Zerayaeqob Mengesha

Data-driven decision support to reduce “driving-under the influence of alcohol” offenses

Master’s thesis of Faculty of Information Technology

January 15, 2018

University of Jyväskylä

Faculty Information Technology

Author: Zerayaeqob Mengesha

Contact information: zekeymen@gmail.com

Supervisors: Prof. Miettinen Kaisa and Dr. Markus Hartikainen

Title: Data-driven decision support to reduce "driving under the influence of alcohol" offences.

Työn nimi: Datapohjainen päätöksenteon tuki rattijuoppoustapausten ehkäisyssä

Project: Master's thesis

Study line: Web Intelligence and Service Engineering

Page count: 74+2

Abstract: Extracting valuable knowledge from data to support decision making is a widely practiced trend. Data-driven decision support (DDDS) provides insight for decision makers by exploring and extracting underlying patterns within a dataset. This thesis covers the process of DDDS in reducing driving under the influence of alcohol (DUI) offenses by introducing proposed prison sentences. In this thesis, DDDS is applied to a DUI dataset by analyzing patterns in the dataset and by introducing proposed prison sentences for offenders to reduce the number of DUI cases. Background theories in data mining, machine learning, optimization and decision science that are related to the thesis project are also covered. Furthermore, the thesis presents the application of data analysis and multiobjective optimization, in formulating and optimizing objective functions representing DUI reduction. The results obtained from the analysis show that, by grouping individuals with similar DUI patterns and by introducing different proposed prison sentences for each group, it is possible to provide decision support that can reduce the number of DUIs at certain time intervals.

Keywords: Data-driven optimization, Decision-support, DUI, Multiobjective optimization

Suomenkielinen tiivistelmä: Abstract in Finnish…

Avainsanat: Datapohjainen optimointi, päätäksenteon tuki, rattijuoppous, monitavoiteoptimointi

Glossary

BBN	Bayesian Belief Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
DDDS	Data Driven Decision Support
DM	Data mining
DUI	Driving under the influence of alcohol
ID	Influence Diagram
MOO	Multiobjective Optimization
SOO	Single Objective Optimization

List of Figures

List of Tables

Contents

1 Introduction

We generate data in our daily activities whenever we browse the internet, use our Smartphone or visit social media sites. The data generated online through web browsers can easily be used by advertising companies for targeted marketing. Most companies collect and store data for different purposes. According to a report by the McKinsey Global Institute (MGI) [1], most American companies with more than 1000 employees have an average of 200TB stored data. The report has also predicted that the global generation of data will increase by 40% annually.

Big data is defined as data with the following properties: large in volume which requires a special storage consideration, which consists of multiple types and sources, generated at a high rate and contains noise [2]. The trend of extracting valuable knowledge from big data is attracting various sectors to utilize the big data they have stored in their systems for decision support [3]. Accurately and carefully collected clean data could serve as a basis for big-data analysis, in order to launch new services or improve existing ones. According to Larose, data-driven decision support has benefited institutions to reduce cost and detect frauds [4].

There are different types of data analysis techniques and Smith [5] in his article mentioned some of the common types of data analysis.

- Descriptive data analysis to quantitatively describe the main features of the data.
- Exploration data analysis to discover new connections to serve as a basis for future studies.
- Inferential data analysis to generalize some properties about the population based on sample data.
- Predictive data analysis to make predictions for unknown data based on the available data.

- Causal data analysis to investigate the effect of a variable when the value of another variable is altered.

The data analysis technique which is directly related to decision support is prescriptive analytics [6] that enables decision-makers to benefit from the results of descriptive and predictive analysis by taking the necessary course of actions in a timely manner. One of the key attributes for data-driven decision support is the data itself, and for data collection, digital data collection is the preferred method whenever it is applicable. Traditional data collection methods are associated with collecting data through paper questionnaires and interviews, and this method has some drawbacks compared to the digital data collection. Traditional data collection techniques are expensive both in terms of time and money, difficult to do at a larger scale and they are susceptible to human errors [7].

Once the data is collected and stored in a repository, data mining techniques can be applied to discover valuable knowledge from the data, which can be used as an input for decision support. Data mining techniques can be applied to a variety of sectors for predicting future outcomes or classifying a problem set into groups. According to Sasha Issenberg [8], during the 2012 American election campaign, the Obama campaign used data-mining techniques to identify prospective voters and encourage them to head to the polls during the Election Day. They have also used predictive models to predict the outcome of the election in some swing states, in order to allocate scarce resources to these states. In one state the model predicted the outcome of the election with a 99.98% accuracy [8]. This is one good example of the importance of data-mining in predicting future outcomes and supporting decision making.

According to a report by the Finnish road safety authority [9], alcohol accounts for approximately one-quarter of the deaths and one tenth of the injuries occurred in European road traffic. In Finland in 2012, 41 deaths and 617 injuries were caused by driving while intoxicated [9]. Driving under the influence of alcohol (DUI) is a problem for law enforcement

officials and they want to deal with the problem more seriously. Currently, there are three types of punishments for DUI offenders based on the severity of the crime and the level of alcohol found in the offender's blood. The punishments are: paying fines, serving community service and prison sentences. There are drivers who keep repeating the crime and therefore there is a need to restructure the sentencing of repeated DUI offenders to reduce the number of DUI offenses committed by repeated DUI offenders. We consider in this thesis those DUI offenses which are punishable by prison sentences. Moreover, we are interested in dealing with repeated DUI offenders, i.e. drivers who recommit DUI offenses once they have served their punishment. We aim at coming up with a proposed prison sentence for repeated DUI offenders based on the data related to their past DUI related offenses.

The data we are working on is from two datasets obtained from the University of Helsinki, Institute of Criminology and Legal Policy. The data had been treated before we received it to conceal the identity of the individuals. The first dataset contains over 50,000 records along with 40 attributes and the second dataset contains over 4000 records with 32 attributes. The second dataset was already preprocessed and records that were not relevant to the thesis project had been filtered out. It only contains individuals whose first prison conviction was from 2004 to 2007. Two attributes that were not included in the first dataset were included in the second dataset: the exact number of days an offender had served in prison and information if the offender was dead or had moved to another country.

The objective of this thesis is to investigate whether there is a possibility to support the decision making of re-sentencing repeated DUI offenders to reduce the number of DUIs committed by repeated offenders without significantly increasing the expenses of the punishments incurred to the society. By doing so, we try to come up with optimal proposed prison sentences that would reduce the number of DUIs committed by repeated offenders by taking into account the cost associated with keeping them in prison for longer periods of time. The approach we will follow is first to examine if there is a correlation between

committing the next DUI and the other attributes in the data. Specifically, if there is a correlation between the length of a prison sentence and committing the next DUI so that we can easily predict the next DUI after serving time in prison for different proposed sentences.

The second approach will be to group offenders based on their past DUI offenses and propose different prison sentences for each group. This means that the group with a high probability of committing a new DUI in the first few weeks after the release will get longer prison sentences and subsequently, the number of DUIs committed in the first few weeks will be reduced because those high-risk groups will be held in prison for a longer time. To balance the cost of extending the sentences of high-risk groups, groups with a minimum rate of committing a new DUI will get reduced proposed sentences. In both approaches, there are two conflicting objectives: reducing the number of DUIs by proposing additional prison sentences and reducing the cost of incarcerating offenders for the newly added proposed sentence. To handle the tradeoff between these two conflicting objectives, multiobjective optimization will be used. The approaches will be discussed in details in chapter 6 after the analysis of the DUI data. The objective functions for the multiobjective optimization will be defined from the dataset. For the number of DUIs, we will use a checkpoint at a certain period of time. When a proposed sentence reduces the length of the original sentence, the number of DUIs at the checkpoint tends to increase and vice versa. Similarly, the cost function will be defined as the number of offenders multiplied by the number of days in the proposed sentencing and when the proposed sentence reduces the original sentencing, the cost function decreases and vice versa.

The thesis is structured in the following way: Chapters 2 to 5 give a general overview of the different fields covered in the thesis. They include background research of data mining, machine learning, optimization and analytics. Chapter 6 gives an insight into the data analysis techniques used and provides a description of the data and the applied data analysis and optimization techniques to produce the results. Chapter 7 discusses the results obtained and their interpretations. Finally, Chapter 8 presents the conclusions of the thesis.

2 Data Mining

2.1 Introduction

Data mining (DM) is the process of extracting valuable knowledge by discovering new correlations, patterns and trends by examining large amounts of data stored in repositories, through pattern recognition, statistical and mathematical techniques [4]. DM focuses on finding data patterns that provide insight or enable fast and accurate decision-making [10]. It can also be viewed as finding a small set of precious results from a large volume of raw materials [11]. DM can be used either to verify a stated hypothesis or to discover new knowledge [12]. The discovered underlying patterns through DM are represented using models. Models are expressed in terms of mathematical expressions and models under similar groups share similar forms. DM can be applied to any dataset as long as the desired patterns to be discovered reside in the dataset.

Terms and concepts used in this chapter are defined as follow. Record refers to a full set of values of data properties in a data collection. In a tabular data, a record is represented by a data row. Attributes are data properties which describe records, in a tabular data they are represented by data columns. Dataset refers to a collection of data and it usually refers to a single database table in tabular data. Training and testing data are subsets of a dataset used to train and test a model with known input and output attributes. Training a model refers to adjusting the model parameters based on patterns discovered from the training dataset. The performance of a model is assessed using test data. Hence the outcomes of the test data are known, the predicted outcomes of the test data from the model will be compared with the actual outcomes of the test data to determine the performance of the model. Data exploration is the process of describing the data statistically and visually. Data exploration is an important analysis phase to decide which data attributes need further analysis. Evolution analysis tries to discover trends whose values change over time.

2.2 Data mining process

Data mining process includes different phases to attain the final goal of DM i.e. to discover hidden patterns. Cross-Industry Standard Process for Data Mining (CRISP-DM) is a standard data mining process which defines the different phases of DM processes [13].

- Business understanding defines what needs to be accomplished. This phase includes investigating the business objectives and requirements, setting data mining goals, assessing the current situation and preparing a project plan. In the context of this thesis project, we consider business understanding phase as defining the goals to discover patterns related to DUI offenses.
- Data understanding phase considers data requirements. This phase may include establishing initial dataset, data description, data exploration and verification of data quality. If the quality of the data is not good enough, collecting new data might be required.
- Data preparation phase includes the process of data cleaning, data integration, data reduction and data transformation.
    - Data cleaning: Data in its original state is usually incomplete and noisy. To deal with these issues data needs to be preprocessed. Data cleaning tries to handle data values that are expired or missing. Careful analysis of the situation is essential to determine the steps needed to be taken to deal with missing data values. According to Larose [4], the easiest way to deal with missing values is to omit all the records with missing values. But omitting a record for some missing values is not a wise decision and it could lead to a biased subset of the data and instead of omitting the record for missing values, there are suggested ways to replace them with some other values. Larose [4] described three approaches. The first approach is replacing the missing value with a constant. The second approach is to set the average or the mode, a value which appears most frequently, of the data attribute which contains the missing value in place of the missing value. The third one is to randomly generate a value among the existed values of the data attribute which contains the missing value.

    - Data integration is the process of integrating different data sources such as multiple databases, data cubes, which is a three or higher dimensional array of values; or files.
    - Data reduction is the process of reducing the dimension of the data into a much smaller volume that could produce almost identical results to the original data. Data reduction saves processing time by reducing the volume of the data.
    - Data transformation is the process of formatting the data into a format suitable for DM. It includes removing noise from data, aggregation and normalization of the data. Normalization is the rescaling of different data attributes into a common scale.
- Modeling is the phase where selecting and applying of model types come to effect. In the modeling phase, various model types will be applied and their parameter values will be tuned up to be optimal. It also involves the assessment of models and revision of parameter settings.
- Evaluation is the phase where the data mining results will be assessed against the defined business objectives. Evaluation involves reviewing the construction of the model to assert whether the business objectives are achieved. The final output of this phase is making a decision whether to use the data mining results.
- Deployment: the final phase of the data mining process is to organize and present the data mining results so that it can be applied to business operations. Depending on the area of operation, models need to be monitored. As some models might be valid only for a certain period of time and if there is a significant change in the operation, the model must be revised.

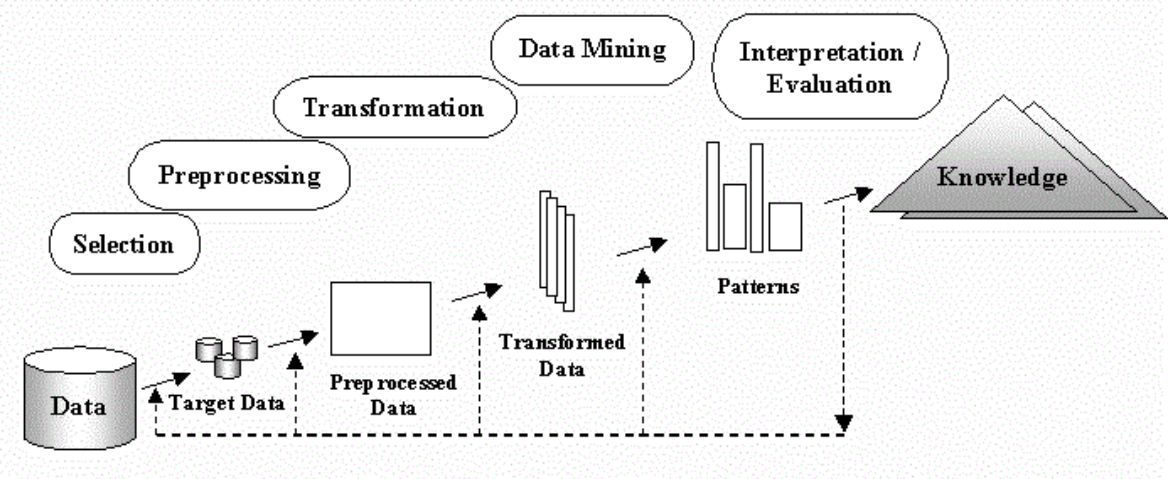

Figure 1. Data mining processes [14]

Figure 1 shows the different stages of data mining processes from selection to interpretation and evaluation.

Characteristics of data mining tools

DM tools need to possess a set of characteristics in the form of scalability, versatility, capability to predict outcomes accurately and automatic implementation [12]. Scalability refers to the capability of the DM tool to handle large scale datasets. Versatility means the DM tool needs to work with different model types. Automatic implementation is not always applicable as human decision making might be required to conduct procedures on how the implementation should be carried out.

2.3 Discovered patterns

DM functionalities are used to determine the kind of patterns that can be discovered from the dataset to be mined (the target dataset). Usually DM is categorized as descriptive or predictive mining [11]. Descriptive mining is used to describe the properties of the target data, while predictive mining is used to forecast future values or missing values based on the analysis of the target dataset.

Class description

Class description belongs to descriptive mining and tells the relationship of a data record to a class or a category. There are two approaches of descriptive mining: data characterization and data discrimination. Data characterization produces a summary of the different features of the target data. On the other hand, data discrimination produces a summary of features by comparing the target class, the class under consideration for data discrimination, with other classes in the dataset [11].

Classification

Classification in DM is the process of assigning classes or categories to records in a dataset. Classification is used with categorical data whereas regression is used with continuous valued functions, numeric function defined over an interval [11]. Classification rules can be represented based on if-then conditions, decision trees or neural networks. If – then rules are textual rule representations to determine the class of an item in a dataset. Decision trees are tree-like data structures and they can be used for both regression and classification. Decision tree internal nodes can represent attribute tests, whether an attribute has certain properties, tree branches represent the result of the attribute test and leaf nodes represent classes. Neural network can also be used for classification with weighted connections between nodes. The weight in a neural network is a number that controls the signal between two nodes.

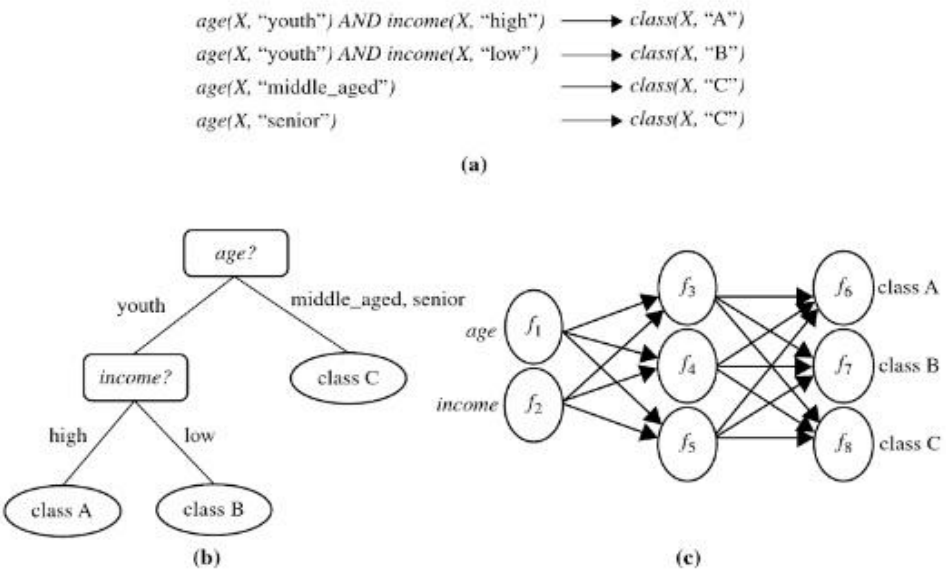

Figure 2. Representation of classification a) if –then b) decision tree c) neural network [11]

Figure 2 shows the different methods to carry out a classification task for a classification example based on age and income attributes.

Regression

Regression models are used for continuous functions to predict missing or unavailable values. For a regression model, independent attributes that are considered to be important for predicting the target outcome need to be identified. Once the independent attributes are identified, training data is used to train the regression model in order to estimate the parameters of the model. To measure the quality of the regression model, the residual sum of squares, the squared sum of the differences between the actual and the predicted target values can be used.

