Faculty of Mathematics and Informatics

VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

MODELLING AND DATA ANALYSIS

MASTER'S STUDY PROGRAMME

Mokesčių vengėjų klasifikacija naudojant socialinių tinklų analizę
Classifying Tax Evaders By Means Of Social Network Analytics

Master's thesis

Author: Lukas.Daraganas
VU email address: Lukas.Daraganas@mif.stud.vu.lt
Supervisor: (Asist. Prof., Dr. Dmitrij Celov)

Vilnius, 2022

Abstract

This thesis uses Logistic Regression, Decision Tree and Random Forest methods to classify companies into tax evading firms and tax compliant firms. Based on Lithuanian VAT law two relationships between persons are important – ownership and kinship. Based on these relationships' networks are then constructed for each year in a dataset. From each of these networks their network features are then extracted. These include - network size, community structure and others. Regarding the model accuracy of all the models, it is clear that Random Forest model produced best results in all examined models. If these results would persist with different dataset, Random Forest model with AUC score around 0.7 could be useful when selecting potential auditing targets for Lithuanian State Tax Inspectorate**.**

Keywords: Logistic regression, Decision Trees, Random Forests, Tax evasion, classification

Santrauka

Šis magistro darbas naudodamas Logistinę regresiją, Sprendimų medžius bei Atsitiktinių miškų metodus klasifikuoja įmones į mokesčių vengiančias įmones bei mokesčių nevengiančias įmones. Vadovaujantis Lietuvos PVM įstatymu dviejų tipų santykiai yra svarbūs – nuosavybės ir giminystės. Remiantis šia informacija, kiekvienais metais duomenų rinkinyje yra kuriami socialiniai tinklai, kuriose pagal šiuos ryšius formuojamos grupės. Iš kiekvieno iš šių ryšių sudarančių grupę yra išgaunamos tinklų statistinės charakteristikos. Jos apima - grupės dydį, struktūrą ir kt. Kalbant apie visų modelių tikslumą, akivaizdu, kad Atsitiktinių miškų metodas davė geriausius rezultatus visuose tirtuose modeliuose. Jei šie Atsitiktinių miškų rezultatai, su AUC balu siekiančiu apie 0.7, išliktų su kitu duomenų rinkiniu, ši klasifikacija galėtų būti naudinga įgyvendinant Lietuvos Valstybinės mokesčių inspekcijos audito tikslus.

Raktažodžiai: Logistinė regresija, Sprendimai medžiai, Atsitiktiniai miškai, Mokesčių vengimas, klasifikacija.

Contents

1. Introduction

In December 2021 European Commission (2021) released a report assessing the actual Valued added tax (VAT) gap among EU Member States for 2019. Report measured difference between potential and actual VAT tax amount. It is estimated that EU countries lost € 134 billion in VAT revenue in 2019. Lithuania (21,4%) was among four countries with the largest VAT gap. Only Romania (34,9%), Greece (25,8%) and Malta (23.5%) had a larger VAT gap. In 2018 Lithuania's VAT gap was - 25.9% and was among three largest in European Union. This shows that although progress has been made regarding results from previous year Lithuania is still among top 4 countries with a biggest VAT gap. There is a clearly a need in Lithuania to decrease this gap, in order to have a sufficient tax base for further economic development. One way how this can be accomplished is identifying more tax evasion cases.

As Slemrod (2007) states no government with a tax regime can rely on taxpayers' conscience to pay their fare share. It seems then more effective auditing practices are needed to bridge the gap between Lithuania and other European Union (EU) countries. This factor is recognized in academia as well. According to Hallsworth (2014) taxpayers deterrence approaches do significantly raise compliance, while the evidence for non-deterrence approaches is less conclusive. Bott et al. (2019) state this sentiment in even stronger words – in influencing taxpayer behavior perceived detection probability plays a crucial role.

Above mentioned authors provide strong evidence that tax deterrence is important. However, analyzing one taxpayer behavior is not enough. Better approach is also to consider taxpayer relationship networks. Networks of relationships play prominent roles in a wide variety of social interactions, to the extent that most social interactions are conducted within social networks and not within firms, markets or institutions. These networks can involve a variety of social contracts. Social networks can also be used when detecting tax evasion among taxpayers. According to Gamannossi and Rablen (2020) network information, information received from statistical network analysis, allows controlling institution to have a higher prediction probability in predicting the likely benefits from conducting an audit of a selected taxpayer.

Therefore, the purpose of this master thesis is to examine network characteristics' influence on potential tax evading companies and to check if accurate classification of companies as tax evaders can be done by using network characteristics.

This thesis follows this methodological approach. Network construction in this thesis is based on Lithuanian VAT law. There are two relationships between persons that are important in this context – ownership and kinship. Based on these relationships network is then constructed.

These relationships between natural persons and/or companies are examined by creating undirected graphs between connected individuals or between connected individuals and companies. Network characteristics are then extracted from each of these graphs for every year in analyzed period from 2014 to 2019. These include - Network size, Community structure and others. After this was accomplished, this thesis followed Lismont et al. (2018) approach and used Logistic Regression, Decision Trees and Random Forests models to classify companies into tax evading and compliant companies. For tax evader status five artificial trigger rules were created – therefore there were 15 models examined in total.

Regarding the model accuracy of all the models Random Forest classifier produced the best results in all examined models. Random Forest model AUC (Area Under the ROC Curve) score in four of five models was around 0.7. If these results would persist with Lithuanian STI's (State Tax Inspectorate) actual trigger rules, this thesis as mentioned above used artificial trigger rules, this classification method could be useful when selecting potential auditing targets for Lithuanian STI.

