

Sales Forecast and Churn Prediction for the International Health Insurance Company

FINAL REPORT

Course: Data Science Continuous Mar23

Jonathan Leipold | Christian Hirning | Rumiya Al-Meri

09.11.2023

# Introduction

## About the company and the product

[BDAE](https://www.bdae.com/en/bdae-group/about-bdae-group) is a specialist for international health insurance products. It is an insurance broker in co-working with the big health insurance companies. BDAE has its’ own sales & product development as well as claims handling service. Main target group are expatriates with the longer international assignments (from at least 1 year).

Company offers own health insurance products. There are several products depending on the area where the insured country of stay is situated. For this particular project one product (insurance type) was chosen to reduce the complexity. The product price is a premium amount which is paid by a client on a monthly, quarterly or a yearly basis.

## background

The sales forecast and churn predictions are inevitable for the company. Based on it BDAE takes decisions on new products development and clients’ acquisition to ensure a continuous growth and avoid existential threat. Quick access to forecast by product allows quick decision-making and resources optimization which in turn reduce costs and increase overall profit.

All data is collected in the ERP-system (SAP) based on a SQL-database. The forecast of future sales is still partly based on manual estimations. The sums and ratios of historical sales data are used combined with manual estimations of summed sales amounts to predict future sales. Data is grouped by (insurance-)product / category, time and some other variables. Churn predictions are not yet implemented into the forecast procedure by company. The goal of this project from a technical point of view to minimize the manual work with the reliable ML-models implementation.

From a scientific point of view there are several challenges to overcome by choosing the best ML-model for sales forecast. It is described later in details in the chapter IV. Due to the fact that there is no information about future behaviour such as the number of contracts is unknown as well as the profile of future clients, it is challenging to find the right and reliable ML-model to predict future sales. Once the model is found, it might be applicable to the similar problem definition, e.g. mobile phone contracts’ sales forecast.

Context

* Context of the project's integration into your business.
* From a technical point of view.
* From an economic point of view.
* From a scientific point of view.

## Objectives

The **main objective** is to create **the best performing model for sales predictions**, in particularly premium amount prediction. Moreover, due to a limited number of features which are known for the future, it was decided to add another objective, namely **churn predictions**. It will be discussed later in this report in details.

1. Find the best model for forecasting / predicting the premium amount

2. Find out how premium adjustments impact the value of premium amount

3. Find the premium adjustments which maximize the premium amount

## Contribution

In this project group Johnathan Leipold is a representative of the BDAE company, an industry expert, the data owner and the initiator of a current project. Johnathan was consulting on the data during the project, set up the initial data base and mostly all pre-processing steps. Christian Hirning and Rumiya Al-Meri have no experience in the insurance industry but rather in statistics and Christian as well a deeper knowledge in programming. Raphael (DataSceintest) contributed as a project tutor.

What are the main objectives to be achieved? Describe in a few lines.

For each member of the group, specify the level of expertise around the problem addressed?

Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions.

(Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?).

# Data exploration and visualisation

## Framework

Three data sets were taken for the analysis for the period of 2014 – 2023 YTD:

* Sales Data in form of transactions, in total about 230 000 transactions
* Premium adjustments data, #transactions?
* Contract termination behaviour, #transactions?

The data sets are owned by BDAE Group and are not available to the public. Each group member signed the confidentiality agreement with BDAE. The personal information was replaced with the created contractID.

* Which set(s) of data(s) did you use to achieve the objectives of your project?
* Are these data freely available? If not, who owns the data?
* Describe the volume of your dataset?

The project is splited in two sub-projects: sales (premium amount) prediction and churn prediction and will be presented in this part by sub-project for a better understandability.

## II.1. Sales prediction

## Relevance

* Which variables seem most relevant to you with regard to your objectives?
* What is the target variable?
* **Target variable** is a premium amount grouped by month.
* What features of your dataset can you highlight?
* Are you limited by some of your data?

## Pre-processing and feature engineering

* Did you have to clean and process the data? If yes, describe your treatment process.
* Did you have to carry out normalization/standardization type transformations of your data? If yes, why?
* Are you considering dimension reduction techniques in the modeling part? If yes, why?

## Visualizations and Statistics

* Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?
* Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)
* Present the statistical analyzes used to confirm the information present on the graphs.
* Draw conclusions from the elements noted above allowing them to project themselves into the modeling part

## II.2. churn prediction

## Relevance

* Which variables seem most relevant to you with regard to your objectives?
* What is the target variable?
* **Target variable** is a premium amount grouped by month.
* What features of your dataset can you highlight?
* Are you limited by some of your data?

## Pre-processing and feature engineering

* Did you have to clean and process the data? If yes, describe your treatment process.
* Did you have to carry out normalization/standardization type transformations of your data? If yes, why?
* Are you considering dimension reduction techniques in the modeling part? If yes, why?

## Visualizations and Statistics

* Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?
* Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)
* Present the statistical analyzes used to confirm the information present on the graphs.
* Draw conclusions from the elements noted above allowing them to project themselves into the modeling part

**Assessment methods:**

**Reconstituted professional situation: from a set of company data, the candidate must implement various pre-processing and data augmentation to make them usable through machine learning techniques.**

# Modelling

## Classification of the problem

* What kind of machine learning problem is your project like? (classification, regression, clustering, etc)
* What task does your project relate to? (fraud detection, facial recognition, sentiment analysis, etc)?
* What is the main performance metric used to compare your models? Why this one?
* Did you use other qualitative or quantitative performance metrics? If yes, detail it.

## Model choice and optimization

* What algorithms have you tried?
* Describe which one(s) you selected and why?
* Did you use parameter optimization techniques such as Grid Search and Cross Validation?
* Have you tested advanced models? Bagging, Boosting, Deep Learning… Why?

## Interpretation of results

* Have you analyzed the errors in your model?
* Did this contribute to his improvement? If yes, describe.
* Have you used interpretability techniques such as SHAP, LIME, Skater… (Grad-CAM for Deep Learning…)
* What has (or not) generated a significant improvement in your performance?

**Assessment methods:**

**Professional scenario: based on a proposed solution, the candidate will have to produce a summary report including: the explanation of the choices of AI solutions implemented, the interpretation of the results, the evaluation of the reliability of the algorithms and an optimization proposal.**

# Conclusions

## challenges

* What was the main scientific obstacle encountered during this project?
* For each of the following points, if you encountered difficulties, detail how they slowed you down in setting up your project.
* Forecast: tasks that took longer than expected, etc.
* Datasets: acquisition, volumetry, processing, aggregation, etc.
* Technical/theoretical skills: timing of skill acquisition, skill not offered in training, etc.
* Relevance: of the approach, model, data, etc.
* IT: storage power, computational power, etc.
* Other

## Report

* Detail what was your main contribution to achieving the project's goals.
* Have you changed the model since the last iteration? If yes, provide details.
* Present the results obtained and compare them to the benchmark
* For each of the project's goals, detail how they were achieved or not.
* If they have been reached, in which process(es) can your model fit? Detail.

## further steps

* What avenues for improvement do you suggest to increase the performance of your model?
* How has your project contributed to an increase in scientific knowledge?

# Bibliography

* <https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments>
* <https://thecleverprogrammer.com/2021/05/19/sales-prediction-with-machine-learning/>
* <https://towardsdatascience.com/5-machine-learning-techniques-for-sales-forecasting-598e4984b109>
* What bibliographical elements (research articles, blog, books, etc.) did you rely on to carry out your project?

# Appendices

* Gantt diagram.
* Description of code files.