LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

Master in

Actuarial Science

Master's Final Work

Internship Report

Healthcare Provider efficiency in Workers' Compensation - An approach with Machine Learning

Francisco Fernandes Correia do Canto Moniz

October - 2019

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Master in

Actuarial Science

Master's Final Work

Internship Report

Healthcare Provider efficiency in Workers' Compensation - An approach with Machine Learning

Francisco Fernandes Correia do Canto Moniz

Supervisors:

João Manuel de Sousa Andrade e Silva

Delminda Luísa Rangel Amado

October - 2019

Acknowledgments

Undertaking this thesis was truly a journey, throughout which the support given was unbelievable. I would first like to thank my faculty advisor João Andrade e Silva for his helpful feedback and overall good disposition to be part of this project. His advice was truly valuable and I always came out of our discussions with a clearer head.

I would like to acknowledge my colleagues at Fidelidade for their good disposition to help newcomers and willingness to share their knowledge. Your support is greatly appreciated. In particular I would like to thank my supervisor at Fidelidade, Delminda Amado, for giving me the orientation and tools necessary to complete this project. A special thank you goes out to the head of the department, Rui Esteves, for inviting me to join this project as well as for the invaluable advice provided throughout.

In addition, I would like to thank my friends, with who I shared my successes and my frustrations, you were always ready to provide a helping hand or coming up with numerous distractions to take my head off this research.

At last, there is my family who supported me unequivocally. Thank you for providing a sympathetic ear and for your wise counsel. It is safe to say, all of this wouldn't have been possible without the opportunities you gave me.

Abstract

Workers' Compensation is a mandatory and very competitive Line of Business (LoB) for Insurance Companies. Companies cannot raise premiums too much from fear of losing market share, but they also cannot lower them as it needs to be financially viable. With the growing popularity of Data Science models, internal processes are being adapted to more precise and advanced models.

Following a work accident, a healthcare provider is recommended to the injured workers.

It is our opinion that the recommendation system in production is too rudimentary and can be optimized. It was our main objective in this internship to optimize this recommendation system. Our solution provides an estimate of medical and transportation cost which depends on the healthcare provider. With this project, claim managers can have access to the most efficient healthcare unit, as well as an estimate of the corresponding liability.

Models for the cost were developed using Extreme Gradient Boosting (XGB) as an alternative to the staple, Generalized Linear Models (GLM). By changing the loss function we applied XGB to both regression and classification problems and achieved more precise predictions.

To encode categorical variables in numerical values, we developed an algorithm that groups costs according to each level of a variable for the past 3 years and then computes its average. This encoding technique is similar to target encoding.

To assess the added value of this model, we compute the costs for the current recommendation applied. The new recommendation recognizes cheaper alternatives and predicts savings in total expenses of up to 1,7 million Euros.

Key-Words: Machine Learning, Regression, Boosting trees, Recommendation System, Patient Attribution, Workers' Compensation

Resumo

O ramo de Acidentes de Trabalho é uma linha de negócio obrigatória e com bastante competitividade. Nos últimos anos, temos observado um crescimento na popularidade de Data Science e esta transformação passa também por atualizar os modelos e processos internos aplicados em seguros.

Após um Acidente de Trabalho, é recomendado ao beneficiário um prestador clínico para acompanhar o seu tratamento.

Usando várias variáveis sociais e patológicas modelamos custos médicos e de transportes, dependendo estes do prestador clínico principal do lesado. Esta metodologia permite que os gestores de sinistros tenham acesso não só à melhor recomendação como também a uma estimativa final de custos.

Os modelos de custo esperado e frequência foram desenvolvidos usando Extreme Gradient Boosting em vez de modelos mais tradicionais como os GLM. XGB é um modelo de Machine Learning útil para previsão tanto em regressão como em classificação. Para problemas com muitos dados, este modelo tende a prever com maior precisão e rapidez.

Para uma utilização eficaz do modelo as variáveis categóricas são codificadas em numéricas através de target encoding modificado. Isto é, as observações são agrupadas de acordo com os níveis da variável e com o ano de ocorrência, e é calculada a média da variável de resposta para cada nível com as observações dos 3 anos antecedentes.

Por fim, para avaliar o valor acrescentado do modelo desenvolvido, calculamos os custos incorridos caso optássemos pela recomendação em prática. A nova recomendação consegue poupar até 1,7 milhões de euros por ano entre despesas médicas e de transportes.

Palavras-Chave: *Machine Learning*, Regressão, *Boosting trees*, Sistemas de Recomendação, Atribuição de Pacientes, Acidentes de Trabalho

List of Figures

3.1 Scoring of observations, by summing predictions of individual trees to calculate the final prediction. 14

4.1 Diagram of expenses in Workers' Compensation 20

4.2 Distribution of medical expenses. 21

4.3 R output from Expected Medical costs with XGB 24

4.4 R output from Expected Medical costs with GLM 24

4.5 R output from Physiotherapy costs 25

5.1 Adjustment factor and derivation of driving distance between injured workers' residence and health clinic 27

A.1 R output from Expected Frequency of Medical Expenses in XGB 36

A.2 R output from Expected Frequency of Medical Expenses in GLM 37

A.3 R output for the proportion of physiotherapy with XGB 38

A.4 R output from Expected Frequency of Physiotherapy Sessions in XGB 38

A.5 Variable Importance of Expected Cost of Medical Expenses in XGB 39

A.6 Variable Importance of Expected Cost of Medical Expenses in GLM 40

A.7 Variable Importance of Expected Frequency of Medical Expenses in XGB 40

A.8 Variable Importance of Expected Frequency of Medical Expenses in GLM 41

A.9 Variable Importance of Expected Cost of Physiotherapy Expenses in XGB 41

A.10 Variable Importance of Expected Frequency of Medical Expenses in XGB 42

A.11 Variable Importance of the proportion of physiotherapy in XGB 42

List of Tables

3.1 Meta parameters tested and values chosen for training models 18

6.1 Global Summary of scenario I 30

6.2 Global Summary of scenario II 31

Contents

Acknowledgments **i**

Abstract **ii**

Resumo **iii**

1 Introduction **2**

1.1 Context . . . 2

1.2 Motivations and Goals . . . 3

1.3 Document Structure . . . 4

2 Data **5**

2.1 Introducing the main datasets . . . 5

2.2 Variables . . . 6

2.2.1 Response variables . . . 6

2.2.2 Explanatory variables . . . 7

2.2.3 Healthcare providers . . . 9

2.3 Main Constraints . . . 10

3 Extreme Gradient Boosting **11**

3.1 Models leading to Extreme Gradient Boosting . . . 11

3.2 Training and Validation loss . . . 12

3.3 Reviewing Tree Boosting . . . 13

3.4 Split Finding . . . 16

3.5 Meta parameters . . . 17

3.6 Tweedie . . . 17

3.7 Generalized Linear Models . . . 19

4 Modelling Treatment costs **20**

4.1 Medical Aggregate Expense 22

4.2 Physiotherapy 25

5 Transportation Costs **26**

5.1 Number of Kilometers 26

5.2 Number of trips 28

5.3 Price per Kilometer 28

6 Recommendation Systems **29**

6.1 Scenario I - no restrictions 30

6.2 Scenario II - Conditional recommendation 30

7 Conclusion and Future Work **32**

7.1 Conclusions 32

7.2 Future Works 32

A Models and Variable Importance **36**

Chapter 1

Introduction

1.1 Context

Workers' Compensation is a mandatory insurance in Portugal. It covers accidents arising from work-related phenomena by ensuring that, after a potential accident, the worker in question is able to perform its usual work as fast and as efficiently as possible; or by ensuring the worker gets compensation for potential permanent disabilities. A work accident is one that happens at the work place during the work time and causes direct or indirect bodily injury, functional disturbance or disease that leads to disability for usual work, gain or death. By definition, a work place is understood any place where the worker has to go in virtue of his/her job and where he/she is directly or indirectly under the supervision of the employer. Work time comprises any time preceding the usual work time in preparation or related to it and any time succeeding in acts related to it. Usual and forced interruptions are also contemplated [APS, *Associação Portuguesa de Seguradores* (2017*b*)].

In exchange for guaranteeing the aforementioned conditions for its clients, insurance companies charge a value per insured worker, commonly referred to as a premium. Premiums are set such that they are fair for both insurer and employing company. Decreasing premium values increased the competitiveness for this line of business for many years, and insurance companies needed to keep up in order to keep market share. Eventually, the loss ratio turned "negative", which is to say the premiums collected were lower than the liabilities. To counteract this effect, premiums have to increase and/or liabilities decrease. Within the context of this project we approached this problem by investing in optimizing internal models such as a recommendation for healthcare providers.

After a work accident, a healthcare provider is recommended to the injured worker based on his residency's municipality and the observed pathology. By analyzing expected costs for each healthcare provider and expected transportation costs the recommendation can be improved. This means that a less-than-optimal recommendation will happen less often therefore reducing costs inherent to this operation.

A recommendation based on costs would usually be obtained from the application of Generalized Linear Models. The GLM based approach provides an easy to interpret structure and displays the influence of each variable on the final prediction. This model provides a greater visibility which is crucial because insurance companies are required by law to disclose the factors that lead to oscillations in premiums. However, for internal models this explainability is unnecessary and GLMs may be replaced by better performing models. Within the context of this project we use Extreme Gradient Boosting. Although it is a black-box model it is less susceptible to over-fitting which means that it behaves better than GLM when exposed to new an unknown combination of factors. The analysis was conducted using R, a programming environment for statistical computing and plotting. All these factors made this project significantly relevant, not only for the industry in question, but for the personal development of the author of this report.

The developed model considers several social, demographic and pathological variables to estimate the average cost and frequency variables. Perhaps the most important explanatory variable is the healthcare provider. By stressing it we simulated the cost for each medical facility, and thus we can compare and choose the best option for each patient. Transportation is also an important factor and contributes for a more accurate choice of the final cost. A more precise prediction of transportation costs contributes to a more accurate choice of the healthcare provider.

1.2 Motivations and Goals

The present report summarises the analysis developed from February to July at *Fidelidade*. I joined the *Direção de Estudos Técnicos de Não Vida e Estatística* (DET) where I focused on the line of business of Workers' Compensation (WC).

The project was organized in a progressive manner where three main goals were defined:

- Modelling the main cost components that can be dependent on the healthcare provider;
- Evaluating the impact of the distance between the injured worker and the healthcare provider;
- Finding the healthcare provider that minimizes the total costs of the injured.

The scope of the internship is to start building a model for patient attribution, as well as claim estimation. By obtaining a preliminary average of the incurred costs by the treatment of a workplace injury a more accurate provisioning can be made at the start of each process.

Transportation from the residency of the injured worker to the treatment facility must be provided by the insurer and its cost is estimated assuming taxis are used. The prices for taxis are fixed leaving only the driving distance and the number of hospitals visits to be estimated.

For medical expenses this dissertation explores two models - Generalized Linear Models and Extreme Gradient Boosting. Generalized Linear Models have been the standard in insurance, results are easily explained and have a solid mathematical background. By drawing a comparison with Extreme Gradient Boosting, we can understand the differences and verify whether the increase in accuracy outweighs the loss in explainability.

