Die approbierte Originalversion dieser Diplom-/Masterarbeit ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this diploma or master thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

MASTER THESIS

An Empirical Approach to Risk Modeling in Brownfield Regeneration

A Master's Thesis submitted for the degree of
MASTER of SCIENCE

Under Supervision of
Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Walter S.A. SCHWAIGER, MBA

by
NAGHMEH JAFARI
1227893

Institut für Managementwissenschaften
Finanzwirtschaft und Controlling

E330 Institute of Management Science
Technical University of Vienna, Vienna, Austria
July 2017

Acknowledgments

I would first like to thank my thesis advisor, Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Walter Schwaiger, of the Institute of Management Science at TU Wien, for his guidance, support, and enthusiasm during my thesis and study, and giving me the opportunity to grow and learn as much as possible along the way. I could not have imagined having a better advisor and mentor for my master's study.

My deepest thanks to other colleagues at TU Wien who provided me with useful information and suggestions.

Last but not least, I would like to express my deepest gratitude to my husband, Arash, who has endlessly supported and inspired me all the way through my studies; my parents for providing me with continuous support and encouragement throughout my life, and my sister who is my biggest source of inspiration. This would not have been possible without them.

Thank you.
Naghmeh Jafari

Abstract

Over the last decade, regeneration of derelict and underused sites with varying degrees of contamination (also known as *Brownfield sites*) has gained popularity as a sustainable land use strategy. However, redevelopment of contaminated fields is a complex and multidimensional problem that entails many risks and uncertainties. The objective of this thesis is to construct, calibrate and validate a risk assessment model that can assist investors and decision-makers in evaluating and classifying brownfield sites to two categories : suitable for redevelopment / not suitable for redevelopment. The three-step model building process is adopted from the methodology of credit risk modeling used in banks and credit rating agencies. The proposed models utilize two machine learning algorithms, namely Classification And Regression Trees (CART), and Random Forest algorithms. The first part of the thesis provides a point of reference in brownfield regeneration risk modeling and describes the current research gaps in this field. The following chapter describes the credit risk model building methodology. Finally, Chapter 4 describes the implementation of risk model building methodology in the field of brownfield risk modeling using programming language R. Appendix A includes the commented R-code for interested readers and can serve as a guideline in implementing the Classification And Regression Tree, and Random Forest algorithms in various fields of study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last several decades, extensive de-industrialization and land use changes of former military, industrial, and commercial sites across Europe have resulted in a large number of derelict and underutilized lands with varying degrees of contamination also known as *Brownfields* [37]. On the other hand, high demands for land in and around cities has caused urban sprawl to become one of the major challenges facing Europe [36]. During the last decade, brownfield site remediation and revitalization has gained increasing attention as a sustainable land use strategy to combat urban sprawl [6]. Several brownfield risk assessment tools and prioritization models have been suggested in the literature in order to help stakeholders evaluate the inherent risks involving brownfield regeneration with the main focus on various aspects of it, such as uncertainty assessment, environmental and health risk assessment, remediation cost assessment, etc. [4, 15, 40]. However, all the models are either developed on a case-by-case basis or lack a multidisciplinary approach [15] and further, all fail to assess their predictive power based on the goodness of the model outcomes against realizations of brownfield regeneration in a model-validation step.

One major project that merges most existing models into one and follows a multidisciplinary approach is the **T**ailored **Im**provement of **B**rownfield **R**egeneration in **E**urope **(TIMBRE)**, which assists stakeholders to rank brownfield sites based on their redevelopment potential by using multi-criteria decision analysis methodology by computing a proritization or ranking score, through a hierarchical structure, which includes dimensions, factors and indicators [40]. Until now, however, the validity and

the prediction accuracy of the suggested model has not been reported in the literature. The validation step according to Gass (1983, p.11) refers to "all activities that establish how closely the model mirrors the perceived reality of the model". In this case, validation of the risk model assesses how accurately the prioritization tool forecasts the successful regeneration of brownfield sites. The lack of empirical evaluation of existing models in brownfield regeneration is the missing link in risk modeling studies that needs to be addressed.

The aim of this thesis is to use the TIMBRE scoring model as the reference point, as it is the state of the art in the field of brownfield regeneration risk models, and develop and validate a risk model for brownfield regeneration. By following the three step *Construction-Calibration-Validation* model building process proposed and implemented by Altman (1968), which is now widely practiced in the field of financial risk management, we attempt to bridge the gap between brownfield regeneration scoring models and risk assessment models [1]. Our goal here is to construct new brownfield regeneration risk classification models by using the decision tree analysis methodology and its extension to random forest algorithm, and later validate them with historical data on redevelopment of brownfield sites is Austria. The main aim of scoring and rating models is to act as a classification tool by assigning a score that best separates the "good" candidates (within the scope of this work: successful regeneration of brownfield sites) from the "bad" (not regenerated sites) in a procedure commonly known as *classification*. Assessing the predictive power of the scoring model and its calibration is a major task of the validation step. By incorporating different construction, calibration and validation methods we close the missing link in brownfield regeneration tools, where mostly heuristic and semi-quantitative risk models have been taken into account without validating the model results. The produced models can further serve as a guideline in risk model building using the classification and regression tree and random forest algorithms in different fields of study. The broader aim of using the CART and random forest algorithms within the framework of three-step risk model building in a new field of research is to illustrate the capability of implementation of such risk models in diverse branches of industry and fields of study.

The remaining thesis consists of the following parts. Chapter 2 continues by introducing the existing TIMBRE scoring model as the state of the art in brownfield regeneration scoring models, followed by the existing research gaps and limitations of the TIMBRE model. Chapter 3 offers a broad review of the state of the art and methodologies commonly used in credit risk scoring field. It further describes the tree decision analysis method and its extension, namely the random forest algorithm, in detail. Credit risk modeling process described in Chapter 3 follows the three-step credit risk model-building methodology which is widely practiced in the banking sector, first introduced by Altman (1968) [1]. In Chapter 4, following the three Construction, Calibration, Validation steps, we develop three new classification models in brownfield regeneration, based on the machine learning methods in credit scoring models, namely the decision tree analysis and random forest algorithm with the help of dataset describing brownfield sites in Austria. We follow by assessing the models' predictive power and accuracy and provide a brief comparison of the developed classification models. Finally, Chapter 5 summarizes the thesis findings and conclusions of the thesis. The commented R-code used for the three-step model building process is further attached in Appendix A to guide interested readers and researchers with the implementation of the CART and random forest algorithms within the risk model building framework.

The extensive dataset used in this thesis is gathered and provided by the Federal Environmental Agency of Austria (Umweltbundesamt - UBA) and Statistik Austria for the ENTEKER *(ENTwicklung Eines Kostenlosen ERkundungsservice)* project, funded by Klima - und Energiefonds within the framework of *"SMART CITIES - FIT for SET"* program. All the statistical computing within this thesis is performed with the aid of programming language R.

Chapter 2

TIMBRE Brownfield Prioritization Model

For decades, there has been a trend toward de-industrialization of industrial, military and mining sites across Europe and North America, which has lead to large number of derelict and unused sites with real or perceived degrees of contamination. Such sites are commonly referred to as *Brownfield Sites* in the literature [37]. As of 2013, the European Environment Agency (EEA) estimated that there are up to 3 million brownfield sites across Europe, mostly located within urban boundaries [12], while the United States Environmental Protection Agency (EPA) estimated somewhere between 500,000 to 1 million brownfields, typically in urban areas [2]. Redevelopment of brownfield sites has become a common practice within the last decade, particularly in areas close to city centers [15]. The reason on one hand is free and undeveloped sites (also called *greenfields*) are limited and becoming more and more scarce and on the other hand, environmental policies, particularly in Europe, encourage the regeneration processes through various grant systems.

However, the inherent risks involved with revitalization of brownfield sites necessitates the use of risk assessment tools that help to evaluate the various aspects of the uncertainties. Some notable brownfield regeneration risks include: risk of liability claims, investment, usability, and marketability risk and stigma surrounding contaminated fields. Such risks affect the potential revitalization of brownfield sites from various fronts such as environmental, social, and economic levels [4]. The following sections provides a brief overview of existing risk assessment models and tools.

2.1 Brownfield Regeneration Background

The first step towards developing a brownfield risk assessment model is to identify factors that determine successful brownfield regeneration. Thornton, Franz, Edwards, Pahlen, and Nathanail (2007) investigate the advantages and deficiencies of the current financial, fiscal, legal, regulatory and policy incentives regarding the sustainable brownfield redevelopment and find that the incentives alone to be only partially effective in the brownfield regeneration process [46]. Dixon (2007) analyses the importance of property development industry in the sustainable regeneration of brownfield sites. His findings show that as much as the attitude of the property development industry affects the process, it is significantly more critical to interact with various stakeholders in order to achieve successful regeneration of brownfield sites [16]. Frantal, Kunc, Klusacek, and Martinat (2015) aim to identify and classify success factors by conducting an international comparative survey [19]. They find that there are "common themes" that drive the successful regeneration of brownfield sites, such as site, local, and economic factors , but any "universal" solution might be too general to function on an international level. Based on the identified success factors, prioritization tools and methodologies have been proposed in the literature. Majority of these tools are based on a case-by-case approach and focus solely on one of the following aspects: health and environment, financial incentives, uncertainty assessment, geology, past or present use, etc. [15].

One assessment tool that adopts a multi-disciplinary approach to brownfield site remediation is the **T**ailored **Im**provement of **B**rownfield **R**egeneration in **E**urope (TIMBRE), which is a European Union funded project within the 7^{th} Framework Programme (FP7) aimed to support the end-users in overcoming existing problems. The projct has resulted in developing a prioritization tool that helps to rank brownfield sites in a portfolio based on their relative redevelopment potential [40]. Following sections present an overview of state of the art TIMBRE prioritization model in the brownfield risk assessment field.

2.2 TIMBRE Model Framework

The TIMBRE prioritization model utilizes a Multi-Criteria Decision Analysis (MCDA) methodology in order *"to assist stakeholders to identify which brownfield sites should be preferably considered for redevelopment or further investigation, taking into account a set of success factors properly identified through a systematic stakeholder engagement procedure"* [40]. The model incorporates the three main pillars of sustainability, i.e. economic, social and environmental dimensions. The risk assessment model is developed by constructing a hierarchical structure in order to enumerate a prioritization or ranking score for the redevelopment potential of a brownfield and calibrating the model with expert-estimated weights. In the following subsections the TIMBRE ranking methodology proposed by Pizzol, Zabeo, Klusáček, Giubilato, Critto, Frantál, Martinát, Kunc, Osman, and Bartke (2016) is described further in detail.

2.2.1 TIMBRE Model-Building Methodology

The construction of the TIMBRE model is based on a hierarchical structure, depicted in Figure 1, that includes dimensions, factors, indicators, and their respective weights. On the highest level, **dimensions** account for specific aspects of brownfield redevelopment potential, such as *"local development potential, site attractiveness and marketability, environmental risks, and/or other specific criteria"* [40]. Each dimension is explained through one or several **factors** that characterize conditions, circumstances, actors, etc. that significantly affect the successful regeneration of a brownfield site. On the lowest level, **indicators** are used to quantify the factors through measurable (continuous, discrete, ordinal, and/or categorical) variables. Furthermore, weights, determined by experts, are assigned to each of the dimensions, factors, and indicators, denoting their relative importance.

After assigning the relative weights by the experts, indicators are aggregated into fac-

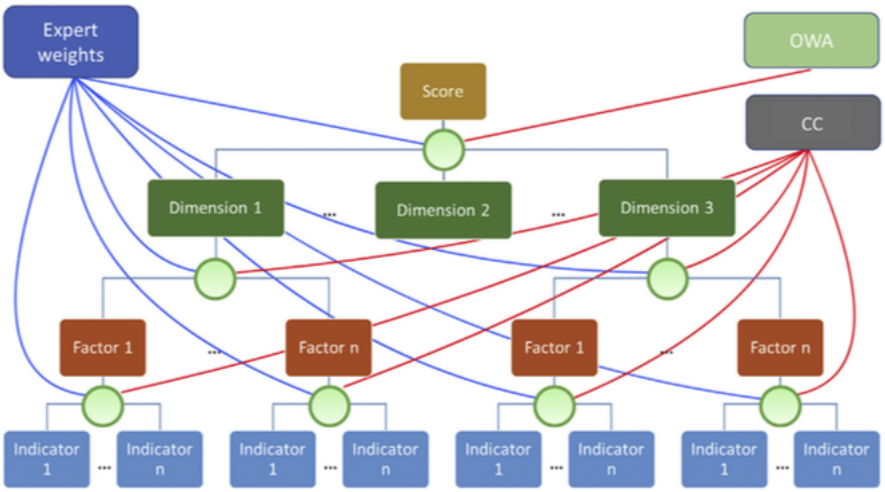

Figure 1: Hierarchical structure at the basis of the MCDA methodology applied in the TIMBRE. CC stands for Convex Combination and OWA for Ordered Weighted Average [40].

tors by the convex combination, which is a specific type of weighted average where the sum of all weights are equal to 1, which are then aggregated once again into dimensions with respect to their relative weights, and finally the dimensions are aggregated to a final *prioritization score* representing the redevelopment potential of a brownfield.

2.2.2 Identification of Relevant Variables

A significant step in the TIMBRE prioritization model is identifying the essential dimensions, factors, and indicators that play a role in successful regeneration of brownfield sites. However, the model does not require a predetermined set of variables, and takes the following points into account when identifying the model variables:

1. Explanatory variables proposed in the literature from previous studies, projects, interviews and surveys with stakeholders and experts from different countries and regions, and statistical data analysis;

2. Availability and comparability of data from existing databases, inventories, registers of brownfield sites and other statistical databases;
3. Measurability of data of known quality, updated at regular intervals in accordance with reliable procedures.

The above-mentioned points allow the model a certain degree of freedom and thus, render a more user-friendly assessment tool. However, this flexibility can lead to considerable variations in model outputs, depending on the variables used. This lack of objectivity in variable selection process could affect the performance of the model.

2.2.3 Normalization of Selected Variables

The selected variables, discussed in the previous section, can be of different orders of magnitude. In order to allow a sensible comparison and usage of data, the indicators and variables need to be comparable, which can be achieved by re-scaling the variables into a common numerical range. This procedure is commonly referred to as *Normalization* [50]. The TIMBRE approach proposes the closed interval $[0, 1]$ as the normalization domain for all variables. That includes all **numerical**, **ordinal**, and **categorical** data. Before describing normalization methods for each data type, the difference between the three data types need to be shortly explained. The value for each ordinal or categorical variable is selected from a finite number of categories. While **ordinal** variables represent a numerical increase or decrease between their discrete values by ordering the values, different categories of categorical variables do not represent any specific numerical significance. An instance for an ordinal variable can be the level of pain a patient is feeling, scaled from 1 to 10. Here, the order of values matter and not the actual value or the difference between the values. **Categorical** data can have two or more classes that do not have any inherent ordering. An example of categorical data is gender with two classes (male and female) with no intrinsic ordering. **Numerical** data, as can be expected, represent actual numbers (discrete or continuous) where both the actual value and difference between the values have

significant meaning, for example, area of a site in $[m^2]$.

The normalization method used for numerical data varies from the method utilized for ordinal and/or categorical variables. For ordinal and categorical data, the normalization is performed by an expert through associating a value from the closed interval $[0, 1]$ to each category of data. The category that most likely leads to the successful regeneration of a brownfield site is set equal to 1, and the least likely category equal to 0, all other categories are given values based on the perception of the expert with regard to the successful brownfield regeneration. Two or several categories can share the same value.

