SALES PREDICTING TELEMEDICINE SERVICES DURING COVID-19 USING MACHINE LEARNING TECHNIQUES

André Luiz Mosko da Cunha (Pontifícia Universidade Católica do Paraná)
Mateus Franco Zuchetti (Pontifícia Universidade Católica do Paraná)
Sarah Cristine Takeute de Oliveira (Pontifícia Universidade Católica do Paraná)
Vinicius Issamu Zanoni Chiba (Pontifícia Universidade Católica do Paraná)
Marcelo Carneiro Gonçalves (Pontifícia Universidade Católica do Paraná)

Telehealth care has been a viable alternative in terms of safety against disease contamination, travel time, and cost reduction for medical consultations during the COVID-19 pandemic. In 2022, Brazil approved the law that authorizes the use of telemedicine services in the country, and since then the services by virtual attendance have grown exponentially. Thus, the demand management process in organizations needed to adapt quickly to this new scenario. Therefore, this research aims to apply machine learning techniques to assist in forecasting sales of remote medical consultation services (telemedicine) during the Covid-19 pandemic in a digital services consulting company. To achieve the objective, first a systematic review of the literature was performed, then collection and analysis of the data needed for the prediction process, and finally choice of the best method from metrics based on minimum error. As a result, it obtained that the technique with the lowest prediction error was Random Forest, which presented only 22% error when applied to the test data set.

Keywords: Sales Predicting, Machine Learning, Health Care.
1. Introduction

The world faces challenges in returning to business activities post-pandemic period, and with this a new exercise of extreme uncertainty consists in projecting future scenarios and ensuring resilience in the face of market unpredictability. In the face of this, the search for personalized, quality, and safe services has been the new expectation of companies. According to the World Health Organization (2023), the COVID-19 caused more than 6.8 million deaths worldwide and 697,000 in Brazil alone. On January 30th, 2020 the coronavirus outbreak constituted a public health emergency of international concern (PHEIC). The PHEIC were defined by the International Health Regulations in 2005 to characterize situations of risk of damage to public health worldwide, where international cooperation is essential for its containment, control and mitigation (RSI, 2005).

During the COVID-19 period in Brazil, on April 15, 2020, telehealth services authorized by law no. 13,989/20 on an emergency basis only (CAMARA, 2020). However, resulting from the COVID-19 pandemic, and the aggravating scenario of deaths in the world, law no. 13.989/20 was repealed to authorize and discipline the practice of telehealth throughout the national territory. With this, on December 27, 2022, law 14.510/22 was sanctioned, authorizing telemedicine in Brazil. The law deals with criteria for remote care of services related to all health professions regulated by the competent bodies of the federal executive branch (CAMARA, 2022).

Since the sanctioning of this law in Brazil, the remote services of medical consultations have grown leading to an increase in the need for management and control of the number of accesses, new subscriptions from physicians and patients. In view of this, this study seeks to perform an application of future demand management for a company that offers telemedicine services in Brazil using machine learning techniques. Machine learning is used to make predictions from statistical models and is defined as a set of methods that identify patterns in data, using them later to predict future data, in order to assist in decision making in cases of uncertainty (MURPHY, 2012). Tools such as Big Data, Machine Learning, Internet of Things are increasingly incorporated with statistical algorithms to meet the needs, as an alternative in the analysis and integration of data in order to assist in solving problems without much effort, in order to avoid errors and waste, and to focus efforts where it is really necessary and will bring result (SILVA, 2021).
The managers of the company, the object of study, as well as collaborators of different companies in the digital industry, have difficulties in analyzing, forecasting, planning, and measuring the efforts required for the greatest audience reach, as well as for their retention. Most studies of demand forecasting and sales control are focused on industries or companies with physical products, and there is a lack of methods aimed at more complex markets of intangible products. Based on this, this work seeks to answer the following problem: "Is it possible to obtain results capable of assisting in decision making regarding the planning process and sales control of telemedicine services from Machine Learning techniques with minimal errors?"

To achieve the objective of the work, the following steps have been taken: (i) literature review of articles published in international journals in the last 5 years that used supervised machine learning techniques; (ii) mapping of the company's sales process; (iii) data collection, mining, and analysis; (iv) tests on machine learning methods to perform demand forecasting using supervised learning techniques, namely: K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM) and Gradient Boosting Machine (GBM); (v) Choosing the best forecasting method based on error metrics; (vi) Discussing the results and proposing improvements with the organization.