1.3 Document Structure

The structure of this report is as follows: in chapter 2 the data used for this project is presented along with variables used for modelling. Additionally, some restrictions coming from the data are addressed. chapter 3 explains Extreme Gradient Boosting theoretically. Further methodologies related to modelling are clarified. chapter 4 and 5 introduce Medical and Transportation expenses and their predictions. chapter 6 presents the main components of the recommendation system and at last chapter 7 contains conclusions and future work.

Chapter 2

Data

In this chapter an introduction is made to the datasets, the methods used for wrangling the data and the main constraints found are discussed.

2.1 Introducing the main datasets

For the construction of the multiple models employed throughout the project there are two main datasets. The first dataset contains all the information regarding the accidents as well as the information about the injured and the second dataset contains the information of the receipts - amount, entity providing care, relevant dates and nature of the expenses.

The first dataset includes all accidents in the line of business of Workers' Compensation except where *Fidelidade* is not the lead co-insurer, meaning, contracts where *Fidelidade* is not the main retainer of risk. The study includes all such contracts from 2007 until 2018. This applies to standard employed workers as well as self employed workers. Only claims that have been deemed closed are considered, which is to say that the cost of the claim is not expected to increase. Over the course of twelve years nearly 600 000 accident reports were accumulated. However due to missing data, opened cases and further restrictions we will consider 280 000 useful for modelling.

The second dataset logs all filed receipts and has over 6 million entries. These are the receipts of the claims, segmented by nature of the expense and include information regarding medical provider. Examples of natures of expense include medical expense, physiotherapy, and imageology exams.

Information was exported from *SAS Enterprise* but pre-processed in *R*. Models were developed in *R* and final results were analysed in *Excel*.

2.2 Variables

Variables can be separated in response variables and explanatory variables. We will introduce them and explain the pre-processing methods used.

2.2.1 Response variables

Response variables are the variables that we are modelling and that we deem important for recommending a healthcare provider. Since the recommendation is to be based on basic health care and physiotherapy we will be modelling frequency of healthcare treatments as well as their costs. In this context frequency is used to represent the number of different days that a patient needs to go to healthcare provider or physiotherapy centre, and so we will be modelling physiotherapy and general medical expenses separately. Our response variables are:

- Total cost of simple medical expenses;
- Total cost of physiotherapy (when physiotherapy is required);
- Proportion of physiotherapy;
- Number of days of out-patient visits to healthcare providers;
- Number of days of out-patient visits to physiotherapy centres.

WHere out-patients are patients who are not hospitalized while in-patients are hospitalized at the hospital.

There were some obstacles raised while building these variables. Due to the internal categorization of expenses, we had to restructure the categories to better fit the purposes of our research. This is further explained in chapter 4. As for the physiotherapy the number of sessions had to sometimes be inferred based on overall price of physiotherapy or on existing information of physiotherapy packages, because several expenses were wrongly logged. On our first approach we tried modelling number of physiotherapy packages, but soon realized that packages can contain different number of sessions. The information would also be distorted by physiotherapy sessions that were not attended as these are not paid.

Training accurate models requires big datasets with meaningful data, as such, we aggregated common medical expenses and built a separate model for physiotherapy. The aforementioned aggregation can be found in chapter 4.

Surgery and other expenses with lower frequency were removed, as models built would not be as reliable. Furthermore, when considering variables for the model, we want variables with complete information (meaningful data) that can explain the model. Thus, variables that were partially complete (missing data) were discarded.

As for the modelling of the transportation cost and frequency of in-patient appointments, a dataset from the portuguese postal office (CTT - *Correios de Portugal. S.A.*) [CTT*, Correios de Portugal, S.A* (2019)] was used for connecting zip codes with municipalities and districts, and from the expenses dataset we count the different days of service provided to estimate a number of visits to healthcare providers. A last dataset was built for distances between healthcare providers and injured's residency using *Google Maps* queries.

Lastly, the claims happen over 12 years and a major concern is the medical inflation. We took a first step by analysing the observed changes in price of each medical expense over the years and surprisingly most expenses had decreasing prices. That is, over the years the contracts drawn with healthcare providers accommodated the same treatments at lower prices. An explanation found is that a contract with *Fidelidade* provides a steady stream of new clients for healthcare providers allowing them to charge lower prices.

2.2.2 Explanatory variables

The explanatory variables can be split into social, demographic and pathological variables. For Extreme Gradient Boosting all input variables need to be encoded into numerical values. With this in mind, we can still use categorical variables, but first we must encode them into numerical ones.

Before going further into the variables that were used, we need to assimilate two basic concepts for transforming categorical variables into numerical variables. One-hot encoding is the process of using dummy variables to represent different levels. Every level of a variable is compared to a fixed level and thus we can encode any categorical variable with n level in to $n-1$ variables [Lantz

(2013)]. An example can be seen in equation 2.1.

$$\text{sex} = \begin{cases} 1 & \text{if sex is male} \\ 0 & \text{if sex is female} \end{cases} \tag{2.1}$$

This is an effective method for variables with few levels and when the steps between levels are the same, if variables are not nominal this should not be used. Thus, encoding the remaining variables requires different methods. We opted for target encoding, where a statistic - in this case the mean - of the target variable is taken for each level, as shown in [*Feature engineering I - Categorical Variables Encoding* (2018)]. In consideration of the extent of years in the study, the grouping for each level considers only the previous 3 years, which also proved useful to provide more consistent predictions in levels with few observations. To combine variables we group observations according to the different combinations of levels of both variables and compute the average value of the past 3 years. The algorithm is only slightly modified to accommodate this.

The social and demographic variables included in the model were sex, age, risk zone and salary of the injured, and the Economic Activity Code (CAE) of the company. The sex is a binary variable so using one-hot encoding we transform it into a numerical variable. In the case of the age variable we left the values as numerical. Salary was grouped according to a factor of the national minimum salary and used only to build combined variables. Risk zone is a variable built for another project at *Fidelidade* which we re-purposed. As for Risk Zone, municipalities were distributed according to the risk companies were exposed to in the LoB of WC. Portugal is divided into 16 Risk Zones. As for the Economic Activity Codes, there are over 1 000 CAE's thus they were organised into 18 clusters. This variable was also built for a different project at *Fidelidade*. Both these variables are good examples of situations where one-hot encoding would create an abundance of variable columns, but mean encoding can synthesise information into one column.

Companies register two main CAE's, even if involved in more activities. For this study we use only one of those CAE chosen at random. This is used to simplify the analysis however we recognise problems can arise. For example, in construction we can have an electrician and an upper management employee who are exposed to very different risks but classified as belonging to the same category. We attempt to mitigate the effect of a single CAE by creating a combined variable which includes salary and activity code, so that we can better identify the risk each worker is exposed to and hence refine the performance of the models.

As for pathological variables we have the pathology and the cause of the accident of the injured party. The pathologies are defined according to the *International Disease Code* (IDC9) [Organization (1978)], but further structured into group according to expert judgement. The cause of the accident is a categorical variable with values such as "falling of object". As in other categorical variables, they are encoded into numerical by target encoding.

2.2.3 Healthcare providers

An initial evaluation of an injured worker's condition is performed over the phone and a healthcare provider is recommended accordingly. This recommendation is a pivotal variable for this project. The first provider is of paramount importance as it evaluates the full extension of the damage and recommends the best fitting treatment.

The first step is finding the first viable healthcare provider for our sample. *Fidelidade* has a network of healthcare providers able to assist any injured worker close to their residency. The contracts are reviewed yearly and *Fidelidade* may change the contracts established with these centres. We were provided with a full list of healthcare providers, which included not only present, but also past medical centres. Furthermore, public hospitals and companies owned clinics (PMT - *Posto Médico Tomador*) are also part of this list, even though these are not possible recommendations as they are only used in particular occasions. To identify the healthcare provider of each injured worker we use the first healthcare provider that treated the patient, however when such institution is a public hospital or a PMT we retrieve next eligible healthcare provider so as to not discard any more data.

In some cases, doctors and special services of an hospital have their own healthcare provider identification number and so we aggregated them under a main healthcare provider. Each healthcare unit is accompanied with a classification according to the services available. When we have less than 1 000 injured workers treated in one health centre we pool them together with other healthcare providers under similar conditions. This reduces the levels of the variable and increases accuracy of predictions for small clinics for which we have less data.

For our recommendation, we considered 219 Physical Rehabilitation Centres and 350 general healthcare providers distributed throughout the Portuguese mainland.

2.3 Main Constraints

Wrangling the data proved to be a challenge as many variables had to be built from the available information. Examples of these are the transportation variables, such as number of physiotherapy sessions, or main healthcare provider. As for working with Healthcare providers and the nature of expenses presented we struggled with many adversities due to misclassifications. Overall, the biggest challenge was understanding which information was available and how to best make use of it.

It is worth noting that the project was developed in a computer with 8GB of RAM, and we were using datasets with several million entries so the whole process of wrangling data, training models and computing final results was time exhaustive and any small correction required long periods to run.

Chapter 3

Extreme Gradient Boosting

Artificial Intelligence is a field of Computer Science studying the creation of intelligence as seen on humans. As an example, machines that are programmed to learn without being explicitly told. Better known as Machine Learning. Machine Learning is a data analysis and algorithm development method that learns from data fed into the model. By continuously learning from new data, the algorithm is able to extrapolate and learn hidden features of the data.

The Machine Learning algorithm we used is Extreme Gradient Boosting. By boosting decision trees, the algorithm can solve regression or classification problems. Comparing to other gradient boosting algorithms XGB shows optimization for big data sets with sparse data, performs out-of-core calculations and increases the model performance which shows faster run times than other similar algorithms [Chen & Guestrin (2016)].

3.1 Models leading to Extreme Gradient Boosting

Extreme Gradient Boosting is one of the most advanced algorithms using decision trees. To fully understand it, we should first review some concepts that lead to its development.

The most basic element of XBG is a decision tree that can be used for classification or regression. Decision trees can be used as a prediction model. Fundamentally, in each node of a tree we make a decision, a variable is chosen and split, separating the observations. After several decision splits we have a decision tree with all observations distributed along the leaves. A statistic, such as the mode, is used to compute the final prediction. A singular tree although easily interpretable is prone to over-fitting [James et al. (2014)]. Tree learning algorithms are also good for handling missing data and ignoring redundant variables.

Single decision trees have predictions with high variance and low accuracy. Thus Breiman suggests a new improvement, Bagging or Bootstrap Aggregating. As the name indicates, random sub-samples are bootstrapped to generate different trees. A statistic such as the mode in classification or the average in regression is used to aggregate predictions. This methodology not only presents more accurate results, as it lowers the variance of the model. The main drawback is the loss in interpretability of the model [Breiman (1996)]. While in a decision tree we can explain the final result by following the path to the leaf, in bagging observations in different trees fall in different leaves so such a definite answer cannot be given. Even though bagging works for reducing the variance, the correlation between trees is still high.

Stochastic modelling presents the perfect solution to high correlation between trees by considering a sub-space of the feature space. That is each tree is built using only part of the available feature space, reducing correlation and improving accuracy [Ho (1995)]. Finally, we can look at a Random Forest as an ensemble of trees giving equal weight to each tree, which suggests an approach where the weights are optimized for a better model. This is the advancement made with boosting.

