Applications*, 29(2), 472-484.
- Lescovec, J. (n.d.). *SNAP: Network Datasets: Social Circles*. Retrieved 10 18, 2016, from SNAP: Stanford Network Analysis Project: <https://snap.stanford.edu/data/egonets-Facebook.html>
- Lescovec, J., & Horvitz, E. (2008). Planetary-Scale Views on a Large Instant-Messaging Network. *Proceedings of the 17th International Conference on World Wide Web* (pp. 915-924). ACM.
- Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting Methods and Applications*. John Wiley & Sons, Inc.
- Milgram, S. (1967). The Small-World Problem. *Psychology Today*, 2(1), 60-67.
- Niraj, R., Gupta, M., & Narasimhan, C. (2001). Customer Profitability In A Supply Chain. *Journal Of Marketing*, 65(3), 1-16.
- Otte, E., & Rousseau, R. (2002). Social Network Analysis: A Powerful Strategy, also for the Information Sciences. *Journal of Information Science*, 28(6), 441-453.
- Reinartz, W., & Kumar, V. (2002). The Mismanagement Of Customer Loyalty. *Harvard Business Review*, 80(7), 86-95.
- RStudio, Inc. (n.d.). *RStudio - Open source and enterprise-ready professional software for R*. Retrieved 10 17, 2016, from RStudio Web Site: <https://www.rstudio.com/>
- Smith, S. (n.d.). *Determining Sample Size: How to Ensure You Get the Correct Sample Size*. Retrieved 01 2017, from Qualtrics: <https://www.qualtrics.com/blog/determining-sample-size/>
- Spreadsheet Software Programs | Excel Free Trial*. (n.d.). Retrieved 10 17, 2016, from Microsoft Excel Web Site: <https://products.office.com/en/excel>
- Stack Exchange. (n.d.). Retrieved 01 2017, from Cross Validated: <http://stats.stackexchange.com/questions/59250/how-to-interpret-the-output-of-the-summary-method-for-an-lm-object-in-r>

- Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social Influence Analysis in Large-Scale Networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 807-816). ACM.
- Taylor, A. D. (2016). Retrieved 10 20, 2016, from Andrew D. Taylor, Department of Zoology, University of Hawaii at Manoa: <http://www2.hawaii.edu/~taylor/z632/Rbestsubsets.pdf>
- The R Foundation. (n.d.). *R: The R Project for Statistical Computing*. Retrieved 10 17, 2016, from R Project Official Web Site: <https://www.r-project.org/>
- Tirenni, G., Kaiser, C., & Herrmann, A. (2007). Applying Decision Trees for Value-Based Customer Relationship Management: Predicting Airline Customers' Future Values. *Journal of Database Marketing & Customer Strategy Management*, 14(2), 130-142.
- Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of Word-of-mouth Versus Traditional Marketing: Findings From An Internet Social Networking Site. *Journal of Marketing*, 73(5), 90-102.
- Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (Vol. 8). Cambridge University Press.
- Watts, D. J. (2004). *Six degrees: The Science of a Connected Age*. WW Norton & Company.
- Watts, D. J., & Strogatz, S. H. (1998). Collective Dynamics of 'small-world' Networks. *Nature*, 393(6684), 440-442.
- Woo, K.-s., & Fock, H. K. (2004). Retaining And Divesting Customers: An Exploratory Study Of Right Customers, "At-Risk" Right Customers, And Wrong Customers. *Journal of Services Marketing*, 18(3), 187-197.

APPENDICES

Appendix A - Descriptive Statistics of the Flight Data

Appendix B - 500 Most Valuable Customers Rankings of the Models

Appendix C - R Source Code



Appendix A - Descriptive Statistics of the Flight Data

Value	value_3	value_6	value_12
Min. : 0.0	Min. : 0	Min. : 0	Min. : 250
1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 4258	1st Qu.: 4718
Median : 0.0	Median : 5802	Median : 9986	Median : 10664
Mean : 905.3	Mean : 6494	Mean : 11340	Mean : 15212
3rd Qu.: 1628.0	3rd Qu.: 7790	3rd Qu.: 14184	3rd Qu.: 18200
Max. : 3329.0	Max. : 36127	Max. : 52427	Max. : 61289

Freq	Freq_3	Freq_6	Freq_12
Min. : 0.0	Min. : 0.0	Min. : 0.00	Min. : 1.0
1st Qu.: 0.0	1st Qu.: 0.0	1st Qu.: 3.00	1st Qu.: 5.0
Median : 0.0	Median : 6.0	Median : 12.00	Median : 15.0
Mean : 1.2	Mean : 6.0	Mean : 10.73	Mean : 15.0
3rd Qu.: 2.0	3rd Qu.: 9.5	3rd Qu.: 15.50	3rd Qu.: 22.5
Max. : 4.0	Max. : 18.0	Max. : 25.00	Max. : 35.0

Av. Tran. Size	Av. Tran. Size_3	Av. Tran. Size_6	Av. Tran. Size_12
Min. : 0.0	Min. : 0.0	Min. : 0.0	Min. : 250.0
1st Qu.: 0.0	1st Qu.: 0.0	1st Qu.: 721.1	1st Qu.: 646.0
Median : 0.0	Median : 710.1	Median : 844.2	Median : 765.0
Mean : 297.9	Mean : 642.7	Mean : 999.2	Mean : 995.8
3rd Qu.: 660.5	3rd Qu.: 915.4	3rd Qu.: 926.3	3rd Qu.: 893.9
Max. : 1028.0	Max. : 2007.1	Max. : 3526.5	Max. : 3526.5

Miles	Miles_3	Miles_6	Miles_12
Min. : 0.0	Min. : 0	Min. : 0	Min. : 216
1st Qu.: 0.0	1st Qu.: 0	1st Qu.: 3079	1st Qu.: 4286
Median : 0.0	Median : 4948	Median : 7431	Median : 9204
Mean : 758.7	Mean : 6334	Mean : 10758	Mean : 14212
3rd Qu.: 1306.5	3rd Qu.: 8332	3rd Qu.: 14418	3rd Qu.: 17933
Max. : 3184.0	Max. : 33038	Max. : 47290	Max. : 55780

Longevity	Recency	Age	Value_New
Min. : 470	Min. : 20.0	Min. : 26.00	Min. : 0
1st Qu.: 848	1st Qu.: 68.5	1st Qu.: 30.50	1st Qu.: 1200
Median : 1623	Median : 98.0	Median : 33.00	Median : 3600
Mean : 1751	Mean : 134.5	Mean : 33.27	Mean : 6505
3rd Qu.: 2536	3rd Qu.: 170.0	3rd Qu.: 36.00	3rd Qu.: 7614
Max. : 3359	Max. : 378.0	Max. : 42.00	Max. : 25850

Appendix B - 500 Most Valuable Customers Rankings of the Models

No	Base ID	Base Value	Prop. ID	Prop. Value
1	920	998,03	656	999,19
2	694	996,56	2554	998,57
3	1284	995,93	68	998,32
4	678	995,24	1560	996,95
5	207	995,06	1685	996,52
6	2152	993,47	3398	996,18
7	235	992,65	2994	996,13
8	2881	992,13	3030	995,99
9	345	991,52	112	995,91
10	2667	991,23	629	995,62
11	1408	990,80	641	995,10
12	3889	990,79	2301	995,03
13	2782	990,73	3539	994,62
14	1456	989,79	2391	994,01
15	894	989,75	3932	993,98
16	1849	988,83	2171	991,89
17	784	988,64	478	991,84
18	1047	986,75	2805	991,77
19	1379	986,07	4017	991,57
20	3105	985,84	3419	988,91
21	3874	985,74	1865	988,31
22	3004	984,49	2653	988,27
23	1275	984,33	1730	988,09
24	3272	980,02	1701	987,93
25	1072	979,49	880	987,82
26	3150	979,03	2541	987,50
27	2669	978,84	3843	987,48
28	3703	975,60	3002	987,45
29	340	973,73	2904	986,41
30	2153	972,75	587	985,81
31	2026	971,73	3970	985,57
32	2474	971,70	1461	985,40
33	1768	971,61	2182	985,18
34	2630	970,76	91	982,68
35	2438	970,33	1087	982,54

