EXPLORATION OF DATA SKILLS & ISSUES

STATISTICAL THINKING FOR DATA SCIENCE

SONUKUMARI RATHORE
25208188

Table of Contents

Introduction

In the highly competitive telecom industry, efficient marketing efforts are critical for obtaining and maintaining customers, especially when advertising subscription services. With considerable marketing costs committed to client acquisition, ineffective targeting can result in lost resources and lower profitability. This research examines data from a telecom company's marketing campaigns to better understand consumer behaviour and determine the aspects that have the most effect on subscription decisions. The purpose of assessing characteristics such as contact communication type, past campaign outcomes, and contact frequency is to give insights into optimising marketing tactics and increasing membership rates.

Problem Formulation

Context Analysis

In today's telecom sector, marketing efforts are very competitive and play a key role in acquiring clients, particularly repeat customers and even young people. According to current industry statistics, client acquisition expenses in telecoms may account for a sizable portion of marketing spend, and poorly targeted campaigns risk wasting resources and lowering total profit.

Problem Statement

The telecommunications company recently launched a new marketing campaign in order to market their new subscription package to clients. They want to understand their customers' behaviour based on prior campaigns. The major goal is to determine the tactics and communication strategies that are most appealing to clients, which will help us enhance membership rates.

Research Questions

1. How does contact communication type (cellular vs. telephone) affect subscription rates?

2. How does the outcome of previous campaigns influence the likelihood of a customer subscribing?

3. Does the number of contacts made with a customer affect their subscription decision?

Hypotheses Addressed

1. Customers contacted through cellular communication have a higher probability of sub scribing compared to those contacted by telephone.

2. Customers with a successful prior campaign outcome are more likely to subscribe aga in than those with failed or no prior outcomes.

3. The frequency of contacts with a customer has a measurable effect, where too few or t oo many contacts may reduce subscription likelihood, while an optimal range increase s it.

Data Preprocessing

Initial Data Import and Cleaning

Before beginning to adequately clean the dataset, a few critical cleaning processes must be completed to ensure that the data is ready for cleaning. The steps were:

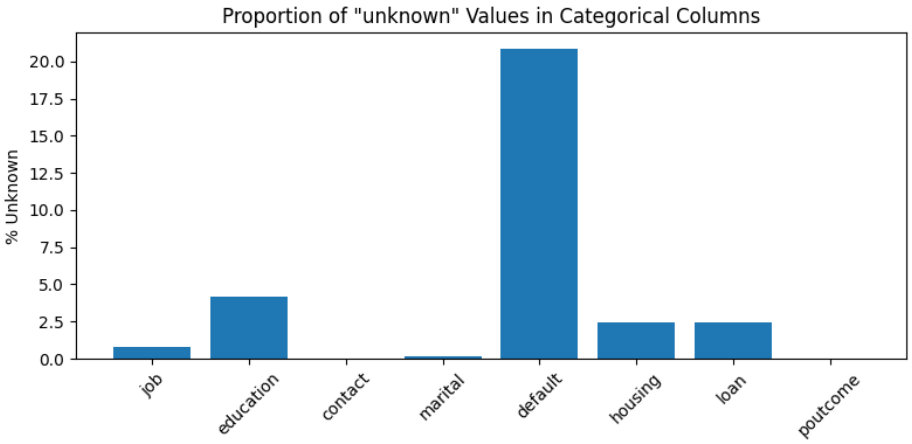
⇒ The dataset was unclean, with special characters in all columns and improper splitting. So, it was critical to reorganise the columns first.

⇒ The original dataset contained a target column named y that indicated whether a client subscribed to the marketing campaign. To improve comprehension and clarity, I changed the column name to subscribed.

⇒ Sorted the dataframe by age (low to high).