Regarding the value of this thesis, it stems from novelty - there are not many scientific papers on this topic, especially with Lithuanian data.

The remainder of this thesis is divided into five parts. Chapter 2 reviews relevant theoretical background. Chapter 3 provided methodology used in this thesis. Chapter 4 examines data used in this thesis. Chapter 5 presents hypothesis and limitations of this thesis. Chapter 6 presents empirical results. Chapter 7 presents conclusions and discusses further research on this subject.

2. Literature review

This chapter provides the theoretical basis for this thesis. First and foremost, this chapter defines tax evasion as the key research subject of the thesis. Secondly, the relevant research on private firm tax evasion is given. Following this the definition of social network analysis is specified. Finally, evidence of existence of social network effects in tax evasion is given.

2.1 Tax evasion

First of all, to understand the purpose of this thesis proper definition of Tax evasion is needed. However, in order to do this one must first understand that almost no one in society pays their fair share of taxes willingly. As Slemrod (2007) states no government with a tax regime can rely on taxpayers' conscience to pay their fare share. According to the author, some people indeed pay their fair share, but many others do not and over time the ratio between the two shrink in non -societal beneficial favor, as tax paying people see how they are being taken advantage of by the others (in economics this is called a 'free-rider' problem). Therefore - paying taxes must be mandatory duty of citizens with parallel and appropriate penalties on noncompliance.

In academic literature previously described phenomenon is defined by a broader tax dodging concept, where tax evasion is a particular case. Kirchler et al. (2003) define three key terms that fall under tax dodging:

1. Tax avoidance - attempt to reduce one's tax payments by legal channels, for example, by taken advantage of legal tax-loopholes.
2. Tax evasion - illegal reduction of tax payments, for example, by under-reporting income.
3. Tax flight - the relocation of businesses abroad in order to save taxes.

This thesis concentrates on private firms' tax evasion. Features of tax evading private firms are given in the following sub-chapter.

2.2 Tax evasion by firms

Harju et al. (2014) have studied the effect of consumption tax rate on tax evasion by firms. Authors utilized a natural field experiment that varied the probability of an audit together with a VAT reform increasing VAT rate from 9% to 23% for hairdressers and not affecting a similar control group. Authors results indicate that firms in experimental group responded by reporting higher VAT relative to the control group. However, higher VAT rate also statistically increased VAT evasion in experimental group.

Gangl et al. (2014) conducted a field experiment focused on novelty of the effect of tax authorities' supervision. There were no positive overall effect of close supervision on tax compliance for examined newly established companies in so called 'high-risk' sectors. However, evidence had showed that for those who are late to pay their taxes, closer tax authority supervision reduces the amount they owe in taxes – it increased compliance.

Morse et al. (2009) state that underpayment of tax on business income can be most commonly linked to the receipt of cash. Authors note that typical 'cash business owners' mostly rely on parallel grey economies in order to underreport their revenue and evade their taxes. Here tax evasion seems to be best explained by an opportunity to cheat the system, linked with a the low perceived probability of detection, small financial penalty, and by peer group norms.

Abdixhiku et al. (2017) state that the low trust in government and judicial system, higher perception of corruption and higher compliance costs increases tax evasive behavior of firms. Authors found that smaller firms and firms in so called risky sectors, less visible to the tax administration, are more likely to evade their taxes.

Results in Alm et al. (2019) indicate that more financially unstable firms are more likely to be involved in tax evasion activities, tax evasion in their case can help them to deal with financing troubles facing their firm. Authors also state that the effects of firms financial situation are heterogeneous across ownership, age, size. Finally, authors note that firm's financial conditions might impact tax evasion by these behaviors - reduction of information disclosure, an increase in cash transactions and an increase in bribe for tax evasion opportunities.

Wang (2009) using cross-country data of firm-level survey found that competition stimulates a pressure for a firm to get involved in questionable tax reporting behaviors (at a decreasing speed). Author states that business obstacles like tax administration and corruption do also play an important role in explaining tax evasion. Other key factors are firm characteristics such as size, age and ownership.

Joulfaian (2000) examined noncompliance with the personal income tax from a sample of corporate income tax returns and found that preferences of firm's managers does play an important role in determining noncompliance with tax code. Author states that non tax compliant firms are three times more likely to be managed by executives who have misreported their personal taxes (the amount of income under-reported is significantly higher in the presence of such executives).

Tedds (2010) using interval regression on a dataset of multi-country firm level data found that firms under-report their taxes in all regions around the world. According to the author government corruption is largest causal effect on under-reporting, taxes have the second single largest causal effect on under-reporting. There is a significant correlation between the following factors - under-reporting, legal organization, business, size, age, ownership, competition intensity and audit controls.

Artavanis et al. (2015) using microdata on household credit from a Greek bank found that 43%-45% of self-employed income is unreported and untaxed. According to the authors primary tax-evading industries are concentrated in professional services. These include - medicine, law, engineering, education, and media. Other industries linked with tax evasion are lodging, restaurants, and business services.

From these reviewed examples, it can be stated that there are common characteristics of firms exposed to tax evasion phenomenon. There are so called ‘high-risk’ sectors or cash businesses in which tax evasion risk significantly increases. Besides, firm’s sector its background characteristics also matter.

2.3. Social Networks

This sub-chapter presents the definition of Social Networks, followed by their importance, presented in next sub-chapter.

