# ANALYZING A URINARY INCONTINENCE DATASET AND BUILDING A CLASSIFICATION MODEL FOR PELVIC TUMORS

CAPSTONE PROJECT PUBLIC SUMMARY

By

Akylbek Subanbekov

Submitted to

Central European University

Department of Economics and Business

In partial fulfillment of the requirements for the degree of

Master of Science in Business Analytics

Supervisor: Jenő Pál



Budapest, Hungary

2020

# Objective

This capstone project analyzes an existing urinary incontinence dataset provided by the Hungarian Incontinence Society (the Client) and builds a classification model to uncover its predictive value in identifying pelvic tumors. The Client's goal is to create a predictive model based on a big pool of medical data that would help doctors identify cancer conditions at the early stage, thereby increasing the chances of recovery for patients.

The Client, who wants to develop a unified online database to which doctors could refer and quickly assess whether a patient with a given condition has a higher risk of developing cancer, can use the project outcome as a reference to build a predictive classification model. The ultimate goal of the online database is to help doctors make better decisions in assessing a patient's condition, decrease the number of misclassifications due to fatigue and human error, and prioritize patients who are at a higher risk of developing cancer.

# Data source

The Client collected the dataset in the second half of 2018 by surveying patients with urinary incontinence complaints. The predictive variables can be classified into two categories: self-reported incontinence related problems, such as complaint intensity due to incontinence, and comorbidities at the time of a survey, such as diabetes, obesity, and urinary tract infections.

The variables on the geographical location of the counselling were removed as certain hospitals and doctors could specialize in cancer treatment, thereby being strong yet misleading predictors for cancer conditions.

The original dataset contained 9,908 entries for male and female patients. After discussion with the Client, 2,345 entries for male patients were dropped due to severe class imbalance that did not represents the real case. The dataset used for model building contained 7,518 entries for female patients with 432 (5.74%) cases of cancer, making this an imbalanced classification problem.

# Methodology and Results

The dataset was cleaned and transformed to a tidy data format for the further modelling in R. The diagnostic tools used to interpret and evaluate the binary classification models were area under the ROC curve (AUC) and area under precision-recall curve (AUCPR).

### *Data visualization*

Interactions between variables were visualized as a part of the exploratory data analysis to better understand the data. The visualizations helped to uncover unexpected interactions such as lower proportion of negative cases among patients with obesity, or among the patients in the age group of 40.

Variable importance plots were used to assess which predictor variables were the most important in predicting the target variable. The top five predictors for both GBM and Random Forest models turned out to be the same. In addition, partial dependence plots were used to illustrate the relationship between input variables and the predictions of the black-box models.

### *Missing values*

Missing values were treated as missing for a reason. To examine whether data imputation could improve the performance of a model, I compared the performance of tree-based models that could account for missing data with the performance of models that were trained on data with imputed missing values. Both pre-processing and data imputation were done after splitting the dataset into training and testing to replicate the situation where the future dataset would be used for prediction. The model trained on the original dataset had higher precision when the recall was over 80%. Therefore, I used the original dataset in the subsequent steps.

### *Class imbalance*

To counteract the negative effects of class imbalance, sampling techniques and alternate cutoffs were used to increase the prediction accuracy of the minority class.

In the first case, a baseline model was trained on 5 different datasets: original, down-sampled, up-sampled, weighted, and Synthetic Minority Oversampling Technique (SMOTE) datasets. The model trained on the original set had the highest AUC on both validation and test sets. Therefore, the original dataset was used in the subsequent steps.

In the second case, the alternate cutoffs were used to increase the proportion of true positive cases (recall) at the expense of specificity. A threshold based on max mean per class accuracy and F2 metric gave recall of 0.8247, specificity of 0.5759 and precision of 0.1201. Further modelling using AUC did not improve these baseline results. As the trade-off between precision

and recall seemed more meaningful than the trade-off between recall and specificity, the next step focused on the precision-recall (PR) curve.

*Precision-recall curve and AUCPR*

The PR curve was more explicit and informative than a ROC curve in differentiating between several models in case of imbalanced classes. First, it focused directly on the minority class. Second, it was more convenient to compare different models and to retrieve a threshold based on a visual data for a required trade-off between precision and recall.

Using PR curve helped to achieve a higher recall value of 0.9437 (up from 0.8247 in a baseline model) with slightly lower precision of 0.0949 (down from 0.1201 in a baseline model).

# Contribution

Data visualization part, which showed a few unexpected results, could help the Client to better understand the data and improve the data collection process in the future. For example, the Client expected the proportion of cancer cases to be higher among patients who gave birth or among patients with obesity, while the visualizations of the dataset showed the opposite.

The model results were visualized using variable importance and partial dependence plots that showed the variables with higher predictive value and illustrated the relationship between input variables and the predictions.

The incontinence related database showed a predictive value in identifying pelvic tumors. However, it is important to test the models on new data sets to confirm or reject these results. In a former case, the Client could use the deliverables of this project as a reference in further research in this area as it represents the complete workflow from data cleaning to model building.

The next potential steps could be exploring the causes behind the unexpected results, validating the model on new datasets, as well as further improving models by applying more advanced techniques and by addressing the issues uncovered during the project.