Normalization of numerical data is slightly more complex and can be achieved in two steps. In the first step, experts indicate whether the variable has a descending or ascending effect on the successful brownfield regeneration. An ascending relationship means that an increase in the value of the variable leads to a higher potential for successful regeneration of brownfield, and a descending relationship implies that with the increase of the variable, the success potential decreases. After assessing the relationship by experts, the variable is normalized by utilizing a linear interpolation between its minimum and maximum values. The normalization functions can be described as:

$$X_i = \begin{cases} \dfrac{x_i - i_{min}}{i_{max} - i_{min}}, & \text{ascending} \\ \dfrac{i_{max} - x_i}{i_{max} - i_{min}}, & \text{descending} \end{cases} \tag{1}$$

where x_i and X_i denote the value of variable i before normalization and after normalization, respectively. i_{min} and i_{max} further represent the minimum and maximum values of variable i. The described normalization process is straight-forward and easy to implement. However, similar to the previous step, the need for experts' assessment of normalization of categorical and ordinal variables, as well as the ascending or descending influence of the variable on brownfield regeneration potential can lead

to variations in model outputs, due to subjectivity of experts' judgment.

2.2.4 Aggregation of Variables into a Final Ranking Score

After normalizing the selected indicators, the last step in computing the final brownfield prioritization score can be performed. As described in Section 2.2.1, the TIMBRE model consists of a hierarchical structure, illustrated in Figure 1, with the measurable indicators at the lowest level of the structure. This final step relies on expert estimations for the relative importance of the indicators that fall under the same factor in the hierarchical structure. These estimations are given as weights to each of the indicators, such that each weight must be in the $[0, 1]$ closed interval, and sum of all weights for each factor should be exactly equal to 1. A convex combination of the indicators results into the factor they describe with a value between 0 and 1. The relation is described in Eq. 2.

$$f_i = \sum_{j \in f_i} w_j X_j \tag{2}$$

where w_j is the relative weight of the normalized variable X_j and f_i is the resulting factor. The convex combination of each factor only includes the indicators that describe that specific factor. The same procedure is then executed for all factors.

Similarly, the same process is utilized again in order to compute the values of dimensions d_k by aggregating the factors belonging to each one through a convex combination, weighted by experts based on their relative importance through W_l.

$$d_k = \sum_{l \in d_k} W_l f_l \tag{3}$$

At this point, the values of each dimension represents the prioritization score of brownfield sites in the dimension's specific aspect, e.g., local development potential,

site attractiveness, or environmental risk. Nevertheless, since the purpose of the TIMBRE methodology is to develop a risk assessment model based on Multi-Criteria Decision Analysis (MCDA), the dimensions need to be aggregated again into one final ranking score that incorporates all the dimensions. In their paper, Pizzol et al. (2016) propose two methods to compute the final ranking score [40]. The first approach is to use the convex combination once again with the expert-estimated weights for each dimension. The second approach is referred to as *Ordered Weighted Average (OWA)*, first introduced by Yager in 1988 [49]. The aggregation technique, first, orders the dimensions based on their importance in a descending order. Using a set of predefined weights, ω_i, the ordered dimensions are then aggregated to a final score. The predefined weights for the TIMBRE model is selected such that as the value to be aggregated decreases, its weight is divided by two. As before, each weight must be in the $[0, 1]$ closed interval, and the sum of all weights must be exactly equal to 1, i.e. for a model with three dimensions, the predefined weights are: $\omega_1 = 0.571$, $\omega_2 = 0.286$, and $\omega_3 = 0.143$. By incorporating the predefined weights, ω_i, with the weights estimated by experts for each dimension based on their relative importance, w_i, a new set of weights can be calculated such that:

$$W_i = w_i.\omega_i \tag{4}$$

$$W_i' = \frac{W_i}{\sum W_i} \tag{5}$$

Eq. 5 is utilized in order to normalize the weights and forcing their sum to be equal to one. Finally, using the weights W_i', the prioritization score, representing the redevelopment potential of a brownfield site can be computed using Eq. 6.

$$S = \sum_{i=1}^{n} d_i.W_i' \tag{6}$$

where d_i denotes the value of i^{th} dimension of the brownfield site. Since all indicator,

factor, and dimension values as well as their respective weights are between 0 and 1, the final ranking score will also be a number in the $[0, 1]$ closed interval. The higher the brownfield prioritization score is, the more likely it is for the brownfield site to be successfully redeveloped in the future. The aim of the TIMBRE prioritization tool is to assess the brownfield sites in a portfolio and rank the sites based on their relative attractiveness using a score value. The model does not define a cut-off value for redevelopment of the brownfield sites, such that if the score is higher than the cut-off value, the site is suitable for redevelopment and if not, the site should not be considered for redevelopment. Instead, TIMBRE model provides a ranking of sites relative to one another in a portfolio of brownfield sites at hand.

Up to now, the hierarchical structure of the TIMBRE brownfield prioritization model and its properties have been described as the state of the art in brownfield risk assessment modeling. The following section briefly describes the limitations of the model and the research gaps in the existing literature related to the TIMBRE methodology, and brownfield assessment models in general, that need to be addressed.

2.3 Research Gaps and Limitations of TIMBRE Model

The TIMBRE prioritization tool makes a significant contribution to the previous literature in brownfield regeneration risk assessment models by adopting a Multi-Criteria Decision Analysis (MCDA) approach and summarizing the findings of previous studies, surveys and interviews with stakeholders from various regions in one single model [40, 50]. Moreover, the method proposed is user-friendly with a straight-forward implementation and easy interpretation of model outcomes. The assessment tool further avoids setting strict requirements on model inputs (dimensions, factors, and indicators, as well as their respective weights). Such characteristics allow the model to have the flexibility to be utilized by a wide range of end users form field owners, to land developers, or even government agencies.

The TIMBRE model is, however, susceptible to several issues in different model building steps. In the process of the TIMBRE model building, variable identification and selection is set as the foundation, upon which the model structure is built later. The model's authors take data availability and measurability, as well as personal judgment of stakeholders and experts into account and therefore, do not require a fixed, predetermined set of variables to be used in the model. The problem with such a variable selection method is that no statistical measures are used to test the significance of relationship between the predictor variables the response variable. For instance, among a large dataset of available variables, not all variables necessarily demonstrate a significant effect on the output variable, which is brownfield regeneration potential. These relationships need to be tested before the variables are selected to be used in the model. Some methods of variable selection include the Gini Coefficient, Akaike Information Criterion (AIC), as well as the Bayesian Information Criterion (BIC) [7, 14]. Lack of a statistical test for variable selection in TIMBRE model is regarded as a research gap in the existing model building methodology.

Moreover, each indicator, factor, and dimension receives a relative weight estimated by experts based on their respective importance for the hierarchical structure of the

model. The final brownfield prioritization score depends on the values of the weights, and is therefore susceptible to experts subjectivity. Depending on the areas of expretise of the decision-makers, widely different weights can be assigned to the variables, which can change the model outcomes as a result. Here again, the model building process needs to include statistical tests such as Maximum Likelihood Estimation in order to estimate model coefficients accurately and objectively [30].

Last but not least, until now the evaluation of model's predictive power has not been reported in any study. The model performance evaluation is achieved through validation of model outcomes against realizations of successful brownfield regeneration. According to Gass (1983) validation refers to *"all activities that establish how closely the model mirrors the perceived reality of the model"* [21]. In this case, TIMBRE model validation assesses how accurately the prioritization tool forecasts the successful regeneration of brownfield sites. Pizzol et al. (2016) perform four case-studies of the TIMBRE model in the Czech Republic, Germany, Poland and Romania to rank the brownfield sites listed in the regions' databases [40], but fail to test the model outputs against realizations of brownfield redevelopment with an in-sample or out-of-sample validation. The validation step is essential for achieving and later maintaining a certain level of prioritization and predictive quality over time [32]. The lack of validation measures for the existing model is another research gap in brownfield regeneration risk modeling studies that should be addressed.

The following chapter continues with an overview of credit risk model building methodology. Since one of the major activities of banks and credit rating agencies is to classify their customers based on their creditworthiness in order make decisions whether to grant or reject loans to them, a variety of classification models have been proposed in the literature, and many have been used and continuously improved in practice over time in the banking sector. Thus, we use this opportunity and refer to credit risk model building methodology in order to obtain a generic classification model methodology that can be implemented in a new field of research, which is brownfield regeneration risk assessment.

Chapter 3

Credit Risk Modeling

This chapter is concerned with the credit risk model building methodology used in the banking industry and credit rating agencies as a cornerstone of their risk management practice. Since one of commercial banks' main business activities is to grant loans to individuals and businesses, they depend on credit scoring and classification techniques to assess the credit risk, depending on the type of borrower (private individual, Small and Medium sized Enterprise, corporate, etc.). There are several advantages to using credit scoring models: First, credit scoring models provide an objective assessment and evaluation of credit risk, since they are based on statistical techniques and not opinions. Second, the credit scoring models can be validated, which allows banks and credit rating agencies to assess the accuracy snd predictive power of their models and predictions. Furthermore, continuous validation with the help of new data collected over time allows the model to maintain a certain level of quality over its life cycle.

In the following sections, a brief overview of relevant research literature in the banking sector is presented that includes several commonly used methods and techniques in this field, followed by the methodology used in the three-step process-based credit risk model building (*Model-Construction, Model-Calibration, Model-Validation*), which was first proposed and implemented by Altman (1968) and is now widely practiced in the industry. After providing a brief description of most commonly used statistical models, we focus on two machine learning techniques in this thesis, i.e. decision tree analysis and random forest machine learning algorithms.

3.1 Review of the Relevant Research Literature

The concept of using statistical techniques for credit risk scoring originated in the 1930s and 1940s when Fisher (1936) first introduced linear discriminant analysis in an effort to discriminate between two groups in a population [17]. Prior to that, banks and financial institutes relied on credit analysts to make decisions for credit management purposes solely based on their subjective judgment. However, as the number of private individuals and businesses applying for loans and credit cards increased, the need for an automated and standardized procedure free of personal subjectivity of an analyst became apparent. William Fair and Earl Isaac founded the first consultancy firm in 1956 to assist banks and financial companies measure the creditworthiness of their customers based on divergence statistics [18].

Over the last several decades, various quantitative methods and techniques have been proposed for credit risk scoring and classificaion in the literature based on the concept of classification of several groups in a data sample [43]. These methods can be divided into parametric and non-parametric or data mining models. The construction step of parametric models involves developing a mathematical formalization with explicit functional specification based on sound assumptions as the theoretical foundation of the model. Non-parametric models on the other hand, do not rely on mathematical relations based on assumptions made about the type of mapping functions. Instead, they use training datasets in order to find less structured, data-driven relations among the observations in a dataset [26]. In the following section, some of the most commonly used methods in credit risk scoring literature are discussed briefly.

Linear Discriminant Analysis (LDA)

Beaver (1966) proposed using a uni-variate model in order to discriminate between failing and non-failing companies based on a single financial factor [5]. Altman (1968) further developed the concept by using a set of financial ratios in multivariate discriminant analysis (MDA) in order to classify businesses to two groups (failing/non-failing)

and produced the so-called *z-score* [1], described as:

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p \tag{7}$$

where Z represents the *z-score*, α is the intercept, X is the matrix of explanatory variables, and β is the vector of coefficients. The concept behind the linear discriminant analysis is basically to find a linear combination of explanatory variables that best separates the subsets (or classes of data) within a dataset. In the case of bankruptcy prediction, for example, the goal is to separate failing from non-failing obligors by finding a linear combination of explanatory variables such that the difference in the means of the two subsets are maximized. Initially, the following assumption were made for the analysis (1) the covariance matrices must be equal for both subgroups in the data. (2) Each classification subgroup must be normally distributed. These assumptions have been relaxed over time as the computation capability has increased dramatically. However, the model can only handle numerical data, and not categorical or ordinal data. Over the years, multivariate discriminant analysis has been widely used in financial sector in order to predict defaults and further used by many researchers.

Although many statistical models have been introduced ever since Altman proposed LDA that are more flexible and require less strict assumptions, the foundation for credit risk model building methodology that has been widely used ever since in practice still remains the one implemented by Altman in 1968. The methodology proposed consists of a process-based structure for credit risk model building, that starts with mathematical model specification as **model-construction**, afterwards, with the help of historical or current data, the key model parameters are estimated within **model-calibration**, and finally, the accuracy of the model is tested in order to evaluate the model quality and potentially improve its predictive power in a **model-validation** step. As the computing power has improved drastically over time, many statistical and machine learning methods have been developed for credit risk modeling that

perform much better than previous mathematical methods including linear discriminant analysis, but the underlying methodology used in credit risk models remain the three-step process proposed by Altman [1].

Linear Regression

The linear regression model is commonly used to investigate the relationship between several explanatory variables and a specific response variable and to find significant explanatory variables related to the response variable by using a linear relationship between the explanatory set of variables, $X = X_1, ..., X_p$, and the outcome variable, Y, as depicted,

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p + \varepsilon \tag{8}$$

where ε represents the random error and β is the vector of coefficients estimated by using the method of ordinary least squares [35]. Many authors have proposed utilizing linear regression models for credit scoring [3, 23, 35]. The model is specially suitable when the response variable is continuous numerical. Since the regression outcomes are not limited to a range, i.e. $x'\beta \in [-\infty, \infty]$, the values cannot be interpreted as a probability.

Logistic Regression

Ohlson (1980) made an extension to the existing linear regression models by using conditional logit models to predict corporate bankruptcy by predicting a binary dependant outcome, $Y = \{y_1, y_2\}$, based on a set of independent variables, $X = \{X_1, ..., X_p\}$ [38]. With the help of Maximum Likelihood Estimation method a linear combination of X is estimated based on logit transformation of Y, such that

$$\pi_i = P(Y_i = y_1 | X_i) = \frac{exp\{X_i\beta\}}{1 - exp\{X_i\beta\}} \tag{9}$$

where $0 \leq \pi_i \leq 1$ represents the probability that instance i belongs to category y_1 conditional to X_i, and β represents the vector of estimated coefficients [30]. Unlike multivariate discriminant analysis, logit regression methodology does not require many restrictive assumptions initially assumed for the linear discriminant analysis. Moreover, categorical data can also be used as dummy variables for each category of data in this method. Since the dependent variable is dichotomous (default/non-default), the use of logit regression is apt for bankrupcy classification. Due to its flexibility and advantages, the logit regression remains as one of the most widely used methodologies in credit scoring to this day [43].

Artificial Neural Networks

An artificial neural network (ANN) is a non-linear machine learning model that has been increasingly used in credit scoring models. The data mining technique attempts to mimic the decision process of the human brain, which functions by sending electronic signals between a large number of neurons [42]. The structure of a network contains one input layer (explanatory variables), one or several hidden layers, and one output layer, consisting of several neurons on each level. Neurons on the input layer receive a certain amount of stimuli from the input variables, process them and subsequently generate output values that are transmitted to the neurons in the following layer. Various types of networks differ based on the number of hidden layers and the activation functions applied to them [35, 48].

k-Nearest Neighbor Classifier

The k-nearest neighbor algorithm is a non-parametric classification model, which examines the similarities between the identified patterns of the training set and the input [25]. Based on the metric chosen, k-nearest neighbor from the input data are chosen, such that they are nearest considering the specified metric. Classification of

a new applicant takes place by finding the class that the majority of its neighbors belong to [48]. The choice of the metric, and the number of nearest neighbors significantly affect the performance of the model [25, 24].

Decision Tree Analysis

Decision tree analysis refers to a non-parametric classification method which uses historical data in order to develop decision rules which have a tree-like structure, as the name suggests [8]. The general aim of this approach is to construct a set of if-then logical conditions so that predictions or classification of instances can be made. There are different types of decision trees, based on the criterion used to build the tress. Generally, all tree models attempt to minimize the impurities in their leaf nodes [8, 22]. The Classification And Regression Tree methodology (CART), first introduced by Breiman, Friedman, Olshen, and Stone (1984) is a type of decision tree that utilizes binary trees and separates a dataset into a finite number of classes by using the Gini index, which makes the model suitable for credit scoring where the task is to classify default and non-default cases contained in the data. The logical relationships derived with the CART method are easy to interpret and quite flexible. Moreover, decision trees are ideal for dealing with big datasets, containing large number of variables, since the method inherently performs variable selection [31].

