CASE STUDY IN DECISION AIDING AND ARTIFICIAL INTELLIGENCE

Predicting the consumption of healthcare products: A hands-on experience with Supervised Learning

"The goal of Machine Learning is never to make "perfect" guesses because Machine Learning deals in domains where there is no such thing. The goal is to make guesses that are good enough to be useful."

(George E.P.Box, British mathematician and professor of statistics)

Doga Yilmaz
Maria Fernanda Padilla
Tom Nikolas Schmidt

Supervisor: Zakaria Yahouni, Denis Koala

21.10.2023

GRENOBLE INP UGA Génie industriel

Table of Contents

Introduction 3

Case Study 3

Methodology 3

Data preparation 3

Data Engineering 4

Categorical Input Variables: 4

Numerical input variables 5

Data normalization 6

Data splitting 6

Results 6

First approach results: 7

Second approach results 7

Conclusion 8

Introduction

Machine learning is a transformative field of artificial intelligence enabling computers to make predictions or decisions without explicit programming. It's widely applied across finance, healthcare, marketing, and more.

Supervised learning, a fundamental technique, uses labeled training data to make predictions or categorize data. The algorithm learns from datasets containing input features (predictors) and corresponding output labels (target variables). Its goal is to create a model that accurately predicts target variables for new data. Supervised learning models come in various forms, like linear regression for continuous value prediction and classification methods, such as logistic regression or decision trees, for categorical outcome prediction.

Case Study

Pharma Gi has provided us with data which includes the consumption of medicines in different time periods. In addition, the data contains information about the characteristics of the hospital that ordered the medication, the environment in which the hospital is located, and much more. Our goal is to provide PharmaGI with a model to forecast the monthly demand of a specific drug for next periods, based on the relationship between the features and the labels. A machine learning process consists of various steps that are carried out one after the other: Data Acquisition, Data Cleaning or Data Engineering, Model Training, Model Testing and Deployment. These steps were followed on the way to finding a suitable prediction formula and are presented below.

Methodology

The approach of this study case was divided as follows:

1. **Hospital-Specific Medicine Prediction Model:** This model will predict the medicine usage for each individual hospital.
2. **Medicine-Specific Hospital Prediction Models:** These models focuses on predicting the usage of each type of medicine for every hospital.

Fitting the data to Linear Regression, XGB, and Random Forest models; selecting the one that yields superior performance indicators.

Data preparation

In the real world, data is often unorganized, redundant, or contains missing elements. Therefore, the first step is to understand, clean, prepare, and manipulate the data to ensure the development of a more accurate model. This process not only helps in avoiding overfitting but also in identifying features with low explanatory power. Additionally, examining the correlation of individual features can make the model easier to understand.

During the data cleaning process, we made several decisions to prepare the dataset for modeling. First, we discarded the 'Unnamed: 0' column from the DataFrame since it served as a numeric sequence index and did not provide meaningful information. Additionally, considering that machine learning models often struggle with missing data, and the number of NaN values in proportion to the entire dataset was small, we chose to remove rows with missing values. The task's target variable is defined in terms of monthly consumption, which led us to prioritize other features over 'WEEK' and 'DATE_MOUV.' As a result, we excluded these columns from the model generation. Upon gaining a better understanding of the features, we observed that the variable 'ID_REF' is a combination of 'ID_SITE_RATTACHE' and 'HOSPI_CODE_UCD.' This combination can result in the same 'ID_REF' for different medicines across different hospitals. Consequently, we decided not to use the 'ID_REF' feature in our model.

The next step involved modifying the data types of certain features to objects, integers, and datetimes. The objective was to standardize the data to ensure consistency and suitability for machine learning models. Since machine learning models typically require numeric values, we converted the corresponding columns to numeric data types. One notable challenge we encountered was initially identifying 42 different types of 'HOSPI_CODE_UCD.' However, upon closer examination, we discovered that half of the data was in a different data format. As a result, there were only 21 distinct 'HOSPI_CODE_UCD' values.

After identifying that the dataset encompassed four distinct hospitals, we made the decision to handle them separately. As a result, we created four separate datasets based on their 'ID_SITE_RATTACHE.' Subsequently, we grouped each DataFrame by 'CODE_ATC,' 'HOSPI_CODE_UCD,' and 'DATE.' The choice of aggregation operation (sum or average) for each group was determined based on the nature of the other features. With these steps, our data preparation process was completed, resulting in four distinct and appropriately prepared datasets. These datasets are now ready for use in the model engineering phase.

