HEALTHCARE PREDICTION CANCER RELATED USING DATA SCIENCE TECHNIQUES

Non scientific project within the context of IBM data science certification

Jean-Francois Moy

February 2020

1. Introduction

1.1 Background

Nowadays, cancer is getting in more and more people’s life directly or indirectly. The share of population having cancer in developed country has reached an historical high when this rate is still quite low in developing countries.

It implies to increase the capacity of healthcare related to this disease. Building a hospital takes time. Training doctors and nurses takes times as well. Being able to diagnose and heal people, each country should be able to forecast its need.

The goal of this project is to try to find some forecastable drivers correlated to cancer.

1.2 Problem

First of all, it has to be mentioned that this article is not a scientific publication. It is not claiming to explain scientifically some root cause of cancer. The idea is to identify some patterns that seem reasonable to help countries structuring their healthcare system.

It is generally admitted that cancer is due to: 1/3 genetics, 1/3 environment, 1/3 randomness. As you can imagine, it would be difficult to get some data about genetic material, and way more difficult about randomness. Thus, we will mainly consider data about people’s environment.

Another issue that we could face would be to find a model composed by features that are difficult to forecast.

1.3 Interest

The interest of this project is idealist but laudable. This is about trying to forecast healthcare needs to try to save some lives or at least, extend life expectancy. It could interest developing countries that doesn’t have a structured healthcare system.

2. Data acquisition, cleaning and use

2.1 Data acquisition and feature selection

Most of the data that have been used in this project are coming from <https://ourworldindata.org> . As mentioned in its ‘About’ page, this website gives free access to data provided by ‘researchers at the University of Oxford, who are the scientific editors of the website content; and the non-profit organization [Global Change Data Lab](https://global-change-data-lab.org/), who publishes and maintains the website and the data tools that make our work possible. ‘

The first step has been to determine the kind of data that could be relevant. After reading different articles about cancer, I made the decision to use the features presented in the table below. As said before, the environment is supposed to be one of the factors explaining the development of tumorous cells. I tried to gather different features related to the environment of a human being. The difficulty being to find relevant and measurable data.

|  |  |  |
| --- | --- | --- |
| # | Feature name | Description |
| 1 | Age\_standardized\_neoplasms\_percent | This correspond to the target value. This means this is the value that we want to be able to forecast thanks to the other features. |
| 2 | Population | number of inhabitants by country |
| 3 | 65andover\_percent | Share of the population having more than 65 years old |
| 4 | fruit\_gram\_per\_day\_per\_capita | Quantity of fruit eaten by one person on a daily basis. |
| 5 | HDI | Human Development Index. The United Nations describe it as follows: “The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.” |
| 5 | Life\_expectancy | Life expectancy |
| 6 | nb\_hospital | Number of hospital by country |
| 7 | Urban\_percentage | share of the population living in an urban area |
| 8 | GDP\_per\_capita | Gross Domestic Product (GDP) is the monetary value of all finished goods and services made within a country during a specific period. This index is here calculated by inhabitant. |
| 9 | calories\_day\_capita | average of calories eaten by one person in one day |
| 10 | pesticide\_kg\_ha | average of quantity of pesticides used by ha per year |
| 11 | milk\_kg\_year\_capita | average of quantity of milk drunk by one person in one day |
| 12 | protein\_g\_day\_capita | average of quantity protein eaten by one person in one day |
| 13 | meat\_kg\_day\_capita | average of quantity of meat eaten by one person in one day |
| 14 | fat\_capita\_day | average of quantity of fat eaten by one person in one day |
| 15 | cigarette\_per\_day | average of number of cigarettes smoked by one person in one day |
| 16 | vegetable\_kg\_pcapita\_year | quantity of vegetables eaten by one person on a yearly basis |

These features have been selected for a period of time starting from 1990 to 2017. This period has been chosen because it allows us to have a continuous dataset for a large number of countries. Developing countries generally have few data before 1990.

The second step has been to build only one data frame from different data sets, each feature coming from a distinct file. The material was quite clean and uniform, but values were missing for some dates. When the data was missing for 20% or less, I chose to default the values with the first value available right after. If the missing part was 20% or more, I chose to delete the country from the dataset. This happened for small isolated countries, with few inhabitants.

2.3 Data cleaning

The dataset being produced, then comes the time for data analysis. From the initial selection of features, some have been removed, like life expectancy, which is redundant with HDI, cigarette per day or pesticides per hectare which were not covering a scope large enough.

Also, as I don’t have scientific knowledge of cancer, I shared this project to have some feedback on my methodology and features. Results are still pending and will be exposed in the final report.

2.4 Methodology – Use of the data

Data will be used to feed an algorithm to find a model able to predict cancer. Feature selection will be done either manually or automatically depending on the chosen algorithm.

Then, a comparison will be done between the forecasted needs and the current available resources and the forecasted available resources.

Then Foursquare will be used display the results of this comparison. A color code will be created to quickly determine which countries are going to cover their needs and which ones will have more difficulties to do so.

