Differentiation detection in insurance pricing models

B.J.E. van Walree

Master's thesis Business Analytics

Differentiation detection in insurance pricing models

Bernt van Walree (2577277)

Master's thesis

Vrije Universiteit Amsterdam
Faculty of Science
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

Host organization:
Aegon nv
Aegonplein 50
2591 TV Den Haag

Supervisors:
d.r. B.T. Szabo (first supervisor)
d.r. R. Bekker (second supervisor)
d.r. R.J.D. Potter van Loon (supervisor Aegon)

July 31, 2021

Preface

This thesis is written as the conclusion of the Master's degree in Business Analytics at the Vrije Universiteit Amsterdam. Business Analytics is a multidisciplinary program revolving around mathematics, computer science and economics. The program is concluded by following an internship at a company. The host company presents a problem they are dealing with, to which the student must find a solution. For at least six months, the student works at the company on the problem they are dealing with. In addition, a report on the internship is written. The internship report serves as the master's thesis of the student.

Artificial intelligence (AI) is increasingly used in multiple business areas. As more companies use AI, the concerns about possible differentiation in AI are also increasing. This research revolves around the creation of a tool that can detect and measure possible differentiation of insurance pricing models.

This internship has taken place at Aegon at the Analytics & Datascience (A&D) department. The A&D department performs various tasks related to data science for the different departments of Aegon. In the department, I was assigned to the DINNO team. This team is responsible for handling compliance and risk issues, concerning data science.

First of all, I am grateful to Aegon and the DINNO team for giving me the opportunity to write my thesis there and for helping me with this research. I especially would like to thank Rogier, my supervisor at Aegon, for his help and advice. Furthermore, I would like to thank my supervisors of the university. I would like to express my gratitude to Botond for being my first supervisor and providing me with different aspects to look into in this research. Finally, I also would like to thank René for being the second reader of this thesis.

Abstract

This thesis describes the creation of a tool that should detect any kind of bias or differentiation in insurance pricing models used by Aegon. AI has become widely adopted in many business fields. One of these fields is insurance. With AI becoming more widely adopted, concerns about this technology also grow. Aegon wants a tool that can detect possible differentiation in the models they use. This way, Aegon can explain to customers and experts how their models behave and they can also show whether their models are differentiating. Furthermore, Aegon can use it to determine if a model works as expected. The tool measures differentiation for every variable in a given data set used by a given model. It can decompose differentiation into direct differentiation and indirect differentiation. This can be further dissected into categories if the feature of interest is categorical. The tool can also show the flows of differentiation through the other variables in the case of indirect differentiation. Other tools have been made by companies and scientists in the last couple of years. However, the tool created in this study is easier to use and gives more narrowed-down information to the user. The tool is tested on six models. Four of these models are created in a simulation. The other two models are created by Aegon and serve as examples of the practicality of this tool. The tests show that differentiation or bias can be found. Both direct and indirect differentiation is found. However, the tool has problems with step functions and it can pick up ”noise differentiation” due to random correlation. In addition, causality is implied in indirect differentiation, which does not always hold. Still, the tests on the Aegon models do show that the tool can be of use in the industry.

keywords: AI, Aegon, Pricing models, Insurance, differentiation, Tool

Contents

1 Introduction

In recent years, Artificial Intelligence (AI) has become widely adopted in a lot of industries [21]. Companies use AI to solve complex and, otherwise, time-consuming issues. Instead of humans performing manual work, this new technology can automatize a lot of operations nowadays. The automation increases efficiency and reduces costs in the applied fields. This makes AI very valuable and explains its growing use.

One of the fields in which AI is applied is insurances [11]. There are several ways in which insurance companies apply the technology. An example is to predict insurance premia with machine learning models [15]. However, there are also downsides to the use of machine learning methods in the pricing process. Using AI can lead to optimization through personalizing prices, but it can also lead to biased and differentiating pricing [7]. The main objective of this research is to find ways to measure and detect possible bias or differentiating in pricing models used by insurance companies.

There are different causes for the possible differentiating in the algorithmic predictions. First of all, the creator of a model that makes the predictions can be biased. Second, the data itself may be biased. The models used are fed with lots of data. This data can be biased which can lead to biased model outcomes. A third possible downside is the algorithm with which the model is created. If the goal of the algorithm is to achieve an accuracy as high as possible, it can amplify certain biases in the data to achieve that goal [9].

The mentioned sources of bias are hard or maybe even impossible to detect by humans. The pricing models used by insurance companies are not simple functions where the outcome can easily be derived and explained. They are rather black-boxes. Even the creators of the models can hardly explain the behavior of their creations. Data is given as input to the model and it produces an outcome. However, how the outcome came to be is unknown. The hidden behavior of the model means that the causes of possible differentiating are hard to detect and to prevent. It must be ensured that the technology used by insurance companies is not differentiating.

An important question to ask in this research is: what is differentiating? The Oxford dictionary gives the following definition for differentiating: ”The practice of treating someone or a particular group in society less fairly than others” [23]. Insurers use groupings of people to determine insurance premia [2]. An insurance company treats some groups of people differently, or unfairly, compared to other groups of people. So, insurance companies differentiate based on certain features and they are allowed to do so. However, they are not allowed to differentiate or group people based on every characteristic. Therefore, it is important to know which features are used by a pricing model and how important these features are.

A possible solution for stopping differentiating in the pricing process is to take protected features out of the data set used by the pricing model. A protected feature (or variable) is a feature on which it is illegal to differentiate against people. Unfortunately, it is not a definitive solution to take the protected variable out of the data set. The variable can still influence the model predictions indirectly, through another variable [18]. In this case, the second variable is dependent on the protected variable.

This is in line with the legal definition of differentiating. differentiating consists of direct differentiating and indirect differentiating [32]. Direct differentiating occurs when, for example, two nearly

identical persons pay different insurance premia. Their only difference is their gender. There is a direct way in which people of a certain gender are treated differently. An example of indirect differentiating is the following. Men pay a higher car insurance premium than women since on average men drive more kilometers with their car. Directly, people who drive a lot are treated differently compared to people who drive less. However, in an indirect manner men are affected more than women. That is because people who drive a lot are more often men. This distinction between direct differentiating and indirect differentiating is also important because direct differentiating on a protected feature like gender is illegal. However, indirect differentiating on a protected feature may be legal. This form of differentiating is legal if the differentiating can be justified.

The goal of this research is to create a tool that can detect possible differentiating of insurance pricing models. The commissioner of this research project is Aegon. Aegon is a Dutch multinational with activities in several financial areas. One of these areas is insurance. Aegon uses various AI applications to calculate insurance premia for customers. The insurance company wants a differentiating detection tool that measures the effect a variable has on the outcome of a pricing model. With this tool the company can explain to clients and experts how Aegon determines insurance premia; is there any differentiating going on? If there is differentiating going on Aegon wants to be able to tackle this problem. Therefore, it is important to know what the detected differentiating consists of. Not only must the tool be able to detect differentiating, but it must also be able to measure it and decompose it into direct and indirect differentiating. Furthermore, if the tool works appropriately, it will be shared with the Verbond van Verzekeraars. The Verbond van Verzekeraars is an interest association of Dutch insurance companies. The Verbond van Verzekeraars will share the tool with other interested insurance companies.

The tool consists of two components; programming code which can be used to detect and measure differentiating and an app that can visualize the results of the code. This thesis will mostly describe the differentiating detection and measuring methods used in the code. The tool is coded in the programming language R.

1.1 Research Questions

To summarize, this thesis revolves around the creation of a tool that can measure and detect differentiating of models. Therefore, this research focuses on the following problems:

Q1. Is it possible to detect differentiating in insurance pricing models?

Q2. Is it possible to measure differentiating in insurance pricing models?

1.2 Thesis Outline

The thesis is structured as follows. In section 2, the related literature is treated. Papers that describe methods to detect differentiating are discussed, but also similar already existing tools are researched. Section 3 describes the data sets used by the models on which the tool is tested. Multiple data sets are used as each model is of course trained on its own data set. Section 4 describes differentiating in the context of this research. The definitions stated in this introduction are used, but these definitions have to be translated to a mathematical setting. In section 5, the basis methodology of differentiating detection and measuring is explained. The algorithms and methods used by the tool are given in section 6. Experiments are conducted to see if the methodology works as expected. Simulation

models are created on which the tool is tested. The test setup is described in section 7. The results of the tests are given in section 8. The workflow of the tool is described in section 9. It is also explained how the app works that can be used to analyze the results of the tool. Lastly, section 10 brings the conclusion of this research, and suggestions for future work are given.

2 Related work

With the growing popularity of AI in daily life, the research area concerning the ethics of AI has also become more popular. Since humans are ever more in contact with AI there is a rising interest in the decision-making of the tools using this technology. Naturally, there will be more academics looking into this field too.

Researchers of Pennsylvania State University have already created a tool that can detect differentiation in algorithms [16]. Their research focuses on detecting differentiation against groups of people. In the paper, two fairness definitions are introduced. Fair on average causal effect (FACE) and fair on average-age causal effect on the treated (FACT). Both definitions look at the expected difference between two sets of model outcomes for sub-groups in a variable.

In "Hunting for Discriminatory Proxies in Linear Regression Models" [31] the authors look at potential differentiation of linear regression models through proxy variables. A proxy variable is a variable that is causally influential on the model's output and correlated with the protected variable. These proxies are identified by solving a second-order cone program [1].

Lu Zhang et al., [32] try to remove differentiating data from data sets in their paper. To remove differentiation, differentiation first has to be detected. This is done by making use of the causal network to model the causal structure of the data. Furthermore, the researchers make the distinction between direct and indirect differentiation. Both types of differentiation are modeled as path-specific effects in the causal network. Lindholm et al., [18] use the same definitions to construct a differentiation-free insurance pricing model.

The topic of this thesis is closely related to explainable AI (XAI) [25]. XAI aims to make opaque models easily interpretable. The two most popular XAI techniques related to this research are LIME and SHAP [4],[17]. SHAP is based on Shapley values [26]. It aims to explain individual predictions through feature relevance [19]. This is done by comparing model outcomes with and without a variable. LIME creates local surrogate models to explain predictions individually. It tests what happens to the predictions when you give variations of input data to a model. An interpretable model is trained on the resulting output data.

"Big Tech" has also entered the field of AI-fairness. Companies like Google (Fairness Indicators) [30], Amazon (Clarify) [14], Microsoft (Fairlearn) [5] and IBM (AIF360) [3] have created tools which can detect differentiation and remove it. Amazon Clarify uses Shapley values to detect differentiation. Fairness Indicators and Fairlearn compare model quality metrics, like model accuracy and the number of false positives, for the model outcomes of different groupings. An example is, how accurate are the model predictions for men and how accurate are the model predictions for women. AIF360 compares the number of favorable results of different groups of people. The user must define what favorable results are. An example here is to compare the number of favorable results for men to the number of favorable results for women.

This research uses the same formal definitions of direct and indirect differentiation as given in [32] and [18]. differentiation, and hence differentiation detection, is defined in another way. In this research, differentiation is measured by comparing the regular model outcome to the model outcome when a variable in the input data has changed. This is similar to LIME. However, LIME is only able to explain

classifying models. The methods used by the tool of this thesis work for any kind of model. What this tool further adds, compared to earlier work, is the amount of information that is returned to the user. differentiation is measured regarding every variable in the data set. differentiation measurements can not only be decomposed into direct and indirect differentiation but each differentiation measurement can be further decomposed into categories if the variable of interest is categorical. In addition, the tool can show through which variables the indirect differentiation flows. Moreover, differentiation can also be measured for numeric variables. This is something a lot of the current tools have difficulties with. The tool described in this research is easier to use than its competitors. The user only needs to give its model as input and a data set. The tools of the big tech companies require a lot of data and/ or model preparation before they can be used. The output these tools give is also unclear and not easy to understand. The differentiation detection tool created in this internship contains an app that can be used to visualize the results. The app makes it easy to understand the results and to share the results with others. Only Amazon provides an easy-to-use app.

3 Data

The differentiation detection methods used in the tool are tested on six models. Four of those models are simulation models which use the same data set. The other two models are Aegon models which use the same base data set. The first Aegon model is an XGBoost model and the other model is an insurance model created by Aegon themselves. Both Aegon models predict insurance premia for customers. They use the same base data set, however, the XGBoost model can not use categorical data. Therefore, the base data set is altered for the XGBoost model. Categorical variables are turned into multiple binary variables. Each binary variable corresponds to a category. This alteration is already performed by Aegon. So, this will not be explained further in this section. The data of the Aegon models is also enriched by data of the Centraal Bureau voor de Statistiek (CBS). In this section, the preparation of the data sets is described and the data sets are explored. Three data sets are explored; the simulation data set, the base Aegon data set and the CBS data set.

3.1 Preparation

The data provided by Aegon is synthesized using GAN's [12]. Synthetic data is data generated by a model based on an actual data set. Anonymizing data has the problem of either deleting important variables or otherwise not being actually anonymous. This is not a problem with synthesizing data. Furthermore, synthesized data can accurately resemble the original data. The data synthesizing was done by mostly.ai [20]. The data had to be synthesized to comply with GDPR [8].

The base data set of the Aegon models is provided by Aegon. This data set is enriched with data of CBS [28]. CBS is the Dutch institute that collects and publishes data on Dutch society and does research on the collected data. By enriching the data sets, more features are incorporated. This enriching is important since insurance companies do not collect certain sensitive data of customers. This includes features like ethnic background. However, insurance companies can still be differentiating based on these features without knowing. As explained in the introduction. Models can be differentiate on features it does not know indirectly. So, more possible differentiation can be detected by enriching data with CBS data. The two data sets are joined on the postal code variable.

Regarding the CBS data set, the location variable first has to be converted to postal codes. The location variable in the CBS data set is denoted in units of Rijks-Driehoek (RD). This is a Dutch-specific location unit. With the library of Simple Features (SF) [22] this unit can be converted to Dutch postal codes. The CBS data set consists of aggregated data of the residents per postal code. Due to privacy reasons, data can be missing. Every line where data is missing is dropped. This is done because it doesn't make sense to impute values at the missing spots, as these values are postal code specific; different groups of people live at different postal codes. If values were imputed in the places of missing data this could lead to misleading differentiation detection results. Furthermore, only the variables of interest for Aegon are chosen to work with further. Every residential variable is turned into percentages of the number of residents living on a postal code.

More preparation is undertaken for each data set. First, each variable in the data set is classified as either a numeric variable or a categorical variable. Next, it is checked if there are any missing values or error values in the data set. If such a value is detected a new value is imputed. If the variable of interest is numeric, the median value of the variable is imputed. If the variable of interest is categorical, the most common value of the variable is imputed.

3.2 Exploration

3.2.1 CBS data

The data set used from the CBS is the ”Statistiche gegevens per vierkant 2020” file. The CBS carves up the whole Netherlands into squares of 100 x 100 meters. The institute collects data on every resident inside these squares and aggregates this data. The features of the data set are given in Table 16 in appendix B.1. Not every variable is used to enrich the other data sets due to unimportance for the users. Only the features displayed in Table 1 are used. The data consists of 132060 rows.

Table 1: selection CBS data

Feature	Type	Description
Man	Numeric	The percentage of men
Vrouw	Numeric	The percentage of women
INW_014	Numeric	The percentage of residents aged -14
INW_1524	Numeric	The percentage of residents aged 15-24
INW_2544	Numeric	The percentage of residents aged 25-44
INW_4564	Numeric	The percentage of residents aged 45-64
INW_65PL	Numeric	The percentage of residents aged 64+
P_NL_ACHTG	Numeric	The percentage of residents with Dutch native background
P_WE_MIG_A	Numeric	The percentage of residents with western background
P_NW_MIG_A	Numeric	The percentage of residents with non-western background
P_KOOPWON	Numeric	The percentage of owner-occupied houses
P_HUURWON	Numeric	The percentage of rental houses
WOZWONING	Numeric	The average value of the properties
UITKMINAOW	Numeric	The percentage of residents with welfare
Postcode	Category	The postal code

Next, the data is explored. The graphs are found in appendix B.1. The percentage of men and women seem to be evenly distributed. Looking at Figure 7, the population seems to be evenly distributed with regards to gender. The older age groups seem to be better represented than the age groups below 25. Figure 8 shows that there are more rental houses than owner-occupied houses. According to Figure 9, there are more people with a native Dutch background than people with a foreign background, which is not surprising of course.

