MACHINE LEARNING IN HEALTHCARE MANAGEMENT FOR MEDICAL INSURANCE COST PREDICTION

Thais Carreira Pfutzenreuter (PUCPR)
thais.pfut@gmail.com

Edson Pinheiro de Lima (PUCPR)
e.pinheiro@pucpr.br

Machine learning projects have been providing a better patient experience in care services. Healthcare has many issues and the cost of it is an essential indicator for insurance providers. In this context, the purpose of the present paper is to offer a comparison of machine learning approaches for the prediction of American medical insurance cost provided by Kaggle community with 1.338 instances. The focus did not consist of winning any competition, but developing a preliminary investigation of algorithms' performance assessment. Linear Regression regularizations were compared with more sophisticated algorithms, KNR, SVR, Simple Tree, Random Forest and XGBoost in terms of accuracy with $R^2$ and RMSE along with computational time. Linear Regression and its regularizations presented a good accuracy with a five-fold cross validation. However, GridSearchCV selection for best parameters achieved superior performance for more advanced algorithms, except Support Vector Machine that did not exhibit competitive accuracy. Computational time revealed to be an interesting assessment and depending on the organizational context, simple tree, $R^2$ 0.88, would occasionally overcome the others, since it had a competitive computational time comparing to XGBoost and Random Forest, the ones with the highest accuracy. The present study has contributed on proving machine learning value for health insurance price prediction and the importance of applying comparative performance metrics for the algorithms not only in accuracy, but also in computational time.

Key-words: Healthcare management, predictive analytics, machine learning, cost prediction, data mining.
1. Introduction

Medicine has incorporated machine learning to contribute to early treatment and disease prevention (PURBA et al., 2019) Research assessments of predictive analytics technology in medical applications have been acknowledged with particular emphasis on how hospitals have integrated predictive analytics into their routine healthcare services to improve the quality of care and disease preventions, although challenges have been faced in developing countries (ALHARTHI, 2018). With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces challenges in essential areas such as electronic record management, data integration and computer aided diagnoses. Machine learning offers a wide range of tools, techniques and frameworks to address these challenges. (NITHYA AND ILANGO, 2017).

In recent years there is an increase of industry initiatives to invest on artificial intelligence projects in health segment in order to extend human life, improve facilities of doctors, nurse teams and other professionals involved in the process of patient care and disease diagnostic assertiveness (PURBA et al., 2019). Bhardwaj et al. (2017) applied a systematic literature review to investigate machine learning multiple applications in healthcare management, identifying companies that have invested on predictive analytics solutions in this field. Among the companies encompassed by Bhardwaj et al. (2017) there is Aprioxi, a pioneer in the creation of algorithms for natural language processing that analyze unstructured data indexed in electronic health records, digitized notes, facsimiles and handwritten notes to produce high-quality forecasts for measurement, care and discovery improvements, Butterfly Network, a company that transforms diagnostic and therapeutic images with deep learning and Flatiron Health, an organization that provides an oncology platform that integrates the entire spectrum of clinical data, allowing cancer care providers to make smart decisions based on clinical datasets.

Organizations have achieved a better patient experience in care services. Healthcare has many issues and the cost of it is a critical indicator for insurance providers. In this context, the purpose of the present paper is to offer a comparison of machine learning approaches for the prediction of American medical health insurance cost provided by Kaggle community with 1.338 instances. The focus is not on winning any competition, but on developing a preliminary investigation of algorithms performance assessment. Linear Regression regularizations are compared with more sophisticated algorithms, KNR, SVR, Simple Tree, Random Forest and XGBoost not only in terms of accuracy with R² and RMSE but also in computational time.
2. Literature Review: Machine Learning in Healthcare Environment

Machine learning has been improving diseases diagnosis and healthcare treatments in a variety of distinct segments, helping to build sophisticated predictive analytics models for chronic kidney disease predictions (CHALEONNAN et al., 2016; PITCHUMANI, 2019), diabetic treatments (KALYANKAR et al., 2017), obstructive sleep detection (SARADHI AND NAGESWARA, 2019) precision psychiatry (BZDOK AND MEYER-LINDENBERG, 2018), effectiveness of music therapy (RAGLIO et al., 2020) and even emergency department revisits (BEN-ASSULI AND VEST, 2020).

