Roche US

Roche is a Swiss multinational healthcare company that is the world's largest biotech company. They are a leading provider of in vitro diagnostics, a global leader in cancer treatments and a frontrunner in personalized healthcare. This project is working with a Roche U.S. team.

Project Description

The goal of this project is to examine trends in Roche instrument repair case notes, using natural language processing and unsupervised machine learning methods. Broadly, the project HDAG will engage with Roche would consist of two main stages.

1) **Feature Extraction:** The HDAG team will first extract useful features from the raw free-form text data. This will be done using natural language processing (NLP) methods, including sentiment analysis using language models, topic modeling with methods such as Latent Dirichlet allocation, keyword extraction using statistical methods and machine learning, and/or other related NLP methodologies.

2) **ML Modeling/Statistical Analysis:** The HDAG team will employ statistical methods, in particular clustering techniques to perform unsupervised text categorization and identify trends in the instrument repair case notes data.

Internal Partners: Data Science Manager for Business Analytics and Strategic Effectiveness (BASE)

Preferred Coding Languages: Python

Specific Skills

1. Natural Language Processing
2. Machine Learning Modeling
3. Clustering Techniques

Expected Technical Difficulty: **Advanced**

Roche Swiss 1

Roche is a Swiss multinational healthcare company that is the world's largest biotech company. They are a leading provider of in vitro diagnostics, a global leader in cancer treatments and a frontrunner in personalized healthcare.

Project Description

Roche contracts vendors known as contract research organizations (CRO) to conduct clinical trials for them. The cost of these contracts are negotiated by procurement based on the requirements of each individual study (e.g. number of patients, frequency of patient visits, therapeutic area). Currently, there is no well-defined, strategic rationale for this negotiation from Roche's side, meaning there is no clear monetary amount that can be assigned to certain requirements within a contract. At the same time, the cost of contracts varies significantly; thus, it is expected that there is large potential here for cost saving through better informed negotiations.

The goals of this case will be to characterize, quantify, and rank the factors driving the cost of these contracts for clinical trials. The team will also explore combinations of these factors ("composite factors") and their effect on contract cost (e.g. using multivariate regression). Finally, a model will be developed to predict the cost of new contracts given a set of values for the previously identified drivers of cost.

Internal Partners: This project is in collaboration with the Strategic Operational Intelligence team at Roche.

Datasets: Past contract negotiations dataset over 391 (~390 studies) containing the cost of the contract, the CRO that conducted the study, and study features such as therapeutic area, number of patients, etc.

Preferred Coding Languages: Python

Specific Skills

1. Statistical Modeling: developing and fitting models to characterize, quantify, and rank factors driving the cost of study contracts.
2. Machine Learning: training, testing, and deploying models to predict the cost of new contracts given study features.
3. Data Visualization: creating useful visualizations for insights summary and interpretation

Expected Technical Difficulty: **Intermediate**

Roche Swiss 2

Roche is a Swiss multinational healthcare company that is the world's largest biotech company. They are a leading provider of in vitro diagnostics, a global leader in cancer treatments and a frontrunner in personalized healthcare.

Project Description

As a pharmaceutical company, Roche requires government approval to bring its products to market. Each of these filings requires significant documentation: Roche estimates that each submission requires around 1500 documents (~2000 pages) of text. There is significant overlap between separate documents: often, all that needs to be changed is a year or molecule. Roche wants a system to automatically find documents similar to the one they are trying to produce to accelerate the filing creation process.

Determining resuability manually would be difficult: at roughly 45 2000-page filings a year, Roche estimates that it would take 25 years to manually standardize documents. As such, Roche wants to build a system where reusability can be determined algorithmically. Roche wants us to develop this model.

The project has three parts. First, we will construct and visualize a network based off of a document similarity metric of our choice. Second, we will determine a reusability metric, and train a model to identify resusability between different documents in this network. Finally, we will develop an algorithm to identify resuability paths between documents.

Internal partners: This project is in collaboration with the Strategic Operational Intelligence team at Roche.

Dataset: Roche intends to give us access to historical filings from past clinical trials, starting with the past twenty years of documentation for 50 molecules.

Preferred coding language: Python

Specific Skills: Network visualization, machine learning, neural network-driven NLP, discrete optimization

Expected Technical Difficulty: **Advanced**