Random Forests

A single Classification and Regression Tree uses a specific and fixed training set to formulate decision rules that best separate classes of data in the training set. The same rules are then applied to new observations in order to assign a class to each new observation. However, the CART trees can be quite unstable and sensitive to changes in the dataset, such that if the training set used to construct and calibrate the model changes, the resulting decision tree can also drastically change. As an extention to this machine learning algorithm, Ho (1995) proposed developing multiple

trees from randomly selected subsets of the data training sets [27]. Breiman (2001) further expanded the concept of *ensemble learning*, which is referred to a machine learning technique that is based on utilizing a set of individual classifiers which are viewed as "*weak learners*" and combining them to a single, more accurate classifier referred to as a "*strong learner*" with higher predictive power [10, 39].

Breiman (2001) proposed *random forests* based on the concept of generating several classification decision trees and aggregating their outcomes to a final classifier [10]. For construction of the random forest, n_{tree} single maximum (unpruned) decision trees are generated by randomly selecting samples of data out of the original dataset, and m_{try} variables from the matrix of the explanatory variables. The model predicts the response variable by majority vote of the single decision trees for classification and averaging the regression values for regression models. Each time a decision tree is constructed, the remaining observations are used to assess the predictive power of the model by obtaining a measure called out-of-bag (OOB) error estimate [9, 10], which is the aggregate of the single error estimates of the decision trees.

The predictive power of the random forest algorithm is unexcelled among current machine learning algorithms and has gained increasing attention over the past years as a classification and/or regression algorithm. Furthermore, the technique does not overfit the data, as single decision trees tend to do, and easily handles large datasets. The existence of missing values and outliers in the data does not impede the performance of the model, either.

In this thesis, we focus on the CART method and Random Forest algorithm, currently used in commercial banks and credit rating agencies for credit scoring purposes. The following sections explain the steps for building a credit risk scoring model with the help of Classification And Regression Tree methodology and its extension to random forests.

3.2 Credit Risk Model-Building Methodology

This section presents the process of Risk Model Building through *model-construction, model-calibration* and *model-validation*, first proposed and implmented by Altman (1968) in the field of credit risk models and later adopted by the Basel Committe on Banking Supervision (BCBS) [1]. The three step process allows attaining a desired level of predictive quality of the risk model and provides feedback infromation of model performance, which is essential for maintaining its performance quality over its life cycle [32]. The generic model building process discussed here is specified utilizing the Classification And Regression Tree methodology and Random Forest algorithm [8, 9, 10].

3.2.1 Model-Construction

The first step in building a risk assessment model is model-construction. A generic risk model M maps risk factors $X_{i,t}$ to a grade $C_{i,t+h}$ that describes the creditworthiness of an obligor, represented by:

$$M : X_{i,t} \rightarrow C_{i,t+h}. \tag{10}$$

The creditworthiness $C_{i,t+h}$ can represent a relative or absolute risk measure. In the case of classification models, the main function is to act as a sorting tool that best separates the "good" candidates that are liable to pay back their financial obligation from the "bad", who should be rejected due to high probability of defaulting in a *classification* procedure, and thus the model maps risk factors or explanatory variables to a specific class from two or several categories, which the applicant belongs to.

Decision trees are powerful learning techniques that are used for the purpose of classification in many fields of research, including but not limited to credit scoring. In essence, decision trees are a learning method that approximates discrete-valued target functions. The model utilizes binary trees and separates a dataset into a finite num-

ber of classes by constructing a set of if-then logical conditions so that predictions or classification of candidates can be made [20].

Decision Tree Fundamentals

Suppose a dataset contains information of several credit applicants. Each applicant is described through a set of explanatory variables (attributes) $X = \{X_1, X_2, ..., X_n\}$, as well as a binary dependent variable $Y \in \{y_1, y_2\}$, describing the category the applicant belongs to (defaulting/non-defaulting). The algorithm starts at a root node, which contains the dataset with applicants from both classes. By going through all attributes and possible cut-off values, the algorithm attempts to find the binary split, or a so-called *splitting rule*, that best separates the dataset to two subsets with one with most defaulting applicants and the other with most non-defaulting, or in other words, such that the two subsets are as homogeneous as possible. Once the root node is split, the same process takes place for the child nodes in a recursive partitioning procedure until a certain criterion is satisfied for the decision tree [20].

Figure 2 schematically illustrates a simple decision tree flow chart that contains both numerical and categorical explanatory variables, since categorical and ordinal attributes can be handled by the decision trees, as well. For such variables the algorithm splits the sample depending on which elements of the categorical attribute best separate the sample data. For categorical variables, the algorithm goes through all possible combinations of classes of the categorical data in order to find the two subsets of classes that best separate the dataset, meaning that if the variable contains k classes, the algorithm considers all $2^{k-1} - 1$ possible binary splits. For ordinal variable, the tree takes the order into account as well, meaning that an attribute with k states will have $k - 1$ possible binary splits to consider. For numerical variables the tree algorithm finds the best cut-off value c_i and splits the dataset depending on whether the value of the attribute is larger or smaller than c_i [45].

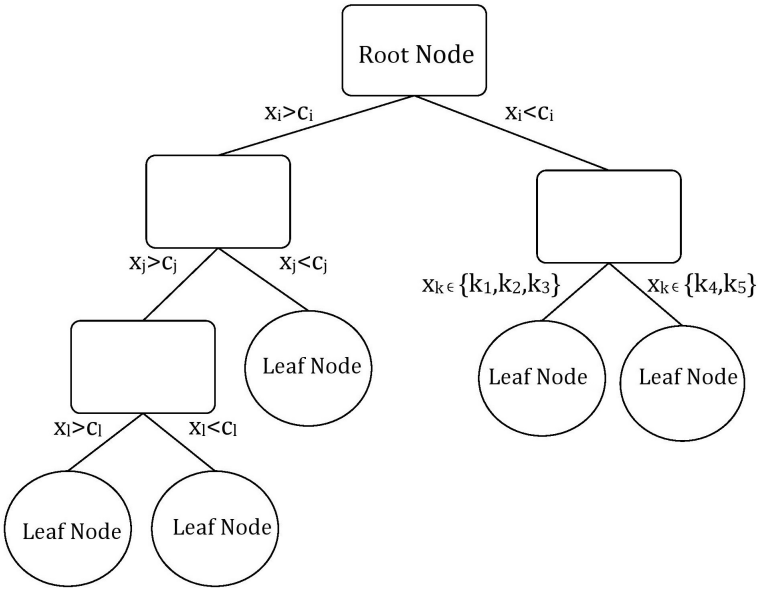

Figure 2: Decision Tree Flow Chart.

The final nodes of the tree, where no more partitioning takes place, are called *leaf nodes*. The algorithm reaches a leaf node when a certain criterion is satisfied. The criterion used to determine the decision or splitting rules is based on the impurity (level of inhomogeneity) of the child nodes versus the parent nodes. Various impurity measures have been proposed to evaluate a decision tree, with the three most commonly used defined as:

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2, \tag{11}$$

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) log_2 p(i|t), \tag{12}$$

$$Classification\ error(t) = 1 - max_i[p(i|t)], \tag{13}$$

such that $p(i|t)$ denotes the fraction of cases belonging to class i at a given node t [45]. In order to determine the splitting rule, a comparison needs to be made between the impurity of a parent node before splitting with the sum of the impurities of the

child nodes after splitting. The gain, Δ, is a function that is used to measure the change in impurity values:

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j) \tag{14}$$

where *I(·)* denotes the impurity measure of a given node, N is the total number of instances at the parent node, k is the number of classifications, and $N(v_j)$ is the number of instances in the child node, v_j. The principle of decision tree algorithms is to determine a splitting condition that maximizes the gain Δ. Since *I(parent)* is the equal for all splitting conditions, maximizing the gain means minimizing the weighted average impurity measures of the child nodes [45]. Figure 3 compares the values of the three impurity measures for a binary classification.

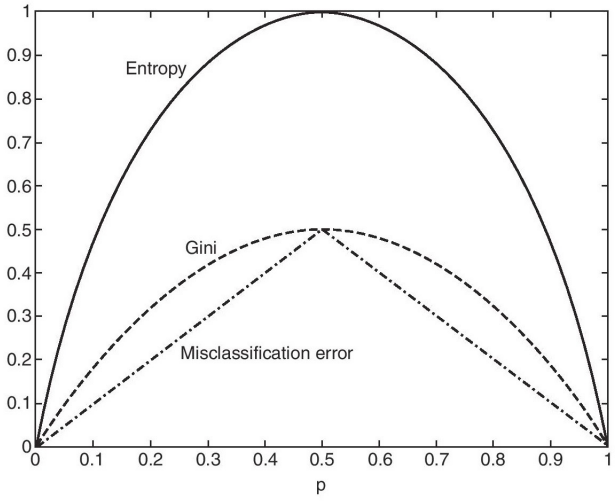

Figure 3: Comparison of impurity measures for binary classification [45].

In Figure 3, p refers to the fraction of applicants belonging to one of the two categories. As can be expected, all three impurity measures reach their maximum value when $p = 0.5$, meaning that there are exactly the same number of records from both categories and thus, the classification algorithm cannot separate the classes at all and the leaf nodes are impure. The minimum values for the measures are attained when

all the records belong to the same category, meaning that p is equal to either 0 or 1 and the tree algorithm has fully separated the classes.

Different types of decision trees are defined depending on the impurity measures used as the criterion for the splitting decision. CART algorithm uses the gini measure (Eq. 11) as the metric used for its splitting decisions. ID3, C4.5 and C5.0 are other type of decision trees that use the entropy measure (Eq. 12) as their splitting function.

Classification And Regression Trees (CART)

Classification and Regression Tree is a type of decision tree used for predictive modelling [8]. One major advantage of the CART algorithm is that the response variable Y, as well as the explanatory variables X can be categorical, nominal, or continuous, making the algorithm incredibly versatile in predictive modeling. If the dependent variable Y is of categorical nature, the tree is referred to as a *Classification Tree* and the model outputs predict discrete values or classes of data, similar to classes of Y. If the model's response variable Y takes continuous numerical value, the model predicts numerical values for the records and the algorithm is referred to as a *Regression Tree*. The splitting criterion used or nominal or categorical dependent variable (classification tree) is the gini index, introduced earlier in Eq. 11. For continuous response variable, the algorithm (regression tree) utilizes Least Squares Deviation (LSD) method as the impurity measure [34]. Another advantage of the CART method is its robustness to outliers. The splitting decision usually isolates outliers in individual node(s). Moreover, the structure of the tree is such that it is invariant to monotone transformations, such as logarithm, or square root of explanatory variables.

Developing a CART tree can be summed up in the following three steps:

1. **Construction of the *maximum tree*.**
2. **Optimizing the the number of nodes by *tree-pruning*.**
3. **Classification of new data with the constructed tree.**

The first step, the construction of the *maximum tree* is the most time consuming step of the model development. In this step, the classification algorithm maximizes the gain function depicted in Eq. 14 for each binary split using the gini index, defined in Eq. 11. Substituting the gini index in the gain function results in:

$$\Delta i(t) = -\sum_{i=0}^{c-1}[p(i|t_p)]^2 + \sum_{j=1}^{k}\frac{N(v_j)}{N}\sum_{i=0}^{c-1}[p(i|t_c)]^2 \tag{15}$$

where t_p and t_c denote parent and child nodes, respectively. Thus, the algorithm needs to solve the following maximization problem at each step :

$$\underset{x_k<x_k^R, j=1,...,p}{\arg\max}\Big[-\sum_{i=0}^{c-1}[p(i|t_p)]^2 + \sum_{j=1}^{k}\frac{N(v_j)}{N}\sum_{i=0}^{c-1}[p(i|t_c)]^2\Big] \tag{16}$$

where x_k refers to the explanatory variables, x_k^R the best splitting or cut-off value of x_k, and $k = 1, ..., p$ is the number of explanatory variables [47]. Other splitting rules have also been proposed in the literature such as Twoing splitting rule. Breiman et al. (1984) show that the maximum tree is insensitive to the choice of the splitting rule criterion, and thus, we use the splitting rule described above. However, tree-pruning procedure, unlike the splitting rule criterion, plays a significant role in the predictive ability of the classification and regression tree model [8].

Random Forests

Up to now, the main focus has been on constructing single Classification and Regression Trees. Each tree uses a specific and fixed training set to formulate decision rules that best separate classes of data in the training set. The same rules are then

applied to new observations in order to assign a class to each record. However, the CART trees developed can be rather unstable and sensitive to changes in the dataset, such that if the training set used to construct and calibrate the model changes, the resulting decision tree can also drastically change. On the other hand, no methods are known that can increase the classification accuracy both on training and test sets. As a solution to the this problem, Ho (1995) proposes developing multiple trees from randomly selected subsets of the data training sets [27]. Breiman (2001) further expands the concept of *ensemble learning*, which is referred to a machine learning technique that is based on utilizing a set of individual classifiers which are viewed as "*weak learners*" and combining them to a single, more accurate classifier referred to as a "*strong learner*" with higher predictive power [10, 39]. Breiman (2001) proposes *random forests* based on the concept of generating several classification decision trees and aggregating their outcomes to a final classificaier [10].

The algorithm of a random forest for classification and regression based on Breiman (2001) is as follows [33]:

- First, n_{tree} bootstrap sub-samples are drawn form the original dataset.
- For each bootstrap sub-sample an unpruned maximum tree is generated. However, at each splitting node, instead of selecting a variable from all the explanatory variables in the dataset, a random sample of m_{try} is selected and then the splitting variable is selected from the random sample of predictors.
- The final step in predicting the response of an observation is aggregating the prediction of the n_{tree} trees by using the majority vote of the respinse variable for classification models and average value of response for regression models.

Random forests differ based on the number of bootstrap samples n_{tree}, and the predictor sample m_{try}. A special case can be viewed when $m_{try} = p$, the number of explanatory variables, which is referred to as *tree bagging* [9]. Tree bagging includes

only one random sampling of the dataset for the number of decision trees generated, as apposed to two sampling (bootstrap samples of dataset, and random sampling of the explanatory variables) in the random forest method.

3.2.2 Model-Calibration

The initial or maximum trees are usually highly complex and consist of a large number of branches on different levels. Decision trees that are too large are usually a result of *overfitting* the data. In the presence of noise in the training sample, a complex model overreacts to any fluctuations in the training data that are not necessarily underlying patterns in the data. As a result, the predictive ability of the model can be undermined when using the model to classify new, previously unseen observations [34]. On the other hand, trees that are not developed sufficiently are susceptible to *underfitting*, meaning that the algorithm cannot fully capture the trends in the data. Therefore, optimizing the size of the decision trees is an important and necessary step in building a classification model with a high predictive power. The optimization of a decision tree is usually referred to *tree-pruning*. There are two methods, commonly used in decision tree algorithms, described in the following paragraphs, to prune the tree model: 1. Optimizing size of the tree 2. Cross-validation [47]. Cross-validation per se is not a tree-pruning method, but rather a measure used to compare the predictive power of the model in order to choose the best decision tree among a set of trees.

Random forests, on the other hand, do not need tree-pruning or cross-validation, since the classifier, by Breiman's account does not overfit the data [10]. By aggregating several trees, each containing a different set of variables, the chances of overfitting is reduced tremendously. The forest algorithm inherently computes an error measure referred to as *out-of-bag (OOB) error estimate*, which will be further described in detail in Section 3.2.2.

Optimizing Size of the Tree

The maximum tree branches out by default until the impurity in the leaf nodes is either zero or cannot be reduced any further. That leads to the number of cases in each leaf node to be small or equal to one, and the decision tree to be extremely large and difficult to comprehend. One method to prune the maximum tree and optimize the tree structure is to set a minimum number of cases, N_{min}, for each leaf node. As expected, the smaller N_{min} is, the more complex the pruned tree will be. As a rule of thumb, N_{min} can be set to 10% of the training set [34]. For an optimal tree size, the trade-off between predictive power of the model on the test set and impurity of the leaf nodes, which is basically the error on the training set, should be taken into account. Cross validation can be used as a method to quantify the trad-off problem, thus, can be used as a measure for tree optimization.

