Brain Tumor Genetic ID

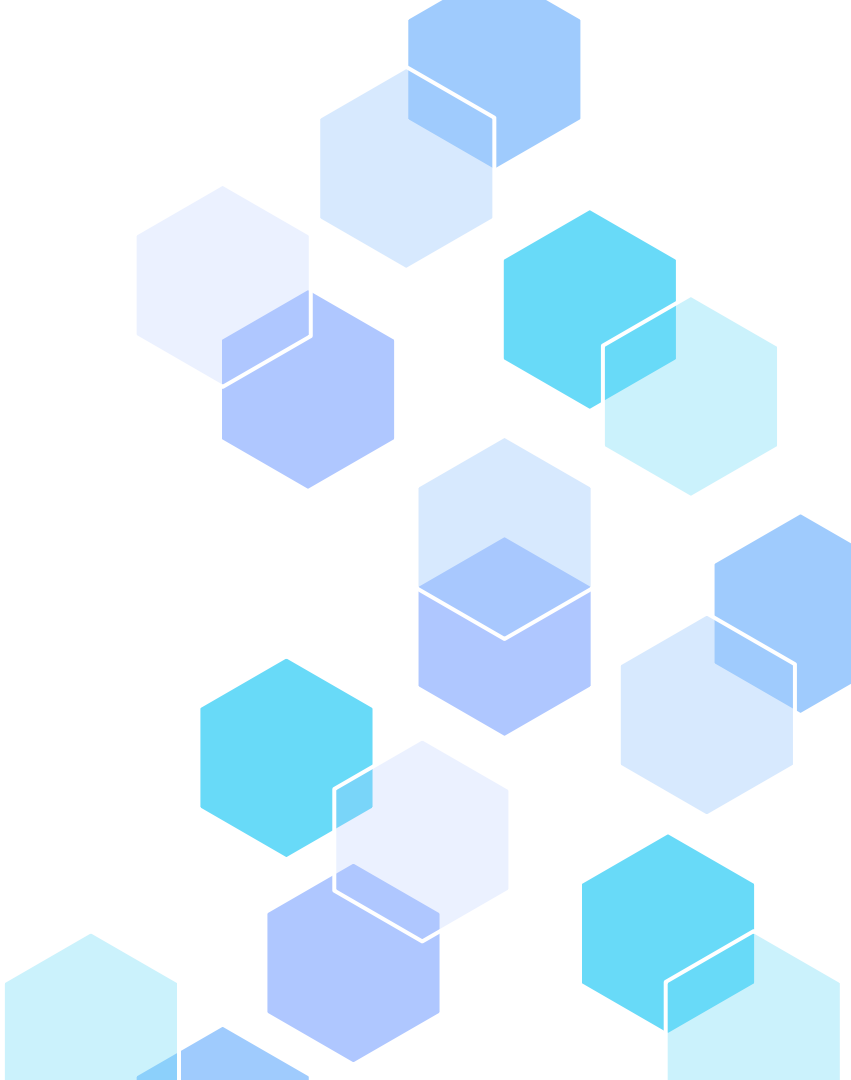
Glioblastoma Multiforme & Low Grade Glioma

Noah C. Strevell

Table of contents

01 Landscape of Oncology

And why it matters.



Introduction

Cancer rates are rising faster than ever. An estimated 41 out of 100 people the the US will develop cancer during their lifetime. As new technologies emerge, it is imperative to have a robust understanding of cancer to better aid in the fight against it.

- **1.4%** of cancers are brain tumors
- **65%** 5-year mortality rate for patients with a malignant brain tumor
- **+18,000** annual deaths from brain tumors
- **<5%** of adults with malignant brain tumors have a family history

+2,000,000
new cases in 2024

Genetics

The way forward

Who benefits?

Patients / Families

Genetic testing for cancer helps patients by enabling early detection, personalized treatment, improved risk assessment, proactive prevention, and informed family planning.

Drug Manufacturers

By enabling the development of targeted therapies, expanding precision medicine markets, and improving treatment efficacy, leading to increased demand for specialized drugs.

