TU Dortmund

Case Studies

Project II: BTA Deep Hole Drilling Process

Lecturers:
Prof. Dr. Uwe Ligges,
M.Sc. Leonie Schürmeyer

Author: Aymane Hachcham

Group number: 3

Group members: Muhammad Raafey Tariq
Farrukh Ahmed
Amirreza Khamehchin Khiabani

June 30, 2023

Contents

1 Introduction

One common application of deep drilling processes is to manufacture holes with a significant ratio of length to diameter (400:1). The tools employed in this process frequently feature an asymmetric cutting edge configuration that generates extremely accurate and smooth holes. The configuration allows for the use of guide pads to enable the self-guidance of the tool within the newly generated bore (Haddag et al., 2013).

The BTA (*Boring and Trepanning Association*) deep hole drilling technique is a frequently employed method for boring holes with diameters ranging from approximately 18 to 2000 mm. The method is commonly utilized in the production of precision tubes and for the hollow drilling of turbine rotors, among other applications (Webber, 2006).

In order to create holes with the significant length-to-diameter ratio, long-shaft tools are essential. However, with longer tools comes increased dynamic compliance, making deep drilling processes more vulnerable to dynamic process disturbances (Webber, 2006).

The most noticeable disturbances are the so-called *chatter vibrations* or *chattering* and *spiralling*. Chattering primarily results in increased tool wear and may leave marks on the discarded bottom of the hole. On the other hand, spiraling can cause a multi-lobe shape deviation of the cross section of the hole, which can significantly impair the quality of the workpiece by deviating from the desired roundness. These disturbances are caused by the interaction between the tool and the workpiece, and may cause defects that do not meet the required standards of the final product (Weinert et al., 2004)

The data used to assess the BTA deep drilling process is extracted from the project C5 (Analyse und Modellbildung des Tiefbohrprozesses mit Methoden der Statistik und Neuronalen Netzen, SFB 475), a reasearch project conducted by the Institut für Spanende Fertigung, FB Maschinenbau at the University of Dortmund.

The aim of this project is to model the drilling process and detect early signs of chattering and spiralling, with especial focus on the dynamic aspects of the process.

Section 2.1 gives a detailed presentation of the measurement technicalities and the data collection process.

In section 3 we present the statistical methods used throughout the analysis and discuss the assumptions and limitations of each method in the context of the data at hand.

The Analysis part is discussed in section 4.

Section 5 holds the interpretation and conclusions on the results obtained, summarizing the main findings and discusses further possible analyses on the data.

2 Problem statement

2.1 Data Assessment

The data presented hereafter provides information on blood pressure measurements taken during the *Wege zur Gesundheit* exhibition in Bruck an der Mur in 2006.

Variable name	**Data type**	**Data scale**
Id	Discrete	Ordinal
Time	Discrete	Ordinal
Terminal	Discrete	Cardinal
Postal code	Discrete	Cardinal
Municipality	Discrete	Nominal
District	Discrete	Nominal
Federal state	Discrete	Nominal
Felt health condition	Discrete	Ordinal
Date of birth	Discrete	Nominal
Gender	Discrete	Nominal
Is smoker	Discrete	Nominal
Is diabetic	Discrete	Nominal
Has cholesterol	Discrete	Nominal
In treatment	Discrete	Nominal
Measured_sys_bp	Continuous	Cardinal
Measured_dia_bp	Continuous	Cardinal
Self-reported_sys_bp	Continuous	Cardinal
Self-reported_dia_bp	Continuous	Cardinal

Table 1: Variables in the dataset and their data types.

The dataset comprises 16386 entries of self-reported and measured systolic and diastolic blood pressures. Table 1 lists the 18 variables included in the dataset, that describe the

physiological attributes and the geographic location of the subjects. The data collection was conducted using a planned experiment.

The *Id* variable is a unique identifier for each entry and is auto-incremented. The *Time* variable is a timestamp of the measurement, sorted by date and time. The variables *Postal code*, *Municipality*, *District*, and *Federal state* describe the participant's place of residence. These 4 variables represent the same information in different levels of granularity,

The *Terminal* indicates the terminal number (1, 2, or 3) at which the measurement was taken. The *Felt health condition* variable is the participant's self-reported health condition, rated from 1 for "very good" to 5 for "very poor". *Date of birth* records the year of birth of the participant, and thus their respective age.

From the list of physiological characteristics of the subjects we have the *Gender* variable indicating the biological sex (male or female), *Is diabetic* (known to have diabetes, yes or no) and *Has cholesterol* (known to have high cholesterol, yes or no). *Is smoker* variable refers to whether the participant is a smoker (true or false) and is also a complementary information to the health status of the subjects. *In treatment* variable informs on the medical treatement that the subjects would receive in case of hypertension (true or false).

Lastly, *Self-reported_sys_bp* and *Self-reported_dia_bp* are the participant's self-reported systolic and diastolic blood pressures respectively, that are contrasted with the measured systolic and diastolic blood pressures in the *Measured_sys_bp* and *Measured_dia_bp* variables. The systolic blood pressure is the pressure in the arteries when the heart contracts (beats), while the diastolic is the pressure in the arteries when the heart is at rest (between beats) (Flint et al., 2019). The blood pressures are give in units of millimeters of mercury (mmHg).

We also observe 135 total surveys conducted outside the data collection window with 132 observations before April 2006 and 3 observations after October 2006. We have a total of 381 missing values in the dataset summarized per variable name in Table 2.

Variable name	Nr. Missing Values
Postleitzahl	334
Gemeinde	331
Bezirk	331
Bundesland	331
Befinden	23
Geburtsjahr	23
Gesclecht	23
Schaetzwert_bp_sys	45
Schaetzwert_by_dia	56

Table 2: Missing values per variable within the collection window from April to October 2006.

About 331 rows in the dataset are records belonging to foreign visitors of the exhibition. Among the observations, there were 22 records with missing values for the variables *Felt health condition*, *Date of birth*, and *Gender*, which were associated with individuals who were foreign residents. The overall data quality is good.

2.2 Project Objectives

The aim of this project is to analyze self-estimated and measured blood pressures of different individuals. We endeavor to develop a robust linear regression model that can accurately predict systolic and diastolic blood pressures using the variables at hand.

Additionally, we plan to employ regression trees and random forests as alternative modeling techniques and benchmark the best model using mean squared error (MSE) on both training and testing sets. To improve the performance of the models, we employ cross-validation and best subset selection algorithms.

We project to identify the most significant predictors of blood pressure measurements and develop a reliable model that can help understand the relationship between the various covariates and the target.

3 Statistical methods

This part focuses on the methods used throughout the report to conduct the analysis.

3.1 The Linear Regression Model

Below, we delve deeper into the concept of linear regression models, discussing various aspects of the model including its assumptions, parameter estimation techniques, confidence intervals for the parameters, and procedures for diagnosing the model.

Model and Assumptions

Let y be a variable of interest also called response variable conditioned on a set of covariates $x_1, x_2, ..., x_p$, the aim is to model a linear relationship between both such as $y = f(x_1, x_2, ..., x_p)$ (Fahrmeir et al., 2013, p.73).