No	Base ID	Base Value	Prop. ID	Prop. Value
36	1194	970,00	4019	982,29
37	2186	968,68	871	982,18
38	1391	968,64	3513	981,82
39	2530	967,73	2729	981,21
40	3467	967,14	3185	981,14
41	2083	965,34	431	980,24
42	1062	964,69	1273	979,87
43	1271	963,76	2499	979,29
44	1567	963,61	3108	978,98
45	2972	963,31	3270	978,93
46	2195	961,06	3864	978,87
47	2347	958,27	2763	978,55
48	3832	956,75	2824	978,45
49	1494	954,05	3716	978,07
50	1189	953,91	1718	977,80
51	1786	953,87	673	976,78
52	3403	953,31	3771	976,60
53	2683	951,98	8	975,98
54	146	951,46	3019	974,80
55	1179	950,99	593	974,63
56	3758	950,75	2767	973,25
57	3583	950,40	3953	973,16
58	2457	947,91	1181	973,03
59	2066	947,05	202	972,48
60	2637	946,46	2630	970,76
61	237	946,33	3908	969,66
62	170	944,97	1151	969,62
63	1898	944,10	515	967,94
64	1263	943,71	4003	966,86
65	2469	942,38	2869	966,46
66	2129	941,42	1523	965,92
67	1485	941,04	3106	964,18
68	1383	940,63	349	962,80
69	2841	939,55	2441	962,17
70	2665	939,51	1326	961,35

No	Base ID	Base Value	Prop. ID	Prop. Value
71	688	939,31	1080	961,26
72	553	938,86	1346	961,11
73	2140	935,75	25	960,01
74	4015	935,64	877	959,12
75	1025	935,31	2347	958,27
76	3092	934,59	3952	958,17
77	79	933,27	2679	957,72
78	2128	932,08	1571	956,84
79	3449	930,89	2780	956,09
80	3033	930,86	3744	955,80
81	2115	930,08	667	955,03
82	2486	928,66	3033	954,04
83	2690	926,72	859	952,48
84	2317	923,91	3879	952,05
85	1184	922,76	1130	952,01
86	2107	922,51	2857	951,89
87	3610	922,39	3451	950,58
88	523	920,42	2781	950,49
89	1623	920,25	3974	950,11
90	3623	919,65	1270	949,84
91	89	918,84	215	949,77
92	2933	918,69	2975	949,35
93	3579	918,52	2392	949,19
94	3639	915,34	1782	948,92
95	2267	914,70	1941	948,83
96	2512	912,93	4028	947,38
97	2199	912,70	624	947,01
98	3972	912,51	4024	945,81
99	3140	912,35	3709	944,39
100	3218	911,59	2773	944,04
101	2651	909,75	957	939,89
102	1274	909,60	719	938,73
103	921	909,16	942	937,71
104	347	909,14	3345	937,49
105	2836	908,96	3570	937,00
106	939	907,73	2140	935,75
107	1735	906,23	3330	935,61
108	1767	906,23	1119	934,67

No	Base ID	Base Value	Prop. ID	Prop. Value
109	1583	906,07	2594	934,35
110	713	905,28	3684	933,93
111	2610	904,80	1285	932,37
112	1621	904,59	668	932,28
113	1198	902,91	2945	926,29
114	3345	900,29	1995	925,72
115	997	900,12	2018	924,32
116	13	896,83	2931	921,73
117	2170	896,03	3838	916,72
118	1685	896,01	1772	915,64
119	2263	895,60	1199	914,19
120	2811	895,55	513	912,50
121	2770	895,33	122	911,44
122	1923	893,85	1021	909,44
123	1153	892,80	2006	908,23
124	2523	889,21	3087	907,98
125	3207	888,97	436	907,07
126	3930	888,18	2985	906,57
127	1495	888,07	1284	904,61
128	1100	886,75	2465	904,41
129	1245	886,25	404	901,09
130	492	886,19	3201	898,20
131	134	885,32	1513	898,07
132	1795	884,51	2258	895,58
133	987	884,33	1819	894,22
134	865	884,05	3793	890,83
135	1922	882,94	171	888,61
136	3282	882,29	2075	886,99
137	1535	881,67	188	883,43
138	1299	881,66	2198	879,92
139	3949	881,16	2890	879,29
140	1664	880,86	2765	878,82
141	3094	875,17	981	875,68
142	1518	875,10	2365	875,41
143	1337	874,44	1432	872,37
144	3003	873,00	2927	871,30
145	1000	871,89	3966	871,03
146	2719	871,51	1859	870,68

No	Base ID	Base Value	Prop. ID	Prop. Value
147	3070	871,05	3077	869,99
148	900	870,23	1881	868,21
149	3797	868,86	3365	866,43
150	1178	868,84	1465	866,02
151	2912	868,69	1054	865,12
152	1712	868,00	2749	864,63
153	967	867,94	373	864,62
154	932	867,88	2881	863,23
155	1844	867,79	3391	863,22
156	1004	866,77	1998	859,54
157	3255	864,85	1584	859,25
158	2981	863,39	2009	858,54
159	3846	863,38	2031	856,27
160	1060	862,59	272	854,12
161	3293	861,23	2743	851,10
162	3145	860,32	1751	847,38
163	1081	857,14	2191	844,80
164	1798	856,71	2145	844,75
165	2433	856,58	3327	843,93
166	2159	856,19	1120	843,67
167	2617	853,08	1851	842,32
168	3551	850,91	1374	841,79
169	3224	849,65	1872	841,77
170	214	849,52	1807	840,01
171	2297	848,71	795	839,65
172	42	847,30	2429	838,24
173	1810	846,34	3731	835,03
174	590	843,46	1720	834,66
175	2496	843,46	1991	833,07
176	3503	842,97	2147	832,43
177	3518	842,48	1827	831,95
178	1504	840,27	1068	831,56
179	795	839,65	355	830,62
180	273	837,33	1293	829,74
181	1224	835,99	1843	828,65
182	22	832,91	1673	827,81
183	3386	831,66	3501	825,45
184	2745	831,31	2516	825,24

No	Base ID	Base Value	Prop. ID	Prop. Value
185	864	830,77	1931	822,31
186	3000	829,06	752	822,09
187	895	828,06	1485	821,90
188	1323	827,59	199	821,56
189	774	825,72	3873	820,66
190	3114	823,35	3448	820,48
191	1642	823,17	820	819,75
192	1427	822,13	3672	819,61
193	1367	821,89	1355	817,86
194	2393	821,69	1911	817,82
195	221	820,68	397	817,62
196	3332	820,61	3072	817,08
197	3667	817,72	3629	816,71
198	2779	816,57	3295	814,89
199	3909	814,94	555	814,59
200	2967	811,11	3000	814,34
201	3919	810,28	703	810,29
202	3588	809,78	455	808,21
203	1585	805,50	1392	807,50
204	1450	804,81	1746	805,26
205	140	804,80	2751	804,73
206	2799	803,65	1626	804,40
207	441	803,40	3020	804,03
208	2546	801,14	1126	802,88
209	145	800,13	332	802,80
210	2959	799,61	773	800,97
211	1827	798,62	971	799,04
212	3395	796,44	3022	798,54
213	3768	794,99	3023	798,29
214	491	794,83	2209	797,12
215	969	793,73	1635	794,01
216	3065	793,00	1226	792,17
217	1529	791,63	3369	791,42
218	2533	790,64	1350	789,85
219	1228	789,69	3902	789,56
220	2822	788,23	428	788,85
221	101	787,97	3640	788,74
222	4012	787,59	1318	787,17