Not only disease predictions and diagnostics, but machine learning has also played a remarkable role on predicting in-hospital length of stay among diabetic patients (MORTON et al., 2014) and among cardiac patients (DAGHISTANI et al., 2019). These studies reinforce that artificial intelligence provides not only improvements in disease diagnoses, but also in the prediction of hospital stay, serving as a strategic source for managing clinical beds and forecasting the allocation of necessary hospital resources. Moreover, machine learning has become a strategic support to protect patients interests in big data era (BALTHAZAR et al., 2018) and has also enabled the improvement of medical education itself (ARORA, 2018).

Health sector has progressively developed its operations with the emergent artificial intelligence. The cloud-to-cloud enables considerable improvements in service performance, although the selection of ideal virtual machines for processes remains complex. This strategic selection of virtual machines promotes performance gains over time and (ABDELAZIZ et al., 2018) suggests a new model for healthcare services based on a cloud environment using Parallel Swarm Optimization to optimize the selection of virtual machines. A diagnostic prediction for chronic kidney disease was tested using two phases to assess the performance of the proposed model. Initially the authors introduced linear regression to determine feature influence on disease, forming a basis with strategic predictor variables. Thirteen relevant factors were identified, such as age, anemia, blood glucose, blood urea and potassium. Some others that might have been influential, such as appetite and diabetes, have been rejected. At the second stage, neural networks performed the model for forecasting, being 70% of the data used for training and the remaining 30% for testing. The study was able to provide a 50% faster executional time than the current models with an accuracy of 97.8%, with a 64% superiority in relation to related works of the last generation, increasing the reliability of the diagnoses of health operational project.
Applications of machine-learning have also been explored to assist decisions about the need for surgical interventions and even the performance of cesarean sections has been encompassed (MOHAMMADI et al., 2012). A case study was examined at a health post in Tabriz, capital of Iran and the C4.5 Decision Trees method was applied to pregnant women dataset with the aim of predicting the need to take normal deliveries to cesareans. A direct relevance was found between the rate of heart failure and deliveries ending in cesarean section, with the results pointing out that 75% of women with inept cardiac status did not have a normal delivery and more than 65% of women with inept cardiac condition also had pressure abnormal blood. The results indicated that the proposed decision tree algorithm reached assertiveness for more than 86.25% test cases (MOHAMADDI et al., 2012).

Srividya et al. (2018) have developed a structure through machine-learning to determine an individual's mental health status and assess health metrics of different social groups and assist in the diagnosis of mental diseases such as anxiety and depression. The framework combined K-means clustering techniques to identify similarities between groups with supervised learning algorithms to train classifiers. Techniques such as SVM, KNN and Random Forest performed in an equivalent way while the use of ensemble classifiers was found to significantly improve the performance of mental health prediction with 90% accuracy. The authors defend the use of this structure to perform behavioral modeling of a target population and the usage of additional physiological parameters such as electrocardiogram and respiratory rate should also be included as predictors for mental status, although the interpretation and accessibility of these characteristics represents a big challenge.

Predictive analytics has also supported underdiagnosed diseases such as delirium offering preventable early signs with a delirium risk assessment during hospital admission process. Veeranki et al. (2018) showed that delirium prediction using machine learning algorithms is possible based on the patients’ health history and complementarily compared the influence of nursing assessment data on prediction models with both clinical and demographic data. Before recommending a predictive algorithm for clinical practice, however, the authors acknowledge the importance to know whether and for whom it works well.

Van Calster et al. (2019) defend some best practices of machine learning in healthcare environment, including: i) predictions should discriminate between individuals with and without the disease, implying balanced datasets, ii) risk predictions should be also accurate and iii) algorithm development must be prevented from overfitting, which usually results in poorer discrimination when classifying new instances.
3. Research Design

The selected case study consists on a comparison of machine learning approaches for the prediction of American medical health insurance cost provided by Kaggle community with 1,338 instances. The dataset encompasses the following characteristics:

- **age:** numerical and discrete variable, consisting of the age of the main beneficiary
- **sex:** categorical variable female/male
- **body mass index:** continuous numeric variable, representing the body mass index, providing an understanding of the body, weights that are relatively high or low in relation to height, objective body weight index (kg/m²) using the relationship between height and weight, ideally between 18.5 and 24.9
- **children:** Numerical and discrete variable, consisting of the number of children covered by health insurance, adding the dependent number included
- **smoker:** categorical variable smoker/non-smoker
- **region:** categorical variable, identifying the beneficiary's residential area in the United States in 4 possibilities: northeast, southeast, southwest, northwest regions.