Cross-Validation

Cross-validation deals with the problem of finding the optimum of the trade-off between misclassification error (impurity in leaf nodes) also called *training error* versus the predictive power of the model on a test set, disjoint from the training set, or the so-called *generalization error* [34]. A highly complex tree has lower misclassification of records in training set until the maximum tree is reached, where the training error is equal to zero. However, highly complex trees perform poorly on disjoint datasets (high generalization error) that the algorithm has not seen before. The following equation describes this optimization problem:

$$R_\alpha(T) = R(T) + \alpha(T') \rightarrow min_T \tag{17}$$

where $R(T)$ denotes the misclassification or training error of the decision tree T, $\alpha(T')$ is a complexity measure, which is a function of the number of leaf nodes T' and

denotes the generalization error based on in-sample cross-validation [47].

The cross-validation procedure takes place by partitioning the training dataset to k equal-sized subsets. Each time, the algorithm takes $k-1$ subsets in order to construct a tree and uses the last subset for testing and evaluating the misclassification error. This process is repeated k times until each subset is used for testing exactly once. The total error is then calculated by aggregating the error of all k runs. The total error can be used as a measure to find the optimal tree size by comparing the error value of different trees. Minimizing the generalization error leads to finding the optimal tree size.

3.2.3 Model-Validation

The final step in credit risk model building process is validation of the model. The validation step according to Gass (1983) refers to "all activities that establish how closely the model mirrors the perceived reality of the model" [21]. Assessing the predictive power and overall accuracy of a classification model is essential in maintaining a desired level of quality over the life time of the risk model by providing feedback information that can be used to optimize the performance of the model [32]. Backtesting and benchmarking are two validation approaches that can be employed depending on the existence of historical data. Backtesting is a retrospective approach that is based on statistical tests of historical realizations of data against model outcomes. Benchmarking on the other hand, is a prospective approach that is based on comparing different risk estimates of various models.

In decision tree analysis, backtesting takes place by classifying a new, previously unseen and disjoint set of observations (test dataset) with the help of the constructed and calibrated tree model and comparing them to the actual classes of data, which the observations belong to. There are various methods to measure the accuracy and predictive power of a tree model. Two common approaches, namely the Receiver Operating Characteristics (ROC) and the confusion matrix are described in the fol-

Table 1: Confusion Matrix

		Prediction		
		{Class 1 (P)}	{Class 2 (N)}	
True Values	{Class 1 (P)}	TP	FN	PPV
	{Class 2 (N)}	FP	TN	NPV
		SEN	SPE	Total

lowing subsections.

Confusion Matrix

The task of a classification model is to predict the class of response, which an observation in the data sample belongs to. If the true classes of the observations are known in the historical dataset, a comparison of the predicted response by the model and the actual categories can be performed. By setting the model predictions against the actual classes from the test dataset the *confusion matrix* can be created. Table 1 illustrates the common structure of a confusion matrix and the measures that are commonly obtained with the help of the matrix for a binary classification model.

The terms TN, FP, TP, and FN used in Table 1 are defined as *True Negative*, *False Positive*, *True Positive*, and *False Negative*, based on whether class 1 (Negative class) and class 2 (Positive class) of data are accurately (True) or not (False) predicted by the model. The sum of all for subgroups is equal to the number of observations in the dataset. With the help of the confusion matrix, various measures for the performance of the predictive model can be defined, such as accuracy, specificity, sensitivity, positive predictive value, and negative predictive value.

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + TN + FN + FP} \tag{18}$$

$$Specificity\ (SPE) = True\ Negative\ Rate = \frac{TN}{TN + FP} \tag{19}$$

$$Sensitivity\ (SEN) = True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{20}$$

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \tag{21}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \tag{22}$$

Other measures for credit scoring analysis have also been proposed in the literature, such as *F-measure*, which are out of scope of this thesis, and thus we do not go further into detail.

Receiver Operating Characteristics (ROC)

Receiver Operating Characteristics (ROC) is a visualization method, utilized to illustrate the predictive or discriminatory power of classification models [51]. The ROC curve is obtained such that the x-axis of the ROC curve illustrates $1 - specificity$ and the y-axis depicts *sensitivity* of the model.

Figure 4 depicts three different models, namely a perfect classification model, a random model, and an arbitrary classification model. The perfect classification model is illustrated with the horizontal line with sensitivity equal to 1. Since the model predicts the class of all records accurately, sensitivity of the model, calculated by Eq. 20, is exactly equal to 1. On the other hand, the random model has no discriminatory power as it just randomly guesses the classes. Thus, the sensitivity increases equally with $1 - specificity$, meaning for each correct prediction, there is exactly 1 wrong prediction. The diagonal depicts a random classification model. For any arbitrary

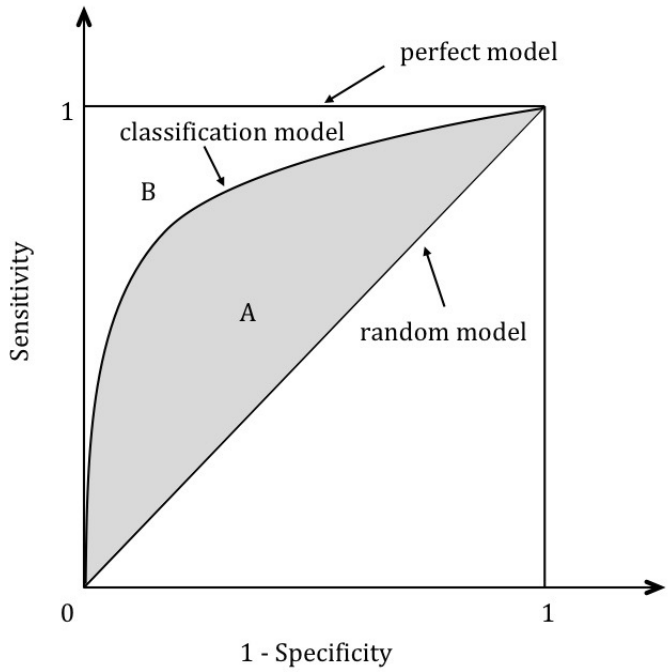

Figure 4: Receiver Operating Characteristics (ROC) Curve.

classification model, the closer the curve is to the perfect model, the higher the discriminatory power of the model will be. Different models can be visually compared with the help of ROC curve.

Furthermore, measures such as Accuracy Ratio, Area Under the ROC curve (AUROC), and Gini Coefficient (not to be confused with the Gini index of the decision tree criterion, defined in Eq. 11) are used to quantify the discriminatory power of the model. The computation of these measures are based on the two areas A and B depicted in Figure 4, such that:

$$Accuracy\ Ratio\ (AR) = \frac{A}{A+B} \tag{23}$$

$$Area\ Under\ the\ ROC\ Curve\ (AUROC) = 0.5(AR+1) \tag{24}$$

$$Gini = 2A \tag{25}$$

The Accuracy Ratio and the Gini coefficient can take a value in $[0, 1]$ closed interval, where the random model has $AR = 0$ and $Gini = 0$, and for the perfect model $AR = 1$ and $Gini = 1$. AUROC values range from 0.5 to 1. For all three measures, higher values demonstrate higher predictive power of a classification model.

Out-of-Bag (OOB) Error Estimate

As mentioned previously, the random forest algorithm does not require a tree-pruning or cross-validation step, as the model inherently calculates a measure for classification error, referred to *out-of bag (OOB) error estimate* by Breiman [10]. The calculation of the random forest involves sampling the training set n_{tree} times. Each time, the remaining observation, not included in the bootstrap sample (thus, "out-of-bag" observations), can be used as a test set to predict the response variable, and compare to the actual response variable. The next step is to aggregate the error estimates of the OOB predictions, referred to as OOB error estimate, which is regarded as an accurate estimate of the generalization error, if enough trees are generated in developing the forest [11]. Using the OOB estimates helps to remove the need for a disjoint and independent test set.

The three *Model-Construction*, *Model-Calibration*, and *Model-Validation* steps, as well as the literature review, described in this chapter, provides a brief description of risk modeling methodology commonly used in the field of credit risk scoring in the banking sector. The next chapter continues with the implementation of the Classification And Regression Tree, as well as random forest algorithms in brownfield regeneration risk assessment field by following the three step *Construction-Calibration-Validation* modeling methodology discussed in Chapter 3.

Chapter 4

CART Analysis in Brownfield Regeneration

In this chapter, three new classification models for successful brownfield regeneration are proposed based on the Classification and Regression Tree (CART) and Random Forest machine learning techniques [8, 9, 10, 27]. In order to develop a valid risk classification model that can attain a desired level of predictive capability and maintain it over time, we follow the three step *Model-Construction*, *Model-Calibration*, *Model-Validation* process, described in detail in Chapter 3. Our model aims to bridge the research gaps and limitations in the existing brownfield redevelopment risk assessment models, as indicated in Section 2.3 and further serve as a guideline for future implementations of CART and random forest algorithms in diverse fields of research within the framework of risk model building. The commented R-code developed and described in this chapter is attached in Appendix A.

Before implementing the CART and random forest algorithms in brownfield regeneration risk classification models, it is important to note some of characteristics and advantages of the two methodologies over other statistical classification methods, briefly discussed in Section 3.1.

- **CART is a non-parametric model.** Thus, it does not require any mathematical formalization for the classification model, which allows more flexibility, as apposed to a linear combination of variables used in LDA, linear regression, or convex combination used in TIMBRE model.

- **CART inherently performs variable selection.** The existing TIMBRE model does not propose any objective methods for selecting variables from large databases, instead it merely depends on experts' suggestions and data availability and includes all available variables in the model. The CART method as well as the Random Forest algorithm intrinsically find the most effective variables through maximizing the gain function, defined in Eq. 14. Adding variables that are insignificant to successful brownfield regeneration does not change the structure or the splitting-rules of the algorithms.

- **CART handles numerical, ordinal, and categorical data.** Since the model does not depend on a functional specification, existence of categorical and ordinal variables with several classes of data does not cause any limitations. Moreover, the model does not make any restrictive assumptions regarding the distribution of variables or the variance/covariance matrices of the classes of the data. However, up to now, due to limitations on computing power, categorical variables with large number of classes cannot be analyzed.

- **CART method is invariant to monotonic transformations of the explanatory variables.** Using the logarithm, square root or any other sort of normalization function does not affect the structure of the tree. It can lead to different cut-off values for the splitting-rules but the variables selected for each binary split, as well as the overall tree structure do not change with variable transformation.

- **CART is not sensitive to outliers in explanatory variables.** By isolating the outliers in a separate node, CART algorithm easily handles noisy data. Moreover, CART can automatically handle missing data.

- **CART can easily be adjusted over time.** With new observations, the model can be continuously re-calibrated over time in order to adjust to the current conditions and maintain a high level of predictive power.

As noted in Section 3.2.1, there are disadvantages to using the CART algorithm. For one, decision trees can be unstable and quite sensitive to changes in the training set, meaning that if observations are added to the training set or omitted from it, the tree structure can change drastically. The CART algorithm also tend to be susceptible to overfitting and overreacting to noisy datasets. The random forest algorithm, on the other hand, covers those issues and has the following features and advantages:

- Random Forests generally do not overfit. Aggregating several single decision trees, and sampling the predictors significantly reduces the chances of overfitting the data. This is an important advantage compared to the decision tree algorithm, which has a habit of overfitting the training data.
- It is unparalleled in predictive power and classification accuracy among current algorithms. At the same time, the computation time is relatively short.
- The algorithm easily handles large datasets, consisting of thousands of observations and hundreds of explanatory variable.
- The model computes estimates of importance for the explanatory variables.
- Random forest internally computes error estimates (OOB estimate) of the model without needing a disjoint test dataset. Also, there is no need to perform cross-validation or tree-pruning, since the model inherently selects bootstrap samples from dataset for tree calibration and the remaining records for validation.
- The algorithm handles outliers and missing data with ease.

The following sections describe the decision tree and random forest classifiers developed for brownfield regeneration classification. All three models follow the three-step credit risk model building methodology, proposed by Altman (1968) [1].

4.1 Model Data Specification

4.1.1 Data Sample

In this thesis, the CART and Random Forest methods are used to develop brownfield regeneration classification models using dataset provided for ENTEKER *(EN-Twicklung Eines Kostenlosen ERkundungsservice)* project, within the framework of *"SMART CITIES - FIT for SET"* program. The extensive dataset was created by combining several separate datasets. The base of the data sample was provided by the Federal Environmental Agency of Austria (Umweltbundesamt), which includes all brownfield sites listed in Austria. The records are limited to sites, where the previous industry in the field was founded before and up to 1989, meaning that if an industry was registered after 1989 and the site has turned into a brownfield, the site is not included in this database. The brownfield dataset provided by Umweltbundesamt (UBA) contains the following attributes:

Table 2: Brownfield Dataset

Attribute	Type of Variable
Geometry	continuous numerical $[m]$ $(\frac{Area}{Circumference})$
Area	continuous numerical $[m^2]$
Circumference	continuous numerical $[m]$
Distance to City Center	categorical $\{< 1km, 1-5km, 5-10km, > 10km\}$
Period of Industry Operation	continuous numerical $[years]$
Contamination Level	ordinal $\{1, 2, 3\}$[a]
Industry Size	categorical $\{small, medium, large\}$
Current Usage	categorical[b]

[a] 1, 2, and 3 denote high, medium, and low contamination level, respectively.

[b] Categories for current usage include residential, industrial, agriculture, train station, etc.

All the attributes listed in Table 2 describe the brownfield sites. The items are selected based on the proposed variables in TIMBRE model, previous studies in the field, and the availability of data in the UBA database. Variables area and circumference describe the size of the brownfield, and geometry denotes its shape. The more compact the brownfield site is (round shapes), the smaller the value of geometry will be. An increase in length to width ratio (extended rectangular shapes) increases the geometry. Geometry, area, circumference, period of industry operation, contamination level, and industry size are regraded as *explanatory variables* (independent) that are brownfield-specific.

The last variable, current usage, is regarded as the *response variable* (dependent). Within the scope of this project, successful regeneration of a brownfield is regarded as redevelopment of a brownfield for residential purposes. In order to use this variable in the analysis, current usage is transformed to a binary variable by splitting all categories into two groups: 1. residential, 2. other. For the remainder of this work, residential category (redeveloped for our purpose) is set equal to 1, and all other categories (not redeveloped as residential) are set to 0. In total, the dataset includes over 46.000 brownfield sites in Austria. By excluding the industrial sites that remain partially in business, which can distort the model outcomes, the final dataset includes 25.324 brownfield sites.

As explained in Chapter 2, local and regional characteristics where the brownfield is located plays an important role in the redevelopment potential of the site, as well. TIMBRE model includes these attributes within the *local development potential* dimension. Thus, our dataset was extended in order to include attributes that describe the municipality and region of the brownfield sites. The main data source is Statistik Austria, which is the Austrian statistical office. Another source used is the *open.data.gv*, which provides public records of open government data. Other data sources include GIS, and Wikipedia. The data covers various characteristics of the

regions, from spacial features of a municipality such as transport connections, number of train stations, and availability of different categories of road transport infrastructure. Socioeconomic features such as real estate prices, population growth, and share of graduates in the municipality are included, as well. Table 3 describes the variables obtained, as well as their respective data type and unit of measurement.

Unlike the brownfield attributes, the municipality variables are repeated through the dataset depending on the location of the brownfield. The municipality with lowest number of brownfield sites contains only 1 site, while the largest municipality has 1819 records of brownfield sites registered. From the beginning of the project, it was decided that Municipality of Vienna should be separated from the rest of the dataset, since the infrastructure and characteristics of this federal state is fundamentally different from the rest of the registered municipalities, and thus, can distort the overall predictive performance of the model. In order to account for geographical and infrastructural differences between the eight federal states of Austria, a categorical variable indicating the state of the brownfield site is included in the analysis, as well.

In total, the dataset includes 49 explanatory variables, 4 of which are categorical, 1 ordinal, and the remaining 44 variables are numerical. The response variable is a categorical variable with two classes, $\{1, 0\}$ for developed, and not developed brownfield sites as residential area.

