CUSTOMER DEMOGRAPHIC SEGMENTATION BASED ON TELECOM BEHAVIORAL DATA

Dimitar Georgiev Trichkov
D. A. Tsenov Academy of Economics – Svishtov
Department of Marketing
e-mail: dimitar.trichkov@gmail.com

Abstract: In the modern world, digitalization becomes ubiquitous and covers almost every aspect of the business and daily life. Telecom services providers have a major role in these processes due to their involvement in collecting, storing and processing enormous amounts of customer data. This also includes personal telecom services usage data, which if correctly interpreted, might be used for many different purposes. Using telecom data to predict certain demographic characteristics of the customers is helpful in more than one aspect: 1) It could add the acquired knowledge into customer segmentation to better target different customer groups. 2) Such data could be used in cases where traditional historic data is not available- the potential strength of predicting customer credit worthiness based on behavior data is still not fully explored. 3) Last but definitely not least, is the use of data for verifying customer identification in fraud detection.

In this paper, an overview of some successful use of telecom data for non-telecom services is shown, as well as with a set of real telco data, statistical techniques are used to demonstrate the relation between mobile telecom services usage and subscription owners' age. Use of alternative customer data could have enormous implication both on traditional predictive models and could alter the role of the telecoms, making them one of the most important information sources for financial institutions, which operate with sensitive customer data.

Key words: telecom data, decision trees, metric variable prediction
JEL: D4, M31.

ДЕМОГРАФСКА СЕГМЕНТАЦИЯ НА КЛИЕНТИТЕ ВЪЗ ОСНОВА НА ПОВЕДЕНЧЕСКИ ДАННИ ОТ ТЕЛЕКОМИТЕ

Димитър Георгиев Тричков
Стопанска академия Д. А. Ценов“ – Свищов
Катедра „Маркетинг“
e-mail: dimitar.trichkov@gmail.com

Резюме: В съвременния свят цифровизацията става повсеместна и обхваща почти всички аспекти на бизнеса и ежедневието. Телекомите имат основна роля в този процес поради участието им в събирането, съхраняването и обработката на огромни количества клиентска информация. Тя включва използването на телеком услуги от клиентите и, обработена по коректен начин, може да бъде използвана за различни цели. Използването на телеком данни за предвиждане на демографски характеристики е обосновано от: 1) То може да участва в сегментацията на клиентите, подобрявайки таргетирането на различни клиентски групи. 2) Може да се използва например, когато исторически финансови данни не са налични – все още е ограничено използването на поведенчески данни за оценка на кредит риск. 3) И на последно, но не и по значение място е използването на тези данни за идентификация на клиенти в случаи на потенциални опити за измама.

В тази студия се демонстрират успешни модели за използване на телеком данни за реализиране на услуги извън обхвата на телекома, както и с набор от реални телекомуникационни данни и използване на определени статистически инструменти се демонстрира връзката между използването на мобилни телекомуникационни услуги и възрастта на собствениците на абонаментната услуга. Използването на алтернативни клиентски данни може да окаже огромно влияние върху традиционните предиктивни модели и така също да промени ролята на телеком операторите, превръщайки ги в основен източник на информация за финансовите институции, опериращи с чувствителни клиентски данни.

Ключови думи: телеком данни, дърво на решенията, прогноза на метрична променлива.

JEL: D4, M31.

1. Introduction

Mobile communications are now dominating our world. With more than 5 billion unique mobile subscribers and almost 4 billion mobile internet users in 2020, according to GSMA (GSMA, 2020), penetration rates have already surpassed 50% from the world's population. Usage data, generated from all these customers on daily basis is enormous and require complex and powerful solutions for its manipulation: this data will be manipulated and explored in this paper. The objective of this research is the use of telecom traffic and registration data as alternative predictor of customer demographics. Telecoms use network generated and collected data in order to deal with different topics in their business: churn prediction, customized targeting; bundling, cross- and up-sell of related services but they are not generally keen on sharing their data with other industries. The goal of this paper is to demonstrate the predictive power of big data analytics on customer demographics, using statistical tools and procedures for telecom clients' data manipulation. It shows some successful alternative use of telecom data for processes and cases, related to other businesses, where demographic predictions were used for non-telecom related services. Tasks that this paper is aiming to consider are:

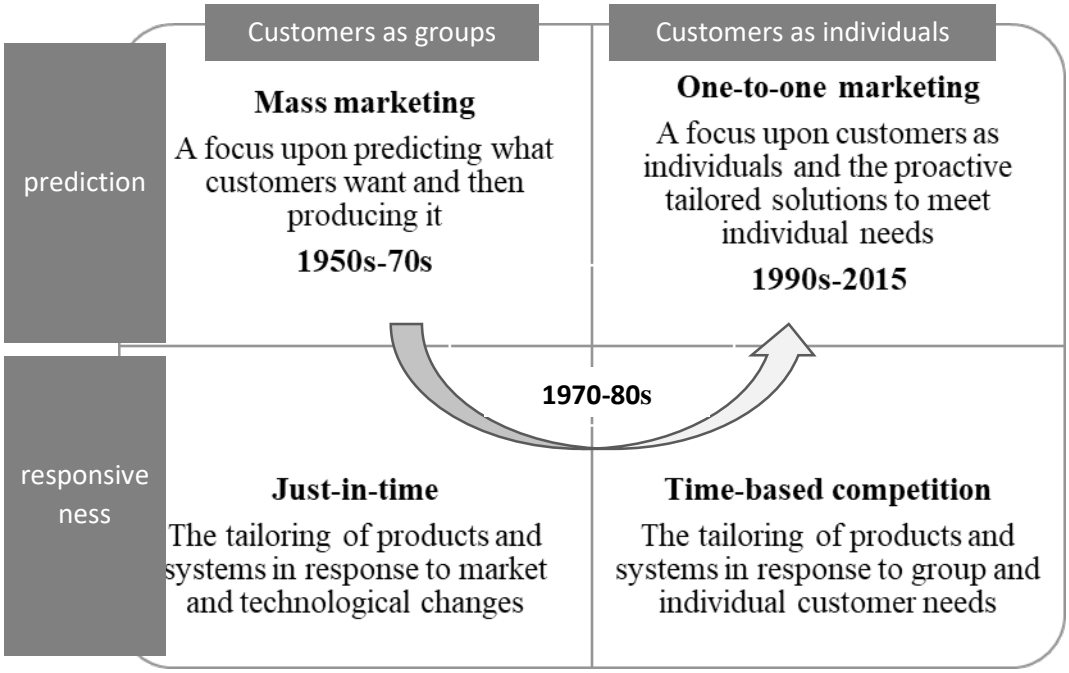
- Show real cases, where large-scale telecom data is successfully used as alternative predictor for solving non-telecom related tasks
- Testing with statistical methods the power of predicting customer age from a set of real telecom usage data
- Summarize the future potential and benefits of using statistical techniques and models for revealing valuable insights hidden in telecom data