The database provided by Kaggle Community suggests a Linear Regression approach. However, the present study proposes to go a step further Linear Regression and Ridge, Lasso and Elastic Net regularizations and compare them to more sophisticated algorithms, KNR, Simple Tree, Random Forest and XG Boost. GridSearchCV was performed using Scikit Learn library to find the best parameters for each algorithm. Figure 1 details Research Design procedures.

![Figure 1 - Research Design](image)

Source: the authors (2021).

The selected performance assessment metrics for the algorithms were $R^2$, RMSE and computational time in seconds.
4. Results

Among the categorical variables, instances are divided between 676 men and 662 women. With reference to regions, the proportion is also fair, since it is divided between 27.2% of the southeast region, 24.3% for both the northwest and southwest regions and 24.2% for the northeast. On the other hand, in relation to smokers, only 20.5% of the base is smokers, since only 274 are smokers in relation to the total of 1338 people. Figure 2 details regions and smoking features.

![Figure 2 – Region and Smoking Categorical variables](image)

Source: the authors (2021).

Analyzing the other variables, the minimum age is 18 years and maximum 64, which presents a dataset with both young and mature adults, but no representants of advanced age. Regarding the number of children, 75% have 2 kids or less and a small proportion of 3 to 5 children’s parents. Body mass index, named bmi by Table 1, reveals mean and median almost equal with 30 kg/m². The medical insurance cost, which is the class, presents a mean of 13.270 USD, however the price has a considerable variability, with a minimum of 1.121.87 USD and maximum of 63.770 USD.

<table>
<thead>
<tr>
<th>Table 1 - Numerical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>age</strong></td>
</tr>
<tr>
<td>count</td>
</tr>
<tr>
<td>mean</td>
</tr>
<tr>
<td>std</td>
</tr>
<tr>
<td>min</td>
</tr>
<tr>
<td>25%</td>
</tr>
<tr>
<td>50%</td>
</tr>
<tr>
<td>75%</td>
</tr>
<tr>
<td>max</td>
</tr>
</tbody>
</table>

Source: the authors (2021).

For the treatment of the features, it was observed that the database had 3 categorical variables: sex, smoker and region. Since applying **one-hot-encoding** for all could greatly
increase the dimension of the dataset and consequently damage the performance of algorithms, the next step was the association of numerical values for sex and smoker features because they are variables with only two possibilities: female/male and smoker/non-smoker. In relation to region, which has four possibilities, it was considered pertinent to perform one-hot-encoding so the feature could be broken down into four variables, as presented by Table 2.

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>bmi</th>
<th>children</th>
<th>smoker</th>
<th>charges</th>
<th>region_northeast</th>
<th>region_northwest</th>
<th>region_southeast</th>
<th>region_southwest</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>27.900</td>
<td>0</td>
<td>1</td>
<td>1.688.492.400</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
<td>33.770</td>
<td>1</td>
<td>0</td>
<td>172.555.230</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>33.000</td>
<td>3</td>
<td>0</td>
<td>444.946.200</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>22.705</td>
<td>0</td>
<td>0</td>
<td>2.198.447.061</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>28.880</td>
<td>0</td>
<td>0</td>
<td>386.685.520</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: the authors (2021).

Linear Regression was initially performed with a five-fold cross-validation. Linear Regression and Lasso regularization presented an R² of 0.7304, while Ridge and Elastic Net regularizations presented a minimally superior performance, with R² 0.7305 both. A Dummy approach, R² 0.0103, was also performed to defend the value of linear regression.

Figure 3 - Linear Regression and Regularizations

Source: the authors (2021).

After Linear regression results exhibited by Figure 3, the following step consisted of KNR, SVR, Simple Tree, Random Forest and XGBoost approaches. GridSearchCV of Scikit learn library was performed to identify the best parameters for each algorithm.
The selected performance assessment metrics for the algorithms were $R^2$, RMSE and Computational time in seconds. Table 3 exhibits the metrics along with KNR, SVR, Simple Tree, Random Forest and XGBoost best parameters for each algorithm enabled by Scikit Learn library with GridSearchCV.