2. Methodology

The following steps were performed: (i) initially, a literature review of articles published in international journals in the last 5 years, from the Scopus database. The terms used were: "Machine Learning", "Health Care", "K-Nearest Neighbors", "Random Forest", "Support Vector Machine" and "Gradient Boosting Machine". (ii) The process mapping was obtained from meetings with the sales process managers using the Bizagi tool (it was suppressed in the research due to page limitation). (iii) For data collection, mining and analysis, the company's database was used and with the help of the Excel tool, the data was grouped in a single file and organized in columnar forms. Then, a data mining analysis was performed in order to identify patterns for clustering and outlier elimination. (iv) The model tests were performed using supervised machine learning methods, namely: "K-Nearest Neighbors", "Random Forest", "Support Vector Machine" and "Gradient Boosting Machine", with the aid of PyCharm software, version 2022, using the Python programming language. The hyperparameters of each model were adjusted to improve the performance of each technique.
The choice of machine learning method to perform the management of the company's future demands were defined from error metrics, namely: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The discussion of the results and proposals for an action plan for the organization occurred with the help of the continuous improvement tool 5W2H (it was suppressed in the research due to page limitation).

3. Theoretical Background

3.1. Production Planning and Control (PPC)

Production Planning and Control (PPC) is one of the most common activities in companies, being related to the main operations of the business. According to Slack et al. (2009), "planning and control activities provide the systems, procedures and decisions that bring together different aspects of supply and demand". This means that planning and control has the versatility to from data such as raw materials required, supply time, production time, production cost, historical demand for products or services, inventory and other data, plan how, in what order and quantity the production should be conducted, so that the company has a better performance in operations activities.

The main activities of PPC are connected with the planning of the supply network, demand forecasting, process flow, process technology, and control of capacity, inventory, supply chain, and production, in a short, medium, and long-term period, and may vary from activities that must be performed in the next few hours to strategic decisions for years.

According to Tubino (2017), the Planning and Production Control (PPC) is a truly relevant strategic tool for an industry or service company that wants to improve its logistics and increase its competitiveness in the market, since it brings benefits such as better results, cost minimization, waste reduction, among other short or long term advantages. PPC has contact with all areas of the organization, and this contact is essential for deliveries to be made. And when it comes to programming the demand forecast, for the purchase of materials and delivery, it is important that everything is aligned. According to the same author, the demand forecast is relevant information in the management of a production system, for all sectors of the company, and especially for Production Planning and Control (PPC).
3.2. Machine Learning

According to Goodfellow et al. (2016), machine learning is an algorithm that can learn from data. Besides that, Machine Learning enable us to tackle tasks that are too difficult to solve with fixed programs written and designed by human beings. This means that many activities can be performed by algorithms in a more performant way than humans. Currently, these algorithms are gaining relevance due to their popularity in many areas of business, such as medicine, marketing, sales, and the financial sector. These algorithms reduced the workload of these business areas and brought agility to the processes. The intense use of these algorithms is due to the development of computer capacity and the quantity and quality of data available.

Machine learning is being increasingly used in several areas and functions in industries and service companies. In Brazil, this concept is still little known/used, as many companies are unaware of its benefits or do not yet have the necessary maturity to implement it. For its use, a system/method of control and data management is required, and these are often not present in the companies. Moreover, there are problems when the opposite occurs, that is, the company generates data simultaneously, making it difficult to choose the model and limiting the processing power of the available computer hardware (LEE et al, 2022).

From the content analysis of the most recent articles, the most frequently used supervised methods were K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM) and Gradient Boosting Machine (GBM).

3.2.1 K-Nearest Neighbors (KNN)

This algorithm compares features of an entity with the training base and relates the new entity to the closest ones from a distance function, and through this approximation, builds classes (neighborhoods). The algorithm categorizes the closest K values in the data set (KRAMER, 2013). Examples of recent papers that have used the KNN technique include works in the area of spreading fake news in social networks Roy et al (2023) and in health care in predicting breast cancer diagnosis (JAISWAL, 2023).