Insurance Providers

Insurance providers benefit from genetic testing for cancer by reducing long-term healthcare costs through early detection, targeted treatments, and preventive care, ultimately minimizing expensive late-stage interventions.

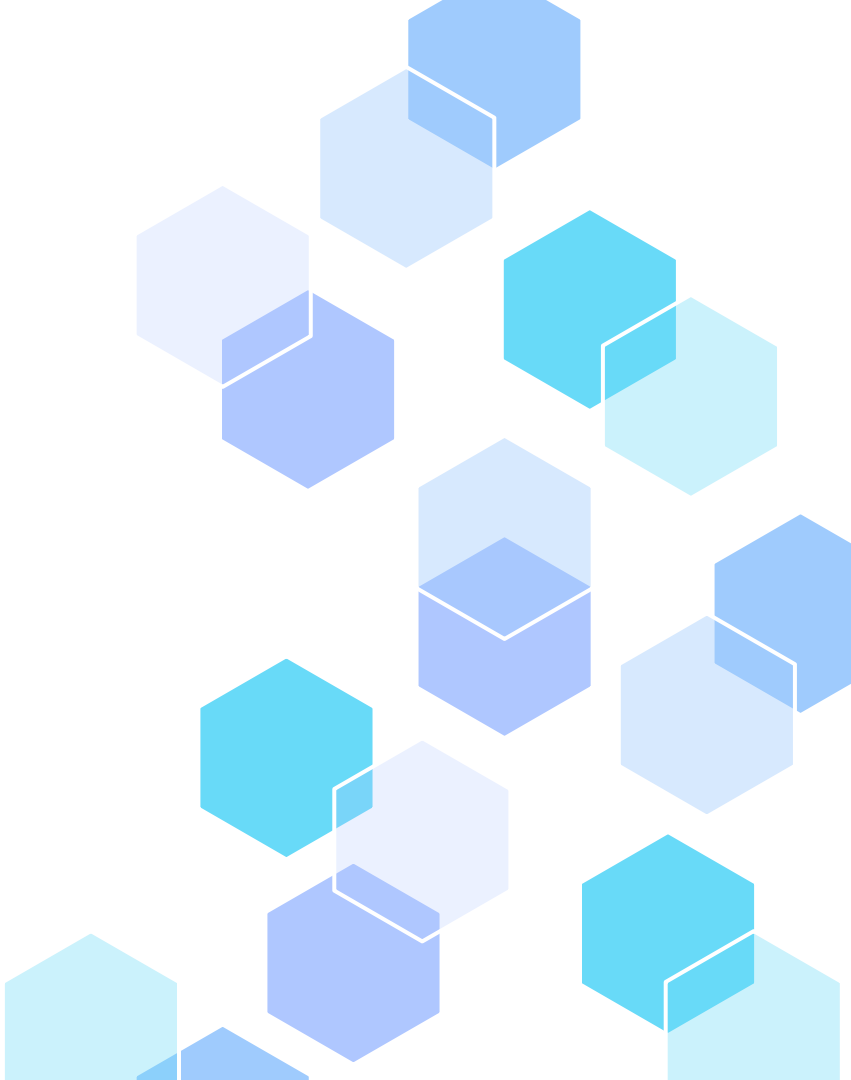
Biotech Companies

By advancing innovation in diagnostics, developing personalized medicine solutions, and expanding market opportunities for genetic testing technologies and targeted therapies.

02

The Data

Source and features



The Cancer Genome Atlas Pan-Cancer Atlas (2018)

The Data

2.5 petabytes of data

- Genomic
- Epigenomic
- Transcriptomic
- Proteomic

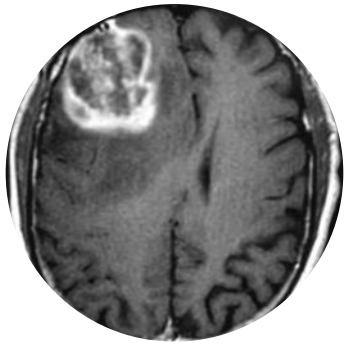
33 Cancer Types

Publicly Available

TCGA Mission

Increase scientific understanding of the molecular basis of cancer and apply this information and apply this information to improve our ability to diagnose, treat, and prevent cancer.