Boosting like the previous algorithms generates thousands of different trees, each tree on its own providing a bad estimator, but by carefully selecting the impact of each tree a more robust predictor is built [Friedman et al. (1998)] and [Schapire & Freund (2012)]. Boosting selects a base classifier, that is defined as any classifier better than random guessing which usually is weak by itself. This assumption, that a base classifier is weak by itself, is the weak learning assumption.

After building a classifier and analysing which observations were badly predicted we can build the next classifier giving more weight to the input space that had bad predictions. Before explaining boosting we will introduce two concepts, the training and validation loss.

3.2 Training and Validation loss

When building a model, an important step is the evaluation of the results. To verify the accuracy we compare the predictions to the observed values. Furthermore, we may want the results to follow a business objective and balancing these goals is part of the model building process. Thus, a first step in prediction is always to partition the data. By setting two samples, we can train the model in the first data and then use new data (unknown to the model) to predict. This methodology uses two losses, the training loss when training the model on the majority of the data and the validation loss when applying the model to unseen data.

Training loss, also known as the objective function, matches the business objective. For example, when modelling costs the distribution of the predictions has to be similar to the distribution of the data. In Actuarial Science, it is usually assumed that costs follow an heavy tail distribution, such as a Gamma, and frequency follows a discrete distribution such as Negative Binomial or Poisson Bahnemann (2015). For our data in particular we found that costs were closely represented by a compound Poisson-Gamma and that frequency of medical sessions was similar to a Generalized Poisson. For the current project, Training loss is optimized by maximizing the likelihood functions of the corresponding distributions.

Validation loss evaluates the performance of the model. During training we have instructed our model to follow a distribution function and in testing we want to make sure it resembles the sample as closely as possible. The loss applied varies according to the type of problem at hand. In classification the Area Under the Curve (AUC) is usually the standard, and for regression the Root Mean Square Error (RMSE) is used.

3.3 Reviewing Tree Boosting

The premise for a typical supervised learning problem is a sample of n observations with m features, $\mathbf{x}$ is the feature vector, and the response or target variable is y. That is $(\mathbf{x}, y)$, where $\mathbf{x} \in \Re^m$ and $y \in \Re$. During training we build a predictor with a part of the sample leaving the rest for assessing the accuracy of the predictions in a sample unknown to the model.

Boosting is an additive model [Friedman (2000)], where the predictor is obtained by an iterative improvement over the last predictor. Suppose our tree ensemble model has K additive functions, then:

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_t(\mathbf{x}_i),\ y = 1, \ldots, n \tag{3.1}$$

And summing over all k,

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i) \tag{3.2}$$

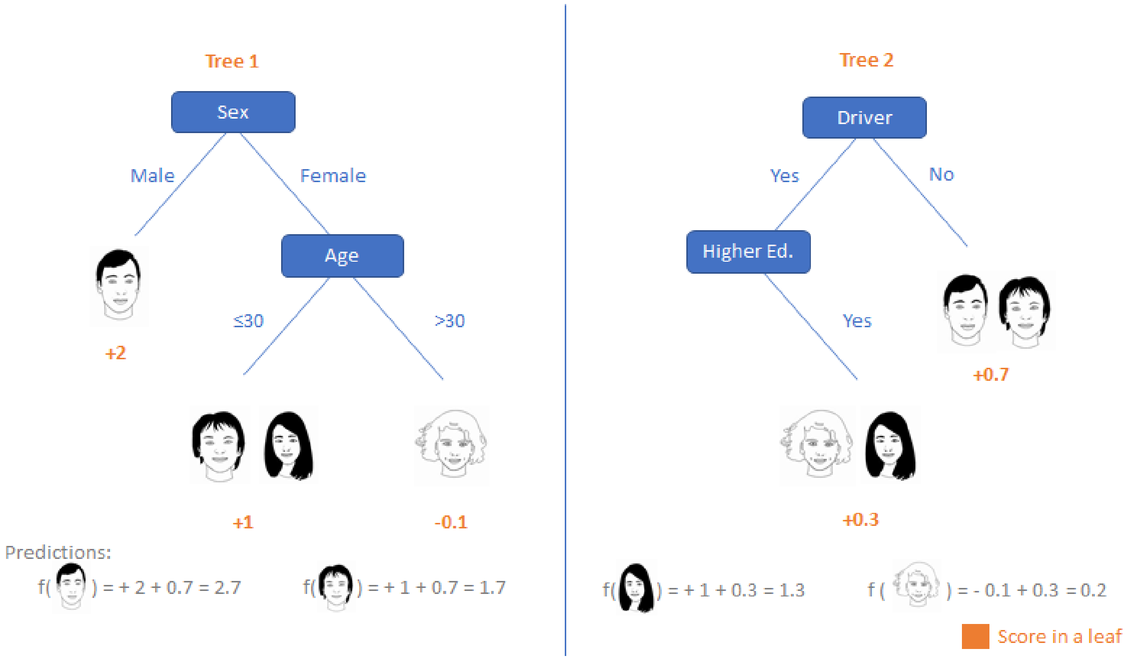


Figure 3.1: Scoring of observations, by summing predictions of individual trees to calculate the final prediction.

Each f_k is an independent decision tree with T leaves and weights ω on the leaves. For regression, the leaves contain a continuous score, $\omega_i \in \Re$. Each observation is classified by the tree according to the decisions at each stump and a final prediction is obtained for each observation by summing the score of the leaves in each tree, as shown in figure 3.3.

The set of rules of functions used by the model is determined initially by minimizing the following objective.

$$Obj = \sum_{i} l(y_i, \hat{y}_i), \tag{3.3}$$

where l is a fitting convex loss function. This measures the accuracy of the model by comparing the predictions $\hat{y}$ with the target variable y. Afterwards a penalty to the model, $\Omega(f_k)$, is applied to avoid complex set of functions. For this model, L2 regularization is applied [Ng (2004)], where a penalty is added to the loss function as the square of the weights/scores. This means larger scores on a leaf have a bigger penalty, which spreads the decision more evenly among all trees to avoid over-fitting. Without this regularization we would have traditional gradient tree boosting problem.

$$\begin{aligned} Obj &= \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) && (3.4)\\ &= \sum_i l(y_i, \hat{y}_i) + \sum_k \gamma T + \frac{1}{2}\lambda ||\omega||^2, && (3.5) \end{aligned}$$

where γ represents a penalty that is added to the model to prevent using complex trees with lots of leaves.

The variables for this model are equations and as such, the optimization is done stepwise, $y_i^t = {y_i}^{(t-1)} + f_t\left(\mathbf{x}_i\right)$, then at step t, we have that the objective is given by:

$$Obj^{(t)} \;=\; \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t)}\right)\right] + \Omega\left(f_t\right) = \sum_{i=1}^{n}\left[l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))\right] + \Omega\left(f_t\right) \tag{3.6}$$

Applying a second order Taylor Series approximation to the loss function, equation 3.6 can be written as:

$$Obj^{(t)} \;=\; \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right)\right] + \Omega\left(f_t\right), \tag{3.7}$$

where g_i and h_i are, respectively, the Gaussian and the Hessian functions of the loss function. Since we are minimizing the loss function the constant terms can be dropped, arriving at:

$$Obj^{(t)} \;\propto\; \sum_{i=1}^{n}\left[f_t(\mathbf{x}_i)\right] + \Omega\left(f_t\right) \tag{3.8}$$

Let j represent the leaves of a tree. Each observation i falls in one and only one leaf j of a tree. The value of the predictor for a observation i in a tree is going to be the weight of the leaf j. Let I_j denote the set of possible leaves for each observation. Then:

$$\bar{O}bj^{(t)} \;=\; \sum_{j=1}^{T}\left[\omega_j\left(\sum_{i\in I_j} g_i\right) + \frac{1}{2}\omega_j^2\left(\lambda + \sum_{i\in I_j} h_i\right)\right] + \gamma T. \tag{3.9}$$

To compute the optimal weights in each leaf, we need only minimize.

$$\frac{\partial \bar{O}bj}{\partial w_j^*} = 0 \iff \sum_{i \in I_j} g_i + \omega_j^* \left(\lambda + \sum_{i \in I_j} h_i \right) = 0 \tag{3.10}$$

$$\iff \omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\lambda + \sum_{i \in I_j} h_i} \tag{3.11}$$

For a fixed tree structure, the optimal value is obtained by substituting (12) in (10). Then:

$$\bar{O}bj^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\lambda + \sum_{i \in I_j} h_i} + \gamma T. \tag{3.12}$$

To compute the loss reduction obtained by an individual split, we split the set of leaves I. Each node, can either go to left, I_l, or to the right I_r, such that $I = I_l \cup I_r$. By applying the following formula, we evaluate possible split candidates.

$$L_{split} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_l} g_i \right)^2}{\lambda + \sum_{i \in I_l} h_i} + \frac{\left(\sum_{i \in I_r} g_i \right)^2}{\lambda + \sum_{i \in I_r} h_i} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\lambda + \sum_{i \in I} h_i} \right] - \gamma \tag{3.13}$$

3.4 Split Finding

On equation 3.13 several splits are considered to minimize the gradient, but we have not seen how the split points are chosen. Although computationally expensive, the best method is a greedy search going over all the points. If the computer does not have enough memory to perform the greedy algorithm, then we can use a quantile search. For either algorithm, the data in each feature column is sorted and stored separately. These feature columns are equal for all trees, so by storing this information the model runs faster. For each feature space, quantiles that give the lowest gradient increase are selected.

A major road block in split finding is sparse data. Sparse data can be mainly caused by missing data or, in some cases, frequent zeros in data. XGB handles sparse data by choosing a default direction to follow in a split with missing data. For each feature, the algorithm learns which direction brings the lower increase in gradient.

3.5 Meta parameters

The model gives the user a lot of flexibility in the adjustment of meta parameters. There are two main categories of parameters explored throughout the modelling phase - tree boosting and learning task. Boosting parameters help control over-fitting and overall accuracy while learning task define the objective to be optimized and the evaluation metric (Training and Validation loss). Table 3.1 presents the final parameters chosen.

Tuning requires analysing variations in several parameters and be careful with dependencies and relationships between different parameters. As such we restricted our tuning to eta, maximum depth, gamma, sub-sample, column sample by tree and lambda. Eta is the learning rate of tree, this is every prediction from a tree is multiplied by a factor to reduce its impact. Max depth sets the depth of trees and is used to control over-fitting, if a tree goes to deep it will learn relationships that are too specific to a particular sample. Gamma specifies the minimum loss for a split; if a split does not reduce the gradient by the specified value then that is a terminal node. It is worth noting the algorithm does prune the tree at the end removing any split that did not add enough value. Column sample by tree and sub-sample are the proportion of the feature and sample space used in each tree, respectively. Lambda controls L2 regularization on terms, as seen on equation 3.5.

As for comparing different models two main evaluation metrics were tested, Root Mean Square Error (RMSE) Chai & Draxler (2014) and Area Under a Receiver Operating Characteristic Curve Fawcett (2006). RMSE is advised for Regression Models while AUC is appropriate for Classification.