Let $x_1, x_2, \ldots x_n$ be the independent attributes, $y_1, y_2, \ldots y_n$ be the target output attributes in the training data and $\hat{y}_1, \hat{y}_2 \ldots \hat{y}_n$ be the predicted target values for the same independent attributes $x_1, x_2, \ldots x_n$.

The Residual Sum of Square (RSS) is $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

There are different types of regression.

1. **Linear regression**
   Linear regression is the association of one or more independent attributes with one target attribute. Simple linear regression has only one independent attribute and multiple linear regression has more than one independent attributes. The general form of linear regression is: $Y = aX + b + e$

   where *Y* is an estimate of the target attribute, *X* the independent attribute, *a* weight of the attribute *X*, *b* the constant term and *e* the prediction error.

2. **Nonlinear regression**

Nonlinear regression is a regression analysis model in which the model is based on a nonlinear function, a function whose graph is not a straight line. To design a nonlinear regression model, the type of nonlinear model will be determined first and then the parameters must be estimated and the error should be minimized using different techniques. As described in [15], the most common methods to minimize error are Least Square, Maximum Likelihood and Quasi-Newton method. Least squared method tries to minimize the residual sum square. Maximum Likelihood is an estimation method used to estimate the parameters of a statistical model for a given dataset. Quasi-Newton method is used to find the local minima or maxima of a function and it can be used to estimate nonlinear regression model parameters.

Clustering analysis

Class labels are not always present at the beginning of data analysis in a dataset to determine to which group a record belongs to. So, in those situations clustering analysis is a good method to mine patterns, to form groups of records or clusters. Clusters are formed by grouping similar items in one cluster and dissimilar ones in other clusters. The resulting clusters could serve as a basis to define classes and formulate rules [16]. In clustering there are four steps to form clusters. The first step is feature/attribute selection, determining features that distinguish one cluster from the other. The second step is determining the appropriate clustering algorithm to form the clusters. The third step is validating the results. Once the clusters are formed the result has to be validated whether it gives a meaningful answer for the selected features and clustering algorithm. The final stage of clustering is result-interpretation i.e. interpreting the results for its intended use. The detailed use of clustering and clustering techniques can be viewed in [16].

Outlier analysis

In a dataset, there might be records which lie outside the regular ranges of fields. Usually, these values are considered as noises in the dataset and they could be discarded during a data mining process. But these rare happenings are important for fraud detection and anomaly analysis, which is an analysis of unexpected patterns [11]. Therefore the concept

of outlier analysis highlights the analysis of records which deviate from the general behavior of the dataset to discover meaningful results.

2.4 Challenges in data mining

The authors of [11] grouped the challenges of data mining into five categories. The authors also mentioned that some of the mentioned challenges are already addressed but the other challenges are still being researched.

Mining methodology

To discover knowledge from data for various disciplines using DM techniques, DM should incorporate a variety of mining methodologies. This is due to the fact that in different disciplines users have different interests in the knowledge to be discovered. For the purpose of generating a wide variety of results when conducting DM tasks on a dataset, different mining methodologies from data description to evolution analysis should be considered. As a result, identifying the right methodology which produces the intended outcome is a challenge in DM.

Interaction issues and data visualization

DM results are not known at the start of a DM process, therefore a portion of the dataset is analyzed to explore the data first. Analyzing portion of the dataset gives an insight for exploring the data and to generate initial results. Intermediate results need to be refined by user interaction with the DM system, to view discovered patterns and to produce results from different perspectives. In order to make DM results usable and easily understandable, data presentation is necessary. Especially if the user has to interact with the DM system in the data mining process, data presentation helps the user to make decisions easily. For visualization purposes results are presented in the form of trees, tables, graphs and charts. Consequently, providing interactivity and visualization for multidimensional dataset is a challenge for DM systems.

Performance issue

To discover knowledge from large datasets, DM algorithms need to be effective and scalable. Scalability is the capability to deal with large volumes of data. Effectiveness of a DM algorithm refers to the efficiency of the running time needed to deliver the expected results. The running time of a DM algorithm should be predictable and acceptable [11]. To reduce the running time, parallel and distributed algorithms can be used. Parallel and distributed algorithms partition the dataset into smaller volumes and process the partitions in parallel. Finally, the algorithms merge the results of the partitions together to reduce the running time. Researches in cloud and cluster computing are also advancing to deliver distributed solutions to handle large-scale computation be mined in parallel [11].

Diversity of data formats and perception of data mining in society

Datasets contain diverse data formats which mean that a single DM system cannot work with all the different data formats. Therefore different data formats may require different DM systems. Similarly, the data might come from a distributed heterogeneous datasets and working with heterogeneous data is a challenge for DM systems.

According to [11], DM has been used to discover knowledge from data mostly for scientific and business purposes. From the society point of view, there are still concerns about misuse and disclosure of personal data. Therefore, [11] suggested more work need to be done on DM projects by addressing the concerns of the society to guarantee privacy-preservation.

3 Machine Learning

3.1 Introduction

This chapter covers the basics of Machine Learning (ML), classification of ML and types of ML algorithms related to the thesis project. ML is not a new field and it has been around for decades. In the past, different computer programmers experimented how to teach computers [17]. Based on the definition of learning, the ability to acquire facts, skills and abstract concepts, researchers tried to implement the same notion in computers [18]. According to [17], ML is the process of programming computers to model associations of attributes based on example data or past experience, by trying to optimize the parameters of the model, to represent the underlying relationship, for the purpose of predicting future outcomes or understanding the data. ML focuses on designing efficient learning algorithms and hence the success of learning algorithms depends on the data, ML has a strong association with data analysis and statistics [19].

Supervised learning

Supervised learning is a learning environment where labeled data, a data record along with a class name or with the target outcome value, is provided to the ML algorithm to discover the association between the labels and the attributes of the data. Then it uses the discovered association to predict the labels of the unlabeled data. The two most common supervised learning tasks are classification and regression, and both of them have been already introduced in chapter 2. There are two terms associated with supervised learning: training data and test data. Training data is a sample of the data used to train the model to discover the known associations in the data. On the other hand test data is a sample of the data which is not available during the training, when the model is trained with the training data, and it is used to evaluate the performance of the model [19]. Using these two data subsets a particular model is first trained and then tested to evaluate how it captures the existing associations.

Unsupervised learning

Unsupervised learning in the context of ML is the process of learning associations and patterns without knowing the outcome/label of data. The ML algorithm tries to find the underline patterns by itself without the presence of predefined labels/outcomes. The most common unsupervised learning task is clustering. As introduced in Chapter 3, clustering is the process of forming groups based on a similarity measure; items similar to each other and dissimilar to other groups are clustered together [20].

Semi-supervised learning

Semi-supervised learning is a mix of both supervised and unsupervised learning containing supervision information on the subset of the data [21]. The data can be divided into two parts, one containing labeled data and the other part containing unlabeled data [21]. In semi-supervised learning, the ML algorithm formulates the underline patterns based on the available labeled data.

Reinforcement learning

In ML, reinforcement learning got the fundamentals from Markov Decision Processes (MDP). The main goal of reinforcement learning is to learn how to take the right actions in an environment where there are only feedbacks for the right actions in the form of rewards [22]. Further details of Markov Decision Processes can be found in [22]. The decision making machine or agent takes an action in an environment and gets a penalty or a reward based on the action it made. Consequently, the machine tries to solve a problem by adopting best policies. The sequences of actions are usually followed by a cumulative reward [20]. In reinforcement learning, exploration refers to trying out new actions to learn from their consequences. On the other hand, exploitation refers to the utilization of already learned set of actions. Even if exploration may slow down the performance of the machine as some new actions might result in penalties, it is the best way for the machine to improve its performance by learning new actions [22].

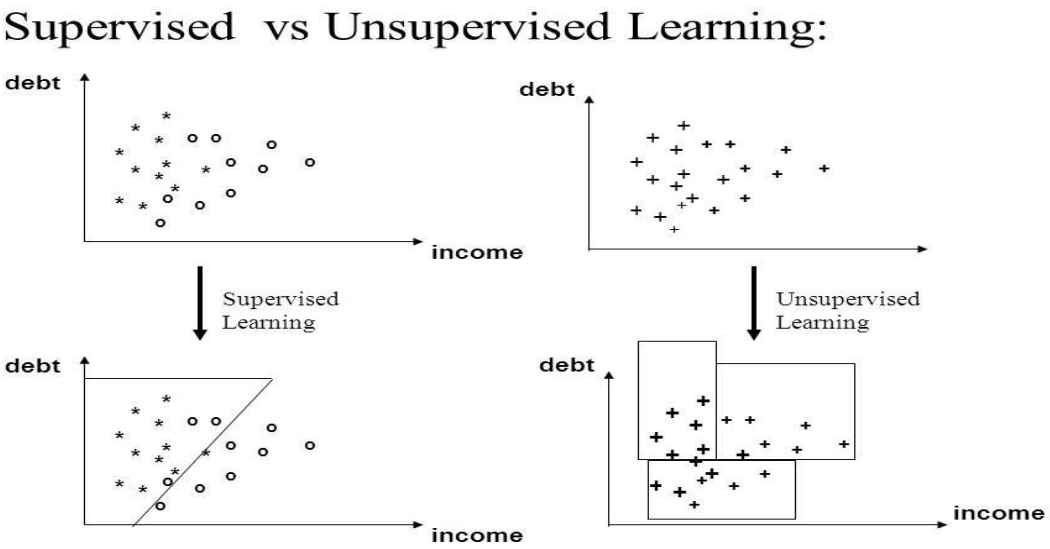

Figure 3. Supervised versus unsupervised learning [23]

Figure 3 shows the distinction between supervised and unsupervised learning. In supervised learning the items are labeled while in unsupervised learning the items are not labeled and the learning algorithm groups items based on how the item attributes are similar to each other.

3.2 Machine learning algorithms

ML has different algorithms to handle different ML problems. The most common algorithms for ML are used to carry out tasks in regression, classification and clustering. As most of those concepts are already covered in the previous section, decision tree, clustering and deep learning will be discussed in this chapter.

Decision tree

Decision tree is used for discrete problem sets of supervised learning. It uses a set of rules or tests to categorize the input data. A decision tree has nodes, branches and leaves. Nodes represent a function to test the input attribute(s). The outcomes of the tests are marked on the branches and finally the labels are represented by the leaves [17].

The way a decision tree works is that it has no fixed structure and it creates branches and leaves based on the complexity of the problem set during the learning process. Its hierarchical structure makes it perform better by eliminating unnecessary comparisons i.e. once a comparison is made at a node and if successive comparisons are required, they will be made only at the nodes that are related to the outcome of the test. Furthermore, complex rule sets will be broken down into simple decisions and eventually, depending on the type of the given problem set, the leaf nodes represent class labels or numeric values [17]. In their book Witten and Frank [24] described in details how a divide and conquer algorithm was derived to develop decision trees and how C4.5, which is a decision tree algorithm, has emerged as a popular decision tree algorithm.

Clustering

Clustering has already been defined in chapter 2. In this chapter clustering will be discussed from the perspective of how it can be implemented. In [25], Bell described in details how clustering can be applied to different application areas. Clustering is used in unsupervised learning to form groups of similar items. The most common algorithm for clustering is KMeans [25]. To perform clustering using KMeans, first the cluster size needs to be determined. At times, it can be difficult to determine the numbers of clusters intuitively. To overcome this problem there are methods to estimate the number of clusters [25]. Once the number of clusters is defined, the algorithm sets points (centroids) as the center of each cluster. Each time a new data item is processed, it reshuffles the clusters by moving the centroids as the mean value of the items in the cluster which contains the centroid, and iteratively the algorithm assigns items to the closest cluster.

KMeans algorithm:

1. Place K number of points into the n-dimensional space represented by the items' attributes that are being clustered. These K points represent initial cluster centroids.
2. Assign an item to a cluster based on the minimum Euclidean distance between the item and the cluster's centroid.

3. Recalculate the positions of the K centroids as the mean value of the items' attributes in the clusters.
4. Repeat steps 2 and 3 until the centroids no longer move.

Deep learning

Deep learning is a branch of ML that deals with artificial neural networks. The study of identifying components of a learning system, which are responsible for the success and failure of a system, has been around for decades [26]. Schmidhuber [26] in his article explained how a standard neural network operates by introducing neurons as simple connected processes, each capable of activating other neurons. He also mentioned that input neurons get activated through environment-perceiving sensors while other neurons are activated by previously activated neurons. Some neurons may trigger actions that influence the environment. Nielsen [27] described the structure and application of artificial neural network and the application areas of deep learning. From the structure point of view, there can be multiple layers of neurons in an artificial neural network. The application areas of deep learning are vast and it is currently applied to for example in image recognition, speech recognition and natural language understanding [27].

4 Optimization

4.1 Introduction

This chapter covers the basics of optimization, types of optimization problems, optimization algorithms and brief overview of multiobjective optimization. Optimization tries to find the best solutions that minimizes or maximizes an objective function in an optimization problem. The optimization problems could be for instance, to reduce wastage, maximize profit or to improve the performance and efficiency of operations. Optimization can be applied to a range of different areas: logistics, manufacturing, resource allocation, sales and marketing, and optimization provides significant benefits both for the industrial and scientific world [28].

4.2 Types of optimization problems

Linear programming:

Nemhauser et al. [29] mentioned that even if linear programming has a longer history, it was emerging in the late 1980s and its development was fuelled by its wide applicability. The basic characteristic of linear programming or linear optimization problem is the linearity of the objective function i.e. the function that will be minimized or maximized, and its constraints are linear. The constraints can be either equality or inequality constraints.

Mathematically a linear optimization problem can be expressed in either of the two forms.

The inequality form:

$$\begin{aligned} &\text{minimize } z = c^T x \\ &\text{subject to } \mathrm{A}x \leq b, \\ &x \geq 0 \end{aligned} \qquad (1)$$

and the standard form:

$$\begin{aligned} &\text{minimize } z = c^T x \\ &\text{subject to } Ax = b, \\ &x \geq 0 \end{aligned} \qquad (2)$$

where A is an (m x n) matrix, z $\in \Re$, c$\in \Re^n$, b $\in \Re^m$, x$\in \Re^n$, m and n are the dimensions of the matrix.

Integer linear programming

A pure integer linear programming problem has the same form as (2) with the exception that x $\in Z^n$. However, in mixed-integer linear programming at least one of the variables involved is an integer [30].

The standard matrix form of mixed-integer programming is:

$$\begin{aligned} &\text{maximize } z = c^T x + d^T y \\ &\text{subject to } Ax + Gy \leq b, \\ &x \geq 0 \\ &y \geq 0 \end{aligned} \qquad (3)$$

where m = number of constraints

n = number of continuous variables
p = number of integer variables
c^T = is a row vector of n elements
d^T = is a row vector of p elements
A = is an m x n matrix
G = is an m x p matrix
b = is a column vector of m constraints
x = is a column vector of n nonnegative integer variables
y = is a column vector of p nonnegative real value variables

Nonlinear optimization

Nonlinear optimization problem is generally defined as follows: for a given set $D \subset \Re^n$ and for a nonlinear continuous function f: D ->$\Re$,

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g_i(x) \geq 0, \\ &\qquad h_j(x) = 0 \end{aligned} \qquad (4)$$

where x is a vector and $x \in D$, g_i and h_j for i=1,2,..p and j=1,2,... m are real valued functions defined in D [31].

In addition to the above mentioned optimization problems, there are other optimization problems including convex, concave and nonconvex functions.

Convex functions are defined over a convex set to have a convex problem. The formal definition of a convex set is: A subset C of the Euclidean space R^n is said to be convex if for every x, y $\in$ C and $\lambda \in \Re$, $0 < \lambda < 1$, the point $\lambda x+(1-\lambda)y \in C$ [32].

Geometrically, a set C is said to be a convex set if for every two element x, y in C, the line segment joining these two points lies entirely in C.

Similarly, a function f:C→R where $C \subset R^n$ and C is a convex set is said to be a convex function if $f((1-\lambda)x+\lambda y) \leq (l-\lambda)f(x)+ \lambda f(y)$ for x, y $\in$ C and $\lambda \in$ [0,1] [32].

In other words a function f is said to be convex if for any two points x and y in C, the line segment connecting x and y lies above or on the graph of f.

In the same way, a function f is called concave if and only if -f is a convex function i.e. for any two points x and y in C, the line segment connecting x and y lies below or on the graph of f. Nonconvex functions are neither convex nor concave functions and they might have multiple local minimum points. One example of nonconvex function is the sin function

g(x) = sin(x), for x $\in [-2\pi,2\pi]$.

Classification of optimization algorithms

Optimization algorithms can be classified as deterministic and stochastic ones. Deterministic optimization depends heavily in the mathematical computation of the gradient or Hessian of the objective function whereas stochastic uses randomness in the optimization algorithm. Deterministic optimization algorithms converge quickly with fewer numbers of computations and the solutions found from deterministic optimization algorithms are replicable. Even if randomness is involved in stochastic optimization algorithms, the results obtained from stochastic optimization algorithms can also be replicated with fixed random seeds. The main limitation with deterministic optimization algorithms is that, globally optimal solution is not guaranteed [33].

4.3 Optimization terminologies

Optimization algorithms and problem types can be categorized based on their characteristics. Cavazzuti [33] listed the most common optimization algorithms with their brief descriptions:

Deterministic optimization problem does not involve random elements and the optimization problem is solved purely based on a mathematical formulation.
Gradient-based optimization methods use the gradient and sometimes the Hessian matrix of the objective function to solve the optimization problem.
Stochastic optimization problem differs from deterministic optimization problem and such algorithms use randomness in the search procedure.
Evolutionary optimization is a stochastic population based optimization algorithm family which implements the concept of evolution. In each generation, the fittest individuals are the ones which give rise to offspring in order to improve the performance of each successive generation.
Single objective optimization (SOO) problems have a single objective function.
Multiobjective optimization (MOO) refers to optimization problems with more than one objective function
Local optimum refers to the minimum or maximum value of the objective function in the domain subset.
Global optimum refers to the minimum or maximum value of the objective function in the whole domain.
Local optimization algorithm refers to an optimization algorithm that is capable of finding the local optimum.