According to Wasserman and Faust (1994) social environment (social networks) can be expressed as patterns or regularities in relationships among interacting units. In this environment set of nodes (representing network members) are linked with each other by one or more types of relations.

Marin & Wellman (2011) claim that social network units are most commonly persons or organizations, however any connected units can be studied as nodes. Also, according to the authors, successful social network analysis requires more than knowing how to measure some characteristics of the networks.

These requirements are:

1. A set of assumptions about how best to describe and explain the social phenomena of interest.
2. Not having false assumption that environments, attributes or circumstances affect actors, in this thesis case actors- firms, independently. Moreover, the user of social networks

approach should not assume the existence of uniformly cohesive and discretely bounded groups. Dempwolf & Lyles (2012) also state that a biggest challenge of network analysis is that the relationships between linked units are assumed as being dependent on each other (G has a relationship with L that relationship is not considered independent of actor G's other relationships with different actors).

3. Context should be taken seriously, network relations themselves are often analyzed in the context of other relations.

2.4. Network effects

Bramoullé et al. (2014) state that geography and social links shape economic interactions. Authors claim that equilibria depend on a single network measure - the lowest eigenvalue. The lowest eigenvalue depends on the two-sidedness of the network (agents can be subdivided into two sets with few links within the sets but many links between them). A network most amplifies substitutability when agents are divided into two distinct sets, and they have links to agents in the other set, but not their own. Actions then rebound from one side to another.

Boning et al. (2020) found that personal visits by tax auditors have a large direct effect on visited firms' tax deposits. However, this effect persists to other clients of visited firms' tax accountants, they also pay more tax. This suggest that a network effect might be present. In this case this network effect accounts for 1.2 times as much revenue as the direct effect.

Gamannossi and Rablen (2020) claim that network information, when examining its benefits to tax authority, better predicts the likely revenue collection benefits from an audit of a particular taxpayer. Authors also state that there is a link between network centrality on a social network and tax evasion.

Lismont et al. (2018) in their paper Predicting tax avoidance by means of social network analytics have identified three potential prediction approaches based on Logistic Regression, Decision Trees and Random Forest models. These techniques were applied on firms specific characteristics, network characteristics and on different combinations of both. Authors connected firms by means of shared board members, currently and in the past resulting in a network (or graph) of firms.

To sum up, the reviewed papers that covered network effects show that these effects are present and could be used to analyze and classify firm tax evasion. Therefore, although this thesis utilizes the similar approach as Lismont et al. (2018), the novelty, in this case, is achieved by controlling the social networks for ‘high-risk’ segments of economic activity to account for cash businesses in which tax evasion risk significantly increases, using different country’s Lithuania’s data and using different variables as social connectors.

3.Methodology and feature extraction

This chapter covers methodological approach of the thesis. The first part explains social network construction from the dataset. The second – classification methods used in the thesis. The third – dependent variable construction. Finally, the fourth part examines how to compare classification performance of different models.

3.1 Social network construction

This thesis uses two types of characteristics:

1. Individual characteristics, alluded in theoretical chapter of the thesis, like the sector in which firm operates. These characteristics are more commonly referred as local features.
2. Network Characteristics, these variables can indicate how knowledge is transferred between members of the network through links between them.

It also important to state that members of the particular network can be connected with each other through various means. Therefore, it is important to construct networks properly. Regarding this thesis and network construction it is based upon Lithuanian VAT law. According to INFOLEX (2021) commentary of Lithuanian VAT law two relationships between persons are important – ownership and kinship.

Ownership relationship which is defined as having ownership status in one or more persons, this usually means owning shares in a company. This relationship can be defined from natural person to company or from company to company (Natural Person → Company or Company → Company).

Kinship relationship can only be from natural person to natural person. (Natural Person → Natural Person).

Based on these relationships social network is then constructed. Due to, nature of the data used in this thesis, mainly its anonymous, relationships between natural persons and/or companies are examined by creating undirected graphs between connected individuals or between connected individuals and companies.

Connected persons here are simply connected components, nodes in a subgraph, in which each pair of nodes links with each other via a path, where any node from the set of nodes can reach any other node by traversing edges - all the nodes in a subgraph are always reachable from each other. Relationships to connect them as mentioned earlier are ownership and kinship that are reexamined every year. If a person stops owning shares or he or she dies, then the relationship is no longer valid when moving to the next year. To sum up, every connected subgraph represents one social network in examined year.

After doing this step, network features or Network statistics are extracted from the analyzed annual data on social networks. These features according to Chiesi (2015) fall under categories described in Figure 1.

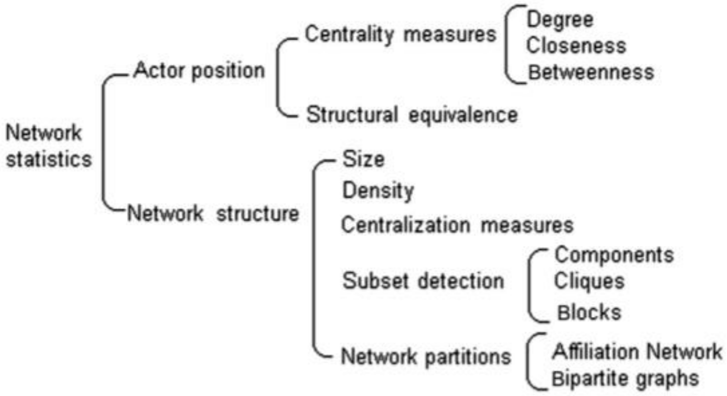

Figure 1. Network Features. Adapted from " Network Analysis. International Encyclopedia of the Social & Behavioral Sciences," by Chiesi, A. M. 2015, Encyclopedia of the Social & Behavioral Sciences, p.522. Copyright 2015 by Encyclopedia of the Social & Behavioral Sciences.

