10-01-2021

Data Analysis Report Iteration 2

Pharmacy deliveries

# 1. Introduction

This document is made after a business proposal for a client Informa.

We were given a dataset of 22 million instances of their medical data and were told to derive value out of it, by using different applied-data-science techniques.

This is the second iteration of our data analysis. So far, we have made a business proposal regarding our understanding of the objective and the data.

This report contains research about the data which we are using, we have added data and preprocessed data in the previous steps already but will also explain what we did in this step to enhance the quality of the data. We will also explain the methods we applied and our gathered results. The above will be supported by proper visualizations. Also, in the end, we will give advice regarding this data analysis.

The main objective of the data analysis is to try to predict the ordered units of a specific medication type for pharmacies in Belgium. We use these features because in the exploratory data analysis, which you can find in the project proposal (see appendix), it became clear that some medication types have a strong relationship with age, the example we showed in the EDA were antidepressants (Psycho-analeptics), antidepressants seemed to be ordered almost only from people between the age of 55 and 75. Of course this does not go for all medication types, take for example vitamins. That is why we have to make sure that the model works for the medication type before using it in a real-life situation. In the previous iteration we have looked at the monthly totals of whole Belgium, in this iteration we are looking at the data for the provinces of Belgium.

In the next chapter, we will go through the data that we are working with to come up with a possible solution.

Table of Contents

[1. Introduction 1](#_Toc61208374)

[2. The Data 3](#_Toc61208375)

[2.1 Pharmacies 3](#_Toc61208376)

[2.2 Population 4](#_Toc61208377)

[3. Features and Preprocessing 5](#_Toc61208378)

[4. Method and Approach 7](#_Toc61208379)

[4.1 Multiple Linear Regression (MLR) 7](#_Toc61208380)

[4.2 Decision Tree Regression 7](#_Toc61208381)

[5. Results and Discussion 8](#_Toc61208382)

[5.1 Multiple Linear Regression (MLR) 8](#_Toc61208383)

[5.2 Decision Tree Regression 9](#_Toc61208384)

[5.3 Random Forest Regression 10](#_Toc61208385)

[5.4 Gradient Boosting Regression 11](#_Toc61208386)

[5.5 Evaluation 12](#_Toc61208387)

[5.5.1 Cross Validation 12](#_Toc61208388)

[5.5.2 Using other medicine types 12](#_Toc61208389)

[6. Ethical considerations 13](#_Toc61208390)

[7. Conclusion 14](#_Toc61208391)

[8. Appendix 15](#_Toc61208392)

[8.1 Business proposal 15](#_Toc61208393)

[8.2 Population dataset 15](#_Toc61208394)

[8.3 Notebook 15](#_Toc61208395)

# 2. The Data

The datasets used for this project consists of the following:

* Pharmacies dataset from Informa, enriched with external resources.
* Data on the population of Belgium by age and province

## 2.1 Pharmacies

The pharmacies dataset contained the following data:

* Delivery date, this is the date when the product was delivered to the customer.
* Delivery time, this is the time when the product was delivered to the customer.
* Pharmacy number, this is the internal system number from the pharmacy. This number can only be used to group data from the same pharmacy.
* Pharmacy Postcode (2) is a part of the postcode from the pharmacy. It contains the first 2 numbers from the postcode.
* Year of birth contains the year of birth of the customer
* Gender contains the gender of the customer (1=male,2=female)
* CNK, the CNK is the unique product code that is standardized within Belgium.
* The product name is the name of the product in the Dutch language.
* ATC code, this is the ATC code which is an international standardized. Every medicine has a unique code. This code is built out of several portions.
* Units, number of product units in a package
* Price, the price of the delivery
* Contribution, the contribution of the customer for this delivery Using the previous data and some external sources the following data was added:
* Age, based on the delivery date and year of birth of the customer
* Province, based on the partial pharmacy postcode
* ATC Classification, based on the first letter of the ATC code
* Medication Type, based on the first letter and first numbers of the ATC code
* Province, based on the first 2 digits of the pharmacy postcode the full list of ATC classification and medication types can be found here

## 2.2 Population

We have selected data from Statbel, the Belgium statistics bureau. They have a collection of data about Belgium, including the population.