Global optimization algorithms are algorithms which are capable of attaining global optimum by overcoming local minima/maxima.
Discrete optimization refers to optimization algorithm that deals with distinct separate variable values.
Decision space is the space where the objective functions are defined.
Objective space is the space formed by the values of the objective functions.
Decision maker is a person who is responsible to decide which objective values he/she prefers over the others in multiobjective optimization problems.
Preference information in multiobjective optimization problem is information regarding which objective function results will be favored over the others.

4.4 Multiobjective optimization

Multiobjective optimization problem has the general form of

$$\begin{aligned} &\min \{ f_1(x), f_2(x), \ldots, f_k(x) \} \\ &\text{subject to } x \in X \subset \Re^n \end{aligned} \qquad (5)$$

and $k \geq 2$ conflicting objective functions with each objective $f_i: X \rightarrow \Re$ [34]

Because problem (5) consists of at least two conflicting objectives, there will not be a single solution which optimizes every objective. For this reason, the notion of Pareto optimality is introduced. A solution $\hat{x}$ is said to be a Pareto optimal solution if there is no $x \in X$ such that

$$f_i(x) \leq f_i(\hat{x}) \quad for\ all\ i = 1,2,\ldots,k \quad and \quad f_j(x) < f_j(\hat{x}) \quad j = 1,2,\ldots,k$$

for at least one objective j.

The Pareto optimal set refers to all Pareto optimal solutions in the decision space. On the other hand, Pareto optimal front refers to all Pareto optimal solutions in the objective space. There are two vectors in the objective space formed from the values of the objective functions in the Pareto optimal set. These two vectors are ideal and nadir. Ideal vector contains the optimal solution values of each objective function and serves as a lower bound for the Pareto optimal front. The nadir vector contains objective function values and serves as the upper bound of the Pareto optimal front. It is not easy to calculate the nadir values and

therefore, nadir vector values can be estimated using payoff table. According to [35], payoff table is constructed using optimal solutions of individual objective functions and objective function values are evaluated for each optimal solution and the nadir points are estimated by taking the worst values of the objective functions. Payoff table does not guarantee true nadir points and it may underestimate or overestimate nadir points [35]. The ideal and nadir vectors are expressed mathematically as follow.

For multiobjective optimization problem of the form (5) the ideal vector is defined as:

$$z_i^{ideal} = \min_{x \in X} f_i(x) \ for\, i = 1,2,...,k$$

Similarly, for multiobjective optimization problem of the form (5) the nadir vector is defined as:

$$z_i^{nadir} = \max_{x \in \rho} f_i(x) \quad for\, i = 1,2,...,n$$, where ρ is Pareto optimal set.

Weak Pareto optimality is a type of optimality in which there is no other vector in the decision space for which all the objectives are better [36]. A decision vector $x^* \in X$ is weakly Pareto optimal if there does not exist another decision vector $x \in X$ such that $f_i(x) < f_i(x^*)$ for all i = 1, 2, …, k.

Based on the role of a decision maker, MOO methods are classified as no preference methods, a posteriori methods, a priori methods and interactive methods [36].

A no preference method is used when a decision maker or preference information is not available. An a posteriori method is used when the decision maker provides his/her preferences after Pareto optimal solutions are generated for the MOO problem. An a priori method is used when the decision maker provides preference information before Pareto optimal solutions are generated for the MOO problem. An interactive method involves generating initial Pareto optimal solutions followed by the decision maker providing pref-

erence information and getting the resulting Pareto optimal solutions repeatedly, until the decision maker got the desired optimal solution.

Multiobjective optimization methods

There are different approaches to solve a MOO problem but many of them involve converting the MOO problem into one or several SOO problem(s). The approach of converting MOO into one or several SOO problem(s) is referred to as scalarization [36].

These two scalarization methods are discussed in this section because they are the most widely used methods for multiobjective optimization problems and usually they are referred as basic methods [37].

The ε-constraint method and the weighing methods are two methods that utilize scalarization approach to solve MOO problems. The weighing method converts the MOO problem of the form (5) into a SOO problem of the form

$$\begin{aligned} &\min \ \sum_{i=1}^{k} w_i f_i(x) \\ &\text{subject to } x \in X \subset \Re^n \\ &\text{where } w_i \in \Re, \sum_{i=1}^{k} w_i = 1 \text{ and } w_i \geq 0 \text{ for all } i = 1,2,...,k \end{aligned} \tag{6}$$

On the other hand, the ε-constraint method makes one of the objectives of the MOO problem as the primary objective function and the other objective functions as constraints. For a MOO problem of the form (5), the corresponding ε-constraint problem is:

$$\begin{aligned} &\min\{f_j(x)\} \\ &s.t. \quad x \in X \subset \Re^n \\ &and \quad f_i(x) \leq \varepsilon_i \quad \text{for all } i \neq j \\ &\text{and for some } j, \text{ where } i, j \in \{1,2,...,k\} \end{aligned} \tag{7}$$

If a solution x^* is an optimal solution to problem (7) for all $j = 1, \ldots, n$ with $\varepsilon_i = f_i(x^*)$ for all $i = 1, \ldots, k$, then the solution x^* is a Pareto optimal solution to problem (5) [36].

Determining suitable values for ε and generating Pareto optimal solutions for the selected ε values is challenging and furthermore, the method may not have a feasible solution for some ε values [37].

5 Analytics

5.1 Introduction

This chapter covers the basic introduction of analytic methods. Analytics is the process of transforming raw data into knowledge for the purpose of delivering decision support [6]. Analytics is grouped into three categories: descriptive, predictive and prescriptive analytics. Descriptive analytics explores the current and historic data. Predictive analytics predicts future trends based on historic data. Prescriptive analytics defines how to take advantage of the results of predictive analytics by taking the necessary course of actions in a timed manner.

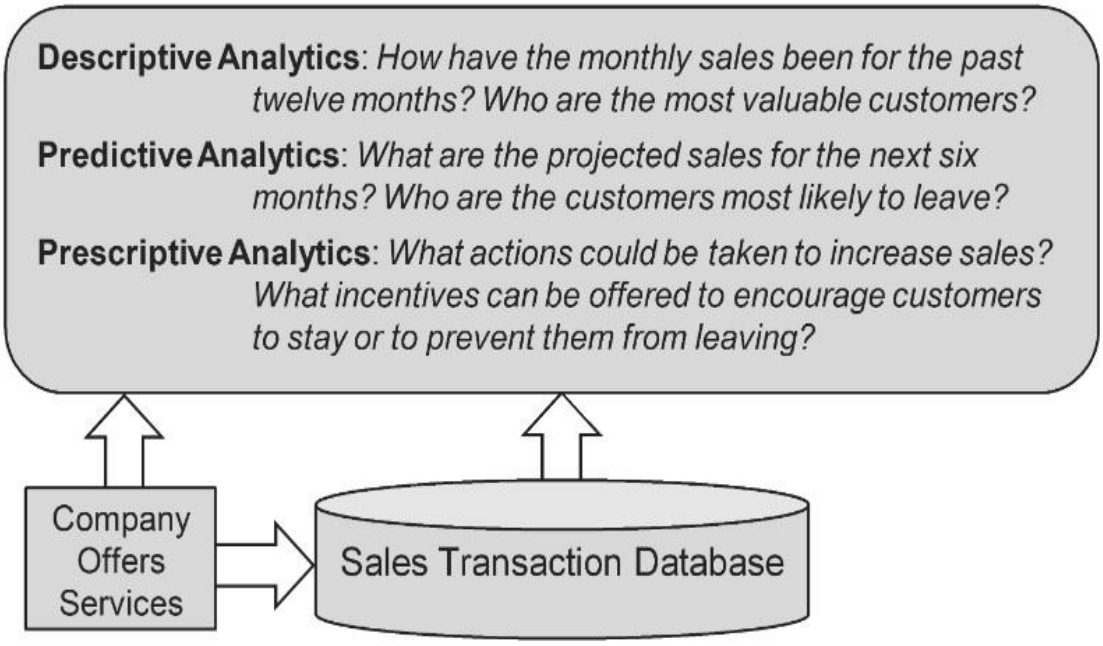

Figure 4. Analytics example [6]

Figure 4 presents how the three types of analytics could be applied in the context of a company which sells products to its customers.

5.2 Descriptive Analytics

As it was mentioned in the introduction part of this chapter, descriptive analytics looks into the current and historic data. Descriptive analytics provides maximum insight of the data by revealing the underlying structure. Current and historic data are presented by descriptive statistics, which describe the summary of basic features of the data. Central tendency which shows the center of the distribution can be used to describe the basic summary of the data. The common measures of central tendency are the mean, median and mode. Mean is the average value of the data, median is the middle term of the data after the data is sorted and mode is the data which appears most frequently. When each item appears an equal number of times in a particular set of data, then the distribution is called unimodal. Similarly, when multiple items have the same highest frequency, the distribution is called multimodal. The variation in data values or dispersion is measured through range and variance. Range is the difference between the maximum and minimum values in the data. Variance is the difference squared sum of the mean and each item in the data, divided by the number of items in the data. Standard deviation is the square root of variance.

Predictive analytics

Predictive analytics forecasts future trends based on the current and historic data. Classification models and regression models are the two types of predictive analytics models which were covered in the ML and DM chapters.

5.3 Prescriptive analysis

Bayesian belief network

In decision-making methods Bayesian Belief Network (BBN) is a probabilistic graphical model consists of nodes, which are random variables representing features, and links representing the relationship between the nodes. Conditional probability, the probability of an event knowing that another event has occurred, is used in constructing BBN. Each node-link in BBN represents conditional dependency based on the conditional probability table

which determines the relation associated with the link. Links are arrowed to show causality. The node from which the link originated represents the cause and the node pointed by the link arrow represents the effect [6]. The state of a node is called belief which is the probability of occurring states.

Decision-making methods

Das [6] described two types of decision support methods: influence diagram (ID) and symbolic argumentation. Decision making based on an influence diagram involves choosing an action or hypothesis from a set of alternative actions/hypotheses, along with their conditional probability values, that provides the best desired result. Decisions based on ID are represented as diagrams with the chance node, decision node and state/value. Chance nodes in ID represent possible outcomes of an action with their probability of occurrences. Decision nodes provide alternative actions/hypothesis. State nodes represent the final result after subsequent decisions. BBN needs to be converted to ID to be used for decision support [6]. The second decision support method, symbolic argumentation, is represented with logical if-then statements.

6 Analysis of DUI data

6.1 Introduction

The background theories related to the thesis topic are covered in the previous chapters; this chapter will present the data analysis of the DUI data based on the covered background theories. This chapter describes how the data was preprocessed to form DUI groups/clusters and how the optimization problems were formulated based on the DUI-data. It also presents how the optimization methods were used to produce results.

The DUI dataset was explored to have an insight of the data and to experiment different data analysis procedures for modeling the association between prison sentences and committing a new DUI. Using descriptive statistics, which serves to describe the content and features of the dataset, important data attributes of the dataset are presented for the purpose of exploring the data. For data description, the data attributes can be grouped into three categories: attributes that describe offenders, attributes that show committed offenses and attributes that show actions taken against the committed offenses.

Tools

The tool we used for data analysis is Python Jupyter notebook. According to the official website [38], Python's Jupyter notebook is an open source web application suitable for data manipulation and data visualization. We have utilized different Python modules to carry out both data manipulation, optimization and data visualization tasks. For data manipulation Pandas, Numpy and Sklearn modules were used along with Matplotlib and Scipy modules for data visualization and optimization, respectively.

6.2 Analysis of the dataset

In the dataset, all individuals convicted to prison from 2004 to 2007 year are included. The dataset contains 1533 individuals convicted to prison in 2004, 1201 from 2005, 972 from 2006 and 801 from 2007. Among the individuals 19 individuals were dropped as their re-

lease date is too late to follow up. In the data set, five years offenses or punishments are calculated as, committed offenses or handed punishments for an individual that occurred during the five years period prior to prison conviction from 2004 to 2007. The description of the dataset with the attribute names, the representation of the data-attributes and their translation is attached in the Appendix A.

During the extraction of the datasets from the main repository, offenders younger than 21 years were already filtered out. Therefore, in the dataset we have ages ranging from 21 to 77 with mean age value 37.9 and standard deviation value 11.02. For the gender attribute, out of 4507 offenders 4339 or 96% are male offenders, and 168 or 4% are female offenders. Therefore, the gender attribute shows that DUI offenses are mostly committed by male offenders.

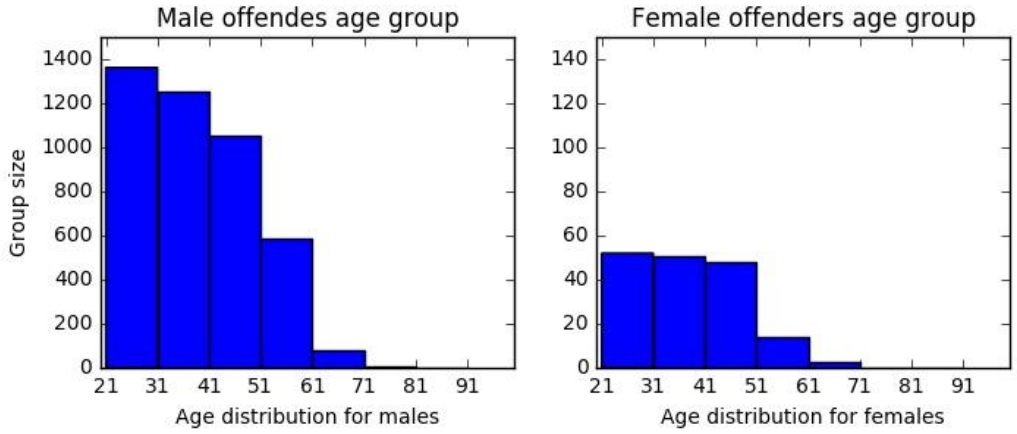

Figure 5. Histogram of age groups for each gender

From Figure 5, by observing the two histograms, in both gender groups the number of individuals who committed DUI offences punishable by imprisonment decreases as they get older. After the age of 51, there is a significance drop in the number of individuals in both genders.

Offenses committed by individuals

Measurement	**Property crimes in 5y**	**DUI offenses in 5y**	**Traffic Crimes 5y**	**Next DUI**	**Violent Crimes 5y**

				Days	
Min	0	0	0	0	0
Max	101	22	78	1096*	18
Mean	4.37	2.38	5.9	-	1.21
Zero values	50.6%	13.6%	9.5%	0.1%	56.8%

Table 1 Description of offenses

*for Next DUI Days, the 1096 code indicates that the offender has not committed any DUI again. This makes it difficult to calculate the mean as the choice of the code influences the mean. Removing individuals with the next DUI code 1096 and calculating the mean value will produce a biased result by not considering all the individuals in the dataset.

Table 1 columns represent data attributes from the DUI dataset. Property crimes in 5 years, DUI offences in 5 years and Traffic crimes in 5 years represent the number of property crimes, the number of DUI offenses and the number of traffic crimes committed by an individual in five years time respectively. Next DUI days for an individual represents the number of days between the last two DUIs. Finally, violent crimes in 5 years represent the number of violent crimes committed by an individual in 5 years time.

Table 1 rows show the minimum, maximum, average and zero values for the column attributes. Zero values indicates the number of records which have zero as their value for the indicated data attribute. Maximum value indicates the maximum number of crimes committed by an individual among all the offenders. Similarly, the minimum value indicates the minimum number of crimes committed by an individual compared to all the other individuals.

From Table 1, one can observe that traffic crimes in 5 years is the most committed crime as only 9.5% of individuals didn't commit Traffic crimes in 5 years. On the other hand, property crimes in 5 years and violent crimes in 5 years have been committed by around 49.4%

and 43.2% of individuals, respectively. But the average number for property crimes in 5 years is greater than that of violent crimes in 5 years.

Punishments given to individuals

For the punishment description, the attributes which could give us good insights are number of prison sentences in five years, number of probation sentences in five years and number of community service sentences in five years.

Measurement	**Community sentences 5y**	**Probation Sentences 5y**	**Prison Sentences 5y**
Min	0	0	0
Max	7	8	25
Mean	0.822	0.62	2.02
Zero values	51.05%	59.88%	45.15%

Table 2. Punishment description

In Table 2 the columns Community sentences 5y, Probation sentences 5y and Prison sentences 5y represent the number of community sentences, probations and prison sentences handed to an individual in 5 years time respectively.

The rows min, max, mean and zero values have similar interpretation as that of Table 1. The only difference is that in Table 2 the values represent punishments instead of crimes.

From the Table 2, by comparing the zero values of the data attributes, prison sentence is the most exercised punishment type while probation sentence is the least practiced one as around 40% of the total individuals were given probation sentences.

6.3 Data cleaning

In the descriptive analytics part of the dataset, the description of the individuals in the dataset, the offenses they have committed and the punishment they have received were pre-

sented. In this phase, attributes with less importance in the analysis phase, for instance data attributes related to court appearances and appeals, were filtered out. In the dataset, the 'next DUI' attribute which indicates the number of days between the last two DUIs contains a code '1096' which indicates the individual has not committed any new DUIs during the follow up time. Originally, records with the next-DUI value of 1096 were filtered out, as these individuals will not affect in reducing DUIs, but later it was found out that these groups affect the cost of the proposed sentencing as they stopped committing DUIs after serving their sentences. As a result, these individuals were included in the data analysis process.

6.4 Data transformation

From what we observed so far, not all individuals behave the same way, as there are differences based on age groups and gender. A good approach to handle repeated DUI offenders is to group them based on the number of DUI offenses they have committed in the past five years. It is the best attribute in our dataset that describes how frequently one is engaged in committing DUI offences. For the process of forming groups two approaches were adopted, the first one is to form predefined DUI groups based on the number of five years DUIs and the second one is using a classifier based on the data attributes age and five years DUIs.

