Predicting Cancer Survival

Abstract

The goal of this project was to use classification models to predict the survival of cancer patients to assist the oncologists choose a better treatment plan for the patients. I worked with data provided by cBioPortal portal, leveraging numerical and categorical feature engineering along with a logistic regression and gradient boosting models to achieve promising results for this binary class problem. The exploratory data analysis was communicated using python visualization libraries.

Design

This project originates from the SADIA bootcamp for data science. The data was obtained from the cBioPortal for cancer genomes, and presents a binary-class survival status of survived and deceased patients for different type of cancer patients. Classifying status accurately via machine learning models would enable the Oncologists to make better descriptions regarding the treatment plans that is chosen for the patients.

Data

The dataset contains 10,945 observations with 26 features for each, 19 of which are categorical. A few feature highlights include clinical, demographic and genetic. Nearly half of the individual features is not important to built predation models and was taken down.

Algorithms

Feature Engineering

1. Convert categorical features to binary dummy variables
2. Scale numerical data.

Models

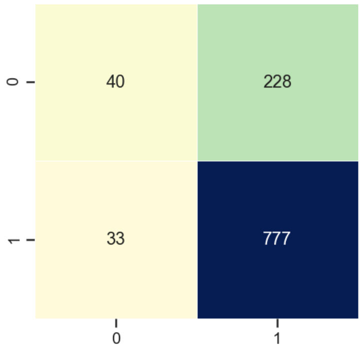
Logistic regression and gradient boosting classifiers were used before settling on gradient boosting as the model with strongest performance.

Model Evaluation and Selection

The entire training dataset of 5,897 records was split into 80/20 train vs. test.
Predictions on the 20% test were limited to the very end, so this split was only used and scores seen just once.
The official metric for DrivenData was classification rate (accuracy), F1 score and confusion matrix of the test split.

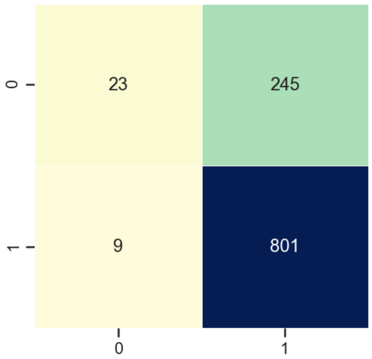
Logistic Regression:

- Accuracy: 0.758
- F1: 0.856
- Confusion matrix:



Gradient Boosting:

- Accuracy: 0.764
- F1: 0.863
- Confusion matrix:



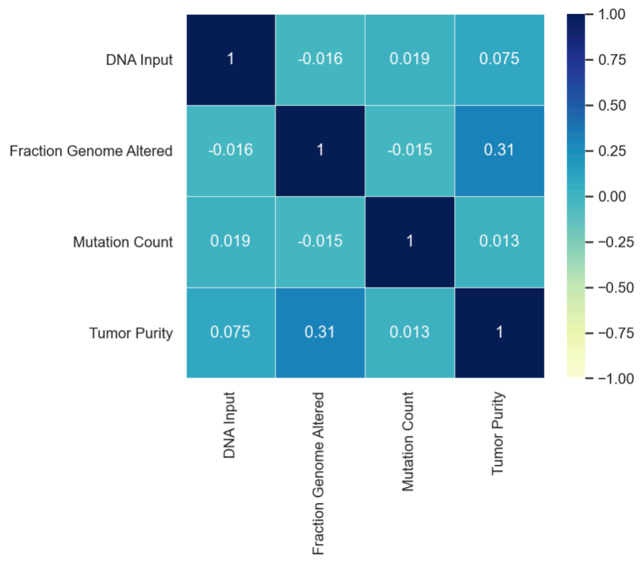
Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib, Seaborn and Plotly for visualization

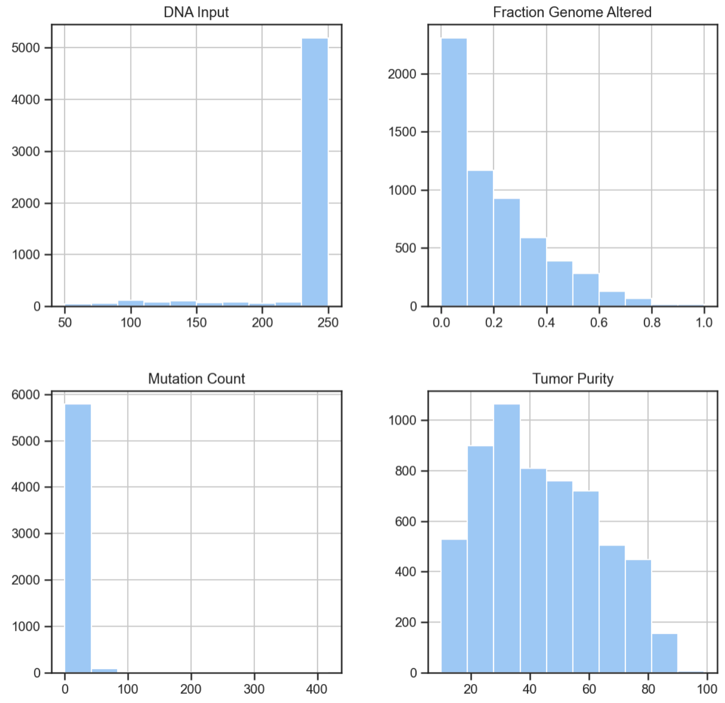
Communication

In addition to the provided slides, below are some visualizations conducted during the EDA.