Table 3 – $R^2$ RMSE and Computational Time Performance Comparison of Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>Best Parameters</th>
<th>Computational Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.7575</td>
<td>5.783</td>
<td>()</td>
<td>0.01</td>
</tr>
<tr>
<td>KNR</td>
<td>0.8166</td>
<td>5.029</td>
<td>{'n_neighbors': 7, 'weights': 'distance'}</td>
<td>0.46</td>
</tr>
<tr>
<td>SVR</td>
<td>0.5087</td>
<td>8.231</td>
<td>{'C': 10, 'kernel': 'linear'}</td>
<td>6.62</td>
</tr>
<tr>
<td>Simple Tree</td>
<td>0.88</td>
<td>4.067</td>
<td>{'max_depth': 5, 'splitter': 'best'}</td>
<td>0.26</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.8643</td>
<td>3.994</td>
<td>{'max_depth': 5, 'n_estimators': 60}</td>
<td>55.48</td>
</tr>
<tr>
<td>XGBOOST</td>
<td>0.8673</td>
<td>3.942</td>
<td>{'max_depth': 3, 'n_estimators': 50}</td>
<td>39.15</td>
</tr>
</tbody>
</table>

Source: the authors (2021).

Performance assessment metrics $R^2$ and computational time in seconds are also presented by Figure 4 for a visual benchmarking analysis.

Figure 4 - $R^2$ and Computational Time Metrics

Source: the authors (2021).

XG Boost presented the best $R^2$ performance, followed by Random Forest and Simple Tree. KNR is a little bit behind but still revealed superior performance compared to Linear Regression. Linear Regression presented superior performance over SVR, the worst algorithm, which could be possibly justified because Support Vector Machines are normally more advantageous in bigger and more complex databases.

Random Forest and XG Boost are the slowest algorithms, taking longer than 55 and 39 seconds, respectively. Simple Tree could possibly be considered the best performing model in case of a trade-off of performing time and accuracy of results, once Simple Tree $R^2$ 0.88 is not far behind XGBoost and Random Forest and it takes a fast time of 0.26 seconds for results, enabling faster decision-making and possible real-time creation of what-if scenarios.
5. Discussion

Linear Regression presented a good accuracy, $R^2$ 0.73 with a five-fold cross validation. However, the regularizations did not provide considerable improvements. On the other hand, additional advanced algorithms such as KNR, Simple Tree and XGBoost performed much better using GridSearchCV for the selection of best parameters. Support Vector Regression, however, did not present an improvement comparing to linear regression, demanding curiosity for a further research investigation. Simple tree presented an outstanding computational time faster than a second and its accuracy is not far behind comparing to the best ones, XGBoost and Random Forest.

Findings reinforce machine learning value is beyond disease preventions (BHARDWAJ et al., 2017; BALTHAZAR et al., 2018) for strategic healthcare management, since health insurance cost is an essential insight for healthcare price policies. Best practices in machine learning were also supported in the present application (ABDELAZIZ et al., 2018; VAN CALSTER et al., 2019; SRIVIDYA et al., 2018), using balanced datasets to avoid overfits and training-test segmentation.

6. Conclusion

Machine learning projects have been improving not only medical diagnostics, but also healthcare management with more assertiveness and faster decision-making processes. Health insurance cost is an important issue for medical corporations’ benchmarking with competitors to offer a competitive, fair and well-founded price for customers. The present study has contributed on proving machine learning value for health insurance price prediction and highlights the importance of applying comparative performance metrics for the algorithms not only in accuracy, but also in computational time.

Linear Regression and its regularizations achieved a good accuracy with a five-fold cross validation. However, GridSearchCV for best parameters selection enabled exceptional performance for more advanced algorithms, such as XGBoost, Random Forest, Simple Tree and KNR. Computational time revealed to be an interesting assessment and depending on the organizational context, Simple Tree with $R^2$ 0.88 would occasionally be the chosen algorithm, since it had a competitive computational time comparing to XGBoost and Random Forest, the ones with the highest accuracy. An important limitation to be addressed is a missing feature selection to test a performance improvement on all algorithms and examine if the ranking remains the same or if algorithms present unexpected performance results.
REFERENCES


ARORA, V. M. Harnessing the Power of Big Data to Improve Graduate Medical Education: Big Idea or Bust? Academic Medicine, v. 93, n. 6, p. 833–834, 2018.


