Data Analysis and Model Building Support for the Development of a Future mHealth Solution for Musculoskeletal Diagnosis

Public Project Summary

Laszlo Szilagyi – 15.12.2018.

Table of Contents

Abstract

This capstone project for the CEU Msc in Business Analytics program aims to provide data analysis and model building support for the development of Sportklinika's Mobile Motion Lab, the basis of a future mHealth solution for musculoskeletal diagnosis.

Background

SPORTklinika is a start-up established in Budapest in August 2016. The company is involved in mHealth, specifically they are developing solutions using smartphone devices' accelerometer sensors for musculoskeletal diagnosis to help rehabilitation.

The start-up's planned product, Mobile Motion Lab will be an mHealth solution, a platform free application for mobile phone or wearable devices which generates musculoskeletal diagnosis with machine learning algorithms. With a smartphone in your pocket the app will track the change of your physical condition and help to identify problems in early phase. The vision is to have the first smartphone app to self-check musculoskeletal symptoms and help find a medical doctor.

SPORTklinika has already managed to develop the minimum viable product (MVP), which is made up of a desktop application and commercially available hardware components. The company has conducted over 130 examinations on individuals including medical anamnesis and a 6-minutes protocol that includes walking and running on a treadmill. The MVP demonstrates the viability of the concept and is now about to be validated in order to allow professional medical use (to cooperate with medical institutions, the diagnostical performance of the MVP has to be validated).

However, having the MVP at hand only demonstrates that accelerometer data can be obtained under controlled conditions. It was the capstone project task to verify that the data can be processed and effectively utilized to build machine learning (ML) models e.g. for segmentation of the individuals based on the health status of their lower limbs.

Task for the capstone project

The objective of the capstone project was to provide data analysis and model building support for the development of SPORTklinika's Mobile Motion Lab. The task was to investigate and help better understand for the start-up's experts how accurate Machine Learning (ML) predictions can be given on an individual's health status regarding the health condition of his/her lower limbs' joints, which features contibute most to the predictions,, and whether ML algorithms could be used to develop the planned application based on the currently available dataset.

Methodology

The dataset

The input data set for the capstone project consisted of calculated metrics derived by SPORTklinika from the raw accelerometer data, complemented with personal and anthropometric data on the individuals examined.

The target variable for supervised learning was *benchmark_1*, a binary target (1 = healthy; 0 = not healthy regarding the condition of the individual's lower limbs' joints) defined by the start-up, based on the combination of the medical anamnesis and a medical expert judgement based on the shape and comparison of the curves produced from accelerometer data. Therefore, the task was methodologically to perform a use case of supervised binary classification on a balanced dataset.

However, most of the time consumed for the project was spent on data exploration and feature engineering, as the dataset were problematic in more ways. To address the problems, we have discussed the distribution and potential outliers of all 145 (mostly numeric and categorical) variables personally with the experts of SPORTklinika. Variables were classified into groups based on their characteristics. The most important from the aspect of future application were the calculated metrics derived from the raw accelerometer data by SPORTklinika (like the average length of a step with the left leg in milliseconds, or standard deviation of the length of steps with the right leg), and the flags indicating if a calculated metric is out of 1 standard deviation from the mean.

The final cleaned and engineered dataset for the capstone project covered only 80 individuals (examinations) and 59 variables, who were professional sportsmen playing ice hockey, basketball, futball or volleyball.

The Machine Learning workflow

All phases of the usual ML workflow were performed during the capstone project, namely: data exploration, data cleaning, feature engineering, feature selection, model training and model testing. However, instead of trying to achieve the best possible performance for prediction, the objective of the capstone project was to help SPORTklinika, the client start-up to better understand their data on

the one hand, and the ML workflow and the possibilities (but also the limits) offered by using ML algorithms for their specific use case on the other.

A number of statistical and ML methods were used to accomplish the task. To name a few more specific, feature selection was supported by the use of Recursive Feature Elimination (RFE) function of the caret package; the trial of the Boruta package providing a feature ranking and selection algorithm based on random forest as default; and also the trial of the InformationValue package. To help better understanding which features contribute most to prediction of the target, a Logit model was built, also with the addition of LASSO regularization.

In the next phase of the exercise, different ML algorithms were tested to explore the possible prediction accuracy given the limited amount of data. Caret package, the “Swiss army knife of R ML tools” was selected for the task, using Logit, LASSO, Random Forest, Neural Network (nnet) and XGBoost (xgbTree) algorithms with some hyperparameter tuning.

Finally, to be able to provide additional information and visualizations on how the algorithms utilize information in the data, I tried to “look in the black box” of Machine Learning by using various methods from visualizing variable importance, to supplementing the analysis with Principal Component Analysis (PCA provides some powerful tools to visualize important variables and also groupings of observations within the data), and also experimenting with packages built to provide insight and explanation on how ML models work, like LIME and DALEX.

Summary and findings

During the analysis, I could not define a really accurate model with any of the applied ML methods. It is not sure that even 80% accuracy in the prediction of the target can be reached consistently by any of the defined ML models.

However, 80 observations are clearly not enough to draw clearly decisive results. The code is written so that it tries to ensure reproducibility and easy scaling at the same time, so more data or even new variables can be fed easily to the R code. Therefore I think SPORTklinika could and should carry on with the exercise. (Of course, larger dataset size would require the introduction of parallel computing, but the current data set was easy to handle on a laptop.)

On the positive side, it seems that there is a real relationship between the features originating from accelerometer data (calculated etrics more directly, and flags indirectly, as the flags represent a medical expert’s feature engineering approach; good news that in practice the flags were applicable for prediction with satisfying results).

The many different visualizations performed during the project hopefully contribute to the start-up’s understanding of the data and the underlying patterns.

Some further recommendations were to reconsider the applied definition of the target variable, until a really practical definition is forged; and to go back to the basis of accelerometer data, and try to engineer further useful variables based on fequency theory.