No	Base ID	Base Value	Prop. ID	Prop. Value
223	1933	787,15	1508	786,82
224	1896	785,60	1963	786,11
225	805	784,00	342	785,95
226	3486	781,78	2770	785,58
227	1983	781,28	1183	785,43
228	3871	779,98	1983	781,28
229	3253	779,34	3947	780,95
230	1269	775,05	3841	780,62
231	803	775,00	547	778,76
232	2259	773,81	696	777,52
233	2051	773,80	1774	775,49
234	1010	773,51	3079	773,25
235	49	772,46	323	772,66
236	1108	772,25	1109	770,89
237	2133	768,20	3124	770,84
238	3631	768,08	1957	770,17
239	1771	766,10	1451	770,11
240	3073	764,11	1001	768,92
241	819	763,65	1322	768,86
242	2142	762,48	1795	765,93
243	2785	762,09	1300	764,44
244	2640	760,44	3921	763,56
245	2322	758,94	760	762,80
246	1496	758,76	2142	762,48
247	163	758,74	1348	760,18
248	113	758,69	2434	759,53
249	3132	755,72	3219	759,53
250	2771	755,14	2844	759,34
251	3008	754,31	1167	759,26
252	2013	753,97	3692	758,60
253	3479	752,58	1216	758,20
254	1498	752,12	1548	757,94
255	3200	750,41	1682	757,36
256	728	748,92	1170	757,10
257	477	747,98	3198	757,10
258	1970	746,33	313	756,63
259	2560	744,73	1975	755,65
260	1074	744,34	1044	754,02

No	Base ID	Base Value	Prop. ID	Prop. Value
261	1477	743,36	2673	753,47
262	2499	740,55	1517	751,39
263	2626	740,42	1440	751,16
264	2883	740,10	169	750,71
265	2432	738,14	2214	749,06
266	1502	737,55	2231	747,89
267	2590	736,15	2761	747,54
268	3462	736,01	726	747,49
269	2819	735,69	3225	747,01
270	3574	735,63	1106	746,95
271	2143	735,00	915	746,31
272	3095	733,17	109	745,68
273	2206	729,64	3176	745,32
274	3927	729,35	2406	744,95
275	3038	724,32	2560	744,73
276	3081	723,73	2845	743,59
277	361	722,63	2566	743,32
278	367	719,80	3529	742,91
279	1363	719,13	369	742,85
280	2436	718,88	387	742,12
281	1584	717,76	1803	740,33
282	3578	717,63	3129	739,50
283	2487	717,35	1833	738,56
284	1754	712,14	977	737,50
285	325	707,19	359	737,11
286	3247	706,74	2590	736,15
287	2996	706,71	2274	734,32
288	2976	703,98	82	733,91
289	3071	703,40	1648	733,64
290	3597	700,25	1306	732,98
291	1093	699,77	1408	732,79
292	3682	699,34	3617	732,54
293	162	698,65	2764	732,22
294	872	697,85	2067	732,12
295	1882	697,43	3123	732,06
296	4034	696,76	963	731,45
297	1892	692,83	2206	729,64
298	2045	692,57	2141	729,52

No	Base ID	Base Value	Prop. ID	Prop. Value
299	1364	692,24	312	728,70
300	2513	691,85	3540	728,45
301	357	690,65	3417	727,82
302	655	690,34	2501	727,79
303	473	689,16	2254	727,38
304	1848	687,69	1836	727,07
305	3433	685,84	1671	726,81
306	3187	685,38	1559	726,75
307	2085	684,25	1458	725,18
308	3683	683,81	2978	725,09
309	2417	683,06	3782	722,77
310	352	682,27	3255	722,05
311	3564	681,67	1606	721,51
312	1260	681,18	1769	720,80
313	1322	679,30	1095	720,57
314	2548	679,12	1131	719,60
315	651	678,12	2050	717,83
316	3331	676,01	2208	717,71
317	1690	673,73	2713	717,70
318	326	673,72	1928	717,66
319	1558	673,10	3497	717,08
320	3211	672,94	792	715,65
321	1579	671,76	992	715,58
322	875	670,49	1406	714,90
323	3306	669,59	3377	714,81
324	385	667,39	1526	714,55
325	873	666,69	3264	714,02
326	1777	666,39	92	713,81
327	2314	664,33	1484	712,05
328	2329	658,39	3203	711,66
329	949	657,56	1805	711,61
330	2515	656,04	3415	710,54
331	1229	655,73	1412	710,41
332	626	654,93	2531	710,20
333	1944	653,93	1779	710,07
334	1658	653,28	1747	709,71
335	3562	652,02	3918	708,88
336	1030	651,02	2111	708,37

No	Base ID	Base Value	Prop. ID	Prop. Value
337	1018	650,45	2558	708,28
338	3861	649,83	1621	707,15
339	2073	648,94	3018	705,06
340	2179	647,83	2687	704,92
341	1574	646,88	645	703,41
342	3458	646,61	2036	703,27
343	3748	646,36	3711	703,05
344	1429	644,86	3276	701,51
345	1597	644,04	1497	701,12
346	531	643,55	1187	701,02
347	259	643,12	148	699,12
348	3417	642,29	3728	699,12
349	1035	636,69	606	698,80
350	1333	635,59	3786	698,27
351	625	634,50	3764	698,09
352	1622	633,47	1707	697,76
353	2111	633,01	3469	697,55
354	3005	631,74	4023	695,62
355	1376	631,62	3132	695,30
356	2755	631,10	2962	694,24
357	638	631,08	426	694,19
358	1785	629,38	2701	693,04
359	2134	626,99	1698	692,90
360	863	626,73	2813	692,71
361	1743	626,68	2045	692,57
362	3111	624,95	510	691,90
363	488	624,65	2696	691,62
364	1920	621,68	2587	691,48
365	489	621,37	3590	690,58
366	2658	620,66	1376	689,60
367	993	615,29	2065	688,55
368	1430	613,62	1381	688,52
369	607	613,19	3635	686,53
370	3257	612,75	3273	686,24
371	1976	611,76	2248	686,06
372	2039	611,45	223	685,95
373	3044	610,63	2917	684,21
374	2191	609,42	1078	684,02