The full list of network features used in this thesis is given in an appendix. The short definitions of network features are available in Igraph (2021), more detailed definitions are available in Boccaletti et al. (2006).

Applied network feature extraction allowed to use these features by non-relational predictive analytic techniques discussed in the next section.

3.2 Classification methods used

Following Lismont et al. (2018), this thesis applies the following techniques that are very common in solving a classification task: Logistic Regression, Decision Trees, and Random Forests.

Hastie and Friedman (2017) give the following definitions of these techniques :

1) Logistic Regression can be defined as a supervised learning classification algorithm, it is used to model the posterior probabilities of the K classes by using linear functions in x, while at the same time ensuring that the value sum to one and remain in [0, 1] range.
2) Decision Tree is a supervised learning technique that can be used both in classification and regression tasks. It has a pre-defined target variable and is widely used in classification problems. Decision Tree predicts the value of target variables by learning decision rules inferred from data used in its training. This method is conceptually simple yet powerful.
3) Random Forest is an ensemble learning technique which constructs and uses multiple Decision Trees and combines them into one model which in theory should provide better performance in classification task.

This thesis uses multiple approaches to provide different perspectives on predictive models and to gain a better understanding which method is better for analyzing a social network classification task.

3.3 Dependent variable construction

Regarding the classification task it is reasonable to assume that STI won't waste its time and resources on an audit unless there's a good chance that the STI can collect additional money. It is also reasonable to assume that STI like its counterparts abroad like IRS in United States likely has their triggers – rules which specify that its agents must do an audit. One of such triggers is usually income, for example, in US, individual is the safest from the attention of audit authority, if her reported income is between $25,000 and $200,000 (Internal Revenue Service 2017).

Moreover, it also very reasonable to assume that income related triggers are very confidential information, which are not to be shared with a public and possible tax evaders.

According to Lithuanian VAT law (STI 2021), individuals must pay VAT if their annual income is higher than 45 thousand EUR, this includes income from controlled (directly or indirectly) companies as well. Using this information and assuming that most individuals legally making above 45 thousand EUR are more likely to immerse themselves in tax avoidance and not tax evasion, this thesis uses 45 thousand Euros as a benchmark and creates five binary dependent variables that are used for this thesis classification problem:

1) All persons with income under 45000 EUR as 1 all others as 0;

2) Persons with income under 45000 EUR, but larger than 0 (there are companies' in a sample with 0 revenue) as 1 all others as 0;

3) Persons with income over 10000 EUR, but under 45000 EUR as 1 all others as 0;

4) Persons with income over 20000 EUR, but under 45000 EUR as 1 all others as 0;

5) Persons with income over 30000 EUR, but under 45000 EUR as 1 all others as 0.

All these different dependent variables help to examine classification task more robustly accounting for different size income triggers that might be applied for income of this size - 45 thousand EUR and below. Therefore, there are 15 models in total., 5 for each classification method used.

3.4 Comparing performance of different models

All of the above-mentioned classification models have been tested on a training sets. Between each other models were compared on their accuracy and their area under the ROC curve (AUC). This was also the approach recommended in Lismont et al. (2018).

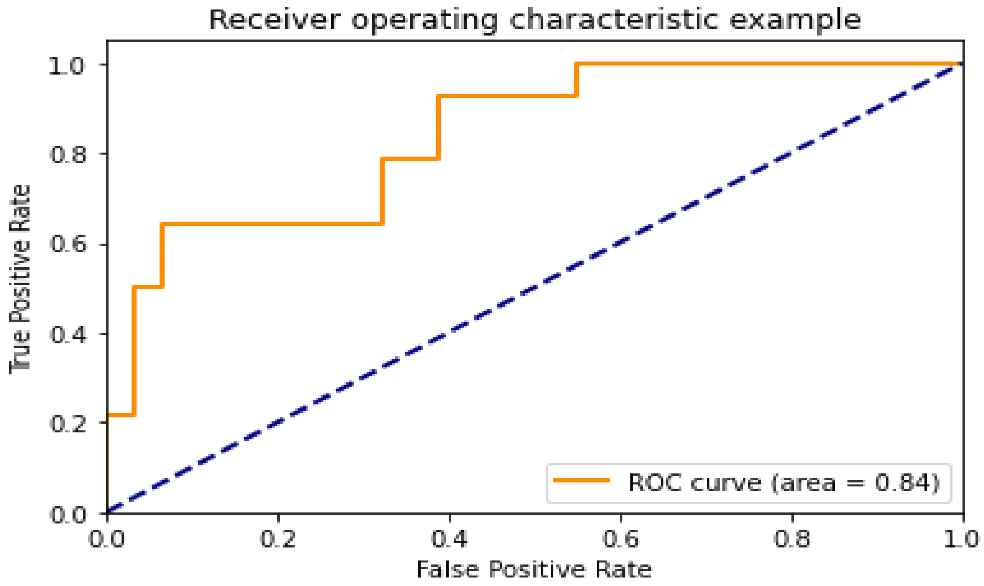

Figure 2. ROC curve drawn by the author.

Muschelli (2020) gives the following definitions :

- Accuracy or overall accuracy is sensitivity and specificity measure which considers both true positives and True negatives in the measured variable.