Telecoms, being at the edge of mobile penetration boundaries must investigate alternative means for growth and potential new business. If the modern mantra: "Data is the new gold" is correct, telecoms possess huge amount of it and must find a way to monetize it. Big data and Artificial Intelligence opportunities for data analysis and manipulation are to be considered from the telecoms as

strong and stable revenue generators from all kind of core telecom businesses and other industries collaboration.

2. Digital marketing and customer based on data demographics

In its evolution, marketing is always keeping up first with technological progress (Figure 1): in the first half of 20 century, it was the industrial revolution that was behind marketing strategy and nowadays it is the internet adoption and digital transformation that shape marketing concepts and activities. The essence of marketing concepts and main focus have changed, based on the changes in the environment: From early days mass marketing to one-to-one marketing personalization in the modern digital environment:



Source: (Colin Gilligan, 2009, r.34)

Figure 1. The shift from mass marketing to one-to-one marketing

Personalization and individualization of marketing activities are now possible mainly due to increasing penetration of connected devices and the huge amount of data generated by them. Big data analytics and data science are the key success factors for understanding customer behavior and improving customers offers. The 5 "V" characteristics behind Big Data environments described in Big Data Fundamentals: Concepts, Drivers & Techniques (Thomas Erl, 2016) are: volume, velocity, variety, veracity, and value. Companies and businesses collect and process extremely large customers' databases (volume) with their preferences and purchasing behavior. In order to deliver immediate insights data processed in real-time (velocity), enabling marketing scientists to customize marketing instruments to consumers at any given moment. Big data comes not only as simple

numerical data but in many other formats (variety), that might include audio, video or text files. Increasing amounts of big data generated boosts the development of such disciplines as data science, machine learning, text-processing, audio-processing, and video-processing and these are responsible for the biggest challenge namely the accuracy of the data analysis and its applications (veracity). Significant improvement in corporate, performance dependent on Big Data integration is the best proof of the importance and strength of the scientific approach to large datasets (value).

Customer demographics data plays an important role in residential marketing and sales. As described by ironFocus data agency (ironfocus, 2020) age as part of the demographic information, depending on company's product/service, could have a vital role in the market segmentation and microtargeting. Elders care services have a specific age target group and channels for communication as well as cosmetics products have their specific marketing segments based on customers' age. Knowing customers' age helps defining social media platforms to be used in engaging desired target audience. Younger demographics, bellow 24, are more likely to use TikTok, Snapchat or Instagram for example. On the contrary Twitter is mainly used by customers between the ages of 18 and 29. LinkedIn, on the other hand, is a platform for professionals aged between 30 and 49. Facebook and YouTube, although used by customers of any age, are used also by seniors in parallel with email clients etc. Age groups also differ as how they respond to advertising. Generation X response is better to informational advertising, whereas Generations Y and Z, who tend to value friends' opinions and influencers' reviews as more realistic.

Another important basis for segmentation is customer gender. It is natural that men and women tend to have different needs, preferences, interests and buying habits. In most of the households, women typically do family shopping and men are not the typical makeup buyers. In the recent years gender equality policies, adopted by most of developed countries, prohibit advertising based on gender type and new gender stereotypes are defined not based on physical sex. Despite these specifics, gender gap is identified as a major issue in low- and middle-income countries, especially when it comes to women financial inclusion, which is one of the topics on CGAP's list (https://www.cgap.org/, 2020). Closing women cap in telecom services like mobile internet and m-money might unlock socio-economic benefits for women and market opportunities for the mobile industry as a whole. (Dalberg Group, 2018).

Telecom operators collect and process huge amount of data mainly in real time on daily basis. Analysing demographics from this data could reveal different patterns that lead to opportunities how to target different customers groups and help tracking the impact of segment-focused initiatives. Such analysis is also vital for monitoring the effect of targeted marketing communication and campaigns results. Despite its potential value, mostly in the countries with no regulation for services registration, mobile operators do not have access to demographics data. Apart from public regulations, mostly used for security reasons, customers might

have multiple SIM registrations (one person in the family "owning" all the SIMs), muli-SIM ownership (one person with more than one SIM- usually one personal and one provided by the employer) or multi-person SIM (commonly used in the developing countries: one SIM and/or device are shared by more than one person) and others. Telecoms have their internal CDR[1] (call detailed record) and behavioral data in complex DWH systems but yet attempts this to be used for demographics prediction is very limited. Data related to number of contacts, number of calls, calls length, time of call, used handsets, mobile wallet and money transactions, applications and many others could reveal useful information about customers' personal profile. Telecom operators should review the data available and investigate innovative ways of its transformation in order to find new opportunities to reach different customer segments with personalized and demographics-based products and services. In addition, telecom operators could improve their analytics and data modelling by collecting demographic data during the registration process or via call center surveys and cross-validate and improve the statistical results from their internal analytical models.

3. Telecom usage data used for marketing purposes

In the telecom industry, there are several attempts behavioral telco data to be used for predicting customers' demographics. GSM Association together with Dalberg Data Insights have developed and is offering a tool: Connected Women's *Gender Analysis and Identification Tool* (GAIT) that uses machine learning to predict mobile user's gender form his/her CDRs (Dalberg Group, 2018). The main issue addressed by the tool is the digital and financial exclusion of women in the developing countries. Its solution is most effective in countries with predominant prepaid services, that do not require SIM ownership registration and demographic data is expensive to collect on a large scale through surveys or other general means. The lack of reliable information about customers' gender data is the main and most significant barrier to addressing the women exclusion in mobile ownership and financial instruments use (Dalberg Group, 2018). As said in The Mobile Gender Gap Report 2019 (Rowntree, 2019) the mobile industry values closing the gender gap in mobile ownership and usage as important for the societies and the economies in general. In low- and middle-income countries, women typically remain unconnected, digitally and financially excluded, representing a huge marketing segment with a great potential for the telecom operators. Report shows that South Asia women are 26% less likely to own a mobile service than men, and are 70% less likely than men to use mobile internet.