3.2.2 Random Forest

Random Forest (RF) is a method that has a simple but remarkably effective concept. This algorithm consists in creating probability trees using the training database, and then averaging all tree results and finally predicting. Usually, to make a prediction for an observation, the
average or mode of a training observation in the region to which it belongs is used. Since the splitting rules used to segment the prediction space can be summarized into a tree, these approaches are known as decision tree methods (JAMES et al, 2013). As simple as it sounds, their effectiveness proves to be extremely high, this is because a large number of trees showing close results have a higher predictive ability than models using individual data. As examples of recent articles that used the Random Forest technique, we can mention in the medical field, the prediction of Alzheimer's diagnosis Agarwal et al (2022) and in the technology and security sector, the detection of anomaly attacks in IoT networks. (TYAGI et al, 2021)

3.2.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) analyzes the data and recognizes patterns that can be used for classification or regression. For this to be possible, the algorithm searches for an ideal hyperplane that separates the two classes, which is the plane that has the greatest possible distance between the analyzed classes. According to James et al (2013), Support Vector Machine is a machine learning technique that aims to recognize patterns in data. Despite presenting a good generalization ability and robustness in the application of large databases, SVM has a high sensitivity in the choice of values, leading to a difficulty in interpreting the generated model.

As recent examples of articles that have used the SVM technique, we can mention in the area of technology and information security, the detection of malicious links Patgiri et al (2019) and the monitoring of the electricity consumption of electrical appliances. (CHEA et al, 2022)

3.2.4 Gradient Boosting Machine

The Gradient Boosting Machine (GBM), unlike other machine learning methods, can be considered an "ensemble" method, i.e., it uses several other prediction methods that are considered weaker, seeking to obtain more robust results.

One of the biggest attractions for using GBM is the possibility of modeling the cost function according to the user's needs, i.e., its flexibility means that it can be used in any data-driven environment. (NATEKIN et al, 2013)

As recent examples of articles that used the Gradient Boosting Machine technique, we can mention in the automotive industry, the prediction of electric vehicles purchase expectation,
Shri et al (2023) and in the medical field, the creation of a model to predict the development of diabetes. (MOHAN et al, 2023).

3. Results
3.1. Content Analysis
The content analysis regarding the Systematic Literature Review of this study used the following filters: (i) first, the search terms involving supervised machine learning methods, machine learning and health care were selected; (ii) next, only the last 5 years, from 2018 to 2022 were selected; (iii) finally, only articles in English were selected. The database used was Scopus. As a result of the search, forty-eight (48) articles were obtained. The titles and abstracts of the articles were read in order to identify the applied area of the article, the main contributions, and the models or techniques applied.

As one of the main results of the content analysis, it was obtained that 52% of the sample (25 articles) were from the field of medicine, followed by biotechnology and computer science. Although the area of concentration was medicine, most of these studies are related to cardiovascular diseases (TIWARI et al, 2022; LEE et al, 2022; NUSINOVISI et al, 2020), health implications as a result of COVID-19 (SANTANA et al, 2021; AKTAR et al, 2021; BAIERLE et al, 2020), clinical studies of patients (MAO et al, 2022; SHAH et al, 2022; MAO et al, 2022; WIE et al, 2022; ALFI et al, 2022; PATEL et al, 2022; JEN et al, 2021; SHIM et al, 2020; HUSSAIN et al, 2021; DEIF et al, 2021; VERMA et al, 2020; SHIM et al, 2021; CHO et al, 2019; KENDALE et al, 2018; ZHANG et al, 2019) and protein substances (AGRAWAL et al, 2021; SCHAEFER et al, 2021; WU et al, 2021; NORA et al, 2016). No work has been identified using a database similar to the case of this study related to demand management for telemedicine services.

3.2. Data collect
The company under study focuses on easy accessibility and communication between patients and physicians, whether for in-person or remote consultations. Through it, doctors can advertise their work and even share their agendas, and patients can search for professionals for their needs and choose according to location, availability, and financial viability. The company currently has the largest healthcare platform in the world. Of Polish origin, it was initially developed in
the European market. In Brazil, it has been operating for 6 years with state-of-the-art technology, developing the health tech sector in the country.

In the company analyzed in this case study, sales are separated in two strands: Inbound and Outbound. Inbound refers to the sales in which the customer contacted the company, and Outbound refers to the sales in which the sales team contacted the customers, and they bought the packages offered.

To carry out this project, we considered the data on the number of sales to health professionals in the Outbound segment from September 2021 to the beginning of June 2022. The data collected were the sales per day during this period. A total of 522 data were collected, organized according to the days of the months in question, with the help of the Excel tool.