Link to the website: <https://statbel.fgov.be/nl>

The dataset we used can also be found in the appendix.

The Population dataset contained the following data:

* Age Group, age of population in Belgium grouped in 5-year interval group
* Province, the province of the measured population
* Total Population, contains the total population per age group per year and divided based on their gender in Belgium
* Year, the year the census being held

After explaining the dataset that is being used in this project, next we are going to explain on how we are going to use the features and what features that we use for our prediction model complete with its justification.

# 3. Features and Preprocessing

In this chapter, we are going to see the features that we use to make this model and why we use these features.

The features that are used to predict the units are age group and the population of the age group at the time of ordering. These features were selected after doing an exploratory data analysis which you can find in the appendix. We selected these features because some medicine types have a strong relationship with age.

The model will predict the units per medicine type of a specific province, so first we had to filter on a medicine type. We wanted to try out a model that works with the monthly totals, so we grouped the data by month and summed the units (see figure below).

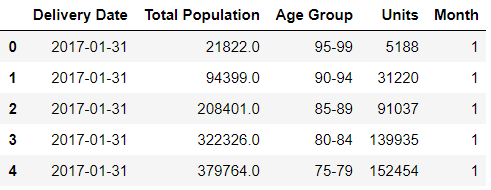


Figure 1 Features used in this process

After that we also had to filter on the province which we wanted to use to try out our models

For demonstration purposes, we use the province of Antwerp as a filter since it is the most populated region on Belgium and this way, the data can be more represented. By having this province filter, we also could see how every region is doing in terms of their sales and could see in terms of the whole population of the region on how much people is buying a particular medicine type. **Add more elaboration?**

Table

Description automatically generated

Figure 2 Updated features used in this process

To use the age group in our model we had to convert it to a number instead of a string. Age 0-4 became 1, 5-9 became 2, and so on. After doing that we also normalized the data (see figure below).

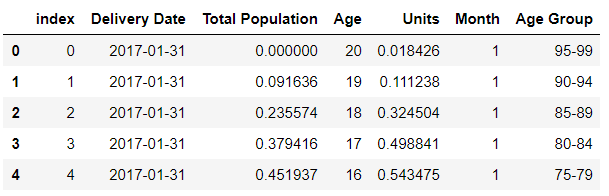


Figure 3 Features used in model after normalized units

Table

Description automatically generated

Figure 4 Updated features used in after being normalized

# 4. Method and Approach

In this chapter, we are going to explain the machine learning method. We tried out different types of regression algorithms in the previous iteration, the first one (MLR) gave the lowest scores and the other three all scored equally. In this iteration we have used the same methods, by using the same methods we can compare the scores with our previous iteration and see how much the provinces impact the models. The methods are listed down below:

## 4.1 Multiple Linear Regression (MLR)

MLR uses several explanatory variables to predict the outcome of a response variable, it is an extension of regular linear regression, which uses only one variable. This was the first algorithm we tried because it is pretty fast to set up and an easy way to compare to the previous iteration.

## 4.2 Decision Tree Regression

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. We tried this algorithm since we think that its multi-layer decision tree for making the final results makes it more accurate to predicting the output. Even though it’s not mimicking how the human brains work while making a decision, the prediction will be more accurate than a standard regression since it has several layers of decision-making points to make sure it has achieved the best results from it. The final prediction is the average of the value of the dependent variable.

4.3 Random Forest Regression

A Random Forest makes use of multiple decision trees and a technique called *bagging*, which involves training each decision tree on a different data sample where sampling is done with replacement. We used Random Forest because it suits the huge dataset we have – Random Forest has the ability to handle a large data set with higher dimensionality.

In the RF classifier, every decision tree forecasts a response for an occurrence and the endmost response is decided through voting. On contrary, in classification, the response received by majority voting of Decision Tree is the final response and in **regression**, the final response is the average of all the responses.