3.2.2 Simulation model data

The data is simulated. 14 features are created. The numeric variables are generated under a normal distribution and the categorical variables are generated completely random. For some features dependencies are built in to simulate interactions. The data set used in the simulation models contains the following features (Table 2). The data consists of 100000 rows.

Table 2: Simulation data

Feature	Type
Age	Numeric
Ethnicity	Category
Education	Category
Income	Numeric
Gender	Category
Income category	Category
House value	Numeric
Family size	Category
Random variable 1	Numeric
Random variable 2	Numeric
Correlated variable 1	Numeric
Correlated variable 2	Numeric
Correlated variable 3	Numeric
Correlated variable 4	Numeric

The distributions of the numeric variables and the proportions of the categorical variables are given in Table 17 and Table 18 in appendix B.2. In this data set, there are several dependencies or correlations. House value and income category are dependent on income. Correlated variable 1 is dependent on age and ethnicity. Correlated variable 2 is dependent on ethnicity. Correlated variable 3 is dependent on gender. Correlated variable 4 is dependent on age. The correlations and dependencies are given in appendix B.2.

The visualizations of the exploration are also found in appendix B.2. According to Figure 10, there are more men than women. Dutch is the most dominant ethnic group and the other ethnicities seem to have around even percentages. Education seems quite evenly distributed except for category 3. The lower-income groups are better represented and there are more single people than big families. Figure 11 shows the box plots of the numeric variables. The numeric data seem to have nice even distributions except for income and house value. There are quite a lot of heavy extreme values in these variables.

4 Differentiation in this research

In the introduction, it was explained that differentiation consists of direct differentiation and indirect differentiation. This means that a model can differentiate on a variable in two ways. Or rather, the variable can influence the outcome of a model in two ways. With this in mind, the following question can be formulated. How much does the outcome of the model change when the input variable changes? Through this notion, the following definitions for differentiation can be formulated.

(1) differentiation in this research is defined as the change in the model outcome due to a change in the variable of interest.

(2) Direct differentiation is defined as the change in the model outcome due to a change in the variable of interest, while all other variables stay constant.

(3) Indirect differentiation is defined as the change in the model outcome due to a change in a second variable caused by a change in the variable of interest.

So, the goals in this research are the following:

- Measure the influence the variable of interest has on the model outcome while the other variables stay constant.
- Measure the influence the variable of interest has on the model outcome through another variable.

Consider the following simple example:

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * z_i + \varepsilon_{i,1}$$

$$z_i = \gamma_0 + \gamma_1 * x_i + \varepsilon_{i,2},$$

$$\varepsilon_{i,1} \sim N(0,1),\ \varepsilon_{i,2} \sim N(0,1),\ i = 1,2,..,n.$$

y_i is the model outcome that is dependent on the variable of interest x_i and another variable z_i. It is quite easy to measure the effect each variable has on the model outcome and thus it is easy to measure differentiation. In this case, β_1 is the measure of direct differentiation and $\beta_2 * \gamma_1$ is the measure of indirect differentiation. The following figure depicts this relationship.

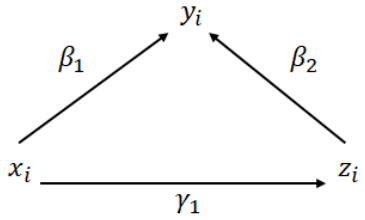

Figure 1: The relationship between a model and its input variables

Figure 1 shows how the model outcome is dependent on x_i and z_i. It also shows the dependency of z_i on x_i. x_i can directly influence the model outcome and it can also indirectly influence the model

outcome through z_i. Indirect differentiation is measured by multiplying the effect of x_i on z_i with the effect of z_i on y_i. Total differentiation is calculated by adding up direct differentiation and indirect differentiation.

An important aspect to explain is how differentiation is measured for a variable that is not in the model training data set. This corresponds to variables in the CBS data set. The model does not observe the variables in the CBS data set, but it can still be differentiating on these features.

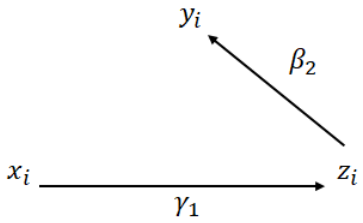

Figure 2: The relationship between a model and an unobserved variable and an input variable

Figure 2 depicts the situation where the variable of interest x_i is not observed by the model. y_i is not dependent on x_i in this situation. There is no way in which x_i can directly affect the model outcome. However, it can affect another variable in the data set z_i which can influence the model outcome y_i. So, there is no direct differentiation measurement possible for features outside the model training data set, but indirect differentiation can be found.

5 Methodology

This chapter explains the methodology behind the tool. First, some general methodology will be explained. Second, further methodology used by the detection algorithms are described.

5.1 General methodology

In this section, an example will be used to explain the methodology. This example is in the form of a linear model.

$$y_i = f(x_i, z_i) + \varepsilon_{i,1}, \tag{1}$$

$$\varepsilon_{i,1} \sim N(0,1),\ i = 1,2...,N.$$

Furthermore,

$$z_i = g(x_i) + \varepsilon_{i,2}, \tag{2}$$

$$\varepsilon_{i,2} \sim N(0,1),\ i = 1,2...,N.$$

$f(x_i, z_i)$ and $g(x_i)$ are linear functions in this particular example. i represents the index of an observation. Here y_i is the outcome of a model. x_i is the variable of interest for which differentiation is to be measured. z_i is another variable in the data set through which differentiation can flow from x_i, as shown in Figure 1. In this model z_i is dependent on x_i. $g(x_i)$ describes this relationship.

In practice, a lot is unknown about a model. A model is trained on a certain data set and this model makes use of a prediction algorithm to produce outcomes. The model owner knows on which variables the model outcome is dependent in a direct sense. Those variables are the input variables. However, how important these variables are in the prediction is unknown. Furthermore, it is also unknown if there are any interactions between the input variables. So, $g(x_i)$ is unknown. This means that possible indirect differentiation is unknown. This does not matter for the differentiation measuring methods. The methods use $f(.)$, which is the function/ algorithm with which the model makes predictions, and $\hat{g}(x_i)$, which is the estimation of $g(x_i)$, to measure direct and indirect differentiation. Models 1 and 2 are used to illustrate how the differentiation detection methods work in simple terms. However, in practice, these models are opaque and difficult to interpret.

differentiation is measured for each variable in the data set. In the following sections, only the differentiation measurement for one variable (x_i) is considered. However, the tool created in this research will measure differentiation for every variable. So, there is no difference in importance between x_i and z_i. These notations are only used in this thesis to distinguish between the variable for which differentiation is measured (x_i) and one other variable in the data set (z_i). Usually, a model is trained on more than two variables. $f(.)$ can be dependent on 10 variables for example. differentiation is measured for each of these 10 variables.

The research makes a distinction between direct differentiation and indirect differentiation. Direct differentiation is the direct effect of x_i on y_i. Indirect differentiation is the indirect effect of x_i on y_i, through z_i. When there are more than two variables in the data set the indirect differentiation of x_i is measured by summing the indirect differentiation measurements of x_i through each of the other variables. For example, a data set could harbor the following variables: x_i, $z_{i,1}$, $z_{i,2}$ and $z_{i,3}$. Indirect differentiation on x_i is measured through each of $z_{i,1}$, $z_{i,2}$ and $z_{i,3}$. These measurements are added

together to get the ”total” indirect differentiation measurement of x_i.

Each variable is defined as either numeric or categorical. The methods to measure and detect differentiation are defined differently for either type of variable. In numeric data, the values have an intrinsic meaning. The values in a categorical variable represent a group to which the observation belongs. For numeric data, differentiation is measured by how much the outcome of the model changes when the variable of interest changes with an (infinitely small) step. For categorical data, differentiation is measured by looking at the difference in model outcomes for the different categories of the variable. First, the methodology behind differentiation detection in numerical data is described in 5.2. and 5.3. Next, the methodology behind differentiation detection in categorical data is described in 5.4.

5.2 Derivative & difference quotient

The derivative and the difference quotient play an important role in the differentiation detection methods. The differentiation detection methods measure differentiation by comparing the regular model outcome, or the fitted values of a model, to the model outcome when the variable of interest has changed with a certain (infinitely small) step. This effect on the model outcome due to a change in a variable can be calculated by the derivative and the difference quotient.

The calculations are fairly easy to compute. Consider the given model 1. When x_i is a continuous numeric variable, the derivative of $f(x_i, z_i)$ with respect to x_i, for observation i, is calculated as

$$D_1 f(x_i, z_i) = \frac{df(x,z_i)}{dx}|_{x=x_i} = lim_{h->0}\frac{f(x_i+h,z_i)-f(x_i,z_i)}{h}.$$

When x_i is a discrete numeric variable, the forward difference quotient of $f(x_i, z_i)$ with respect to x_i, for observation i, is calculated as

$$D_1 f(x_i, z_i) = \frac{df(x,z_i)}{dx}|_{x=x_i} = \frac{f(x_{i+1},z_i)-f(x_i,z_i)}{x_{i+1}-x_i}.$$

Similarly, the derivative or difference quotient of $f(x_i, z_i)$ with respect to z_i, for observation i, can be calculated as either

$$D_2 f(x_i, z_i) = \frac{df(x_i,z)}{dz}|_{z=z_i} = lim_{h->0}\frac{f(x_i,z_i+h)-f(x_i,z_i)}{h},$$

or

$$D_2 f(x_i, z_i) = \frac{df(x_i,z)}{dz}|_{z=z_i} = \frac{f(x_i,z_{i+1})-f(x_i,z_i)}{z_{i+1}-z_i}.$$

$D_1 f(x_i, z_i)$ and $D_2 f(x_i, z_i)$ are the direct differentiation measurements of x_i and z_i. The effect of x_i on z_i can also be measured by the derivative or difference quotient of $g(x_i)$ with respect to x_i.

$$D_1 g(x_i) = \frac{dg(x)}{dx}|_{x=x_i} = lim_{h->0}\frac{g(x_i+h)-f(x_i)}{h},$$

or

$$D_1 g(x_i) = \frac{dg(x)}{dx}|_{x=x_i} = \frac{g(x_{i+1})-g(x_i)}{x_{i+1}-x_i}.$$

5.3 Regression

To measure the indirect differentiation, the relationship between z_i and x_i has to be estimated. i.e. The function $g(x_i)$ has to be estimated. This can be done by fitting a model to the data. Most of the models which are tested in the detection tool are expected to be machine learning models, which can be classified as non-linear. Furthermore, the tool does not know what kind of model is given. Therefore, non-parametric regression estimation is used. A cubic-smoothing spline is used to estimate $g(x_i)$ by $\hat{g}(x_i)$. However, if the interquartile range (IQR) of x_i is 0, a linear model is fitted. The IQR is the difference between the 75th and 25th percentiles of a distribution. There is a tolerance check in the smoothing spline estimation that checks if x_i is varied enough. If the variability is too low, no estimation can be made by the smoothing spline. The tolerance is calculated as a multiplication of the IQR. So, if the IQR is 0, the smoothing spline can not estimate $g(x_i)$ [24]. Any other non-parametric regression estimator seems to have the same problem if the input data is not varied enough.

5.3.1 Linear Regression

A linear model is fit on the data to estimate the function

$$g(x_i) = E(z_i|x_i).$$

It is assumed that the relationship between z_i and x_i is linear. The estimation $\hat{g}(x_i)$ can be used to make predictions to estimate the indirect differentiation. $g(x_i)$ is estimated using the ordinary least squares (OLS) method. OLS minimizes the sums of the squared differences between the predicted data and the observed data, also known as residuals. In the regression function

$$\hat{g}(x_i) = \hat{\gamma}_0 + \hat{\gamma}_1 * x_i,$$

$\hat{\gamma}_0$ and $\hat{\gamma}_1$ are chosen such that the squared sum of the residuals is minimized. They are also known as the least squares estimators

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$\hat{\gamma}_0 = \bar{z} - \hat{\gamma}_1\bar{x}.$$

5.3.2 Non-parametric regression

$g(x_i)$ can be estimated by non-parametric regression estimators. There is no assumption made on the linearity of $g(x_i)$. The two most widely used methods are kernel regression and smoothing splines [13]. Spline and kernel smoothing are asymptotically the same. To every kernel operator, there is an equivalent spline operator. Both Kernel smoothing and spline smoothing have been compared in several ways [27]. Since according to the literature there is not much difference between the two non-parametric regression methods, both methods are tried.

First, kernel regression is explored. The two main kernel estimators are the Nadaraya-Watson (NW) kernel estimator

$$\hat{g}(x) = \sum_{i=1}^{n}(w_i(x)z_i),$$

$$\text{with } w_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n}K(\frac{x-x_j}{h})},$$

and the local linear (LL) estimator

$$\hat{g}(x) = argmin(\alpha)(\sum_{i=1}^{n}(K(\frac{x_i - x}{h})(z_i - \alpha)^2)) \text{ [13].}$$

There are several kernel functions K, but the Epanechnikov kernel

$$K(x) = \frac{3}{4}(1 - x^2)1(|x| \leq 1)$$

is optimal giving the smallest mean squared error [29]. Asymptotically the NW and LL estimators are similar, except for the bias function. In general, the LL estimator has a smaller bias than the NW estimator. In addition, the LL estimator has a better performance on non-constant $g(x)$. This estimator also performs better near the boundary of the support of x [13]. The most important parameter to be defined in kernel regression is the bandwidth variable h. The optimal value for h is usually obtained through cross-validation [29]. Cross-validation is a technique to test the quality of the predictions of a model. This is done by estimating the prediction error of the estimator. There are several R packages that use cross-validation to determine the optimal h. The two Kernel estimators are compared to the spline estimator.

The cubic smoothing spline is the most popular spline estimator [13]. The cubic smoothing spline $\hat{g}(x_i)$ is the minimizer of

$$\sum_{i=1}^{n}(z_i - \hat{g}(x_i))^2 + \lambda \int (\hat{g}''(x)^2)dx.$$

λ is the smoothing parameter. The optimal value for the smoothing parameter is obtained through cross-validation. R incorporates packages which use cross-validation to determine the optimal λ.

So, the literature shows that there is not much difference between a kernel estimator and a spline estimator. Both estimators are tried in this research. The optimal settings for the cubic smoothing spline are computed much faster than the optimal settings for the kernel estimator. Therefore, the cubic smoothing spline is used in the tool.

5.3.3 Cubic Smoothing spline

Consider Model 2. The cubic smoothing spline estimate $\hat{g}(x_i)$ of $g(x_i)$ is the minimizer of

$$\sum_{i=1}^{n}(z_i - \hat{g}(x_i))^2 + \lambda \int (\hat{g}''(x)^2)dx.$$

This minimizing consists of two parts; minimizing the squared error

$$\sum_{i=1}^{n}(z_i - \hat{g}(x_i))^2$$

and minimizing the curvature

$$\lambda \int (\hat{g}''(x)^2)dx.$$

The smoothing parameter λ controls the trade-off between accuracy and curvature. If $\lambda = 0$ the smoothing spline will be an interpolating spline. Every data point is visited and no accurate predictions can be made due to overfitting. If λ becomes infinitely big, the curvature becomes too important in the minimization. This will result in the smoothing spline becoming a linear least squares estimate. The optimal value of λ is determined through the generalized cross-validation method.

5.4 Measuring differentiation in categorical data

In the previous subsections, it was described how differentiation can be measured in numeric data. These methods can not be used for categorical data. Here, the values of a variable represent the group to which an observation belongs. Still, comparable methods to numeric data can be used here. Instead of calculating how much the model outcome changes when the protected variable changes by a small step, the differences between the model outcomes for every category of x_i is measured.