[1] Data record produced by a telecommunications equipment that documents the details of a telecommunications transaction (voice call, text message, internet data session etc.) that passes through that facility or device. The record contains various attributes, such as time, duration, completion status, source number, destination number etc.

The GSMA tool GAIT (Dalberg Group, 2018) purpose is to solve the above issues by successfully forecasting gender based on CDR data. It is designed to be with simple and resource-light implementation with an automated integration process. The two implementation stages are:

1. Performing a survey to generate 'ground truth' about gender data to train the model;

GAIT tool uses statistical supervised model to make predictions about customer base gender based on field survey as an essential first stage in the process. It provides accurate gender flags for sample customer data that is used to train the algorithm to recognize male and female usage patterns.

2. Computing model based on usage patterns of customers and applying the algorithm to estimate subscriber gender.

GAIT is trained with machine learning algorithms on the sample of the customer base for which gender is already known. Parameters, used as predictors are Voice CDRs; SMS CDRs; Data CDRs and Top-up and other telecom traffic data.

GAIT was piloted in Bangladesh with the local telecom Robi Axiata. The true gender of almost 15,000 subscribers was collected via phone interviews. Then the algorithm based on their CDRs data was applied to the whole customer base and the outcome is that the true gender of customers was predicted with over 80% accuracy (63,2% success in male gender prediction and 23,8% in female prediction) as seen in Table 1:

Table 1
GAIT subscriber gender estimates compared to KYC data

% of total customer base (sums to 100%)		Predicted gender (output from GAIT)	
Previous MNO KYC data		MALE	FEMALE
	MALE	63,2%	23,8%
	FEMALE	6,3%	6,7%

Source: (Dalberg Group, 2018, *p .13*)

The analysis of the results revealed significant differences in how different genders use their mobile phones. For instance, the common truth that women's average call duration is much longer than for male customers was proven by the data analyzed. In addition, five features were identified as the most significant subscribers' gender predictors (their importance is included) (Dalberg Group, 2018):

- Average incoming call duration (0.057; on average women accept longer calls);
- Number of different contacts (0.044; women have contact with fewer numbers);

- Average travel distance for an average active day (0.037; women travel less on average);
- Number of distinct cell towers that handled the subscriber's transactions (0.032; women use fewer cell towers on average);
- Radius of gyration for the full study period (0.030; women travel less on average).

This example of how telecom data could be used for not only pure telecom marketing activity, shows the power and potential of both machine learning algorithms and also the hidden information stored in behavioral data of every individual and the immense potential for its use.

Another example for implementation of statistical models on telco data for reaching non-telecom related results is Cignifi credit score (Cignifi , 2020). In their presentation for The World Savings and Retail Banking Institute (WSBI) (Hakim, 2013) they demonstrated the possibility for generating of a personal credit score of a customer based on his mobile phone usage. WSBI and Cignifi wanted to investigate the possible connection/relation between savings behavior and mobile phone usage. They wanted to demonstrate feasibility of using mobile data to identify potential savers. In case of a positive outcome this could enable telecoms to provide banks with an innovative way to qualify and communicate with new market segments. Lessons learned could be also used by other financial institutions and telecoms to reach broader customer segments in their portfolio. The project was performed in Ghana in 2013 with 1 600 common accounts data from the local Airtel telecom's customers and clients of HFC bank. Five customer segments were identified, based on CDR records analysis and different marketing campaigns and channels for communication were then used by both telecom and bank to address these segments.

Conclusions from the project were very promising for the future use of the approach:

Financial institutions specifically in developing countries do not have at the moment efficient way to reach and qualify huge market segments and in such cases mobile phones data could allow them to improve their performance. With smart segmentation based on telecom data financial institutions could be able to target under-served customers, potentially unlocking a dormant customer base, optimizing their portfolios with an innovative product instead of applying mass marketing strategy.

Perhaps Tala (https://tala.co.tz) is the most successful example in the world about innovative and effective use of mobile phone usage customer data for providing tailored financial services. Starting with small credits starting from USD 10 in Tanzania, Tala is now having a growing business in five countries over three continents, reaching a valuation in the $700 million range in 2019 (Shen, 2019). Tala's business model essence is hidden in Android smartphone data collection and analysis. Any potential customer could apply for a loan via mobile app and get an instant decision, regardless his/her financial history. Approved

loans are loaded in person's m-wallet and he starts building his digital credit history. For providing loans to customers with no financial records, Tala browses two categories of data: Android device data and behavioral data. Behavioral data is how customer uses/ moves through our app – for example, whether they read the terms and conditions, whether they provide their real home or work address etc. Handset data include device model, year of production, ID, apps downloaded on the phone, etc. This data is analyzed with machine learning techniques, trained on historic data, in order to determine whether a customer is eligible for a loan. After customer takes a loan, his/her repayment behavior is the most significant predictor of the future lending decisions. Tala follows high ethics standards and explicitly requests users' permission to share their data. Every app access permission in Android handset is individually granted with an explanation of how data will be used. Each person is evaluated regardless his/her gender, race, or other discriminatory factors, same decision criteria is applied in a fully automated algorithm.