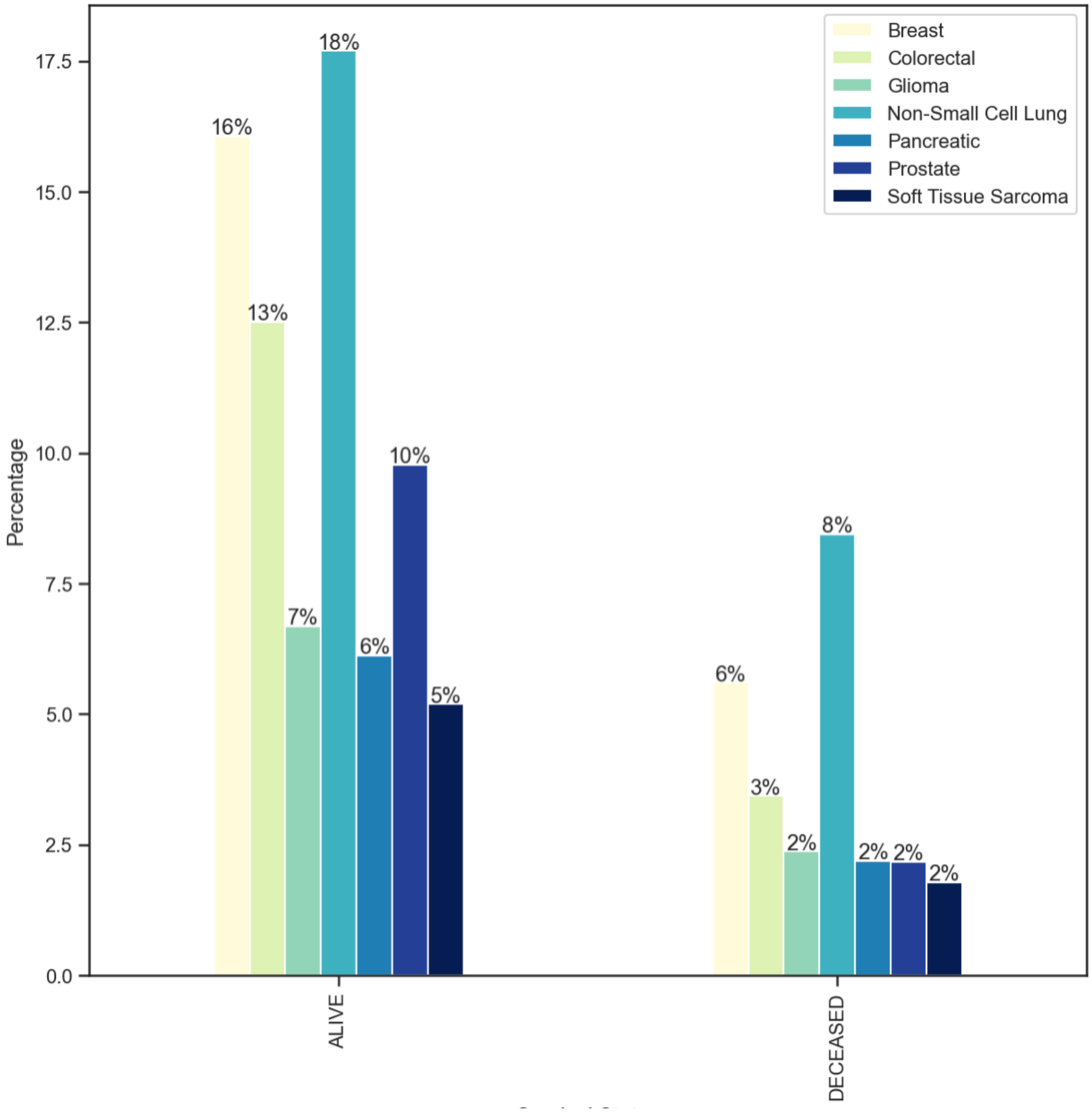
- Numerical features correlation:



- Numerical Features distribution



- Effect of cancer type on survival


Breast
Colorectal
Glioma
Non-Small Cell Lung
Pancreatic
Prostate
Soft Tissue Sarcoma
Percentage
ALIVE
DECEASED