Statistics and Probability
20592 – DSBA

Group Work Guidelines

Objective of the work is to analyze Churn phenomenon on data coming from Telco Sector and propose the best Target for a Retention Commercial Campaign

Dataset to be analysed: https://www.kaggle.com/blastchar/telco-customer-churn

Each row represents a customer, each column contains customer's attributes. The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target.

Guidelines for the Analysis

Describe Churn Phenomenon through univariate and bivariate analysis, managing potential issues related to data: outliers, missing,

Estimate a Scoring Model through Logistic Regression in order to predict Churn Phenomenon. Please, consider the variable Churn in the provided data set as dependent variable and other variables as potential inputs for your model, of course is on charge to you decide which independent variables to use.

Describe the model in term of interpretation and performance.

Use Lift Chart and Gain Chart to solve the Business Case here below described:

- Objective: Identify the Target size (in term of number of customers) to contact through the Retention Commercial Campaign in order to have a Break-even Balance: costs for contacts have to be compensated by revenues obtained by retained customers.
- Inputs to set-up the Business Case Simulation:
  - ✓ Cost for each single contact: 5 euros
  - ✓ Expected Revenues for each retained customer: 25 euros
  - ✓ Expected Retention rate obtained through the Commercial Campaign: 25% (consider the number of people who really churned by target size, we're hypothesizing that 1 out of 4 of them would not have abandoned the Company if contacted by the campaign)

A possible **path** is to perform the following analyses:

- exploratory data analysis (EDA) in order to check the quality and the coherence of the data
  - How variables are codified?
  - What about missing values?
- EDA in order to have a first insight about important patterns (as group of customers)
  - Factorial analysis and/or cluster analysis
  - Data visualisation (of the features and of the target conditioned to the features)
- Statistical models to model the "Churn" behaviour
  - Linear models (i.e. logistic regression)
  - Check if it's possible to use transformation of the variables to get better results
- Resampling methods to asses the quality of the estimates (confidence intervals); comparison with parametric results

The following **sections of the final report** have to be provided:

1. Description of churn with univariate/bivariate analyses, with outliers detection (data audit)
2. Proposal of an optimal model with logistical regression
3. Building of lift and gain charts
4. Cut-off selection, including the economical parameters

The following aspects will be **evaluated**:

- care about details in the EDA phase
- meaningful data visualizations
- checking of the assumptions of the selected statistical model
- tidiness of the code, presence of comments, use of functions or classes, use of proper libraries (pandas, …)

The most important point is to provide description and interpretation of the findings, from a statistical and business point of view.

Deliverables

1. Power Point Presentation reporting your, objective, main insights obtained by preliminary analysis, estimated scoring model and business case simulations. All slide have to report main statistical outputs and your comments just for statistical and business point of view. Power Point Presentation has to be sent to alberto.saccardi@unibocconi,it and davide.posillipo@unibocconi.it by December 24th

2. Jupyter Notebook with the full code used to produce the results (or a .py file).