Module 6

XN PROJECT: EDA & PROJECT PLAN

ALY 6080 Integrated Experiential Learning

Instructor: Prof. Allen Bryce

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA

Submitted By: Group 2

Alok Madamanchi

Kashika Tyagi

Md Tajrianul Islam

Sandeep Srivatsav

Spurthi Patnam

Sweta Mankala

XN PROJECT: EDA & PROJECT PLAN

The purpose of this assignment is to create a roadmap for the final project. The project we are working on is based on Leukemia disease. We aim to analyze the top 10 targets of Leukemia Indication for our sponsor - Silicon Therapeutics.

Analytic approach to answer the project requirements :

We believe, it is important for us to understand the important attributes behind drug discovery and accurately identify the interactions between ligands and target to choose the right models we will be using. ChEMBL is an open-access, large scale drug discovery resource, which provides us with some property calculation values, like aLogP, HBA, HBD, Mwt, Ro5, RTB, HAtoms, pKa, LogD, QED also there is a pChembl value which represents the potency of the small molecules. We aim to find the correlation between these values and a drug-like molecule being currently used in the treatment of Leukemia. But then again, these property calculation values may vary depending on a particular target either molecular (like protein/ nucleic acid) or non-molecular (like cell line/ tissue). We might be able to better understand these correlation as we do some actual EDA.

Initially, we are planning to extract the data from ChEMBL database on mechanism level using web services offered by ChEMBL which can be accessed by any of the programming languages. Thus, we will be using Python language to extract the data from ChEMBL database by installing the standard Python package manager (PyPi) to query the ChEMBL database. Using Python's client code hosted on GitHub, we will be performing descriptive statistics on the extracted data followed by Data Preparation before applying the models.

For this project, we will be using Python and Tableau:

Python - To understand the correlation between the variables and identify the meaningful patterns.

Tableau – We will be using tableau to build interactive charts and graphs to identify the top 10 small molecules and targets in the drug discovery domain.

We will be following the CRISP DM methodology to address our business problem.

Milestones to measure progress :

To successfully address our business problem, we designed a plan to measure the progress of the project based on certain factors as listed below:

Define – Clearly define the business problem so that we can build an effective route map to address it. This involves understanding the data to work with and technologies/resources to be used.

Measure - We understand the dataset and the variables in it. This involves the process of data collection as we must collect features that can help with our target centric analysis. We will also clean the data to ensure it has no missing or null values as cleaner the data, better the data analysis will be.

Analyze – We will perform EDA to identify trends/patterns using either Tableau or Python. We will understand the correlation between variables for feature selection. We will build various classification models to predict the target for Leukemia.

Design - We evaluate the models based on the performance measures and select the best one and test it.

Verify - To check if the predicted target compounds of small molecules can help address the business problem which is to identify the target responsible for the disease Leukemia.

Job assignments if working in a group :

1. Md Tajrianul Islam :

I can be good in preparing EDAs and doing any domain related research. But I am eager to learn when it comes model building and data preparation and will really appreciate, if we can work on group calls.

2. Kashika Tyagi :

I will perform descriptive statistics to understand the data and identify meaningful patterns using Python language. I will then prepare the data (such as handling missing values, scaling data using normalization techniques) to perform Exploratory Data Analysis followed by applying the regression models in Python. I am going to refer ChEMBL documentations and Python's client

code hosted on GitHub to achieve the objective of identifying the targets for Leukemia Indication.

3. Spurthi Patnam :

I will contribute to the exploratory data analysis by performing univariate and multivariate analysis using python. I will be using the pandas package to perform this analysis. Using pandas, I will analyze each column that can help discover the appropriate drug target. Also, I will create interactive tableau dashboards to present this analysis.

4. Sandeep Srivatsav :

Using the processed data, I will perform exploratory data analysis using Tableau/Power BI to understand the patterns of data and perform statistical analysis such as correlation between variables using Python. I will also build classification models that can help identify the best targets for Leukemia and evaluate all the models.

4. Alok Madamanchi :

I am good with exploratory data analysis. Can perform data visualization in PowerBI/tableau to understand the patterns in the data. I can also work on prioritizing the targets that we are working on. I would also like to perform univariate and multivariate analysis graphically.

5. Sweta Mankala :

Data extraction using python API and data cleaning which eliminates any missing/null/duplicate values. I will also perform data preparation to normalize the data, remove outliers and scale the data. Using the processed data, I will build data models using classification techniques to predict the targets for Leukemia. I will also test the models to ensure it aligns with the business goals

Key risks and strategies to mitigate them :

The primary risk would be if the business problem is not well-defined as the entire process of data analysis would not answer our business question. Data gathering is the significant to our analysis to fetch the right variables that help us with our target centric analysis. If the records collected are not enough to help us understand our analysis, we might have to fetch data from external sources such as PubChem.

Another key risk is in this project would be properly understand the attributes and their importance, as none of us is a medicinal chemistry expert. But fortunately, we do not need to depend on any mediator, our professor is also a directly stakeholder of the project. So regular and effective communication would be the key behind mitigating the risk of getting derailed from our project goal.

We are also concerned about how a parent molecules' or similar druglike molecules' property may or may not affect the possibility of small molecule being effective for the target. Again, more extensive research and communication with our professor is the ultimate way of overcoming this concern.

Measure of success :

Measure of success would be being able to provide our stakeholders with actionable information. For that, we aim to predict the right target compounds that cure the disease Leukemia. With the predicted targets found for the data that is processed, we will have to ensure that the predictions align with the goals of our sponsor as this should be beneficial to detect the right target compounds for any additional data.

Presentation method and delivery of your proof of concept :

We will be using the storytelling style to present our analysis and deliver our proof of concept. Storytelling intends to confer concrete, numerical answer, and communicate it compellingly, turning complex analyses into actionable insights. We will tell a story regarding the drug targets that can be used to discover targets for leukemia using the CHEMBL database, including CHEMBL Id, Parent molecule type, Parent molecule name, Max phase indication, etc. The data story presents a great lead to the drug discovery of leukemia.

References :

1.) chembl/chembl_webresource_client
Chembl
https://github.com/chembl/chembl_webresource_client