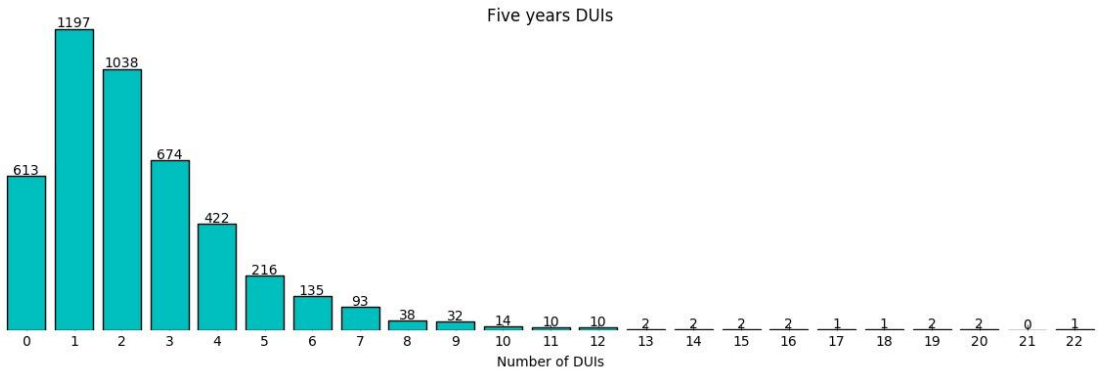

Figure 6. Five years DUIs

From Figure 6, one can see that except for individuals with zero number of DUIs in the past five years, the number of individuals drop as the number of DUIs increases from 1 to 11. For the number of DUIs greater than 11 there is no clear pattern and these groups contain fewer numbers of individuals.

6.5 Group formation based on number of DUIs

To have groups with a reasonable number of individuals, we grouped the individuals with five or more DUIs together to form 6 DUI groups. Therefore, for the predefined groups, we have six groups with individuals who committed 0, 1, 2, 3, 4 and above 4 DUIs in the past five years, respectively. As it was explained in Section 6.2, all the included individuals have committed DUI offences punishable by prison sentences. Therefore, Group1, individuals with zero DUIs in five years time, was included because these individuals have next-DUI attribute value for committing new DUI and the next DUI attribute contains the number of days between the last two DUIs. The number of individuals in group 1, with zero number of DUIs in five years time, who committed new DUI offences, can be checked from Table 3 below.

	Group-1	Group-2	Group-3	Group-4	Group-5	Group-6
Total size	613	1197	1038	674	422	563
Quit DUI	60.52%	57.98%	52.22%	42.14%	41.23%	29.66%
Repeated DUI	242	503	496	390	248	396
Average Age	37.05	38.93	38.59	38.83	36.58	35.32
Prison duration $\leq$ 14 days	18	24	17	18	6	16

Table 3. DUI group size and quitting ratio for predefined groups

The columns in Table 3 represent the predefined groups which contains individuals who committed 0, 1, 2, 3, 4 and above 4 number of DUIs in five years time. For the rows in Table 3, quit DUI indicates the percentage of individuals in each group who stopped committing DUI offences. Repeated DUI indicates the number of individuals who committed new DUI. Average age row indicates the average age of each group and finally, Prison duration less than or equal to 14 days indicates the number of individuals in each group who served less than 15 days prison sentence.

From Table 3, almost 60% of the individuals in Group 1 and Group 2 have not committed any DUIs again and as we go along the groups, the percentage of individuals who stopped committing DUIs decreases. For the age attribute, Group 6, individuals with the highest number of DUIs, is the youngest group.

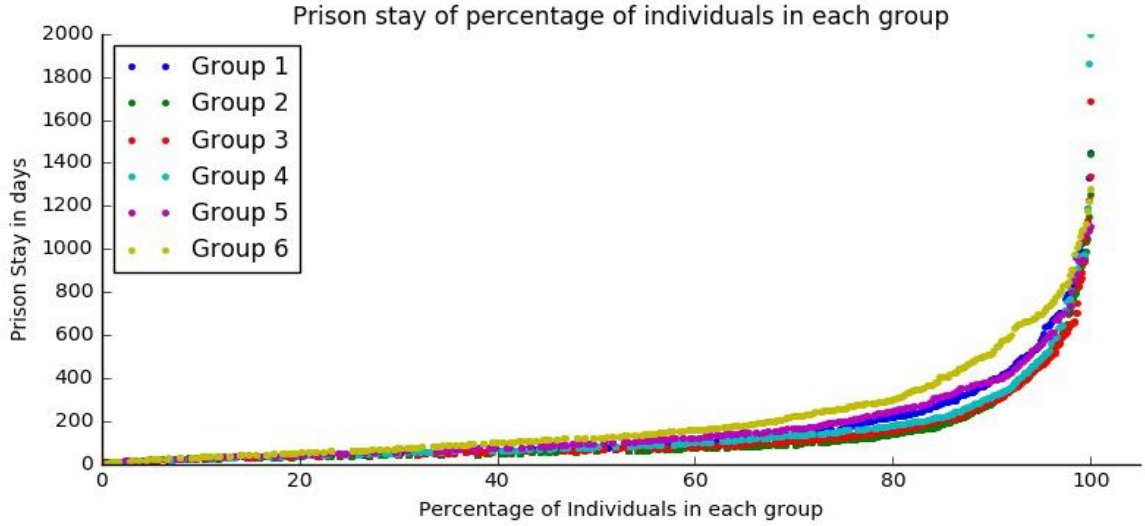

Figure 7. Prison stay of individuals in each group

The horizontal axis in Figure 7 represents percentage of individuals from each group and the vertical axis represents the number of days held in prison. The figure presents an insight on how prison stays are distributed among DUI groups.

Figure 7 shows individuals in Group 6 served longer prison sentences than the rest of the DUI groups, as around 20% of the individuals in group 6 served more than 300 days.

6.6 DUIs groups based on a classifier

The second approach to form DUI groups is using KMeans classifier, based on age and five years DUI attributes. The KMeans classifier was explained in Chapter 3. To have the number of clusters equal to the number of predefined groups, cluster size was set to 6. KMeans classifier starts with random centroids to form clusters, thus it makes it difficult to create fixed set of clusters on subsequent cluster formations. To deal with this issue we will set a fixed random seed to replicate the generated clusters for future use. As it was mentioned in Chapter 3, the KMeans algorithm utilizes the Euclidian distance to determine how two data items are close to each other. Therefore, to reduce the effect of scale differences we will normalize the two attributes using the following formula:

Normalized(x) = (x-*min*(x)) / (*max*(x)-*min*(x)).

where x is the attribute value to be normalized, min(x) is the minimum value of the attribute and max(x) is the maximum value of the attribute.

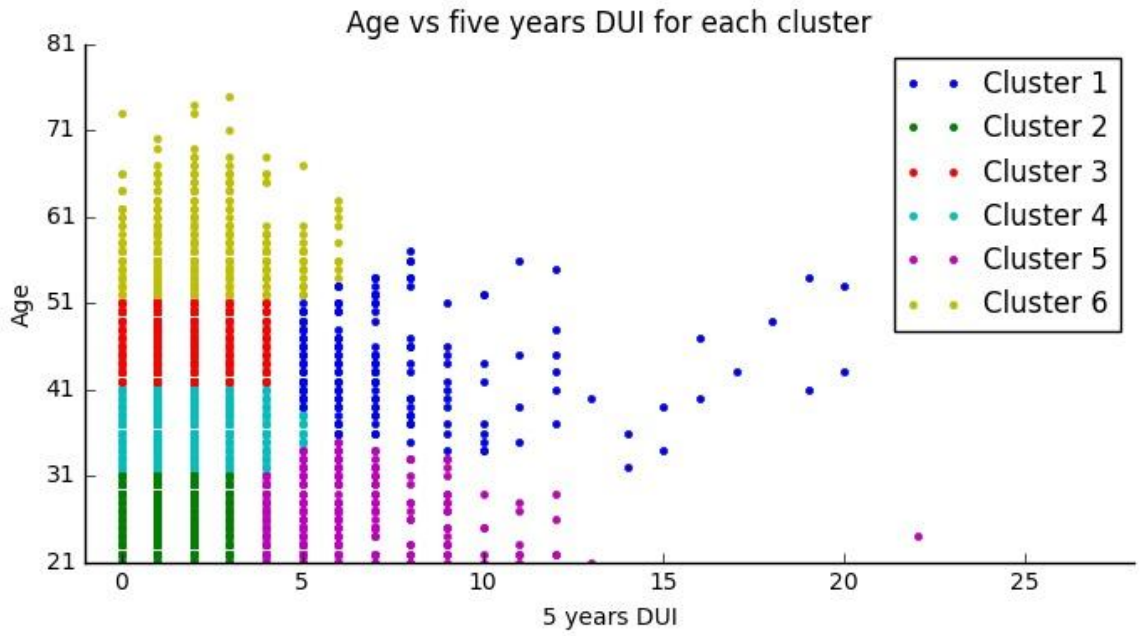

Figure 8. Age and 5 years DUIs for clusters

Figure 8 provides a better visualization on how DUI clusters are formed using age and 5-years-DUI data attributes, as individuals who share similar age and 5-years-DUI values are clustered together.

	Cluster-1	**Cluster-2**	**Cluster -3**	**Cluster -4**	**Cluster -5**	**Cluster-6**
Total size	220	1124	950	1190	450	573
Quit DUI	31.36%	46.89%	57.68%	50.17%	28.0%	63.7%
Average 5yrs DUI	7.27	1.49	1.77	1.79	5.59	1.99
Average Age	43.97	26.18	46.41	36.53	26.09	56.63

Table 4. Cluster description

Table 4 columns indicate the 6 clusters formed by the KMeans classifier. The rows indicate the total number of individuals, the percentage of individuals who stopped committing new DUI offences after their last DUI conviction, the average number of five years DUIs and the average age of individuals in each cluster. Centeroid values of age and number of DUIs values for each cluster are the following: (43.97, 7.28), (26.18, 1.49), (46.41, 1.77), (36.53, 1.79), (26.09, 5.59), (56.63, 1.99).

Table 4 shows that Cluster-6 has the highest percentage of individuals who stopped committing new DUIs and it also has the oldest individuals. Contrary to that, Cluster-5 contains the lowest percentage of individuals who stopped committing DUIs. Cluster-2 and Cluster-5 contain the youngest individuals. Cluster-1 contains the highest average number of five years DUIs.

Association between next DUI and prison sentence

Two groups have already been formed using predefined settings and a classifier. The correlation between the attributes prison sentence and next DUI will be assessed to determine whether a regression model could be used to capture the underline pattern.

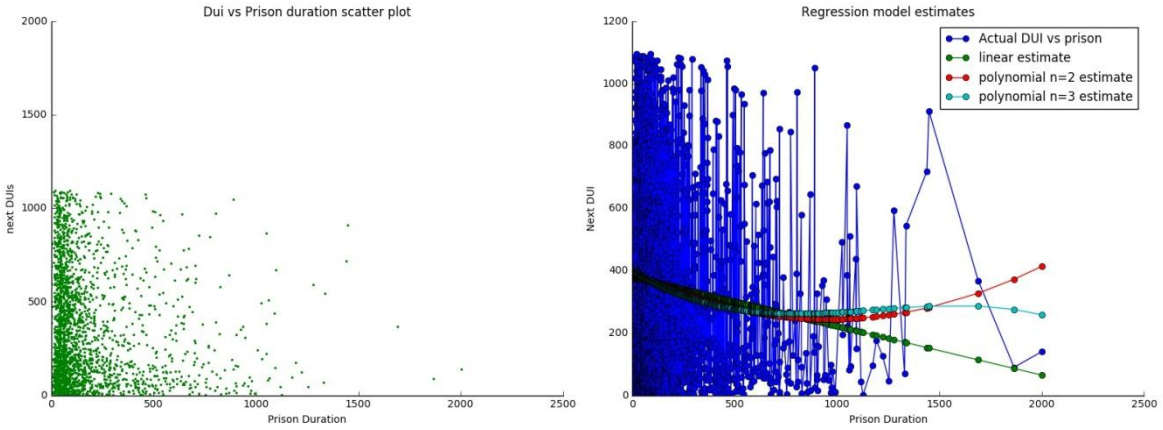

Figure 9. Prison duration vs. next DUI for imprisoned individuals

From the two subplots of Figure 9, one can observe that there is no clear association between the two attributes prison duration and next DUI values. Because there is no clear association between the two attributes, the regression model used was unable to capture the pattern.

The correlation coefficient which determines how well attributes are related to each other is used to test the correlation of prison duration and next DUI attributes. Correlation coefficient value of 1 indicates there is a strong direct relationship, -1 indicates there is a strong inverse relationship and 0 indicates there is no relationship between the attributes. The correlation coefficient for the attributes prison duration and next DUI equals to -0.1759 which is close to zero.

Therefore, both Figure 9 and the correlation coefficient show that the two attributes are barely correlated to each other and it is not possible to model the relationship using a regression model. Thus, the option of using regression models to predict next DUI values for proposed sentencing is not applicable to the data. It can also be stated that the length of the prison sentence has no direct influence on the action of committing new DUIs.

6.7 Alternative problem solving approach

The alternative approach is to see how many individuals from each DUI group/cluster recommit DUI offenses after they served their prison sentences, at certain checkpoints.

Checkpoints are date marks or follow up periods to check whether a new proposed sentencing has reduced or raised the number of DUIs at that follow up period. The main idea is that many individuals recommit DUI offenses once they are released from prison and if they are being kept in prison for a longer time, there is a possibility to reduce the number of DUIs that would be committed by the imprisoned individuals for the extended duration. Similarly, if the individuals are released from prison earlier than when they are supposed to be released, there is a higher possibility that the number of DUIs increase due to the reduced prison stay.

There are also some facts that need to be considered concerning when the next DUI will be committed. There is a possibility that an offender may not be caught on the first day he/she committed a DUI offence after serving a prison sentence unless they are stopped by an officer for a traffic violation. Offenders could also commit multiple DUIs while their case is in the court process and in this case they might get a combined sentence.

Giving an additional prison sentence to offenders will raise the cost in keeping the offenders for an extended duration in prison and to get the desired outcome the tradeoff between cost and the resulted number of DUIs from the proposed prison sentence needs to be examined.

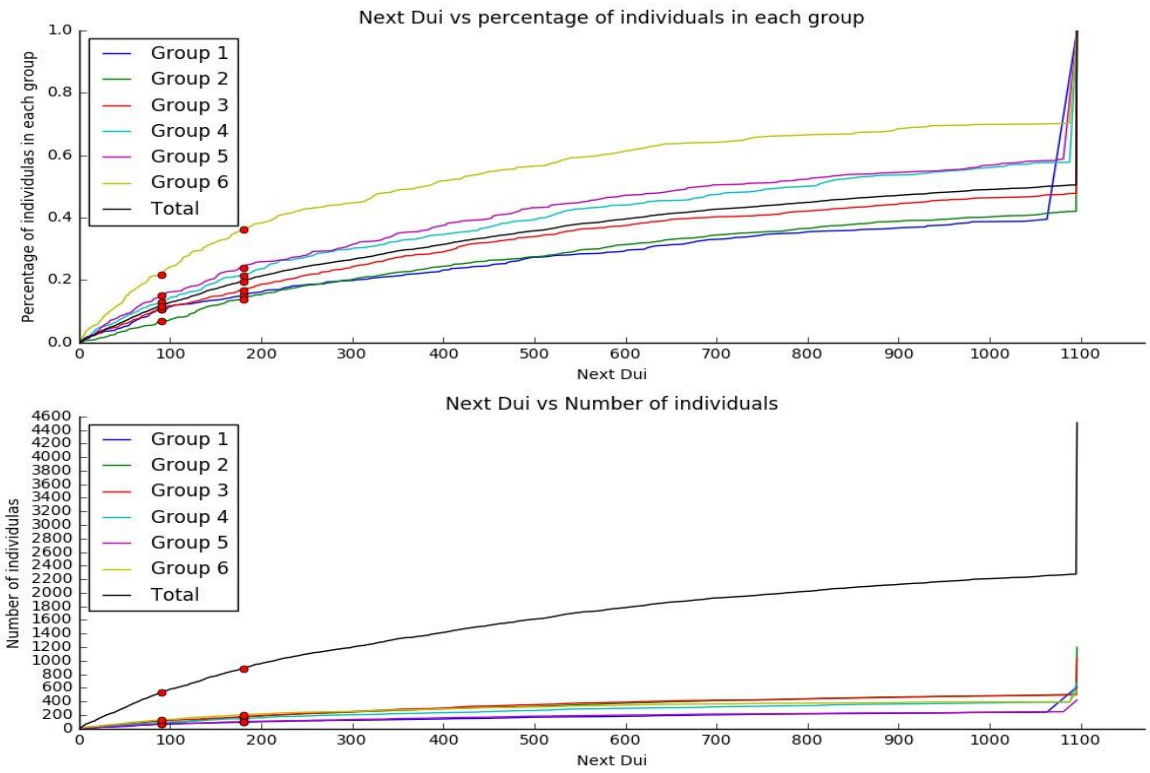

Figure 10. Cumulative DUI graph

Figure 10 displays two subplots, the first subplot shows the cumulative next DUI ratio and the second subplot shows the cumulative number of individuals for next DUI attribute. From Figure 10, one can observe that more than twenty percent of the total offenders in each group committed their next DUI in less than 400 days. The two chosen checkpoints for the follow up period are 3 months and 6 months time intervals and they marked as red for referencing. By selecting a next DUI value one can locate that the percentage/number of individuals from a group who recommitted DUI offence.

Figure 10 was constructed in such a way that, first the minimum and maximum next DUI attribute values are obtained for each DUI group. Then for each group, for an arbitrary day d between the minimum and the maximum next DUI values, the cumulative DUI value is calculated as the ratio of the number of individuals whose next DUI value is less than or equal to d, to the total number of individuals in that particular group.