4.1.2 Data Splitting

Based on the three-step risk modeling methodology used in this thesis, the dataset for the CART tree needs to be split to two subsets: *training set* used to construct the maximum tree and calibrate the optimized or pruned tree and *test set* used for validating the final model in order to evaluate its predictive power. Given that the dataset available is large enough, the two training and test sets can be easily created. By randomly selecting records of data and not replacing the record in the dataset, the training subset is created. The remaining records make up the test set. Normally, 60%

Table 3: Municipality Dataset

Attribute	Type of Variable
Road Transportation Infrastructure	continuous numerical $[m]$
Area of the Municipality	continuous numerical $[m^2]$
Municipal Taxes	continuous numerical [€]
Real Estate Prices	continuous numerical $[€/m^2]$
Trend of Real Estate Prices	categorical $\{ascending, descending\}$
Average Household Size	continuous numerical
Number of Families	discrete numerical
Number of Private Households	discrete numerical
Number of Residential Houses	discrete numerical
Number of Bureaus	discrete numerical
Number of Workplaces	discrete numerical
Number of Employees	discrete numerical
Number of Train Stations	discrete numerical
Number of Traffic Lights	discrete numerical
Number of Hotels	discrete numerical
External Migration	discrete numerical
Internal Migration	discrete numerical
Share of Academics	continuous numerical
Population Density	continuous numerical $(\frac{Population}{Areaof Region})$
Population Growth	discrete numerical $(BirthRate - MortalityRate)$
Federal State	categorical $\{B, K, N, O, Sa, St, T, V\}$[a]

[a]Federal State of Vienna is excluded from the dataset, due to fundamental differences with the rest of the dataset.

to 80% is used for construction and calibration of the model, and the remaining 40% to 20% is used for validation of the model performance. For this thesis, we use 25%/75% splitting ratio for calibration and validation of the CART models, respectively.

4.1.3 Data Exploring

Before constructing the analysis, it is helpful to provide descriptive statistics for the explanatory variables, such as mean, median, standard deviation, and range of values. Table 4 offers a great insight to the attributes used to describe the brownfield sites and their distributions. Since many variables in our dataset are used to describe another attribute further in detail, (for instance, total number of residential apartments, buildings with $1-2$ apartments, buildings with 3 or more apartments), only the total numbers are listed in Table 4.

4.2 Model-Construction

In this section, we describe the methodology used in order to construct the decision trees and random forest. The *partykit* package is used in this thesis for construction of the maximum classification tree and the pruned trees in the following sections. Various packages are available to use in programming language R in order to construct decision trees. The superiority of the partykit package lies in the permutation tests implemented in the algorithm in order to statistically determine the most important variables and splitting rules [44].

The algorithm functions as follows: first, a global null hypothesis is set and tested for independence of explanatory variables and the response variable. If the hypothesis cannot be rejected the algorithm stops, meaning that a significant decision tree cannot be constructed with the set of explanatory variables. Otherwise, the explanatory variable with strongest association to the response variable is selected. The asso-

Table 4: Descriptive Statistics

Attribute	Mean	St. Deviation	Median	Min	Max	Range	Skew	Kurtosis
Geometry	11.68	10.93	8.62	0.59	279.74	279.15	5.55	66.47
Area of Brownfield	5.81e03	3.32e04	1.52e03	6.64	2.88e06	2.88e06	43.42	2.86e03
Circumference	256.60	296.67	177.06	11.20	1.03e04	1.02e04	8.64	165.87
Area of the Municipality	5.30e07	4.79e07	3.62e07	1.1e05	4.67e08	4.67e08	1.91	5.93
Municipal Taxes	1.53e07	3.39e07	1.43e06	3000	1.37e08	1.37e08	2.59	5.26
Real Estate Prices	1.38e03	803.35	1.11e03	680	4.77e03	4.09e03	2.80	8.69
Average Household Size	4.56e03	1.55e04	2.36	1.77	6.23e04	6.24e04	3.24	8.72
Number of Families	1.24e04	2.16e04	1.62e03	25	6.46e04	6.46e04	1.74	1.29
Number of Private Households	1.95e04	3.63e04	2.39e03	34	1.29e05	1.29e05	2.15	3.40
Number of Residential Houses	4.79e03	9.02e03	1.30e03	34	3.51e04	3.51e04	2.74	6.26
Number of Bureaus	175.31	370.92	23	0	1.36e03	1.36e03	2.57	5.19
Number of Workplaces	2.78e03	5.94e03	404	11	2.21e04	2.21e04	2.60	5.33
Number of Employees	2.12e04	5.14e04	1	0.00	1.79e05	1.79e05	2.51	4.78
Number of Train Stations	1.09	1.27	1.00	0.00	5.00	5.00	1.34	1.03
Number of Traffic Lights	1.14	1.64	0.00	0.00	6.00	6.00	1.55	1.38
Number of Hotels	53.64	89.78	17	0	444	444	2.34	4.34
External Migration	501.22	1138.58	40	−64	4323	4387	2.40	3.99
Internal Migration	122.40	453.93	10	−4286	1484	5770	0.19	16.30
Share of Academics	14.32	7.46	12.10	2.80	40.40	37.60	1.24	0.63
Population Density(100)	52.75	17.39	55.90	14.60	93.80	79.20	−0.46	−0.57
Population Density(10)	227.77	165.26	180.70	7.30	896.20	888.90	1.17	1.42
Population Growth	41.01	184.38	−1.00	−167.00	690.00	857.004	3.07	8.04

ciation is computed by the p-value of the partial null hypothesis test of the single explanatory variable and the response. The type of the test can also be specified for the model construction. After determining the variable, a binary split is executed for the variable, such that the impurity measure of the leaf nodes are minimized, using the Gini index as defined in Eq. 11. The binary split involves finding the cut-off value for the splitting rule. This procedure is then recursively repeated. The permutation tests, implemented in the algorithm were first developed by Strasser and Weber (1999) [44, 28, 29]. Function *partykit* :: *ctree* is used to develop the initial tree. No control parameters are used for the maximum tree in order to allow the extension of tree branches until the algorithm stops.

Other tree structures can also be developed and optimized by using the *partykit* :: *ctree_control* command to set various types of control parameters, such as minimum number of observations in leaf nodes, minimum number of observations in splitting nodes, the type of the statistical test executed, the significance level for the statistical test used, etc. In the following section, we develop and calibrate the unpruned maximum tree, and a pruned tree based on optimizing the tree size using the training dataset.

The final step is to develop a random forest based on a set of unpruned maximum trees on bootstrap samples of the dataset. Unlike the two previous single trees (maximum and pruned), the random forest algorithm does not require two disjoint datasets, one for calibration of the model, and one for validation of the model's predictive power. The only parameters needed to develop the random forest classifier is the number of maximum trees generated n_{tree} and the number of explanatory variables m_{try}, which are randomly selected from the set of predictors each time for the tree to be constructed. The algorithm further computes the OOB error estimate that can be viewed as the generalization error of the model and demonstrates its predictive power. For this purpose, we use the *randomForest* package in R, which implements Breiman's random forest algorithm [33]. Function *randomForest* :: *randomForest* generates the random forest. Control parameters can further be set for the forest,

such as setting number of observations in leaf nodes of each decision tree and defining a vector of importance for the predictor variables. For our random forest model, we follow the parameters originally proposed by Breiman (2001) in his original paper [10].

4.3 Model-Calibration

4.3.1 Maximum Tree Calibration

In this section, the maximum tree is created using the training dataset, which contains 18.739 records of brownfield sites, described through 49 explanatory variables. Function *partykit* :: *ctree* is used as previously described for the construction and calibration of the model, with no control parameters set for the model. The maximum tree for the CART model branches out until the Gini index, define in Eq. 11 cannot be reduced any further with no control parameters to stop the algorithm beforehand. The resulting tree for this dataset consists of 44 inner nodes and 45 leaf nodes. Figure 5 illustrates the complex structure of the maximum tree. The tree begins at the root node, and with each indentation the tree branches to child nodes. The nodes are numbered in order to ease comprehension of the tree. For each node, the splitting rule is indicated with the variable selected, and the cut-off value used for the binary split. The recursive partitioning progresses until the algorithm reaches a leaf node. The leaf nodes are indicated by the class of response (0/1), number of observations in the nodes (n), and the misclassification error (err). The class of the leaf node is decided by the majority of observations (class 0 or 1) in the node.

Model formula:
Usage ~ Federal State + Municipality Taxes + Academic Ratio + Household +
  Household Size + Families + Total Buildings + 1-2 Apartments +
  3 - More Apartments + Shared Apartments + Hotels + Bureau + Population +
  Population Growth + Internal Migration + External Migration + Workplace Total +
  Workplace 0-4 + Workplace 5-19 + Workplace 20-99 + Workplace 100-250 +
  Workplace 250plus + Employee Total + Dependent Employee + Length
Municipality + Area Municipality + Distance + Road_FRC_0 + Road_FRC_1 +
Road_FRC_2 + Road_FRC_3 + Road_FRC_4 + Road_FRC_5 + Road_FRC_6 +
Road_FRC_7 + Train Station + Traffic Stop + Loading Station + Trend + Real Estate
Prices + Area + Circumference + Geometry + Distance to Center + Contamination
Level + Contamination Period + Business Size + Population Density 100km +
Population Density 10km

Fitted Tree:
[1] root
| [2] Geometry <= 12.02
| | [3] Federal State in B, K, O
| | | [4] Business Size in large, medium
| | | | [5] Families <= 20101: 0 (n = 1091, err = 44.8%)
| | | | [6] Families > 20101
| | | | | [7] Geometry <= 8.13: 1 (n = 82, err = 12.2%)
| | | | | [8] Geometry > 8.13: 1 (n = 57, err = 49.1%)
| | | [9] Business Size in small
| | | | [10] Federal State in B, O
| | | | | [11] Geometry <= 7.65
| | | | | | [12] Distance to Center in 1 - 5 km, > 10 km
| | | | | | | [13] Real Estate Prices <= 891: 1 (n = 281, err = 27.0%)
| | | | | | | [14] Real Estate Prices > 891: 1 (n = 620, err = 39.5%)
| | | | | | [15] Distance to Center in 5 - 10 km, < 1 km: 1 (n = 710, err = 26.5%)
| | | | | [16] Geometry > 7.65
| | | | | | [17] Road_FRC_1 <= 6097.25952: 1 (n = 1031, err = 41.9%)
| | | | | | [18] Road_FRC_1 > 6097.25952: 0 (n = 159, err = 41.5%)
| | | | [19] Federal State in K
| | | | | [20] Population Growth <= -80
| | | | | | [21] Contamination Level <= 1: 1 (n = 31, err = 38.7%)
| | | | | | [22] Contamination Level > 1: 0 (n = 93, err = 18.3%)
| | | | | [23] Population Growth > -80
| | | | | | [24] Trend in ascending: 1 (n = 89, err = 27.0%)
| | | | | | [25] Trend in descending: 0 (n = 272, err = 50.0%)
| | [26] Federal State in N, Sa, St, T, V
| | | [27] PopulationDensity10 <= 302.3
| | | | [28] Trend in ascending, fix
| | | | | [29] Business Size in large, medium: 1 (n = 1049, err = 36.1%)
| | | | | [30] Business Size in small

| | | | | | [31] Geometry <= 2.94
| | | | | | | [32] Geometry <= 2.39
| | | | | | | | [33] Real Estate Prices <= 1626.7
| | | | | | | | | [34] Road_FRC_2 <= 9517.48922: 0 (n = 91, err = 6.6%)
| | | | | | | | | [35] Road_FRC_2 > 9517.48922: 0 (n = 13, err = 46.2%)
| | | | | | | | [36] Real Estate Prices > 1626.7: 1 (n = 20, err = 35.0%)
| | | | | | | [37] Geometry > 2.39
| | | | | | | | [38] Contamination Level <= 1: 1 (n = 12, err = 0.0%)
| | | | | | | | [39] Contamination Level > 1: 0 (n = 86, err = 37.2%)
| | | | | | [40] Geometry > 2.94
| | | | | | | [41] Circumference <= 146.69
| | | | | | | | [42] Geometry <= 4.32
| | | | | | | | | [43] Distance <= 30520.10493: 1 (n = 370, err = 27.0%)
| | | | | | | | | [44] Distance > 30520.10493
| | | | | | | | | | [45] Workplace 5-19 <= 15
| | | | | | | | | | | [46] Hotels <= 11: 0 (n = 20, err = 0.0%)
| | | | | | | | | | | [47] Hotels > 11: 0 (n = 9, err = 44.4%)
| | | | | | | | | | [48] Workplace 5-19 > 15: 1 (n = 23, err = 30.4%)
| | | | | | | | [49] Geometry > 4.32: 1 (n = 1394, err = 20.7%)
| | | | | | | [50] Circumference > 146.69
| | | | | | | | [51] Distance to Center in 1 - 5 km, < 1 km: 1 (n = 318, err = 39.6%)
| | | | | | | | [52] Distance to Center in 5 - 10 km, > 10 km: 1 (n = 801, err = 27.2%)
| | | | [53] Trend in descending
| | | | | [54] Real Estate Prices <= 1957.1: 1 (n = 892, err = 44.1%)
| | | | | [55] Real Estate Prices > 1957.1
| | | | | | [56] Area <= 1416.93: 1 (n = 306, err = 22.5%)
| | | | | | [57] Area > 1416.93: 1 (n = 136, err = 42.6%)
| | | [58] PopulationDensity10 > 302.3
| | | | [59] Area <= 1048.49
| | | | | [60] Geometry <= 2.97: 1 (n = 71, err = 38.0%)
| | | | | [61] Geometry > 2.97
| | | | | | [62] Traffic Stop <= 1: 1 (n = 589, err = 19.4%)
| | | | | | [63] Traffic Stop > 1: 1 (n = 1098, err = 11.0%)
| | | | [64] Area > 1048.49: 1 (n = 1220, err = 29.3%)
| [65] Geometry > 12.02
| | [66] Business Size in large, medium
| | | [67] Federal State in B, N, St, V
| | | | [68] Contamination Period <= 53
| | | | | [69] Household Size <= 2.08: 0 (n = 155, err = 48.4%)
| | | | | [70] Household Size > 2.08: 0 (n = 1041, err = 30.5%)
| | | | [71] Contamination Period > 53
| | | | | [72] Geometry <= 17.08: 1 (n = 241, err = 40.2%)
| | | | | [73] Geometry > 17.08: 0 (n = 379, err = 39.8%)
| | | [74] Federal State in K, O, Sa, T

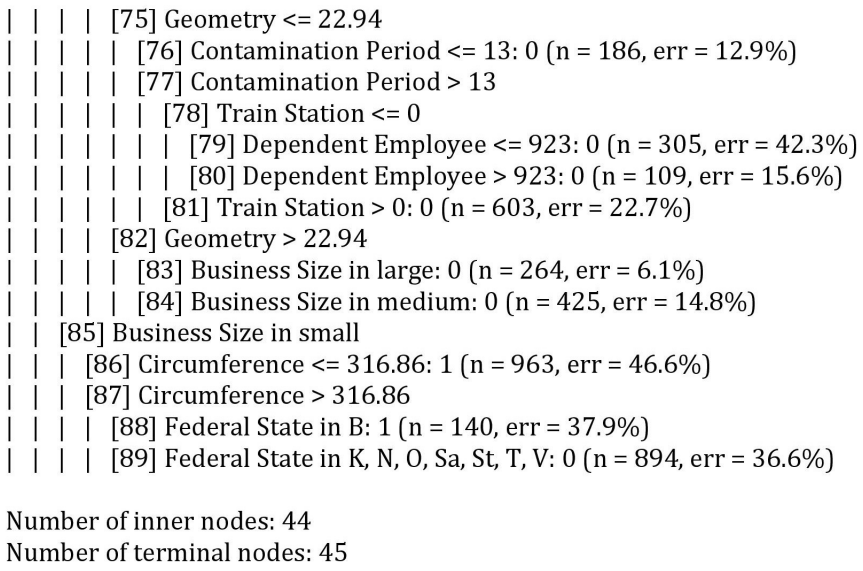

Figure 5: Maximum Tree for Classification of Brownfield Regeneration

As mentioned previously, the *partykit* package uses statistical tests in order to choose the explanatory variable for each splitting rule. The p-value associated with the tests are provided as an output of the model. For all binary splits in our maximum tree, the $p - value < 0.001$, and thus the the choice of variables are statistically significant. It is clear based on Figure 5 that the CART algorithm has selected the most important explanatory variables for the tree structure, and has omitted the rest altogether. This is a direct result of the variable selection process inherently performed in the CART method.