No	Base ID	Base Value	Prop. ID	Prop. Value
375	1885	607,43	2663	683,50
376	2234	606,23	1212	683,03
377	3535	605,54	1556	682,68
378	1593	603,96	1781	682,44
379	1910	603,70	2979	682,08
380	1862	602,03	3762	681,75
381	647	601,74	1494	681,42
382	1668	601,08	617	681,29
383	3992	599,87	2341	681,06
384	550	599,87	424	678,93
385	3630	598,85	3173	678,78
386	3622	598,20	772	677,61
387	1262	595,18	3283	676,64
388	2099	594,63	3068	676,46
389	610	591,01	2682	676,45
390	572	589,71	3035	675,77
391	2306	586,77	3478	675,00
392	1321	583,31	486	675,00
393	1172	581,92	238	673,93
394	3722	581,49	3157	673,90
395	1654	581,16	2614	673,82
396	848	578,42	1700	673,60
397	786	577,98	3440	673,22
398	3062	576,86	1421	673,02
399	423	576,75	1121	672,89
400	2736	576,61	3050	672,65
401	519	575,79	3348	672,26
402	3340	575,62	3891	671,77
403	1956	575,11	116	670,98
404	280	573,88	1835	670,05
405	922	573,79	2454	668,93
406	1967	573,75	3361	667,44
407	243	573,45	85	666,98
408	393	573,19	3210	666,69
409	3495	572,69	1758	666,68
410	3161	568,19	599	666,55
411	821	566,91	1710	666,36
412	3364	566,85	251	665,55

No	Base ID	Base Value	Prop. ID	Prop. Value
413	2639	563,73	3058	665,49
414	3727	557,99	3726	664,86
415	3010	555,78	1143	662,16
416	983	552,87	1391	661,87
417	1695	551,05	1870	661,58
418	2835	546,36	3459	661,54
419	2725	546,11	84	661,19
420	1986	545,94	2727	660,89
421	3432	544,73	5	660,16
422	319	542,70	3675	659,72
423	2048	541,55	3013	659,70
424	1168	540,84	3527	659,57
425	121	540,35	3150	658,72
426	2366	540,14	3778	658,02
427	496	539,96	807	655,40
428	2357	535,74	2468	654,44
429	2224	534,96	416	651,33
430	1417	534,11	196	650,92
431	458	532,32	2709	650,89
432	2597	530,43	1045	650,54
433	802	528,22	2205	650,36
434	420	527,12	3454	649,91
435	1587	525,53	1401	649,71
436	3881	525,02	961	649,53
437	2924	524,78	2073	648,94
438	1165	524,31	3736	648,89
439	2563	524,13	2935	648,34
440	779	523,80	3510	647,88
441	45	523,50	2260	647,11
442	908	519,32	543	646,99
443	3956	518,27	1622	646,65
444	3845	517,70	3138	646,19
445	149	517,36	532	645,72
446	2636	517,08	3254	645,13
447	2758	516,52	2019	644,82
448	711	516,28	2635	644,71
449	3525	515,59	3788	644,24
450	21	515,44	36	643,51

No	Base ID	Base Value	Prop. ID	Prop. Value
451	809	513,96	318	643,10
452	2886	513,93	2889	641,75
453	3710	512,42	3356	640,82
454	3807	512,10	2439	640,60
455	1612	508,15	3594	640,12
456	813	507,24	168	640,07
457	3387	506,31	1317	639,26
458	1879	505,55	2134	638,70
459	3652	504,63	2789	638,52
460	1713	502,05	4000	638,43
461	118	501,96	3309	638,34
462	3739	499,80	565	637,76
463	2254	498,90	2452	637,19
464	1357	497,54	2	637,05
465	1546	495,63	1272	636,70
466	3402	495,50	1840	635,29
467	3434	495,12	3092	634,68
468	1980	494,22	3580	634,24
469	690	493,23	3515	634,04
470	60	492,94	2804	634,03
471	3125	492,13	222	633,05
472	1410	491,15	3308	632,98
473	2123	489,38	4009	632,64
474	3907	488,17	3712	631,11
475	858	486,35	2695	630,98

No	Base ID	Base Value	Prop. ID	Prop. Value
476	2247	485,98	2834	630,92
477	2865	482,58	1218	630,17
478	139	481,28	706	630,14
479	2244	481,25	10	629,30
480	2501	480,72	666	629,24
481	3353	480,36	257	628,94
482	2240	478,33	375	628,74
483	3512	475,02	3563	628,66
484	3886	472,95	466	628,57
485	2418	471,76	2605	628,39
486	1850	470,97	225	628,32
487	782	467,34	2152	628,25
488	2562	466,07	2186	628,07
489	3511	464,93	3122	627,12
490	3382	463,08	3572	626,89
491	453	462,74	3015	625,81
492	292	462,10	2964	625,46
493	2772	460,43	1506	624,04
494	2037	459,90	3194	623,72
495	3806	458,73	3492	623,17
496	3429	456,23	1977	622,20
497	2239	453,77	882	621,71
498	293	451,91	2199	620,31
499	3725	449,72	3222	620,08
500	3231	449,50	2175	619,71

Appendix C - R Source Code

```
library(leaps)
library(car)
library(MASS)
library(msm)
library(igraph)
library(forecast)
library(stats)

#===== START OF FUNCTIONS =====
addValueNewDisplayIndex = function(combined_data) {
  index_col = rep(0, length(combined_data[,1]))
  combined_data[["ValueNewDisplayIndex"]] = index_col
  c = 0

  for (i in 1:length(category_data[,1])) {
    vIndex = combined_data[i, "Value_New"] / 1000000

    if (vIndex < 2.0) {
      c = 1
    } else if (vIndex < 10.0) {
      c = 2
    } else if (vIndex < 15.0) {
      c = 3
    } else if (vIndex < 20.0) {
      c = 4
    } else if (vIndex < 25.0) {
      c = 5
    } else if (vIndex < 30.0) {
      c = 6
    } else {
      c = 7
    }

    combined_data[i, "ValueNewDisplayIndex"] = c
  }

  return (combined_data)
}

calculateMonthlyRecency = function(mf) {
  mr = NULL
  mr[["Id"]] = mf[, "Id"]
  mr = as.data.frame(mr)
  mr[["Recency"]] = rep(1, length(mf[,1]))
  colnames(mr) = c("Id", "Recency")

  for (i in 1:length(mf[,1])) {
    recency = 1
    for (j in length(mf[1,-15]):2) {
      if (mf[i, j] == 0) {
        recency = recency + 1
      } else {
        break
      }
    }
    mr[i, "Recency"] = recency
  }
}
```

```

    }

    return (mr)
}

addCategory = function(combined_data, category_data) {
  category_col = rep(0, length(combined_data[,1]))
  combined_data[["Category"]] = category_col

  for (i in 1:length(category_data[,1])) {
    combined_data[which(combined_data$SN_Id==category_data[i, "Id"]), "Category"] =
category_data[i, "Category"]
  }

  return (combined_data)
}

addProfitability = function(flight_data) {
  profitability_col = rep(1.0, length(flight_data[,1]))
  flight_data[["Profitability"]] = profitability_col

  for (i in 1:length(flight_data[,1])) {
    flight_data[i, "Profitability"] = flight_data[i, "Value_12"] / flight_data[i,
"Miles_12"]
  }

  return (flight_data)
}

combineSNandFlightData = function(flights, sn_scores)
{
  combined = NULL

  for (i in 1:length(flights[,1])) {
    sn_row = sn_scores[which(sn_scores$SN_Id==flights[i, 1]),]
    a_row = cbind(sn_row, flights[i,-1])

    colnames(a_row) = c(colnames(sn_row), colnames(flights[,,-1]))
    if (is.null(combined)) {
      combined = a_row
    } else {
      combined = rbind(combined, a_row)
    }
  }

  return (combined)
}

# Update Value_New wrt social scores
updateValueNew = function(people_value_new, sn_scores)
{
  new_values = people_value_new

  for (i in 1:length(people_value_new[,1])) {
    score = sn_scores[which(sn_scores$SN_Id==people_value_new[i, 1]), 2]
    new_values[i, 2] = (score * 1000 * people_value_new[i, 2]) + people_value_new[i,
2]
  }
}