The systematic component f is designed as a linear combination of the covariates. This assumption constitutes the fundamental basis of our linear model.

$$f(x_1, x_2, ..., x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3.1}$$

where $(\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ are the coefficients to be estimated.

The classical linear model assumes that the errors are independent and identically normally distributed with mean zero and constant variance σ^2 (Fahrmeir et al., 2013, p.76):

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

We jointly assume that the errors are homoscedastic with constant variance, $Var(\varepsilon_i) = \sigma^2$ and are generally not correlated between each other, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ (Fahrmeir et al., 2013, p.75).

Therefore, incorporating the error term into the model yields the following equation for each observation $i = \{1, 2, \ldots, n\}$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \tag{3.2}$$

A more compact matrix notation that encompasses all the observations is given by:

$$\boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.3}$$

where $\boldsymbol{y}$ is the vector of responses $(y_1, y_2, \ldots, y_n)$, $\boldsymbol{X}$ represents the *design matrix*,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$$

$\boldsymbol{\beta}$ the vector of coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ and $\boldsymbol{\varepsilon}$ the vector of errors $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$.

The *design matrix* $\boldsymbol{X}$ should be of full column rank, the number of observations n should be equal or superior to the number of covariates (Fahrmeir et al., 2013, p.75).

Parameter Estimation and Residuals

The aim is to model the expected value of the response conditioned on the covariates, following the distributional assumptions explained above. In order to do so, we need to estimate the model coefficients using the observations (Fahrmeir et al., 2013, p.77).

A direct estimator $\hat{y}_i$ of the response y_i given by the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ can be expressed as follows:

$$\widehat{\mathbb{E}(y_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip} \tag{3.4}$$

The difference between the true value y_i and the estimated one $\hat{y}_i$ is called residual, $\hat{\varepsilon}_i = y_i - \hat{y}_i$ (Fahrmeir et al., 2013, p.77).

The coefficients are estimated using the *Least Squares* method. This is akin to estimating the Maximum Likelihood of parameters while assuming a normal distribution of the errors. (Fahrmeir et al., 2013, p.105-106).

$$LS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.5}$$

With a careful manipulation of the above expression we obtain the Least Squares estimator of $\hat{\beta}$ as the following expression (Fahrmeir et al., 2013, p.107):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3.6}$$

From equation 3.6, we can derive the predictor $\hat{y}$ as:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{3.7}$$

The $n \times n$ matrix $\boldsymbol{H}$ is the hat matrix determined by the expression: $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ (Fahrmeir et al., 2013, p.108).

Due to the presence of heteroscedastic and correlated residuals, it becomes challenging to validate the assumption of homoscedastic errors. Heteroscedastic residuals do not necessarily imply the presence of heteroscedastic errors in general, making it difficult to ascertain the true nature of the errors (Fahrmeir et al., 2013, p.124). Thus, we prefer using standardized residuals, which are defined as follows:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}$$

where h_{ii} is the diagonal entry of the Hat matrix at observation i, $\hat{\varepsilon}_i$ is the normal residual, and $\hat{\sigma}$ is the estimated standard deviation of the residuals (Fahrmeir et al., 2013, p.124).

Assessing the Model's accuracy

To assess the goodness of fit of the linear model, the coefficient of determination, denoted as R^2, is often utilized (Fahrmeir et al., 2013, p. 112). It is defined as the proportion of variance in the response that is explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3.8}$$

While the R^2 statistic provides a useful measure of model fit, it has a limitation in that it does not adequately account for the increase in model parameters. To address this issue, the adjusted R^2 statistic is preferred. It includes a correction term that effectively handles this problem, as highlighted by (Fahrmeir et al., 2013, p.148).

The adjusted R^2 statistic is defined as follows:

$$\bar{R}^2 = 1 - \frac{n-1}{n-p} \cdot (1 - R^2) \tag{3.9}$$

where n is the number of observations and p is the number of covariates.

Test of Significance and Confidence Intervals

A test of significance is carried out to verify which estimated parameters $\hat{\beta}_j$ are contributing to the model. The dual relationship between statistical tests and confidence intervals allows us to derive confidence intervals for the regression parameters as well (Fahrmeir et al., 2013, p.126).

The hypothesis test is defined as follows:

$$\begin{aligned} &H_0 : \beta_0 = 0 \quad \text{against} \quad H_1 : \beta_0 \neq 0 \\ &H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0 \\ &\vdots \\ &H_0 : \beta_p = 0 \quad \text{against} \quad H_1 : \beta_p \neq 0 \end{aligned}$$

The test statistic is given by:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim t_{n-(p+1)} \tag{3.10}$$

t_j follows a Student's t distribution with $n-(p+1)$ degrees of freedom (Fahrmeir et al., 2013, p.130-131).

The p-value

The p-value is the probability of observing a test statistic at least as extreme as the one computed from the sample data, given a specific level of significance α (Fahrmeir et al., 2013, p.130-131). It states the minimal significance level at which the null hypothesis can be rejected.

The p-value is computed as follows:

$$p = 2 \cdot P(T_{n-(p+1)} > |t_j|) \tag{3.11}$$

The Confidence intervals

The $(1-\alpha)$ confidence interval, by definition, encompasses the corresponding estimated parameter $\hat{\beta}_i, i = 0, ..., p$, with a probability of $1 - \alpha$.

For high sample sizes, the estimator $\hat{\beta}_i$ is asymptotically normally distributed for random observations of y_i. Therefore, the confindence interval for $\hat{\beta}_j$ with $1-\alpha$ level is given as:

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2) \cdot se_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2) \cdot se_j] \tag{3.12}$$

where $t_{n-p}(1-\alpha/2)$ is the critical value of the t test at the significance level α, and se_j is the standard error of $\hat{\beta}_j$: $se_j = \sqrt{\widehat{Var(\hat{\beta}_j)}}$ (Fahrmeir et al., 2013, p.131).

3.2 Best Subset Selection

Best Subset Selection is a method that aims to find the best model among all possible models that can be constructed from a given set of p predictors. It fits a linear model for each possible combination of the covariates and selects the best one upon a given criterion (James et al., 2013, p.255).

The algorithm is as follows:

Algorithm 1 Best Subset Selection

Input: Design matrix $\mathbf{X}$, response vector $\mathbf{y}$, maximum number of predictors p

Output: Best subset selection model $\hat{f}(x)$

Define M_0 as the constant model with zero regressors. M_0 simply predicts the sample mean for each observation.

for $p = 1, 2, ..., k$ **do**

  for all $\binom{k}{p}$ model combinations with p regressors **do**

    Fit the model using least squares regression

    From all model combinations, choose the model that minimizes the information criterion as the best model and name it M_p.

  end for

end for

After iterating over all best models $M_0, ..., M_k$, choose the best one among them with respect to an information criterion, such as the Mean Squared Error (MSE).

3.3 Pearson correlation

The Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It is defined as the covariance between X and Y divided by the product of their standard deviations (Fahrmeir et al., 2013, p. 105):

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.13}$$

where $cov(X,Y)$ is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively.