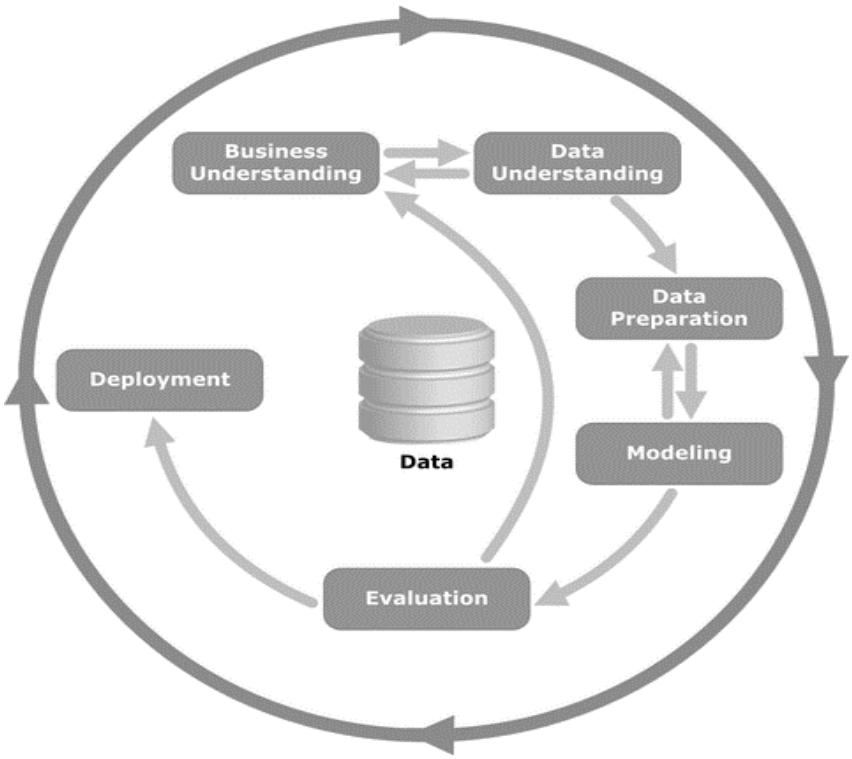
4. Predicting age from telco data- process, data and techniques

In order to test whether telecom usage data is suitable for predicting the age of the customer, we are using an anonymized sample of real data from existing Bulgarian telecom database from 2014 year. The sample is with 106 000 unique customer details[2] and the averages of the monthly totals of their mobile phone usage for 6 months. As stated in their report (World Bank Group, 2017), World Bank Group estimated that more than one billion people around the world lack proof of their legal identity thus their demographic characteristics are unknown to both public authorities and business. This anonymity automatically excludes these individuals from all the activities and benefits provided by the society such as services in the fields of education, healthcare, finance, travel etc. all described in Integrated Biometrics's report from 2017 (Integrated Biometrics, 2017). As shown above in the previous bullet, GSMA offers a tool for successful identification of customers' gender from telecom data. Here it will be tested the next level- to predict customers' age from telecom traffic data. Methodology used is the standard one in data mining community:

- *CRISP-DM Methodology*:

The common model, used for performing tasks of this kind is, as shown on Figure 2, the commonly accepted by data mining specialists CRISP-DM (Cross-industry standard process for data mining):

[2] Dataset used in the research is a selection of anonymized real customer data of existing telecom operator in Eastern Europe



Source: (Hipp, 2000)

Figure 2. CRISP-DM Methodology[3]

Bellow in Figure 3 is shown the deployment process and the tasks, related to each of the phases of the project:

Table 2
Detailed CRISP-DM process and related tasks

Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
• **Determine Business Objectives** • **Assess Situaltion** • **Determine DM Goals** • **Produce project plan**	• **Collect Initial Data** • **Describe Data** • **Explore Data** • **Verify Data Quality**	• **Data Set** • **Select Data** • **Clean Data** • **Construct Data** • **Integrate Data** • **Format Data**	• **Select Modeling Technique** • **Generate Test Design** • **Build Model** • **Assess Model**	• **Evaluate results** • **Review Process** • **Determine Next Steps**	• **Plan Deployment** • **Plan Monitoring and Maintenance** • **Produce Final Report** • **Review Project**

Source: (Hipp, 2000)

In order to be able to predict the dependent variable, our goal should be clear and measurable, data should be formatted, clean and structured in order to be

[3] CRISP-DM is an open standard process methodology for Data Mining analytics conceived in 1996 by companies Integral Solutions Ltd, Teradata, Daimler AG, NCR Corporation and OHRA,

manipulated by the desired model and the results should be evaluated against the initial data provided.

➢ *Data structure, strategy and task*

Choosing analysis strategy depends on the data availability. In case of a lack of previous information about variables correlation, unsupervised model is used, where the training algorithm contains a mechanism for adapting model parameters to their most appropriate values by using appropriate criteria for this purpose. An example of such a task is clustering, where there is no initial information about the allocation of the observed variables to any group (Иванов, 2016).

For the dataset in this paper, supervised model will be used: there the dependent values are defined on the basis of predefined samples with the analyzed object/variable. The values of the parameters of the model are determined by a computational algorithm (process) and they are compared with the values of the variable in the known samples and are corrected so that the error is minimal. The dataset with predefined values of the dependent variable is called training set or training sample. The learning process of the model usually involves two stages: in the first stage the training set is manipulated with statistical procedures so that dependences are detected and described and in the second stage the model is tested for its reliability and accuracy. In the second stage also a set that did not participate in the learning process called "test set" is used together with the original set. In the test procedure of the second stage the calculated dependent variable values of the computed model are compared with the original values of the datasets (Иванов, 2016).

Telecom data, used in this paper includes averages of customers' 6 months bills in 2014of a Bulgarian telecom provider. Monthly totals of bills, traffic data in minutes, number of SMS sent and used MB internet are included but not detailed CDR on daily basis.

- Minutes of use are split into different columns based on numbers dialed– to same provider (onnet), to other mobile provider in the country (offnet) and the share of this minutes in the sum of all minutes used, to fixed numbers (fixed), international and to family members.
- Tariffs' groups are presented as numbers from 1 to 5 depending on related fixed fees. Flag for tariffs with included MB internet is called "Smart" (1-with internet, 0 without)
- Year of SIM first registration is included. In case the owner is changed-still the first year of registration remains in the set.

Data, used here have some specifics, although for all the records (100%) we have customers' age i.e., data is clean and without missing values, it does not correlate 100% with subscription ownership: Europe has strict security and fraud requirements when it comes to sensitive telecom data SIM cards registration. Mandatory registration means that users must provide personal information such as names, ID card number and address in order to subscribe to a mobile service. As of January 2020, GSMA counted 155 countries, where SIM registration is

mandatory (Erdoo Yongo, 2020). In addition, according to Bulgarian telecom providers terms and conditions, SIMs could be registered to business and/or residential customers. One person might have all the family SIMs on his name I order to receive one monthly invoice or have/use both company and private numbers. In countries where telecom providers have strong community offers, dual usage is popular: customers use more than one telecom provider and if possible dual SIM smartphones. In order to avoid or minimize the effects of dual of family SIMs usage in, the analyzed dataset customers with only one SIM per personal ID are included. Such restriction assures with great confidence the fact that the owner and the actual user of the SIM is one and the same person. For the customers in the database, information about their gender and region of birth is extracted, based on what they provided upon registration.

