ROCL Group - Data Science Specialization (Healthcare Drug Persistence)

Members:

Reem Aboelsoud, Germany, niologic GmbH
reem.abulsoud@gmail.com

Oana Cucu, United Kingdom, University of Bristol
omcucu@gmail.com

Leslie Hadaway, Saint Vincent and the Grenadines, Trinity School of Medicine
lihadaway@outlook.com

Cindy Huynh, United States of America, University of Texas at Austin
cindytrinhhuynh@gmail.com

Problem description

The ABC pharmaceutical company aims to conduct an investigation into the determinants that affect the longevity of a drug, specifically focusing on the persistence rates among patients who are prescribed the treatment for a particular medical condition. The data available for analysis encompasses patient demographics, risk factors, pre-existing conditions, and comorbidities, along with limited details about the attending physician (such as specialization in a specific field or being a general practitioner). No specific information pertaining to the drug under investigation is provided.

To build a model of persistence predictors based on the available data, a classification approach can be applied. The data can be divided into input features (predictors) and the target variable (persistence).
A possible classification of the data:

1. Input Features (Predictors):
    - Patient demographics: Age, gender, ethnicity, etc.
    - Risk factors: Smoking status, obesity, family history, etc.
    - Underlying conditions and comorbidities: Diabetes, hypertension, cardiovascular diseases, etc.
    - Physician information: Specialist or general practitioner, experience level, etc.
    - Psychosocial factors: Patient beliefs, attitudes, social support, healthcare provider-patient interactions, etc.
2. Target Variable:
    - Persistence: This can be represented as a binary variable, indicating whether a patient continued or discontinued the prescribed treatment. For example, "1" can represent continued use, and "0" can represent discontinuation.

Once the data is classified into predictors and the target variable, we can apply various classification algorithms to build a model – a commonly used algorithm is logistic regression.
The model-building process involves training the model using labelled data (where persistence outcomes are known) and then evaluating its performance using appropriate metrics such as accuracy, precision, recall, and F1 score. Additionally, techniques like feature selection, cross-validation, and hyperparameter tuning can be employed to optimize the model's performance.

By applying this classification approach to the data, we can develop a model that predicts drug persistence based on the provided predictors, which can provide valuable insights and assist in decision-making processes related to treatment adherence.

Business understanding

Drug persistence is a critical factor in determining the success and effectiveness of a given drug. It is commonly recognized that a comprehensive understanding of drug persistence requires examining a combination of internal (participant-related) and external (e.g., drug- or physician-related) factors. However, in this particular case, our investigation focuses on understanding mainly the internal, health-related factors that influence patients' continued use of the drug, with some analyses on social factors such as physician data and participant demographics.

The aim of our analysis is to gain insights into the underlying motivations and barriers that influence medication adherence. This understanding can help identify target consumer groups who are more likely to benefit from the specific drug developed by ABC pharma company. While considering drug-related factors, such as side effects, is important in evaluating drug persistence, this investigation primarily emphasizes the participant's perspective and the influences that shape their decision to continue using the drug. By focusing on the participants' health-related aspects, we can develop patient-centred strategies, interventions, and support systems to promote better medication adherence and improve overall treatment outcomes.

By aligning our research with the common approach in investigating drug persistence, we aim to contribute to the growing body of knowledge in this field and provide valuable insights for pharmaceutical companies, healthcare providers, and researchers working towards enhancing patient compliance and maximizing the potential of drug therapies.

Project lifecycle/deadline

May 26th: Complete Exploratory Data Analysis (EDA) and reach a consensus on data visualization techniques, incorporating valuable contributions from all participants.

June 10th: Finalize the model development phase by building a logistic regression model and validate it using AUC-ROC analyses. Each participant should work individually on this task.

June 30th: Discuss and agree upon the common results and procedures derived from the individual model developments. Finalize the code implementation and prepare the report for submission.

The revised plan maintains the key milestones and deadlines while providing clearer instructions regarding the tasks to be completed. It emphasizes the collaborative nature of the EDA and visualization phase, while acknowledging that the model development and validation steps should be conducted individually. Lastly, it highlights the importance of reaching a consensus on the results and procedures before finalizing the code and report for submission.

Data Intake Report

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	922 KB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Mention approach of dedup validation (identification)

The present data contain no duplicates or missing values; however, we will still be performing the following steps for the purposes of feature engineering:

1. Consistency checks: Ensuring that the data is internally consistent and follows predefined rules or standards. This can involve checking for logical inconsistencies or discrepancies within the dataset, such as contradictory values or illogical relationships between variables.
2. Accuracy assessment: Verifying the accuracy of the data by comparing it with external sources or known reference data. This can involve cross-referencing the data with authoritative databases or conducting data validation against established benchmarks or ground truth information.
3. Outlier detection: Identifying and examining outliers or extreme values that may be unexpected or potentially erroneous. Outliers can be indicative of data entry errors or other data quality issues, and their identification can help in validating the data.
4. Data format and type validation: Verifying that the data is in the expected format and adheres to the specified data types. This includes checking for formatting errors, data truncation, or mismatches between the defined data types and the actual values.

- Mention your assumptions (if you assume any other thing for data quality analysis)

For the proposed logistic regression model, we will make the common relevant assumptions:

1. Binary outcome: Logistic regression assumes that the dependent variable or outcome variable is binary or dichotomous. It should represent two mutually exclusive and exhaustive categories, such as "success" or "failure," "yes" or "no," etc.
2. Linearity of the logit: Logistic regression assumes that the relationship between the independent variables and the log odds of the outcome variable (logit) is linear. In other words, the effect of the independent variables on the logit is additive and proportional.

3. Independence of observations: Logistic regression assumes that the observations in the dataset are independent of each other. There should be no correlation or dependence among the observations.
4. No multicollinearity: Logistic regression assumes that there is no perfect multicollinearity among the independent variables. Perfect multicollinearity occurs when there is a perfect linear relationship between two or more independent variables, making it difficult to estimate their individual effects.
5. Absence of influential outliers: Logistic regression assumes that there are no influential outliers in the dataset that could unduly influence the estimation of model parameters.
6. Large sample size: Logistic regression performs well with a relatively large sample size to ensure stable parameter estimation and accurate inference.

Github repo link:

https://github.com/lihadaway/data-glacier-internship/tree/c191138ff6dae50782971f90893e0b8791ef70d1/drug-persistency-classification