Data Engineering

Please note that all the steps that will be mentioned in this section were applied to all the four hospitals Data Frames.

Categorical Input Variables:

Since it is not possible to compute the correlation coefficient between a numerical and a categorical variable, we couldn't calculate the correlation matrix for all independent inputs, as was done in the exercise. Therefore, we had to adopt a different approach to determine which categorical independent variables significantly influence the dependent variable 'Quantity' and whether they are independent of each other.

We investigated which categorical independent variables:

1. Influence the dependent variable 'Quantity': This is achieved by first introducing (n-1) dummy variables for the n different values of the categorical variable. Subsequently, we conducted a regression with the dummy variables and the dependent variable, and we examined the R^2 score of this regression. The

R^2 score allows us to quantify how much of the dependent variable's variance is explained by the categorical variable.

Table 1 R-squared scores by categorical variable of HOSPITAL 3.

Categorical variable	R-squared score
CODE_ATC	0.553827711577664
HOSPI_CODE_UCD	0.9509409408838944
DATE	0.004323109081771914

2. Are independent of each other: We performed a pairwise Chi-squared test to determine which of the relevant categorical variables are independent from each other.

Table 2 P-values of categorical variables relation of HOSPITAL 3.

Chi-squared test for columns	P-value
CODE_ATC' and 'HOSPI_CODE_UCD'	0.0
CODE_ATC' and 'DATE'	1.0
HOSPI_CODE_UCD' and 'DATE'	1.0

This strategy aims to examine the impact of each categorical variable on the target variable 'QUANTITY' using linear regressions. One-hot encoding is employed to enable the use of categorical data in a format comprehensible by linear regression models. The R-squared value indicates how effectively the model explains the variability in the target variable. From this analysis we understood that “HOSPI_CODE_UCD” has the greatest influence on the setted target due to its high R^2. After the correlation analysis there was a strong relation between 'HOSPI_CODE_UCD' and 'CODE_ATC', so said that the decision was to just maintain “HOSPI_CODE_UCD”.

Numerical input variables

The correlation of the numerical features was analyzed as seen in class, utilizing the correlation matrix and heat map. The first realization was that the Number of hospital facilities “N_ETB” was the same for all the hospitals so it was dropped.

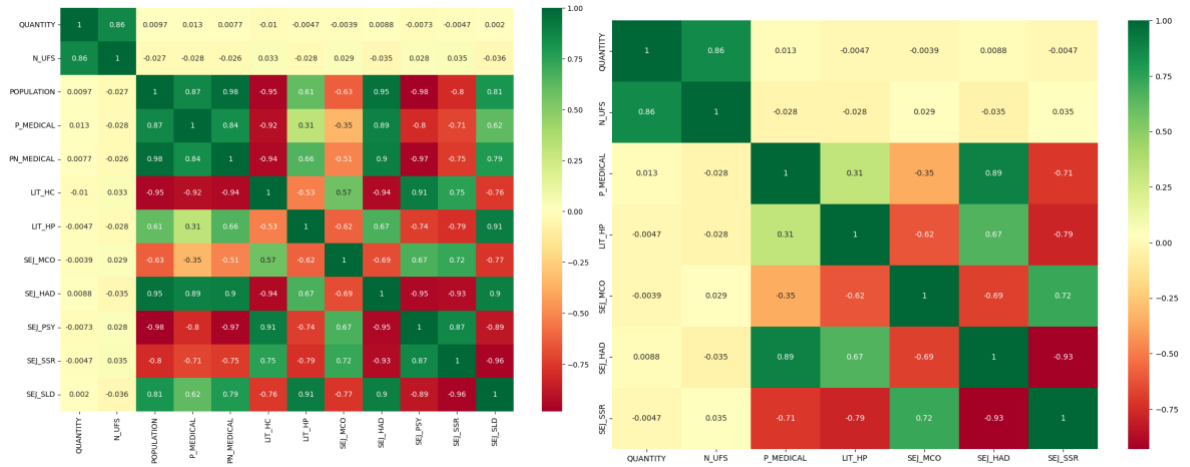

Figure 1 and 2. Before and after features choice

Please refer to the Figure 1 and 2, they show the behavior of the data before last feature cleaning. We decided to carry on to the model engineering with the following variables:

Table 3 Features chosen and type of variable.