To demonstrate the approach of introducing proposed sentencing a bit further, let us say that we have proposed to extend the prison stay of Group 1 offenders from Table 3 by 90

days i.e. introducing a 90 days proposed sentencing. This will result in zero DUIs in the first three months for Group 1, which means that we push the cumulative DUIs graph in Figure 10 to the right direction by 90 units. As it was decided to maintain a minimum prison sentence of 14 days, individuals who served less than 14 days will not get a reduced sentence when a reduction of a prison sentence is considered. Similarly, reducing the sentences for Group-1 by 90 days, keeping in mind the mandatory minimum 14 days sentencing, means the same as pulling the cumulative next DUI graph to the left by 90 units and that increases the number of DUIs committed in the first 90 days.

The difference in the reduction or increase of DUIs at the follow up periods depends on the slope of the cumulative DUI graph. If the slope of the graph is zero at a checkpoint, reducing or increasing sentencing will not result in a significant change in the number of DUIs at the checkpoint. Similarly, if the cumulative DUI graph is steep at a checkpoint, the introduction of proposed sentencing will result in a significant change by reducing or increasing DUIs at the checkpoint.

As it has already been mentioned in the introduction section, introducing proposed prison sentencing also affects the cost associated with keeping DUI offenders in prison. The actual cost associated with keeping an offender in prison for a day is not available. In order to compute the tradeoff between cost and the number or DUIs, cost is defined as the number of prisoners multiplied by the sentence duration in days.

To assess the effect of proposed sentencing two checkpoints will be used: 90 days 180 days after the individuals served their prison sentences. As it was explained in the earlier section, how the proposed sentencing works when the original sentencing is extended or reduced by 90 days, a table is presented below to show how each DUI group/cluster will react to the newly proposed sentencing. When a proposed sentencing extends the original sentencing by 90 days, the number of DUIs that can be reduced from each group lie in the first 90 days. Similarly, when a proposed sentencing reduces the original sentencing by 90 days the number of DUIs increases based on the number of individuals in that group whose next DUI attribute value lies between the interval 90 days and 180 days.

	size	Individuals with next DUI < 90 days		next DUI between 90 and 180 days		next DUI between 180 and 270 days	
		Tot	Percentage	Tot	Percentage	Tot	Percentage
Group 1	613	65	10.6%	28	4.57%	23	3.75%
Group 2	1197	81	6.77%	86	7.18%	57	4.76%
Group 3	1038	111	10.69%	62	5.97%	62	5.97%
Group 4	674	86	12.76%	58	8.61%	49	7.27%
Group 5	422	63	14.93%	38	9.0%	22	5.21%
Group 6	563	123	21.85%	80	14.21%	42	7.46%

Table 5. Next-DUI distributions for groups

The columns in Table 5 represent the size of each predefined group, the number of individuals whose next DUI data attribute values lie between 0 to 90 days, between 90 days to 180 days and between 180 days and 270 days. The rows represent the predefined DUI groups.

Table 5 shows next-DUI values at three months interval. From the table, one can observe that Group 6 is a group which has the highest percentage of individuals who committed DUIs in all the three intervals.

Similar to what has been done to the predefined groups, a table for next-DUI values for the intervals of 3 months, 6 months and 9 months is also constructed to the DUI clusters.

	Size	Individuals with next DUI < 90 days		next DUI between 90 and 180 days		next DUI between 180 and 270 days	
		Tot	percentage	Tot	Percentage	Tot	Percentage
Cluster 1	220	43	19.55%	34	15.45%	17	7.73%
Cluster 2	1124	157	13.97%	87	7.74%	73	6.49%

Cluster 3	950	76	8.0%	47	4.95%	43	4.53%
Cluster 4	1190	122	10.25%	93	7.82%	63	5.29%
Cluster 5	450	106	23.56%	60	13.33%	28	6.22%
Cluster 6	573	25	4.36%	31	5.41%	31	5.41%

Table 6. Next-DUI distribution for clusters

In Table 6, the rows represent the clusters and the columns in Table 6 represent the size of the clusters, the number of individuals whose next DUI data attribute values lie between 0 to 90 days, between 90 days to 180 days and between 180 days and 270 days.

From Table 6, Cluster-2 has the highest number of individuals and Cluster-5 has the highest percentage of individuals who committed DUI in the first three months. For the intervals from 3 months to six months and from six months to nine months, the number of individuals who committed DUIs dropped for all but cluster 6.

6.8 Definition of objective functions

To solve the multiobjective optimization problem of reducing the number of DUIs and cost values, we will use three objectives. The objectives are the two DUI checkpoints, DUIs after 90 days and DUIs after 180 days, and cost. Because we have 6 DUI groups and clusters, we will have six different proposed sentences for each group/cluster i.e. each group/cluster will get a different proposed sentencing.

The general format of the objective functions is shown as follow:

$$F: \quad f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + f_5(x_5) + f_6(x_6)$$
$$G: \quad g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) + g_5(x_5) + g_6(x_6)$$
$$C: \quad c_1(x_1) + c_2(x_2) + c_3(x_3) + c_4(x_4) + c_5(x_5) + c_6(x_6)$$

where:

$f_1, f_2, ..., f_6$ represents the number of DUIs after 90 days for groups/ clusters 1,2,...,6

$g_1, g_2, ..., g_6$ represents the number of DUIs after 180 days for groups/ clusters 1,2,...,6

$c_1, c_2, ..., c_6$ represents the cost function for groups/ clusters 1,2,...,6

$x_1, x_2, ..., x_6$ represents proposed sentencing variable for groups/ clusters 1,2,...,6

depending on the type of sentencing, which is discussed in Section 6.9, the proposed sentencing variables could represent different values and the values of the functions F, G and C are the result of the sum of individual results for each DUI cluster/group.

6.9 Types of sentencing

Two sentencing approaches will be explored: the first one is based on a fixed proposed sentencing of 90 days i.e. a maximum of 90 days will be added or reduced to the original sentencing. The second one is based on a percentage of previously served sentences. There is a two-week mandatory prison sentence and to keep this minimum prison sentence in place, when prison terms are reduced, the reduction only considers prison sentences beyond the minimum 14 days. To have a realistic proposed sentencing to observe how the introduction of a proposed sentencing will reduce or raise the number of DUIs, the maximum proposed sentence that will be added or reduced is set to 90 days. According to our assumption, if we are going to give additional 90 days sentencing for any DUI group/cluster, there will be zero DUIs for the first three months resulted from that particular DUI group/cluster.

For proposed sentencing between -90 and 90 days, which means reducing or increasing the original sentencing by 90 days, the optimization problem will have the form: $\min\{F(X), G(X), C(X)\}$ where $X = \lfloor x_1, x_2, ..., x_6 \rfloor$ and $x_i \in [-90, 90]$ *for* $i = 1,2,...,6$

The proposed sentencing is not based on the original sentencing therefore individuals in a particular group/cluster will get similar proposed sentencing.

For percentage proposed sentencing based on served prison sentences, the optimization problem will be:

$$\min\{F(\lambda Y), G(\lambda Y), C(\lambda Y)\} \ .$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_6]$ for $\lambda_i \in [-1, 1]$ and $Y = S_{ij}$ representing the served prison sentence of the j^{th} individual in the i^{th} DUI group/cluster for $i = 1,2,\ldots 6$ and $j = 1,2,\ldots, \eta_i$ where η_i is the size of the i^{th} DUI group/cluster .

With percentage proposed sentencing, individuals in each group will get separate sentences based on the percentage of the prison terms they served in the past.

6.10 Cost and number of DUIs for proposed sentencing

In the following section, two graphs will be presented for both DUI groups and DUI clusters. The first graph contains two sub plots one for the number of DUIs and the other one for cost. These graphs are used to analyze how proposed sentencing affects the number of DUIs and cost for both DUI groups and DUI clusters. The second graph, DUI versus cost graph is presented to demonstrate which DUI groups or DUI clusters provide the best ratio i.e. when we intend to reduce cost, by reducing the original sentencing, to find out which cluster/group yields maximum cost to number of DUIs ratio and when we intend to reduce number of DUIs, by extending the original sentencing, which group/cluster yields the best number of DUIs to cost ratio.

To interpret DUI and cost subplots for a proposed sentencing, first we check the proposed sentencing values. If the proposed sentencing value is positive, it means the original sentence is extended. If the value is negative it indicates that the original sentence is reduced and if it is zero the original sentence is unchanged.

For the DUI versus cost graph, for both number of DUIs and cost, negative value indicates that the amount is reducing and positive value indicates that the amount is rising. The de-

sired values are negative DUIs with small cost and negative cost values with fewer number of DUIs.

Predefined groups DUI and cost subplots

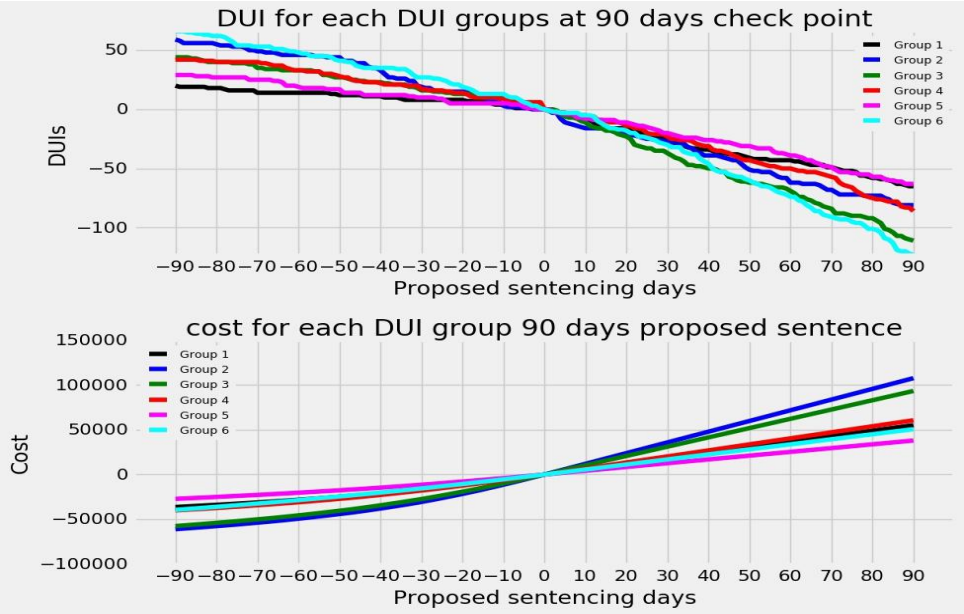

Figure 11. DUI and cost functions at 90 days checkpoint

Figure 11 shows that by extending the original sentencing by additional 90 days, it's possible to reduce over 200 DUI offences from Group 3 and 6 combined. However, the same proposed sentencing has resulted the number of DUI offenses to be dropped by just over 120 DUI offenses from DUI groups 1 and 5 combined. This shows that, compared to DUI groups 1 and 5, more individuals from DUI groups 3 and 6 committed DUI offenses within the first 90 days after they had committed their last DUI offense.

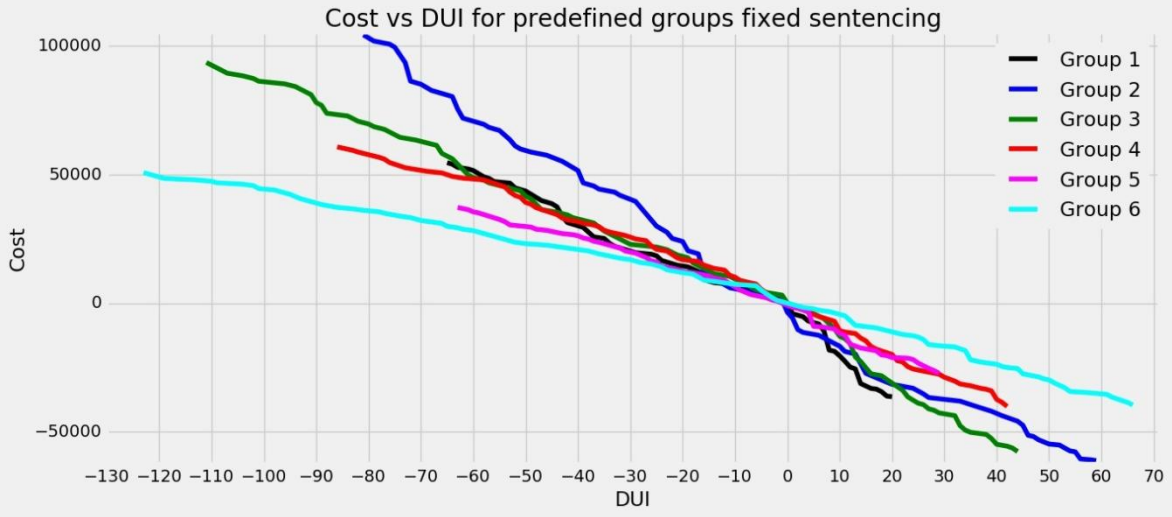

Figure 12. Cost vs. DUI for predefined groups fixed sentencing

From Figure 12, one can observe that to reduce the number of DUIs Group 6 is the cheapest to reduce the number of DUIs and conversely reducing DUIs for Group2 is expensive. On the other hand, when considering reducing cost, Group1, Group3 and Group2 provide the best ratio to save cost with lower number of DUIs compared to the other DUI groups.

Predefined group percentage sentencing

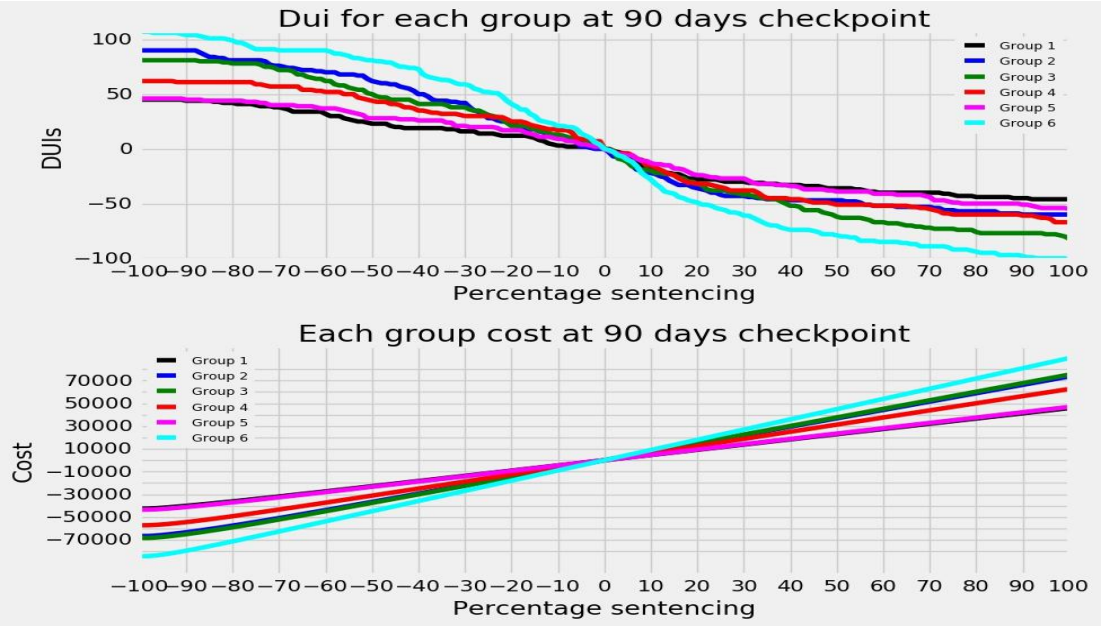

Figure 13. Cost and DUI plots based on percentage sentencing

By comparing Figure 11 and Figure 13, proposed sentencing based on a percentage of the served prison sentences differs from the fixed sentencing. Compared to the fixed sentencing, in the percentage proposed sentencing, there is a significant difference in the number of DUIs increased by extending the original prison sentences. Another thing to observe from Figure 13 is that, when reducing the original sentencing, the percentage sentencing at 90 days checkpoint favors groups that have individuals who served longer prison sentences with a next-DUI value less than 6 months as it has already been explained through pulling the cumulative DUI graph to the left. This means if a group has a lot of individuals with next-DUI values less than 6 months but they have served less than 14 days, then the percentage proposed sentencing will not have a significant effect in reducing or increasing the number of DUIs.

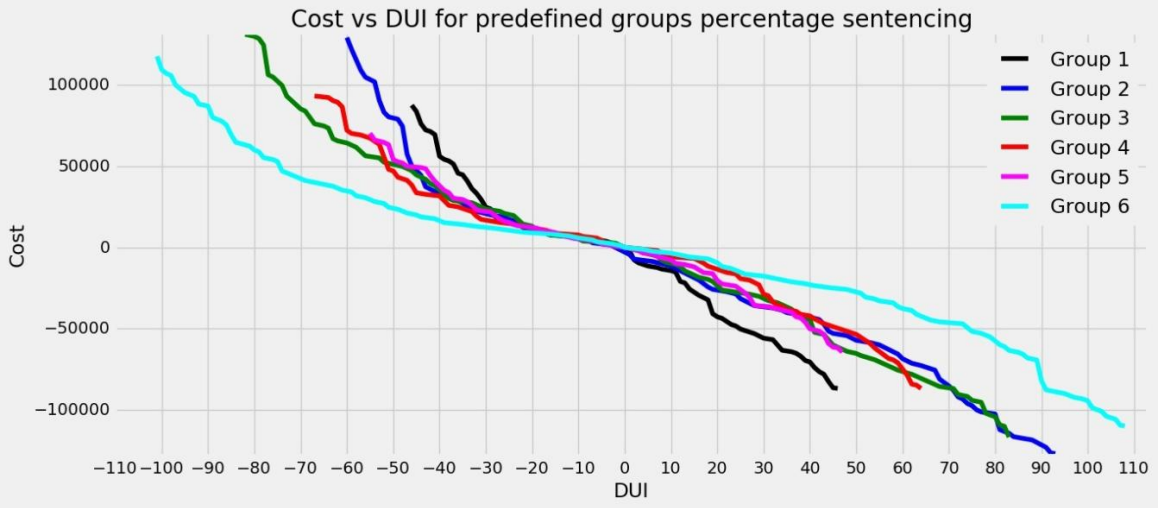

Figure 14. Cost versus DUI plot for percentage sentencing

From Figure 14, one can observe that Group 6 is the cheapest group to reduce DUIs compared to the other groups. To reduce cost, Group 1 is the favorable group among the other groups to cut cost without increasing the number of DUIs significantly.