The Pearson correlation coefficient is bounded between -1 and 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation and 0 indicates no linear correlation between the variables (Fahrmeir et al., 2013, p. 105).

3.4 Normal Probability QQ plots

The QQ-plot is a graphical tool used to compare a data sample with the standard normal distribution (Hay-Jahans, 2019, p.147).

Considering a sample $(y_{(1)}, y_{(2)}, ..., y_{(n)})$ of n observations, the QQ plot is constructed by rearranging the sample in ascending order and plotting the probability points p_i for $i = 1, 2, ..., n$, where:

$$p_i = \begin{cases} (i - 3/8)/(n + 1/4) & \text{if } m \leq 10 \\ (i - 1/2)/n & \text{if } n > 10 \end{cases}$$

The probability points p_i are used to find the theorical quantiles x_i. For Normal probability qq plots we find x_i such as $p_i = P(X \leq x_i)$ where $X \sim N(0,1)$ (Hay-Jahans, 2019, p.148).

3.5 Residual plots

The residual plot is used to assess the assumption of constant variance in the linear model. It plots the residuals against the fitted values of the response variable (Hay-Jahans, 2019, p.149). By doing so, we check for heteroscedasticity, which is the non-constant variance of the residuals.

3.6 Mean Squared Error

The Mean Squared Error (MSE) is a measure of the average squared difference between the observed values and the predicted values (James et al., 2013, p.50).

It is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{3.14}$$

3.7 Regression Trees

Regression trees are non-parametric methods used to model the relationship between a response variable and a set of predictors. They are a powerful tool to model non-linear relationships (James et al., 2013, p.372). Below we explained the different steps involved in the algorithm.

Stratification of the Feature space

We first start by stratifying the feature space into J distinct and non-overlapping regions $R_1, R_2, ..., R_J$. For an observation x_i, the prediction is given by the mean of the response y_i in the region R_j to which x_i belongs (James et al., 2013, p.372).

$$\hat{f}(x) = \sum_{j=1}^{J} c_j \cdot I(x \in R_j) \tag{3.15}$$

where c_j is the mean response for the training observations in the jth box.

Recursive Binary Splitting

The algorithm used to construct the regression tree is called CART (*Classification and Regression Trees*). It is a *greedy* algorithm, meaning that at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree (James et al., 2013, p.372).

This is how the CART algorithm works:

Algorithm 2 Recursive Binary Splitting

Input: Design matrix $\mathbf{X}$, response vector $\mathbf{y}$

Output: Regression tree $\hat{f}(x)$

Divide the predictor space into J distinct and non-overlapping regions $R_1, R_2, ..., R_J$

For each $j = 1, 2, ..., J$, compute the mean response c_j for the training observations in R_j

For any j and any predictor X_j, define the pair of half-planes

$$R_1(j, s) = \{X | X_j < s\} \text{ and } R_2(j, s) = \{X | X_j \geq s\} \tag{3.16}$$

where s is a threshold value for X_j

Choose the predictor j and the threshold value s such that the error

$$\sum_{i:x_i \in R_1(j,s)} (y_i - c_1)^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - c_2)^2 \tag{3.17}$$

is minimized

Repeat the process on the two regions $R_1(j, s)$ and $R_2(j, s)$ until a stopping criterion is reached.

The hyperparameters that are considered to generate the best regression model are the following:

- **The max depth**: This hyperparameter specifies the maximum depth or the maximum number of levels in the tree.
- **The max features used**: This hyperparameter determines the maximum number of features (predictors) considered for each split
- **The minimum sample leaves**: This hyperparameter determines the minimum number of samples required to be at a leaf node.
- **The splitter**: This hyperparameter determines the strategy used to choose the split at each node.

3.8 Random Forests

Random Forests are an ensemble method that combines many regression trees into a single model. Like in bagging, a number of decision trees are built on bootstrapped training

samples. However, in Random Forests, only a subset of the predictors is considered at each split (James et al., 2013, p.385-386).

Bootstrap sampling is the process of taking repeated samples from the training data with replacement. The bootstrap sample is the training set for growing the tree. The remaining observations not included in the bootstrap sample are called *out-of-bag* (OOB) observations.

The algorithm is as follows:

Algorithm 3 Random Forests

Input: Design matrix $\mathbf{X}$, response vector $\mathbf{y}$, number of trees B

Output: Random Forests model $\hat{f}(x)$

for $b = 1, 2, ..., B$ **do**

Draw a bootstrap sample Z^* of size n from the training data

Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached:

1. Select m variables at random from the p variables
2. Pick the best variable/split-point among the m
3. Split the node into two daughter nodes

end for

Output the ensemble of trees $\{T_b\}_1^B$

The steps of the algorithm are the following:

1. A bootstrap sample is drawn from the training data. This means that a sample of size n is drawn from the training data with replacement.
2. A random forest tree is grown to the bootstrapped data. This means that a decision tree is grown to the bootstrapped data, but when splitting a node, only m randomly selected variables are considered as split candidates. The split is allowed to use only one of those m variables.
3. The process is repeated until the minimum node size n_{min} is reached.
4. The number of trees B is usually chosen to be large enough so that the error has stabilized.

5. The algorithm is repeated B times, and the final prediction is given by the average of the predictions of all the trees.

3.9 Fine Tuning using Cross-Validation

In order to fine tune the random forest model and find the optimal values for the hyperparameters, cross validation is used. The idea behind cross-validation is to split the available data into two separate sets: a training set and a testing set. The training set is used to fit the model, while the testing set is used to evaluate its performance (James et al., 2013, p.291)

The process for implementing cross-validation is as follows:

1. The data is split into K folds.
2. For each fold $k = 1, 2, ..., K$:
    a) The model is trained on all the folds except the kth fold.
    b) The model is tested on the kth fold.
3. The average mean squared error across all K folds is computed.
4. The process is repeated for different values of the hyperparameters.
5. The hyperparameters that give the lowest average error are chosen.

4 Statistical analysis

The following analysis is conducted using the Python programming language and the following libraries:

- `Pandas, Version 2.0.0` (Pandas Development Team, 2021)
- `Numpy, Version 1.24.23` (Oliphant, Travis E. and The NumPy Community, 2022)
- `Sci-kit learn, Version 1.2.1` (Pedregosa, F. et al., 2021)
- `Statsmodels, Version 1.24.2` (Statmodels API, 2021)
- `Seaborn, Version 1.2.1` (Waskom, Michael and Seaborn Development Team, 2021)

4.1 Descriptive analysis and data transformation

Feature engineering

Before proceeding to the exploratory analysis part, we performed on the data a set of preprocessing operations that are relevant for further analysis assessment.

We started by feature engineering[1] some of the categorical variables. We created the variable *Age*, based on the *Date of birth* and the date of the blood pressure measurements (April 2006). We also generated new variables for the month, day, and hour based on the *Time* variable to investigate whether the time of day or month of the year had any effect on blood pressure. Additionally, we split the *Terminal* 3 variable into two subgroups, "3a" and "3b," after the measurement device was changed.