3.6 Tweedie

For a given variance-power parameter, p, a Tweedie distribution belongs to the family of the exponential dispersion models. The Tweedie distributions have a special mean-variance relationship. Given a random variable $Y \sim Td(\mu, \sigma^2)$, that is mean and dispersion parameter σ^2, its variance is given by

$$Var(Y) = \sigma^2 \mu^p \tag{3.14}$$

Meta Parameter	Average Cost Physio	Average Cost Medical	Proportion of Physio	Number of Physio Sessions	Number of Medical Sessions
Max Depth	4	3	3	3	3
Col-sample by tree	0,8	0,8	0,8	0,8	0,8
Sub-sample	0,8	0,8	0,8	0,8	0,8
Eta	2	2	1	2	2
Gamma	2	2	2	2	2
Objective	Tweedie	Tweedie	Bernoulli	Poisson	Poisson
Evaluation Metric	RMSE	RMSE	AUC	RMSE	RMSE
Tweedie Variance Power	1.25	1.35	NA	NA	NA

Table 3.1: Meta parameters tested and values chosen for training models

Where $p \in$ is the variance-power parameter. By changing p different distributions are obtained. Namely:

- $p = 0$, Normal distribution
- $p = 1$, Poisson distribution
- $1 < p < 2$, Compound Poisson-Gamma distribution
- $p = 2$, Gamma distribution
- $p = 3$, Inverse Gaussian distribution

We take special notice of the Compound Poisson-Gamma, we can use it to model total claim cost where it is assumed Poisson distributed arrival of claims and and Gamma distributed claim amounts [Jørgensen & C. Paes De Souza (1994)].

3.7 Generalized Linear Models

To be able to assess if employing more advanced models is necessary, we build the models for medical costs also using GLMs. As GLM is a well known topic we will not present a theoretical introduction. However, for those interested we recommend consulting [McCullagh & Nelder (1989)] for a theoretical approach and [Goldburg & Tevet (2016)] for a practical overview.

Chapter 4

Modelling Treatment costs

In this chapter we will apply the model described previously to predict costs incurred from treatments. We start by understanding how receipts are kept and how to best use them for modelling. Once the costs are aggregated, a model is built using GLM and XGB and an analysis of the results obtained in each model is provided.

Fidelidade organizes work accident expenses into categories according to their nature. These can be:

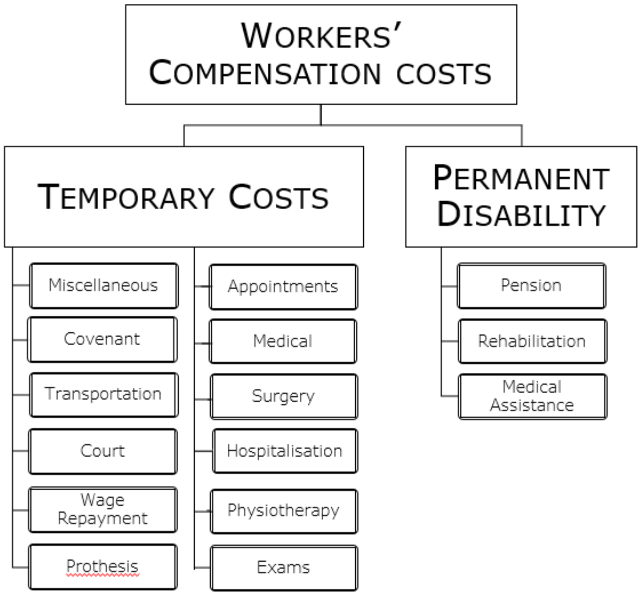


Figure 4.1: Diagram of expenses in Workers' Compensation

Covenants are fees paid to the healthcare provider for a package of treatments. There are three main aspects in a Covenant: period of coverage, extension of treatments and premium. The covenant for each healthcare provider is unique, updated on a yearly basis and it is assumed the premium is designed as an average of the cost of the covered treatments.

To each new injured worker, *Fidelidade* attributes a ”process number” under which we can aggregate receipts to obtain the full cost of each claim and this is the key for connecting with other data tables. We are interested in costs related to medical expenses and since we are creating a general recommendation system, we will predict only common medical costs. Figure 4.1 decomposes medically related expenses in its several components.

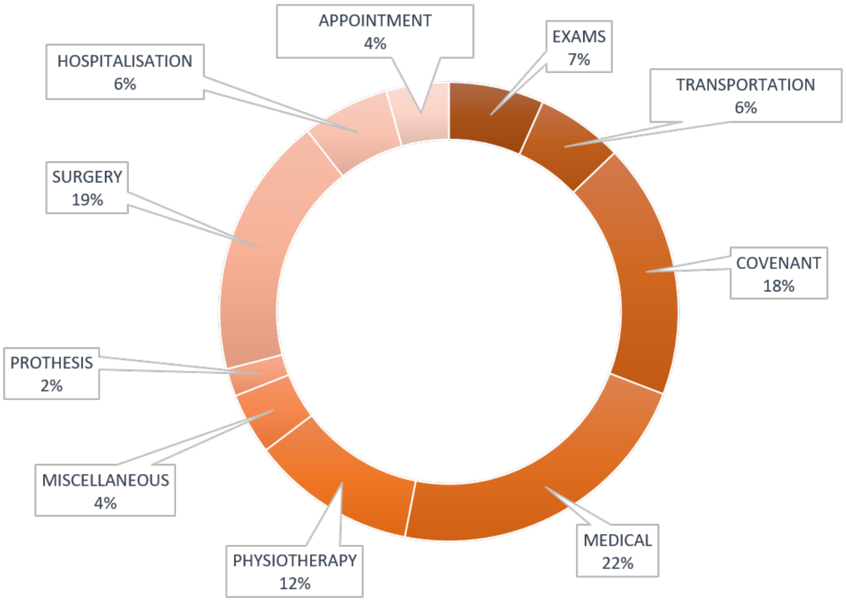


Figure 4.2: Distribution of medical expenses.

Besides all the medical costs, another big influence on the total cost is the wage repayments from a work leave. After an accident, the worker might have to rest or attend treatments during the day. This situations are classified, respectively, as absolute and partial temporal disabilities. During this project these costs are not considered, however, they are acknowledged and present a possible improvement on the developed models.

4.1 Medical Aggregate Expense

In order to correctly model each cost category, a main assumption concerning the treatments covered in each covenant contracts has to be set. If coverages varied we could not draw a comparison of costs between different healthcare providers. Thus, to set a fair and equal starting ground, all medical expenses that are covered in the most in-depth contract are aggregated under a new expense. This new category is the Medical Aggregate Expense and can be computed as:

$$\begin{array}{c} Medical \\ Aggregate \\ Expense \end{array} = Covenant + Appointments + Medical\ Expenses + Exams, \tag{4.1}$$

where Exams expense exclude X-rays and CT scans which are never covered in covenant contracts.

For the modelling of the Medical Aggregate Expense we used Extreme Gradient Boosting with a Tweedie distribution as the training loss and Root Mean Square Error as the validation loss.

During the modelling process we run tests on several meta parameters adjusting according to the observed errors. The final parameters can be seen on table 3.1. As for the variables we performed a final test by consulting one of the outputs of the model, the importance of the variables, which is measured as the total sum of the reduction in likelihood of using a variable in each split. The following figure is an example of the outputs, the remaining model summaries and plots of variable importance can be explored in the Appendix A.

Before analysing the results we should first understand what each figure shows. On the left side, the plots compare observations and predictions to understand overall fitness, while on the right an analysis of the errors is performed. We would like to draw the attention of the reader to the first plot (top-right), a scatter plot of observation-predictions where we can observe the variance in predictions. On the bottom right of this figure 4.3 different error measures are displayed (Adjusted R-squared, RMSE and Mean Absolute Error). These metrics are the main criteria for choosing the final model. On the bottom left, the plot displays the densities of predictions and observations. It is desirable that the shape of the densities are similar so that a new observation can be accurately predicted. As for the right side of the figure, we can study the residuals of the models. Perhaps the most useful is the error density map. If the curve is asymmetric then predictions are biased to be higher or lower than observations.

We are now ready to compare models built using GLM and XGB methodologies. Figures 4.3 and 4.4 can be used to compare both models. The shape of densities in the values distribution plots are very similar, thus the decision was made based on the error measures.

Figures 4.3 and 4.4 model the Expected Medical Costs using XGB and GLM,respectively. In concrete terms, XGB provide a 100% better R-squared, 12,5% lower RMSE and 11% Mean Absolute Error. All these measures prove that for this problem XGB is the best choice.

During training we selected the explanatory variables with higher importance as these have more explainability of the results and discard the remaining. This importance plots for the variables of each model can be consulted on the Appendix. Figure 4.3 and 4.4 present the predictions from the models.

