Proposal for Capstone Project II

Personalized Medicine: Redefining Cancer Treatment

Data Set:

https://www.kaggle.com/c/msk-redefining-cancer-treatment/data

Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers). Currently this interpretation of genetic mutations is being done manually. This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

Both, training and test, data sets are provided via two different files. One (training/test_variants) provides the information about the genetic mutations, whereas the other (training/test_text) provides the clinical evidence (text) that our human experts used to classify the genetic mutations. Both are linked via the ID field. Some of the test data is machine-generated to prevent hand labeling.

Problem to resolve:

To develop algorithms to classify genetic mutations based on clinical evidence (text). There are nine different classes a genetic mutation can be classified on. Since interpreting clinical evidence is very challenging even for human specialists. Therefore, modeling the clinical evidence (text) will be critical for the success of my approach.

Clients and Audiences

Biotech Company, pharmaceutical company.

Approach

1. Downloading training set and test set from Kaggle website.
2. Be familiar with the variables in the data set.
3. Explanatory analysis about the distribution of genetic mutation class, genes which have highest occurrence in each class and so on.
4. Applying NLP techniques such as Tf-idf which is known as one good technique to use for text transformation and get good features out of text for training our machine learning model.
5. Building classification model using tree based methods.

Project Deliverables

The project deliverable will contain following:

- This document explaining approach of the project.
- Presentation created on this project.
- Graphs and visualizations created for this project.
- Python code for model building with comments and details.