Clusters with fixed sentencing

In the following sections we will consider clusters.

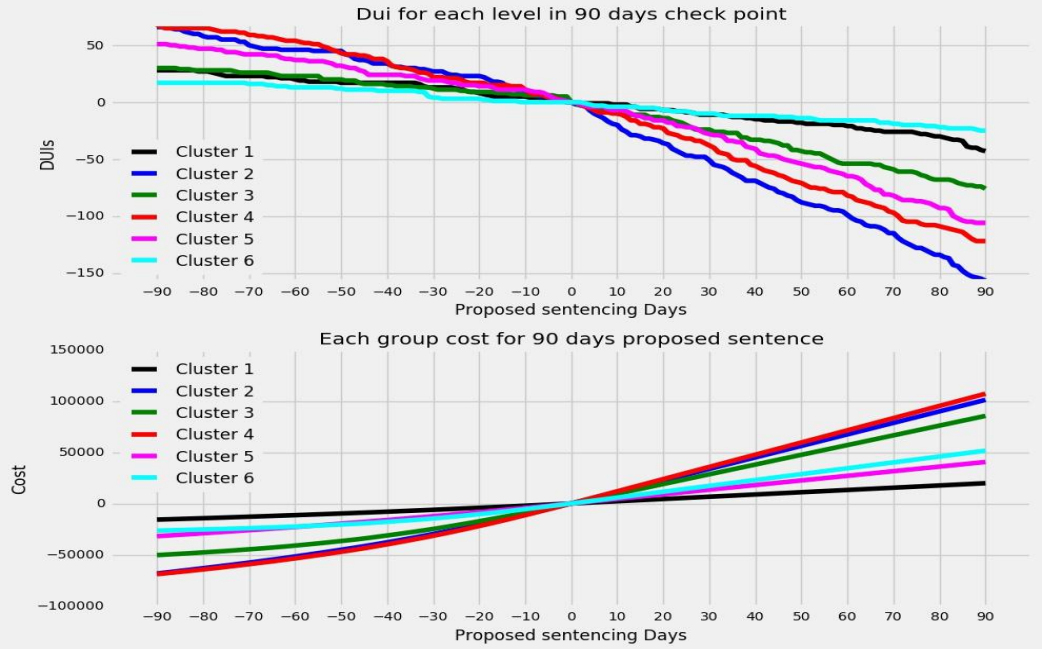

Figure 15. Fixed sentencing number of DUIs and cost plots for clusters

From Figure 15, compared to the other clusters the introduction of proposed sentencing has minimal effect on Cluster 6. The introduction of proposed sentences has a significant change in the number of DUIs for Cluster 2. From the cost subplot, Cluster 4 and Cluster 2 are the two clusters with the largest number of offenders and their associated cost dominated the cost of other clusters by being the minimum for reduced sentencing and by being the maximum for increased sentencing.

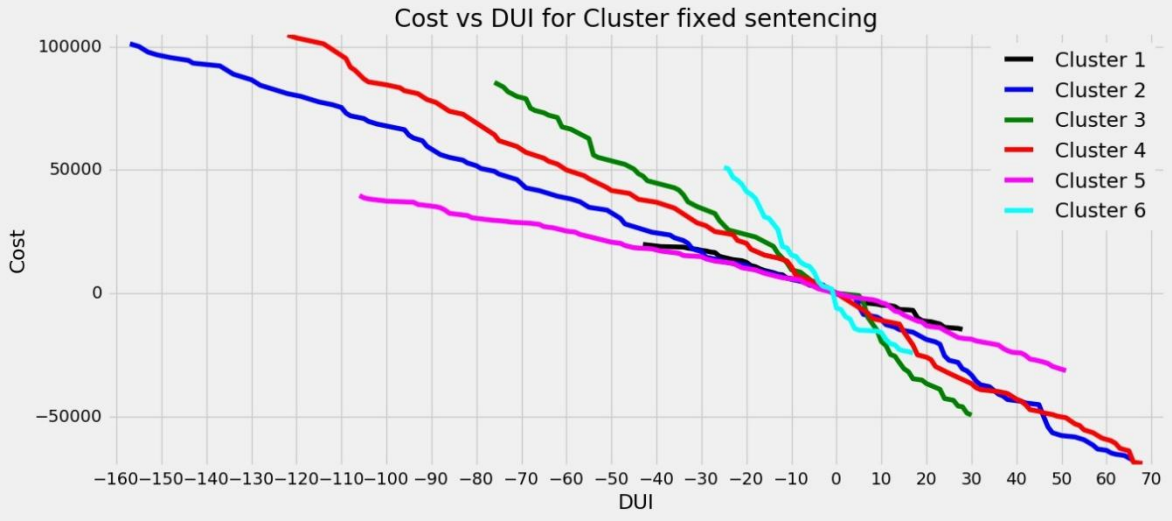

Figure 16. Cost vs. DUI for fixed sentencing DUI clusters

From Figure 16, Cluster 5 has the lowest cost to reduce DUIs and Cluster 6 has the highest cost to reduce the number of DUIs by 25. Cluster 6 provides best DUI versus cost ratio to reduce cost for less than 10 DUIs. If one has a limit to allow the number of DUIs to rise from 10 to 30 to save cost, then Cluster 3 is the best choice compared to the other clusters.

Clusters percentage sentencing

In this section the two plots for the DUI clusters will be presented. The first plot depicts number of DUIs and cost values for percentage proposed sentencing. It shows how proposed sentencing affects in reducing or increasing DUI offenses or cost values for each DUI cluster. The second one depicts the relationship between number of DUIs and cost for DUI clusters when percentage proposed sentencing is used. It shows the best DUI offenses

versus cost ratio i.e. how much cost is associated when we reduce or extend number of DUI offenses.

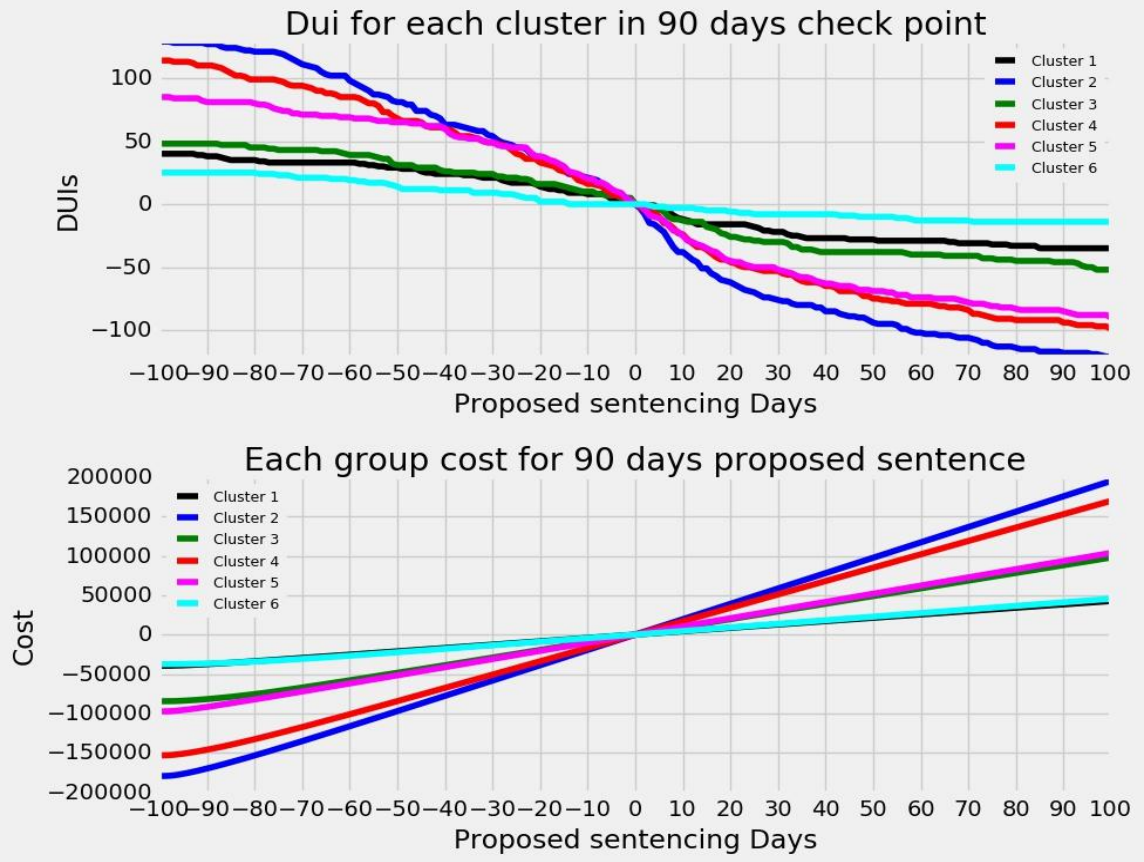

Figure 17. Percentage sentencing for clusters

Comparing the figures Figure 17 and Figure 15, in Figure 17 the number of DUIs reduced from Cluster 2, by increasing the length of prison sentences, is slightly smaller than the one in Figure 15. Similarly, when the original sentence is reduced, the number of DUIs from Cluster 2 in Figure 17 showed a significant increase compared to the same cluster in Figure 15. The reason is that, reducing DUI at 90 days checkpoint depends on the number of individuals in a DUI cluster/group whose next DUI value is between 90 days and 180 days. In addition to that for percentage proposed sentencing, the length of served prison sentences plays a major role in increasing and decreasing the number of DUIs.

From the cost subplot of Figure 17, Cluster 2, which has 1124 offenders, has the highest reduction and increase of cost compared to all the other clusters. The cost for Cluster 4,

which has 1190 offenders and has more individuals than cluster 2, did not get the highest cost reduction or increase because the sentencing is based on the served prison sentences. This indicates the average prison sentence for cluster 2 is greater than the average prison sentence of cluster 4.

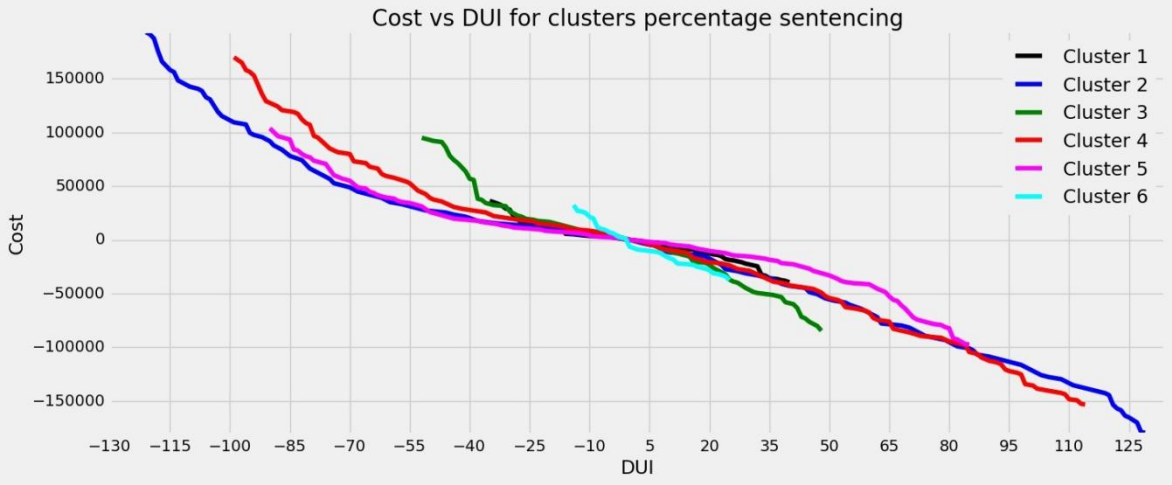

Figure 18. Cost vs. DUI for clusters percentage sentencing

From Figure 18, one can observe that Cluster 2 is the best cluster to reduce the number of DUIs with minimum cost. To reduce cost with minimum DUI increase, Cluster 3 is the best choice provided that the limit on the number of DUIs is less than 45 i.e. the maximum number of DUIs that can be raised by reducing the original sentencing is around 45. If we allow the number of DUIs to rise between 45 to 110 in order to save cost, Cluster 4 is the preferred cluster which provides the best number of DUIs to cost ratio.

6.11 Applying optimization methods

We used an a posteriori method to solve the optimization problems formulated in Section 6.8 and to present the results for the decision maker so that he/she can make the decision based on his/her preferences. The MOO method used to generate solutions is the ε-constraint method which was introduced in Section 4.5.1. Using the ε-constraint method, the cost function is used as an objective to be optimized; number of DUIs at 90 days and number of DUIs at 180 days are used as constraints. The main reason to use the ε-constraint method was because it is suitable to generate a set of Pareto optimal solutions

when a decision maker is not available. In addition to that compared to the other scalarization methods the ε-constrained method is easy to understand and to interpret the association between the constraints and the optimal solution.

After the important data attributes related to DUIs are identified, the original dataset was preprocessed to extract only the needed data attributes. Hence during the court sentencing process criminal history is considered for DUI offenders, predefined DUI groups were formed based on committed number of DUI offences in five years. Then the dataset was divided into sub-DUI groups based on the number of DUI offences an offender has committed in five years. For the DUI clusters, first the age and the number of DUIs in five years data attributes were normalized to eliminate the effect of scale differences. Then the number of clusters (6) and the random seed were provided to the KMean classifier to produce six DUI clusters.

Once the DUI groups and clusters are formed, three objective functions were defined: the number of DUIs at 3 months checkpoint, the number of DUIs at 6 months checkpoint and the cost functions. After the objective functions were defined on the dataset, polynomial regression model was used to capture the relationship of proposed sentencing and the corresponding number of DUIs at 3 months checkpoint, number of DUIs at 6 months checkpoint and cost. The right polynomial regression model was selected by comparing the residual sum of square (RSS) values. Further, the actual and estimated polynomial regression models were plotted to analyze how well the model captured the underlying patterns. For the optimization process Python's Scipy module solver was used to generate the ideal and nadir vectors. To estimate the nadir vector values, objective function values were calculated for each solution points which has yielded the ideal value of an objective function, then among these calculated objective function values the maximum objective function values for each objective function was selected to form the nadir vector. Once the ideal and nadir vectors were generated, the objective function values were normalized using the ideal and nadir vectors to produce objective function values between 0 and 1.

For the multiobjective optimization using the ε-constraint method, 100 random epsilon values were generated to create constraints on the number of DUIs at 3 months checkpoint

and 6 months checkpoint, and cost was used as the primary objective. The tolerance parameter for the solver was 10^{-40} and the maximum iteration was 10^{5}.

For the zero cost solution, setting the cost value to zero and by finding the maximum number of DUIs that can be reduced without any additional expense, the number of DUIs at 3 months checkpoint was used as an objective and cost equals to zero as a constraint.

7 Optimization results

The optimization result will be presented according to the two DUI categories and sentencing types. In the result tables, Table 7 to Table 10, columns indicate the three objectives used: DUI after 90 days, DUI after 180 days and cost. The rows indicate the ideal vector, the nadir vector and the zero cost solution.

1. Predefined DUI groups

This section includes results for the predefined DUI groups. The first table displays the result when fixed proposed sentencing is used and the second table displays the result when percentage proposed sentencing is used.

i. Fixed proposed sentencing

Table 7 shows the optimization results for predefined DUI groups based on fixed sentencing.

	DUI after 90 days	DUI after 180 days	Cost
Z ideal	-534	-334	-255987
Z nadir	255	178	403265
Zero Cost	-120	-65	0

Table 7. Predefined group fixed sentencing result

Zero cost solution: X = (-5, -47, -89, 21, 90, 89) ,

where $X = [x_1,x_2,\ldots,x_6]$ and $x_i \in [-90, 90]$ for $i=1,2,\ldots,6$.

From Table 7, which displays the optimization result of the predefined DUI groups based on fixed sentencing, the ideal vector for the three objectives is Z ideal = [-534,-334, -255987]. This shows that the maximum numbers of DUIs that can be reduced with fixed sentencing are 534 numbers of DUIs at 90 days checkpoint and 334 numbers of DUIs at the 180 days checkpoint. For the cost, the maximum cost that can be saved with fixed pro-

posed sentencing is 255987. Based on the nadir vector, z nadir = [255,178,403265] the maximum number of DUIs that can be added to the existing number of DUIs is 255 at 90 days checkpoint and 178 number of DUIs at 180 days checkpoint. Similarly, the cost value can be raised to a maximum value of 403265.

From the zero cost solution, by setting cost value to zero the maximum number of DUIs that can be reduced at 90 days checkpoint is 120 and at 180 days checkpoint the maximum number of DUIs that can be reduced is 65.

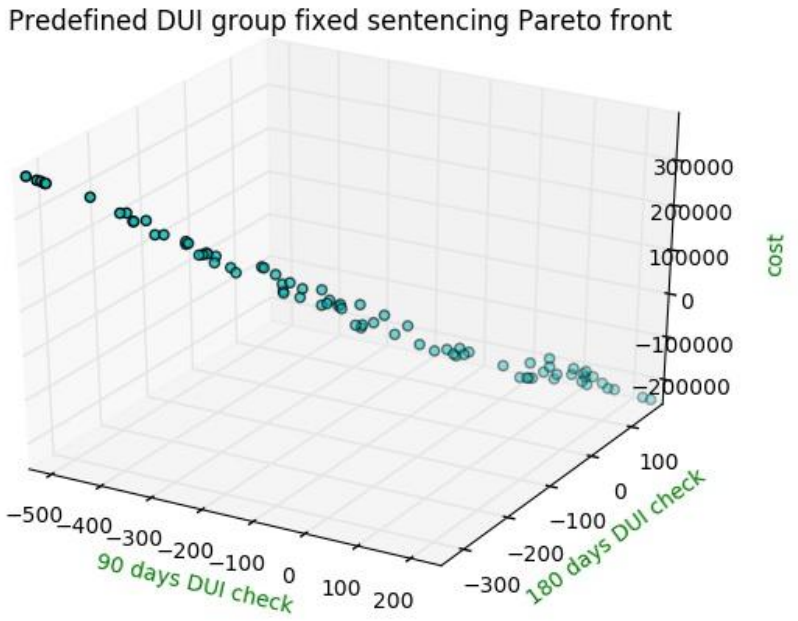

Figure 19. Predefined DUI group fixed sentencing Pareto front

Figure 19 shows the Pareto optimal front for the predefined DUI groups based on fixed proposed sentencing.