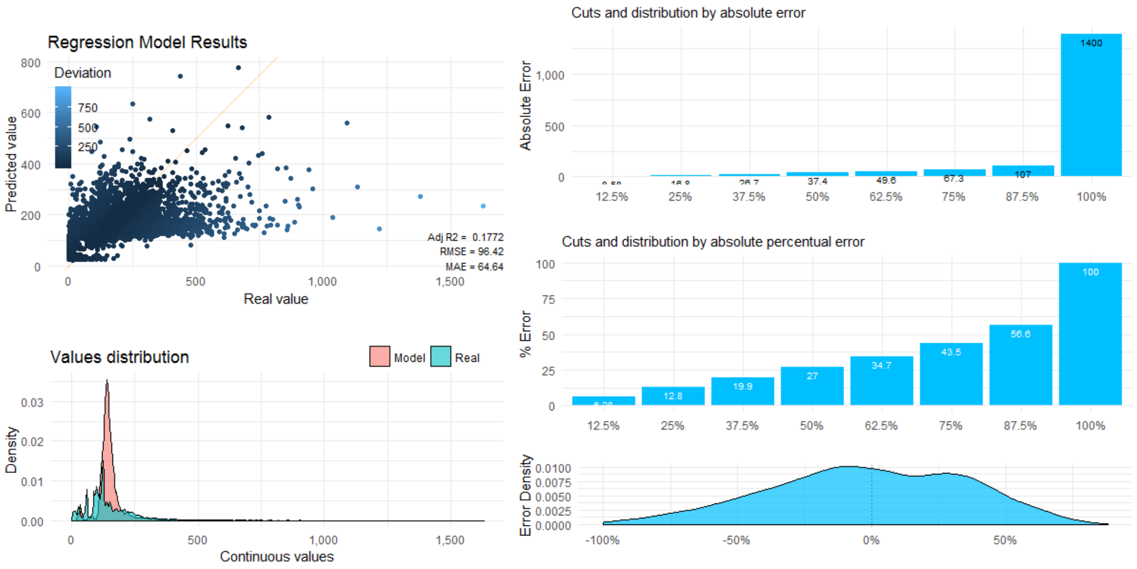


Figure 4.3: R output from Expected Medical costs with XGB

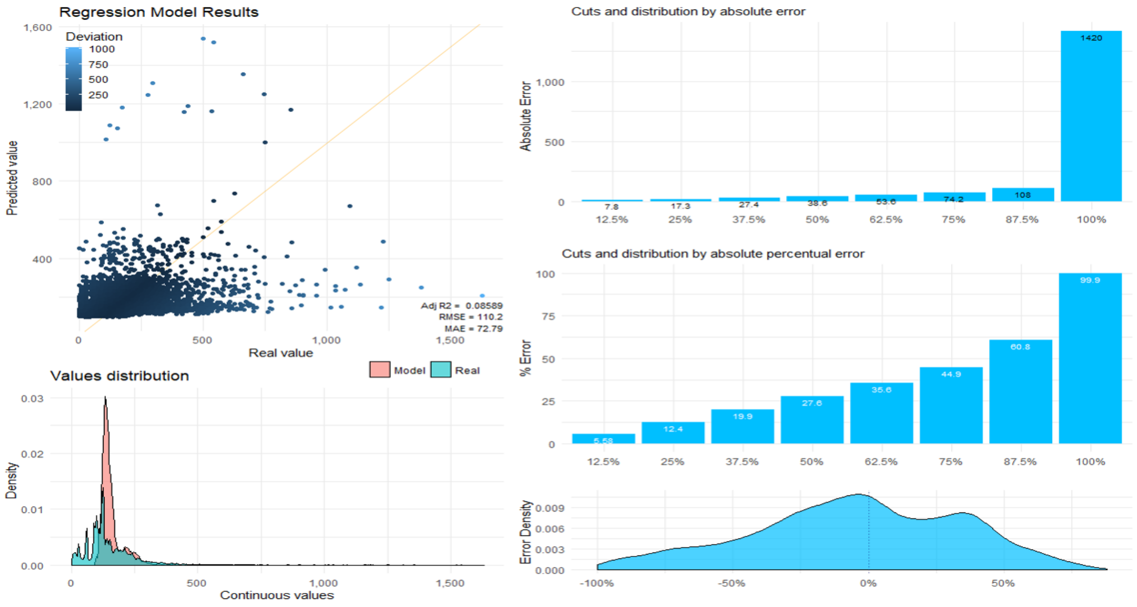


Figure 4.4: R output from Expected Medical costs with GLM

4.2 Physiotherapy

Physiotherapy is the physical rehabilitation of the injured, which includes hydrotherapy, and its taken in sessions up to 20 sessions. For a full recovery the injured worker might need to undertake physiotherapy more than once, in which case the cost is aggregated to the the first package of costs. Physiotherapy is an easy addition to the model, the proportion of physiotherapy is high (around 25%) and the physiotherapy centres are usually independent from other health clinics which means the recommendation system for physiotherapy is separate.

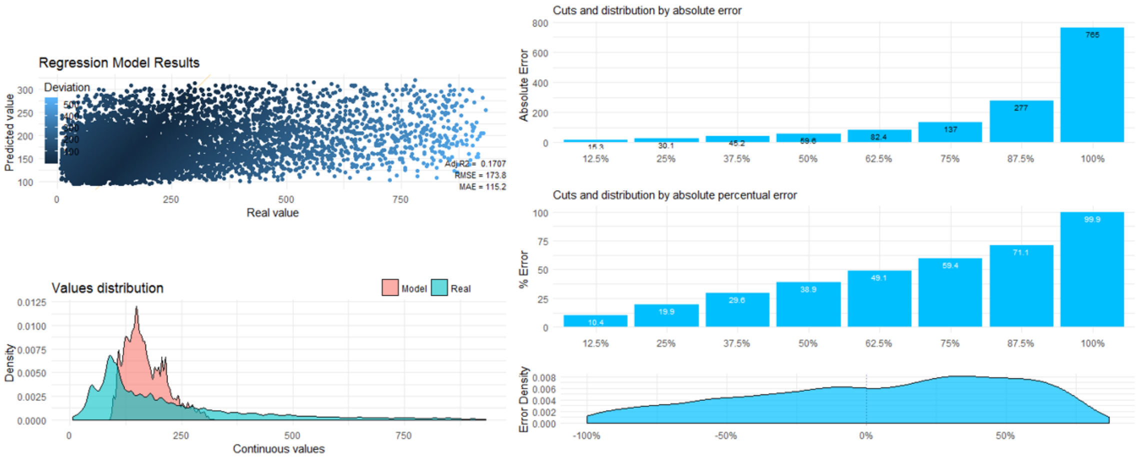


Figure 4.5: R output from Physiotherapy costs

For physiotherapy we are building two models - the proportion of the physiotherapy and the average cost of the treatment. The models applied were Extreme Gradient Boosting with a validation loss of, respectively, a Bernoulli and a Tweedie and the testing losses were Area Under the Curve and Root Mean Square Error.

The procedure for modelling is similar to the Aggregate Medical Costs, the meta parameters can be consulted in table 3.1. and the results on figure 4.5.

Chapter 5

Transportation Costs

Transportation costs are incurred everytime an injured requires medical attention and it is the insurance companies responsibility to facilitate it. Under Article 40 of the Workers' Compensation Legal Regime, transportation has to be made available to and from the treatment centre [APS*, Associação Portuguesa de Seguradores* (2017*a*)].

The typical means of transportation are buses, taxis, private cars, ambulances and, in rare occasions, airplanes or helicopters. Ideally, workers use public transportation or travel on their own car and are later refunded for their expenses, however the prices per km are unknown. Transportation by taxi is provided when the injured worker is unable to drive himself. For the remaining of the study we assumed transportation by taxi since prices are fixed, so estimates will be consistent even though they are probably disproportionate to reality.

The cost of transportation will be estimated as a product of the number of kilometers between injured and healthcare provider by the number of necessary trips and the price per kilometer.

$$\text{Cost of Transportation} = \text{Number of km} \times \text{Number of trips} \times \text{Price per km} \tag{5.1}$$

Over the next sections we will analyse each factor and explain its modelling.

5.1 Number of Kilometers

The number of kilometers between two points is given by the driving distance. While querying google for the distance between two points is feasible for a small sample, when we considered the size of our sample deriving a formula to estimate this distance was more realistic.

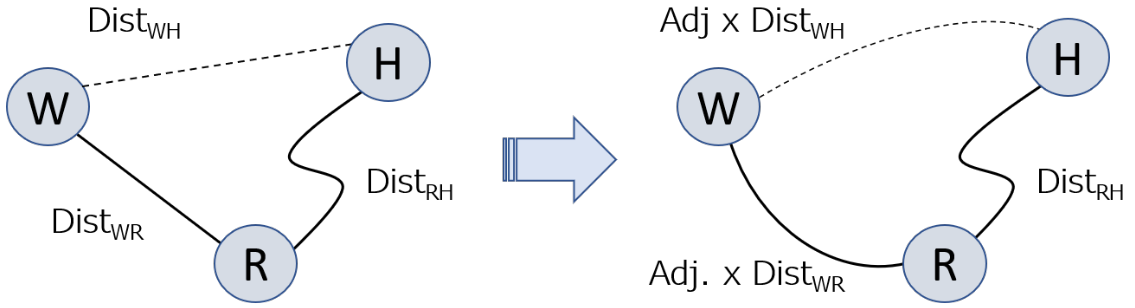


Figure 5.1: Adjustment factor and derivation of driving distance between injured workers' residence and health clinic

Then the idea is to use a reference point in each municipality as a halfway point from a new injured to the healthcare centre, as shown in figure 5.1. The distance is then estimated as an average between the adjusted euclidean distance and the driving distance. Let W represent the injured worker residence, R the reference point and H the healthcare provider.

$$\text{Estimated Distance} = \frac{Dist_{WH} \times \text{Adj. factor} + (Dist_{WR} \times \text{Adj. factor} + Dist_{RH})}{2}, \tag{5.2}$$

where $Dist_{ij}$ represents the distance between i and j. The adjustment factor is given by the following average:

$$\text{Adj. Factor} = E\left[\frac{\text{Driving Distance}}{\text{Euclidean Distance}}\right] \tag{5.3}$$

Distances are computed from all injured workers' residences to all healthcare providers locations, these are the healthcare providers described in chapter 2. Healthcare providers and injured in the Autonomous Regions of Açores and Madeira were removed as they are not connected to Mainland Portugal by road.

5.2 Number of trips

Two models had to be built - one for the medical sessions and another for physiotherapy. Medical expenses will englobe medical appointments, speciality appointments and out-patient expenses. Adding all bills registered in separate days we have the total number of necessary rides. On the other hand, as for physiotherapy the response variable is just the number of sessions.

Extreme Gradient Boosting will again be used, this time we are modelling frequency, so the loss function is a generalized Poisson. Input variables need to be numeric so the same encoding is applied.

5.3 Price per Kilometer

The price per kilometer depends on the number of kilometers driven and the injured workers' home district.

A fee will be added according to the waiting time of the taxi driver. This fee is paid in periods of 30 minutes. It is assumed that the average waiting is 1 hour and 30 minutes, corresponding to an increase of 10 euros per trip.

Chapter 6

Recommendation Systems

A Recommendation System provides a suggestion for a user. In our project we are interested in suggesting healthcare providers to injured workers. By creating a scoring function we obtain a ranking of healthcare providers personalized to each user. Currently any injured worker is recommended to a partner facility based on initial assessment of pathology and residency's municipality. By considering an estimation of the medical costs and adding a more accurate prediction of travel expenses we rank health centres according to their costs. The model gives a score to each healthcare provider and recommends to each injured worker an optimal healthcare provider. Currently any injured worker is recommended to a partner facility based on initial assessment of pathology and residency's municipality. It was observed that for more serious injuries an ambulance is called taking the injured to a public hospital and only when the injured is stable is he/she then transferred to a partner healthcare provider.

Our objective is to find the healthcare provider with the lowest cost. To get an estimate of the cost of each one, we estimate the costs of each healthcare provider with the previously trained model. The estimated cost of travel is added to obtain a final cost estimate. For each injured worker we minimized the cost and found the optimal healthcare provider.

From a business perspective *Fidelidade* built central medical units to assist work accidents. These units have invested in doctors specialized in many pathologies, thus providing better diagnostic, assessment of injury and ultimately lower incidence of Permanent Disability. There is higher attention to identifying chronic disease to avoid future liabilities. These are all advantages in choosing the central units that are hidden from the model.

Two scenarios were developed, unrestricted recommendation and segmented in hospital/health clinics. An unrestricted recommendation can suggest an injured to a clinic even when the pathology is severe. The second scenario segments the healthcare sample into hospital and clinic and recommends within the same class that the current model would recommend.

6.1 Scenario I - no restrictions

In the first scenario we apply no restrictions, injured can be recommended anywhere. Initially, we compare our models recommendation to the current recommendation. However, to understand the degree of difference between predictions we relax our restrictions and start recommending the same facilities as the current model if the variation between predictions is lower than some percentage. Table 6.1 displays the results for variations of 5, 10, 15, 20 and 25%. This contains data from 178 642 previous work accidents. The actual recommendation is unknown, so a proxy is used based on what the current system would recommend. First column informs the reader of the main statistics when our model would give the same recommendation. Afterwards, each column considers a different variation of prices for which the recommendation would remain the same. That is, in the second column where a variation of 5% is considered, if we predicted hospital A to be only 4% cheaper than the current recommendation of hospital B, then we would still recommend hospital B.