A brief overview of the misclassification error shows that the error rates in the leaf nodes range form $0.0\% - 50.0\%$, indicating a range from perfect classification in some nodes to random classification in others. Overall, the misclassification rate of the maximum tree (Eq. 18) for this training set amounts to 31.44% ($Accuracy = 68.56\%$). Moreover, the true positive rate, or sensitivity (Eq. 20) is equal to 0.81156 and true negative rate or specificity (Eq. 19) is 0.5189. The values show that the maximum

tree has high discriminatory power when it comes to developed brownfield sites (class 1 of response), but when it comes to not developed brownfield sites (class 0) the maximum tree is close to a random classification model with low discriminatory power.

4.3.2 Pruned Tree Calibration

The next step after constructing and calibrating the maximum tree is to optimize the tree size in order to ease comprehension of the model and avoid overfitting the data, which the decision tree tend to do. A quick look at the nodes of the maximum tree shows that some leaf nodes contain as little as 9 observations. Such a fine and complex tree construction is susceptible to $overfitting$ the data, which causes overreaction to the noise in the data and fluctuations that do not necessarily represent the underlying drivers of brownfield regeneration. Overfitting can reduce the predictive power of the tree model on previously unseen observations. A simple method used to optimize the tree structure and increase the generalization capability of the model is to set a minimum number of observations N_{min} in the leaf nodes. As a rule of thumb, many authors suggest using 10% of total number of observations in the training data as N_{min} [34].

The optimization method is implemented in R by defining the control parameters through restricting the minimum "bucket" size N_{min} in $ctree\_control$. The resulting calibrated tree is depicted in Figure 6. The pruned tree now consists of 7 internal nodes and 8 leaf nodes, which is considerably smaller than the maximum tree, and hence, much easier to interpret and visualize. Similar to the maximum tree, all the variables selected for splitting rules are statistically significant. Moreover, the minimum number of observations in a node is 1816, which satisfies the 10% rule. The general expectation is that the misclassification or training error should increase with the tree-pruning step. Construction of the confusion tree for the training set shows that the training error is 33.93% ($Accuracy = 66.07\%$). Compared to the maximum tree the training error has increased only slightly (3.6%). The sensitivity of the

```
Model formula:
Usage ~ Federal State + Municipality Taxes + Academic Ratio + Household +
    Household Size + Families + Total Buildings + 1-2 Apartments +
    3 - More Apartments + Shared Apartments + Hotels + Bureau + Population +
    Population Growth + Internal Migration + External Migration + Workplace Total +
    Workplace 0-4 + Workplace 5-19 + Workplace 20-99 + Workplace 100-250 +
    Workplace 250plus + Employee Total + Dependent Employee  + Length
Municipality + Area  Municipality + Distance + Road_FRC_0 + Road_FRC_1 +
Road_FRC_2 + Road_FRC_3 + Road_FRC_4 + Road_FRC_5 + Road_FRC_6 +
Road_FRC_7 + Train Station + Traffic Stop + Loading Station + Trend + Real Estate
Prices + Area + Circumference + Geometry + Distance to Center + Contamination
Level + Contamination Period + Business Size + Population Density 100km +
Population Density 10km

Fitted party:
[1] root
|   [2] Geometry <= 12.02
|   |   [3] Academic Ratio <= 25.6
|   |   |   [4] Business Size in large, medium: 1 (n = 2795, err = 42.7%)
|   |   |   [5] Business Size in small
|   |   |   |   [6] Federal State in B, N, Sa, St, T, V
|   |   |   |   |   [7] PopulationDensity10 <= 116.3: 1 (n = 2618, err = 37.1%)
|   |   |   |   |   [8] PopulationDensity10 > 116.3: 1 (n = 3212, err = 24.3%)
|   |   |   |   [9] Federal State in K, O: 1 (n = 2435, err = 41.5%)
|   |   [10] Academic Ratio > 25.6: 1 (n = 1974, err = 20.4%)
|   [11] Geometry > 12.02
|   |   [12] Business Size in large, medium
|   |   |   [13] Federal State in B, N, St, V: 0 (n = 1816, err = 37.8%)
|   |   |   [14] Federal State in K, O, Sa, T: 0 (n = 1892, err = 20.4%)
|   |   [15] Business Size in small: 0 (n = 1997, err = 46.5%)

Number of inner nodes: 7
Number of terminal nodes: 8
```

Figure 6: Tree Model 1 optimized through minimum number of observations N_{min}.

calibrated tree is equal to 0.8126, which means that the pruned tree can classify the developed brownfield sites with high precision and just as well as the maximum tree. The specificity, however, is equal to 0.4594, which shows a 11.5% decrease compared to the maximum tree. This means that the pruned tree performs worse than a random model when it comes to classifying the not developed brownfield sites.

4.3.1 Random Forest Calibration

In this section, the final brownfield regeneration classification model, which is the random forest is calibrated using the entire dataset. As previously mentioned, calibration and validation of the random forest can be achieved using a single dataset, since each time the model uses a bootstrap sample for the calibration of the single decision trees, and the remaining observations for estimating the predictive power of the model, also known as the OOB error estimate. For calibration of the random forest, two values need to be selected, which are the number of maximum trees generated for the forest, n_{tree}, and the number of variables sampled from the entire predictor matrix, m_{try}. Generally, there are no rules for selecting the number of decision trees used for a random forest. Increasing n_{tree} results in higher model performance and predictive power, but increases the computation time, on the other hand. As long as the processor can handle the computation, it is generally better to increase n_{tree} as much as possible. For this thesis, we set $n_{tree} = 500$.

For the selection of the number of features sampled, m_{try}, there are several suggestions in the literature, but no strict rules. A special case of random forest is the Braiman's tree baging algorithm, which is when the number of selected features is equal to the number of predictors. Breiman (2001) later uses the first integer less than $log_2 p + 1$, where p is the number of explanatory variables [10]. Other suggestions include square root of number of variables, ($\sqrt{p}$), or 20% of all predictors, etc. There is, however, no clear advantage in using any of the values mentioned. Using a fraction of the predictors allow an increase in individuality of the generated trees, but decreases the

overall predictive accuracy of the model. Selecting the number of variables can thus be viewed as an optimization process. For our model, we select the value proposed by Breiman, which equals $int(log_2 p + 1) = 5$ based on our dataset. We do not set any further controls in construction and calibration of the random tree and follow the methodology used by Breiman [9, 10]

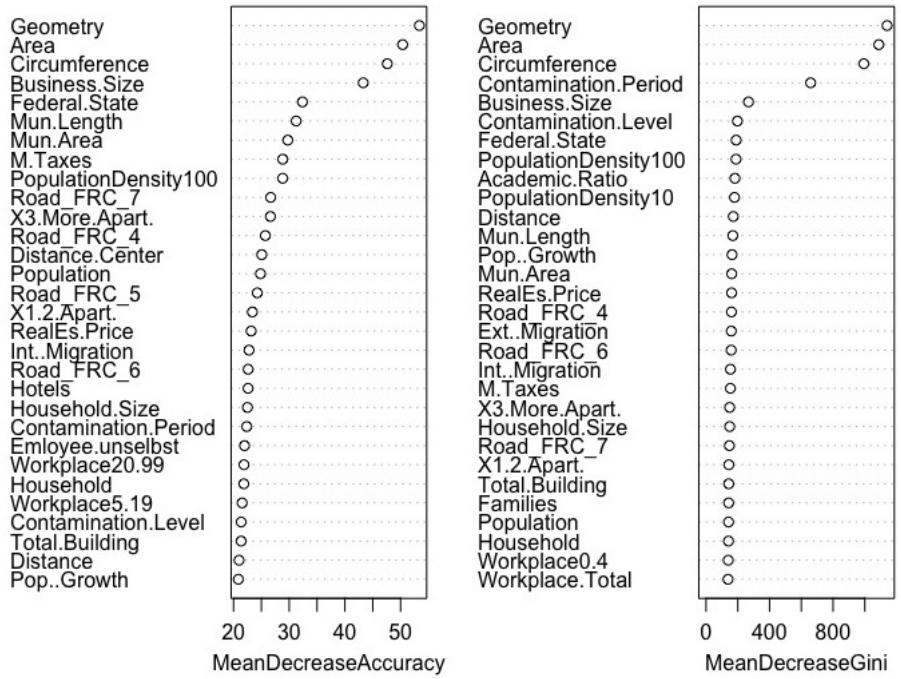

Figure 7: Variable Importance Measures Obtained by the Random Forest Classifier.

Generally, since the random forest consists of an ensemble of decision trees, the model structure cannot be plotted in a similar fashion to single decision trees. Function $randomForest :: getTree$ plots a specifically selected tree from the random forest, but the benefits of plotting a single tree from the algorithm are rather limited, and thus, we leave out the visualization of the last model. However, the $varImpPlot$

plots a dotchart of variable importance as measured by a random forest. The random forest developed in this section is constructed and calibrated based on the following variable importance measures, depicted in Figure 7.

The next and final section of this chapter evaluates the predictive power of the two trees developed earlier by utilizing an independent validation dataset. For the random forest model, out-of-bag (OOB) error estimate of the random forest classifier will be computed and discussed.

4.4 Model-Validation

The final stage in risk model building process is the validation of the constructed and calibrated model. This step helps to evaluate the predictive power of the model by using a disjoint set of observations, which the model has not seen before in order to predict the classes of response the new observations belong to. This procedure helps to assess the model performance, and further provide feedback information needed to maintain and improve a certain level of quality for the model.

We start by validating the maximum tree, developed in Section 4.3.1. As previously described, 75% of the dataset is used as the training set to construct the models. The remaining 25% of the dataset (*test set*) is used for validation and is entirely disjoint from the training set. The test set includes over $6,580$ observations from both classes combined. We use the *predict* $\{stats\}$ function in R, which is a generic function that can be used to make predictions based on the results of model fitting functions [41]. Here, the function predicts the classes of response (0 or 1) for the test set based on the results of the tree model algorithms.

4.4.1 Maximum Tree Validation

The first method used to illustrate the predictive power of the maximum tree is to set the model predictions against the actual classes of response variable from the test set with the confusion matrix. Table 5 illustrates the results of validating the maximum tree with the help of the confusion matrix.

Table 5: Confusion Matrix of the Maximum Tree

		Prediction		
		{Class 0 (ND[a])}	{Class 1 (D[b])}	
True Values	{Class 0 (ND)}	1492	780	NPV = 0.6567
	{Class 1 (D)}	1384	2929	PPV = 0.6791
		SPE = 0.5188	SEN[c] = 0.7897	

[a] ND stands for Not Developed Brownfield Sites.

[b] D stands for Developed Brownfield Sites for residential purposes.

[c] PPV, NPV, SPE, and SEN, are acronyms for Positive Predictive Value, Negative Predictive Value, Specificity, and Sensitivity, respectively.

Based on the confusion matrix the following measures can be computed. The overall accuracy of the model, computed with Eq. 18, is equal to 0.6714, meaning that the maximum tree classifies 67.14% of the new observations correctly based on the class they belong to. The sensitivity (or the *true positive rate*) of the maximum tree is 0.7897, which indicates that from all the brownfield sites predicted as developed as residential area in the test set, the maximum tree model has detected 78.97% of them correctly. This high detection rate for the developed sites in the test data set shows that the maximum tree has identified the underlying patterns of developed sites. On the other hand, the *positive predictive value* of the model, calculated in Eq. 21, is equal to 0.6791. PPV indicates that from all the sites that are actually developed for residential purposes, more that $\frac{2}{3}$ are predicted accurately and $\frac{1}{3}$ are predicted falsely.

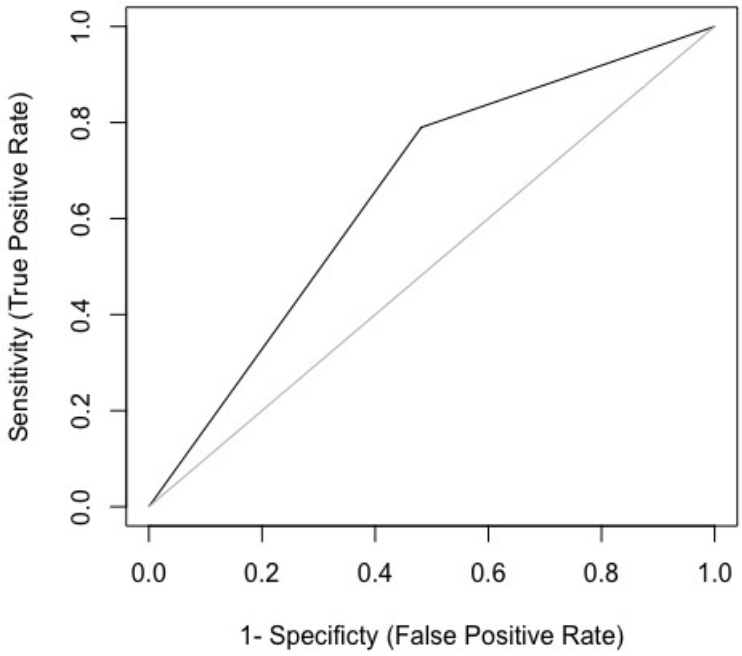

Figure 8: Receiver Operating Characteristics (ROC) Curve of the Maximum Tree.

The specificity (*true negative rate*) of the model is equal to 0.5188. The low true negative rate implies that the maximum tree model can identify the not developed brownfield sites (class 0) only slightly more than 50% of the time, which is basically similar to a random model. The *negative predictive value*, furthermore indicates that from all the brownfield sites that the model has predicted as not developed, around 66% are not developed in reality and 33% are predicted falsely, meaning they are actually regenerated and used as residential area, even though our model classifies them as not developed. The above-mentioned figures imply that the model faces difficulty in detecting the negative class compared to the developed fields. One reason could be that from the many brownfield sites listed as not developed in our dataset, many still have the potential to be regenerated, but up to the time of data collection, no

one has seized the opportunity to revitalize the sites.

The second approach used to evaluate and visualize the predictive power of a classification model is the ROC curve. Figure 8 plots the $Sensitivity$ against $1-Specificity$ of the maximum tree model. Since the maximum tree is a binary classification model that predicts discrete values for the response outcome, the ROC curve consists of two straight lines for the two classification categories. The area under the curve (AUROC), a measure commonly used to evaluate the predictive power of the model as defined in Eq. 24 is equal to 1 for a perfect classification model and equal to 0.5 for a random model. The AUROC measure computed for the maximum tree is equal to 0.6542, which is consistent with the results from the confusion matrix and implies that the overall predictive power of the model is average.

4.4.2 Pruned Tree Validation

After constructing the maximum tree, the pruned tree was constructed and calibrated in Section 4.3.2 in order to avoid the overfitting phenomenon that tends to happen in model construction step of decision trees. The validation approaches described above are calculated again for the pruned tree model in order to be able to compare the predictive capabilities of the two models. The confusion matrix for the test set of the pruned tree is illustrated in Table 6.

Table 6: Confusion Matrix of the Pruned Tree

		Prediction		
		{Class 0 (ND)}	{Class 1 (D)}	
True Values	{Class 0 (ND)}	1349	720	NPV = 0.6520
	{Class 1 (D)}	1527	2989	PPV = 0.6619
		SPE = 0.4691	SEN = 0.8059	

The values illustrated in Table 6 show that the sensitivity of the pruned tree is 0.8059, which is slightly higher than the maximum tree (2% increase), but the positive predicted value decreases around 2.5% and is equal to 0.6619. Moreover, the specificity of the pruned tree has decreased to 0.4691 from 0.5188. The pruned model now detects less than half of the negative class (not developed sites) accurately, which is worse than a random binary classifier. The negative predicted value remains almost constant with the tree-pruning step with 0.6520 from previous value 0.6567. The overall accuracy of the model is also slightly less than the maximum tree and is equal to 0.6588. In general, the pruned tree model has a marginally lower predictive power compared to the previous model, but the model structure is substantially smaller and easier to comprehend.