Target Variable



Glioblastoma Multiforme

Grade IV tumor. Grow rapidly and are the most aggressive type of glioma. App. 50% of all primary malignant brain tumors

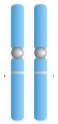
Low Grade Glioma

Grade I or II tumors. Grow slowly and not as aggressive.

Some Features

Chromosome

Which chromosome is affected, chromosome length



Allele

Reference Allele

Mutation Description

The specific type of mutation / related info

Gene Location

Where the mutated gene is located in the genome

Patient Info

Gender, smoking & drinking habits

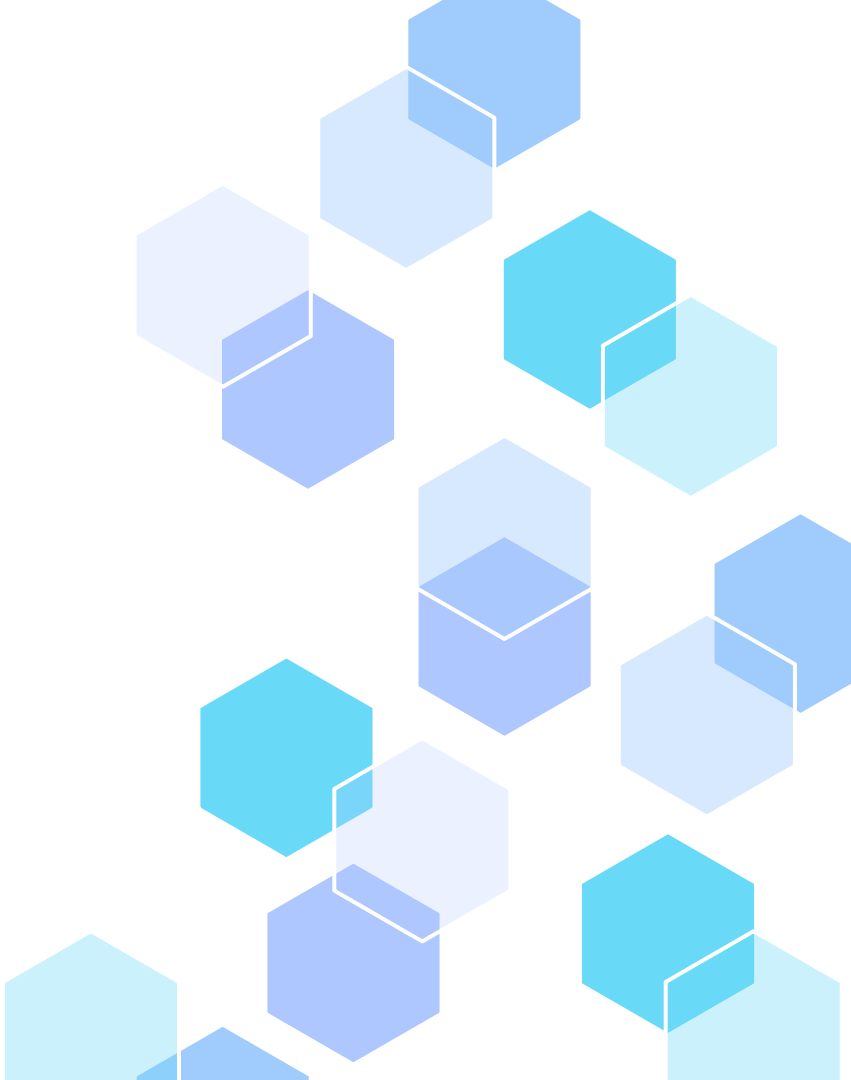
Variant Type

Describes what type of mutation

03

Modelling

Models used & results



Modelling Process

Round 1

- After research / domain knowledge, complete feature engineering
- Decide what models to use for the binary classification problem
- Complete initial series of modelling using gridsearch to obtain optimal hyperparameters
- Evaluate model performances

Round 2

- Address class imbalances via SMOTE, undersampling, and class weights
- Based on most successful resampling, do another round of grid searches.
- Based on most successful resampling, do another round of grid searches.
- Evaluate model performances