We incorporated meteorological data, such as temperature, humidity, and weather conditions, into our analysis. Specifically, we included the meteorological readings taken on the day of the measurements. For temperature we included the maximum, minimum, and average temperature readings in degrees Celsius. The humidity variable is the average relative humidity in percent. The additional variables can help to account for potential confounding factors that may influence an individual's blood pressure.

In addition we identified foreign visitors based on their postal code, district, and municipality. They amount to 331 observations, which is 0.2% of the total number of observations.

Table 3 shows the results of the dummy coding process for the categorical variables.

[1] `https://en.wikipedia.org/wiki/Feature_engineering/`

Variable name	**Factor levels**
Time	187
Terminal	3
Postal code	813
Municipality	992
District	99
Federal state	9
Felt healt condition	5
Gender	2
Is smoker	2
Is diabetic	2
Has cholesterol	2
In treatment	2
Month	7
Day	7

Table 3: Lis of the categorical variables in the dataset and their respective factor level after applying dummy coding.

We clearly notice that the variables *Postal code*, *Municipality*, *Time*, and *District* have a high cardinality. And therefore, can pose challenges in modeling because including such variables can lead to overfitting, high variance, and poor model performance.

Moreover, the variables *Postal code*, *Municipality*, and *District* contain the same information as the variable *Federal state* (geographical information). Keeping them in the analysis would lead to multicollinearity issues. Hence, we decided to drop the variables *Postal code*, *Municipality*, and *District* and keep only the variable *Federal state* with 9 factor levels.

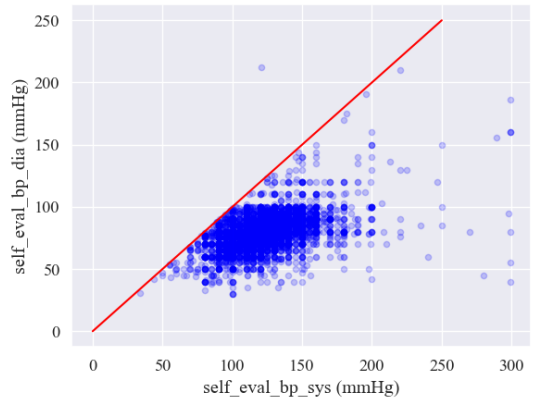
Further to this, we investigated discrepancies in some specific variables. The variable *Age* for example contains observations with age values superior to 110 years, inferior to 15 years (considering the fact that children under 15 years old are not allowed to donate blood), or even some observations with zero values. We decided to trim any observation above 100 years old and below 15 years old. The variable *Id* breaks off quite a few times with huge jumps at some moments. The ids are supposed to be unique and consecutive, but this is not the case. We decided to drop the variable *Id*.

Ultimately, the number of variables increased to 26 after the feature engineering process, taking into account the categorical variables and their factor levels.

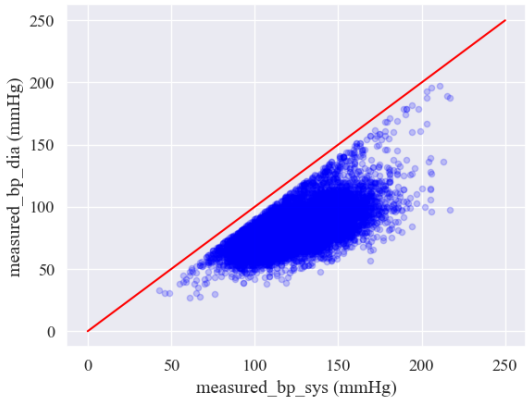
Lastly, we removed all null values from the dataset. These preprocessing steps were necessary to prepare the data for further analysis and ensure that the data was clean and complete.

Descriptive analysis

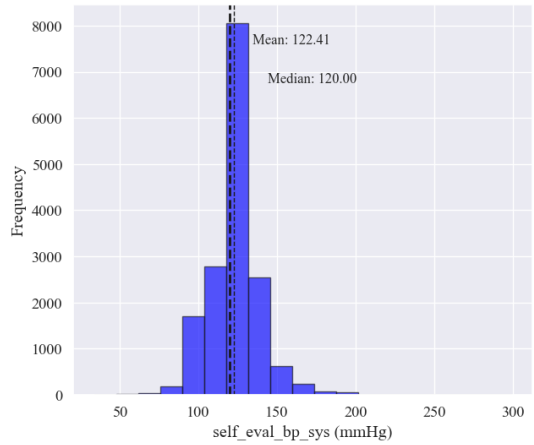
In the first instance, we inspect the distribution of the target variable *Systolic blood pressure* and *Diastolic blood pressure* in their two forms: measured and self-reported.



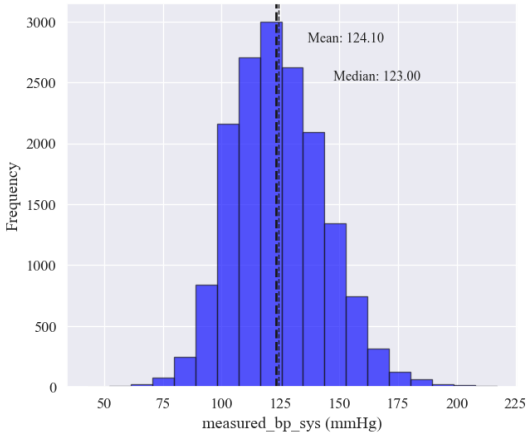
(a) Self reported systolic and diastolic bp.



(b) Measured systolic and diastolic bp.



(c) Histogram of self reported systolic bp.



(d) Histogram of measured systolic bp.

Figure 1: Histograms and scatter plots for the self reported and measured bp.

The scatter plot in Figure 1(a) compared to Figure 1(b) displays a substantial number of observations that deviate from the general trend, indicating the presence of a significant number of outliers. To confirm this hypothesis, the histogram in Figure 1(c) shows a non-normal distribution, which suggests that the data may not conform to a Gaussian distribution. In contrast to the histogram in Figure 1(d) which approximates to a normal distribution.

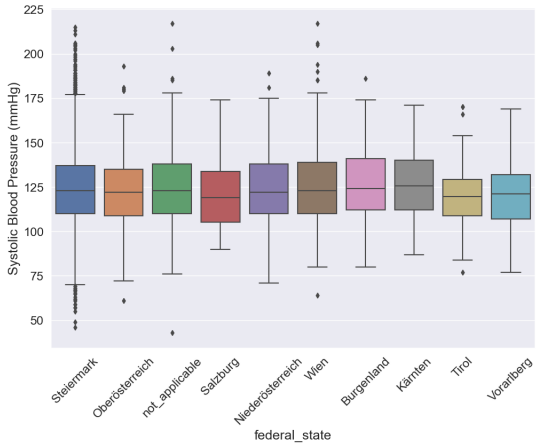
These findings may indicate the presence of influential observations or underlying data generating processes that are not well captured by the model assumptions. Which may indicate that the self-reported blood pressure is not a reliable measure of the true blood pressure. Hence, we decide to not further consider the variables *Self-reported systolic* and *Self-reported diastolic* blood pressures any longer in our analysis.