- Receiver operating characteristic (ROC) shows how a predictor compares to the true outcome, how true positive rate (sensitivity) is changing with varying true negative rate or specificity for different thresholds. The predictive capabilities of a variable in ROC are summarized by the area under the curve (AUC),this area is calculated by integrating areas under the line segments.

Lismont et. al (2018) also adds that the closer the ROC curve is to the top left, and thus the higher the area under this curve, the better the model performs. AUC gives an aggregate measure of model's performance across all thresholds. AUC is the probability that the model ranks a random positive example more highly than a random negative example.

Example of ROC curve is showed in Figure 2.

4. Data

The following chapter presents data used in this thesis. The first section describes the data, while the second section summarizes characteristics of the dataset.

4.1.Dataset description

The dataset used in this thesis is provided by STI and originates from various financial reports and other information available on STI databases. It should be noted that before providing the data used in this thesis STI applied anonymization procedure, protecting sensitive personal information by encrypting identifiers that connect particular individuals to stored data. The dataset covers annual observations for the period from 2014 to 2019. It consists of the three types of information:

1. Company characteristics (Revenue, Number of Employees, Sector).
2. Anonymized kinship relationship'.
3. Ownership (shareholder) information (who owns which shares).

Anonymized kinship relationship and Ownership (shareholder) information are only used for network construction. Company characteristics and extracted network characteristics are used for classification task.

4.2. Descriptive characteristics of the dataset

In total after network construction total number of observations in this thesis is 4312. Descriptive statistics of selected variables are provided in Table 1.

Table 1. Selected descriptive statistics.

	Number of employees	**Revenue**	**Degree**	**Nodes**	**Number of Communities**
mean	7.27	150581.01	3.02	22.68	3.33
std	153.45	308663.94	2.97	49.61	3.90
min	0.00	0.00	1.00	3.00	1.00
25%	0.00	0.00	2.00	4.00	1.00
50%	1.00	15585.50	2.00	7.00	2.00
75%	3.00	135052.00	3.00	17.00	4.00
max	10000.00	1864880.00	86.00	462.00	39.00

Descriptive statistics of these variables show that there are large differences between analyzed firms. Also, it should be noted that due to data anonymization it cannot be checked if the data provided by STI is correct. Hence, assumption is made that the data is correct.

Out of selected variables two of the most important one's are firm revenue (from which dependent variables are constructed) and network size (which is calculated by a number of nodes in a network) annual changes are shown below in Figure 3 and Figure 4 respectively. Annual median values are taken to better account for variation in a data.

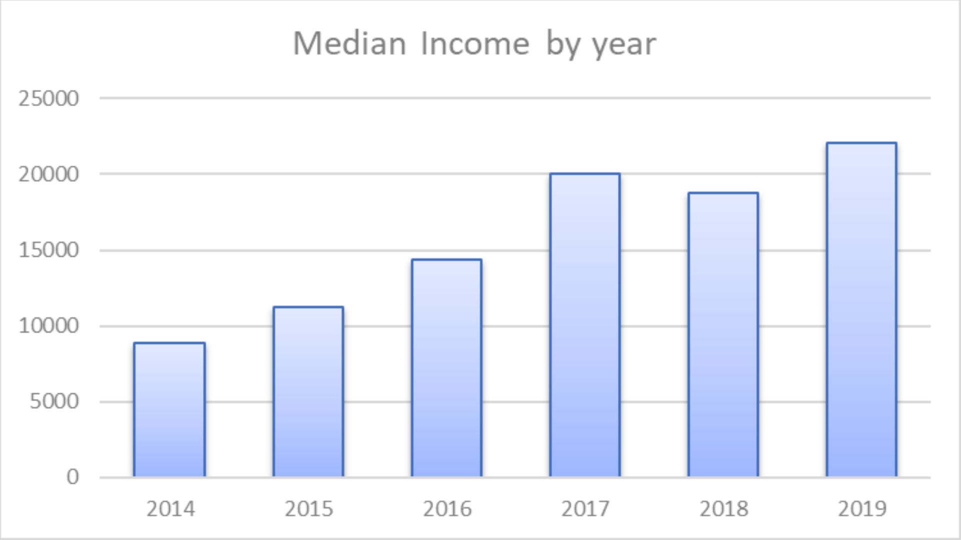

Figure 3. Median Income by year

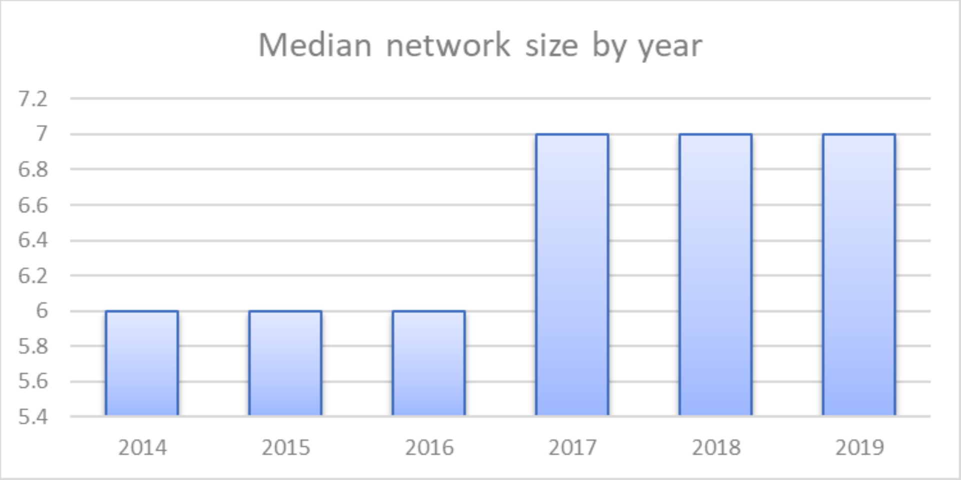

Figure 4. Median network size by year

In general, median firm income increased in all but one-year 2018. Size of the median network increased in 2017 and remained the same in 2018 and 2019. It can be seen that by the end of the period both median network size and company revenue increased from 2014.