Summary Table of Missing/Unknowns

	variable	missing_count	missing_%	unknown_count	unknown_%
0	age	0	0.0	0	0.00
1	job	0	0.0	330	0.80
2	marital	0	0.0	80	0.19
3	education	0	0.0	1730	4.20
4	default	0	0.0	8595	20.88
5	housing	0	0.0	990	2.40
6	loan	0	0.0	990	2.40
7	contact	0	0.0	0	0.00
8	month	0	0.0	0	0.00
9	day_of_week	0	0.0	0	0.00
10	duration	0	0.0	0	0.00
11	campaign	0	0.0	0	0.00
12	pdays	0	0.0	0	0.00
13	previous	0	0.0	0	0.00
14	poutcome	0	0.0	0	0.00
15	emp.var.rate	0	0.0	0	0.00
16	cons.price.idx	0	0.0	0	0.00
17	cons.conf.idx	0	0.0	0	0.00
18	euribor3m	0	0.0	0	0.00
19	nr.employed	0	0.0	0	0.00
20	subscribed	0	0.0	0	0.00
21	age_zscore	0	0.0	0	0.00
22	duration_zscore	0	0.0	0	0.00
23	campaign_zscore	0	0.0	0	0.00
24	_camp_bin	0	0.0	0	0.00



Key Observations

⇒ **Missing Values:** The dataset is readily apparent, with zero missing values, suggesting great data completeness and integrity for analysis.

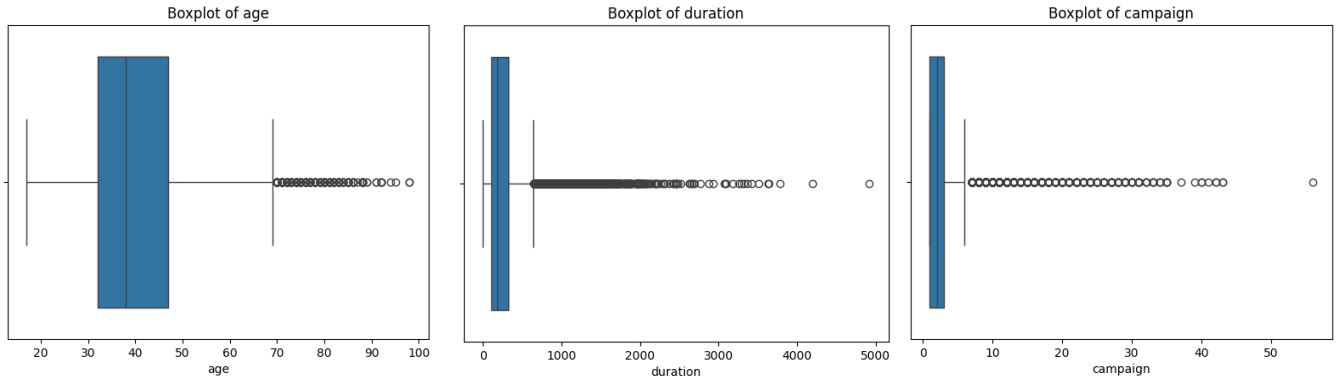
⇒ **Unknown Values:**

- The proportion of unknown values varies between features
  - Default: 20.88% of values are “unknown”, represents a significant amount of the data.
  - Education: 4.20% “unknown”
  - Job: 0.80% “unknown”
  - Housing and loan: 2.40% “unknown”
  - Marital: 0.19% “unknown”

⇒ **Result for Analysis:**

- Absence of missing data improves the validity of future analyses and models.
- Consider "unknown" items in key category columns, particularly the large proportion in "default," since these might affect model accuracy and interpretation. Appropriate treatment or stratification of "unknowns" may be required in later analytical processes.