Define x_i as a categorical variable, such that $x_i \in C$. C consists of K categories such that $C = \{c_1, c_2, ..., c_K\}$. Then y_i can be rewritten as:

$$y_i = f(x_i, z_i) + \varepsilon_{i,1} = \beta_0 + \beta_1 * I_{x_i=c_1} + \beta_2 * I_{x_i=c_2} + ... + \beta_K * I_{x_i=c_K} + \beta_{k+1} * z_i + \varepsilon_{i,1}.$$

The methods for numeric data can not be used here as it does not make sense to compare $f(x_i)$ to $f(x_i + h)$. Instead, the predictions for each category are compared to one another. The predictions are not directly compared to one another as this comparison will differ based on which category is taken as a base category. Instead, they are compared to the weighted average of these predicted values. This can be written in the following way.

The direct effect of x_i on y_i, for observation i is measured by:

$$\sum_{k=1}^{K} w_k * |f(c_k, z_i) - \mu_i|,$$

$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k, z_i),$$

$$w_k = \frac{1}{N} \sum_{i=1}^{N} I_{x_i=c_k}.$$

w_k is the weight corresponding to category c_k. μ_i is the weighted average of the predictions at observation i. The difference is taken between the predictions where the value of x_i is set to c_k and the weighted average of these predictions. It is taken into account how often a category occurs in the data. This way a category with a high influence on the model outcome, but a low amount of occurrences does not have too much influence on the differentiation measurement of x_i.

In a similar way, the effect of x_i on z_i can be measured. Consider z_i to be redefined as

$$z_i = g(x_i) + \varepsilon_{i,2} = \gamma_0 + \gamma_1 * I_{x_i=c_1} + \gamma_2 * I_{x_i=c_2} + ... + \gamma_k * I_{x_i=c_k} + \varepsilon_{i,2}.$$

Instead of using regression estimation to obtain $\hat{g}(x_i)$, the average value of z_i given each category of x_i is used. So, there is no effect measured for every observation of x_i, but an average effect is measured.

$$\sum_{k=1}^{K} w_k * |\hat{g}(c_k) - \nu|,$$

$$\nu = \sum_{k=1}^{K} w_k * \hat{g}(c_k).$$

$$\hat{g}(c_k) = \frac{\sum_{i=1}^{N} I_{x_i=c_k} * z_i}{\sum_{i=1}^{N} I_{x_i=c_k}}$$

$$w_k = \frac{1}{N} \sum_{i=1}^{N} I_{x_i=c_k}.$$

ν is the weighted average of z_i given x_i. $\hat{g}(c_k)$ is the average of z_i given c_k. The difference is taken between the average value of z_i given a category of x_i and the weighted average of z_i given each category of x_i, while taking the weightings into account.

6 Algorithms & methods

This chapter describes the algorithms and methods that are used in the tool to measure and detect differentiation. Model 1 and 2, as given in section 5, are used as examples in this section. As explained before, the data consists of numeric and categorical variables. Both types of variables correspond to different ways to measure differentiation.

differentiation is measured in two ways; in a regular manner and an absolute manner. Both are measured since differentiation is measured for every observation of a variable. This can lead to incorporating both negative and positive measurements which can cancel each other out. So, when for example the variable gender consists of the values man and woman it could be the case that man has a "positive" effect on the model outcome and woman has a "negative" effect on the model outcome. These effects can cancel each other out and a differentiation measurement of 0 is found. This is not a problem when the absolute calculations are performed. However, the direction of differentiation is unknown for man and woman in this case. Since this is valuable information and the regular measurements do show this direction, both methods are used to measure differentiation. The methods below depict the absolute manner and the regular differentiation measurement methods are given in Appendix A.1 - A.4.

Furthermore, the differentiation measurements are standardized. For numeric variables, this means that the measurements represent how much the model outcome changes when the input variable changes by 1 standard deviation, in percentages of the average fitted values of the model. For categorical variables, this means that the measurements represent how much the model outcome changes when the input variable changes by a category, in percentages of the average fitted values of the model.

A point to note is that when a model incorporates discrete steps in its decision-making, like decision tree-based models, the numeric variables are treated as discrete variables in the differentiation measuring. This means that the difference quotient is taken of $f(.)$ with respect to that variable instead of the derivative. This only relates to direct differentiation, not indirect differentiation.

The final subsection contains the algorithm as used in the tool, A basic version is shown, because otherwise, the algorithm becomes too big and hard to explain. It is explained how this algorithm works in practice.

6.1 Situation 1: x_i numeric, z_i numeric

When both x_i and z_i are numeric, simple derivatives are taken to measure differentiation. First, the direct effect of x_i on $f(x_i, z_i)$ is measured by the derivative $D_1 f(x_i, z_i)$. Second, $g(x_i)$ is estimated by $\hat{g}(x_i)$. This is preferably done by using a cubic smoothing spline, but if the IQR is 0, a linear model is fitted. Next, the effect of x_i on $\hat{g}(x_i)$ is measured by $D_1 \hat{g}(x_i)$. This is multiplied with the effect of z_i on $f(x_i, z_i)$: $D_2 f(x_i, z_i)$.

differentiation in the model on variable x_i at observation i can be measured by the following methods.

Direct differentiation:

$$|D_1 f(x_i, z_i) * \frac{\sigma_x}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|. \tag{3}$$

Indirect differentiation:

$$|D_1\hat{g}(x_i) * \frac{\sigma_x}{\sigma_z} * D_2 f(x_i, z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|, \tag{4}$$

$\hat{g}(x_i)$ is the estimation of the function $g(x_i) = E(z_i|x_i)$.

Furthermore, differentiation in the model based on variable x_i is present when:

$$\text{Direct differentiation} = \frac{1}{N}\sum_{i=1}^{N} |D_1 f(x_i, z_i) * \frac{\sigma_x}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

$$\text{Indirect differentiation} = \frac{1}{N}\sum_{i=1}^{N} |D_1\hat{g}(x_i) * \frac{\sigma_x}{\sigma_z} * D_2 f(x_i, z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

6.2 Situation 2: x_i numeric, z_i categorical

If x_i is numeric and z_i categorical, indirect differentiation is measured differently compared to method 1. Since z_i is not numeric, no derivative $D_2 f(x_i, z_i)$ can be calculated. Instead, to measure the effect of z_i on $f(x_i, z_i)$, the difference is taken between the predictions where the value of z_i is set to c_k and the weighted average of these predictions (μ_i). The weighting of each category is taken into account (w_k). This difference is denoted by $E_{i,k}$. Furthermore, for each category of z_i a (nonparametric) regression is made on x_i to estimate the relationship between z_i and x_i.

Consider $z_i \in C$ with K number of categories such that $C = \{c_1, c_2, ..., c_K\}$.

differentiation in the model on variable x_i at observation i can be measured by the following methods.

Direct differentiation:

$$|D_1 f(x_i, z_i) * \frac{\sigma_x}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|. \tag{5}$$

Indirect differentiation :

$$\sum_{k=1}^{K} |D_1\hat{g}_k(x_i) * \sigma_x * E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|, \tag{6}$$

$$E_{i,k} = w_k * |f(x_i, c_k) - \mu_i|,$$

$$\mu_i = \sum_{k=1}^{K} w_k * f(x_i, c_k),$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{z_i = c_k},$$

$\hat{g}_k(x_i)$ is the estimation of $g_k(x_i) = E(c_k|x_i)$.

Furthermore, differentiation in the model based on variable x_i is present when:

$$\text{Direct differentiation} = \frac{1}{N}\sum_{i=1}^{N} |D_1 f(x_i, z_i) * \frac{\sigma_x}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

$$\text{Indirect differentiation} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} |D_1\hat{g}_k(x_i) * \sigma_x * E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

6.3 Situation 3: x_i categorical, z_i numeric

Now, the direct effect of x_i on the model outcome and on z_i are measured differently. The predictions for x_i set to c_k are compared to the weighted average of these predictions (μ_i). This difference is denoted by $E_{i,k}$. The weight per category, w_k, is taken into account. Since z_i is numeric, the effect of z_i on the model outcome can be calculated through $D_2 f(x_i, z_i)$. The effect of x_i on z_i, $U_{i,k}$, is measured in a similar manner as the effect of x_i on $f(x_i, z_i)$, but now the average value of z_i for a given category of x_i is used ($\hat{g}(c_k)$), instead of predictions.

Consider $x_i \in C$ with K number of categories such that $C = \{c_1, c_2, ..., c_K\}$.

differentiation in the model on variable x_i at observation i can be measured by the following methods.

Direct differentiation:

$$\sum_{k=1}^{K} |E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|, \tag{7}$$

$$E_{i,k} = w_k * |f(c_k, z_i) - \mu_i|,$$

$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k, z_i),$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i = c_k}.$$

Indirect differentiation:

$$\sum_{k=1}^{K} |U_{i,k} * \frac{1}{\sigma_z} * D_2 f(x_i, z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|, \tag{8}$$

$$U_{i,k} = w_k * |\hat{g}(c_k) - \nu|,$$

$$\nu = \sum_{k=1}^{K} w_k * \hat{g}(c_k),$$

$$\hat{g}(c_k) = \frac{\sum_{i=1}^{N} I_{x_i = c_k} * z_i}{\sum_{i=1}^{N} I_{x_i = c_k}},$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i = c_k}.$$

Furthermore, differentiation in the model based on feature x is present when:

$$\text{Direct differentiation} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} |E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

$$\text{Indirect differentiation} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} |U_{i,k} * \frac{1}{\sigma_z} * D_2 f(x_i, z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

6.4 Situation 4: x_i categorical, z_i categorical

Earlier subsections already showed how to measure the direct differentiation here for a categorical variable. Regarding indirect differentiation, the differentiation is now measured for every category of z_i, given every category of x_i. So, first, a category of z_i is picked, then the average of the number of occurrences of the category of z_i given the category of x_i is taken ($\hat{g}_l(c_k)$). This is compared to the weighted average v_l, taking into account the weight of every category of x_i. This results in the effect of x_i on z_i, denoted by $U_{k,l}$. This is multiplied by the effect of z_i on $f(x_i, z_i)$. This effect is measured similarly as the direct differentiation on x_i and is denoted by $R_{i,l}$.

Consider $x_i \in C$ with K number of categories such that $C = \{c_1, c_2, ..., c_K\}$.
Consider $z_i \in O$ with L number of categories such that $O = \{o_1, o_2, ..., o_L\}$.

differentiation in the model on variable x_i at observation i can be measured by the following methods.

Direct differentiation:

$$\sum_{k=1}^{K} |E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|, \tag{9}$$

$$E_{i,k} = w_k * |f(c_k, z_i) - \mu_i|,$$

$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k, z_i),$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i=c_k}.$$

Indirect differentiation:

$$\sum_{l=1}^{L} w_l * \sum_{k=1}^{K} |U_{k,l} * R_{i,l} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}|, \tag{10}$$

$$R_{i,l} = w_l * |f(x_i, o_l) - \upsilon_i|,$$
$$i = 1, 2, ..., N,\ l = 1, 2, ..., L,$$

$$\upsilon_i = \sum_{l=1}^{L} w_l * f(x_i, o_l),$$

$$w_l = \frac{1}{N}\sum_{i=1}^{N} I_{z_i=o_l},$$

$$U_{k,l} = w_k * |\hat{g}_l(c_k) - \nu_l|,$$
$$k = 1, 2, ..., K,\ l = 1, 2, ..., L,$$

$$\nu_l = \sum_{k=1}^{K} w_k * \hat{g}_l(c_k).$$

$$\hat{g}_l(c_k) = \frac{\sum_{i=1}^{N} I_{x_i=c_k} * I_{z_i=o_l}}{\sum_{i=1}^{N} I_{x_i=c_k}},$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i=c_k}.$$

Furthermore, differentiation in the model based on feature x is present when:

$$\text{Direct differentiation} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} |E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

$$\text{Indirect differentiation} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} w_l * \sum_{k=1}^{K} |U_{k,l} * R_{i,l} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}| \neq 0.$$

6.5 Algorithm

The basic algorithm is given. It only depicts the absolute differentiation measurements as explained earlier. The algorithm does not include the regular differentiation measurements as the algorithm would become hard to follow. Furthermore, measurement extractions as indirect differentiation per variable and direct/ indirect differentiation per category are also not displayed for the same reason. Bootstrap confidence intervals can also be extracted. This is not shown in the displayed algorithm for the same reasons. However, in the actual algorithm, all these measurements are extracted. The functions in the algorithm refer to the formulas in each of the four situations described above. Note that $function\_3$ is the same as $function\_5$ and $function\_7$ is the same as $function\_9$.

Algorithm 1 differentiation measuring

$x_{i,w}$ is the variable in the input data set at (column) index $w = 1,2,...,W$ and with $i = 1,2,..,N$ observations (rows).
for $w = 1,2,\ldots,W$ **do**
  if $class(x_{i,w})$ =categorical **then**
    $direct_w = function\_7(x_{i,w})$
  else if $class(x_{i,w})$ =numeric **then**
    $direct_w = function\_3(x_{i,w})$
  end if
  for $j = 1,2,\ldots,W-1$ and $j \neq w$ **do**
    $z_{i,j} = x_{i,j}$
    if $class(x_{i,w})$ =categorical **then**
      if $class(z_{i,j})$ =categorical **then**
        $indirect_j = function\_10(x_{i,w}, z_{i,j})$
      else if $class(z_{i,j})$ =numeric **then**
        $indirect_j = function\_8(x_{i,w}, z_{i,j})$
      end if
    end if
    if $class(x_{i,w})$ =numeric **then**
      if $class(z_{i,j})$ =categorical **then**
        $indirect_j = function\_6(x_{i,w}, z_{i,j})$
      else if $class(z_{i,j})$ =numeric **then**
        $indirect_j = function\_4(x_{i,w}, z_{i,j})$
      end if
    end if
  end for
  $Indirect\_differentiation_w = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{W-1}(indirect_j))$
  $Direct\_differentiation_w = \frac{1}{N}\sum_{i=1}^{N}(direct_w)$
end for

The algorithm works as follows. The data enters the first loop of the algorithm. As described earlier, differentiation is measured for each variable in the data set. The first loop picks the variable for which differentiation is measured. First, direct differentiation is measured, depending on the type of variable it is. Next, the second for-loop begins. Here all the other variables in the data set are considered. Indirect differentiation is measured for each combination of $(x_{i,w}, z_{i,j})$. So, indirect differentiation on x_i is measured through every other variable in the data set. Depending on the data type of $x_{i,w}$ and

$z_{i,j}$, the corresponding method is picked to measure differentiation. At the end of the inner loop, the distributions of differentiation measurements are added up and the mean is taken forward. The mean value of the direct and indirect differentiation measurements are the results given in the tool. These measurements are multiplied by 100% to get the percentages.

Indirect differentiation per variable can easily be retrieved by storing $indirect_j$ separately. An example formula is given in Appendix A.5.

differentiation measurements per category can also be retrieved in a similar way. When x_i is categorical, differentiation measurements are calculated per category. These measurements per category can be stored separately. Example formulas are given in Appendices A.6 and A.7. Note, that the weight per category is not taken into account when taking the difference between the predictions (or averages) per category and the weighted average of these predictions (or averages). This way the exact effect per category can be obtained.

Bootstrap confidence intervals around the differentiation measurements can be created. Confidence intervals of 95% strength are created. Furthermore, the percentile interval method is used to create an interval and 399 simulations are performed. 399 simulations are chosen as this does not take too long to compute, which is preferable for the practicality of the differentiation detection tool. In addition it is a good number to use according to research [10]. The percentile interval is used as it is one of the most popular methods to use and no assumptions are made on the underlying distribution [6]. The bootstrap algorithm is given in Appendix A.8.

7 Experimental Setup

In this section, the experimental setup is explained. The goal of the experiments is to test the differentiation detection methods used by the differentiation detection tool; test if the methods works as intended and see if the tool can be of use in the industry. Four simulation models are created to test the methods. Simulations are used because in a simulation it can easily be derived what the expected results are. Thus, it can easily be concluded if the methods work as intended. Besides the simulation models, two Aegon models are used to test the differentiation detection methods. It can not be concluded if the detection methods work well based on the tests on the two Aegon models. It is impossible to derive the expected results. However, the tests do show if the detection tool would function on models used in real life. The testing on the two Aegon models serves as an example to show how the tool can be used in practice.