	Base Rec.	**Rec. with 5% var.**	**Rec. with 10% var.**	**Rec. with 15% var.**	**Rec. with 20% var.**	**Rec. with 25% var.**
Count	55 956	69 965	81 341	94 214	102 594	111 267
Percentage (%)	31,46	39,34	45,74	52,98	57,69	62,57
Portfolios Expected Gain (€)	6 926 834,98	6 882 867,49	6 766 952,45	6 550 689,59	6 345 736,84	6 069 785,21
Annual Expected Gain (€)	2 287 722,17	2 273 201,05	2 234 917,85	2 163 492,83	2 095 803,19	2 004 664,79

Table 6.1: Global Summary of scenario I

6.2 Scenario II - Conditional recommendation

The base scenario has an intrinsic flaw, each healthcare provider is limited by the number of patients a health clinic can accommodate. Smaller clinics do not have the dimension to be one of the main health providers. To reduce the probability of overburdening the medical units the patient attribution is similar to the current model - if an hospital would be recommended, then we will recommend an hospital. Otherwise, we recommend a health clinic.

For this scenario we classified the recommendation in production in hospital/Clinic and recommend Hospitals if it would usually recommend hospitals and clinics otherwise. The main problematic with recommending too hany injured to clinics is the patient capacity of medical units. While an hospital might be able to support a big surge in injured workers, a small clinic can easily be overrun. Table 6.2 displays the results from the aforementioned variations in prediction values. Overall, we can see that an increase in similarity with the current recommendation comes at a small reduction in gain.

	Base Rec.	**Rec. with 5% var.**	**Rec. with 10% var.**	**Rec. with 15% var.**	**Rec. with 20% var.**	**Rec. with 25% var.**
Count	58 709	71 337	81 024	91 108	98 333	107 528
Percentage (%)	35,00	42,52	48,30	54,31	58,62	64,10
Portfolio's Expected Gain(€)	6 030 780,75	5 987 883,06	5 887 623,13	5 711 750,91	5 522 338,13	5 218 406,28
Annual Expected Gain (€)	1 991 782,81	1 977 615,01	1 944 502,21	1 886 416,99	1 823 859,73	1 723 480,31

Table 6.2: Global Summary of scenario II

Chapter 7

Conclusion and Future Work

7.1 Conclusions

After analysing and pre-processing all information available we had to decide which cost categories would be modelled. All medical expenses with high frequency were aggregated for modelling. We used Extreme Gradient Boosting which showed clear advantages over GLMs.

Following the medical costs, we model the transportation costs, by estimating the distance Injured Worker - Healthcare Provider and the number of trips. The transportation is assumed to be provided by taxis where the price per kilometer is fixed. When the healthcare providers recommended by both models are different, there is on average a 5 kilometer decrease from choosing our recommendation which translates into saving approximately 14 euro per injured worker in transportation alone.

Two recommendation systems are built. On a first attempt, no restrictions are applied to patient attribution. The scoring is based only on the cheapest overall cost. However, we recognise that not all healthcare centres are equipped to perform, for example, surgery or accommodate burn victims. Thus a second model is created, where healthcare providers are split according to size and functions in hospitals and health clinics. The cost predicted on this second model are more realistic, the expected yearly gain is still high and estimated at 1,7 million €.

7.2 Future Works

Many improvements, tests or even methodologies have been left out due to time constraints. Running the currents models is a time consuming process and the methodologies applied were thought out with possible improvements in the future.

First and foremost, a global estimate of expenses incurred from a Temporal Disability has to be computed. For this work we focused on basic medical expenses, physiotherapy, imageology and transportation costs leaving out salary repayments for recovery days, pharmacy expenses, surgery/ hospitalisation and prostheses. For salary repayments we would like to model the number of days with partial and/or absolute disability and build an estimate by multiplying by the repayment factor and the daily wage. As for the other expenses we would model them separately and each cost would be added to a final estimate.

Models for regression and classification are always being updated and there are some models that could have been tested even in the area of boosting that for some problems have reported better results than XGB, namely *Catboost* [Dorogush & Gulin (2018)]. *Catboost* as in XGB applies gradient boosting with decision trees to build a predictor but provides an in-model categorical variable pre-processing which might be more adequate than our current encoding.

During model tuning, a last improvement could be made by adding external information from national databases. The idea is to combine the information in our entries, such as location or profession, with national statistics so that we can capture some explanation as to how our response variables work.

Bibliography

APS*, Associação Portuguesa de Seguradores* (2017*a*), '*Regime Jurídico de Acidentes de Trabalho Anotado - Article nº 40*'.
URL: *https://www.apseguradores.pt/Portal/ContentResourceDownload_Entry.aspx?ResourceId=15832*

APS*, Associação Portuguesa de Seguradores* (2017*b*), '*Regime Jurídico de Acidentes de Trabalho Anotado - Article nº 8*'.
URL: *https://www.apseguradores.pt/Portal/ContentResourceDownload_Entry.aspx?ResourceId=15832*

Bahnemann, D. (2015), 'Distributions for actuaries'.

Breiman, L. (1996), 'Bagging predictors', *Mach. Learn.* **24**(2), 123–140.

Chai, T. & Draxler, R. (2014), 'Root mean square error (rmse) or mean absolute error (mae)?– arguments against avoiding rmse in the literature', *Geoscientific Model Development* pp. 1247–1250.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, ACM, New York, NY, USA, pp. 785–794.
URL: *http://doi.acm.org/10.1145/2939672.2939785*

CTT*, Correios de Portugal, S.A* (2019), '*Códigos Postais*'.
URL: *https://www.ctt.pt/feapl$_2$/app/restricted/postalCodeSearch/postalCodeDownloadFiles.jspx*

Dorogush, Anna V Ershov, V. & Gulin, A. (2018), 'Catboost: gradient boosting with categorical features support'.

Fawcett, T. (2006), 'An introduction to roc analysis', **27**(8), 861–874.

Feature engineering I - Categorical Variables Encoding (2018). accessed on 23/07/2019.
URL: *wrosinski.github.io/fe_categorical_encoding/*

Friedman, J. (2000), 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics* **29**, 1189–1232.

Friedman, J., Hastie, T. & Tibshirani, R. (1998), 'Additive logistic regression: a statistical view of boosting', *Annals of Statistics* **28**, 2000.

Goldburg, Mark, K. A. & Tevet, D. (2016), 'Generalized linear models for insurance rating', *CAS monograph series* (5), 2–26.

Ho, T. K. (1995), Random decision forests, *in* 'Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1', ICDAR '95, IEEE Computer Society, Washington, DC, USA, pp. 278–.
URL: *http://dl.acm.org/citation.cfm?id=844379.844681*

James, G., Tibshirani, R., Witten, D. & Hastie, T. (2014), *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated.

Jørgensen, B. & C. Paes De Souza, M. (1994), 'Fitting tweedie's compound poisson model to insurance claims data', *Scandinavian Actuarial Journal* **1994**(1), 69–93.
URL: *https://doi.org/10.1080/03461238.1994.10413930*

Lantz, B. (2013), *Machine Learning with R*, Packt Publishing Limited.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Springer.

Ng, A. Y. (2004), Feature selection, l1 vs. l2 regularization, and rotational invariance, *in* 'Proceedings of the twenty-first international conference on Machine learning', ACM, p. 78.

Organization, W. H. (1978), 'International classification of diseases: ninth revision, basic tabulation list with alphabetic index'.

Schapire, R. E. & Freund, Y. (2012), *Boosting: Foundations and Algorithms*, The MIT Press.

Appendix A

Models and Variable Importance

In this Appendix, we display outputs from models introduced during the main body but not included on it. These are the models of frequency for Medical Expenses and Physiotherapy as well as for the proportion of physiotherapy.

Expected frequency of medical expenses was also modelled with GLM. By using boosted trees we were able to achieve a 13% reduction in RMSE.