Figure 2 (the following page) shows a series of boxplots for systolic and diastolic blood pressures w.r.t different variables. Figure 2(a) and Figure 2(b) show how the systolic and diastolic blood pressures vary with the *Federal state* variable. The main interest angle for this comparison is to see if the prevalence of hypertension (high blood pressure) can vary by geographic region, thus leading to different blood pressure levels. Additionally, factors such as diet, lifestyle, and genetic predisposition could also contribute to regional differences in blood pressure. We observe that the overall median for diastolic blood pressure is around 80 mmHg, which is considered normal. Similarly, the median for systolic blood pressure is around 125 mmHg, which is also in the norm. We discerne a homogeneous distribution of the blood pressure levels across the different federal states, except in the case of the Steiermark state, where there is a higher number of observations that are falling outside the interquarterile range for both blood pressure levels. This might indicate that the prevalence of hypertension is higher in this region.

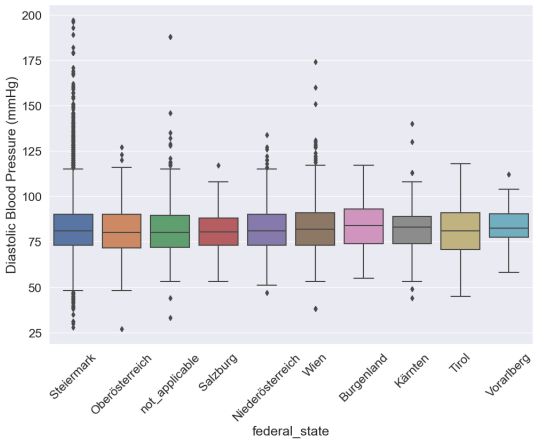
In Figure 2(c) and Figure 2(d) we want to see if there is enough evidence to suggest that the blood pressure may vary seasonally. Recent studies have shown that blood pressure levels tend to be higher in the winter and lower in the summer (Flint et al., 2019). Figure 2(c) indicates the presence of a seasonal pattern for the systolic blood pressure, where the median is lower in the summer months (June, July, August) w.r.t October and November. Even though we only have 26 observations in November.

Blood pressure can also vary depending on the day of the week. It tends to be higher on weekdays than on weekends. Seemingly, the stress of the work week can contribute to higher blood pressure levels (Juhanoja et al., 2016). Figure 2(d) tries to capture this effect. We observe that the median for systolic blood pressure is slightly higher on

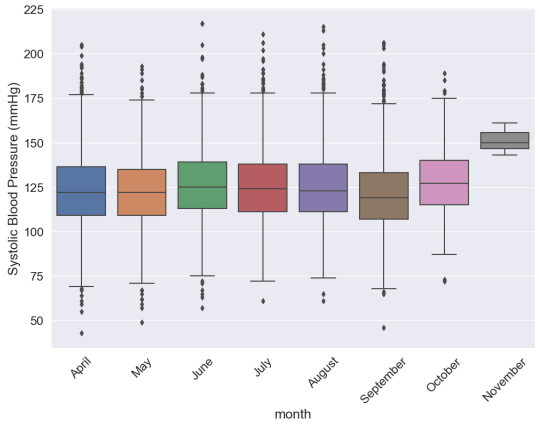
weekdays. However, the difference is not significant enough to conclude that there is a difference in blood pressure levels.



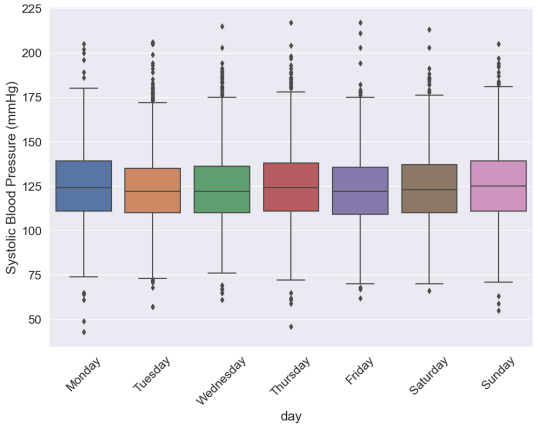
(a) Systolic bp by federal state.



(b) Diastolic bp by federal state.



(c) Systolic bp by month.

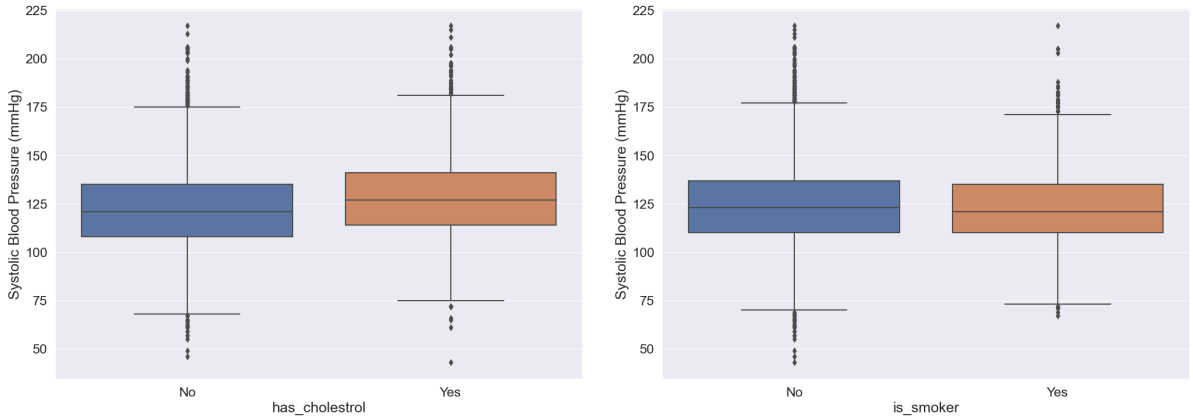


(d) Systolic bp by day of the week.

Figure 2: Boxplots for systolic and diastolic blood pressure w.r.t Autrian federal state, month, and day of the week.

Furthermore, we assessed how the systolic blood pressure varies with the *Cholesterol level* and *Smoking status* variables. Figure 3(a) shows that the median for systolic blood pressure is higher for individuals with high cholesterol level. This suggests that subjects with high cholesterol may have higher systolic blood pressure levels on average compared to those with normal cholesterol. This is consistent with previous research that has shown a positive association between high cholesterol and elevated blood pressure levels (Sakurai et al., 2011).

Conversely, Figure 3(b) shows that there is not enough evidence to conclude that smoking status has an effect on systolic blood pressure levels (at least for the current sample size).



(a) Systolic bp by cholesterol level.