This section is divided into two subsections. The first subsection describes the simulation models and the experimental setup of the testing. The second subsection revolves around the two Aegon models and the experimental setup there.

7.1 Simulation models

Four simulation models are created to test the differentiation detection methods. The data set used by these models is described in section 3. The four models are different from one another. This is to see how the tool functions when confronted with different types of models.

For simplicity's sake in writing down the models, abbreviations of the variables in the data set are used. The abbreviations are given in Table 3.

Table 3: Abbreviation simulation model data

Variable	letter
Age	a_i
Gender	g_i
Ethnicity	t_i
Education	d_i
Income	v_i
Income category	o_i
House value	h_i
Random variable 1	$r_{i,1}$
Random variable 2	$r_{i,2}$
Correlated variable 1	$c_{i,1}$
Correlated variable 2	$c_{i,2}$
Correlated variable 3	$c_{i,3}$
Correlated variable 4	$c_{i,4}$

Some models incorporate a variable going through step function. This is an extra test for the tool to simulate how it reacts to models which make use of discrete steps in their decision making (decision-

tree based models). The step-function $s(x_i)$ is defined as:

$$s(x_i) = \begin{cases} 1 & \text{if } 16 \leq x_i < 18 \\ 2 & \text{if } 18 \leq x_i < 20 \\ 3 & \text{if } 20 \leq x_i < 22 \\ 4 & \text{if } 22 \leq x_i < 24 \\ 5 & \text{if } 24 \leq x_i < 26 \\ 6 & \text{if } 26 \leq x_i < 28 \end{cases}$$

7.1.1 Simulation model 1: A simple linear model with direct differentiation

$$y_i = 0.5 * g_i + 0.01 * a_i + I_{t_i=Finnish} + 0.5 * d_i + 2 * r_{i,1} + 3 * r_{i,2}$$

Simulation model 1 is used to test if direct differentiation can be picked up in a simple linear model. This model differentiates on the following variables: gender, age, ethnicity, education, random variable 1, and random variable 2. In this case, it is expected that the tool will pick up direct differentiation for these variables. Furthermore, since it is a simple linear model the coefficients of every variable give a clue to the differentiation measurement that should be measured. The differentiation measurements are standardized. If this wasn't the case, the differentiation measurements returned should be equal to the model coefficients. The highest direct differentiation measurement should be picked up for random variable 2 and the lowest measurement should be picked up for age.

7.1.2 Simulation model 2: A non-linear model with direct differentiation

$$y_i = 0.001 * a_i^2 + 4 * log(r_{i,1}) + 1e^{-5} * (r_{i,2})^3 + 0.2 * s(c_{i,1})$$

In simulation model 2 it is tested if direct differentiation can be picked up in a non-linear model. The model differentiates on the variables age, random variable 1, random variable 2, and correlated variable 1. Furthermore, it is also tested if differentiation can be picked up for a variable that goes through a step function. Moreover, although it is not the main objective here, indirect differentiation for age and ethnicity should also be picked up; correlated variable 1 is dependent on age and ethnicity.

7.1.3 Simulation model 3: A simple linear model with indirect differentiation

$$y_i = c_{i,1} + 2 * c_{i,2} + 3 * c_{i,3} - 4 * c_{i,4}$$

In this simple linear model, the expected results for the direct differentiation should easily be found for the four correlated variables. Model 3 differentiates on correlated variable 1, correlated variable 2, correlated variable 3, and correlated variable 4. The correlated variables are all dependent on other variables. These are age ($c_{i,1}$ & $c_{i,4}$), ethnicity ($c_{i,1}$ & $c_{i,2}$) and gender ($c_{i,3}$). Therefore, indirect differentiation should be found in these three variables. So, it is tested if differentiation is detected for variables that are not observed (directly) by the model, but which have an influence on the model through other variables.

7.1.4 Simulation model 4: A non-linear model with interaction terms

$$y_i = 0.2 * s(c_{i,1}) + 0.01 * a_i * d_i + 0.1 * log(v_i) * s(c_{i,2}) + log(h_i) * (v_i > 5000) * (I_{d_i=3}) + 0.1 * s(c_{i,4})$$

Simulation model 4 is a non-linear model with a step function. The model differentiates on correlated variable 1, age, education, income, correlated variable 2, house value and correlated variable 4. Furthermore, there are interactions between the input variables. This model is closer to reality as machine learning models can create opaque relationships between variables. This makes it hard to understand how the model behaves and how important each variable is in the predictions of the model. Still, some expectations can be made. The input variables are expected to harbor direct differentiation. Indirect differentiation is expected to be found in income, age and ethnicity due to dependencies. House value is dependent on income. The three correlated variables are dependent on (one of) age and ethnicity.

8 Results

In this section, the results are given for the experiments described in section 7. The results given in the tables are related to the expectations made in section 7 and the two research questions. The two research questions are: ”Can differentiation be detected in insurance pricing models?” and can ”differentiation be measured in insurance pricing models?”

The measurements in the tables for direct differentiation represent how much the model outcome changes when the variable of interest changes (by a standard deviation/ category), in percentages of the mean of the fitted model values. The indirect differentiation measurements should be read in the same manner. Indirect differentiation measurements are given for the row variables coming through a column variable. The row variables are the variables for which differentiation is measured.

The subsections below show only parts of the results. The full result tables are given in Appendix C. The direct differentiation measurements are given in a regular manner and in an absolute manner, to give a full picture of the differentiation measurements.

Lastly, the standard deviations of the numeric variables are given in Appendix B.2. in Table 23. The mean fitted values of the four simulation models are given in the same Appendix in Table 24. With this information, the un-standardized results can be calculated.

8.1 Simulation Models

8.1.1 Simulation model 1

Table 4 displays the direct differentiation measurements for the input variables of simulation model 1. As a first example of what the measurements mean: The mean of the model outcome changes by 5.915% if random variable 1 changes by one standard deviation. The mean of the model outcome changes by 0.129% if ethnicity changes by a category.

Direct differentiation is only measured for the input variables of simulation model 1. The regular measurements always have a differentiation measurement of 0 for categorical variables, but the absolute measurements do show significant results. The size of the differentiation measurements also correspond to the coefficient every variable has in the model. This can also be checked by recalculating the measurements to their un-standardized versions. The biggest differentiation measurement is picked up for random variable 2 and the lowest measurement is picked up for age. In the case of ethnicity, the methods have picked up that Finnish people pay more compared to other ethnicities (Table 5). There were no expectations made about indirect differentiation. However, indirect ”noise” differentiation is picked up for every variable as given by the tables in Appendix C.1.

Table 4: Direct differentiation measurements

	Gender	Age	Ethnicity	Education	Random var 1	Random var 2
Regular	0.000 %	0.049 %	0.000 %	0.000 %	3.963 %	5.915 %
Absolute	0.207 %	0.049 %	0.129 %	0.429 %	3.963 %	5.915 %

Table 5: Direct differentiation measurements for ethnicity

	Belgian	Dutch	Finnish	German	Other
Regular	-0.070 %	-0.070 %	0.910 %	-0.070 %	-0.070 %
Absolute	0.070 %	0.070 %	0.910 %	0.070 %	0.070 %

Referring back to the two research questions. differentiation can be detected and measured as expected, given simulation model 1.

8.1.2 Simulation model 2

Table 6 shows that direct differentiation is picked up for all the input variables of simulation model 2. differentiation is found for correlated variable 1 which goes through a step function. However, the measurement is not as expected. According to the measurement in a regular manner, a change in correlated variable 1 has a negative effect on the model outcome.

Table 6: Direct differentiation measurements

	Age	Random var 1	Random var 2	Corr var 1
Regular	2.775 %	2.740 %	0.176 %	-0.910 %
Absolute	2.775 %	2.740 %	0.176 %	1.690 %

Table 7 shows the indirect differentiation measurements for ethnicity and age coming through age, random variable 1, random variable 2 and correlated variable 1. As an example: The table shows that if ethnicity changes by a category the mean of the model predictions changes by 0.021% through age. differentiation is picked up for ethnicity coming through age and correlated variable 1 and for age coming through correlated variable 1. Other indirect differentiation measurements are also picked up. However, these measurements are much smaller. Still, noise is picked up regarding indirect differentiation for every variable as given in the Tables in Appendix C.2.

Table 7: Absolute indirect differentiation measurements for ethnicity and age

	Age	Random var 1	Random var 2	Corr var 1
Ethnicity	0.021 %	0.004 %	0.000 %	0.022 %
Age	-	0.002 %	0.001 %	0.016 %

Referring back to the two research questions. differentiation can be detected as expected but not measured as expected, given simulation model 2.

8.1.3 Simulation model 3

Direct differentiation is measured for all the input variables. The measurements also correspond to the coefficients each variable has in model 3; correlated variable 1 has the lowest differentiation measurement and correlated variable 4 has the biggest differentiation measurement.

Table 8: Direct differentiation measurements

	Corr var 1	Corr var 2	Corr var 3	Corr var 4
Regular	4.582 %	10.232 %	13.589 %	-18.345 %
Absolute	4.582 %	10.232 %	13.589 %	18.345 %

Table 9 contains the absolute indirect differentiation measurements for age, ethnicity and gender. Again, noise differentiation is picked up. For age, the biggest differentiation measurements flow through correlated variable 1 and correlated variable 4. These are also the variables that are dependent on age. Correlated variable 2 is only dependent on ethnicity. This is also shown in Table 9. Indirect differentiation of ethnicity through correlated variable 2 has a big measurement. Gender only has a dependency in correlated variable 3. Table 9 shows that the biggest indirect differentiation measurement of gender goes through correlated variable 3. The full tables are given in Appendix C.3.

Table 9: Absolute indirect differentiation measurements for age, ethnicity and gender

	Corr var 1	Corr var 2	Corr var 3	Corr var 4
Age	0.044 %	0.031 %	0.022 %	0.443 %
Ethnicity	0.059 %	3.379 %	0.034 %	0.074 %
Gender	0.010 %	0.030 %	1.381 %	0.061 %

Taking age as en example: when the measurement through correlated variable 4 is un-standardized and the effect of correlated variable 4 on the model is taken away, the result is the effect of age on correlated variable 4. This effect closely resembles the factor 0.01 by which correlated variable 4 is dependent on age, as given in section 3. The same holds for the male gender when it comes to correlated variable 3. This is given in the following table.

Table 10: Indirect differentiation measurement for gender

	Female	Male
Regular	-2.319 %	1.062 %
Absolute	2.453 %	1.005 %

Referring back to the two research questions. differentiation can be detected and measured as expected, given simulation model 3.

8.1.4 Simulation model 4

Table 11 shows that direct differentiation is picked up for all the input variables of the model, except for correlated variable 4. Particularly high measurements are found for education, income and correlated variable 2.

Table 11: Direct differentiation measurements

	Age	Education	Income	House value	Corr var 1	Corr var 2
Regular	1.838 %	0.000 %	15.344 %	1.396 %	-1.804 %	5.896 %
Absolute	1.838 %	32.042 %	15.344 %	1.396 %	5.477 %	23.488 %

	Corr var 4
Regular	0.000 %
Absolute	0.000 %

Not all the expected indirect differentiation measurements are found. No indirect differentiation measurement is found flowing through correlated variable 4, regarding age. Moreover, the indirect differentiation measurement of age through correlated variable 1 also is not that big compared to the noise measurements through other variables. For ethnicity, the biggest indirect differentiation measurement is found through correlated variable 2 and the second highest measurement is found through correlated variable 1. Maybe the most surprising results are those of income category and house value. It shows relatively big indirect differentiation measurements coming through income. Income is not dependent on those two variables. On the contrary, income category and house value are dependent on income. All the results can be found in Appendix C.4.

Table 12: Absolute indirect differentiation measurements

	Age	Education	Income	House value	Corr var 1	Corr var 2
House value	0.029 %	0.133 %	15.085 %	-	0.017 %	0.270 %
Age	-	0.042 %	0.041 %	0.004 %	0.052 %	0.070 %
Ethnicity	0.013 %	0.018 %	0.037 %	0.004 %	0.070 %	7.760 %
Income category	0.009 %	0.077 %	7.244 %	2.294 %	0.011 %	0.087 %
Income	0.017 %	0.124 %	-	1.372 %	0.012 %	0.217 %

	Corr var 4
House value	0.000 %
Age	0.000 %
Ethnicity	0.000 %
Income category	0.000 %
Income	0.000 %

Referring back to the two research questions. differentiation can not be detected and measured as expected, given simulation model 4.

9 The tool

This section explains how the tool works that incorporates the differentiation detection methods described in this research. As explained in the introduction, the tool consists of two components: programming code that detects and measures differentiation and an app that can make visualizations of the results. The two components of the tool are programmed in R.

The script in which the detection methods are programmed is shared with a model owner. When the script is shared, the user gives a data set and a model (that has a prediction function). Next, the user has to define a couple of fields. An example of such a field is whether the given data set should be enriched with CBS data. After these fields are declared, the script can be run. The code includes the same data pre-processing as described in section 3. The script produces a resulting data set with differentiation measurements for each variable in the original data set. The measurements are given in Table 15.

Table 13: The differentiation measurements performed by the tool

	Description
1	Direct differentiation (regular)
2	Direct differentiation per category (regular)
3	Direct differentiation (absolute)
4	Direct differentiation per category (absolute)
5	Indirect differentiation (regular)
6	Indirect differentiation per variable (regular)
7	Indirect differentiation per category (regular)
8	Indirect differentiation (absolute)
9	Indirect differentiation per variable (absolute)
10	Indirect differentiation per category (absolute)
11	Total differentiation (regular)
12	Total differentiation per category (regular)
13	Total differentiation (absolute)
14	Total differentiation per category (regular)

The user gives the resulting data set to the app. When the data set is uploaded to the app, the user can make several visualizations. These visualizations display the different measurements. The biggest differentiating features can be displayed. The differentiation measurements for certain variables of interest can be displayed and the differentiation measurements on a micro level can be displayed. The next figures show how the app works.

First, a data set is uploaded with the upload button. Next, the user can choose what kind of graph he or she wants to be displayed. The selection consists of the choices ”Every variable” and ”Distinct selection”. ”Every variable” refers to the whole data set. A differentiation measurement and a sorting can be chosen. The graph will display a number of variables sorted regarding the chosen differentiation measurement. This is displayed in Figure 3.

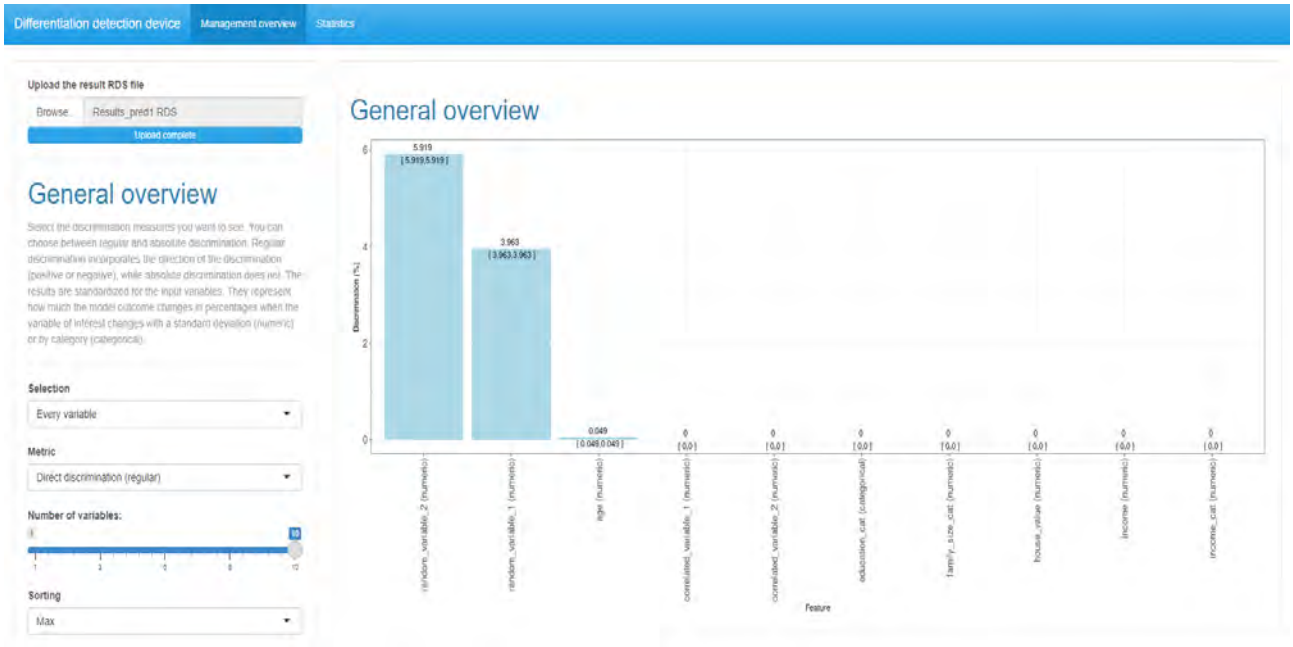

Figure 3: ”Every variable” graph

”Distinct selection” gives a new input option where the user can choose up to five features to display. The user can choose what kind of metric, absolute or regular, the graph should display. For every variable, the direct, indirect and total differentiation measurements are displayed for the chosen metric. Figure 3 shows the result when ”distinct selection” is picked.