END OF MATERIAL FOR WEEK 4

What I need:

- What indicator to measure the needs in terms of health care needs related to cancer: human/machines

- Opinion about feature: selected ones, new ones

For the second week, the final deliverables of the project will be:

1. A link to your Notebook on your Github repository, showing your code. (**15 marks**)
2. A full report consisting of all of the following components (**15 marks**):

* Introduction where you discuss the business problem and who would be interested in this project.
* Data where you describe the data that will be used to solve the problem and the source of the data.
* Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
* Results section where you discuss the results.
* Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
* Conclusion section where you conclude the report.

1. Your choice of a presentation or blogpost. (**10 marks**)

3. Exploratory Data Analysis

3.1 Calculation of target variable

Player improvement year over year was not a feature in the dataset, and had to be calculated. I

chose to calculate the difference of win shares between two consecutive years as the target

variable. Win shares were chosen out of a few metrics because it is the most interpretable, after

all, we play basketball to win. Calculated player improvement had a normal distribution centered

around 0, with most values between -6 and 6. To verify if this calculation is consistent with

people’s eye-test of player improvement, I plotted the rank of improvement of past Most

Improved Players winners among all players, and found that in most cases, they were among the

most improved players (Figure 1). This suggested that the chosen metric of player improvement,

was a reasonable one.

3.2 Relationship between improvement and age

It is widely accepted that younger players are more likely to improve than older players, and it

was indeed supported by our data. Players’ median improvement declined as players’ age

increased (Figure 2), and the mean improvement of different age groups (<25, 25-29, 30-34,

>35) were all significantly different from each other (z-test, p<0.001, except for 30-34 vs. >35,

p=0.002).

Figure 1. Rank of delta-win-share of Most Improved Players winners among all players of each year

Figure 2. Box plot of improvement of players of different ages.

3.3 Relationship between improvement and overall ability

The hypothesis here is that players who are already stars don’t have much room to improve,

while a mediocre player can still improve. Our data were consistent with this hypothesis. Using

win share per 48 minutes (WS/48) as a measure of a player’s overall ability, I observed a

negative relationship between a player’s overall ability and his improvement next season (Figure

3). The mean improvement of star players (WS/48 > 0.2), solid players (WS/48 between 0.1 and

0.2), rotational players (WS/48 between 0 and 0.1), and “scrubs” (WS/48 below 0) were

significantly different from each other (z-test, p<0.001) (Figure 4).

3.4 Relationship between improvement and minutes played

I hypothesized that players with less playing time might be more likely to improve. If a team

recognizes a player's positive contribution during his limited time, he is likely to get more

playing time, and therefore increase his production and/or improve his skills. On the other hand,

if a good player is already a starter, he is already playing a lot of minutes and can't get more

playing time. After inspecting the data, it was true that players who played less than 25 minutes a

game had statistically higher improvement than those who played more than 25 minutes a game

(z-test, p<0.001). However, the actual difference of mean between the two groups was small

(~0.7).

3.5 Relationship between improvement and games played

I observed a negative relationship between player improvement and the games played (Figure 5).

If a good player missed significant numbers of games, it was probably because of injury, which

might have negatively impacted his performance. He might return to his former form next

season, and therefore improve. Players who played fewer than 50 games were more likely to

improve than those who played more than 50 games. (z-test, p<0.001, difference of mean=1.3).

Figure 3. Scatter plot of improvement and player overall ability (measured by win share per 48 minutes)

Figure 4. Histogram of player improvement separated into 4 groups based on how good a player is.

Figure 5. Scatter plot of player improvement and games played.

3.6 Relationship between improvement and positions

There is this myth among NBA fans that frontcourt players take longer to adapt to the NBA than

backcourt players, therefore they would have smaller improvement in the first few years. I

transformed the feature of player position into a binary feature (frontcourt vs. backcourt players)

and found that there was no difference between frontcourt and backcourt players in their

improvements, even in their first 2 years (z-test, p=0.34)

3.7 Relationship between improvement and last year’s improvement

I hypothesized that a player’s improvement might be correlated with his previous improvement,

because younger players might improve continuously for a few years, and older players might

decline for a few years straight. It turned out that the relationship between improvement and

prior improvement was negative (Figure 6). In other words, more often than not, a player will

“regress to the mean” rather than continuously improve or decline.

Figure 6. Scatter plot of player improvement and that of last season

3.8 Relationship between improvement and draft positions

I, as many other basketball fans, thought that players drafted earlier are generally more talented

and therefore more likely to improve than players drafted later, at least in their early years. It

turned out this was only true for a few really young and talented players (Figure 7) . Players

under the age 20 with different draft positions did not have statistically different improvement

(z-test, p=0.16).

3.9 Relationship between improvement and teams

I engineered two features based on team information: was a player on a good or bad team, and

did the player change team next season. Player improvement and team strength (measured by

total win shares) had a very weak negative relationship. Players that changed teams were slightly

more likely to improve than players that stayed on the same team (z-test, p<0.001, difference of

mean = 0.2).

Figure 7. Box plot of player improvement among different draft groups and ages

4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict player

improvement. Regression models can provide additional information on the amount of

improvement, while classification models focus on the probabilities a player might improve. The

underlying algorithms are similar between regression and classification models, but different

audience might prefer one over the other. For example, an NBA team executive might be more

interested in the amount of improvement (regression models), but a general NBA fan might find

the results of classification models more interpretable. Therefore, in this study, I carried out both

regression and classification modeling.