In Bulgaria is mandatory providing customer id for telecom contract signing where gender, age and address data is available and could be extracted in the dataset. Due to customer obligation to provide this data, there is no missing or incorrect values in the dataset. Although dataset consists of only customer with no other subscription under their id and name, it is possible the real user to be different i.e., customer might have a business subscription and provides the SIM to a family member; Persons bellow 18 years are not allowed to conclude a contract, therefore an adult person must subscribe for them[4].

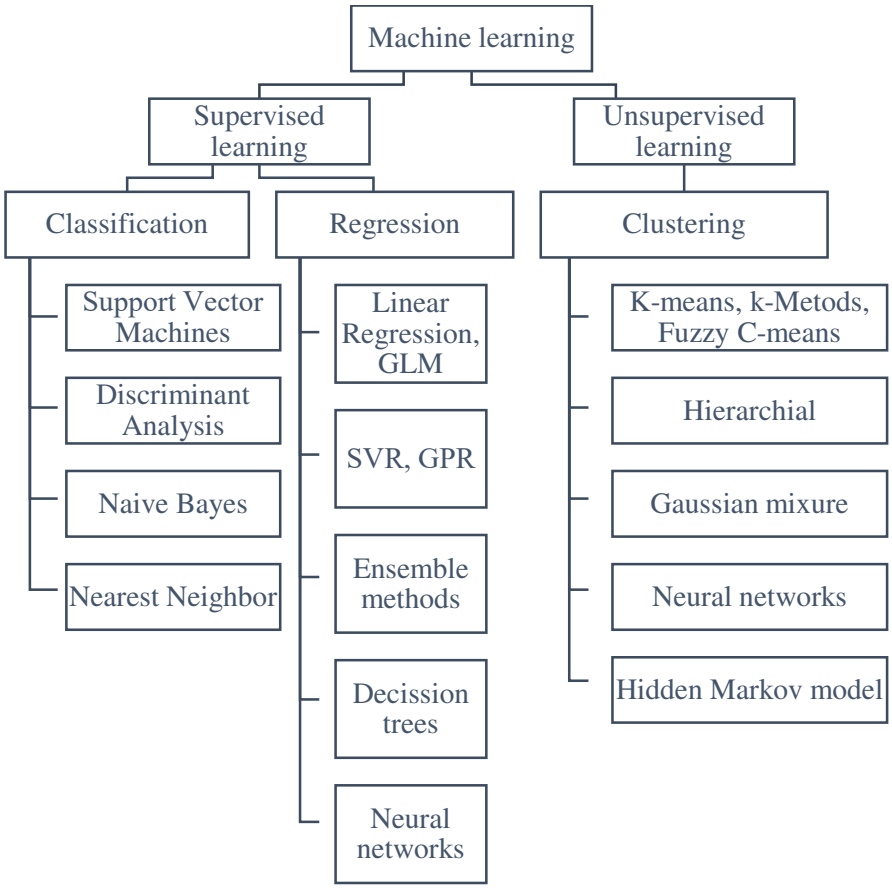
Customer usage data (tariffs and spent minutes/MB/SMS etc.) also might have some imperfections:

- In case of tariff change, only tariff at the moment of data extraction is shown. Traffic data might be related to usage under different tariff.
- Disconnected or reconnected customers are also included in the dataset if they have usage for the monthly period: even if they used only a minute or MB for a single day or even hour
- Missing data for SMS, offnet share, international calls or any other parameter is substituted with zero, as this indicated that this service is not used in the observed period.

➢ Data Mining Techniques

As seen in Figure 3, for supervised tasks, different statistical techniques are used. Classification is that statistical technique, that identifies for a new observation in which of a set of existing categories to be put, on the basis of a training set of data containing instances, whose category membership is already known. This algorithm is suitable for categorical data types of the independent variable and could not be useful for predicting continuous variable. As age is a numerical, thus continuous variable, regression, classification and regression trees, SVM and neural networks algorithms are better to be used for solving this task.

[4] Full list of necessary documents is available at telco's websites:https://www.a1.bg/otgovornost-obshti-uslovia-i-pravila, https://www.vivacom.bg/bg/residential/polezni-syveti/chesto-zadavani-vyprosi



Source: (bhattacharjee, 2017)

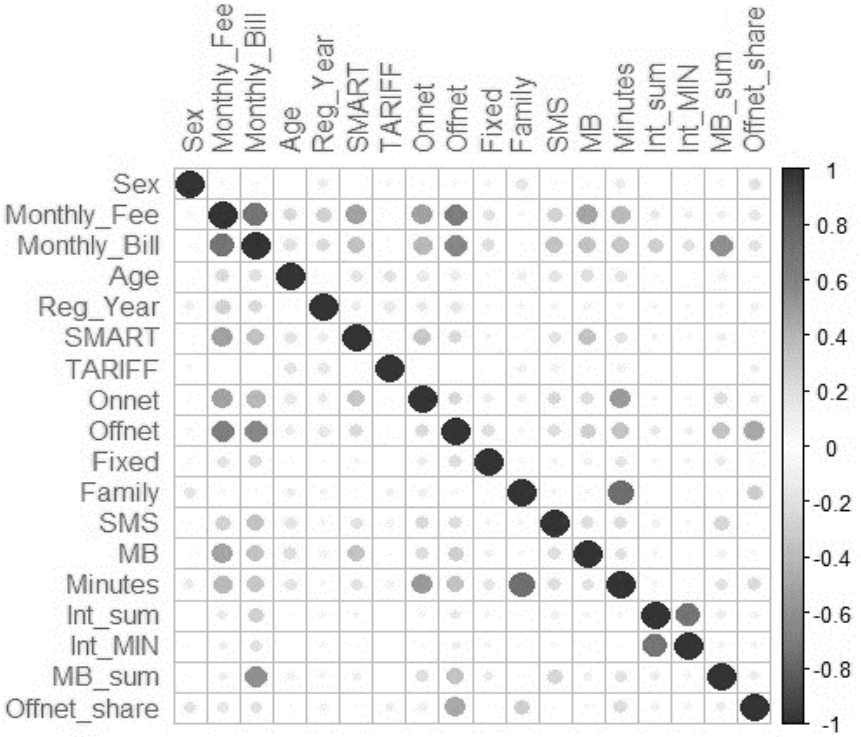
Figure 3. Popular Machine Learning Algorithms

Linear regression assumes linear relationship between the dependent and independent variables, on the contrary logistic regression does not to assume the relation is linear. Logistic regression is a powerful and well-established statistical technique that estimates the probabilities of the target variables. It is analogous to simple linear regression but using classification technique. Another technique is the decision tree: it is a classifier in the form of a tree structure. Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node which is the target value. Neural networks are powerful machine learning algorithms that use complex, nonlinear mapping functions for estimation and classification. These models estimate weights that connect predictors to the output. Models with more complex topologies may also include intermediate, hidden layers, and neurons. The training procedure is an iterative process. Input records, with known outcomes, are presented to the network and model prediction is evaluated with respect to the observed results. Observed errors are used to adjust and optimize the initial weight estimates. They are considered as “black box” solutions since they do not provide an explanation of their predictions. They only provide a sensitivity analysis, which summarizes the predictive importance of the input fields (Иванов, 2016).

