BETTERFLY

Case Interview

Data Science

March 2023

Instructions

- It would be best to deliver your solution to the problems in a GitHub repository.
- Inside this repository, we must be able to find all the files used to resolve the problems.
- You must do it in a Jupyter Notebook. For this, we ask that the language used is Python 3.
- Inside the notebook, you must indicate the details of the solved questions.
- It would help if you clearly described the different assumptions you are assuming in the case resolution. Then, based on this, we can follow your logical structure to develop the solution.
- You must send the link to the repository to fhernandez@betterfly.com
- **Important to note: All information in this Case is fictitious and was artificially generated for the purpose of being used with the interviewees.**

First Problem - Churn Clients

Since Betterfly is a platform that delivers services in a B2B channel, when a customer unsubscribes from a service, multiple users are unsubscribed, directly impacting the company's recurring revenue. Given this, it is critical to predict when a customer approaches churn and the main factors involved in this decision. Currently, there are clustering models that group both customers and users based on their behavior, which will be input for prediction.

Methodology expected

For the above, you will be provided with a .zip file containing the data sources needed to solve this problem. With these datasets, perform the following tasks, describing:

- Review the dataset and variables. Then, make an initial exploratory analysis based on the variable dictionary.
- Perform data cleaning and feature engineering if necessary. Describe the rationale for your decisions.
- Build a churn prediction model that ranks customers from most to least likely to churn. Perform a performance analysis of the resulting model. Justify the choice of metrics and the methodology used to obtain them.
- Considering there is a limited capacity to perform preventive actions on each customer, define a strategy to use the results obtained by the model and describe it to be presented to the corresponding stakeholder.

Second Problem - Regression Analysis

You are in charge of conducting a study regarding cancer death rates. Information was collected from censuses, census data, and cancer foundations. This information is available in the *reg_cancer_deathrate.csv* dataset, which contains information on 3047 communes in the country summarized in 19 variables. This study aims to construct a Multiple Regression Model that explains the variable *DeathRate*, corresponding to the average cancer mortality per capita.

First stage: Data processing.

1) Perform an exploratory analysis of the dataset to detect and clean missing values and possible anomalies in the data. It is important to note that all variables must be positive, so if you identify an abnormal value, interpret it as missing data and eliminate it.
2) Perform a joint behavior analysis between the variable of interest and the covariates to determine which would contribute the most to the regression model design. In addition, it is advisable to perform a correlation analysis and justify the choice.
3) Evaluate making transformations to the variables to improve the linear relationships. Justify your decision.

Second stage: Modeling.

1) Propose three different Multiple Regression models with the variables selected in the previous stage. Then, choose the best model according to the goodness-of-fit criteria. Finally, plot the AIC to make the selection.
2) Study the multicollinearity problem in the variables selected in the previous model. If it exists, build a new model that does not present this problem.
3) For the model constructed in the previous point, carry out a study of outliers and influential points. If they exist, identify them and build the selected model again without these observations. Finally, comment on the results obtained.
4) Explain if the elimination of outliers and influential points has any effect on the model assumptions. Then, perform a residual study on the models in 2) and 3) and compare results. Make your analysis through graphs and hypothesis tests.

Deliverable: Repository with all pertaining files.

POWER YOUR TEAM.
POWER YOUR BUSINESS.
POWER THE WORLD.