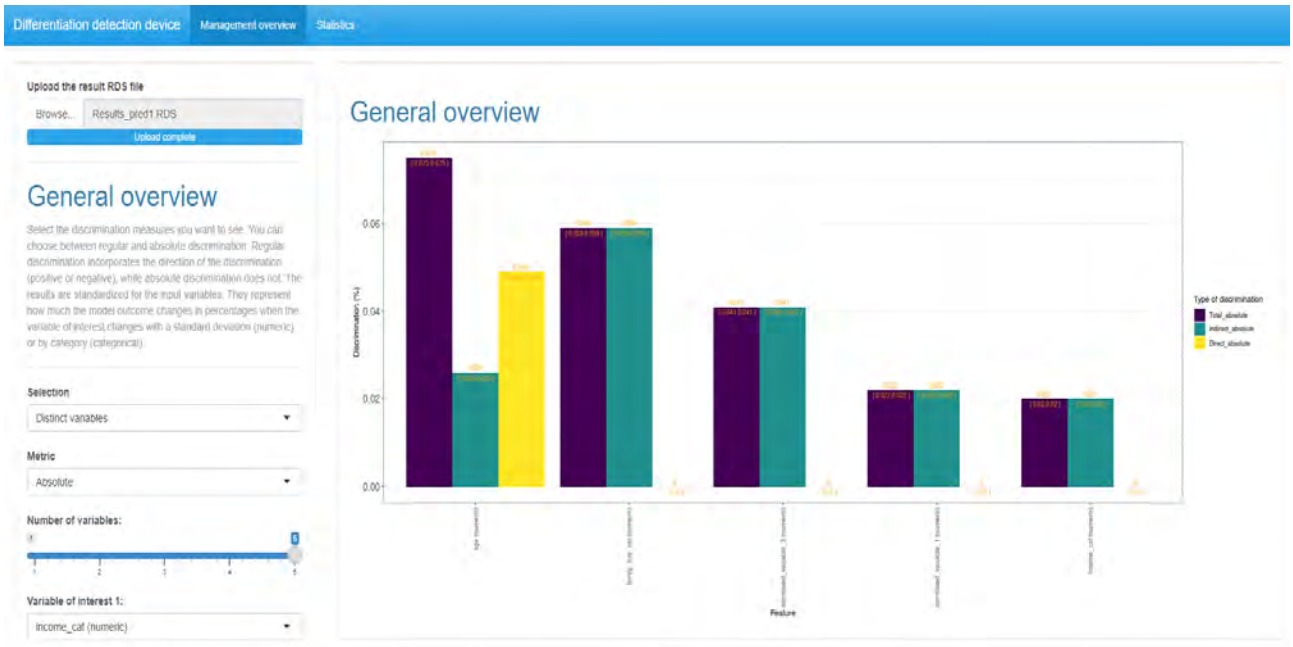

Figure 4: ”Distinct selection” graph

The app also contains a second graph. This graph displays the differentiation measurements per category and the indirect differentiation measurements per variable. Figure 5 displays such a graph.

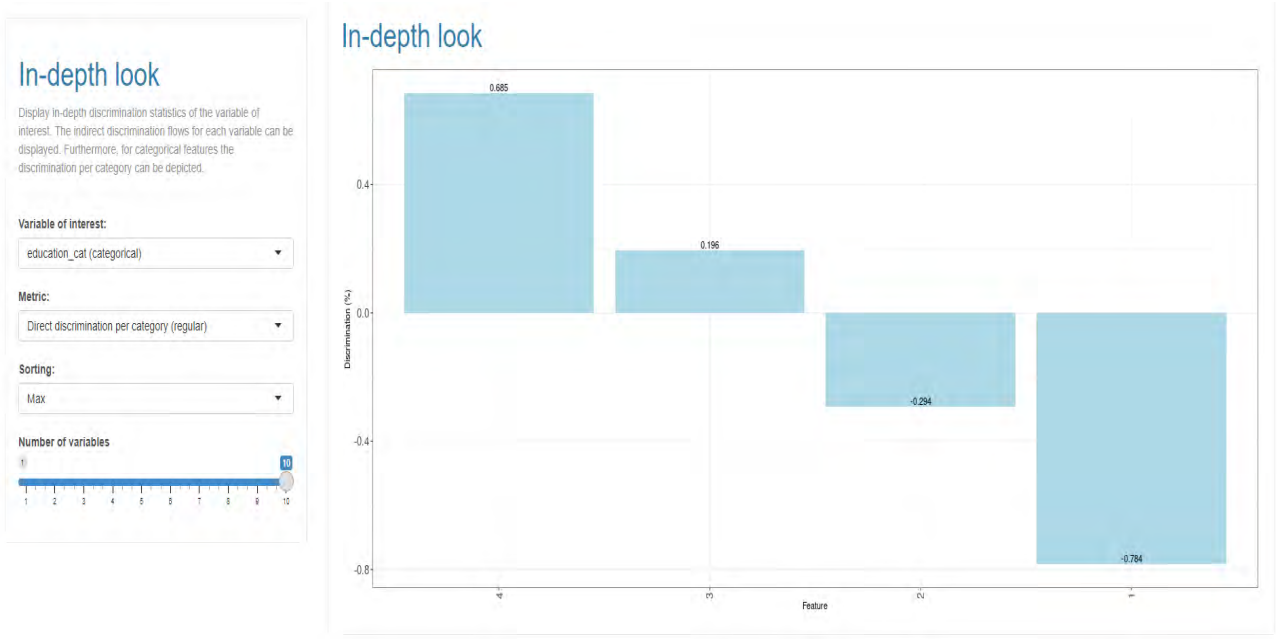

Figure 5: Count per category

Furthermore, the app consists of a second tab. The same measurements as displayed in Figure 5 can be displayed in box plots. This way the user can learn more about the distribution of the differentiation measurements. A visualisation is given in Figure 6.

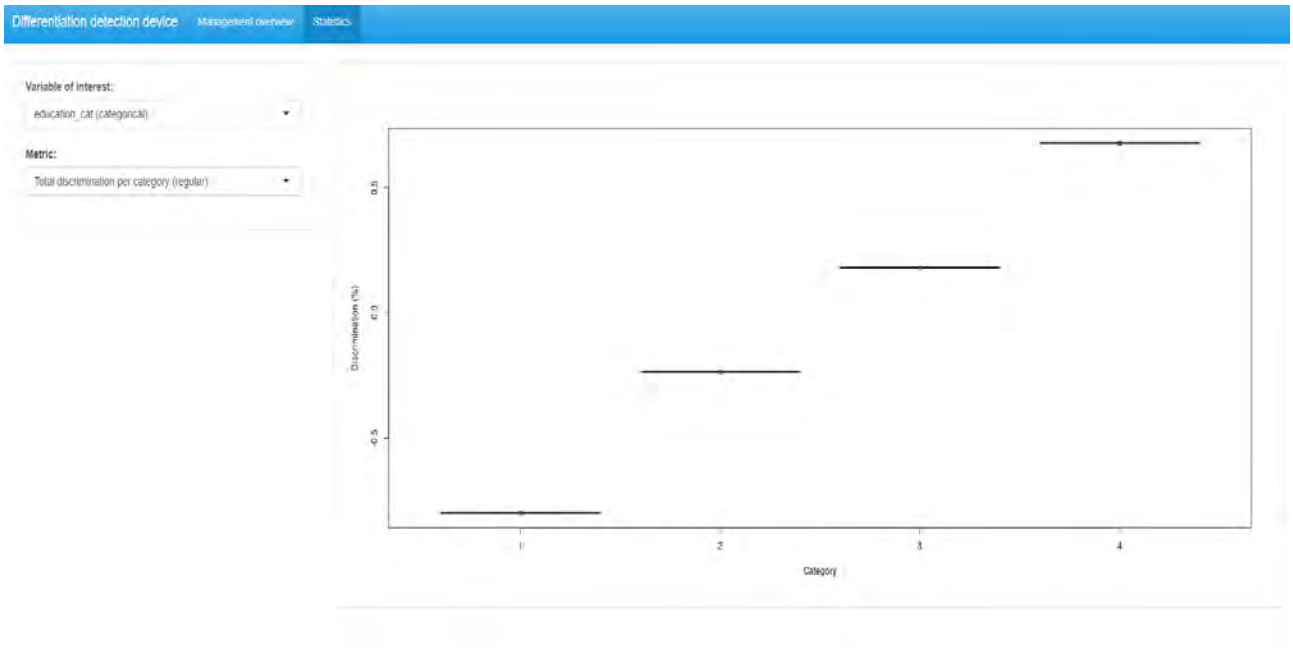

Figure 6: Count per category

10 Conclusion

10.1 Summary

This thesis introduces a new tool to detect and measure differentiation in insurance pricing models. The literature research shows that, although the academic field has only recently become popular, there are already quite some tools that aim to do something similar. The tool in this research distinguishes its self by how easy it is to use and by the information the tool gives. So far, there does not exist an alternative tool or technique which can detect and measure differentiation on the same level as the methods used by the tool of this thesis. The detection tool is tested on four simulation models and two Aegon models. The tests show that the tool can detect and measure differentiation in most models. However, the tool has problems with measuring differentiation in a few situations. The tool does not correctly measure or even detect differentiation when a model incorporates a step function. Furthermore, the tool picks op noise differentiation. differentiation is detected where there should be none found. The tests on the Aegon models showed that differentiation can be detected for variables the models do not even observe. In the XGBoost model, the most differentiating features originated from the CBS data set. This shows why it is important to enrich a given data set with the CBS data set. This way Aegon can gather more information about the behavior of their pricing models.

The biggest difference between this tool and other tools and researches about differentiation detection is the information this tool provides. The existing tools used in the industry do not make the distinction between direct and indirect differentiation and also do not work on numeric data. Furthermore, no other research describes how to dissect the differentiation measurements into measurements per category, for categorical variables, or into measurements per variables in the case of indirect differentiation. So, it can be seen through which variables the variable of interest is differentiated on by a model.

The experiments show that direct differentiation is almost always picked up. The tests in simulation model 4 and simulation model 2 show that there can be difficulties in detecting direct differentiation for variables that go through a step function. However, differentiation can be measured in the XGBoost model, which also incorporates discrete steps in its decision-making. There are also some problems with the indirect differentiation measuring. Noise differentiation is picked up here. Because of this, the indirect differentiation measurements for variables that have a slight dependency on another variable do not stand out compared to the noise differentiation measurements. Overall, the tests do show that for a large part differentiation is detected and measured as expected.

With this tool, Aegon can get a lot of information about the behavior of the models they use. differentiation can be detected and measured. It can be found where the differentiation originates from and with this information the model owners can explain to interested people how their model behaves. The most important part here is that any doubt about possible differentiating models can be taken away. Because, if the tool finds differentiation, this can be negated with the information the model owner gets from the tool. If unwanted direct differentiation is found, the model owner can take the variable of interest out of the data set. The same can be done in the case of unwanted indirect differentiation. In that case, the second variable can be taken out of the data set. This corresponds to taking x_i or z_i out of the relationship depicted in Figure 1 in section 4. If the variables are too important in the model predictions, the model owner can incorporate a contrary differentiation factor. If women pay a higher insurance premium than men, Aegon can give a discount to women for example.

There are some limitations to the methods to measure differentiation. First of all, the tool is very much dependent on the data the user gives. The differentiation found does not necessarily imply that the model is differentiating. It can also mean that there is a certain bias in the given data set. A category within the variable ethnicity can be overrepresented for example. These biases will lead to biased model predictions which are picked up by the tool. The tool will give a differentiation measurement, but it does not say if this originates from the model or the data. The user has to keep in mind that not only the model can be at fault, but there could also be problems in the given data.

Furthermore, the tool incorporates data pre-processing which has some room for improvement. If a data set has a categorical variable with lots of missing values, the tool will perform a very skewed pre-processing. All the missing values are imputed by the median value or the most common value. Then again, this can easily be avoided if the user him- or herself performs the pre-processing. The data preparation of the tool right now works as intended; fill up empty values, such that the tool can run. Preparing a data set is not the objective of the tool.

Lastly, regarding the indirect differentiation measuring: a statistical relationship does not imply causation. The indirect differentiation measurement methods use regression estimations and average values. These techniques will show relationships between variables, but they do not imply that a change in one variable is caused by a change in the other variable. This causal relationship is implied in the way differentiation is defined in this thesis. The test results in simulation model 4 show that high indirect differentiation measurements are found for income category and house value through income. This implies that a change in house value or income category causes a big change in the model outcome through the income variable. This is not correct. Those two variables are dependent on income, not the other way round. These measurements are found, because of the correlations between house value and income, and income category and income. This is also the reason why there is a relatively big differentiation measurement of income category through house value. These two variables are correlated, because they are correlated with income. However, neither house value or income category have a causal relationship with one another.

10.2 Future Work

There are a couple of points which should be improved. First of all, more research should be put into causation theory. As explained before, right now relationships are established where there actually should be none. It is not weird that these correlations are detected, but for this tool to fully work as intended, only relationships should be established where one variable causes the other variable. That way, the indirect differentiation measurements truly represent the theory as introduced in this research.

Secondly, more work could be put into the data pre-processing. This is not one of the objectives of the tool. However, a better data preparation would lead to more accurate results, as the data will not be skewed because of the tool if it is not properly cleaned. This can be improved by incorporating machine learning pre-processing techniques.

Lastly, more tests should be done on real models. The simulation model test shows that the tool has problems with step functions. However, the test on the XGBoost model shows that the tool does detect differentiation in this model which makes use of discrete steps in its decision making. Furthermore, there are other types of models that have not been tested. These include classifying models.

References

[1] Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical programming*, 95(1):3–51, 2003.

[2] Laurence Barry and Arthur Charpentier. Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143, 2020.

[3] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[4] Vaishak Belle and Ioannis Papantonis. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*, 2020.

[5] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[6] James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine*, 19(9):1141–1164, 2000.

[7] Alberto Cevolini and Elena Esposito. From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data & Society*, 7(2):2053951720939228, 2020.

[8] European Commission. 2018 reform of eu data protection rules.

[9] Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. Biased programmers? or biased data? a field experiment in operationalizing ai ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 679–681, 2020.

[10] Russell Davidson and James G MacKinnon. Bootstrap tests: How many bootstraps? *Econometric Reviews*, 19(1):55–68, 2000.

[11] Martin Eling, Davide Nuessle, and Julian Staubli. The impact of artificial intelligence along the insurance value chain and on the insurability of risks. *The Geneva Papers on Risk and Insurance-Issues and Practice*, pages 1–37, 2021.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] B. E. Hansen. *Econometrics*. University of Wisconsin, Department of Economics, 2002.