Round 1 GridSearch

Logistic Regression

```
cachedir = mkdtemp()

estimators = [('normalise', StandardScaler()),
              ('PCA_model', PCA(n_components=15)),
              ('model', LogisticRegression())]

pipe = Pipeline(estimators, memory = cachedir,
                verbose = True)

param_grid = [
             {'model': [LogisticRegression()],
              'normalise': [StandardScaler(), None],
              'model__penalty': ['l1', 'l2'],
              'model__solver': ['saga', 'liblinear'],
              'model__C': [0.001, 0.01, 0.1, 1, 10, 15]}]

grid = GridSearchCV(pipe, param_grid, cv=5, verbose=1)
fittedgrid = grid.fit(X_train, y_train)
```

Decision Tree Classifier

```
estimators = [('normalise', StandardScaler()),
              ('model', DecisionTreeClassifier())]

pipe = Pipeline(estimators, verbose = True)

param_grid = [
             {'model': [DecisionTreeClassifier()],
              'normalise': [StandardScaler(), None],
              'model__criterion': ['gini', 'entropy'],
              'model__max_depth': [10, 20, 30, 40],
              'model__min_samples_split': [2, 5, 10],
              'model__min_samples_leaf': [1, 2, 4]}
]

grid = GridSearchCV(pipe, param_grid, cv=5, verbose = 1)
fittedgrid = grid.fit(X_train, y_train)
```

K-Nearest Neighbors

```
estimators = [('normalise', StandardScaler()),
              ('model', KNeighborsClassifier())]

pipe = Pipeline(estimators, verbose=True)

param_grid = [
             { 'model__n_neighbors' : [5, 7, 9, 11, 13, 15],
               'model__weights' : ['uniform',
                                   'distance'],
               'model__metric' : ['minkowski', 'euclidean',
                                  'manhattan']}
]
grid = GridSearchCV(pipe, param_grid, verbose=1, cv=3,
                    n_jobs=-1)
fittedgrid_knn = grid.fit(X_train, y_train)
```

Round 1 Model Evaluation

Logistic Regression

Accuracy = 70.40%

Precision = 70.59%

Recall = 99.25%

F1 Score = 83%

	Predicted LGG	Predicted GBM
True LGG	147	6879
True GBM	123	16516

Decision Tree Classifier

Accuracy = 70.46%

Precision = 70.69%

Recall = 98.79%

F1 Score = 82.47%

	Predicted LGG	Predicted GBM
True LGG	236	6790
True GBM	200	16439

K-Nearest Neighbors

Accuracy = 70.50%

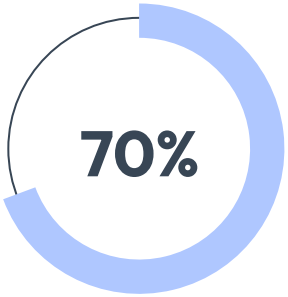
Precision = 70.76%

Recall = 98.93%

F1 Score = 82.50%

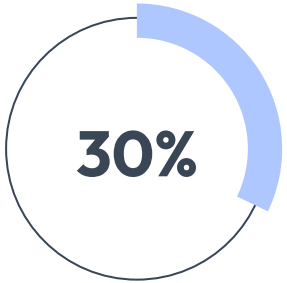
	Predicted LGG	Predicted GBM
True LGG	223	6803
True GBM	178	16461