5.Hypothesis and limitations of the data

This chapter covers hypothesis and limitations of this thesis.

5.1.Hypothesis

This thesis aims to empirically test the following hypotheses:

H1. Network characteristics' are statistically significant explanatory variables when they are used to classify potential tax evading companies. These characteristics provide important information about the particular company and should be used in tax evasion detection classification task.

H2. Network size does matter and have a positive impact when classifying potentially tax evading companies, different network sizes are important when they are use to indicate tax evading company's.

H3. Communities in a networks are important and have a positive impact when classifying potentially tax evading companies, community number is an important variable in successful tax evaders classification task.

H4. Individual connections also matter in a network and have a positive impact when classifying potentially tax evading companies, number of connections are important when detecting potential tax evading company's.

H5. Company specific characteristics are also statistically significant explanatory variables and should be used in classification procedure, these characteristics provide valuable information in tax evaders classification task.

5.2. Limitations of the data

This thesis has three main limitations:

L1. The usage of anonymized data, there is no way to check validity of the data used.

L2. The absence of precise income trigger rules. It would be better to check classification precision with exact income trigger rules STI applies.

L3. Incomplete firm specific characteristics data. There is some important characteristics – firm age, expenses and background of its shareholders that are missing and would have helped in this classification task.

These three limitations might impact classification precision compared with precision reached when using STI working dataset with true income trigger rules.

6. Results and discussion

This section presents and discusses the results after training models covered in Methodology chapter. First results from Logistic Regression, Decision Trees and Random Forest are covered. Next implications of the results and potential changes for future research are discussed.

6.1. Results

All models were trained on the training data sets. Stratified training and test samples were created in order to have balanced output class in both train and test sets and in order to avoid overfitting or underfitting. The final training of the models was done after the feature selection was done on the training sets. This was performed using Recursive feature selection with 5-fold cross-validation. Also, the potential multicollinearity for Logistic Regressions' was checked using Variance Inflation Factor (VIF), there were no variables above 10 threshold value.

The following models are defined by their dependent variables:

- Model 1 - All persons with income under 45000 EUR as 1 all others as 0.
- Model 2 - Persons with income under 45000 EUR, but larger than 0 (there are companies' in a sample with 0 revenue) as 1 all others as 0.
- Model 3 - Persons with income over 10000 EUR, but under 45000 EUR as 1 all others as 0.
- Model 4 - Persons with income over 20000 EUR, but under 45000 EUR as 1 all others as 0.
- Model 5 - Persons with income over 30000 EUR, but under 45000 EUR as 1 all others as 0.

Models' dependent variable distribution is given below.

Table 2. Models' dependent variable distribution

	Model 1	Model 2	Model 3	Model 4	Model 5
1	2667	1227	684	429	226
0	1645	3085	3628	3883	4086

6.1.2 Logistic Regression

Table 3. Performance of the Logistic Regression models in terms of accuracy and AUC.

LOG	Accuracy	AUC
Model 1	0.65	0.65
Model 2	0.72	0.63
Model 3	0.84	0.55
Model 4	0.90	0.57
Model 5	0.95	0.53

In the Logistic Regression case, the first model performs best in terms of AUC. These results indicate that applying different classification rules with different dependent variable distributions does change classification precision in a model.

Logistic Regression results are given in table 4. These results indicate that network characteristics and firm specific characteristics both do play an important role. However, their effects depending on different trigger value can be different, for example Degree characteristic or the number adjacent edges has a negative effect in Model 1 and a positive effect in Models 3 and 4.

Table 4. Logistic Regression models'.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Network characteristics					
The number of adjacent edges.	-0.1245****	-0.0003	0.0519***	0.0562***	not included
Number of communities	0.1098***	-0.036	-0.0264	0.0394	not included
The graph level centrality index	0.5692***	0.0852	-0.0417	0.0234	-2.2043****
Nodes_number	-0.0003	-0.0226***	-0.0098	-0.0254	not included
Edge_number	-0.003	0.0171***	0.0084	0.0164	not included
The ratio of the number of edges and the number of possible edges.	not included	-1.1902***	not included	-1.6507	-0.9052
The values of the first eigenvector of the graph adjacency matrix	0.3081	-0.2358	-0.8973**	-0.184***	0.0994
Firm characteristics					
Number of employees	not included	not included	-0.0676****	-0.0519****	not included

EVRK_section_B	not included	-1.6525***	not included	-1.9442	-1.804
EVRK_section_C	0.2899	-1.1263***	-0.8176***	-1.0657	-1.6138****
EVRK_section_F	1.0229****	-0.6829	-0.8858***	-1.1491	-1.5913****
EVRK_section_G	-0.1587	-1.6061****	-1.0628***	-1.1553	not included
EVRK_section_H	-0.1753	-1.2503***	-0.8843***	-1.2431	-1.0661****
EVRK_section_I	0.2346	-1.3297***	-1.0329***	-0.7918	not included
EVRK_section_L	0.0991	-0.4473	-0.4416	-0.5227	-0.9817****
EVRK_section_N	-0.4138	-0.7631	-0.4173	-0.4496	-0.9038***
EVRK_section_P	-0.4988	-1.1086***	-0.2935	-0.2598	not included
EVRK_section_Q	-1.6034****	-1.3811***	-0.5636	-0.9481	-1.1493***
EVRK_section_S	0.3953	-0.6291	-0.5328	-0.9828	-1.6304***

**p-value < 0.05; ***p-value < 0.01; ****p-value < 0.001

6.1.3 Decision trees

Table 5 provides information about performance of the Decision T ree models in terms of accuracy and AUC.