4.4 Gradient Boosting Regression

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

The name gradient boosting arises because target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error, in the space of possible predictions for each training case.

# 5. Results and Discussion

For comparing the different algorithms, we used the medication type “Psycho-analeptics” for all algorithms, this is because the correlation between age and units is different for every medication type and so the model’s accuracy will be different for all medication types. The province that we used was “Antwerp”.

## 5.1 Multiple Linear Regression (MLR)

The first algorithm we tried was MLR. This gave us the following results:

Text

Description automatically generated

This algorithm had an accuracy of about 82%. This is visualized in the graph below where you can see the difference between what the model predicted and the actual values.

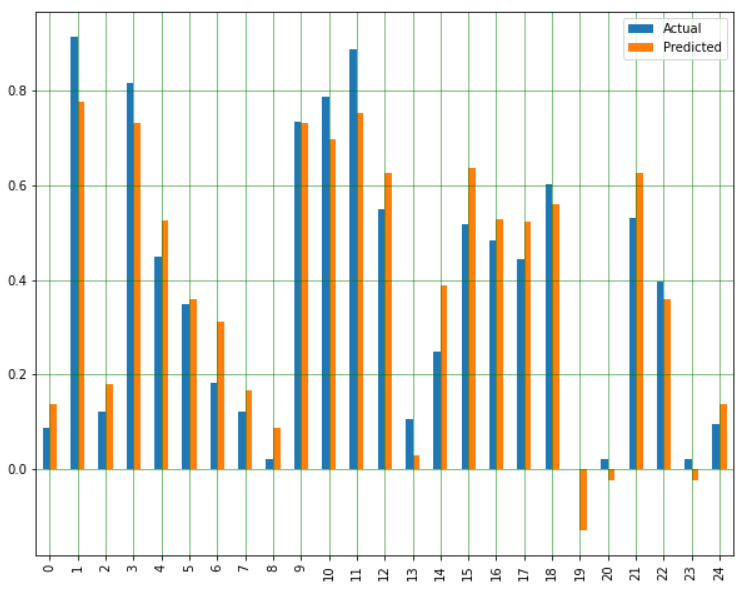


Figure 5 Multiple Linear Regression Results

Compared to the first iteration where we did not filter on province the accuracy is lower. In the first iteration we had an accuracy of about 90%. This means that it is harder to make more detailed predictions, but we will have to see what the other, better suited, algorithms show.

## 5.2 Decision Tree Regression

The second algorithm we tried was decision tree regression. This algorithm gave us the following results

Text

Description automatically generated

As you can see the accuracy is much higher than by using MLR and the errors are much smaller. So, this is a better algorithm to use in our case. The accuracy is visualized in the graph below.

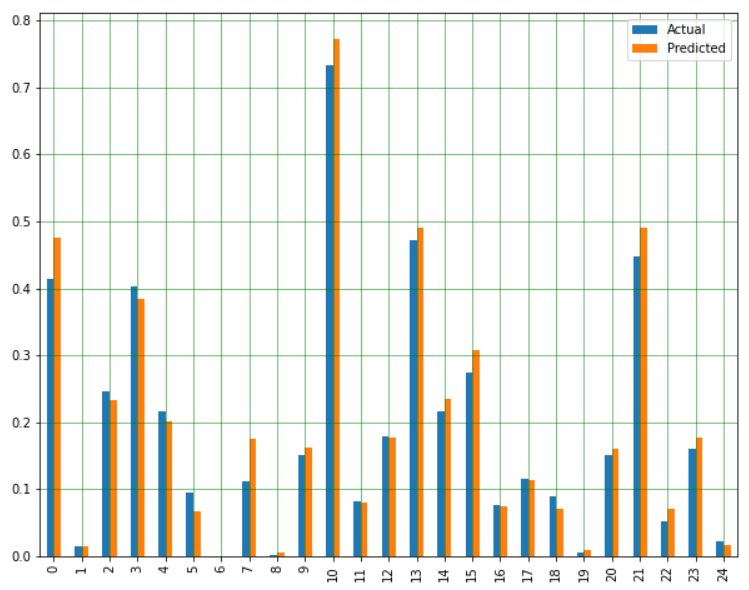


Figure 6 Decision Tree Regression Results

So, this algorithm performed much better than MLR. Now let us compare it to the other two algorithms.