Outlier Detection

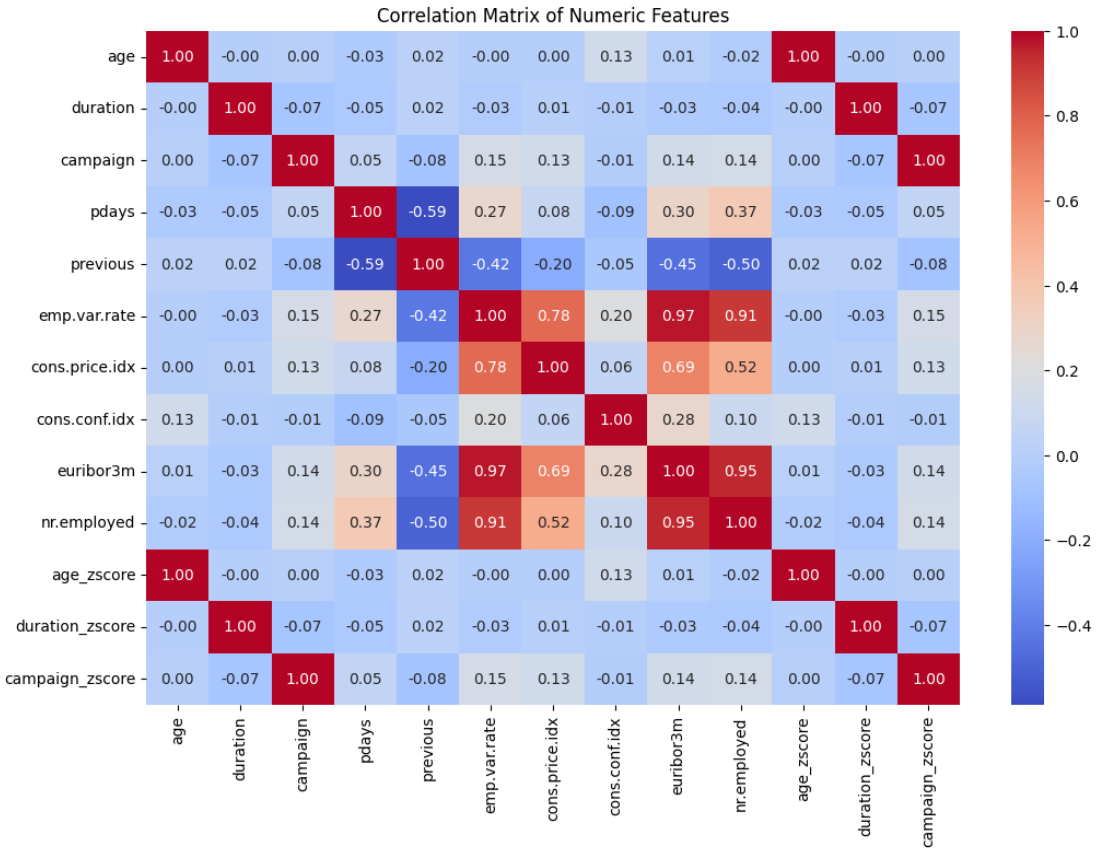


Age: The age range is primarily between 30 and 50 years old, with occasional outliers above 70 signifying older consumers who may require verification.

Duration: Call durations vary greatly, with the majority being less than 1000 seconds but some surpassing 4000 seconds, which can have a substantial influence on model performance and should be treated with caution.

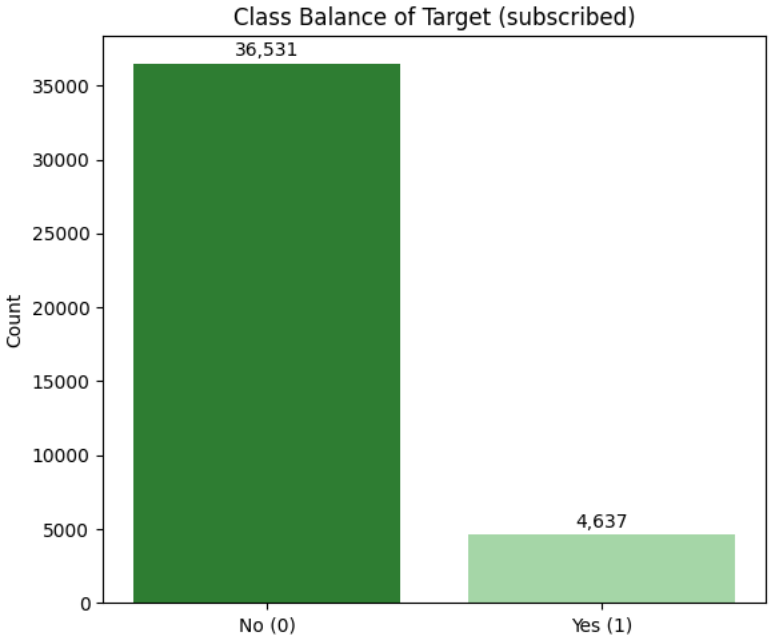
Campaign: Campaign contact counts are substantially skewed, with most consumers reached less than five times, but some receiving over thirty calls, and in extreme situations, over fifty, necessitating particular attention in the study.

Correlation Matrix



The correlation matrix indicates largely moderate correlations, with the exception of high positive correlations (>0.9) between several economic indicators (emp.var.rate, euribor3m, nr.employed) and a significant negative correlation (-0.59) between pdays and prior. This identifies redundant characteristics and directs feature selection for modelling.