Class Imbalance



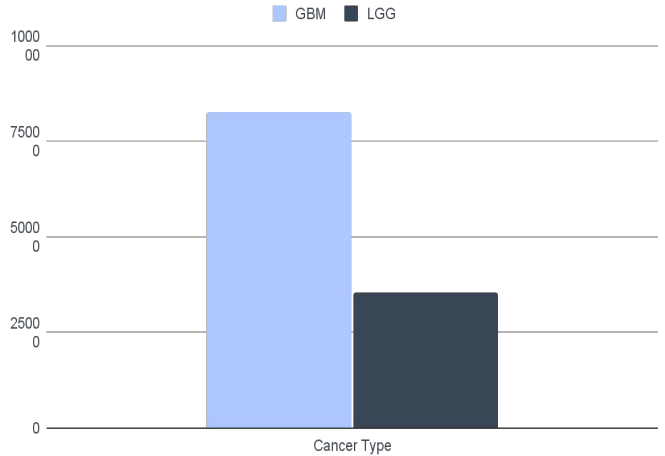
Glioblastoma Multiforme

82,765 instances of GBM



Low Grade Glioma

35,556 instances of LGG



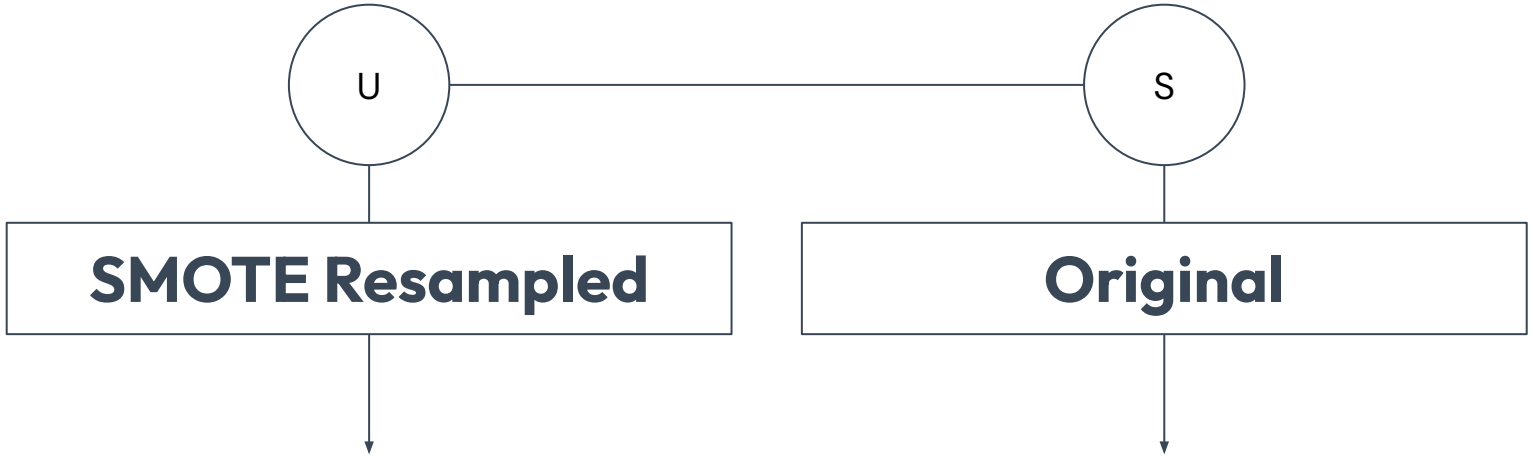
Addressing Class Imbalance

Oversampling (SMOTE)	
Accuracy Score	69.6%
Recall Score	98.6%
Precision Score	70.3%
F1 Score	82.1%
Incorrect Predictions	

Undersampling (RandomUnderSampler)	
Accuracy Score	59.5%
Recall Score	66.5%
Precision Score	73.2%
F1 Score	69.7%
Incorrect Predictions	

Unsampled	
Accuracy Score	70%
Recall Score	99.9%
Precision Score	70%
F1 Score	82%
Incorrect Predictions	7104

