Data analysis for telecommunication services

Marcelo Dante Caiafa[1], Ariel Aurelio[1] Alejandro Bevilacqua[1]

[1]Universidad Nacional de La Matanza,
Florencio Varela 1903, CP1754, Bs As, Argentina
{mcaiafa, aaurelio, abevilacqua}@unlam.edu.ar

Abstract. The digital transformation has brought significant changes in the production of goods and services. The data has become a main source of knowledge. Its adequate treatment allows to obtain valuable information. With data analysis skills the organizations can process big amounts of data providing relevant insights into customer behavior, market trends, and business operations. The professional link between university and industry is the goal of our academic work. The research addresses the need to consolidate results that add value to decision makers. This is about the implementation of an articulation project based on data of telecommunication services. The work is an exploratory data analysis. The goal is to build a customer profile with high dropout potential. Python and its different libraries were used in each stage. The teamwork was made up of students. Their performance was evaluated by analytical rubrics. The objective is to document and develop data analysis project. It´s intended to promote activities that integrate academy with industry. It reflects the value that new tools are capable to offering students, professors and managers in ICT (information & communication technology).

Keywords: Data analysis, ICT, churn, professional skills.

1 Introduction

Digital transformation is generating in the last decades big changes in the way of producing and consuming goods and services [1]. The information management allows organizations to improve their processes, "data is the new currency that sustains fundamental changes in the fourth industrial revolution" [2].

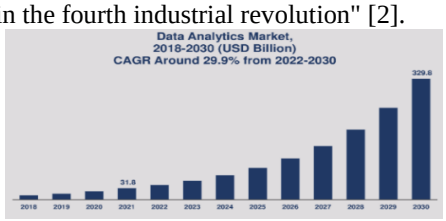

Fig. 1. Data Analytics Market (Acumen R&C 2021)

A data-oriented organization has advantages in decision-making because it is based on evidence [3]. Some studies [4] [5] show organizations with data-driven approach improve productivity and profitability. These indicate global data analytics market

grew to usd 31.8 billion by 2021 and estimate it will reach usd 329.8 billion in 2030.

The chart bellow shows four analysis types. These can be considered stages of the same project, that as complexity progresses the value of contribution increases [6].

a) Descriptive analytics: It illustrates outcome data collected over a time interval.

b) Diagnostic analytics: it looks for the root cause of a problem.

c) Predictive analytics: it uses past data to make forecasts.

d) Prescriptive analytics: it involves machine learning techniques to forecasting.

Fig. 2. Analytical madurity model (Gartner 2012)

The study is based on the articulation of academic and professional worlds. It focuses on exploratory analysis of data from telecommunications services. Different authors [7] [8] [9] are taken as reference. The data analysis was grouped in three stages [10].
Data collection: It was implemented by scraping process using Beautifulsoup.
Data processing: cleaning, wrangling, insigths detection Pandas &numpy were used
Data exploration: information was visualized by Matplotlib and Seaborn.

An EDA (Exploratory Data Analysis) is focused on process. Entry-level techniques (descriptive statistics, correlations, and basic visualizations) can be learned [11]. The research question that guides the work is how to develop data analysis project that links engineer students to telecommunications service sector. The objectives are:

1. To document the process to define typical customer profile with the high churn.
2. To build methodological proposal for EDA project to deploy professional skills.

The churn rate refers to percentage of customers who close their contract/subscription for period of time. High churn rate means inability to retain customers. It inquired about the value which this proposal would provide. Cost to retain customer vs. costs of getting new one, loyalty, cross-sell strategies were studied deeply.

2 Work Development

This project is structured in five stages. Firstly, software tool selection and requirements are performed. The analytics rubrics are made [12]. On the second, data collection is processed by scraping and the indicators are defined. The third instance is about data processing. They are cleaned, ordered and the statistical analysis is executed. The results presentation and correlation matrix are built in stage four. Software tool was Python 3.9 due to: expressive, interpreted, cross-platform, object oriented and open-source language, its extensive libraries among others [13].

2.1 Data Collection

Due to security policy issues, the data had to be accessed through scrap techniques.

```
import requests
import pandas as pd
from bs4 import BeautifulSoup
from urllib3.exceptions import InsecureRequestWarning
from urllib3 import disable_warnings  #soluciona problemas con max ssl certif
disable_warnings(InsecureRequestWarning)
dias = [03012023,03022023,03032023, 03042023, 03052023, 03062023, .....]
def obtener_data(dia):
    web = f"https://intranet/wiki/{dia}CustomerData" #se enmascara nombre real
    response = requests.get(web, verify=False)
    contenido = response.text
    soup = BeautifulSoup(contenido, "lxml")  # lxml será el parser a utilizar
    datas = soup.find_all("div", class_ = "detailsbox")
    MonthlyCharges = []
    Churn = []
    .......# se resume script
    for data in datas:
        MonthlyCharges.append(data.find("th", class_ = "MonthlyCharges").get_text())
        Churn.append(data.find("th", class_ = "Churn").get_text())
        ...........# se resume script
    diccionario_data = {"MonthlyCharges": MonthlyCharges, "Churn": Churn, ........}
    df_data = pd.DataFrame(diccionario_data)
    df_data["dia"] = dia
    return df_data
telco = [obtener_data(year) for dia in dias]  #genero lista para colectar la data histórica
df_telco = pd.concat (telco, ignore_index=True) #genero df concatenando todos los registros
ruta_csv = "\\2023\\UNLAM\\Investig_2023\\dataframe.csv"  #genero csv con df call dataframe
df_telco.to_csv(path_or_buf=ruta_csv, index=False)
```

Fig. 3. Scrapping Process

And were collected from html page on intranet web page which displays CRM´s information (customer relationship management) [14]. During last two months were analyzed and the complete dataset had 7236 records of twelve variables.