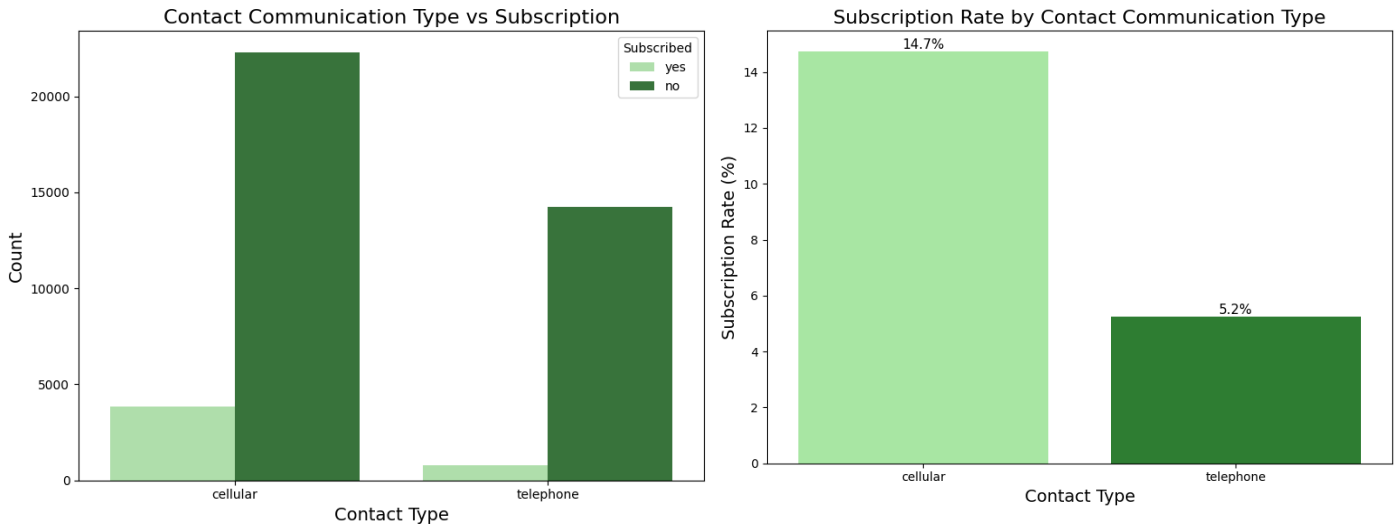
Class Imbalance



The goal variable "subscribed" is extremely lopsided, with 88.74% non-subscribers and 11.26% subscribers. This mismatch may bias models towards the majority group, necessitating approaches such as resampling to assure accurate subscription projections.