[14] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, et al. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. 2021.

[15] Roel Henckaerts, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285, 2021.

[16] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

[17] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2021.

[18] Mathias Lindholm, Ronald Richman, Andreas Tsanakas, and Mario V Wuthrich. Discrimination-free insurance pricing. *Available at SSRN 3520676*, 2020.

[19] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[20] mostly.ai. https://mostly.ai/, 2020.

[21] Avneet Pannu. Artificial intelligence and its application in different areas. *Artificial Intelligence*, 4(10):79–84, 2015.

[22] Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018.

[23] Oxford University Press. *Discrimination*. Oxford University Press.

[24] B. D. Ripley and Martin Maechler. *Smooth.spline: Fit a smoothing spline*.

[25] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[26] LS Shapley. Notes on the n-person game – ii: The value of an n-person game. 1951.

[27] B. W. Silverman. Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*, 12(3):898 – 916, 1984.

[28] Centraal Bureau voor de Statistiek. Statistische gegevens per vierkant en postcode 2020-2019-2018.

[29] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer New York, 2013.

[30] Catherina Xu, Christina Greer, Manasi N Joshi, and Tulsee Doshi. Fairness indicators demo: Scalable infrastructure for fair ml systems, 2020.

[31] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. *arXiv preprint arXiv:1810.07155*, 2018.

[32] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

Appendices

A Algorithms & formulas

A.1 Numeric x_i, numeric z_i

Direct differentiation:

$$D_1 f(x_i, z_i) * \frac{\sigma_x}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}.$$

Indirect differentiation:

$$D_1 \hat{g}(x_i) * \frac{\sigma_x}{\sigma_z} * D_2 f(x_i, z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}.$$

A.2 Numeric x_i, categorical z_i

Direct differentiation:

$$D_1 f(x_i, z_i) * \frac{\sigma_x}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)}.$$

Indirect differentiation :

$$\sum_{k=1}^{K} D_1 \hat{g}_k(x_i) * \sigma_x * E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)},$$

$$E_{i,k} = w_k * (f(x_i, c_k) - \mu_i),$$

$$\mu_i = \sum_{k=1}^{K} w_k * f(x_i, c_k),$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{z_i = c_k}.$$

A.3 Categorical x_i, numeric z_i

Direct differentiation:

$$\sum_{k=1}^{K} E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)},$$

$$E_{i,k} = w_k * (f(c_k, z_i) - \mu_i),$$

$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k, z_i),$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i = c_k}.$$

Indirect differentiation:

$$\sum_{k=1}^{K} U_{i,k} * \frac{1}{\sigma_z} * D_2 f(x_i, z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i, z_i)},$$

$$U_{i,k} = \sum_{k=1}^{K} w_k * (\hat{g}(c_k) - \nu),$$

$$\nu = \sum_{k=1}^{K} w_k * \hat{g}(c_k),$$

$$\hat{g}(c_k) = \frac{\sum_{i=1}^{N} I_{x_i = c_k} * z_i}{\sum_{i=1}^{N} I_{x_i = c_k}},$$

$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i = c_k}.$$

A.4 Categorical x_i, categorical z_i

Direct differentiation:

$$\sum_{k=1}^{K} E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i,z_i)},$$
$$E_{i,k} = w_k * (f(c_k,z_i) - \mu_i),$$
$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k,z_i).$$
$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i=c_k}$$

Indirect differentiation:

$$\sum_{l=1}^{L} w_l * (\sum_{k=1}^{K} U_{k,l} * R_{i,l} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i,z_i)}),$$
$$R_{i,l} = w_l * (f(x_i,o_l) - \upsilon_i),$$
$$\upsilon_i = \sum_{l=1}^{L} w_l * f(x_i,o_l),$$
$$w_l = \frac{1}{N}\sum_{i=1}^{N} I_{z_i=o_l},$$
$$U_{k,l} = w_k * (\hat{g}_l(c_k) - \nu_l),$$
$$\nu_l = \sum_{k=1}^{K} w_k * \hat{g}_(c_k).$$
$$\hat{g}_l(c_k) = \frac{\sum_{i=1}^{N} I_{x_i=c_k} * I_{z_i=o_l}}{\sum_{i=1}^{N} I_{x_i=c_k}},$$
$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i=c_k}.$$

A.5 Indirect differentiation per variable

Absolute:

$$\frac{1}{N}\sum_{i=1}^{N} |D_1\hat{g}(x_i) * \frac{\sigma_x}{\sigma_z} * D_2 f(x_i,z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i,z_i)}|.$$

Regular:

$$\frac{1}{N}\sum_{i=1}^{N} D_1\hat{g}(x_i) * \frac{\sigma_x}{\sigma_z} * D_2 f(x_i,z_i) * \frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N} f(x_i,z_i)}.$$

A.6 Direct differentiation per category

Absolute:

$$\frac{1}{N}\sum_{i=1}^{N} |E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i,z_i)}|,$$
$$E_{i,k} = |f(c_k,z_i) - \mu_i|,$$
$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k,z_i),$$
$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i=c_k}.$$

Regular:

$$\frac{1}{N}\sum_{i=1}^{N} E_{i,k} * \frac{1}{\frac{1}{N}\sum_{i=1}^{N} f(x_i,z_i)},$$
$$E_{i,k} = (f(c_k,z_i) - \mu_i),$$
$$\mu_i = \sum_{k=1}^{K} w_k * f(c_k,z_i),$$
$$w_k = \frac{1}{N}\sum_{i=1}^{N} I_{x_i=c_k}.$$

A.7 Indirect differentiation per category

Absolute:

$$\frac{1}{N}\sum_{n=1}^{N}|U_{i,k}*\frac{1}{\sigma_z}*D_2f(x_i,z_i)*\frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N}f(x_i,z_i)}|,$$

$$U_{i,k}=|\hat{g}(c_k)-\nu|,$$

$$\nu=\sum_{k=1}^{K}w_k*\hat{g}(c_k),$$

$$\hat{g}(c_k)=\frac{\sum_{i=1}^{N}I_{x_i=c_k}*z_i}{\sum_{i=1}^{N}I_{x_i=c_k}},$$

$$w_k=\frac{1}{N}\sum_{i=1}^{N}I_{x_i=c_k}.$$

Regular:

$$\frac{1}{N}\sum_{n=1}^{N}U_{i,k}*\frac{1}{\sigma_z}*D_2f(x_i,z_i)*\frac{\sigma_z}{\frac{1}{N}\sum_{i=1}^{N}f(x_i,z_i)},$$

$$U_{i,k}=\hat{g}(c_k)-\nu,$$

$$\nu=\sum_{k=1}^{K}w_k*\hat{g}(c_k),$$

$$\hat{g}(c_k)=\frac{\sum_{i=1}^{N}I_{x_i=c_k}*z_i}{\sum_{i=1}^{N}I_{x_i=c_k}},$$

$$w_k=\frac{1}{N}\sum_{i=1}^{N}I_{x_i=c_k}.$$

A.8 Bootstrap

Algorithm 2 Boostrap algorithm

1: B is the number of simulations
2: F^* is the empirical distribution
3: $x=\{x_1,x_2,...,x_n\}$
4: $T(x)=\frac{1}{N}\sum_{j=1}^{N}x_i$
5: **for** $i=1,2,\ldots,B$ **do**
6: $x^*=x_1^*,x_2^*,...,x_n^*\sim F^*$
7: $\text{Tboot}_i=\text{T}(\text{x}^*)$
8: **end for**

The bootsrap confidence intervals:

$$C_n=(\theta^*_{\alpha/2},\theta^*_{1-\alpha/2})$$

θ^* is the mean obtained from the bootstrap algorithm.

B Data

B.1 CBS

Table 14: Data of CBS

Feature	Type	Description
Inwoner	Numeric	Number of residents
Man	Numeric	Number of men
Vrouw	Numeric	Number of women
INW_014	Numeric	Number of residents aged -14
INW_1524	Numeric	Number of residents aged 15-24
INW_2544	Numeric	Number of residents aged 25-44
INW_4564	Numeric	Number of residents aged 45-64
INW_65PL	Numeric	Number of residents aged 64+
P_NL_ACHTG	Numeric	Percentage of residents with Dutch native background
P_WE_MIG_A	Numeric	Percentage of residents with western background
P_NW_MIG_A	Numeric	Percentage of residents with non-western background
AANTAL_HH	Numeric	Number of households
TOTHH_EENP	Numeric	Number of households with one person
TOTHH_MPZK	Numeric	Number of households with multiple persons and no children
HH_EENOUD	Numeric	Number of households with one parent and children
HH_TWEEOUD	Numeric	Number of households with two parents and children
GEM_HH_GR	Numeric	The number of residents per household
WONING	Numeric	The number of houses
WONVOOR45	Numeric	The number of houses built before 1945
WON_4564	Numeric	The number of houses built between 1945-1965
WON_6574	Numeric	The number of houses built between 1965-1975
WON_7584	Numeric	The number of houses built between 1975-1985
WON_8594	Numeric	The number of houses built between 1985-1995
WON_9504	Numeric	The number of houses built between 1995-2005
WON_0514	Numeric	The number of houses built between 2005-2015
WON_1524	Numeric	The number of houses built between 2015-2025
WON_MRGEZ	Numeric	The number of multiple-families houses
P_KOOPWON	Numeric	Percentage of owner-occupied houses
P_HUURWON	Numeric	Percentage of rental houses
WON_HCORP	Numeric	Number of houses owned by corporations
WON_NBEW	Numeric	The number of uninhabited houses
WOZWONING	Numeric	The average value of the houses
UITKMINAOW	Numeric	The number of residents with welfare
Postcode	Category	The postal code