## 5.3 Random Forest Regression

This is the third algorithm which we used – Random Forest Regressor. It gave us the following results:

Text, letter

Description automatically generated

Compared to the last time we used the random forest regressor where we did not filter by province and with the older dataset, the accuracy was 97%. So there is a 4-5% difference in the accuracy. The mean squared error has increased since last time. The other metrics about the error are similar.

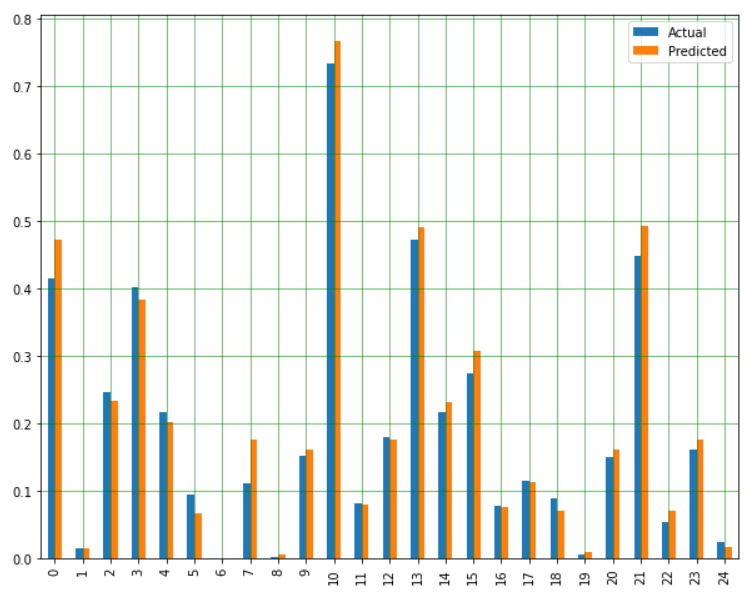


Figure 7 Random Forest Regression Results

## 5.4 Gradient Boosting Regression

This is the fourth algorithm which we used – Random Forest Regressor. It gave us the following results:

Text, letter

Description automatically generated

Last time we used the Gradient Boosting we scored 97%. So there’s a 5% difference in accuracy. There is a small increase in the mean squared error – from 0.0014 to 0.0029 and everything else is the same.

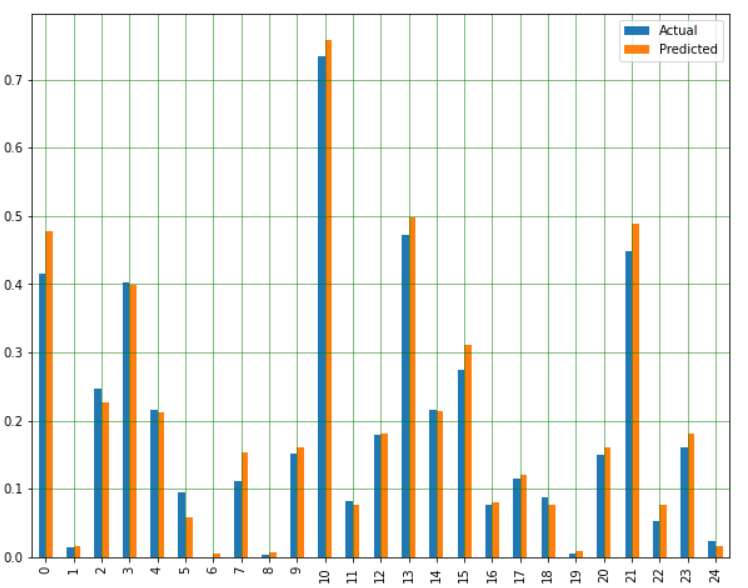


Figure 8 Gradient Boosting Regression Results

## 5.5 Evaluation

Just like in the previous iteration we have used the evaluation metrics below and explained what these metrics mean so the results can be better understood.