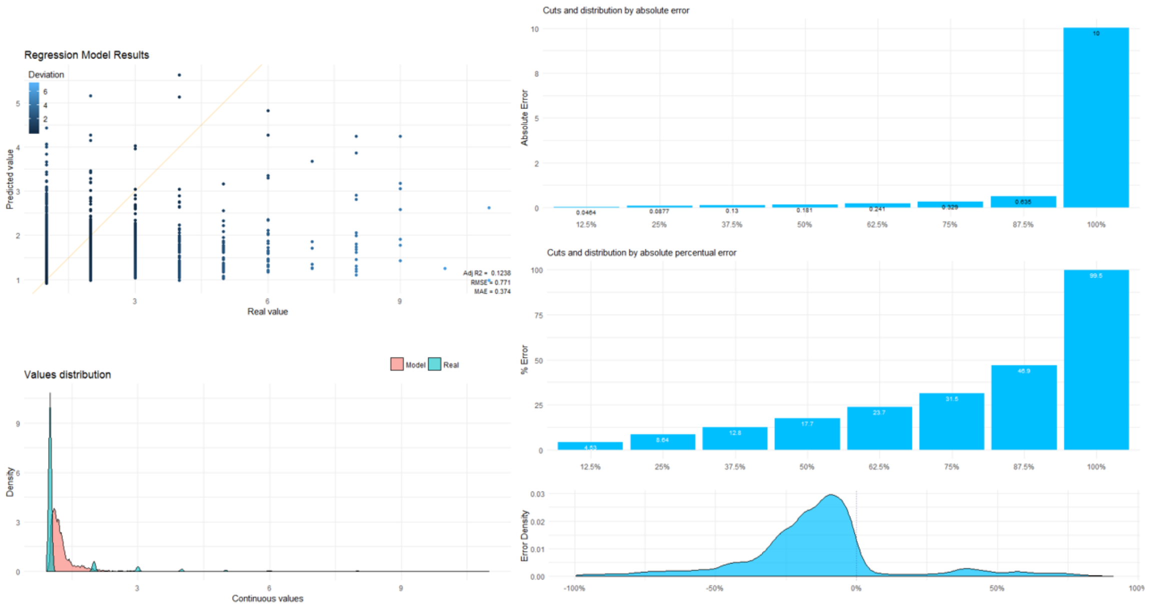


Figure A.1: R output from Expected Frequency of Medical Expenses in XGB

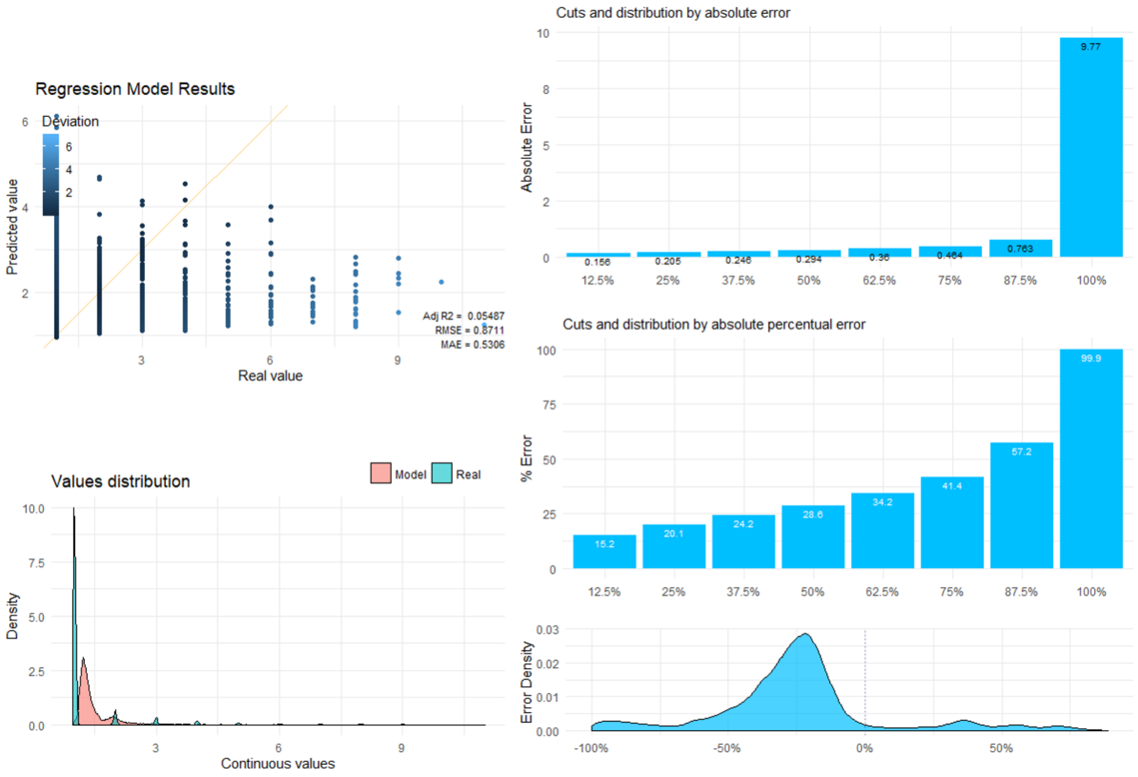


Figure A.2: R output from Expected Frequency of Medical Expenses in GLM

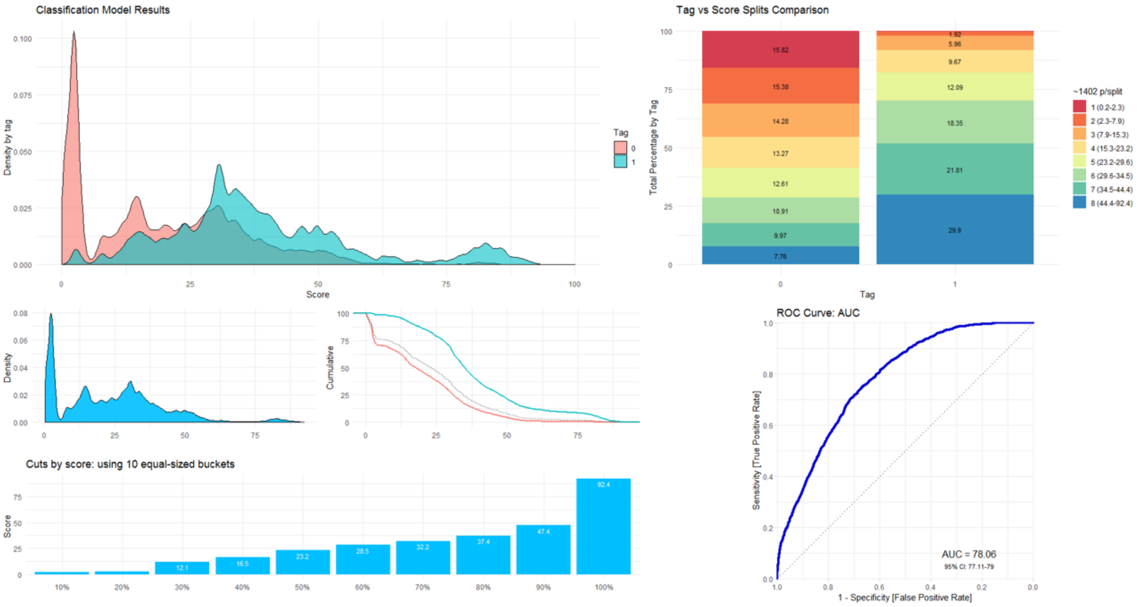


Figure A.3: R output for the proportion of physiotherapy with XGB

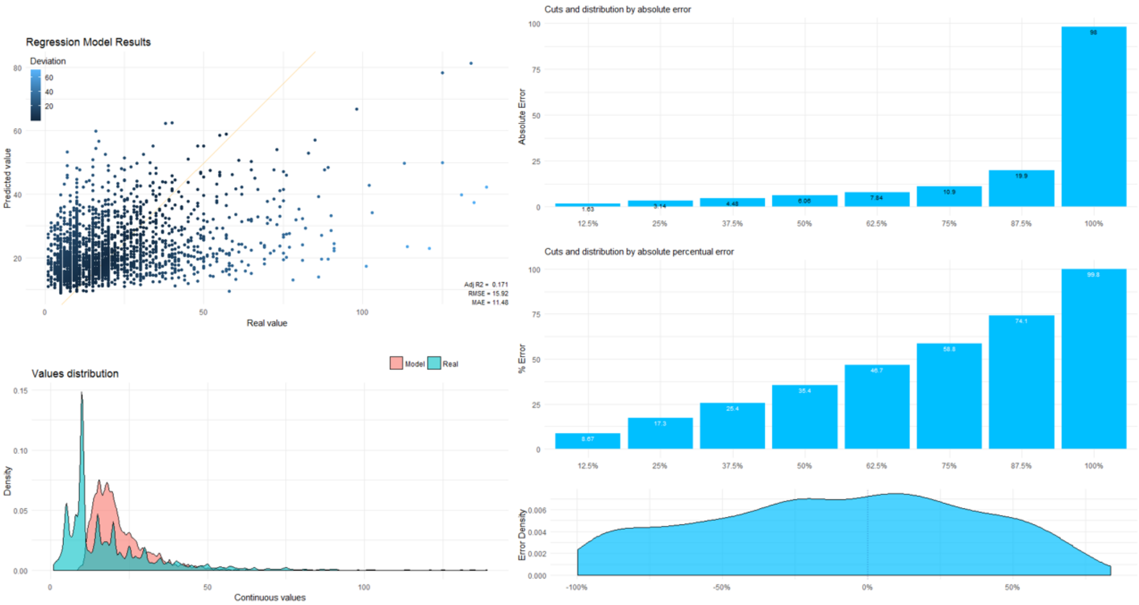


Figure A.4: R output from Expected Frequency of Physiotherapy Sessions in XGB

In Machine Learning, tuning is the process of adjusting variables and meta parameters to control over-fitting and accuracy of predictions. One of the observed outputs of the model is the Importance of each variable, this is calculated as the reduction in the validation loss from applying that variable in each split. The output is given as the percentage of the impact of each variable. Variables with a small percentage or impact can then be removed as their impact on the final prediction is negligible.

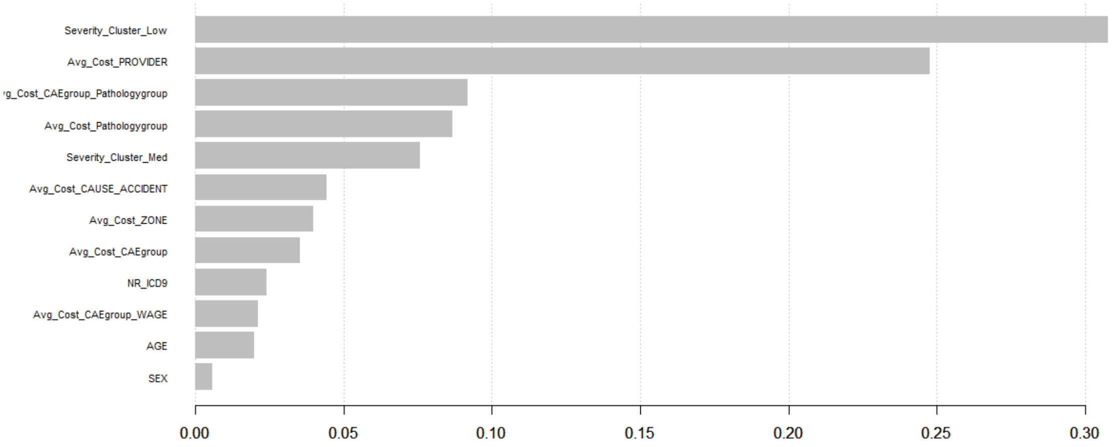
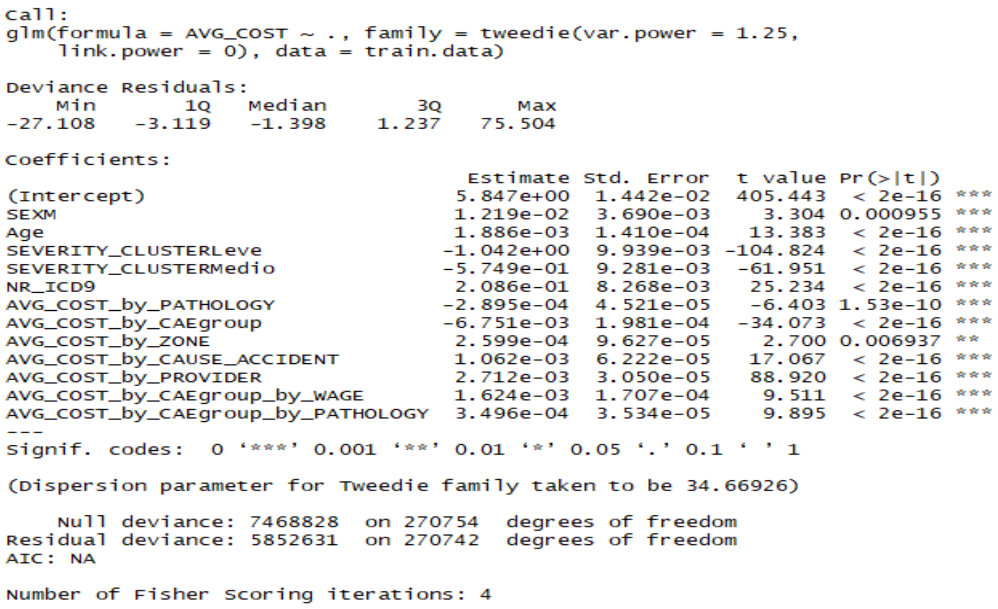
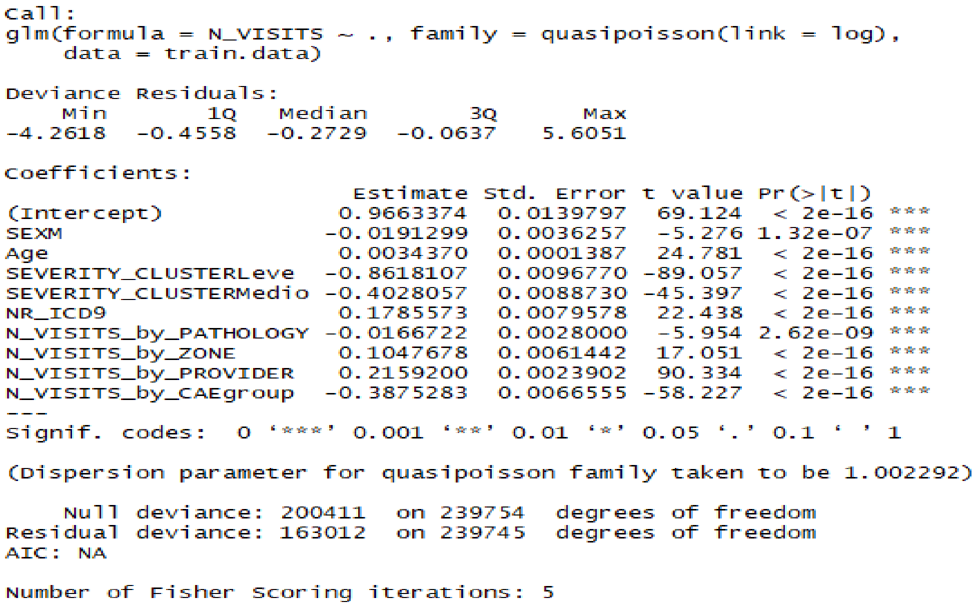