Finally, the ROC curve of the pruned tree is plotted for visualization of the model performance. At first glance, the curve looks quite similar to that of the maximum tree. The AUROC of the curve is equal to 0.6375, which is 2.6% lower than the maximum tree.

Both the confusion matrix and the Receiver Operating Characteristics approach indicate that the pruned tree model not only does not have a higher predictive power than the maximum tree, it actually performs slightly worse than the original model. The tree-pruning step is included in the model development process in order to enhance the generalization capability of the decision tree model and avoid overfitting the data. But in our case, tree-pruning negatively affects the model performance. Since the only measure used to optimize the tree structure for this model is the tree size through setting a minimum number of observations in leaf nodes ($N_{min} \sim 10\% of the training set$), the predictive power of the model did not improve with the tree-pruning step. However, it is worth mentioning that the model structure of the pruned-tree is much more simple, easier to comprehend and utilize for new observations in the data.

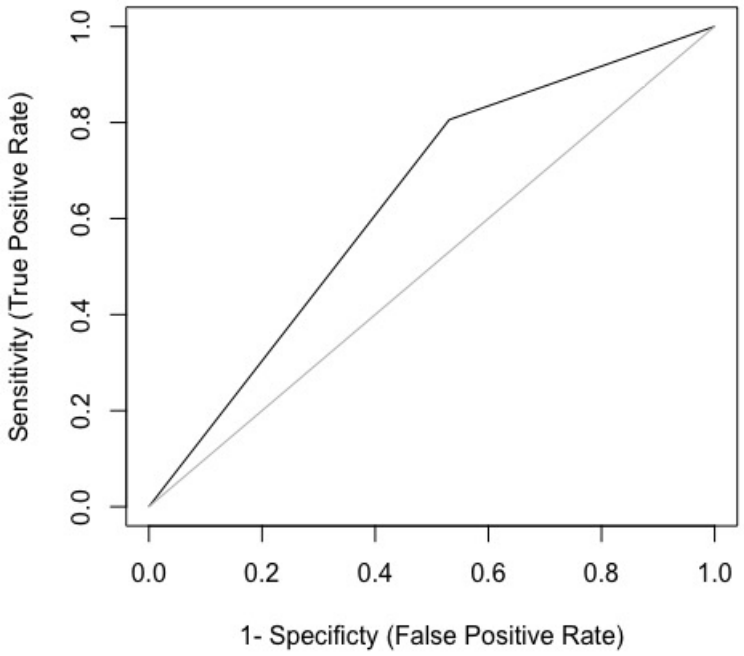

Figure 9: Receiver Operating Characteristics (ROC) Curve of the Pruned Tree.

4.4.3 Random Forest Validation

The final model to be validated in this thesis is the random forest classification model. Validation of this model is slightly different form the two previous models, since the random forest algorithm intrinsically obtains a classification error estimate and does not require an independent set of observations for validation. Each time the algorithm randomly selects samples from the main dataset to generate decision trees, the remaining observations (or the "out-of-bag" data) are used for predictions and error estimation. The aggregation of all sample errors results in the Out-Of-Bag (OBB) error estimate. The OOB can be regarded as an accurate generalization error [11].

The first approach to measure the predictive power of the random forest is construc-

tion the confusion matrix, depicted in Table 7.

Table 7: Confusion Matrix of the Random Forest

		Prediction		
		{Class 0 (ND)}	{Class 1 (D)}	
True Values	{Class 0 (ND)}	6480	4458	NPV = 0.6516
	{Class 1 (D)}	3465	10921	PPV = 0.7101
		SPE = 0.5924	SEN = 0.7591	

The general measure used to assess the predictive power of the model is the accuracy measure, as defined in Eq. 18, is equal to 0.6871, which is higher than both previous tree models. Moreover, specificity of the classifier is now much higher than both tree models and equal to 0.5924. Although this value is still rather low, the detection of the negative class is now better than a random classifier. The positive predictive value (PPV) has significantly increased in comparison to the other models, but the negative predictive value (NPV) has remained more or less similar to the previous two tree models. The next method used to illustrate the predictive power of the random forest model is the Receiver Operating Characteristics (ROC) curve. The area under the ROC curve (AUROC), depicted in Figure 10, for the random forest is equal to 0.6809, which demonstrates a 4% increase from the maximum tree and a 7% increase from the pruned tree.

For the construction and calibration of the random forest, the number of generated decision trees for the development of the random forest, n_{min}, is set to 500. The selection of the number is arbitrary and set as an initial guess that could be optimized later. After constructing, calibrating and validating the model, we can assess the effect of n_{tree} on model error rates. Figure 11 demonstrates the change in three different error rates, namely random forest's OOB error estimate, error rate of the developed

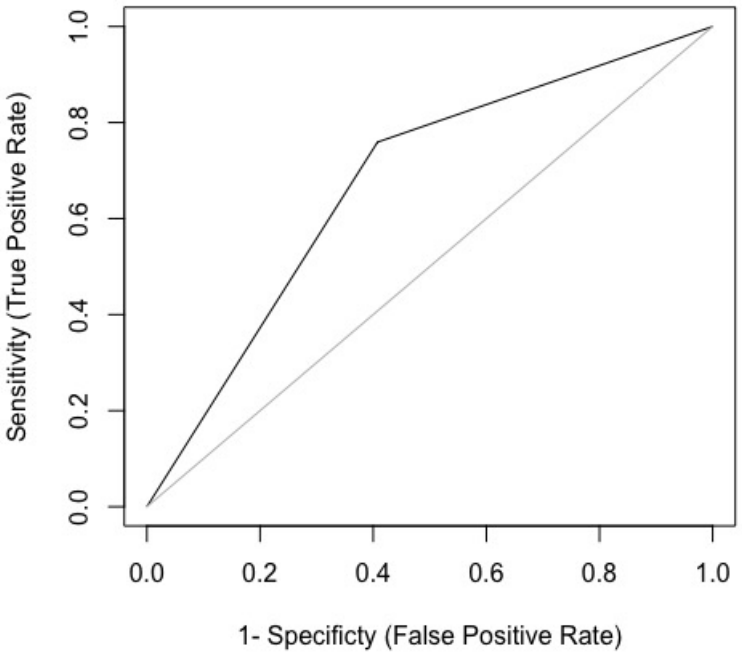

Figure 10: Receiver Operating Characteristics (ROC) Curve of the Random Forest.

brownfield sites (class 1), and error rate of the un-developed brownfield sites (class 0). As seen all throughout this section, the error rate of the negative class is the highest of the three error rates, and error rate of the positive class is the lowest, meaning that regardless of the number of trees used, all single models and combinations of single trees face difficulty when detecting the not-developed brownfield sites. The OOB error estimate, which can be viewed as the generalization error of the model, is the average of the two previous error rates. As demonstrated in Figure 11, all three rates decrease drastically as the number of trees increase to around 100. An increase to 200 trees for the construction of the random forest slightly improves model performance. After that, all three error rates converge to specific values and do not change as the number of trees increase. If computation time is relevant in the model building process and needs to be reduced, the number of trees can be reduced to 200 in the fu-

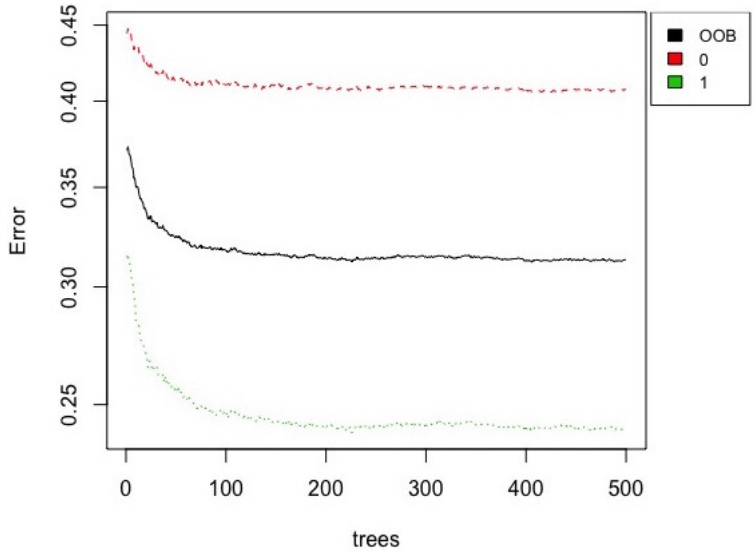

Figure 11: Effect of Number of Decision Trees on the Error Rates of the Random Forest.

ture models in order to save computation time without sacrificing model performance.

4.5 Model Comparison

In the final section of Chapter 4, a brief summary of the performance of the three developed models, namely the maximum tree, the pruned tree, and the random forest, is conducted. The aim of this section is to ease comparison of the models. The following figure, Figure 12, plots the ROC curves of all three developed models. As mentioned previously, the random forest demonstrates the highest predictive power, and has the largest area under the ROC curve. The maximum and the pruned tree follow by a margin, respectively. As previously described, since the nature of the classifiers are binary, the ROC curve consists of two straight lines. Table 8 further provides a figurative comparison of all relevant measures of the models, calculated throughout this chapter.

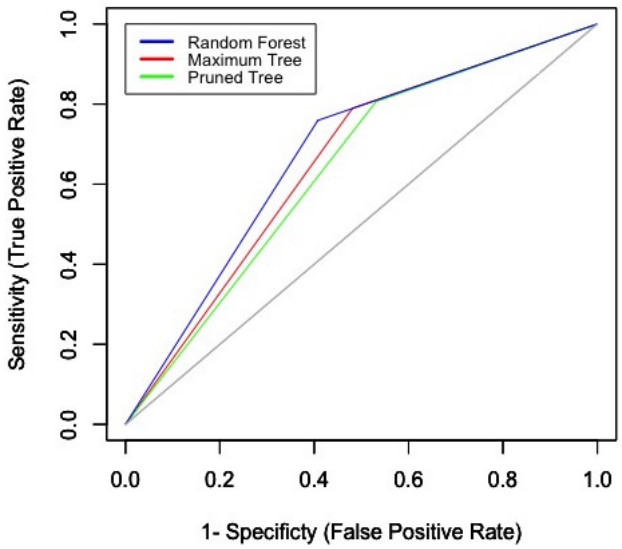

Figure 12: Comparison of Receiver Operating Characteristics (ROC) Curves of All Models.

Table 8: Comparison of Maximum Tree, Pruned Tree & Random Forest.

Performance Measure	Maximum Tree	Pruned Tree	Random Forest
Accuracy	0.6714	0.6588	0.6871
Sensitivity	0.7897	0.8059	0.7591
Specificity	0.5188	0.4691	0.5924
PPV	0.6791	0.6619	0.7101
NPV	0.6567	0.6520	0.6516
AUROC [a]	0.6542	0.6375	0.6809

[a]PPV, NPV, and AUROC are acronyms for Positive Predictive Value, Negative Predictive Value, and Area Under ROC curve, respectively.

Chapter 5

Discussions and Conclusions

Over the last several decades, brownfield site redevelopment and regeneration has received widespread attention as a sustainable land use strategy to fight urban sprawl, which has become a major issue facing Europe [6]. *Brownfield* sites are defined as former military, industrial, and commercial sites, which are currently derelict and underutilized with varying degrees of contamination [37]. Although public interest in revitalizing brownfield sites has increased over the past several years, due to the various types of risks involved with brownfield regeneration, such as social, environmental, and economic dimensions, risk assessment tools are needed in order to assist stake-holders and decision makers with prioritizing and classifying the suitable brownfield sites that have high potential to be redeveloped for various purposes, such as residential, business and operational, schools and playgrounds, etc.

For this purpose, several assessment tools and prioritization models have been suggested in the literature in order to forecast the potential of a ceratin site and evaluate the inherent risks of brownfield regeneration with their main focus on various aspects of it, such as uncertainty assessment, environmental and health risk assessment, remediation cost assessment, etc. [4, 15, 40]. However, all the proposed models are either developed on a case-by-case basis or lack a multidisciplinary approach needed for regeneration assessment [15] and further, all models fail to assess their predictive power based on the goodness of the model outcomes against realizations of brownfield regeneration in a procedure referred to as *validation*. The validation process is an essential step in risk model building methodology that is executed in order to assess

the performance of the model and used to maintain a desired level of model quality over its life cycle.

One major project that merges most existing models into one and utilizes a multidisciplinary approach is the **T**ailored **Im**provement of **B**rownfield **R**egeneration in **E**urope **(TIMBRE)**, which assists stakeholders to rank brownfields based on their redevelopment potential by using multi-criteria decision analysis methodology by computing a proritization or ranking score, through a hierarchical structure, which includes dimensions, factors and indicators [40]. Until now, however, the validity and the predictive power of the suggested model has not been reported in the literature. The lack of empirical evaluation of existing state of the are model in brownfield regeneration is the missing link in risk modeling studies that needs to be addressed.

This thesis aims to use the TIMBRE scoring model as the reference point and develop and validate a risk model for brownfield regeneration. By following the three step *Construction-Calibration-Validation* model building process, widely practiced in the field of financial risk management, the attempt is to bridge the gap between brownfield regeneration scoring models and risk assessment models. Our goal is to construct a new brownfield regeneration risk model using the decision tree analysis methodology, and validate it with historical data on redevelopment of brownfield sites is Austria. The main function of our model is to act as a classification tool by assigning a class to each brownfield that best separates the "good" candidates (within the scope of this work: successful regeneration of brownfield sites) from the "bad" (not regenerated sites) in a procedure commonly known as *classification*. Assessing the predictive power of the classification model and its calibration is a major task of the validation step. By incorporating different calibration and validation methods we close the missing link in brownfield regeneration tools.

We first construct the initial decision tree known as *maximum tree* with the help of a training set within the Classification and Regression Tree (CART) methodology [8]. The *partykit* package in programimng language R is used for construction of the

maximum classification tree, which functions by selecting the explanatory variable with strongest association to the response variable out of the explanatory variable matrix and computing the corresponding p-value of the partial null hypothesis test. After selecting the variable, a binary split is executed for the variable and the procedure is then recursively repeated [44, 28, 29]. No control parameters are used for the maximum tree in order to allow the extension of tree branches until the algorithm stops. The resulting maximum tree consists of 44 inner nodes and 45 leaf nodes as depicted in Figure 5. The overall accuracy of the model is equal to 68.56% on the training set and 67.14% on the test set. The fact that the predictive power of the model on training and test sets are so close shows that the generalization capability of the model through the splitting decisions are high and the overfitting phenomenon does not play a significant role on the model predictive power. Moreover, the sensitivity of the model both on training and test sets are much higher (0.81156, and 0.7897, respectively) than the specificity measures (0.5189 and 0.4691 for training and test set), meaning that the model detects the positive class of data (developed brownfield sites) much better than the negative class (not developed sites).

The second step is optimizing the size of the tree in order to boost the predictive power of the model and avoid overfitting the data that is prevalent in maximum tree construction. As a rule of thumb, the size of the leaf nodes is set to 10% of the training set, which helps to avoid overreaction of the model to noisy data. The resulting optimized tree is depicted in Figure 6, which now consists of 7 internal nodes and 8 leaf nodes, and shows an classification accuracy of 66.07% on the training set, which is a 3.6% decrease from the maximum tree. The optimized tree is validated with a test set with an accuracy equal to 65.88% and a sensitivity equal to 0.8059, which means that the pruned tree can classify the developed brownfield sites with high precision and just as well as the maximum tree. The specificity, however, is equal to 0.4691, which shows a 10% decrease compared to the maximum tree. This means that the pruned tree performs worse than a random model when it comes to classifying the not developed brownfield sites. Overall the maximum tree tree performs slightly better than the pruned tree, which means that the tree-pruning step does not help to

increase generalization capability of the tree model. However, the model structure of the pruned tree is much more simple and comprehensible compared to the complex maximum tree.