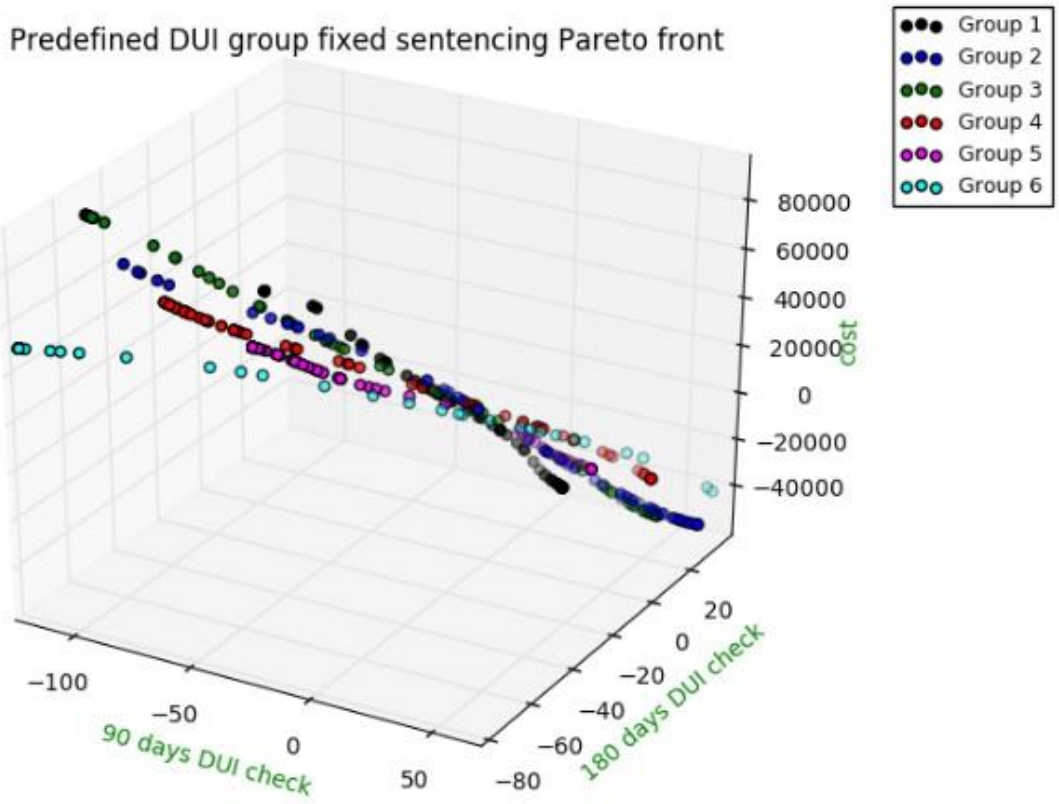

Figure 20 Distribution of the Pareto optimal front among DUI groups for fixed sentencing

Figure 19, 21, 23 and 25 represents DUI after 90 days (F), DUI after 180 days (G) and cost (C) functions for DUI clusters/groups as mentioned in Section 6.8. Hence the number of DUIs and cost functions for the DUI groups/clusters are the sum of number of DUIs and cost for each DUI group/cluster, the figure doesn't show the number of DUIs and cost values for each DUI group/cluster instead it shows the aggregated result of number of DUIs and cost values for DUI groups/clusters for a proposed sentence (x_1, x_2, ..., x_6). From the figures the decision maker can analyze the tradeoff between number of DUIs and associated cost. As described earlier negative value indicates the objective function value is reduced and positive value indicates objective function value is increased.

Figures 20, 22, 24 and 26 show which regions from a particular DUI group/cluster were used to form the Pareto optimal front for the aggregated objective function F, G and C for the corresponding figures of Figure 19, 21, 23 and 25 respectively. Even if these figures do not show the set of solutions that formed the aggregated DUIs and cost values of the Pareto

optimal front, they give an insight on how the solution sets are distributed among the individual DUI groups/clusters. Because of the high dimension, having six DUI groups/clusters, it is not possible to visualize which set of solutions from individual DUI groups/clusters formed the aggregated Pareto optimal front of the objective functions.

ii. Percentage proposed sentencing based on served prison duration

	DUI after 90 days	DUI after 180 days	Cost
Z ideal	-432	-521	-617137
Z nadir	454	338	649390
Zero Cost	-87	-120	0

Table 8. Predefined group percentage sentencing result

Zero cost solution: X = (-0.26, 0.2, -1, 0.55, 0.45, 0.49) ,

where $X = [x_1, x_2, \ldots, x_6]$ and $x_i \in [-1,1]$ for $i=1,2,\ldots,6$

In Table 8, the ideal vector [-432, -512, -617137] indicates the minimum values of the three objectives which serves as the lower bound. The nadir vector [454, 338, 649390] indicates the maximum values the three objectives can attain that can serve as the upper bound. Therefore, the Pareto front solutions will lie between the two bounds of the ideal and nadir vector values.

From the zero cost solution, by fixing the cost to zero the maximum number of DUIs that can be reduced at 90 days checkpoint is 87 and at 180 days checkpoint is 120.

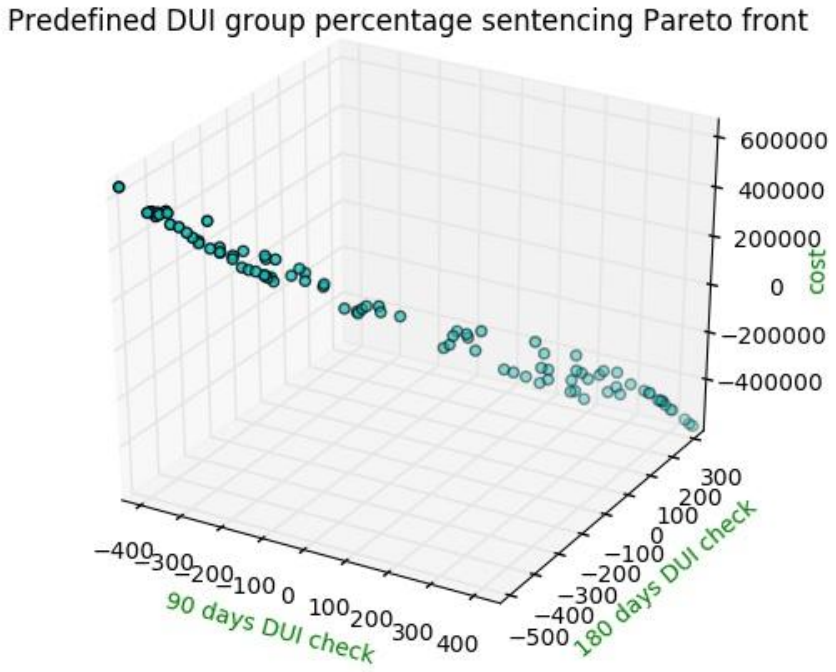

Figure 21. Predefined DUI group percentage sentencing Pareto front

Figure 21 shows the Pareto front solutions for the predefined DUI group based on percentage sentencing.

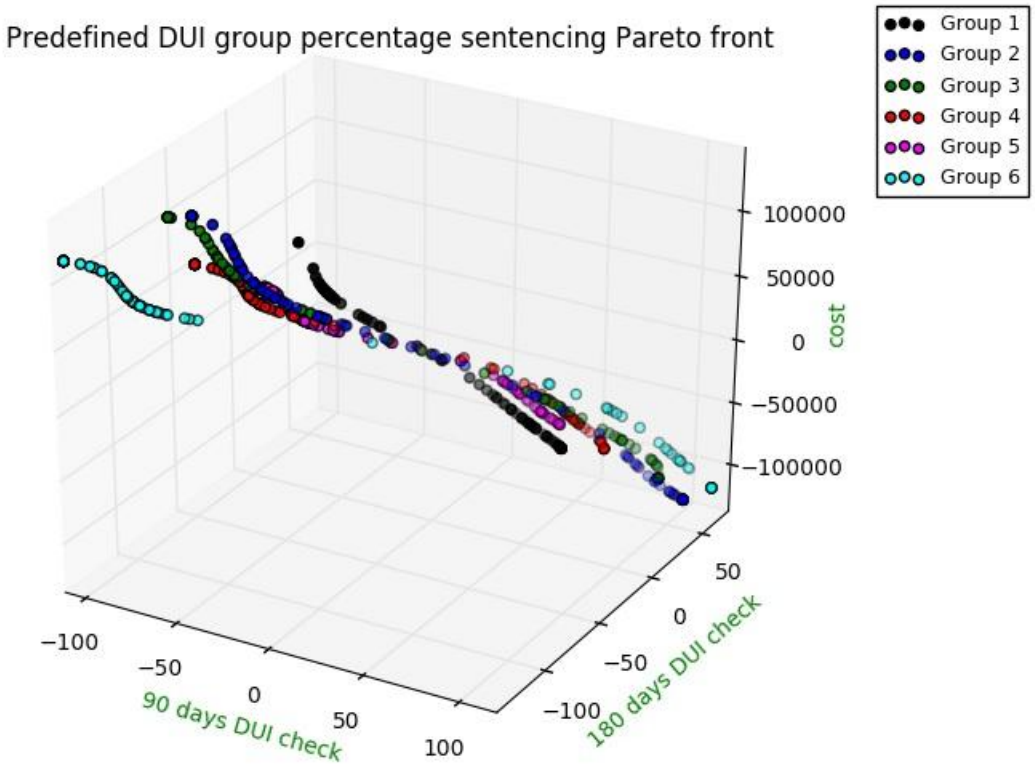

Figure 22 Distribution of the Pareto optimal front among DUI groups for percentage sentencing

2. DUI clusters

This section presents the optimization results for the DUI clusters according to the two types of sentencing. It follows the same format as the results of the optimization for the predefined DUI groups.

i. Fixed proposed sentencing

	DUI after 90 days	DUI after 180 days	Cost
Z ideal	-534	-346	-256220
Z nadir	255	176	403265

Zero Cost	-130	-58	0

Table 9, Fixed sentencing result for clusters

Zero cost solution: X = (90, 14, -90, -3, 90, -60) ,

where $X = [x_1,x_2,\ldots,x_6]$ and $x_i \in [-90, 90]$ for $i=1,2,\ldots,6$

From Table 9, the ideal vector z ideal = [-534,-346,-256220] indicates that the maximum number of DUIs that can be reduced based on DUI clusters fixed sentencing is, 534 number of DUIs at 90 days checkpoint and 346 DUIs at 180 days check point. Similar to the number of DUIs, the maximum cost value that can be saved is 256220. The nadir vector = [255, 176, 403265], indicates that upper bound of the solution. From the nadir vector, the maximum number of DUIs that can be added to the existing number of DUIs is 255 at 90 days checkpoint and 176 at 180 days checkpoint. The cost value can be raised by a maximum of 403265.

For the zero cost solution, by keeping the cost value to zero, 130 number of DUIs can be reduced at 90 days checkpoint and 58 DUIs can be reduced at 180 days checkpoint.

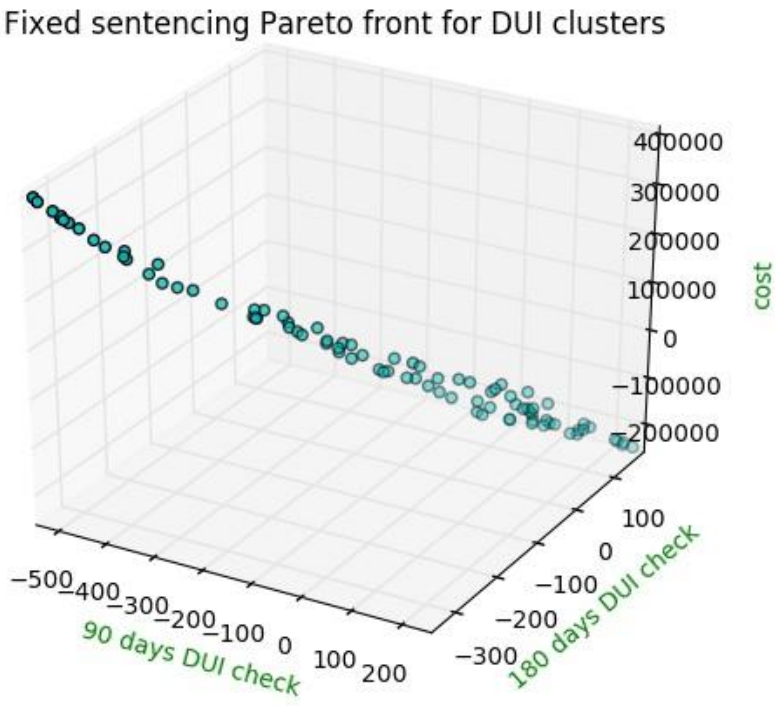

Figure 23. Fixed sentencing Pareto front for DUI clusters

Figure 23 shows the Pareto optimal solutions for DUI clusters based on fixed sentencing.

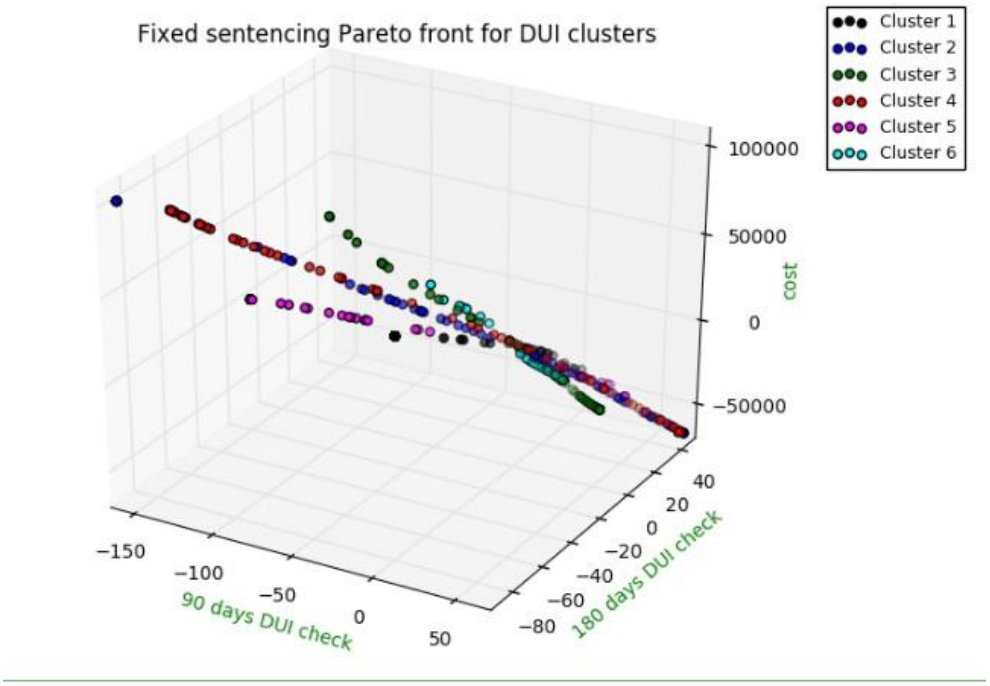

Figure 24 Distribution of the Pareto front among DUI clusters for fixed sentencing

ii. Percentage proposed sentencing based on served prison sentences

	DUI after 90 days	DUI after 180 days	Cost
Z ideal	-433	-521	-617137
Z nadir	454	338	649390
Zero Cost	-68	-84	0

Table 10 Percentage sentencing result for clusters

Zero cost solution: X = (-0.67, 0.36, -0.61, 0.9, 0.36, -1) ,

where $X = [x_1, x_2, \ldots, x_6]$ and $x_i \in [-1,1]$ for $i=1,2,\ldots,6$

From Table 10, by analyzing the ideal vector [-433,-521,- 617137] and the nadir vector [454, 338, 649390] one can observe that for the cost value reducing the original sentencing by 100% and extending the original sentence by 100% produces different result. The rea-

son is that to extend the original sentencing there is no condition to check but to reduce the original sentencing the minimum 14 days prison sentence should be preserved.

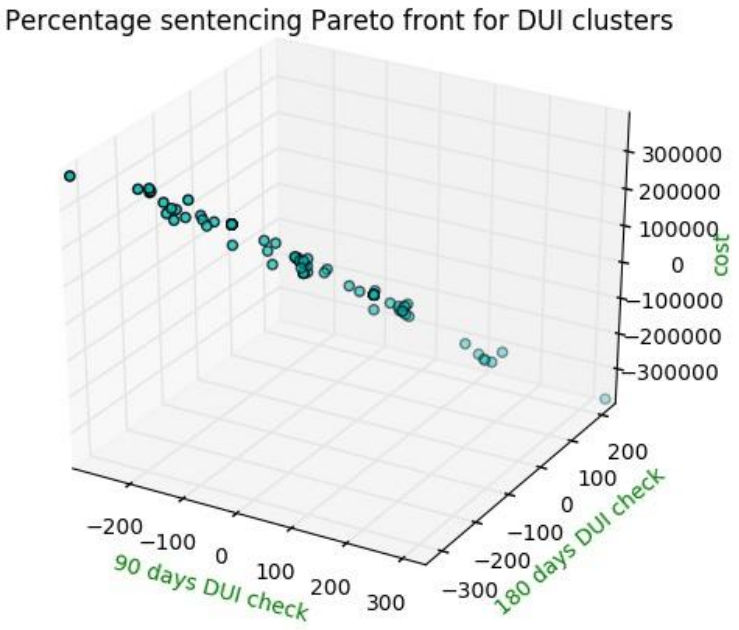

Figure 25. Percentage sentencing Pareto front for DUI clusters

Figure 25 shows the Pareto front for clusters based on percentage sentencing and from the figure one can observe that cost and the two DUI objective function values are conflicting. When the cost value declines the number of DUIs at the two checkpoints starts to rise and vice versa.