5. Evaluation of prediction results

Correlation matrix of the variables in the examined dataset a is shown in Table 3:

Table 3
Dataset correlation matrix



Source: Author's calculation on the examined dataset using R (https://www.r-project.org/)

The obvious dependencies are clear from the above table: monthly bill correlates with monthly fee and high usage, smart flag correlates with used internet (MB), high number of family minutes spent result in overall high number of minutes used but not the bill (as they are usually not charged); on the opposite, high offnet usage affects the bill as these minutes are usually charged higher than minutes in the own network. Variable to be tested- Age has a negative correlation to SMS and MB services: older people normally use new technologies and alternative to voice services less, compared with the younger ones. In addition, older people tend to have older handsets with limited or no internet access.

Using the data analysis tool JMP (www.jmp.com), data was split into training and validation sets. Dependent variable Age is continuous and the relationship with the independent variables is not liner, so regression tree models that work on the principle of reduction of variance are tested: Bootstrap forest, Boosted tree and regression tree.

- Bootstrap forest model, from the JMP Bootstrap Forest Platform (SAS Institute Inc, 2020). The method predicts averages the predicted response value across multiple decision trees. Each tree grows on a bootstrap sample of the training data.

A bootstrap sample, drawn with substitution, is a random sample of observations. Moreover, at any break in the decision tree, the predictors are sampled. Fitting each individual tree is done with bootstrap sample of observations. In case all 100% of the observations are sampled, the expected proportion of unused observations will be 1/e, or approximately 36.8%- these are called out-of-bag observations and the ones used in fitting the tree are called in-bag observations. For continuous variables, in the Bootstrap Forest model predicted values of observations are the sum of their predicted values over the set of individual trees. Results from the model are seen in Table 3 and 4:

Tables 3 and 4
General statistics for Bootstrap forest model in JMP

Individual Trees	RMSE
In Bag	8.34198
Out of Bag	10.88788

	RSquare	RMSE	N
Training	0.30665	10.29252	79760
Validation	0.27493	10.51161	26591

Source: Author's calculation, using JMP(www.jmp.com)

Most important variables, contributing to the model (Table 5) are MB internet used, tariff, SMS, monthly fee and bill and the year of SIM registration. Usage data suggests that younger customers are more engaged in alternative use-SMS and data, while subscription variables (tariff and bill) explain different preferences and financial potential of customers of different age.

Table 5
Most important predictors in Bootstrap forest model

Term	Number of Splits	SS	Portion
MB	1175	577289.13	0.2677
SMS	2462	442287.555	0.2051
TARIFF	349	418924.578	0.1942
Monthly_Fee	1546	144435.198	0.0670
Monthly_Bill	2622	110247.952	0.0511
Reg_Year	2462	108523.49	0.0503
Minutes	2777	102120.899	0.0473
Family	2539	47455.0017	0.0220
Onnet	2475	43297.2423	0.0201
MB_sum	2426	37413.0564	0.0173
Offnet	2372	33499.9756	0.0155
Fixed	2073	32148.6844	0.0149
Offnet_share	1977	29842.8163	0.0138
SMART	155	12022.0966	0.0056
Int_sum	542	8907.75299	0.0041
Int_MIN	438	8366.03236	0.0039

Source: Author's calculation, using JMP(www.jmp.com)

- Boosted tree model in JMP (SAS Institute Inc, 2020)

The process of boosting is constructing a large, additive decision tree by fitting a series of smaller decision trees, named layers. The tree consists of a small number of splits in each layer. Based on the residuals of the previous layers, the tree is fitted, allowing each layer to correct the fit for bad fitting details from the previous layers. For each tree, predicted observation value in a leaf is equal to a mean of all observations in the leaf. The end prediction for an observation over all the layers is the sum of the predictions for that observation. Results from the model are seen in Table 6 and 7:

Table 6
Boosted tree model statistics

	RSquare	RMSE	N
Training	0.264	10.60386	79760
Validation	0.264	10.58881	26591

Table 7
Most important predictors in Boosted tree model

Term	Number of Splits	SS	Portion
MB	19	6876243.21	0.4060
SMS	25	4058114.84	0.2396
TARIFF	31	3967273.22	0.2342
Reg_Year	31	865282.797	0.0511
Minutes	18	665874.308	0.0393
Monthly_Fee	6	162863.86	0.0096
Monthly_Bill	6	129874.361	0.0077
Fixed	7	106612.188	0.0063
Int_MIN	4	62294.8279	0.0037
Offnet_share	3	42281.9257	0.0025
SMART	0	0	0.0000
Onnet	0	0	0.0000
Offnet	0	0	0.0000
Family	0	0	0.0000
Int_sum	0	0	0.0000
MB_sum	0	0	0.0000

Tables' source: Author's calculation, using JMP(www.jmp.com)

As for Boosted tree model – the most important model contributor is consumed internet in MB, followed by SMS, tariff, year of registration and total minutes of use. The model prioritizes the negative correlation between overall

usage and age, also correctly assuming positive correlation between early year of registration and age.