Exploratory Data Analysis

Impact of Contact Communication Type on Subscription

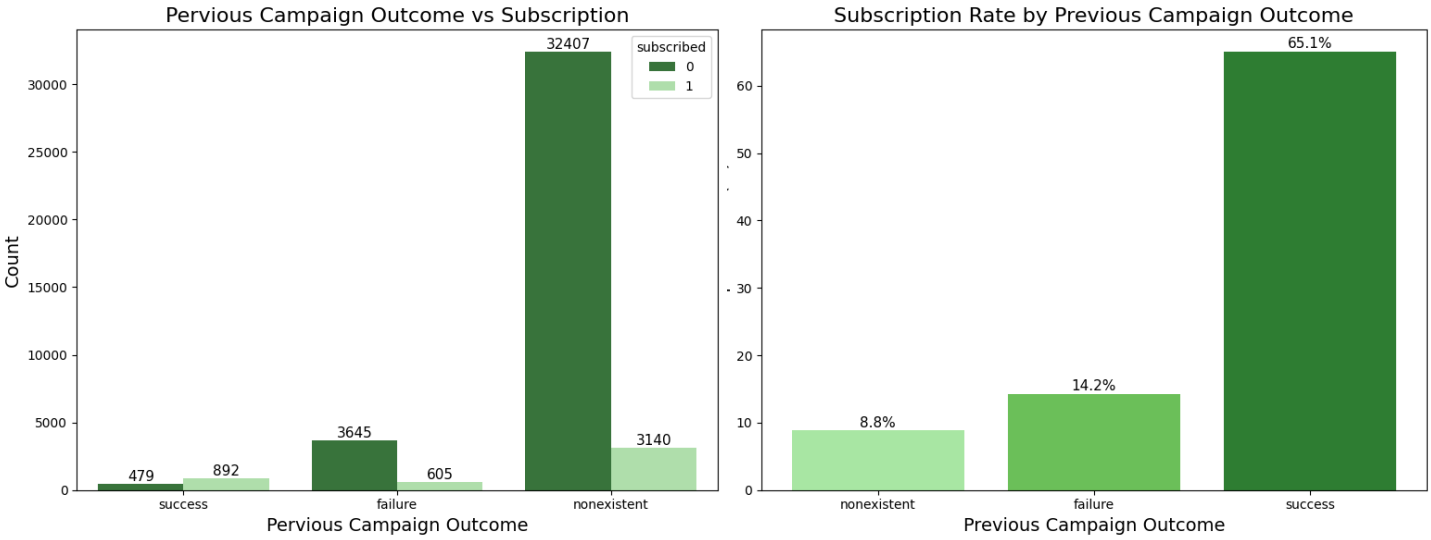


Analysis

⇒ The first figure shows the total number of consumers reached via cellular or telephone, as well as their subscription status. It demonstrates that most clients were contacted by cellular, with a higher number subscribing than telephone interactions. Despite more customers not subscribing in both groups, cell phone interactions resulted in a much larger number of subscriptions.

⇒ The second graphic shows the subscription rates for each contact type. Customers reached by cellular had a much higher subscription rate of 14.7%, compared to only 5.2% for phone calls.