### 3.3. Comparative analysis of machine learning techniques

The parameters used in each model can be seen below (see Table 1):

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>“n_neighbors=20, *, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None”</td>
</tr>
<tr>
<td>RF</td>
<td>“criterion='squared_error', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, RF random_state=1, max_leaf_nodes=None, min_impurity_decrease=0.0, ccp_alpha=0.0”</td>
</tr>
<tr>
<td>SVM</td>
<td>“kernel='rbf', degree=3, gamma='scale', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1”</td>
</tr>
<tr>
<td>GBM</td>
<td>“loss='squared_error', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, init=None, random_state=None, max_features=None, alpha=0.9, verbose=0, max_leaf_nodes=None, warm_start=False, validation_fraction=0.1, n_iter_no_change=None, tol=0.0001, ccp_alpha=0.0”</td>
</tr>
</tbody>
</table>

Source: Authors (2023)

Table 2 summarizes the performance of each technique by metric:

<table>
<thead>
<tr>
<th>Metrics/Model</th>
<th>KNN</th>
<th>RF</th>
<th>SVM</th>
<th>GBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>50.05</td>
<td>15.01</td>
<td>45.16</td>
<td>26.97</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.75</td>
<td>0.22</td>
<td>0.64</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Considering the results, the technique with the lowest error was Random Forest, which presented the lowest errors when compared to the other techniques. Considering the MAPE, the error obtained in the test set for the Random Forest technique was only 22%, therefore classifying it as the chosen method for prediction.

The behavior of the demand forecast for the test set using the selected technique was as follows:

**Figure 1 – Demand forecasting based on the best technique.**

![Demand forecasting based on the best technique](Source: Authors (2023))

The behavior of the error on the test set with the application of the best technique was as follows:

**Figure 2 - Demand Forecast Error based on the best technique**

![Demand Forecast Error based on the best technique](Source: Authors (2023))
In figure 01 it is possible to observe the adherence of the forecast with the test set data. In figure 02 it is possible to verify that the errors were around zero, thus classifying a good solution for the organization.

4. Final Considerations

The paper aimed to apply machine learning techniques to assist in predicting sales of remote telemedicine services with a database covering the Covid-19 pandemic period of a digital services consulting company.

The application of machine learning to predict the number of sales of services is extremely necessary to assist in the dissemination of health professionals of the company object of study focused on promoting accessibility services and communication between patients and doctors (telemedicine), in the face of the pandemic scenario we are experiencing.

Through the application of techniques as KNN, RF, SVM and GBM, the model that proved most efficient for the case was the Random Forest, which had the lowest average percentage error compared to the others, with the value of 22% that can be considered acceptable, considering that this is an initial model and the amount of data collected is not large enough for greater reliability. Thus, the study meets stakeholders' expectations by presenting the possibility of predicting the quantity of future sales, as of the prediction methods using machine learning.

This work contributes to the company's preparation to meet future service demands, since there is the possibility of forecasting through the method suggested above.

As a contribution to academia, the article presented an application of supervised machine learning methods in an innovative dataset, which has not yet been used by other researchers, according to the systematic literature review conducted in this study, presenting a dataset of telemedicine services.

Moreover, as a practical contribution to society, this work presents adherence to the needs for international cooperation (PHEIC), to prevent the proliferation of cases of COVID-19 in Brazil and consequently in the world from the alternative use of telemedicine services. With the approval of law no. 14.510/22 in Brazil [4], telemedicine services tend to grow, and this will lead to the need for demand management of these services. We show in the research that making use of supervised machine learning methods is an efficient alternative, with minimal errors, for predicting the demand for health tech services, which allows the decision maker to anticipate in terms of resource allocation to ensure fast, quality, and safe services.
As a limitation of the research, the aggregation of data that were passed on for the construction of the research stands out, since a limited number of predictors was obtained in the predicting process. For future studies, it is suggested to capture a larger amount of sales data, to bring more reliability and assertiveness to the method used, in future forecasts. Furthermore, the addition of predictive variables to the model, such as sales hours, locations, etc., to reduce errors and increase the accuracy of the model.

REFERENCE


Lee, Y. H., Tsai, T. H., Chen, J. H., Huang, C. J., Chiang, C. E., Chen, C. H., & Cheng,


Wie, J. H., Lee, S. J., Choi, S. K., Jo, Y. S., Hwang, H. S., Park, M. H., Kim, Y. H.,