2.2 Data Processing

Many data manipulations were made. Data type is defined for each variable. Missing data and outliers are detected and consolidated for statistical analysis. Categorical variables are mapped to numerical [15]. Null data are removed.

2.3 Data Exploration

Statistical analysis is carried out to customer profile. Sample of correlation matrix is reproduced below which allowed to identify relevant variables for churn prediction.

	Churn	CargosMensuales	Permanencia	CargosTotales	Adultos	TipodeContrato_Anual	TipodeContrato_Bianual	TipodeContrato_Mensu
Churn	1.000000	0.192858	-0.354049	-0.199484	0.150541	-0.178225	-0.301552	0.4045
CargosMensuales	0.192858	1.000000	0.246862	0.651065	0.219874	0.004810	-0.073256	0.0589
Permanencia	-0.354049	0.246862	1.000000	0.825880	0.015683	0.202338	0.563801	-0.6493
CargosTotales	-0.199484	0.651065	0.825880	1.000000	0.102411	0.170569	0.358036	-0.4467
Adultos	0.150541	0.219874	0.015683	0.102411	1.000000	-0.046491	-0.116205	0.1377
TipodeContrato_Anual	-0.178225	0.004810	0.202338	0.170569	-0.046491	1.000000	-0.288843	-0.5700
TipodeContrato_Bianual	-0.301552	-0.073256	0.563801	0.358036	-0.116205	-0.288843	1.000000	-0.6219
TipodeContrato_Mensual	0.404565	0.058933	-0.649346	-0.446776	0.137752	-0.570053	-0.621933	1.0000

Fig. 4. Correlation Matrix

3 Results

The churn rate was 26%. It is observed that clients with highest unsubscribe request are those with highest monthly charges and lower than average tenure [16]. The most relevant features for churn prediction are selected. This involves identifying the variables that have a significant impact on the probability that a customer will resign

the service. The most representative variables for dropout rate are: monthlycharges, total_charge, OnlineSecurity, Techsupport, PaymentMethod.

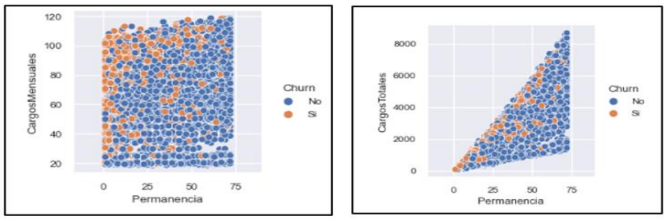

Fig. 5. Correlation charge vs ternure

4 Conclusions

The most representative variables in customer profile with high churn rate were typified: type of contract (monthly), technical support (not contracted), payment method (electronic check) and online security (not contracted). The clients with the greatest tendency to abandon are those with lowest monthly payment and the shortest stay.

This work was a first step in the data science study. This is not just infrastructure implementation, but it is strategic to support decision-making process.

The evaluation of student´s performance with analytical rubrics allowed to determine the development of professional competences in each one of the different stages.

For future work, it´s proposed to continue with the development of the analytical maturity model, based on this descriptive analysis, advancing to the predictive one with machine learning, through SciPy libraries. Additionally, this model could be applied in other careers, that use data sets linked to their specific disciplines, to integrate knowledge and skills from their productive environments.

References

1. S Schwab. La cuarta revolución industrial. World Economic Forum. Ed Debate, (2016)
2. R. Privdeville. Prepare for data Revolution. Data-driven world. Armanino, (2019)
3. M. Schwartz. War & Peace & IT: Business Leadership. Ed IT Revolution Press. (2019)
4. A. McAfee, & Brynjolfsson. Management Revolution. Harvard Business Review. (2012)
5. Acumen, Research & Consulting. Global Data Analytic Market. (2022).
6. W. Jensen. Statistics=Analytics. Quality Engineering, Inc., Flagstaff, Arizona, p 7. (2021)
7. C, O'Neil & R. Schutt. Doing Data Science. O'Reilly Media Inc, California. (2013)
8. J. Saltz & Shamshurin. Exploring the process of data science. IEEE Conference, (2015)
9. Wickham, H., Grolemund, G.: R for Data Science. O'Reilly Media Inc, California. (2016)
10. S. Van Daele & Jansseswillen. Identifying the Steps in EDA. ICPM. Ed. Springer (2022)
11. M. Courtney. Exploratory Data Analysis in Schools. IJEPL. Volume17(4). (2021)
12. J. Del Pozo. Competencias profesionales: herramientas de evaluación. Ed Narcea. (2017)
13. W. Bel, Algoritmos y estructuras de datos en Python. Facultad CyT. p17. Ed Uader (2020)
14. S. Mukhiya & Ahmed. Hands-On exploratory data analysis with Python. Ed. Packt (2020)
15. W. McKinney. Python for data analysis: wrangling Pandas & Numpy. Ed.O´Reilly. (2022)
16. A. Pajankar. Hands-on Matplotlib. Learn Plotting with Python 3. Ed Apress (2021).