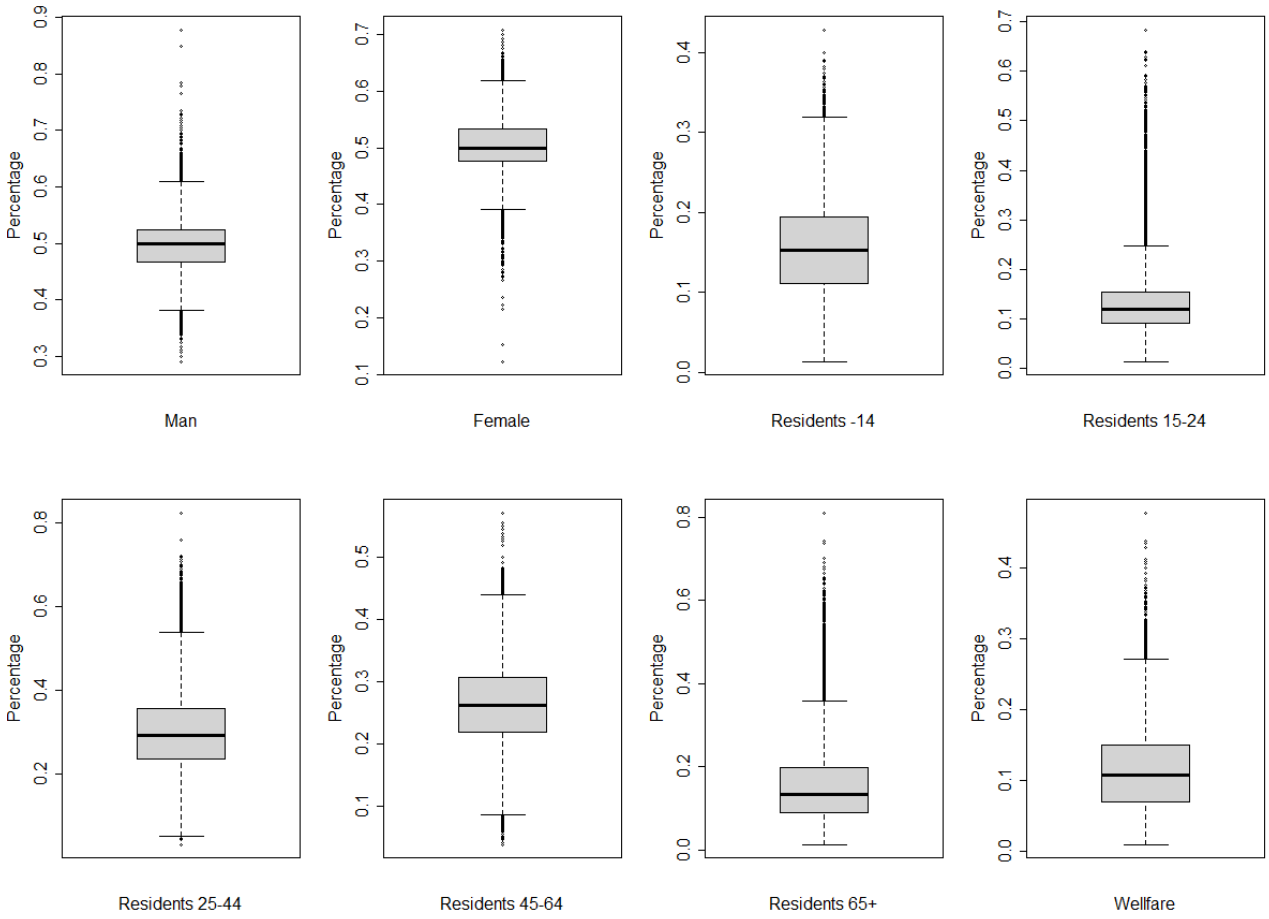

Figure 7: Box plots of the population variables

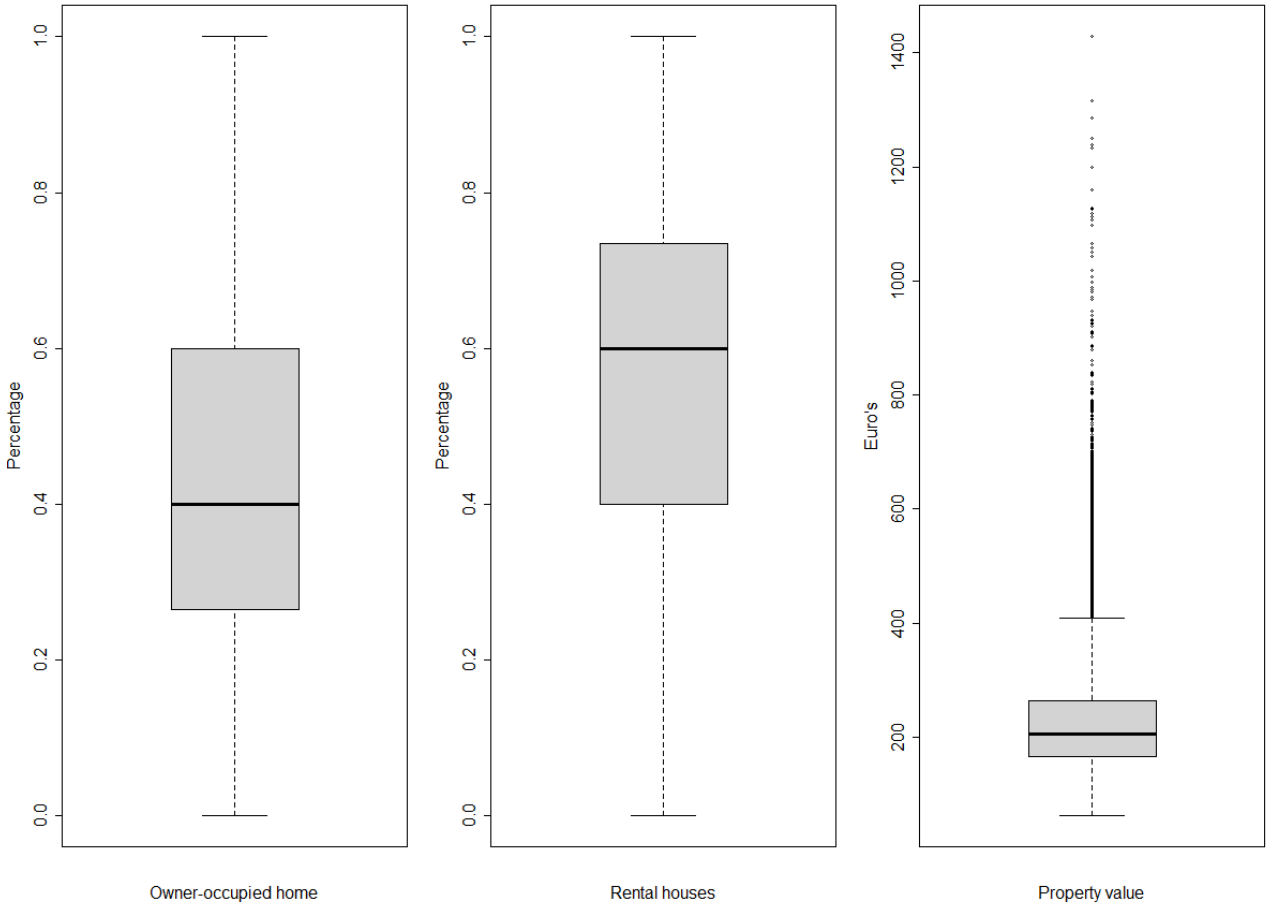

Figure 8: Box plots of the property variables

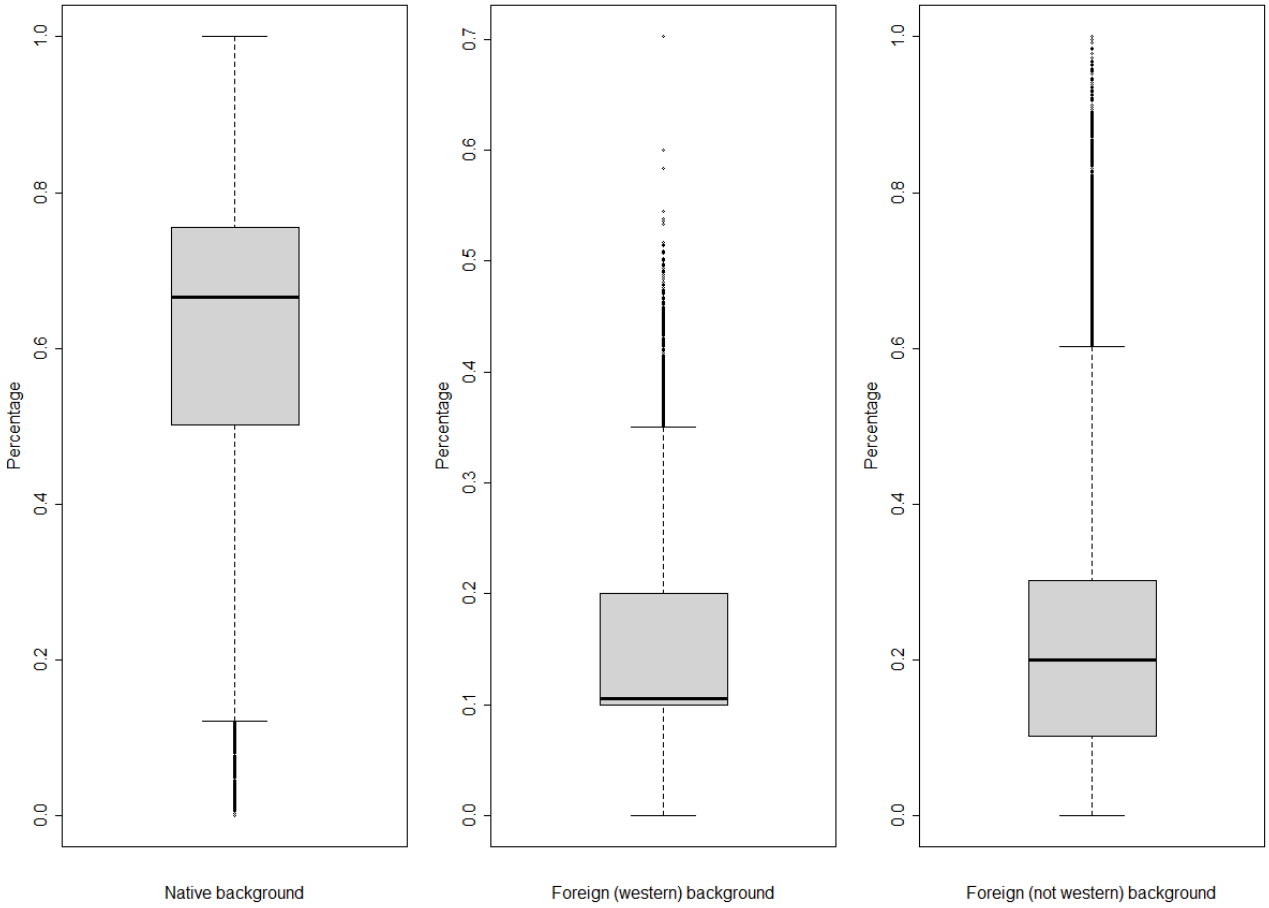

Figure 9: Box plots of the background variables

B.2 Simulation

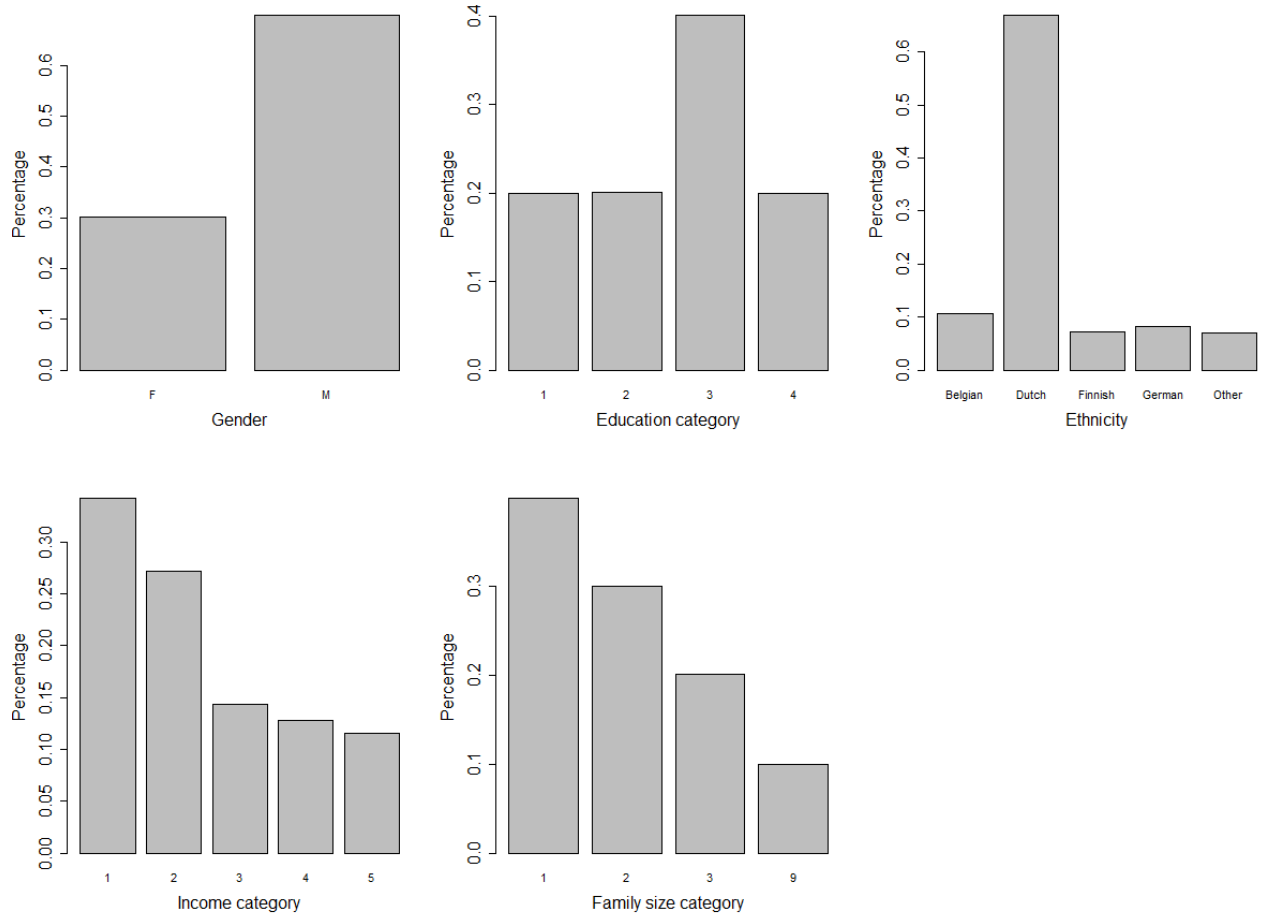

Figure 10: Count per category

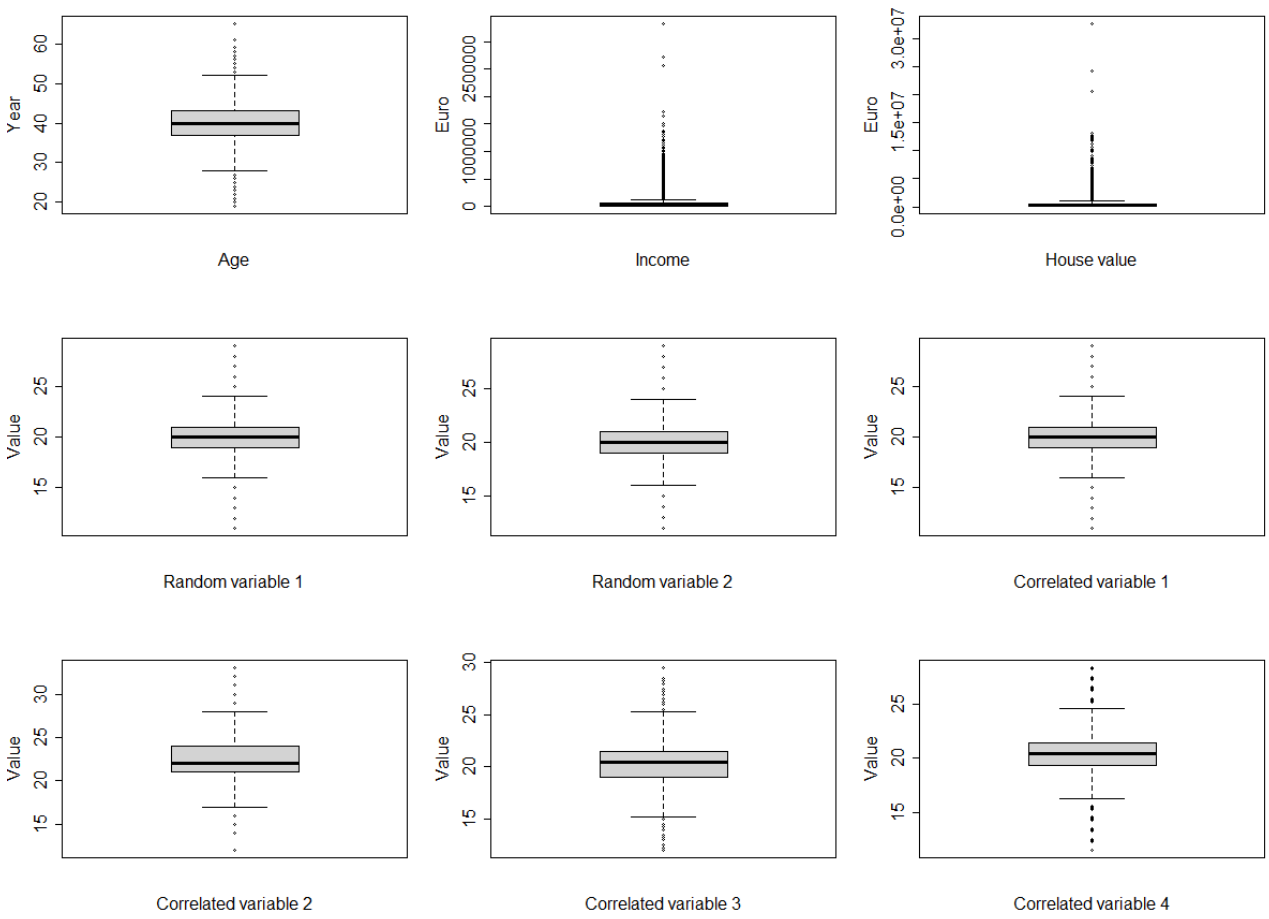

Figure 11: Box plot for numeric data

Table 15: Distribution of numeric variables

Feature	Distribution
Age	N(40,5)
Income	N(49625,65765)
House value	N(403945,535212)
Random variable 1	N(20,2)
Random variable 2	N(20,2)
Corr variable 1	N(20,2)
Corr variable 2	N(22,3)
Corr variable 3	N(20,2)
Corr variable 4	N(21,2)

Table 16: Proportions of the categorical variables

Feature	Proportions
Ethnicity	D: 0.56, G: 0.07, F: 0.06, B: 0.09, O: 0.06
Education	1: 0.20, 2: 0.20, 3: 0.40, 4: 0.20
Gender	M: 0.70, F: 0.30
Income category	1:0.34, 2: 0.27, 3: 0.14, 4: 0.12, 5: 0.11
Family size	1: 0.40, 2: 0.30, 3: 0.20, 4: 0.10

Table 17: Correlation with Income

	Correlation
House value	0.989
Income category	0.736

Table 18: Correlation with Age

	Correlation
Corr variable 1	-0.007
Corr variable 4	0.024

Table 19: Correlation with Ethnicity

	Correlation
Corr variable 1	-0.001
Corr variable 2	0.447

Table 20: Correlation with Gender

	Correlation
Corr variable 3	0.109

Table 21: Standard deviations of numeric variables

	σ
Age	4.993
Income	65764.530
House value	535211.700
Random var 1	2.027
Random var 2	2.019
Corr var 1	2.022
Corr var 2	2.255
Corr var 3	2.031
Corr var 4	2.021

Table 22: Mean of the fitted values of the simulation models

	μ
Model 1	102.109
Model 2	14.572
Model 3	44.073
Model 4	7.063

Correlated variable 1 ($cv_{i,1}$) is dependent on age (a_i) and ethnicity (t_i):

$$cv_{i,1} = u_i + I_{t_i='Finnish'} + I_{a_i<35},$$
$$u_i \sim N(20,2).$$

Correlated variable 2 ($cv_{i,2}$) is dependent on ethnicity:

$$cv_{i,2} = u_i + l(t_i),$$
$$u_i \sim N(20,2).$$

$$l(t_i) = \begin{cases} 1 & \text{if } t_i = \text{'Belgian'} \\ 2 & \text{if } t_i = \text{'Dutch'} \\ 3 & \text{if } t_i = \text{'Finnish'} \\ 4 & \text{if } t_i = \text{'German'} \\ 5 & \text{if } t_i = \text{'Other'} \end{cases}$$

Correlated variable 3 ($cv_{i,3}$) is dependent on gender (g_i):

$$cv_{i,3} = u_i + 0.5 * I_{g_i='Man'},$$
$$u_i \sim N(20,2).$$

Correlated variable 4 ($cv_{i,4}$) is dependent on age:

$$cv_{i,4} = u_i + 0.01 * a_i,$$
$$u_i \sim N(20,2).$$

C Results

C.1 Simulation model 1

Table 23: Regular direct differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	0.000 %	0.000 %
Age	0.049 %	-0.017 %	0.032 %
Ethnicity	0.000 %	0.000 %	0.000 %
Education	0.000 %	0.000 %	0.000 %
Income	0.000 %	-0.024 %	-0.023 %
Income category	0.000 %	0.000 %	0.000 %
House value	0.000 %	-0.031 %	-0.031 %
Family size category	0.000 %	0.000 %	0.000 %
Random var 1	3.963 %	-0.019 %	3.942 %
Random var 2	5.915 %	-0.014 %	5.901 %
Corr var 1	0.000 %	-0.012 %	-0.012 %
Corr var 2	0.000 %	-0.023 %	-0.023 %
Corr var 3	0.000 %	0.015 %	0.015 %
Corr var 4	0.000 %	-0.004 %	-0.004 %