```

```

    return (new_values)
  }

# Assign flight data
assignFlightData = function(alters, friendship_vector, all_categories, homophily,
fitted_model, col_names,
                           monthly_freq_avg, monthly_miles_avg, monthly_miles_sd,
monthly_price_avg,
                           monthly_price_sd, cat_val_new)
{
  # egos_with_cat = subset(friendship_vector, !duplicated(ego_id))
  # egos_with_cat = egos_with_cat[, -1]

  f_all = NULL
  m_all = NULL
  p_all = NULL

  for (i in 1:length(alters)) {
    if (homophily == 0.0) {
      # Choose the flight category of the alter wrt real category proportions of
the real flight data randomly
      # alter_category = sample(egos_with_cat$category, 1)
      alter_category = sample(all_categories, 1)
    } else {
      # Choose the flight category of the alter wrt the flight category of first
ego of the alter's friends
    }

    f_each = NULL
    m_each = NULL
    p_each = NULL

    for (j in 2:length(monthly_freq_avg[alter_category,])) {
      freq = rpois(1, monthly_freq_avg[alter_category,j])
      # categories = rep(alter_category, length(frequencies))

      if (is.null(f_each)) {
        f_each[[j-1]] = alters[i]
        f_each = as.data.frame(f_each)
        f_each[[j]] = freq
      } else {
        f_each[[j]] = freq
      }

      miles = mvrnorm(2, monthly_miles_avg[alter_category,j],
monthly_miles_sd[alter_category,j], empirical = TRUE)
      #miles = rtnorm(1, mean = monthly_miles_avg[alter_category,j], sd =
monthly_miles_sd[alter_category,j], lower = 0.0)

      if (is.null(m_each)) {
        m_each[[j-1]] = alters[i]
        m_each = as.data.frame(m_each)
        m_each[[j]] = miles[1]
      } else {
        m_each[[j]] = miles[1]
      }

      prices = mvrnorm(2, monthly_price_avg[alter_category,j],
monthly_price_sd[alter_category,j], empirical = TRUE)

```

```

#prices = rtnorm(1, mean = monthly_price_avg[alter_category,j], sd =
monthly_price_sd[alter_category,j], lower = 0.0)

  if (is.null(p_each)) {
    p_each[[j-1]] = alters[i]
    p_each = as.data.frame(p_each)
    p_each[[j]] = prices[1]
  } else {
    p_each[[j]] = prices[1]
  }
}

f_each[[length(monthly_freq_avg[alter_category,])+1]] = alter_category
m_each[[length(monthly_freq_avg[alter_category,])+1]] = alter_category
p_each[[length(monthly_freq_avg[alter_category,])+1]] = alter_category

colnames(f_each) = col_names
if (is.null(f_all)) {
  f_all = f_each
} else {
  f_all = rbind(f_all, f_each)
}

colnames(m_each) = col_names
if (is.null(m_all)) {
  m_all = m_each
} else {
  m_all = rbind(m_all, m_each)
}

colnames(p_each) = col_names
if (is.null(p_all)) {
  p_all = p_each
} else {
  p_all = rbind(p_all, p_each)
}
}

flight_data = summarizeFlightInfo(f_all, m_all, p_all)

if (! is.null(fitted_model)) {
  fcast = predict(fitted_model, newdata = flight_data)
  fcast[which(fcast < 0)] = 0
  flight_data[["Value_New"]] = fcast
} else {
  for (i in 1:nrow(flight_data)) {
    cat_index = flight_data[i, "Category"]
    v_new = rtnorm(1, cat_val_new[cat_index, "Average"],
                  cat_val_new[cat_index, "StdDev"],
                  cat_val_new[cat_index, "MinValue"],
                  cat_val_new[cat_index, "MaxValue"])
    flight_data[i, "Value_New"] = v_new
  }
}
return (flight_data)
}

```

```

# Predicted R-Squared
pred_r_squared <- function(linear.model)
{
  lm.anova <- anova(linear.model)
  tss <- sum(lm.anova$"Sum Sq")
  # predictive R^2
  pred.r.squared <- 1 - PRESS(linear.model)/(tss)
  return(pred.r.squared)
}

PRESS <- function(linear.model)
{
  pr <- residuals(linear.model)/(1 - lm.influence(linear.model)$hat)
  PRESS <- sum(pr^2)
  return(PRESS)
}

summarizeFlightInfo = function(mf, mm, mp)
{
  Value = mp[, "Jan-2015"]
  Value_3 = mp[, "Oct-2014"] + mp[, "Nov-2014"] + mp[, "Dec-2014"]
  Value_6 = mp[, "Jul-2014"] + mp[, "Aug-2014"] + mp[, "Sep-2014"] +
    mp[, "Oct-2014"] + mp[, "Nov-2014"] + mp[, "Dec-2014"]
  Value_12 = mp[, "Jan-2014"] + mp[, "Feb-2014"] + mp[, "Mar-2014"] +
    mp[, "Apr-2014"] + mp[, "May-2014"] + mp[, "Jun-2014"] +
    mp[, "Jul-2014"] + mp[, "Aug-2014"] + mp[, "Sep-2014"] +
    mp[, "Oct-2014"] + mp[, "Nov-2014"] + mp[, "Dec-2014"]

  Freq = mf[, "Jan-2015"]
  Freq_3 = mf[, "Oct-2014"] + mf[, "Nov-2014"] + mf[, "Dec-2014"]
  Freq_6 = mf[, "Jul-2014"] + mf[, "Aug-2014"] + mf[, "Sep-2014"] +
    mf[, "Oct-2014"] + mf[, "Nov-2014"] + mf[, "Dec-2014"]
  Freq_12 = mf[, "Jan-2014"] + mf[, "Feb-2014"] + mf[, "Mar-2014"] +
    mf[, "Apr-2014"] + mf[, "May-2014"] + mf[, "Jun-2014"] +
    mf[, "Jul-2014"] + mf[, "Aug-2014"] + mf[, "Sep-2014"] +
    mf[, "Oct-2014"] + mf[, "Nov-2014"] + mf[, "Dec-2014"]

  Av.Tran.Size = Value / Freq
  Av.Tran.Size_3 = Value_3 / Freq_3
  Av.Tran.Size_6 = Value_6 / Freq_6
  Av.Tran.Size_12 = Value_12 / Freq_12

  Av.Tran.Size[which(!is.finite(Av.Tran.Size))] = 0
  Av.Tran.Size_3[which(!is.finite(Av.Tran.Size_3))] = 0
  Av.Tran.Size_6[which(!is.finite(Av.Tran.Size_6))] = 0
  Av.Tran.Size_12[which(!is.finite(Av.Tran.Size_12))] = 0

  Miles = mm[, "Jan-2015"]
  Miles_3 = mm[, "Oct-2014"] + mm[, "Nov-2014"] + mm[, "Dec-2014"]
  Miles_6 = mm[, "Jul-2014"] + mm[, "Aug-2014"] + mm[, "Sep-2014"] +
    mm[, "Oct-2014"] + mm[, "Nov-2014"] + mm[, "Dec-2014"]
  Miles_12 = mm[, "Jan-2014"] + mm[, "Feb-2014"] + mm[, "Mar-2014"] +
    mm[, "Apr-2014"] + mm[, "May-2014"] + mm[, "Jun-2014"] +
    mm[, "Jul-2014"] + mm[, "Aug-2014"] + mm[, "Sep-2014"] +
    mm[, "Oct-2014"] + mm[, "Nov-2014"] + mm[, "Dec-2014"]

  Value_New = rep(0, length(Value))

  mr = calculateMonthlyRecency(mf)
}