The final step in brownfield risk classification model is developing the random forest, which was developed as an extension to decision tree analysis in order to remove some of the problems prevalent in CART algorithm, such as overfitting the data and instability of the model in existence of noisy data [9, 10, 27]. We adopt the methodology proposed by Breiman (2001) and implemented in R in $randomForest$ package [10, 33]. The random forest classifier is developed by aggregating n_{tree} single maximum (unpruned) trees, generated on bootstrap samples of the original dataset, and randomly selected m_{try} predictors. The response is predicted by majority vote of the single classification trees. The out-of-bag data observation are used to compute the OOB error estimate, which can be regarded as the generalization error.

The developed random forest for brownfield regeneration classification demonstrates a predictive accuracy of 68.71%, which is greater than the two previous models. Upon closer inspection, it is clear than the random forest performs much better than the other two models in detecting the negative class of data (not developed brownfields), (SPE = 59.24%), but preforms slightly worse than the other two in classifying the positive class (SEN = 75.91). Overall the random forest model performs better than the other two decision tree models. The predictive accuracy of the three models differ slightly and ranges from 65% to 68%, which is generally regarded as average in predictive modeling. All three models demonstrate higher sensitivity than specificity, meaning that the detection of the positive class is executed better in all three models compared to the negative class. Since the results are consistent in various models, the limited accuracy can be a direct result of the dataset used to train the algorithms. A possible explanation can be as follows: the developed brownfield sites follow the underlying patterns detected by the algorithms, and thus, their detection rate is quite high. The brownfield sites not developed, on the other hand, are detected accurately with a much lower rate, not necessarily because of the algorithm,

but actually because many of the brownfield sites remain to be developed and at the time of data collection, they have not been revitalized, or because the brownfields have been developed in spite of low success potential of the site potentially due to lack of wide-spread knowledge in this field. Alternatively, there could be other drivers that are not included in our dataset.

In summary, by developing three different brownfield regeneration risk classification models, following the three step *Model-Construction, Model-Calibration*, and *Model-Validation* proposed and implemented initially by Altman (1968) in the credit risk modeling sector, we have attempted to bridge the missing gaps in the literature in the field of brownfield regeneration risk assessment modeling. The models are developed by utilizing two machine learning algorithms that are currently viewed as two of the most popular and powerful algorithms, namely Classification And Regression Tree (CART) and Random Forest algorithms. The Random Forest demonstrates higher predictive accuracy, and is quite fast, flexible, and easy to use for predictive modeling, and can handle large datasets containing various types of explanatory variables (continuous, discrete, ordinal, categorical). Moreover, the chances of ovefitting the data is reduced by including and aggregating several individual decision trees. The only disadvantage to the random forest model is that the response values from the algorithm are incomprehensible compared to single decision trees. The structure of CART trees along with the set of decision rules are easy to illustrate and comprehend. The lack of visualization opportunity can be regarded as a disadvantage of the random forest algorithm.

The next step in improving the predictive performance of the model is updating the dataset, where the response variable (status of brownfield regeneration) and the predictors (explanatory variables) are collected contemporaneously. Furthermore, classification of brownfield sites can be collected more in detail, in order to distinguish successful regeneration of brownfield sites from unsuccessful with a finer distinction than whether or not the site is now being utilized as a residential area to improve the predictive power of the model.

Bibliography

[1] Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance*, 23(4), pp. 589 - 611.

[2] Amekudzi, A., Fomunung., I. (2004). Integrating Brownfields Redevelopment with Transportation Planning. *Journal of Urban Planning and Development*, 130(4), pp. 204 - 212.

[3] Banasik, J., Crook, J., Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8), pp. 822 - 832.

[4] Bartke, S. (2011). Valuation of market uncertainties for contaminated land. *International Journal of Strategic Property Management*, 15(4), pp. 356 - 378.

[5] Beaver, W. (1967). Financial Ratios as Predictor of Failure, in *Empirical Research in Accounting: Selected Studies*, Supplement to *Journal of Accounting Research*, 4, pp. 71 - 111.

[6] BenDor, T. K., Metcalf, S. S., and Paich, M. (2011). The dynamics of brownfield redevelopment. *Sustainability*, 3(6), pp. 914 - 936.

[7] Bierens, H. J. (2004). Information criteria and model selection.*Manuscript, Penn State University.*

[8] Breiman, L. Friedman, J. H. Olshen, R. A. Stone, C. J. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

[9] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2), pp. 123 - 140.

[10] Breiman, L. (2001). Random Forests.*Machine Learning*. 45 (1): pp. 5 - 32.

[11] Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48, pp. 287 - 297.

[12] Brownfield Regeneration. (2013). *Science for Environment Policy*, 39, pp. 3 - 19.

[13] Chatterjee, S., Barcun, S. (1970). A Nonparametric Approach to Credit Screening.*Journal of American Statistical Association*, 65(11970), pp. 50 - 154.

[14] Claeskens, G. (2016). Statistical model choice.*Annual Review of Statistics and Its Application*, 3, pp. 233 - 256.

[15] Dasgupta, S., & Tam, E. K. (2009). ENVIRONMENTAL REVIEW: A Comprehensive Review of Existing Classification Systems of Brownfield Sites.*Annual Review of Statistics and Its Application*, 11, pp. 285 - 300.

[16] Dixon, T. (2007). The property development Industry and sustainable urban brownfield regeneration in England: an analysis of case studies in Thames Gateway and Greater Manchester. *Urban Studies*, 44, pp. 2379 - 2400.

[17] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems,*Annals of Eugenic*, 7, pp. 179 - 188.

[18] Foust, D., Pressman, A. (2008). "Credit Scores: Not-So-Magic Numbers". Retrieved 2017-04-02.

[19] Frantaal, B., Kunc, J., Klusaacek, P., Martinaat, S. (2015). Assessing success factors of brownfields regeneration: international and inter-stakeholder perspective.*Transylvanian Review of Administrative Sciences*, No. 44 E/2015, pp. 91 - 107

[20] *Fundamentals of Machine Learning.* (2016). Manuscript, Princeton University, Princeton.

[21] Gass, S. I. (1983). Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis,*Operations Research*, 31(4), pp. 601 - 663.

[22] Hand, D. J., Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring.*Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 160, pp. 523 - 541.

[23] Hand, D., Kelly, M. (2002). Superscorecards. *IMA Journal Management Mathematics*, 13(4), pp. 273 - 281.

[24] Hand, D. J., Vinciotti, V. (2003). Choosing k for Two-Class Nearest Neighbour Classifiers with Unbalanced Classes.*Pattern Recognition Letters*, 24, pp. 1555 - 1562.

[25] Henley, W. E., Hand, D. J. (1996). A k-nearest Neighbour Classifier for Assessing Consumer Credit Risk. *Statistician*, 45, pp. 77 - 95.

[26] Hernandez, M. A.,Torero, M. (2013). Parametric versus nonparametric methods in risk scoring: an application to microcredit. *Empirical Economics*, 46(3), pp. 1057 - 1079.

[27] Ho, T. K. (1995). Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16, pp. 278 - 282.

[28] Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), pp. 651 - 674.

[29] Hothorn, T., Zeileis A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research*, 16, pp. 3905 - 3909.

[30] Hosmer Jr, D. W. and Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

[31] Hooman, A., Marthandan, G., Karamizadeh, S. (2013). Statistical and Data Mining Methods in Credit Scoring. *SSRN Electronic Journal*. doi:10.2139/ssrn.2312067

[32] Lederer, T. (2009). *Context-Specific Consideration of Credit Risk Model Validation* (Doctoral dissertation). Retrieved from Publikationsdatenbank der Technischen Universität Wien. (Accession No. TUW-182150)

[33] Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest.*R News*, 2(3), pp. 18 - 22.

[34] Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining Knowl Discov*, 1, pp. 14 - 23. doi:10.1002/widm.8

[35] Louzada, F., Ara, A., Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall compari- son.Surveys in Operations Research and Management Science, 21(2), pp. 117 - 134.

[36] Ludlow, D. (2006). Urban sprawl in europe: the ignored challenge.

[37] Nathanail, C. P. (2011). Sustainable brownfield regeneration.*In Dealing with Contaminated Sites*, pp. 1079 - 1104. Springer.

[38] Ohlson. J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy.*Journal of Accounting Research* (Spring), pp. 109 - 131.

[39] Opitz, D., Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, pp. 169 - 198.

[40] Pizzol, L., Zabeo, A., Klusacek, P., Giubilato, E., Critto, A., Frantal, B., Martinat, S., Kunc, J., Osman, R., and Bartke, S. (2016). Timbre brownfield prioritization tool to support effective brownfield regeneration.*Journal of environmental management*, 166, pp. 178 - 192.

[41] R Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL *//www.R-project.org/.*

[42] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

[43] Sabato, G. (2010). Assessing the Quality of Retail Customers: Credit Risk Scoring Models,*The IUP Journal of Financial Risk Management*, 7(1 & 2), pp. 35 - 43.

[44] Strasser, H., Weber, C. (1999). On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*, 8, pp. 220 - 250.

[45] Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*, Manuscript, University of Minnesota.

[46] Thornton, G., Franz, M., Edwards, D., Pahlen, G., Nathanail, P. (2007). The challenge of sustainability: incentives for brownfield regeneration in Europe. *Environmental Science & Policy*, 10, pp. 116 - 134.

[47] Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications* (Master's thesis). Retrieved from *https://edoc.huberlin.de/handle/18452/4*.

[48] Vojtek, M., Koèenda, E. (2006). Credit Scoring Methods, *Czech Journal of Economics and Finance*, 56(3-4), pp. 152 - 167.

[49] Yager, R.R. (1988). On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Trans. Syst. Man. Cybern*, 18, pp. 183 - 190.

[50] Zabeo, A., Pizzol, L., Agostini, P., Critto, A., Giove, S., Marcomini, A. (2011). Regional risk assessment for contaminated sites Part 1: Vulnerability assessment by multicriteria decision analysis.*Environ. Int*, 37, pp. 1295 - 1306.

[51] Zweig, M. H., Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 29, pp. 561 - 577.

Appendix A

```
library(caTools)
library(grid)
library(partykit)
library(MASS)
library(pROC)
library(caret)
library(ROCR)
library(randomForest)
library(reprtree)

# Dataset is the set of variables which includes all dependent and
# independent parameters.

# Set.seed function is used so that sample.split function produces
# the same subsets each time the program is run.

set.seed(120)

# Sample.split is used to split data from dataset into two subsets
# (train and test) in predefined ratio, set by SplitRatio while
# preserving relative ratios of different labels in dataset.

sample <- sample.split(dataset, SplitRatio = 0.75)
train <- subset(dataset, sample == TRUE)
test <- subset(dataset, sample == FALSE)
```

```
# Ctree function performs recursive partitioning for nominal,
# ordered, continuous, and multivariate response variables
# in a conditional inference framework. The train set is
# used construct and calibrate the Classification and
# Regression Tree (CART). The dependent variable (Usage)
# is partitioned against the independent variable set.
# The resulting CART is then plotted. In this step, the
#  maximal tree is produced without any trimming.

CART <-ctree(Usage~.,data = train)
plot(as.simpleparty(CART))

# After the construction and calibration of the maximum
# tree, CART, it is validated using the test subset.
# Predict is a generic function for predictions from the
# results of model fitting functions, namely CART.
# Table function produces the confusion matrix for the
# validation procedure results and the overall accuracy.

CART.predict = predict(CART, newdata=test)
Conf_Mat_CART<-table(predict=(CART.predict),actual=test$Usage)
CART_Acc<-(1-mean(CART.predict == test$Usage))

# An approach used to visualize the predictive power
# of a classification model is the ROC curve. Roc function
# from the pRoc package builds a ROC curve.

CART_ROC <-roc(as.numeric(CART.predict),as.numeric(test$Usage))
plot(CART_ROC)
```

```
# The second model is the pruned-tree. Ctree function is
# utilized similar to the previous model. However, ctree_control
# is used, which includes various parameters that control
# aspects of the tree construction and tree pruning.

c <-ctree_control(teststat = c( "max"),
                  testtype = c("Univariate"),
                  mincriterion = 0.95, minbucket = 1800L,
                  minprob = 0.1, stump = FALSE, mtry = Inf,
                  maxdepth = Inf, multiway = FALSE,
                  splittry = 2L, majority = FALSE)
CART_Pruned <-ctree(Usage~.,data = train, control = c)

# Test subset is used to validate the pruned-tree.
# Confusion matrix depicts the validation results.
# ROC curve further visualizes the results.

Pruned.predict <- predict(CART_Pruned, newdata=test)
Conf_Mat_CART_Pruned <-table(predict= (Pruned.predict),
                             actual= test$Usage )
CART_Pruned_Acc <-(1-mean(Pruned.predict == test$Usage))
CART_Pruned_ROC <-roc(as.numeric(CART.predict),
                      as.numeric(test$Usage))
plot(CART_Pruned_ROC)

# Random forest is constructed using the randomForest
# function randomForest package, which implements
# Breiman's random forest algorithm for classification
# and regression. ntree sets the number of trees to grow.
```

```
# This should not be set to too small to ensure that
# every input row gets predicted at least a few times.
# mtry is the number of variables randomly sampled as
# candidates at each split.

Rand_Forest <- randomForest(Usage ~ ., dataset ,ntree=500,
                        mtry=5, importance=TRUE, do.trace=100)
print(Rand_Forest)
Rand_For_CM <- confusionMatrix(data=Rand_Forest$predicted ,
                  reference=dataset$Usage,positive = "1")

# Following graph depicts the effect of number of decision
# trees on the error rates of the random forest.

layout(matrix(c(1,2),nrow=1), width=c(4,1))
par(mar=c(5,4,4,0))
plot(Rand_Forest , log="y")
par(mar=c(5,0,4,2))
plot(c(0,1),type="n", axes=F, xlab="", ylab="")
legend("top", colnames(Rand_Forest$err.rate),col=1:4,
        cex=0.8,fill=1:4)

# varImpPlot visualizes variable importance measures
# obtained by the random forest classifier.

varImpPlot(Rand_Forest)


# ROC curve of all three models depicted for better
# comparison in one graph.
```

```
C_p <- predict(CART, newdata=(test), type="response")
C_pred <- prediction(as.numeric(C_p), as.numeric(test$Usage))
C_Perf <- performance(C_pred, measure = "tpr", x.measure = "fpr")
plot(C_Perf,xlab="1- Specificty (False Positive Rate)",
     ylab="Sensitivity (True Positive Rate)",col="red")
lines(c(0,1),c(0,1),col="grey")
par(new = TRUE)

P_p <- predict(CART_Pruned, newdata=(test), type="response")
P_pred <- prediction(as.numeric(P_p), as.numeric(test$Usage))
P_Perf <- performance(P_pred, measure = "tpr", x.measure = "fpr")
plot(P_Perf,xlab="1- Specificty (False Positive Rate)",
     ylab="Sensitivity (True Positive Rate)",col="green")
lines(c(0,1),c(0,1),col="grey")
par(new = TRUE)

R_pred <- prediction(as.numeric(Rand_Forest$predicted),
                     as.numeric(dataset$Usage))
R_Perf <- performance(R_pred, measure = "tpr", x.measure = "fpr")
plot(R_Perf,xlab="1- Specificty (False Positive Rate)",
     ylab="Sensitivity (True Positive Rate)",col="blue")
lines(c(0,1),c(0,1),col="grey")
legend(0,1,c("Random Forest","Maximum Tree","Pruned Tree")
       ,cex=0.8,lty=c(1,1),lwd=c(2.5,2.5),col=c("blue","red","green"))

# Area under the curve for all three models is calculated.
#For a classification model 1 represents a prefect classifica-
#tion model and 0.5 is a random model.
```

```
Max_Tree_auc <- performance(C_pred, measure = "auc")
Max_Tree_auc  <- auc@y.values[[1]]
Max_Tree_auc

Pruned_Tree_auc <- performance(P_pred, measure = "auc")
Pruned_Tree_auc  <- auc@y.values[[1]]
Pruned_Tree_auc

Rand_Forest_auc  <- performance(R_pred, measure = "auc")
Rand_Forest_auc  <- auc@y.values[[1]]
Rand_Forest_auc
```