Figure A.5: Variable Importance of Expected Cost of Medical Expenses in XGB



```
Call:
glm(formula = AVG_COST ~ ., family = tweedie(var.power = 1.25, 
    link.power = 0), data = train.data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-27.108   -3.119   -1.398    1.237   75.504  

Coefficients:
                                    Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                        5.847e+00  1.442e-02  405.443  < 2e-16 ***
SEXM                               1.219e-02  3.690e-03    3.304 0.000955 ***
Age                                1.886e-03  1.410e-04   13.383  < 2e-16 ***
SEVERITY_CLUSTERLeve              -1.042e+00  9.939e-03 -104.824  < 2e-16 ***
SEVERITY_CLUSTERMedio             -5.749e-01  9.281e-03  -61.951  < 2e-16 ***
NR_ICD9                            2.086e-01  8.268e-03   25.234  < 2e-16 ***
AVG_COST_by_PATHOLOGY             -2.895e-04  4.521e-05   -6.403 1.53e-10 ***
AVG_COST_by_CAEgroup              -6.751e-03  1.981e-04  -34.073  < 2e-16 ***
AVG_COST_by_ZONE                   2.599e-04  9.627e-05    2.700 0.006937 ** 
AVG_COST_by_CAUSE_ACCIDENT         1.062e-03  6.222e-05   17.067  < 2e-16 ***
AVG_COST_by_PROVIDER               2.712e-03  3.050e-05   88.920  < 2e-16 ***
AVG_COST_by_CAEgroup_by_WAGE       1.624e-03  1.707e-04    9.511  < 2e-16 ***
AVG_COST_by_CAEgroup_by_PATHOLOGY  3.496e-04  3.534e-05    9.895  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 34.66926)

    Null deviance: 7468828  on 270754  degrees of freedom
Residual deviance: 5852631  on 270742  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Figure A.6: Variable Importance of Expected Cost of Medical Expenses in GLM



```
Call:
glm(formula = N_VISITS ~ ., family = quasipoisson(link = log), 
    data = train.data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.2618  -0.4558  -0.2729  -0.0637   5.6051  

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            0.9663374  0.0139797  69.124  < 2e-16 ***
SEXM                  -0.0191299  0.0036257  -5.276 1.32e-07 ***
Age                    0.0034370  0.0001387  24.781  < 2e-16 ***
SEVERITY_CLUSTERLeve  -0.8618107  0.0096770 -89.057  < 2e-16 ***
SEVERITY_CLUSTERMedio -0.4028057  0.0088730 -45.397  < 2e-16 ***
NR_ICD9                0.1785573  0.0079578  22.438  < 2e-16 ***
N_VISITS_by_PATHOLOGY -0.0166722  0.0028000  -5.954 2.62e-09 ***
N_VISITS_by_ZONE       0.1047678  0.0061442  17.051  < 2e-16 ***
N_VISITS_by_PROVIDER   0.2159200  0.0023902  90.334  < 2e-16 ***
N_VISITS_by_CAEgroup  -0.3875283  0.0066555 -58.227  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.002292)

    Null deviance: 200411  on 239754  degrees of freedom
Residual deviance: 163012  on 239745  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure A.7: Variable Importance of Expected Frequency of Medical Expenses in XGB

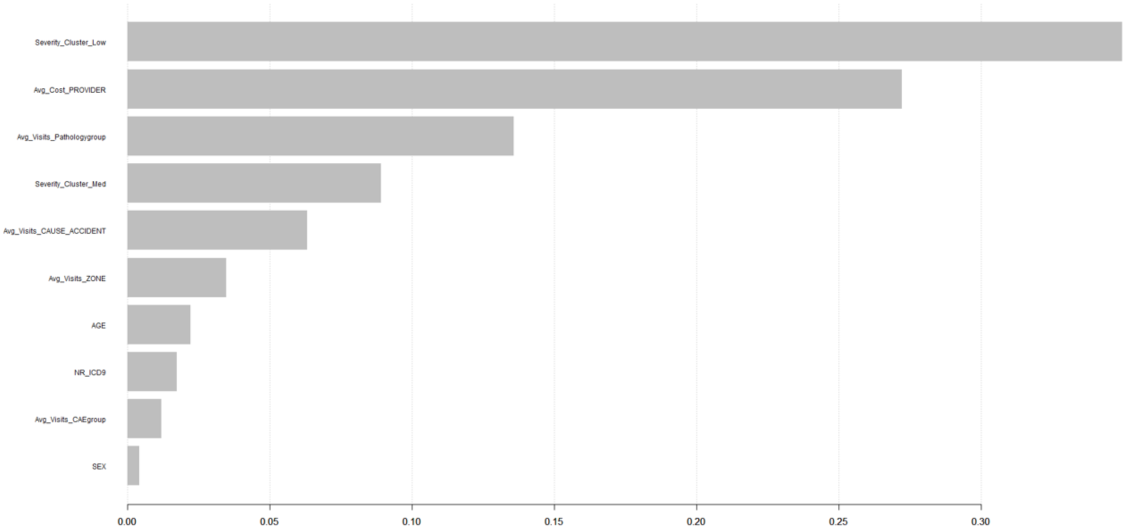


Figure A.8: Variable Importance of Expected Frequency of Medical Expenses in GLM

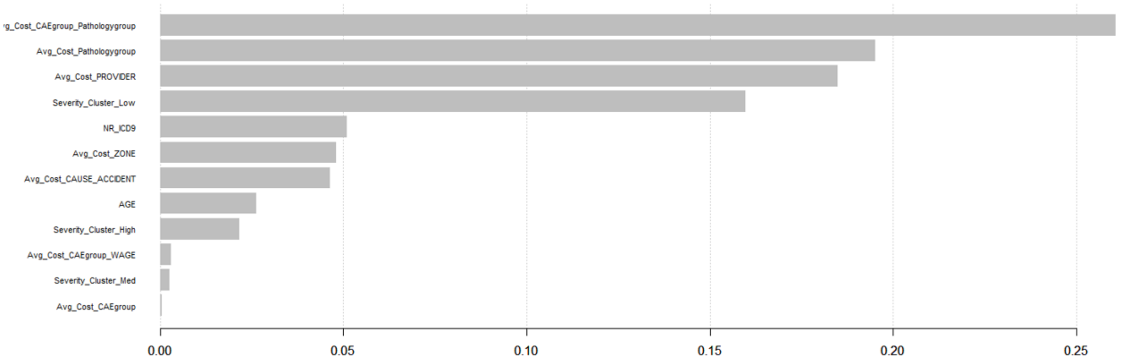


Figure A.9: Variable Importance of Expected Cost of Physiotherapy Expenses in XGB

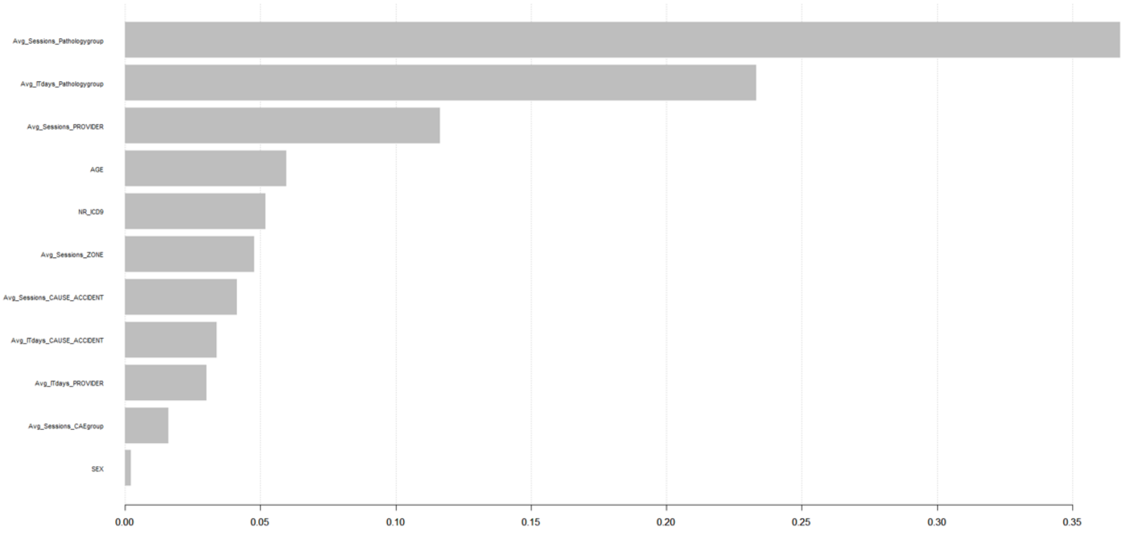


Figure A.10: Variable Importance of Expected Frequency of Medical Expenses in XGB

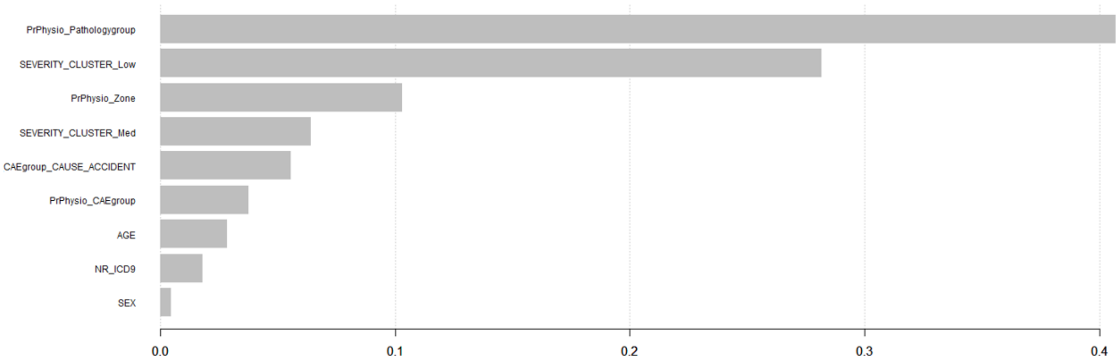


Figure A.11: Variable Importance of the proportion of physiotherapy in XGB