```

```

    flight_data = cbind(mf[, "Id"], mf[, "Category"], Value, Value_3, Value_6,
Value_12, Freq, Freq_3, Freq_6, Freq_12,
                        Av.Tran.Size, Av.Tran.Size_3, Av.Tran.Size_6, Av.Tran.Size_12,
                        Miles, Miles_3, Miles_6, Miles_12, Value_New, mr[, "Recency"])
    colnames(flight_data) <- c("Id", "Category", "Value", "Value_3", "Value_6",
"Value_12", "Freq", "Freq_3", "Freq_6", "Freq_12",
                              "Av.Tran.Size", "Av.Tran.Size_3", "Av.Tran.Size_6",
"Av.Tran.Size_12",
                              "Miles", "Miles_3", "Miles_6", "Miles_12", "Value_New",
"MonthlyRecency")
    flight_data = as.data.frame(flight_data)
    return(flight_data)
}

findSNWithGreatestVariance = function(sn_scores)
{
  max_var = 0.0
  max_index = 0
  for (i in 1:length(sn_scores[1,])) {
    cur_var = var(sn_scores[,i])
    if (cur_var > max_var) {
      max_var = cur_var
      max_index = i
    }
  }

  return(max_index)
}

##### END OF FUNCTIONS #####

##### SOCIAL NETWORK SUPPORTED DATA MINING USING REGRESSION #####

# Read friendship vector
friendship_vector = read.csv2("../flight-data/friendship_vector.csv", sep = ";")

# Read CLV data
clv_data = read.csv2("../flight-data/clv_data.csv", sep = ";")
flight_data_ego = clv_data[c(1,2,3,4,6,9,10,14,15,20), -c(1, 19, 20, 21, 23, 24)]
flight_data_non_ego = clv_data[c(5,11,17,18,19), -c(1, 19, 20, 21, 23, 24)]
clv_data = clv_data[c(1:6,9:11,14:15,17:20), -c(1, 23, 24)]
all_categories = clv_data[, 1]
clv_data = clv_data[, -1]

# Summary of clv_data, univariate statistics
summary(clv_data)

# Variances of all variables except ID
apply(clv_data[,1:21], 2, var)

# standard deviation of all variables except ID
apply(clv_data[,1:21], 2, sd)

# Bivariate statistics
write.csv2(cor(clv_data[,1:21], use="complete.obs"),
file="output/CorrelationMatrix.csv")

# Multiple regression

```

```

fit <- lm(Value_New ~ Value + Value_3 + Value_6 + Value_12
          + Freq + Freq_3 + Freq_6 + Freq_12
          + Av.Tran.Size + Av.Tran.Size_3 + Av.Tran.Size_6 + Av.Tran.Size_12
          + Miles + Miles_3 + Miles_6 + Miles_12,
          #+ Longevity + Recency + Age,
          data=clv_data)
summary(fit)

leaps <- regsubsets(Value_New ~ Value + Value_3 + Value_6 + Value_12
                   + Freq + Freq_3 + Freq_6 + Freq_12
                   + Av.Tran.Size + Av.Tran.Size_3 + Av.Tran.Size_6 + Av.Tran.Size_12
                   + Miles + Miles_3 + Miles_6 + Miles_12,
                   data=clv_data, nbest = 1, nvmax = 10)
summary(leaps)
png(file="output/BestSubsetRegression.png", width=720, height=405)
plot(leaps, scale="adjr2", main = "Results of Best Subsets Regression Using
Adjusted R2")
dev.off()

fit = lm(Value_New ~ Value_3 + Value_12
          + Freq + Freq_6 + Freq_12
          + Av.Tran.Size_6 + Av.Tran.Size_12
          + Miles_6 + Miles_12,
          data=clv_data)
summary(fit)
pred_r_squared(fit)

fit = lm(Value_New ~ Value_12
          + Freq + Freq_6 + Freq_12
          + Av.Tran.Size_6 + Av.Tran.Size_12
          + Miles_6 + Miles_12,
          data=clv_data)
summary(fit)
pred_r_squared(fit)

fit = lm(Value_New ~ Value_12
          + Freq_6 + Freq_12
          + Miles_3 + Miles_6 + Miles_12,
          data=clv_data)
summary(fit)
pred_r_squared(fit)

fit = lm(Value_New ~ Value + Value_3
          + Freq + Freq_3
          + Av.Tran.Size_3 + Av.Tran.Size_6
          + Miles_6,
          data=clv_data)
summary(fit)
pred_r_squared(fit)

fit = lm(Value_New ~ Value_3 + Value_6
          + Freq_6 + Freq_12
          + Miles_12,
          data=clv_data)
summary(fit)
pred_r_squared(fit)

fit = lm(Value_New ~ Value
          + Freq + Freq_6

```



```

        + Av.Tran.Size_3,
        data=clv_data)
summary(fit)
pred_r_squared(fit)

fit = lm(Value_New ~ Value_6
        + Miles + Miles_3,
        data=clv_data)
summary(fit)
pred_r_squared(fit)

fit_miles = lm(Value_New ~ Miles,
               data=clv_data)
summary(fit_miles)
pred_r_squared(fit_miles)

png(file="output/FlightDataFittingPlots.png", width=1920, height=1080)
orig_par = par(mfrow=c(2,2), ljoin=1)
plot(fit, pch=19, cex=1.3, col="blue")
dev.off()
par(orig_par)

# Read category based averages and standard deviations to generate flight data for
Facebook data
monthly_freq_avg = read.csv2("../flight-data/cat_monthly_freq_avg.csv", sep = ";")
monthly_miles_avg = read.csv2("../flight-data/cat_monthly_miles_avg.csv", sep =
";")
monthly_miles_sd = read.csv2("../flight-data/cat_monthly_miles_sd.csv", sep = ";")
monthly_price_avg = read.csv2("../flight-data/cat_monthly_price_avg.csv", sep =
";")
monthly_price_sd = read.csv2("../flight-data/cat_monthly_price_sd.csv", sep = ";")
cat_value_new = read.csv2("../flight-data/cat_value_new.csv", sep = ";")

# Read Social Score Data
sn_score_data = read.csv2("../flight-data/sn_score_data.csv", sep = ";")
sn_score_data = as.data.frame(sn_score_data)
sn_score_data_ego = sn_score_data[which(sn_score_data$SN_Id == 0 |
sn_score_data$SN_Id==107 |
sn_score_data$SN_Id==348 |
sn_score_data$SN_Id==414 |
sn_score_data$SN_Id==686 |
sn_score_data$SN_Id==698 |
sn_score_data$SN_Id==1684 |
sn_score_data$SN_Id==1912 |
sn_score_data$SN_Id==3437 |
sn_score_data$SN_Id==3980),]

all_edges = read.table("facebook-data/facebook_combined.txt", col.names =
c("Source", "Sink"))
combined_data_ego = cbind(sn_score_data_ego, flight_data_ego[,-1])
max_var_ss_index = findSNWithGreatestVariance(sn_score_data[, -1])
sn_score_ordered = sn_score_data[order(sn_score_data[,max_var_ss_index+1]), ]
sn_ordered_id = sn_score_ordered[, 1]

all_nodes = unique(c(all_edges$Source, all_edges$Sink))
node_count = length(all_nodes)
col_names = c("Id", "Jan-2014", "Feb-2014", "Mar-2014", "Apr-2014", "May-2014",
"Jun-2014", "Jul-2014", "Aug-2014", "Sep-2014", "Oct-2014", "Nov-2014", "Dec-2014",
"Jan-2015", "Category")
egos = unique(friendship_vector$ego_id)