* **Mean Absolute Error** is the average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were. The measure gives an idea of the magnitude of the error, but no idea of the direction (e.g., over or under predicting).
* **Mean Squared Error** is much like the mean absolute error in that it provides a gross idea of the magnitude of error. Taking the square root of the mean squared error converts the units back to the original units of the output variable and can be meaningful for description and presentation. This is called the Root Mean Squared Error (or RMSE).
* **R2** metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature, this measure is called the coefficient of determination. This is a value between 0 and 1 for no-fit and perfect fit respectively.

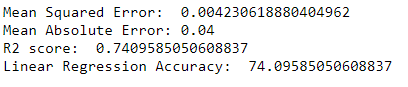
We have applied all of these metrics to verify our work on the machine learning model that we made. With these metrics, we came to a conclusion that most of the regression method deliver more or less the same results, especially with more complex regression algorithm such as Random Forest or Decision tree. For the specific case we tried the models with (Psycho-analeptics, and Antwerp) Decision Tree Regression came out on top with the lowest errors and the highest accuracy. Comparing the results from this iteration to the previous one we can see that the models became less accurate, about 4%, but in return they did get more detailed and probably more useful.

### 5.5.1 Cross Validation

To re-evaluate and validate our results, we did some cross-validation on each algorithm and find that the model is not over or under fitted since it results in a similar score between the initial algorithm with the one validated with cross validation. This result remain consistent throughout the algorithm except on the multiple linear regression, this means that the model is overfitted or it just not suitable for this use case.

### 5.5.2 Using other medicine types

The other thing that we validated was the use of different medicine types. The connection between age and the use of medicine types is the most important thing in this model. The Psycho-analeptics which we used to test the model was almost only used by people between the ages of 55 and 75, this is why the accuracy was so high. Of course this does not go for all medicine types, take for example Vitamins. Vitamins are used by people of all ages so this means the connection between age and the use of medicine isn’t that high. This is shown by the results below



As you can see the accuracy of the model when using vitamins is a bit worse than when using a medicine type like psycho-analeptics.

# 6. Ethical considerations

In current age of information that we live in, the data is continuously generated and used. Information is a very integral part of our lives now and with it comes the need to manage that information. We handled the data with integrity, it was not tampered in any way. We decided to explore the data without prior interpretations of the results that we might find, and same strategy was implemented while visualizing and training the model as well. The Dataset or the insights derived from It during the process were in no way shared outside the team.

We firmly believe that our findings and insights will solely be used to profit the organization/customers without harming anyone’s privacy in the process. The client is free to make changes to the code/models or combine more datasets in order to achieve better results, however the dataset added should not be manufactured to please a certain goal or an entity.

# 7. Conclusion

So, we can use the model we now have to predict the total monthly ordered units per province of a medicine type based on the population per province in Belgium. This works very good for some medicines, but it is not as accurate for all medicine types. This is because not all medicine types, for example vitamins and painkillers don’t have such a strong relationship with the age that they are used at, also the usage of medicines also can be affected by geographical location or environmental factors. So, the best thing to do is to look for the accuracy of all the medicine types and see which ones are usable and their consumptions does not get easily affected by the various factors.  
  
By adding the provinces, the accuracy of the model decreased with about 4%, but in return it now is more detailed and arguably more defined and useful. The accuracy, as predicted by our model can be used in a manner to benefit both client and consumer. By predicting the demand of a certain medicine in a specific province the warehouses can be stocked and hence avoid over supply or shortage of medicine. This ensures that the medicine can be evenly distributed across the country as per the requirements, and hence in return saving time, effort, increase monetizing benefit for the organization and also help consumers save time, and health in return.