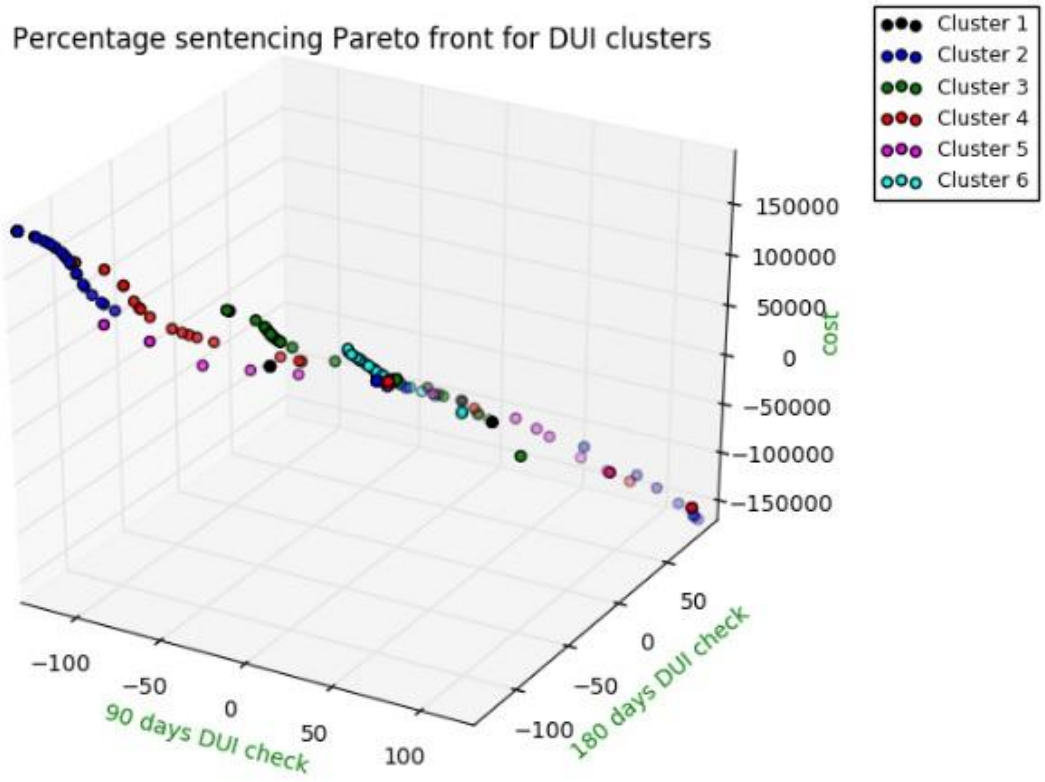

Figure 26 Distribution of the Pareto optimal front among DUI clusters for percentage sentencing

By observing the four result tables Table 7, Table 8, Table 9 and Table 10 one can states that both fixed sentencing and percentage sentencing for DUI groups and DUI clusters have the same ideal and nadir vectors. The reason is that, the ideal and nadir vector values consider the whole DUI groups together therefore, the group formation has no effect on the total outcome. On the other hand, the zero solution gives different proposed sentencing for each DUI group/cluster. As a result the formation of the groups/clusters has a direct effect on the result of the zero cost solution.

Based on the presented results, the decision maker will decide which objective value he prefers over the other i.e. reducing the number of DUIs by extending the original sentence or saving cost by reducing the original sentencing. For all the three objectives, the decision maker can pick available Pareto optimal solutions based on his/her preferences.

For the sentence type, fixed sentencing gives individuals the same number of prison sentences for all individuals in a particular DUI group/cluster and the percentage sentencing gives prison sentences for individuals based on previously served prison sentences. Therefore, percentage sentencing gives different prison sentences for different individuals in a particular DUI group/cluster.

Fourth objective

To visualize the effect of proposed sentences after one year, a fourth objective, the number of DUIs after one year was used. From the obtained result, the number of DUIs reduced or increased after one year based on proposed sentencing is very small in magnitude compared to results obtained for the number of DUIs at 90 days and the number of DUIs at 180 days. The values of the objective functions are normalized, the value 0 indicates the objective function is at its minimum value and the value 1 indicates the objective function attains its maximum value.

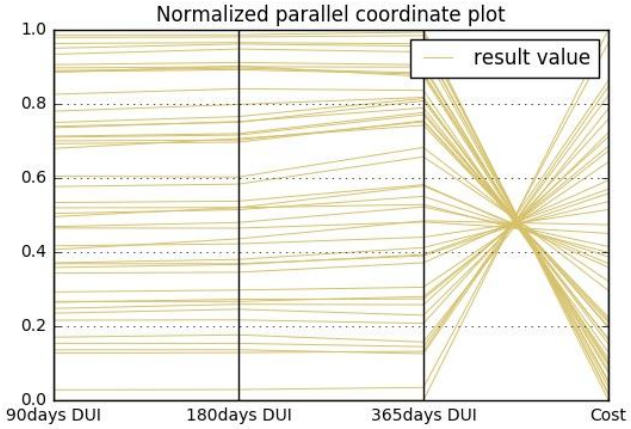

Figure 27.Normalized objective function values for four objectives

Figure 27 shows that the three objectives on the number of DUIs at different checkpoints somehow behave the same. Contrary to the three objectives related to the number of DUIs, cost function behaves differently. When the three objectives, the number of DUIs at different checkpoints, attain a maximum value of 1, cost hits the minimum value of 0. On the other hand, when cost attains its maximum value of 1, the other three objectives hits their

lowest value of zero. Therefore, the plot shows the objectives, the number of DUIs at three different checkpoints and cost, are conflicting objectives.

8 Conclusions

The research question in this thesis was: Is it possible to provide decision support to reduce the number of DUIs by restructuring the sentencing of DUI offenders?

As it was shown in the data analysis phase by analyzing the dataset, human behavior is difficult to predict and a longer prison sentence does not guarantee in discouraging an individual from committing DUIs frequently. We tried to tackle the need of reducing the number of DUIs by analyzing past behaviors of offenders, their prison sentences and the number of days between their last two DUIs.

The multiobjective optimization problem was formed by the conflicting objectives reducing cost by reducing the original sentencing and reducing number of DUIs at 3 months and 6 months checkpoint by extending the original sentencing. The approach we have used to propose a prison sentence was to group individuals with similar past DUI offenses together. The grouping result showed that individuals who had committed more DUI offences in the past tend to recommit DUI offences repeatedly. By grouping individuals with similar past DUI history and applying two types of sentencing: fixed sentencing and percentage sentencing based on served prison sentences, it was possible to come up with optimal set of solutions that reduces the number of DUIs at 3 months and 6 months checkpoints by considering the cost associated in imprisoning the individuals for the extend duration.

Furthermore, objective functions to measure the number of DUIs at checkpoints and the associated cost were constructed based on attributes obtained from the dataset. For the optimization task, an a posteriori multiobjective optimization method called ε-constraint method was used to generate set of optimal solutions.

In conclusion, there is a possibility to reduce the number of DUIs at certain time intervals by proposing new prison sentences for different DUI groups. However, the result does not support the hypothesis that groups with fewer past DUIs should get reduced sentencing and groups with higher number of past DUIs should get additional sentencing. Instead, the results show that groups with many individuals who committed their next DUI in the first three months after their prison sentence should get additional sentencing and the other

groups with fewer individuals who committed their next DUI in the first three months should get reduced sentences. Moreover, extending the prison sentences of offenders could help them to stay clean from alcohol intoxication and by providing them with additional rehabilitation programs; it would be possible to help offenders to become clean and not to recommit new DUI offences after they served their sentences.

Further research needs to be done on repeated offenders more closely by gathering detailed information regarding their criminal history, alcohol and substance abuse and childhood background thoroughly to investigate and address the root cause of committing DUI offences repeatedly.

Bibliography

[1] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela Hung Byers (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[2] Brian Runciman and Keith Gordon (2014). Big Data: Opportunities and challenges. Swindon, UK: BCS, The Chartered Institute for IT.

[3] Sivakumar (2015). How top 10 industries use big data applications(blog). Retrieved from http://www.datascienceassn.org/content/how-top-10-industries-use-big-data-applications. Visited on November 06, 2017.

[4] Daniel Larose (2005). Discovering knowledge in data : An Introduction to Data Mining (1). Hoboken, US: Wiley-Interscience.

[5] Jerry Smith (2013). Six types of analyses every data scientist should know (blog). Retrieved from https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/ Visited on November 06, 2017.

[6] Subrata Kumar Das (2014). Computational business analytics. Boca Raton, FL: Chapman and Hall/CRC.

[7] Gretta Fitzgerald and Mike FitzGibbon (2014). A Comparative analysis of traditional and digital data collection methods in social research in LDCs - Case Studies Exploring Implications for Participation, Empowerment, and (mis)Understandings. Cape Town, South Africa: The International Federation of Automatic Control.

[8] Sasha Issenberg (2012, Dec 17). How Obama wrangled data to win his second term, Retrieved from https://www.technologyreview.com/s/508851/how-obama-wrangled-data-to-win-his-second-term/ Visited on November 06, 2017.

[9] Finnish annual road safety review (2013). Finland, Finnish Transport Safety Agency.

[10] Ian Witten, Eibe Frank, Mark Hall and Christopher Pal (Eds) (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.

[11] Jiawei Han, Micheline Kamber and Jian Pei. (2011). Data mining: concepts and techniques. Saint Louis: Elsevier Science.

[12] David Olson and Dursun Delen (2008). Advanced data mining techniques. Springer-Verlag Berlin Heidelberg.

[13] Rüdiger Wirth and Jochen Hipp (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the fourth international conference on the practical application of knowledge discovery and data mining. P 29 - 39.

[14] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. (1996). From data mining to knowledge discovery: An overview. Advances in knowledge discovery and data mining. AI Magazine 17 (3) 37 - 51.

[15] Robert Nisbet, John Elder and Gary Miner(2009). Handbook of statistical analysis and data mining applications. Amsterdam; Boston: Academic Press/Elsevier.

[16] Rui Xu and Donald Wunsch.(2009). Clustering. Oxford: Wiley-IEEE Press.

[17] Ethem Alpaydin. (2014). Introduction to machine learning (Third edition.). The MIT Press.

[18] Jaime Carbonell, Ryszard Michalski and Torn Mitchell (Eds.) (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[19] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar. (2012). Foundations of machine learning. Cambridge, MA: MIT Press.

[20] Gavin Hackeling. (2014). Mastering machine learning with scikit-learn. Birmingham: Packt Publishing.

[21] Olivier Chapelle, Bernhard Schölkopf and Alexander Zien. (Eds.) (2010).Semi-supervised learning. MIT Press.

[22] Marco Wiering and Martijn van Otterlo (Eds.) (2012). Reinforcement learning: State-of-the-art. Berlin, Heidelberg.

[23] Supervised vs. unsupervised learning: Know the difference (2015). Retrieved from: http://datacafeblog.com/supervised-vs-unsupervised-learning-know-the-difference/ Accessed November 6, 2017.

[24] Ian Witten and Eibe Frank. (2005). Data mining: Practical machine learning tools and techniques (2nd ed.). Morgan Kaufman.

[25] Jason Bell (2015). Machine learning: Hands-on for developers and technical professionals. Indianapolis, Indiana: Wiley.

[26] Jürgen Schmidhuber (2015). Deep learning in neural networks: An overview. Neural Networks, Vol 61, pp 85-117.

[27] Michael Nielsen (2015). Neural networks and deep learning. Determination Press.

[28] Christian Blum (2005). Ant colony optimization: Introduction and recent trend. Elsevier.

[29] G. L. Nemhauser , A. H. G. Rinnooy Kan and M. J. Todd (1989). Optimization. New York: North-Holland.

[30] Giacomo Zambelli, Gérard Cornuéjols and Michele Conforti (2014) Integer Programming. Springer.

[31] Stephen VaVasis (1991). Nonlinear optimization complexity issues. Oxford University Press.

[32] Ben Rosen and Panos M. Pardalos (1987). Constrained Global Optimization: Algorithms and Applications. Springer Berlin Heidelberg.

[33] Marco Cavazzuti (2013). Optimization methods: From theory to design: scientific and technological aspects in mechanics. Berlin: Springer.

[34] Jürgen Branke, Kalyanmoy Deb, Kaisa Miettinen and Roman Slowinski (Eds.) (2008) Multiobjective Optimization: Interactive and Evolutionary Approaches. Springer.

[35] Kalyanmoy Deb and Kaisa Miettinen (2010) Nadir Point Estimation Using Evolutionary Approaches: Better Accuracy and Computational Speed Through Focused Search. In: Ehrgott M., Naujoks B., Stewart T., Wallenius J. (eds) Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems. Lecture Notes in Economics and Mathematical Systems, vol 634. Springer, Berlin, Heidelberg.

[36] Kaisa Miettinen (1999). Nonlinear multiobjective optimization. Boston: Kluwer Academic Publishers.

[37] Kaisa Miettinen and Jussi Hakanen (2009) Why Use Interactive Multi-Objective Optimization in Chemical Process Design? In " Multi-objective optimization : Techniques and Applications in Chemical Engineering" Edited. by G. P. Rangaiah, pp 155 - 160, World Scientific.

[38] Project Jupyter. http://jupyter.org/ (November 7, 2017).

Appendices

A Dataset attributes description

Numero	Nimi	Suomi	In English
1	laskuri	id-muuttuja	id
2	sukupuoli	0=nainen, 1=mies	sex (0=man, 1=woman)
3	rikoskoodi	230301=rattijuopumus, 210401=törkeä rattijuopumus	230301=DUI, 230401=aggravated DUI
4	vankila	ehdottoman tuomion pituus päivinä	length of prison conviction
5	tuomiot_5v	tuomiot 4-5 vuotta sitten (rikosten yhteismäärä)	crimes in convictions 4-5 years ago (n of separate crimes in convictions)
6	tuomiot_4v	tuomiot 3-4 vuotta sitten (rikosten yhteismäärä)	crimes in convictions 3-4 years ago
7	tuomiot_3v	tuomiot 2-3 vuotta sitten (rikosten yhteismäärä)	crimes in convictions 2-3 years ago
8	tuomiot_2v	tuomiot 1-2 vuotta sitten (rikosten yhteismäärä)	crimes in convictions 1-2 years ago
9	tuomiot_1v	tuomiot 0-1 vuotta sitten (rikosten yhteismäärä)	crimes in convictions 0-1 years ago
10	vvtuomiot	väkivaltarikostuomiot 5 vuotta ennen (rikosten yhteismäärä)	violent crimes past 5 years
11	ortuomiot	omaisuusrikostuomiot 5 vuotta ennen (rikosten yhteismäärä)	property crimes past 5 years
12	rjtuomiot	rattituomiot 5 vuotta ennen (rikosten yhteismäärä)	DUIs past 5 years
13	hatuomiot	alkoholi- ja huumetuomiot 5 vuotta ennen (rikosten yhteismäärä)	alcohol and drug offences past 5 years
14	lrtuomiot	liikennerikostuomiot 5 vuotta ennen (rikosten yhteismäärä)	traffic offences past 5 years (n of convictions)
15	ehdottomat_5v	ehdottomat 4-5 vuotta sitten (erillisten tuomioiden määrä)	prison convictions 4-5 years ago
16	ehdottomat_4v	ehdottomat 3-4 vuotta sitten (erillisten tuomioiden määrä)	prison convictions 3-4

Numero	Nimi	Suomi	In English
17	ehdottomat_3v	ehdottomat 2-3 vuotta sitten (erillisten tuomioiden määrä)	prison convictions 2-3
18	ehdottomat_2v	ehdottomat 1-2 vuotta sitten (erillisten tuomioiden määrä)	prison convictions 1-2
19	ehdottomat_1v	ehdottomat 0-1 vuotta sitten (erillisten tuomioiden määrä)	prison convictions 0-1
20	vankeuskaudet_5v	ehdottomat 5 vuotta ennen (erillisten tuomioiden määrä)	prison convictions past 5 years
21	ehdolliset_5v	ehdolliset 5 vuotta ennen (erillisten tuomioiden määrä)	suspended sentences past 5 years
22	ykp_5v	ykp:t 5 vuotta ennen (erillisten tuomioiden määrä)	community services past 5 years
23	time_from_ offence_to_ao	aika teosta käräjäoikeuteen	time from offence to district court
24	time_from_ao_to_ho	aika käräjäoikeudesta hovioikeuteen (puuttuva=ei hovioikeuteen)	time from district court to appeals court (if applicable)
25	time_from_offence _to_ratkaisupvm	aika teosta lainvoimaiseen tuomioon	time from offence to final court decision (district or appeals)
26	age_in_years_floor	ikä tekohetkellä, pyöristetty alaspäin seuraavaan kokonaislukuun	age at the time of the offence
27	rikokset_tuomioissa	rikosten määrä nykyisessä tuomiossa	number of separate crimes in current conviction
28	rattien_maara_tuomiossa	rattien määrä nykyisessä tuomiossa	number of DUIs in current conviction
29	duration	vankeuskauden todellinen kesto	true length of prison term served
30	seur_r_tuomio	aika seuraavan rattituomioon (1096=ei uusinut)	next DUI (1096=no new DUI within 3 years)
31	seur_sv_tuomio	aika seuraavan rattiin tai vakavampaan (1096=ei uusinut)	next DUI or more severe (1096=no new offence within 3 years)
32	move_or_death	seuraava muutto tai kuolema	time of next move or death