```

```

alters = setdiff(all_nodes, egos)
sn_ordered_alters = setdiff(sn_ordered_id, egos)
ms_members = sample(alters, 5, replace = FALSE)
non_ms_members = setdiff(alters, ms_members)
sn_ordered_non_ms = setdiff(sn_ordered_alters, ms_members)

sn_score_data_non_ego = sn_score_data[which(sn_score_data$SN_Id == ms_members[1]),]
for (i in 2:length(ms_members)) {
  sn_score_data_non_ego = rbind(sn_score_data_non_ego,
sn_score_data[which(sn_score_data$SN_Id == ms_members[i]),])
}
combined_data_ms_5 = cbind(sn_score_data_non_ego, flight_data_non_ego[, -1])

#ms_members = sample(non_ms_members, 200, replace = FALSE)
ms_members = base::union(sn_ordered_non_ms[1:100], sn_ordered_non_ms[3925:4024])
non_ms_members = setdiff(non_ms_members, ms_members)
flight_data_ms_200 = assignFlightData(ms_members, friendship_vector, all_categories,
homophily = 0.0, NULL, col_names,
monthly_freq_avg, monthly_miles_avg,
monthly_miles_sd,
monthly_price_avg, monthly_price_sd,
cat_value_new)

# Remove "Category" column from flight_data_ms_200 while combining
combined_data_ms_200 = combineSNandFlightData(flight_data_ms_200[, -2],
sn_score_data)
new_values = updateValueNew(combined_data_ms_200[, c(1, 25)],
sn_score_data[, c(1, max_var_ss_index+1)])
combined_data_ms_200[[25]] = new_values[, 2]
combined_data_ms_15 = rbind(combined_data_ego, combined_data_ms_5)
combined_data_ms_ego = rbind(combined_data_ms_15, combined_data_ms_200)

# Base regression model's coefficients of determination (adjusted and predicted R2)
for 215 people
fit_215 = lm(Value_New ~ Value_6
+ Miles + Miles_3,
data=combined_data_ms_ego)
summary(fit_215)
pred_r_squared(fit_215)

# Coefficients of determination of base regression model updated to include only
Miles factor
fit_215_miles = lm(Value_New ~ Miles,
data=combined_data_ms_ego)
summary(fit_215_miles)
pred_r_squared(fit_215_miles)

fit_15 = lm(Value_New ~ Value_6
+ Miles + Miles_3,
data=combined_data_ms_15)
summary(fit_15)
pred_r_squared(fit_15)

leaps_15_ss <- regsubsets(Value_New ~ Value_6 + Miles + Miles_3
+ SN_Auth_Score + SN_Btw_Cntr + SN_Cls_Cntr + SN_Degree +
SN_HubScore + SN_PageRank,
data=combined_data_ms_15, nbest = 1, nvmax = 12)
summary(leaps_15_ss)
png(file="output/BestSubsetRegressionSS_15.png", width=1920, height=1080)

```

```

plot(leaps_15_ss, scale="adjr2", main = "Best Subset Regression for the Model
Including Social Scores of 15 people Using Adjusted R2")
dev.off()

leaps_ss <- regsubsets(Value_New ~ Value_6 + Miles + Miles_3
                      + SN_Btw_Cntr + SN_Cls_Cntr + SN_Degree + SN_HubScore +
SN_PageRank,
                      data=combined_data_ms_ego, nbest = 1, nvmax = 12)
summary(leaps_ss)
png(file="output/BestSubsetRegressionSS.png", width=1920, height=1080)
plot(leaps_ss, scale="adjr2", main = "Best Subset Regression for the Model
Including Social Scores Using Adjusted R2")
dev.off()

leaps_ss_miles <- regsubsets(Value_New ~ Miles
                            + SN_Btw_Cntr + SN_Cls_Cntr + SN_Degree + SN_HubScore +
SN_PageRank,
                            data=combined_data_ms_ego, nbest = 1, nvmax = 12)
summary(leaps_ss_miles)
png(file="output/BestSubsetRegressionSSMiles.png", width=1920, height=1080)
plot(leaps_ss_miles, scale="adjr2", main = "BSR for the Model Including Miles and
Social Scores Using Adjusted R2")
dev.off()

leaps_ss_nohs <- regsubsets(Value_New ~ Value_6 + Miles + Miles_3
                            + SN_Btw_Cntr + SN_Cls_Cntr + SN_Degree + SN_PageRank,
                            data=combined_data_ms_ego, nbest = 1, nvmax = 12)
summary(leaps_ss_nohs)
png(file="output/BestSubsetRegressionSSNoHS.png", width=720, height=405)
plot(leaps_ss_nohs, scale="adjr2", main = "Best Subsets Regression for the Model
Including Social Scores Using Adjusted R2")
dev.off()

leaps_ss_miles_nohs <- regsubsets(Value_New ~ Miles
                                  + SN_Btw_Cntr + SN_Cls_Cntr + SN_Degree + SN_PageRank,
                                  data=combined_data_ms_ego, nbest = 1, nvmax = 12)
summary(leaps_ss_miles_nohs)
png(file="output/BestSubsetRegressionSSMilesNoHS.png", width=1920, height=1080)
plot(leaps_ss_miles_nohs, scale="adjr2", main = "BSR for the Model Including Miles
and Social Scores Using Adjusted R2")
dev.off()

# Number of factors in the base model and the model with social scores should be
the same,
# so the following model which gives the highest adjusted R2 value using three
factors was chosen
fit_ss_miles = lm(Value_New ~ Value_6
                  + SN_Btw_Cntr + SN_Degree,
                  data=combined_data_ms_ego)
summary(fit_ss_miles)
pred_r_squared(fit_ss_miles)

fit_ss_miles = lm(Value_New ~ Miles
                  + SN_Btw_Cntr + SN_Degree,
                  data=combined_data_ms_ego)
summary(fit_ss_miles)
pred_r_squared(fit_ss_miles)

```

```

# Increase the standard deviations of monthly_miles_sd and monthly_price_sd by
multiplying both matrices with a constant
monthly_miles_non_ms_sd = monthly_miles_sd;
monthly_price_non_ms_sd = monthly_price_sd;
monthly_miles_non_ms_sd[, -1] = monthly_miles_sd[, -1] * 4
monthly_price_non_ms_sd[, -1] = monthly_price_sd[, -1] * 4

flight_data_non_ms = assignFlightData(non_ms_members, friendship_vector,
all_categories, homophily = 0.0, NULL, col_names,
monthly_freq_avg, monthly_miles_avg,
monthly_miles_non_ms_sd,
monthly_price_avg, monthly_price_non_ms_sd,
cat_value_new)
combined_data_non_ms = combineSNandFlightData(flight_data_non_ms[, -2],
sn_score_data)

# Predict Value_new parameter using base model
fcast = predict(fit_215, newdata = combined_data_non_ms)
fcast[which(fcast < 0)] = 0
combined_data_non_ms[["Value_New"]] = fcast

base_model_combined_data = rbind(combined_data_ms_ego, combined_data_non_ms)
write.csv(file="output/base_model_combined_data.csv", base_model_combined_data, sep
= ";")

# Predict Value_new parameter using new model with social scores
fcast = predict(fit_ss_miles, newdata = combined_data_non_ms)
fcast[which(fcast < 0)] = 0
combined_data_non_ms[["Value_New"]] = fcast

combined_data = rbind(combined_data_ms_ego, combined_data_non_ms)
write.csv(file="output/new_model_combined_data.csv", combined_data, sep = ";")

# Add Category column
category_data = as.data.frame(combined_data_ms_15[, "SN_Id"])
colnames(category_data) = c("Id")
category_data[["Category"]] = c(flight_data_ego[, "Category"],
flight_data_non_ego[, "Category"])
category_data = rbind(category_data, flight_data_ms_200[, 1:2],
flight_data_non_ms[, 1:2])
combined_data = addCategory(combined_data, category_data)

# Add Profitability column
combined_data = addProfitability(combined_data)

# Add ValueNewDisplayIndex column
combined_data = addValueNewDisplayIndex(combined_data)

#==== AIRLINE CUSTOMER VALUE ANALYSES =====#
# Create graph data structure
g = graph.data.frame(all_edges, directed = FALSE,
vertices = combined_data[, c("SN_Id", "Value_New",
"MonthlyRecency", "Category", "Profitability", "ValueNewDisplayIndex")])

# Category analysis
V(g)$color = V(g)$Category
V(g)$color = gsub("^1", "#cc1d1d", V(g)$color)
V(g)$color = gsub("^2", "#5b6a77", V(g)$color)
V(g)$color = gsub("^3", "#db9f2a", V(g)$color)