- Regression tree-partitioning (SAS Institute Inc, 2020)

This platform recursively partitions data using the relationship between independent and dependent values, creating a decision tree. The algorithm searches for all possible splits of predictors to make the best prediction. These data partitions are made recursively to form a tree of decision rules. They continue until the desired fit is achieved. Model's parameters are shown in Tables 8 and 9:

Table 8

General statistics in regression trees model

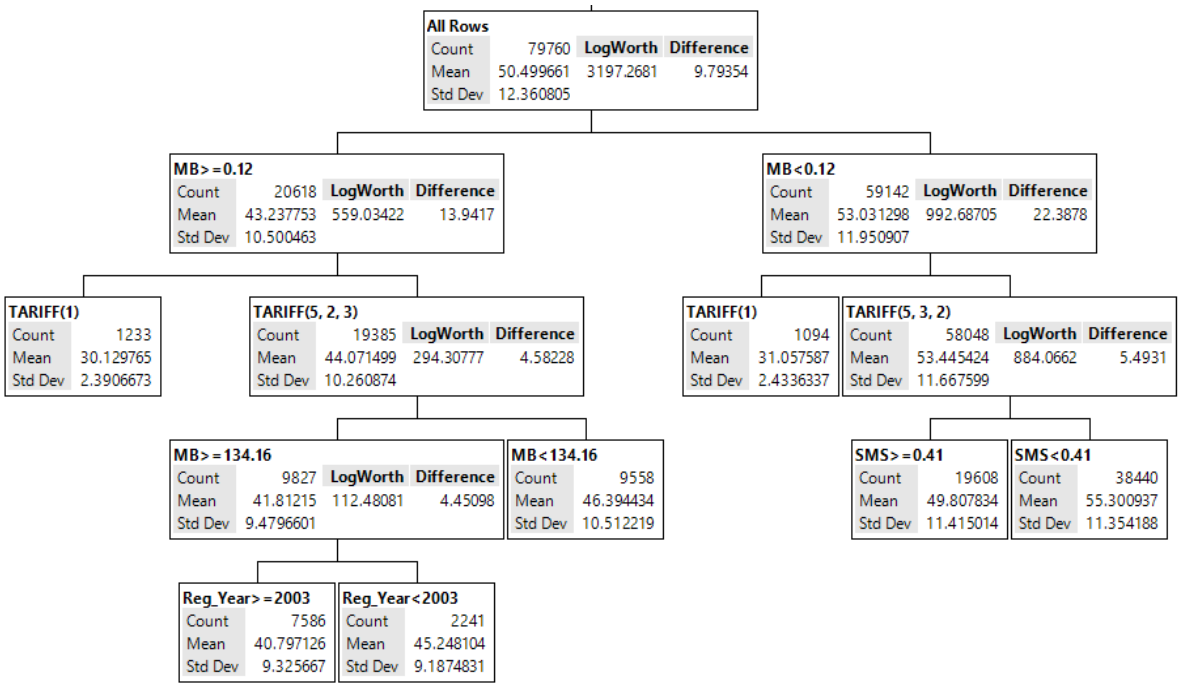
	RSquare	RMSE	N	Number of Splits	AICc
Training	0.264	10.607418	79760	49	603166
Validation	0.260	10.622864	26591		

Table 9

Most important predictors in JMP regression tree model

Term	Number of Splits	SS	Portion
MB	4	1574360.97	0.4902
TARIFF	4	787734.689	0.2453
SMS	12	511527.178	0.1593
Reg_Year	11	155888.868	0.0485
Minutes	5	111882.064	0.0348
Fixed	4	29232.7097	0.0091
Offnet_share	3	15117.6836	0.0047
Monthly_Bill	3	13623.0413	0.0042
Family	3	12588.9171	0.0039
Monthly_Fee	0	0	0.0000
SMART	0	0	0.0000
Onnet	0	0	0.0000
Offnet	0	0	0.0000
Int_sum	0	0	0.0000
Int_MIN	0	0	0.0000
MB_sum	0	0	0.0000

Tables' source: Author's calculation, using JMP(www.jmp.com)



Source: Author's calculation, using JMP(www.jmp.com)

Figure 4. First five layers of regression tree model

Table 9 above shows the most important predictors in the model. Again, the model shows mobile internet use as the major predictor, followed by tariffs, SMS and year of SIM registration and total minutes use. Negative correlation between non-voice services and age is visible in all models, as well as tariffs preferences also vary across generations.

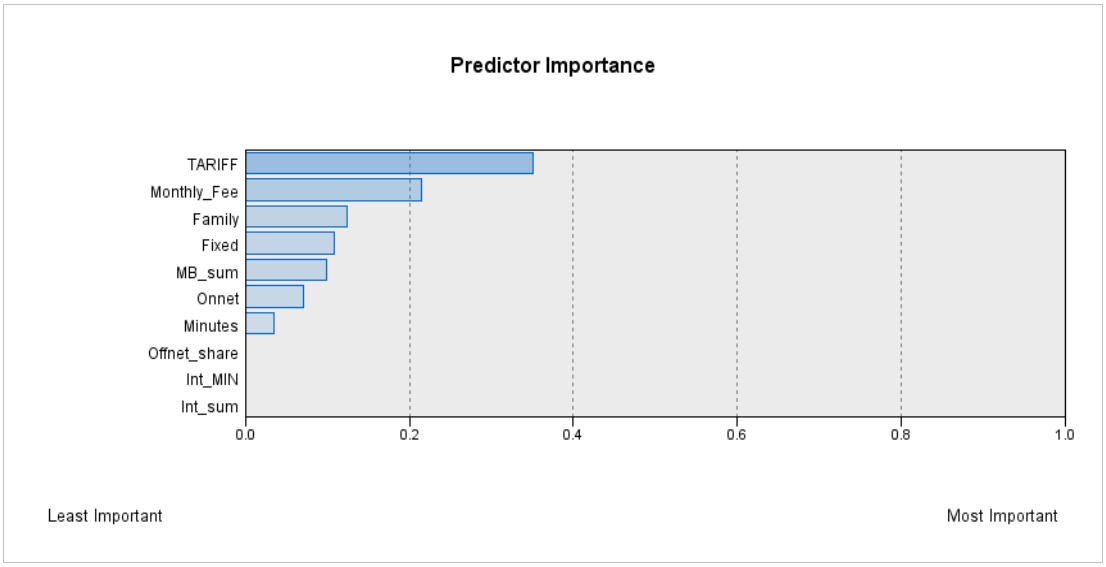
- Comparing predictive models in SPSS Modeler (IBM, 2020) with 50/25/25 training/validation/testing sets (see results in Table 10 and Figure 5)

The model with the highest correlation is C&R trees model. The model uses recursive partitioning for splitting the training observations into splits by minimizing the impurity at each stage. Pure node is one with all of node cases classified in a certain category of the target field. Dependent and independent variables could be numeric or categorical and all splits are binary.

