29-11-2020

Data Analysis Report Iteration 1

Pharmacy deliveries

# 1. Introduction

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

So far, we have made a business proposal to understand the objective and the data and now we have also done a data analysis which is described in this document.

This report contains research about the data which we are using, we have added data and preprocessed data in the previous step (business proposal) 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 and what our next step will be.

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 (Psych 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 this iteration we will be looking at monthly totals of whole Belgium, in the future we could make this more specific.

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](#_Toc57576640)

[2.1 Pharmacies 3](#_Toc57576641)

[2.2 Population 4](#_Toc57576642)

[3. Features and Preprocessing 5](#_Toc57576643)

[4. Method and Approach 6](#_Toc57576644)

[4.1 Multiple Linear Regression (MLR) 6](#_Toc57576645)

[4.2 Decision Tree Regression 6](#_Toc57576646)

[5. Results and Discussion 7](#_Toc57576647)

[5.1 Multiple Linear Regression (MLR) 7](#_Toc57576648)

[5.2 Decision Tree Regression 9](#_Toc57576649)

[5.3 Random Forest Regression 10](#_Toc57576650)

[5.4 Gradient Boosting Regression 10](#_Toc57576651)

[5.5 Evaluation 11](#_Toc57576652)

[6. Ethical considerations 12](#_Toc57576653)

[7. Conclusion and Future 13](#_Toc57576654)

[8. Appendix 14](#_Toc57576655)

[8.1 Business proposal 14](#_Toc57576656)

[8.2 Multiple Linear Regression Notebook 14](#_Toc57576657)

[8.3 Decision Tree Regression Notebook 14](#_Toc57576658)

[8.4 Random Forest and Gradient Boosting Regression Notebook 14](#_Toc57576659)

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, gender and year

## 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

The Population dataset contained the following data:

* Age Group, age of population in Belgium grouped in 5-year interval group
* Male Population, contains the male population per age group per year in Belgium
* Female population, contains the female population per age group per year in Belgium
* Total Population, contains the total population per age group per year 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 type have a strong relationship with age.

The model will predict the units per medicine type, 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).

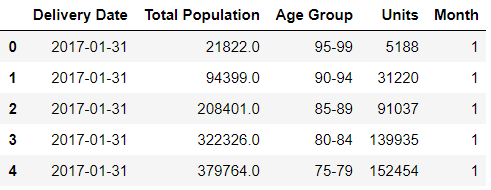


Figure 1 Features used in this process

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).

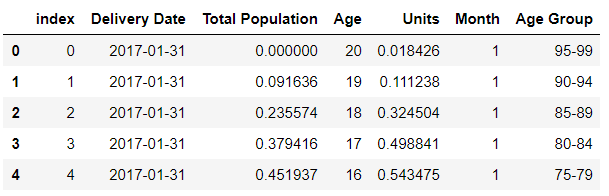


Figure 2 Features used in model after normalized units

# 4. Method and Approach

In this chapter, we are going to explain the machine learning method. We tried out different types of regression algorithms to see which method can give a better result in the end for our project. The method will be 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 different features.

## 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 mimicing 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 RandomForest because it suits the huge dataset we have – RandomForest 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.

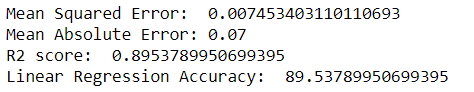
The name gradient boosting arises because target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error, in the space of possible predictions for each training case.

# 5. Results and Discussion

For comparing the different algorithms, we used the medication type “Psychoanaleptics” 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.

## 5.1 Multiple Linear Regression (MLR)

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



This algorithm had an accuracy of about 90% which seems pretty good. This is visualized in the graph below where you can see the difference between what the model predicted and the actual values.