Table 5. Performance of the Decision Tree models in terms of accuracy and AUC.

Decision Tree	Accuracy	AUC
Model 1	0.65	0.64
Model 2	0.70	0.62
Model 3	0.81	0.70
Model 4	0.88	0.61
Model 5	0.95	0.52

Comparing Decision Tree models' results with Logistic Regressions' results it can be stated that Decision Tree models do not universally outperform Logistic Regression models in terms of AUC, however in some cases they do – there is a clear indicator that non-linear effects might exist.

6.1.4 Random Forest

Random Forest results are presented in table 6.

Table 6. Performance of the Random Forest models in terms of accuracy and AUC.

Random Forest	Accuracy	AUC
Model 1	0.68	0.68
Model 2	0.73	0.69
Model 3	0.83	0.74
Model 4	0.90	0.71
Model 5	0.94	0.54

Random Forest clearly outperforms the Logistic Regression models, all five AUC scores are higher in Random Forest models compared with Logistic Regression models. Therefore, it can be stated that – there is a clear indicator that non-linear effects exist.

Regarding the variable importance in random forest models it's given in Table 7.

Table 7. 10 most important features in each Random Forest model.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Network characteristics					
The values of the first eigenvector of the graph adjacency matrix	1	1	2	4	3
Kleinberg's hub centrality score	2	2	4	2	4
The graph level centrality index	3	4	5	5	1
Shortest (directed or undirected) paths between vertices	4	5	6	6	Not included
Modularity	5	6	7	7	Not included
Edge_number	6	8	8	9	Not included
The number of adjacent edges.	7	10	10	10	Not included
Nodes_number	8	9	9	Not included	Not included
Diameter	9	Not included	Not included	Not included	Not included
Number of communities	10	Not included	Not included	Not included	Not included
Access to every other vertex from a given vertex	Not included	3	3	3	Not included
The ratio of the number of edges and the number of possible edges.	Not included	7	Not included	8	2
Firm characteristics					
Number of employees	Not included	Not included	1	1	Not included
EVRK_section_L	Not included	Not included	Not included	Not included	5
EVRK_section_H	Not included	Not included	Not included	Not included	6
EVRK_section_N	Not included	Not included	Not included	Not included	7
EVRK_section_F	Not included	Not included	Not included	Not included	8
EVRK_section_C	Not included	Not included	Not included	Not included	9
EVRK_section_Q	Not included	Not included	Not included	Not included	10

The most interesting feature of this table is the fact that in all but one model, Model 5, network characteristics clearly dominate firm specific characteristics.

6.2 Discussion

It can be stated that in terms of hypothesis this thesis rejected second, third and fourth hypothesis there were no universal impact of Network size (Nodes number), Communities in a network (Number of communities) and Individual connections (The number of adjacent edges) in all of examined models. However, this thesis failed to reject First and Fifth hypothesis, this shows that indeed network characteristics and firm specific characteristics are important when classifying a potentially tax evading firm.

Regarding the precision of all examined models, Random Forest models produced best results in all examined models. If these results would persist with STI's actual trigger rules Random Forest model with AUC score around 0.7 could be useful when selecting potential auditing targets.

Regarding overall achieved precision in all the models, it should be noted that there is a possibility that there are a lot a noise in this data which cannot be checked due to the fact that this data set is anonymized. Also, the fact that this thesis used incomplete firm specific characteristics data could have hindered models' performance as well. These missing firm specific characteristics – firm age, expenses and background of its shareholders could have increased model precision and helped in this classification task. Therefore, this thesis showed that indeed network characteristics and firm specific characteristics are useful in tax evasion classification tasks, but however, to achieve better results more diverse data is needed.

7. Conclusions and further research

The purpose of this master thesis was to examine network characteristics' influence on tax evading companies and to look if classification of companies as tax evaders was possible using network characteristics. This thesis used STI data from 2014 to 2019. There were 4312 observations with group characteristics in total. There were large differences between analyzed firms and their group characteristics. Groups were constructed based on ownership and kinship characteristics (relationships) between persons. Each group's network characteristics were extracted by yearly basis. The following conclusions were made :

1. This thesis rejected second, third and fourth hypothesis there were no universal impact of Network size (Nodes number), Communities in a network (Number of communities) and Individual connections (The number of adjacent edges) in all of examined models.
2. This thesis failed to reject First and Fifth hypothesis, this shows that indeed network characteristics and firm specific characteristics are important when classifying a potentially tax evading firms.
3. Random Forest models produced best results in all examined models. If these results would persist with STI's actual trigger rules Random Forest model with AUC score around 0.7 could be useful when selecting potential auditing targets.
4. This thesis showed that indeed network characteristics and firm specific characteristics are useful in tax evasion classification tasks, but however, to achieve better results more diverse data is needed.

In order to improve these results future research should examine these factors :

- Different group design – connecting groups by different social relationships.
- Examine if relationships persists in every economic sector, this dataset was too small for this purpose.
- Add and examine more firm specific characteristics.
- Add spatial dimension, hypothesis could be made that social links or groups can be different in different cities.
- Use actual STI's trigger rules.
- Do this analysis on pre-validated dataset.
- Compare results from different countries.