Table 9
Regression models comparison in SPSS Modeler

Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
	C&R Tree 1	< 1	0.498	15	0.752
	Neural Net 1	< 1	0.458	16	0.791
	Regression 1	< 1	0.351	16	0.877
	Generalized Linear 1	< 1	0.351	16	0.877

Source: Author's calculation, using SPSS Modeler



Source: Author's calculation, using SPSS Modeler

Figure 5. Predictors importance in SPSS Modeler regression models

In SPSS Modeler, prediction models prioritize financial predictors like tariff and monthly fee, followed by minutes used to different destination, only after these comes MB internet charged (part of the bill, generated by MB used). In these models voice traffic, generated to different directions has the highest share in the prediction power.

6. Conclusions

This and previous researches demonstrate that predicting customer demographics from a limited set of data could be successfully used for marketing purposes and Telecoms should address these applications in order to benefit from this alternative data usage. In case they extract more detailed and richer input data, their results might be significantly improved. Further testing of predictive power of data mining algorithms and techniques could demonstrate even better performance by enriching input customer data with for example subscribers' physical location and movement data, like GPS and BSC log data, used handsets/devices data, mobile applications used on device; mobile financial data-m-wallet and m-transactions used and a combination of data from different digital channels could boost the results of predictive models. Single customer view, identifying unique user across online and offline platforms is the essence of micro-on single personal level behavioral, financial and social segmentation. For sure, in some countries, difficulties of predicting customers' demographic could exist, as usage patterns, registration processes or other local regulations, as well as technological issues could affect the results in a negative way. Hence, predictive model's success rate is heavily dependent on the cases/countries where

results could be tested against existing real market data. It is crucial to understand that predictions, based on telecom data have an unexploited power to overhaul the way we do business in general. In the digital transformation era, data generated by personal devices and IOT connected appliances and machines increases tremendously at a fast pace. According to IDC it will increase almost four times from 2019 to 2025 (David Reinsel, 2018) and its proper use could boost businesses and even improve quality of life in the developed and developing world. Legal and moral implications, related to this knowledge and its "gatekeepers" should be considered and tight control should be put in place as the possible applications of the Big data are unlimited.

References

(2020, 10 20). Retrieved from https://www.cgap.org/: https://www.cgap.org/topics/collections/womens-financial-inclusion

bhattacharjee, j. (2017, 11 10). *Popular Machine Learning Algorithms*. Retrieved from https://medium.com/: https://medium.com/technology-nineleaps/popular-machine-learning-algorithms-a574e3835ebb

Cignifi . (2020, 10 10). *https://www.cignifi.com/*. Retrieved from https://www.cignifi.com/analytics: https://www.cignifi.com/analytics

Colin Gilligan, R. M. (2009). *Strategic Marketing Planning Second edition.* Oxford: Butterworth-Heinemann.

Dalberg Group. (2018). *The Gender Analysis& Identification Toolkit.* London, UK: GSM Association.

David Reinsel, J. G. (2018). *The Digitization of the World.* IDC.

Erdoo Yongo, Y. T. (2020). *Access to Mobile Services and Proof of Identity 2020:The Undisputed Linkages* . London, UK: GSM Association.

GSMA. (2020). *The mobile Economy 2020.* London, UK: GSM Association.

Hakim, J. (2013). *Using mobile data to reach the unbanked.* Retrieved from https://www.wsbi-esbg.org/: https://bit.ly/35sKWCV

Hipp, R. W. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. Retrieved from https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

IBM. (2020, 10 22). *IBM SPSS Modeler 18.1 Modeling Nodes*. Retrieved from https://www.ibm.com: https://ibm.co/2FZGYcc

Integrated Biometrics. (2017). *Identity in a Developing World.* Spartanburg: Integrated Biometrics. Retrieved from https://integratedbiometrics.com/

ironfocus. (2020, 20 10). *Demographic Segmentation: Why is it important?* Retrieved from https://ironfocus.com/: https://bit.ly/3dYrKR9

Rowntree, O. (2019). *The Mobile Gender Gap Report 2019.* London, UK: GSM Association.

SAS Institute Inc. (2020, 8 13). *Boosted Tree Fit Many Layers of Trees, Each Based on the Previous Layer*. Retrieved from jmp.com: https://bit.ly/2HrOtZL

SAS Institute Inc. (2020, 10 20). *Overview of the Bootstrap Forest Platform*. Retrieved from jmp.com: https://bit.ly/37DweeX

SAS Institute Inc. (2020, 8 13). *Partition Models Use Decision Trees to Explore and Model Your Data*. Retrieved from jmp.com: https://bit.ly/35xVL6r

Shen, L. (2019, 8 21). *Tala, a Company That Offers Loans for as Little as $10, Just Raised $110 Million*. Retrieved 2020, from fortune.com: https://bit.ly/2Tk5Iii

Thomas Erl, W. K. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques.* Prentice Hall Press, USA.

World Bank Group. (2017). *Principles on Identification for Sustainable Development:Toward the Digital Age.* World Bank Group.

Иванов, М. (2016). *СЪВРЕМЕННИ МЕТОДИ ЗА ИНТЕЛИГЕНТЕН АНАЛИЗ НА ДАННИ.* София: НБУ.

НАУЧНИ ИЗСЛЕДВАНИЯ НА ДОКТОРАНТИ

ГОДИШЕН АЛМАНАХ

Том XIII, 2020 г.
Книга 16

СТОПАНСКА АКАДЕМИЯ „Д. А. ЦЕНОВ“ - СВИЩОВ

ГОДИШЕН
АЛМАНАХ
НАУЧНИ ИЗСЛЕДВАНИЯ НА ДОКТОРАНТИ

Том XIII, 2020
Книга 16

Академично издателство
„ЦЕНОВ“- Свищов

ISSN 1313-6542

СЪДЪРЖАНИЕ

Студии

Статии

ГОДИШЕН
АЛМАНАХ
НАУЧНИ ИЗСЛЕДВАНИЯ НА ДОКТОРАНТИ
Студии и статии
Том XIII – 2020, книга 16

Даден за печат на 31.08.2021 г., излязъл от печат 09.09.2021 г.
Поръчка № 18780; формат 16/70/100; тираж 65

ISSN 1313-6542

Издателство и печат: Академично издателство „Ценов“
Свищов, ул. „Цанко Церковски“ 11А