Feature	Desctiption	Type of variable
HOSPI_CODE_UCD	Medicine UCD code	Categorical
DATE	combination made of the YEAR and MONTH	Categorical
QUANTITY	Quantity of medicine consumed daily	Numerical
N_UFS	Number of medical units	Numerical
P_MEDICAL	Number of physicians	Numerical
LIT_HP	Number of beds for full hospitalization	Numerical
SEJ_MCO	Number of visits in MCO departments	Numerical
SEJ_HAD	Number of visits in HAD departments	Numerical
SEJ_SSR	Number of visits in SSR departments	Numerical

These decisions were taken based on the nature of the feature, from a more business mindset, responding to the questions “What does it mean in real life for the logistics of the medicine? Does it influence? They were also supported by the correlation between them, there were needed variables as independent as possible to get a proper fit of the model.

Data normalization

Once the outliers were visualized, first we tried to do the modeling by removing outliers. But as we were not sure if there was really an assignable cause so we could remove them, we better proceed to normalize the data.

We scaled (normalized) the data in specific columns of the DataFrame using Min-Max scaling, reminding once again all of this was done by each of the four hospitals DataFrames. The selected columns for scaling are those that are not 'HOSPI_CODE_UCD', 'YEAR', or 'MONTH'. This scaling ensures that the values in these columns fall within a consistent range, making them suitable for certain machine learning algorithms and analyses.

Once we normalized, we made sure we still had information about the 21 UCD codes, once we verified, we proceeded to define the final data frames with which we would be fitting the different models. There were defined two settings of df, an encoded one for the single regressions for each medicine and an uncoded one.

Data splitting

Data was separated in the proportions suggested in class, 80% for the training and 20% for the testing. Also making sure the data was shuffled.

Results

As mentioned earlier in the methodology, the approach in this work was to experiment with different models to understand their behavior and determine the best fit. The results are presented as follows: the first approach involves fitting the model to the overall hospital data, and the second approach involves analyzing the best fit for each of the medicines. It's important to reiterate that we are replicating this process in the four DataFrames dedicated to each hospital.

First approach results:

From this analysis the suggestion we do, setting a constraint of accuracy percentage greater than 95% for both Learning and testing, that in other to get better predictions:

- Hospital 1 should predict with the proposed Overall XGB Regression.
- Hospital 2 should predict with the proposed Random Forest Regression.
- Hospital 3 should predict with the Overall XGB Regression.
- Hospital 4 should predict with the Overall Linear Regression.

Table 4 Accuracy obtained by each model for each Hospital.

Hospital 1	Accuracy	
	Learning	Testing
Overall Linear Regression	86,00%	89,00%
Overall XGB Regression	99,95%	96,00%
Overall Random Forest Regression	98,64%	97,00%

Hospital 3	Accuracy	
	Learning	Testing
Overall Linear Regression	95,00%	99,00%
Overall XGB Regression	99,50%	99,00%
Overall Random Forest Regression	99,30%	99,00%

Hospital 2	Accuracy	
	Learning	Testing
Overall Linear Regression	89,90%	90,00%
Overall XGB Regression	99,98%	93,00%
Overall Random Forest Regression	99,10%	96,00%

Hospital 4	Accuracy	
	Learning	Testing
Overall Linear Regression	95,98%	99,00%
Overall XGB Regression	99,95%	59,00%
Overall Random Forest Regression	99,59%	75,00%

Second approach results

Following the same approach to present results as in the first section, please find below Table 4 that summarizes the model that performs the best prediction among the analyzed models. The decision was taken comparing the R^2 value, choosing the max.

Table 5 Best model fitting by medicine and hospital based on R^2.