4.1 Regression models

4.1.1 Applying standard algorithms and their problems

I applied linear models (linear regression, Ridge regression, and Lasso regression), support

vector machines (SVM), random forest, and gradient boost models to the dataset, using root

mean squared error (RMSE) as the tuning and evaluation metric. The results all had the same

problems. The predicted values had much narrow range than the actual values (Figure 8), and as

a result, the prediction errors were larger as the actual values deviated further from zero (Figure

9). These results were not acceptable, because players with large improvement/decline were

arguably more important for NBA teams to predict than players with little change in

performance. Having larger errors on those predictions was obviously not desirable.

4.1.2 Solution to the problems

The reason behind these problems were the uneven distribution of player improvement, in that

players with little improvement/decline were more common than players with big

improvement/decline (Figure 8). Therefore, the models tried to prioritize minimizing errors on

players with little improvement/decline when RMSE was used as the evaluation metric. My

solution to this problem was to assign weights to samples based on the inverse of the abundances

of target values. In other words, players with large improvement/decline would have higher

weights in model training and evaluation because they were more rare. Using this method, all

models predicted target values with similar range and distribution as the actual target values

(Figure 10).

Figure 8. Distribution of actual and predicted improvement using linear regression with equal weights of

samples.

Figure 9. Scatterplot of prediction errors vs. actual target values using linear regression with equal

weights of samples.

Figure 10. Distribution of actual and predicted improvement using linear regression with different weights

of samples based on inverse of sample abundance.

4.1.3 Performances of different models

Using the new approach of different sample weights, I built linear regression, SVM, random

forest, and gradient boost models using weighted root mean squared error as the evaluation

metric. For each model, hyperparameters were tuned using the same metric and cross validation.

For comparison, I also built a simple linear regression model with just one independent variable

(age) as the benchmark model. SVM had the best performance among all models, which had

~26% less error than the benchmark model (Table 2). The predicted improvements had linear

relationship with the actual improvements (Figure 11).

Table 2. Performance of the regression models.

Benchmark

(one feature)

Linear

Regression

SVM Random Forest Gradient Boost

Weighted

RMSE

3.84 2.98 2.86 2.93 2.96

4.2 Classification models

The application of classification models was much more straightforward. I divided the samples

into two classes (improvement>=0 or <0). The number of samples in each class were about the

same. I chose logarithmic loss as the metric here because the results would probably be presented

with probabilities and logarithmic loss puts more emphasis on the probabilities than other

metrics. Logistic regression, SVM, random forest, gradient boost models and a voting model

were tuned and built. Among the individual models, the SVM model performed the best (~67.5%

accuracy), and voting model performed similarly as the SVM model (Table 3), though the

differences between models were small.

Figure 11. Scatter plot of predicted and actual player improvements of the SVM model.

Table 3. Performance of classification models. Best performance labeled in red.

Logistic

Regression

SVM Random Forest Gradient Boost Voting Model

Log Loss 0.605 0.603 0.612 0.613 0.603

Accuracy 0.675 0.675 0.672 0.672 0.675

No. of True

Positives

835 830 810 815 838

No. of False

Positives

413 406 396 400 416

No. of False

Negatives

438 443 463 458 435

No. of True

Negatives

929 936 946 942 926

Figure 12. A section of ROC curves of different classification models.

I also evaluated the models using their ROC curves. In this particular problem, lower false

positive rate is more important than higher true positive rate. In other words, it is more important

to be sure that a player will improve as predicted, rather than predict all players who will

improve, simply because a team can only have limited number of players. In the ROC curves

with low false-positive rate, the voting model had slightly higher true positive rates than other

models (Figure 12).

5. Conclusions

In this study, I analyzed the relationship between NBA players’ improvement/decline and their

performance and biographic data. I identified age, win share, minutes/games played,

improvement last season among the most important features that affect a player’s improvement

next season. I built both regression models and classification models to predict whether and how

much a player would improve/decline. These models can be very useful in helping NBA team

management in a number of ways. For example, it could help identify players to acquire,

estimate the size of the contract to offer players, plan for performance changes of players already

on the team, etc.

6. Future directions

I was able to achieve ~26% improvement from the benchmark model in the regression problem,

and ~68% accuracy in the classification problem. However, there was still significant variance

that could not be predicted by the models in this study. I think the models could use more

improvements on capturing players’ individual traits. For example, two players might have

similar performance metrics, but one might be more physical and the other might be more

finesse. The future performance of these two types of players might be different. Another

example is that players whose contracts are expiring might play harder/better than players who

just signed hefty contracts. More data, especially data of different types, would help improve

model performances significantly.

Models in this study mainly focused on individual features. However, interactions with

teammates, coaches, might also contribute to a player’s performance. For example, if a player

had a new teammate who is a superstar at the same position, his performance is likely to suffer

because of competition. These interactions data are obviously more difficult to extract and

quantify, but if optimized, could bring significant improvements to the models.