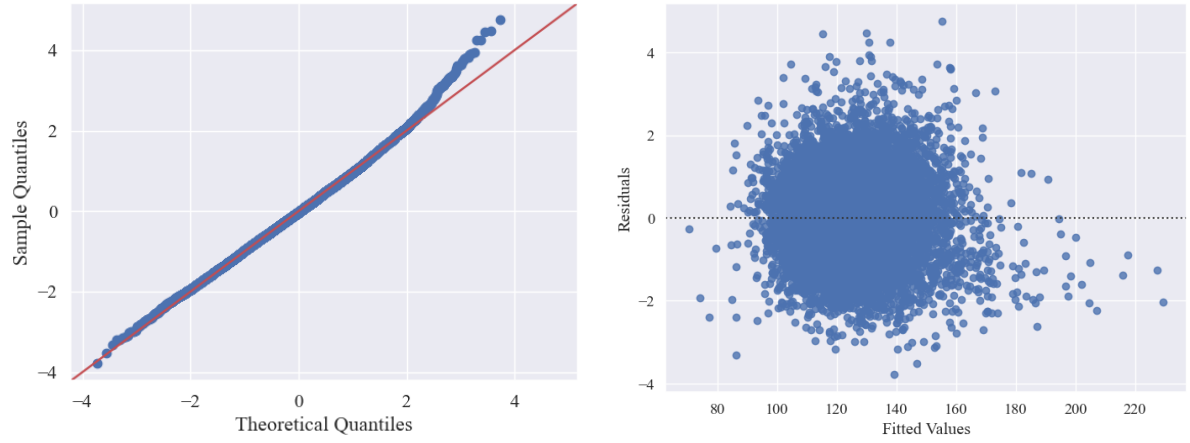
(b) Systolic bp by smoking status.

Figure 3: Boxplots for systolic blood pressure w.r.t cholesterol level and smoking status.

4.2 Linear regression

In this section, we fit two linear regression models to the data, taking the *Measured Systolic blood pressure* and *Measured Diastolic blood pressure* as the response variables.

We initially assess the normality assumption for the response variables using a QQ plot (c.f. section 3.4) shown in Figure 4(a).



(a) Normal qq plot for the systolic bp.

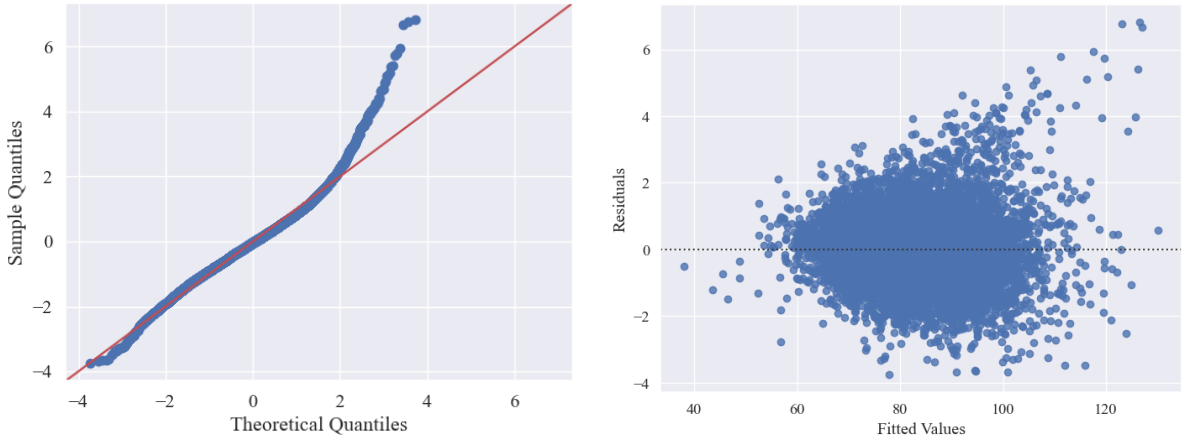
(b) Residual plot for the systolic bp.

Figure 4: Normal qq plot (left) and residual plot (right) for the systolic blood pressure model.

We observe in Figure 4(a) that the data points align well with the theoretical quantiles, except for the upper tail. This suggests that the normality assumption is reasonable for the current sample size.

Figure 4(b) displays the residual plot (c.f. section 3.5) for the systolic blood pressure model. We observe that the residuals are evenly distributed around zero, which leads to conclude that the homoscedasticity assumption is also reasonable.

Conversely, we notice in Figure 5(a) that the data points do not align well with the theoretical quantiles. This suggests that the normality assumption is somewhat alterated for the diastolic blood pressure model. The residual plot in Figure 5(b) similarly indicates the presence of non-homoscedasticity. The residuals exhibit a fan shape, displaying unequal spread across the range of predicted values. This means that the variability of the response variable differs across the levels of the predictor variables.



(a) Normal qq plot for the systolic bp.

(b) Residual plot for the systolic bp.

Figure 5: Normal qq plot (left) and residual plot (right) for the systolic blood pressure model.

Considering that the deviance from normality and homoscedasticity is not severe for both response variables, we proceed to fit a linear regression model to each blood pressure without transforming it. The training set contains 10404 observations against 4460 observations for the testing set.

We employ the linear regression model using the *Diastolic blood pressure* as the response variable to demonstrate the interpretation of the coefficients. We only include the significant variables (c.f. section 3.1).

Table 7 (c.f. Appendix) shows the results of the linear model fitted to the diastolic blood pressure. The intercept coefficient of 18.7946 indicates the estimated mean diastolic blood pressure when all independent variables are equal to zero. The rest of the coefficients are interpreted by an increase or decrease in the diastolic blood pressure while holding the other features constant.

For the categorical variable *Terminal*, terminal 2 has a positive coefficient of 1.94, indicating that individuals in terminal 2 have a diastolic blood pressure that is, on average, 1.94 mmHg higher than those in terminal 1 (reference level). Similarly, terminal 3a has a negative coefficients of -1.248 indicating that measured individuals in terminal 3a have on average a lower diastolic blood pressure that those in terminal 1.

Felt health condition 2 has a negative coefficient of -0.72, showing that individuals who report feeling very good have a diastolic blood pressure that is, on average, 0.72 mmHg lower than those who report feeling excellent. Which shows that the diastolic blood pressure increases as the felt health condition decreases. The gender has a positive relationship with the diastolic blood pressure and males have on average a higher diastolic blood pressure than females.

Individuals who report following a treatment have a lower diastolic blood pressure on average that those not in treatment. The *Systolic blood pressure* is also employed as a predictor for the diastolic blood pressure, and its coefficient is positive indicating that higher measured systolic blood pressure is associated with higher diastolic blood pressure. One unit increase in the systolic blood pressure is associated with a 0.61 mmHg increase in the diastolic blood pressure.

The results of the model also indicate that the diastolic blood pressure decreases as the age increases. On average a one year older individual would lower his diastolic blood pressure by 0.09 mmHg. The hour of the day also plays a role, the model suggests that on average the systolic blood pressure tends to be lower during later hours of the day.

The weather conditions have also an impact on the diastolic blood pressure. Specifically, higher temperature and higher humidity are associated with higher blood pressure. One unit Celsius increment in the temperature is linked to 0.6 mmHg increase in the diastolic blood pressure.

The linear model for the diastolic blood pressure has an R^2 value of 0.473. indicating that approximately 47.3% of the variability in the response can be explained by the predictors. The adjusted R^2 value for the diastolic blood pressure model is 0.471. The linear model for the systolic blood pressure performs better with an R^2 value of 0.542.