```

```

png(file="output/CategoryAnalysis.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(g,
            vertex.label=NA,
            layout=layout.fruchterman.reingold,
            edge.arrow.size=.1,
            vertex.size=ifelse(V(g)$Value_New / 1000000 <= 1.0, 2,
                              ifelse(V(g)$Value_New / 1000000 <= 10.0, 3,
                                      ifelse(V(g)$Value_New / 1000000 <= 20.0, 5,
                                              7))))
par(orig_par)
dev.off()
# Category analysis of the customers whose Value_New / 1000000 is greater than 10.0
sgg10 = induced_subgraph(g, which((V(g)$Value_New / 1000000) > 10.0), impl =
"create_from_scratch")
V(sgg10)$color = V(sgg10)$Category
V(sgg10)$color = gsub("^1", "#cc1d1d", V(sgg10)$color)
V(sgg10)$color = gsub("^2", "#5b6a77", V(sgg10)$color)
V(sgg10)$color = gsub("^3", "#db9f2a", V(sgg10)$color)
png(file="output/CategoryAnalysisFiltered.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(sgg10,
            vertex.label=NA,
            layout=layout_nicely,
            edge.arrow.size=.1,
            vertex.size=ifelse(V(sgg10)$Value_New / 1000000 <= 15.0, 2,
                              ifelse(V(sgg10)$Value_New / 1000000 <= 20.0, 3,
                                      ifelse(V(sgg10)$Value_New / 1000000 <= 27.0,
                                              5, 7))))
par(orig_par)
dev.off()

# Value_New analysis
V(g)$color = V(g)$ValueNewDisplayIndex
V(g)$color = gsub("^1", "#ccff99", V(g)$color)
V(g)$color = gsub("^2", "#ffff99", V(g)$color)
V(g)$color = gsub("^3", "#ffcc66", V(g)$color)
V(g)$color = gsub("^4", "#ff9999", V(g)$color)
V(g)$color = gsub("^5", "#ff9966", V(g)$color)
V(g)$color = gsub("^6", "#cc0000", V(g)$color)
V(g)$color = gsub("^7", "#660000", V(g)$color)
png(file="output/ValueNewAnalysis.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(g,
            vertex.label=NA,
            layout=layout.fruchterman.reingold,
            edge.arrow.size=.1,
            vertex.size=as.integer(V(g)$ValueNewDisplayIndex) + 1)
par(orig_par)
dev.off()

# Value_New analysis of the customers whose ValueNewDisplayIndex is greater than or
equal to 3,
# that is, whose Value_New / 1000000 value is greater or equal to 10.0
sgvn = induced_subgraph(g, which(as.integer(V(g)$ValueNewDisplayIndex) >= 3), impl =
"create_from_scratch")
png(file="output/ValueNewAnalysisFiltered.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(sgvn,

```

```

        vertex.label=NA,
        layout=layout_nicely,
        edge.arrow.size=.1,
        vertex.size=as.integer(V(sgvn)$ValueNewDisplayIndex) - 1)
par(orig_par)
dev.off()

# Profitability analysis
V(g)$color = V(g)$Profitability
V(g)$color = gsub("^0.*", "#ccff99", V(g)$color)
V(g)$color = gsub("^1.*", "#ffff99", V(g)$color)
V(g)$color = gsub("^2.*", "#660000", V(g)$color)
png(file="output/ProfitabilityAnalysis.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(g,
            vertex.label=NA,
            layout=layout_fruchterman_reingold,
            edge.arrow.size=.1,
            vertex.size=ceiling(V(g)$Profitability) + 1)
par(orig_par)
dev.off()

sgp = induced_subgraph(g, which(V(g)$Profitability >= 1.2), impl =
"create_from_scratch")
png(file="output/ProfitabilityAnalysisFiltered.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(sgp,
            vertex.label=NA,
            layout=layout_nicely,
            edge.arrow.size=.1,
            vertex.size=ceiling(V(sgp)$Profitability) + 1)
par(orig_par)
dev.off()

# Churn analysis
V(g)$color = V(g)$MonthlyRecency
V(g)$color = gsub("^1[0-9]+", "#660000", V(g)$color)
V(g)$color = gsub("^1", "#ccff99", V(g)$color)
V(g)$color = gsub("^2", "#ffff99", V(g)$color)
V(g)$color = gsub("^3", "#ffcc66", V(g)$color)
V(g)$color = gsub("^4", "#ff9999", V(g)$color)
V(g)$color = gsub("^5", "#ff9966", V(g)$color)
V(g)$color = gsub("^6", "#cc0000", V(g)$color)
V(g)$color = gsub("^7-9", "#660000", V(g)$color)
png(file="output/ChurnAnalysis.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)
plot.igraph(g,
            vertex.label=NA,
            layout=layout_fruchterman_reingold,
            edge.arrow.size=.1,
            vertex.size=ifelse(V(g)$MonthlyRecency <= 5, 2, V(g)$MonthlyRecency -
2))
par(orig_par)
dev.off()

sgmr = induced_subgraph(g, which(V(g)$MonthlyRecency >= 6), impl =
"create_from_scratch")
png(file="output/ChurnAnalysisFiltered.png", width=1920, height=1080)
orig_par = par(cex.lab=0.5)

```

```
plot.igraph(sgmr,  
            vertex.label=V(sgmr)$SN_Id,  
            vertex.label.color="white",  
            layout=layout_nicely,  
            edge.arrow.size=.1,  
            vertex.size=V(sgmr)$MonthlyRecency * 2)  
par(orig_par)  
dev.off()
```



RESUME

PERSONAL INFORMATION

Name Surname : Ahmet Birol ÇAVDAR
Nationality : Turkish Republic
Place/Date of Birth : Ankara – 16.10.1977
Marital Status : Married
Address : Ahi Mesut Mh. 1872. Cd. No: 11
B-Blok D:35 Etimesgut / ANKARA
E-mail : abcavdar@havelsan.com.tr
Telephone : +90 (532) 6476913



EDUCATION

High School : Keçiören High School (Ankara) - 1994
Undergraduate : Hacettepe University,
Computer Engineering Department (Ankara) - 2000

PROFESSIONAL EXPERIENCE

Bilişim Ltd. Şti., Ankara

Software Developer, June 2000 – September 2003

HAVELSAN, Ankara

Team Leader, September 2003 – Today

FOREIGN LANGUAGES

English