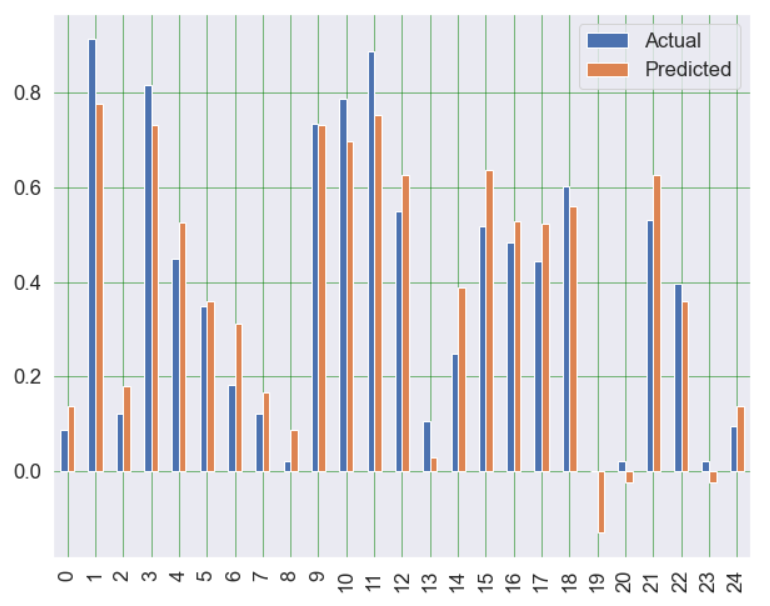


Figure 3 Multiple Linear Regression Results

We also tried adding the month value as a feature because the EDA told us that there was a pattern in the monthly ordered units. But because the pattern wasn’t 100% the same for each year and we only have data of 3 years this made the model less accurate (see accuracy below) so we decided to not use the month value as a feature.



As said before the model depends on the relationship between age and the amount of units ordered. This relationship is different for almost every medicine type. For example, in our exploratory data analysis we found that people only order psychoanaleptics from age 55 to around 75, this would not be the case for painkillers. Painkillers are used at almost every age and therefore the model would be less accurate for painkillers and possibly other medication types. To show this we changed the medication type for the MLR model to a different one.

So, to show the difference in the model for different medicine types we used vitamins, as vitamins are used at almost every age. This gave us an accuracy of about 65% which is considerably lower than the 90% from the psych analeptics.



## 5.2 Decision Tree Regression

Text, letter

Description automatically generatedWe also tried decision tree regression multiple times. The first time we got an accuracy of 77% with 2 layers of depth in this model, which is worse than the MLR. We tried it again with different depths, how many branches the tree contains. This gave us a very good accuracy of about 97% and in this model, we use 5 layers of depth which result in even better than the previous 90% from the MLR algorithm

Chart, bar chart

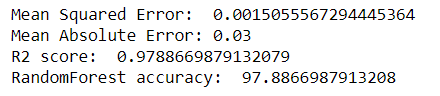
Description automatically generated

Figure 4 Decision Tree Regression Results

In order to get a grasp on why we use this algorithm, we decided to use the same features on all of our models which is the medicine type of ‘psych analeptics’ usage ranging from age 55-75. And as already mentioned above, this algorithm models predicted the units being bought more accurately than the usage of multiple linear regression.

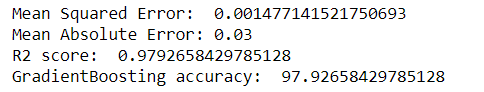
## 5.3 Random Forest Regression

We mainly used this Random Forest to see if there are any improvements for the evaluation metrics and model accuracy. As we can see, the mean absolute error is higher than that of the Decision Tree, but everything else is pretty similar.



## 5.4 Gradient Boosting Regression

We experiment with many different regression models to see if we can get a higher score. The Gradient Boosting Regression was the one with which we scored the highest score – **97.92%**



## 5.5 Evaluation

Following one of our discussions with the teachers we came to the conclusion to include metrics other than the accuracy for each model. After doing some research we found that there are 3 evaluation metrics which are the best for our regression model:

* **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 these metrics in our models including with the accuracy of the model. With these metrics we concluded that three algorithms worked well these are Decision Tree Regression, Random Forest Regression and Gradient Boosting Regression. All of them had pretty similar results so it does not really matter which one we use.

# 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.

# 7. Conclusion and Future

So, we can use the model we now have to predict the total monthly ordered units of a medicine type based on the population 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. So, the best thing to do is to look for the accuracy of all the medicine types and see which ones are usable.

For the next iteration we would like to make the model a bit more detailed by specifying the region and gender. By adding the gender, the model’s overall accuracy could improve, and by specifying the region it would be a lot more useful for our client because we found that the medicine use in different provinces can differ.