The *Measured systolic blood pressure* is the most significant predictor in the model, accounting for approximately 92% of the total variability in the diastolic blood pressure. Which explains why we decided to include it in the model despite the high correlation with the response variable.

4.3 Ensemble Methods

Regression Tree and Random Forest

The set of hyperparameters (c.f. section 4.3) choosen for the regression tree and the random forest model for the systolic blood pressure are the following:

- Max depth: 10.
- Max number of features: 40.
- Minimum sample of leaves per node: 70.
- Splitter: Best algorithm.

We use the same training/testing split as in the linear regression model. We present the results of the regression tree models on the diastolic blood pressure in Table 4.

Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2
Regression Tree (Base)	0.490	217.600	1.000	-0.100
Random Forest (Base)	15.950	114.530	0.920	0.420

Table 4: Results of the base regression tree and random forest models for the systolic blood pressure.

From Table 4 we conclude the following resutls: The average squared difference between the predicted and actual diastolic blood pressure values on the training data is 0.49. This indicates a relatively low level of error in the model's predictions for the training dataset. In comparison the mean squred error on the training dataset for the random forest model is 15.95. Which indicates that the random forest model exhibits a higher level of error in its predictions for the training dataset.

For the test set, the mean squared error for the Random forest is lower than the regression tree model. Which shows that the random forest model performs better than the regression tree model for the test dataset, and therefore, the random forest model is less prone to overfitting and generalizes better.

The R^2 for both models in the training set is very high indicating a high level of overfitting. Nontheless, the random forest model outperforms the regression tree model in the test set, explaining 42% of the variability in the diastolic blood pressure. This number is similar to the previous linear regression model.

4.4 Evaluation of the goodness of fit

Best subset of explanatory variables

Using the best subset selection algorithm explained in section 3.2, we iteratively fit multiple linear models to the data and select the best set based on the MSE criterion.

Table 5 shows the results of the best subset selection algorithm for the diastolic blood pressure. We observe that the best subset selection algorithm does not improve the performance of the base linear model, the adjusted R^2 is approximately the same for both models.

Model	Train Mean Sq Error	Test Mean Sq Error	Train Adjusted R2	Test Adjusted R2
LM (Base)	109.906	108.071	0.455	0.451
LM (Best Subset)	108.053	110.150	0.453	0.454

Table 5: Results of the best subset selection algorithm for the diastolic blood pressure.

Fine tuning with cross-validation

For the regression tree models we performed an additional step to improve on the base model. We used the cross-validation (explained in section 3.9) method to fine tune the hyperparameters of the regression tree model and the random forest model.

Table 6 shows the results of the fine tunned regression tree and random forest models for the diastolic blood pressure. For both models the fine tuning step improves the performance of the base model on the test set. Leading to an Adjusted R^2 of 0.460 for the regression tree model and 0.470 for the random forest model.

Model	Train Mean Sq Error	Test Mean Sq Error	Train Adjusted R2	Test Adjusted R2
Reg Tree (Base)	0.490	217.600	1.000	-0.110
Reg Tree (Fine tunned)	100.350	105.500	0.500	0.460
R Forest (Base)	15.950	114.530	0.920	0.420
R Forest (Fine tunned)	97.310	103.100	0.520	0.470

Table 6: Results for the fine tunned regression tree and random forest models for the diastolic blood pressure.

To draw a comparison between the linear regression and the regression tree models we can observe that both models exhibit similar performance in terms of explaining the variability in the response variable. The adjusted R^2 values for both models are very close, suggesting that they capture a similar proportion of the total variation in the data.

The linear model offers straightforward interpretability as it provides the coefficients that represent the relationship between the predictors and the response variable. Conversely, the interpretation of a regression tree is more intuitive (no parameters to interpret) as it involves hierarchical splits based on predictor values.

The ARIMA model is a time series model that is used to predict future values based on past values. The mathematical formulation of the model is as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{4.1}$$

5 Summary

In this case study we focused on a dataset that contains information about blood pressure measurements taken from visitors to the *Wege zur Gesundheit* exhibition held in the city of Bruck an der Mur in 2006. The event aimed to promote healthy living and featured various health-related activities.

The dataset comprises 16,386 measurements of systolic and diastolic blood pressure, as well as 18 other parameters that describe the geographic and physiological characteristics of the visitors. The aim is to model the systolic and diastolic blood pressure based on the other parameters.

We initially started by a pre-processing phase, where we performed feature engineering on some of the categorical variables in our dataset. Initially, we created a new variable

called *Age* by calculating the difference between the date of birth and the date of blood pressure measurements. We also extracted the variables *Year*, *Month* and *Day* from the date of blood pressure measurements. Furthermore, we split the *Terminal* 3 variable into two subgroups, "3a" and "3b," after the measurement device was changed. Additionally, we incorporated meteorological data, such as temperature, humidity, and weather conditions, in our analysis. Specifically, we incorporated the meteorological readings taken on the day of the blood pressure measurements. Lastly, we removed all inconsistent observations from the dataset, and dummy encoded the categorical variables, amounting to a total of 26 usable features for our analysis.

We thereafter, proceeded to explore the data applying descriptive analysis methods to grasp indepth insight into it. We studied the distribution of the systolic and diastolic blood pressures in their two forms: measured and self-reported. We concluded that the distribution of the self-reported blood pressure suggests the possibility of influential observations or data generating processes not well captured by the model assumptions. Thus, indicating that the measurements may not be reliable. We therefore, decided to focus our analysis on the measured blood pressure.

We also examined the potential correlation between blood pressure and various factors, such as federal state, in order to determine if there was a variation in hypertension prevalence across different geographic regions. The analysis revealed generally uniform distribution of blood pressure levels across all states. We also investigated the relationship between blood pressure and seasonality, and found evidence to suggest that blood pressure levels may vary seasonally. We further explored possible correlations between health-related factors and blood pressure levels. We found that there is a positive association between high cholesterol and elevated blood pressure levels.

In the course of the analysis, we assessed if the normality assumptions of the linear model were met by the response variables: *Measured Diastolic blood pressure* and *Measured Systolic blood pressure*. And then proceeded to fit two linear models for each target blood pressure. Both linear models perfomed in a moderately effective way, with the Systolic model performing slightly better than the Diastolic model, averaging an R^2 value of 0.57 against 0.47 for the diastolic model. We also found out that for the diastolic model, 92% of the variance is explained by the systolic blood pressure as a predictor. This is not surprising as the systolic blood pressure is a strong predictor of the diastolic blood pressure.

Alongside the linear models, we also fitted a regression tree and random forest model for each target blood pressure. We also used fine-tuning algortihms witht the MSE as a criterion to improve on the current state of our models. We used best subset selection to select the best predictors for the linear models, and used cross-validation to fine-tune the regression tree and random forest models. The results showed minimal differences, with only marginal improvements observed. The random forest model performed slightly better than all other models with an adjusted R^2 on the testing set of 0.47 for the diastolic blood pressure and 0.54 for the systolic blood pressure.

Finally, the significant model's parameters were interpreted and the results were discussed.