Impact of Previous Campaign Outcome vs Subscription



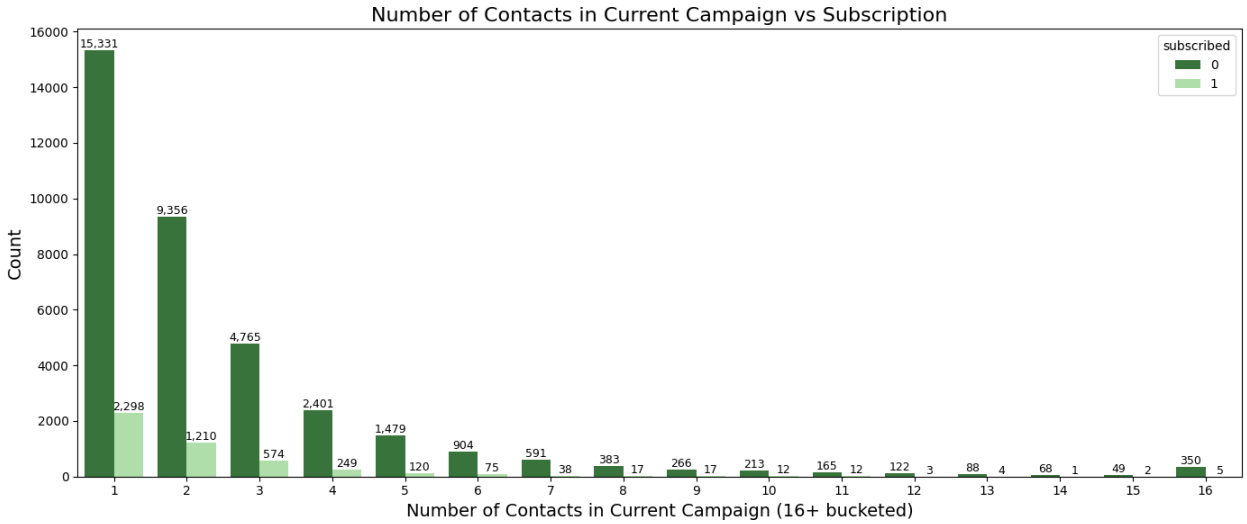
Analysis

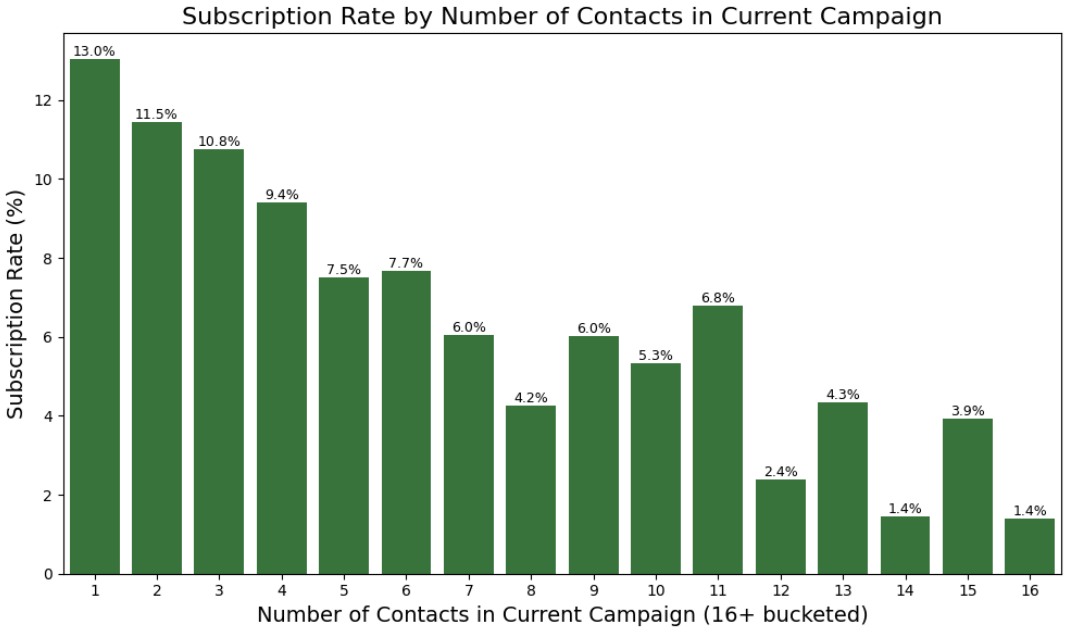
⇒ There is a strong positive correlation between past success and future subscriptions, as customers who had a "success" outcome in the previous campaign have the highest subscription counts and rates.

⇒ Subscription counts are much lower for customers with "failure" or "nonexistent" prior campaign outcomes; the nonexistent group is the largest but has the lowest conversion rate.

⇒ Subscription rates for every previous campaign outcome are shown in the second chart. The subscription rate for customers who have had a successful previous outcome is 65.1%, which is significantly higher than the rates for customers who have had previous failures (14.2%) and those without prior campaigns (8.8%).

Number of Contacts Vs Subscription





Analysis

⇒ The majority of consumers receive one or two interactions during the campaign, with fewer customers reached several times, indicating that outreach focusses mostly on first touchpoints.

⇒ Subscription rates are highest for consumers contacted once (13.0%), and they continue quite robust for two to three interactions before gradually diminishing and falling below 10% after three encounters.

⇒ Excessive contact efforts, particularly 14 or more, result in the lowest subscription rates (1.4%), suggesting that too many follow-ups may reduce a customer's probability to subscribe.

Conclusion

The study found that cellular contact is far more successful than telephone contact in promoting subscriptions, with cellular outreach resulting in approximately three times higher subscription rates. Furthermore, clients with good prior campaign outcomes have much greater subscription probability, emphasising the value of previous positive involvement. Contact frequency is a significant aspect. While most clients receive one or two interactions with a high subscription rate, numerous contact efforts significantly diminish subscription chance. These findings highlight the necessity for targeted, well-planned marketing activities that focus on efficient communication channels, utilise prior campaign accomplishments, and maximise the number of client encounters.

Recommendation

To improve marketing efficacy and membership rates, the telecom firm should prioritise cellular connectivity over phone calls. Marketing efforts should be directed towards engaging clients with a track record of successful campaign outcomes, while establishing focused tactics to convert those with no or failed past engagements. Finally, campaigns should balance contact frequency by focussing on quality initial touches rather than excessive follow-ups, so preventing consumer fatigue and increasing conversion rates.

References

⇒ Moro, S., Laureano, R., & Cortez, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. https://doi.org/10.1016/j.dss.2014.03.003

⇒ Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing, 14,243257. https://doi.org/10.1016/j.eswa.2008.05.021

⇒ Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636. https://doi.org/10.1016/j.eswa.2008.05.021

⇒ Tsiptsis, K., & Chorianopoulos, A. (2009). Customer segmentation techniques in CRM: A literature review. International Journal of Business Intelligence Research, 1(3), 56-68.

⇒ Lin, Y., & Tsai, C. (2020). Effectiveness of multi-channel marketing strategies in telecom industry: An empirical study. Journal of Business Research, 112, 155-164. https://doi.org/10.1016/j.jbusres.2019.10.026

⇒ López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141. https://doi.org/10.1016/j.ins.2013.07.007