HOSPI_CODE_UCD	Best fit for medicin prediction depending on			
	Hospital 1	Hospital 2	Hospital 3	Hospital 4
3400890837149	Random Forest Regression	XGB Regression	Linear Regression	XGB Regression
3400891191226	Random Forest Regression	Random Forest Regression	Random forest	Random Forest Regression
3400891225037	Random Forest Regression	Random Forest Regression	Random forest	XGB Regression
3400891235203	Random Forest Regression	XGB Regression	XGB Regression	Random Forest Regression
3400891996128	Random Forest Regression	XGB Regression	Linear Regression	Random Forest Regression
3400892052120	Random Forest Regression	Linear Regression	XGB Regression	Random Forest Regression
3400892065366	Random Forest Regression	XGB Regression	Linear Regression	Random Forest Regression
3400892075761	XGB Regression	XGB Regression	Linear Regression	Random Forest Regression
3400892088310	Random Forest Regression	XGB Regression	Linear Regression	Random Forest Regression
3400892203645	Linear Regression	Linear Regression	Random forest	Random Forest Regression
3400892508566	Random Forest Regression	XGB Regression	XGB Regression	Random Forest Regression
3400892669236	Linear Regression	Random Forest Regression	Linear Regression	XGB Regression
3400892697789	Random Forest Regression	XGB Regression	XGB Regression	Random Forest Regression
3400892729589	XGB Regression	XGB Regression	Linear Regression	Random Forest Regression
3400892745848	Random Forest Regression	Random Forest Regression	Random forest	XGB Regression
3400892761527	Random Forest Regression	Random Forest Regression	Linear Regression	Random Forest Regression
3400892761695	XGB Regression	Linear Regression	XGB Regression	Random Forest Regression
3400893022634	XGB Regression	Random Forest Regression	Random forest	Random Forest Regression
3400893736135	Random Forest Regression	Random Forest Regression	Random forest	Random Forest Regression
3400893826706	XGB Regression	Random Forest Regression	Linear Regression	Random Forest Regression
3400893875490	Random Forest Regression	Linear Regression	Linear Regression	Random Forest Regression

Although we can provide a model, it represents the best result among those we've analyzed. This serves to underscore the fact that the Key Performance Indicators (KPIs) were not optimal in this approach, resulting in substantial percentages and negative outcomes. Following a thorough analysis of this situation, we formulated the hypothesis that the available data for each UCD (medicine) might not be sufficient to create a robust model. This is why, when we applied the model to the entire dataset, we achieved better results. A hypothesis for this behavior is presented below:

In the large model, we consider the dataset on a per-hospital basis and the model calculates the average for all quantities within that hospital when computing the variance. When drugs are ordered in various quantity ranges, the variance tends to be large. For example, if one drug is consistently ordered in small quantities and another is consistently

ordered in large quantities, the variance becomes significant because the mean falls somewhere between the small and large quantities. In the per-drug regression models, the mean used for variance from the quantities of a single drug within each hospital. In other words, we break down the per-hospital dataset into smaller, more specific datasets for each drug. Tend to have data points that are closer together since they pertain to a single drug. In the scenario described earlier with one drug having low volumes and another with high volumes, the variance in the separate datasets for each drug is much lower.

Furthermore, fewer data points are used in the smaller datasets, as some drugs have very limited data available per hospital (in some cases, only 3-4 data points remain). In very small datasets, it's more likely for the variance to be lower compared to large datasets because small samples may have less variation due to their limited number of observations. Both effects lead to the variance across all these smaller datasets being smaller in total compared to the variance in the large dataset per hospital. This could explain why the R^2 values in the individual regressions are significantly lower than in the regression involving all drugs. This is because the variance in the dataset plays a crucial role in the denominator of the R^2 formula.

Therefore, in the individual regressions where the dataset has a much smaller variance than the large dataset, the denominator is often very small. As a result, the model needs to make very precise predictions to achieve a high R^2score. On the other hand, in the regression involving all drugs, the dataset has a very large variance, and the model doesn't need to make extremely precise predictions to attain a good R^2score.

Conclusion

From this experience, we gain a better understanding of the fundamental principles of predictive modeling. It underscores the importance of grasping the context in which the model operates—of comprehending not just the numbers and algorithms, but the real-world significance of the features. What do these features truly represent? How are they interconnected? How do they collectively influence the system we seek to model?

In practice, achieving a high-quality model fit is intrinsically tied to a genuine understanding of the underlying system and trying different models to compare them. It emphasizes the need to dig deeper, to uncover the intricate relationships between variables, and to acknowledge the interplay of factors that shape the outcomes we aim to predict. In essence, a successful model is an outcome of a robust understanding of the system it represents—a testament to the power of knowledge in the world of data and analytics.

With the help of the model generation and comparison, we hope to be able to give PharmaGI a tool to predict the required monthly amount of medication.