References

1. Abdixhiku, L., Krasniqi, B., Pugh, G., & Hashi, I. (2017). Firm-level determinants of tax evasion in transition economies. Economic Systems, 41(3), 354-366.
2. Alm, J., Liu, Y., & Zhang, K. (2019). Financial constraints and firm tax evasion. International Tax and Public Finance, 26(1), 71-102.
3. Artavanis, N., Morse, A., & Tsoutsoura, M. (2015). Tax evasion across industries: soft credit evidence from Greece (No. w21552). National Bureau of Economic Research.
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. Physics Reports, 424(4-5), 175–308. doi:10.1016/j.physrep.2005.10.009
5. Boning, W. C., Guyton, J., Hodge, R., & Slemrod, J. (2020). Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms. Journal of Public Economics, 190, 104261.
6. Bott, K. M., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2020). You've got mail: A randomized field experiment on tax evasion. *Management Science*, *66*(7), 2801-2819.
7. Bramoullé, Y., Kranton, R., & D'amours, M. (2014). Strategic interaction and networks. American Economic Review, 104(3), 898-930.
8. Chiesi, A. M. (2015). Network Analysis. International Encyclopedia of the Social & Behavioral Sciences, 518–523. doi:10.1016/b978-0-08-097086-8.73055-8
9. degl'Innocenti, D. G., & Rablen, M. D. (2020). Tax evasion on a social network. Journal of Economic Behavior & Organization, 169, 79-91.
10. Dempwolf, C. S., & Lyles, L. W. (2012). The uses of social network analysis in planning: A review of the literature. Journal of Planning Literature, 27(1), 3-21.
11. European Commission, Directorate-General for Taxation and Customs Union, Poniatowski, G., Bonch-Osmolovskiy, M., Śmietanka, A. (2021). VAT gap in the EU : report 2021, Publications Office. https://data.europa.eu/doi/10.2778/447556
12. Gangl, K., Torgler, B., Kirchler, E., & Hofmann, E. (2014). Effects of supervision on tax compliance: Evidence from a field experiment in Austria. Economics Letters, 123(3), 378-382.

13. Harju, J., Kosonen, T., & Ropponen, O. (2014, January). Do honest hairdressers get a haircut?. In Proceedings. Annual Conference on Taxation

14. Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

15. igraph R manual pages. (2021). Igraph.Org. https://igraph.org/r/html/latest/

16. Infolex. (2021).Pridėtinės Vertės mokesčio įstatymo komentaras. INFOLEX.LT . https://www.infolex.lt/teise/DocumentSinglePart.aspx?AktoId=14401&StrNr=2 Accesed Dec 1,2021.

17. Internal Revenue Service (2017). "Tax Audits: Triggers and Tips," Page 5. Accessed Nov. 1, 2021.

18. Joulfaian, D. (2000). Corporate income tax evasion and managerial preferences. Review of Economics and Statistics, 82(4), 698-701.

19. STI (2021). Kada juridiniam asmeniui atsiranda prievolė registruotis PVM mokėtoju? - www.vmi.lt. https://www.vmi.lt/evmi/kada-privaloma-registruotis-pvm-mok%C4%97toju- Accessed Dec. 1, 2021.

20. Kirchler, E., Maciejovsky, B., & Schneider, F. (2003). Everyday representations of tax avoidance, tax evasion, and tax flight: Do legal differences matter?. Journal of Economic Psychology, 24(4), 535-553.

21. Lismont, J., Cardinaels, E., Bruynseels, L., De Groote, S., Baesens, B., Lemahieu, W., & Vanthienen, J. (2018). Predicting tax avoidance by means of social network analytics. Decision Support Systems, 108, 13-24.

22. Marin, A., & Wellman, B. (2011). Social network analysis: An introduction. The SAGE handbook of social network analysis, 11, 25.

23. Michael Hallsworth, The use of field experiments to increase tax compliance, *Oxford Review of Economic Policy*, Volume 30, Issue 4, WINTER 2014, Pages 658–679, https://doi.org/10.1093/oxrep/gru034

24. Morse, S. C., Karlinsky, S., & Bankman, J. (2009). Cash businesses and tax evasion. Stan. L. & Pol'y Rev., 20, 37.

25. Muschelli J. (2020). ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. Journal of classification, 37(3), 696–708. https://doi.org/10.1007/s00357-019-09345-1

pp. 1-32). National Tax Association.

26. Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. Journal of Economic perspectives, 21(1), 25-48.

27. Tedds, L. M. (2010). Keeping it off the books: an empirical investigation of firms that engage in tax evasion. Applied Economics, 42(19), 2459-2473.

28. Wang, Y. (2009). Competition and tax evasion: a cross country study.

29. Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.

Appendix

Network characteristics

- The number of its adjacent edges.
- Number of vertices of a graph
- The length of the longest geodesic
- How many steps is required to access every other vertex from a given vertex.
- The graph level centrality index.
- The ratio of the number of edges and the number of possible edges.
- Number of edges in a graph
- The values of the first eigenvector of the graph adjacency matrix
- the number of geodesics (shortest paths) going through a vertex or an edge.
- The node-level centrality scores.
- Kleinberg's hub centrality scores.
- Kleinberg's authority centrality scores.
- Shortest (directed or undirected) paths between vertices
- Many networks consist of modules which are densely connected themselves but sparsely connected to other modules.
- Number of communities
- Modularity of the graph partitioning