Table 24: Absolute direct differentiation statistics

	Direct	Indirect	Total
Gender	0.207 %	0.026 %	0.226%
Age	0.050 %	0.024 %	0.074 %
Ethnicity	0.129 %	0.019 %	0.148 %
Education	0.429 %	0.027 %	0.456 %
Income	0.000 %	0.029 %	0.029 %
Income category	0.000 %	0.017 %	0.017 %
House value	0.000 %	0.057 %	0.057 %
Family size category	0.000 %	0.581 %	0.581 %
Random var 1	3.963 %	0.020 %	3.983 %
Random var 2	5.915 %	0.012 %	5.927 %
Corr var 1	0.000 %	0.018 %	0.018 %
Corr var 2	0.000 %	0.033 %	0.033 %
Corr var 3	0.000 %	0.026 %	0.026 %
Corr var 4	0.000 %	0.024%	0.024 %

Table 25: Regular indirect differentiation statistics

	Gender	Age	Ethnicity	Education	Random var 1	Random var 2
Gender	-	0.000 %	0.000 %	0.000%	0.000 %	0.000 %
Age	0.000 %	-	0.000 %	0.000 %	0.004 %	-0.021%
Ethnicity	0.000%	0.000 %	-	0.000 %	0.000 %	0.000%
Education	0.000 %	0.000 %	0.000%	-	0.000 %	0.000 %
Income	0.000 %	0.000 %	0.000 %	0.000 %	-0.014%	-0.010%
Income category	0.000 %	0.000 %	0.000 %	0.000 %	0.000%	0.000%
House value	0.000%	0.000 %	0.000 %	0.000 %	-0.009%	-0.022%
Family size category	0.000 %	0.000 %	0.000 %	0.000 %	0.000%	0.000%
Random var 1	0.000 %	0.000 %	0.000 %	0.000 %	-	-0.019%
Random var 2	0.000%	0.000 %	0.000 %	0.000 %	-0.012%	-
Corr var 1	0.000%	0.000 %	0.001 %	0.000 %	0.002 %	-0.016%
Corr var 2	0.000%	0.000 %	0.006 %	0.000 %	-0.009 %	-0.014%
Corr var 3	0.010%	0.000 %	0.000 %	0.000 %	-0.005%	0.010%
Corr var 4	0.000%	0.001 %	0.000 %	0.000 %	0.005%	-0.02%

Table 26: Absolute indirect differentiation statistics

	Gender	Age	Ethnicity	Education	Random var 1	Random var 2
Gender	-	0.000%	0.000%	0.000%	0.007%	0.019%
Age	0.000%	-	0.000%	0.000%	0.004%	0.020%
Ethnicity	0.000%	0.000%	-	0.000%	0.005%	0.014%
Education	0.000%	0.000%	0.000%	-	0.009%	0.016%
Income	0.000%	0.000%	0.000%	0.001%	0.014%	0.014%
Income category	0.000%	0.000%	0.000%	0.000%	0.007%	0.010%
House value	0.000%	0.001%	0.000%	0.001%	0.017%	0.038%
Family size category	0.000%	0.000%	0.000%	0.001%	0.020%	0.037%
Random var 1	0.000%	0.001%	0.000%	0.001%	-	0.018%
Random var 2	0.000%	0.000%	0.000%	0.000%	0.012%	-
Corr var 1	0.000%	0.000%	0.000%	0.000%	0.002%	0.016%
Corr var 2	0.000%	0.000%	0.009%	0.001%	0.009%	0.014%
Corr var 3	0.010%	0.000%	0.000%	0.001%	0.005%	0.010%
Corr var 4	0.001%	0.000%	0.000%	0.000%	0.005%	0.018%

C.2 Simulation model 2

Table 27: Regular differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	0.000 %	0.000 %
Age	2.775 %	0.006 %	2.781 %
Ethnicity	0.000 %	0.000 %	0.000 %
Education	0.000 %	0.000 %	0.000 %
Income	0.000 %	-0.033 %	-0.033 %
Income category	0.000 %	0.000 %	0.000 %
House value	0.000 %	-0.049 %	-0.049 %
Family size category	0.000 %	0.000 %	0.000 %
Random var 1	2.740 %	0.001 %	2.741 %
Random var 2	0.176 %	-0.015 %	0.161 %
Corr var 1	-0.910 %	-0.019 %	-0.929 %
Corr var 2	0.000 %	0.007 %	0.007 %
Corr var 3	0.000 %	-0.004 %	-0.004 %
Corr var 4	0.000 %	0.064 %	0.064 %

Table 28: Absolute differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	0.010 %	0.010 %
Age	2.775 %	0.018 %	2.793 %
Ethnicity	0.000 %	0.051 %	0.051 %
Education	0.000 %	0.038 %	0.038 %
Income	0.000 %	0.039 %	0.039 %
Income category	0.000 %	0.021 %	0.021 %
House value	0.000 %	0.068 %	0.068 %
Family size category	0.000 %	0.018 %	0.018 %
Random var 1	2.740 %	0.004 %	2.744 %
Random var 2	0.176 %	0.023 %	0.198 %
Corr var 1	1.690 %	0.020 %	1.710 %
Corr var 2	0.000 %	0.089 %	0.089 %
Corr var 3	0.000 %	0.014 %	0.014 %
Corr var 4	0.000 %	0.080 %	0.080 %

Table 29: Regular indirect differentiation statistics

	Age	Random var 1	Random var 2	Corr var 1
Gender	0.000 %	0.000 %	0.000 %	0.000 %
Age	-	0.002 %	-0.001 %	0.005 %
Ethnicity	0.000 %	0.000 %	0.000 %	0.000 %
Education	0.000 %	0.000 %	0.000 %	0.000 %
Income	-0.026 %	-0.009 %	0.000 %	0.002 %
Income category	0.000 %	0.000 %	0.000 %	0.000 %
House value	-0.044 %	-0.007 %	-0.001 %	0.003 %
Family size category	0.000 %	0.000 %	0.000 %	0.000 %
Random var 1	0.002 %	-	-0.001 %	0.000 %
Random var 2	-0.009 %	-0.009 %	-	0.003 %
Corr var 1	-0.019 %	0.001 %	0.001 %	-
Corr var 2	0.011 %	-0.007 %	0.000 %	0.003 %
Corr var 3	-0.001 %	-0.004 %	0.000 %	0.001 %
Corr var 4	0.067 %	0.003 %	-0.001 %	-0.005 %

Table 30: Absolute indirect differentiation statistics

	Age	Random var 1	Random var 2	Corr var 1
Gender	0.001 %	0.004 %	0.001 %	0.004 %
Age	-	0.002 %	0.001 %	0.016 %
Ethnicity	0.021 %	0.004 %	0.004 %	0.022 %
Education	0.007 %	0.007 %	0.000 %	0.024 %
Income	0.026 %	0.009 %	0.000 %	0.004 %
Income category	0.013 %	0.005 %	0.000 %	0.003 %
House value	0.044 %	0.012 %	0.001 %	0.005 %
Family size category	0.002 %	0.013 %	0.001 %	0.002 %
Random var 1	0.002 %	-	0.001 %	0.001 %
Random var 2	0.009 %	0.009 %	-	0.005 %
Corr var 1	0.019 %	0.001 %	0.000 %	-
Corr var 2	0.011 %	0.007 %	0.000 %	0.071 %
Corr var 3	0.008 %	0.004 %	0.000 %	0.002 %
Corr var 4	0.067 %	0.003 %	0.001 %	0.009 %

C.3 Simulation model 3

Table 31: Regular differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	0.000 %	0.000 %
Age	0.000 %	-0.443 %	-0.443 %
Ethnicity	0.000 %	0.000 %	0.000 %
Education	0.000 %	0.000 %	0.000 %
Income	0.000 %	-0.356 %	-0.356 %
Income category	0.000 %	0.000 %	0.000 %
House value	0.000 %	-0.574 %	-0.574 %
Family size category	0.000 %	0.000 %	0.000%
Random var 1	0.000 %	-0.054 %	-0.054 %
Random var 2	0.000 %	-0.022 %	-0.022 %
Corr var 1	4.582 %	-0.164 %	4.418 %
Corr var 2	10.232 %	-0.021 %	10.211 %
Corr var 3	13.589%	0.068 %	13.657%
Corr var 4	-18.345 %	- 0.004 %	-18.341 %

Table 32: Absolute differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	1.481%	1.481%
Age	0.000 %	0.539 %	0.539 %
Ethnicity	0.000 %	3.544 %	3.544 %
Education	0.000 %	0.234 %	0.234 %
Income	0.000 %	0.368 %	0.368 %
Income category	0.000 %	0.296 %	0.296 %
House value	0.000 %	0.616 %	0.616 %
Family size category	0.000 %	0.055 %	0.055 %
Random var 1	0.000 %	0.265 %	0.265 %
Random var 2	0.000 %	0.467 %	0.467 %
Corr var 1	4.582 %	0.203 %	4.785 %
Corr var 2	10.232 %	0.238 %	10.470 %
Corr var 3	13.589%	0.080 %	13.669 %
Corr var 4	18.345 %	0.060 %	18.405 %

Table 33: Regular indirect differentiation statistics

	Corr var 1	Corr var 2	Corr var 3	Corr var 4
Gender	0.000 %	0.000 %	0.000 %	0.000 %
Age	-0.027 %	0.031 %	-0.004 %	-0.443 %
Ethnicity	0.000 %	0.000 %	0.000 %	0.000 %
Education	0.000 %	0.000 %	0.000 %	0.000 %
Income	-0.008 %	-0.089 %	-0.235 %	-0.025 %
Income category	0.000 %	0.000 %	0.000 %	0.000 %
House value	-0.014 %	-0.106 %	-0.290 %	-0.164 %
Family size category	0.000 %	0.000 %	0.000 %	0.000 %
Random var 1	0.002 %	0.001 %	-0.018 %	-0.039 %
Random var 2	-0.013 %	-0.020 %	-0.001 %	0.056 %
Corr var 1	-	-0.040 %	-0.018 %	-0.096 %
Corr var 2	-0.015 %	-	0.042 %	-0.006 %
Corr var 3	-0.006 %	0.031 %	-	0.043 %
Corr var 4	0.025 %	0.003 %	-0.032 %	-

Table 34: Absolute indirect differentiation statistics

	Corr var 1	Corr var 2	Corr var 3	Corr var 4
Gender	0.010 %	0.030 %	1.380 %	0.061 %
Age	0.044 %	0.031 %	0.022 %	0.443 %
Ethnicity	0.059 %	3.379 %	0.034 %	0.073 %
Education	0.007 %	0.087 %	0.068 %	0.072 %
Income	0.010 %	0.094 %	0.239 %	0.025 %
Income category	0.009 %	0.038 %	0.133 %	0.116 %
House value	0.014 %	0.117 %	0.302 %	0.182 %
Family size category	0.005 %	0.010 %	0.023 %	0.018 %
Random var 1	0.002 %	0.166 %	0.018 %	0.069 %
Random var 2	0.013 %	0.278 %	0.120 %	0.056 %
Corr var 1	-	0.040 %	0.018 %	0.144 %
Corr var 2	0.191 %	-	0.042 %	0.006 %
Corr var 3	0.006 %	0.031 %	-	0.043 %
Corr var 4	0.025 %	0.003 %	0.032 %	-

C.4 Simulation model 4

Table 35: Regular differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	0.000 %	0.000 %
Age	1.838%	-0.026%	1.812%
Ethnicity	0.000 %	0.000 %	0.000 %
Education	0.000 %	0.000 %	0.000 %
Income	15.344 %	1.354 %	16.698 %
Income category	0.000 %	0.000 %	0.000 %
House value	1.396 %	15.020 %	16.416 %
Family size category	0.000 %	0.000 %	0.000 %
Random var 1	0.000 %	-0.030 %	-0.030 %
Random var 2	0.000 %	-0.025 %	-0.025 %
Corr var 1	-1.804 %	-0.110%	-1.914 %
Corr var 2	5.896 %	-0.096 %	5.800 %
Corr var 3	0.000 %	-0.019 %	-0.019 %
Corr var 4	0.000 %	0.144 %	0.144 %

Table 36: Absolute differentiation statistics

	Direct	Indirect	Total
Gender	0.000 %	0.115%	0.115%
Age	1.838%	0.209%	2.039%
Ethnicity	0.000 %	7.897 %	7.897 %
Education	32.042 %	0.277 %	32.0319 %
Income	15.344 %	1.742 %	17.086 %
Income category	0.000 %	9.720 %	9.720 %
House value	1.396 %	15.534 %	16.920 %
Family size category	0.000 %	0.215 %	0.215 %
Random var 1	0.000 %	0.527 %	0.527 %
Random var 2	0.000 %	0.744 %	0.744 %
Corr var 1	5.477 %	0.214%	5.691 %
Corr var 2	23.488 %	0.478 %	23.966 %
Corr var 3	0.000 %	0.478 %	0.478 %
Corr var 4	0.000 %	0.257 %	0.257 %

Table 37: Regular indirect differentiation statistics

	Age	Education	Income	House value	Corr var 1	Corr var 2	Corr var 4
Gender	0.000%	0.000%	0.000%	0.000%	0.000%	0.000 %	0.000 %
Age	-	-0.009%	-0.041%	-0.004%	0.011%	0.018 %	0.000 %
Ethnicity	0.000%	0.000%	0.000%	0.000%	0.000%	0.000 %	0.000 %
Education	0.000%	-	0.000%	0.000%	0.000%	0.000 %	0.000 %
Income	-0.016%	0.046%	-	1.372%	0.003%	-0.051 %	0.000 %
Income category	0.000%	0.000%	0.000%	0.000%	0.000%	0.000 %	0.000 %
House value	-0.028%	0.019%	15.085%	-	0.005%	-0.062%	0.000 %
Family size category	0.000%	0.000%	0.000%	0.000%	0.000%	0.000 %	0.000 %
Random var 1	0.002%	-0.035%	0.004%	0.000%	0.000%	0.000 %	0.000 %
Random var 2	-0.006%	-0.022%	0.014%	0.002%	0.005%	-0.019 %	0.000 %
Corr var 1	-0.013%	-0.024%	-0.045%	-0.005%	-	-0.023 %	0.000 %
Corr var 2	0.006%	-0.079%	-0.026%	-0.002%	0.004%	-	0.000 %
Corr var 3	0.000%	0.043%	-0.077%	-0.006%	0.002%	0.018 %	0.000 %
Corr var 4	0.044%	0.085%	0.021%	0.002%	-0.010%	0.002 %	-

Table 38: Absolute indirect differentiation statistics

	Age	Education	Income	House value	Corr var 1	Corr var 2	Corr var 4
Gender	0.001%	0.012%	0.018%	0.002%	0.012%	0.069 %	0.000 %
Age	-	0.042%	0.041%	0.004%	0.052%	0.070 %	0.000 %
Ethnicity	0.014%	0.018%	0.037%	0.004%	0.070%	7.760 %	0.000 %
Education	0.005%	-	0.060%	0.006%	0.008%	0.195 %	0.000 %
Income	0.017%	0.124%	-	1.372%	0.012%	0.217 %	0.000 %
Income category	0.009%	0.077%	7.244%	2.294%	0.011%	0.087 %	0.000 %
House value	0.030%	0.133%	15.085%	-	0.017%	0.269%	0.000 %
Family size category	0.001%	0.149%	0.033%	0.004%	0.006%	0.022 %	0.000 %
Random var 1	0.002%	0.139%	0.004%	0.001%	0.002%	0.379 %	0.000 %
Random var 2	0.006%	0.047%	0.034%	0.005%	0.015%	0.637 %	0.000 %
Corr var 1	0.013%	0.058%	0.045%	0.005%	-	0.093 %	0.000 %
Corr var 2	0.007%	0.214%	0.026%	0.002%	0.229%	-	0.000 %
Corr var 3	0.001%	0.258%	0.077%	0.006%	0.007%	0.072 %	0.000 %
Corr var 4	0.044%	0.153%	0.021%	0.002%	0.029%	0.007 %	-