In this case study we assumed that the relationship between the response variable and the regressors is linear. Even though, in most realistic cases the linear model is no longer relevant and for further studies it would be interesting to explore advanced regression models like the Robust regression or even the Generalized Additive Models (GAMs) that can capture non-linear relationships.

Bibliography

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression Models, Methods and Applications*. Berlin ; New York : Springer, first edition, 2013.

Alexander C Flint, Craig Conell, Xiaoxi Ren, Nicole M Banki, Samuel L Chan, Vivek A Rao, Ronald B Melles, and Deepak L Bhatt. Effect of systolic and diastolic blood pressure on cardiovascular outcomes. *The New England journal of medicine*, 381(3): 243–251, 2019. doi: 10.1056/NEJMoa1803180.

Badis Haddag, Julien Thil, Mohammed Nouari, and Claude Barlier. A study of the bta deep drilling process through a quantitative and qualitative analysis of the chip formation process. *Key engineering materials*, 554-557:1992–2008, 2013. doi: 10.4028/www.scientific.net/KEM.554-557.1992.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Chapman and Hall/CRC, 1st edition, 2019.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, second edition, 2013.

EP Juhanoja, PJ Puukka, JK Johansson, TJ Niiranen, and AM Jula. The impact of the day of the week on home blood pressure: the finn-home study. *Blood Pressure Monitoring*, 21(2):63–68, 2016. doi: 10.1097/MBP.0000000000000156.

Oliphant, Travis E. and The NumPy Community. NumPy: A guide to Numerical Python, 2022. URL `\url{https://numpy.org/}`. [Online; accessed 5 Jun. 2023].

Pandas Development Team. pandas: powerful python data analysis toolkit, 2021. URL `\url{https://pandas.pydata.org/}`. [Online; accessed 5 Jun. 2023].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in Python, 2021. URL `\url{https://scikit-learn.org/stable/}`. [Online; accessed 5 Jun. 2023].

Masaru Sakurai, Jeremiah Stamler, Katsuyuki Miura, Ian J Brown, Hiroyuki Nakagawa, Paul Elliott, Hirotsugu Ueshima, Queenie Chan, Ioanna Tzoulaki, Alan R Dyer, Akira Okayama, and Liancheng Zhao. Relationship of dietary cholesterol to blood pressure:

the intermap study. *Journal of Hypertension*, 29(2):222–228, 2011. doi: 10.1097/HJH.0b013e32834069a5.

Statmodels API. Statsmodels: Econometric and statistical modeling with Python, 2021. URL \url{https://www.statsmodels.org/stable/index.html}. [Online; accessed 25. Jan. 2022].

Waskom, Michael and Seaborn Development Team. Seaborn: Statistical data visualization, 2021. URL \url{https://seaborn.pydata.org/}. [Online; accessed 5 Jun. 2023].

Oliver Webber. Investigations on drilling depth-dependent process dynamics in bta deep drilling, 2006. URL https://d-nb.info/983001065. Zugl.: Dortmund, Univ., Diss., 2006.

K. Weinert, O. Webber, A. Gepperth, Y. Zhang, and W. Theis. Time varying dynamics in bta deep hole drilling. In R. Teti, editor, *Proceedings of the 4th CIRP International Seminar on Intelligent Computation in Manufacturing Engineering*, pages 419–424, 2004.

Appendix

Additional Tables

Table 7 (following page) shows the list of regressors and their coefficients for the linear model fitted to the diastolic blood pressure.

Table 7: Summary of the Linear regression model fitted to the Diastolic blood pressure.

Variable name	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	18.7946	2.392	7.856	0.000	14.105	23.484
terminal_2	1.9449	0.247	7.876	0.000	1.461	2.429
terminal_3a	-1.2479	0.419	-2.981	0.003	-2.068	-0.427
terminal_3b	-0.6710	0.290	-2.316	0.021	-1.239	-0.103
federal_state_Kärnten	-2.2591	1.471	-1.536	0.125	-5.143	0.625
federal_state_Niederösterreich	-1.6394	1.302	-1.259	0.208	-4.192	0.913
federal_state_Oberösterreich	-1.7391	1.444	-1.204	0.229	-4.571	1.092
federal_state_Salzburg	-0.4074	1.777	-0.229	0.819	-3.890	3.075
federal_state_Steiermark	-1.4835	1.160	-1.279	0.201	-3.757	0.789
federal_state_Tirol	-1.6314	1.891	-0.863	0.388	-5.337	2.074
federal_state_Vorarlberg	0.5143	2.923	0.176	0.860	-5.215	6.243
federal_state_Wien	-0.9295	1.292	-0.720	0.472	-3.462	1.603
federal_state_not_applicable	-1.8763	1.394	-1.346	0.178	-4.609	0.856
felt_health_condition_2	-0.7195	0.229	-3.140	0.002	-1.169	-0.270
felt_health_condition_3	-0.2442	0.327	-0.746	0.456	-0.886	0.398
felt_health_condition_4	-0.3979	0.969	-0.411	0.681	-2.296	1.501
felt_health_condition_5	3.7336	1.613	2.314	0.021	0.571	6.896
gender_m	1.2023	0.208	5.779	0.000	0.795	1.610
is_smoker_True	0.0357	0.282	0.126	0.899	-0.518	0.589
is_diabetic_True	-0.2325	0.291	-0.800	0.424	-0.802	0.337
has_cholestrol_True	0.1427	0.270	0.529	0.597	-0.386	0.671
in_treatment_True	-1.6391	0.329	-4.985	0.000	-2.284	-0.995
month_Aug	0.7421	0.786	0.944	0.345	-0.799	2.283
month_Jul	-0.5602	0.898	-0.624	0.533	-2.321	1.200
month_Jun	0.4529	0.825	0.549	0.583	-1.165	2.071
month_May	0.5684	0.778	0.731	0.465	-0.956	2.092
month_Nov	21.4579	10.456	2.052	0.040	0.962	41.954
month_Oct	-3.2398	0.782	-4.142	0.000	-4.773	-1.707
month_Sep	-2.7086	0.791	-3.425	0.001	-4.259	-1.158
day_Monday	0.1772	0.426	0.416	0.677	-0.658	1.012
day_Saturday	0.4438	0.383	1.157	0.247	-0.308	1.195
day_Sunday	0.2094	0.362	0.578	0.563	-0.500	0.919
day_Thursday	-0.9265	0.401	-2.312	0.021	-1.712	-0.141
day_Tuesday	-1.1554	0.435	-2.659	0.008	-2.007	-0.304
day_Wednesday	-1.0129	0.418	-2.421	0.015	-1.833	-0.193
measured_bp_sys	0.5290	0.006	90.745	0.000	0.518	0.540
age	-0.0892	0.007	-12.168	0.000	-0.104	-0.075
hour	-0.0898	0.046	-1.952	0.051	-0.180	0.000
temp	0.6937	0.155	4.466	0.000	0.389	0.998
humidity	0.0669	0.016	4.136	0.000	0.035	0.099
temp_min	-0.3059	0.101	-3.031	0.002	-0.504	-0.108
temp_max	-0.3587	0.077	-4.634	0.000	-0.510	-0.207