Unsampled vs Resampled Gridsearch Results



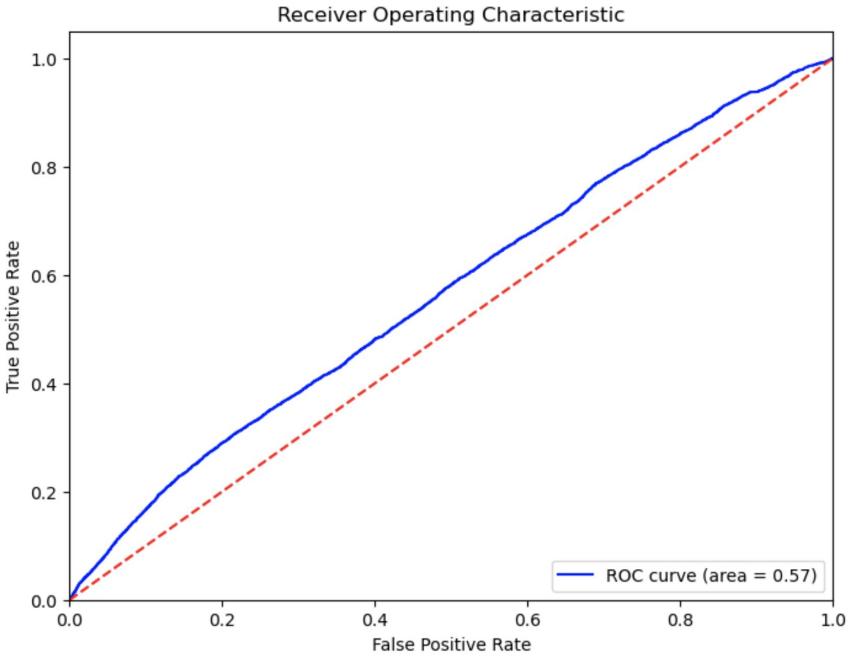
Hyperparameters

PCA n_components: 20
Model C: 15
Model Penalty: l1
Model solver: liblinear
Normalise: None

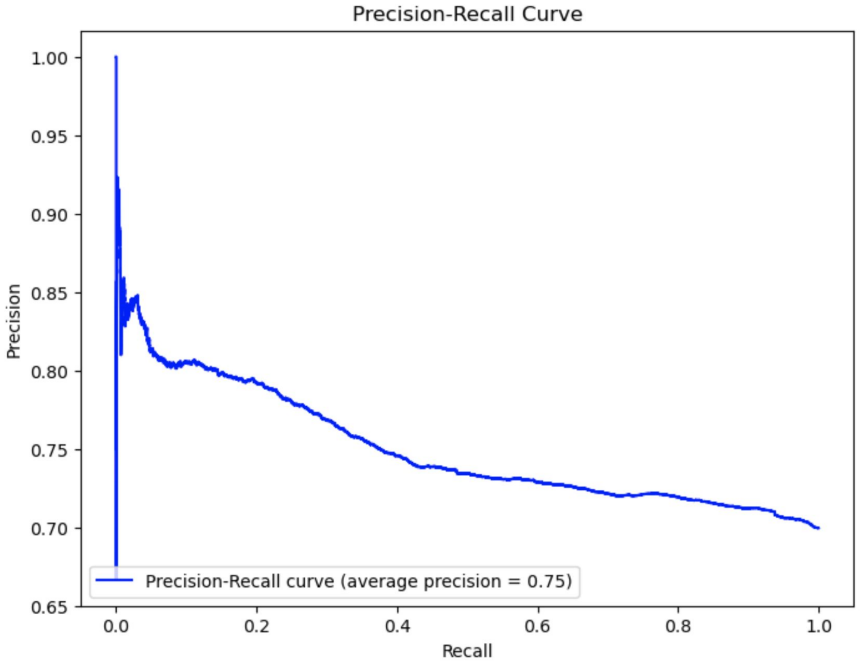
Hyperparameters

PCA n_components: 15
Model C: 1
Model Penalty: l2
Model Solver: Liblinear
Normalise: None

Logistic Regression Round 2 Model Performance



ROC AUC: 0.57



Avg. Precision: 75%

Logistic Regression Round 2 Model Performance

Evaluation Metrics		
Accuracy Score	69.70%	-0.7%
Recall Score	97.87%	-1.38%
Precision Score	70.38%	-0.21%
F1 Score	81.88%	-1.12%
Log Loss	61.29%	
Avg Precision	74.98%	

Top 10 Features for GBM

1. Variant_Classification_IGR (9.21)
2. Mut_Len_55-63 (8.85)
3. mutation_transversion (7.89)
4. mutation_other (7.29)
5. mutation_transition (7.08)
6. Variant_Classification_Silent (6.60)
7. Variant_Classification_In_Frame_Del (5.13)
8. Variant_Classification_In_Frame_Ins (5.08)
9. Variant_Classification_Intron (4.86)
10. Mut_Len_46-54 (4.74)

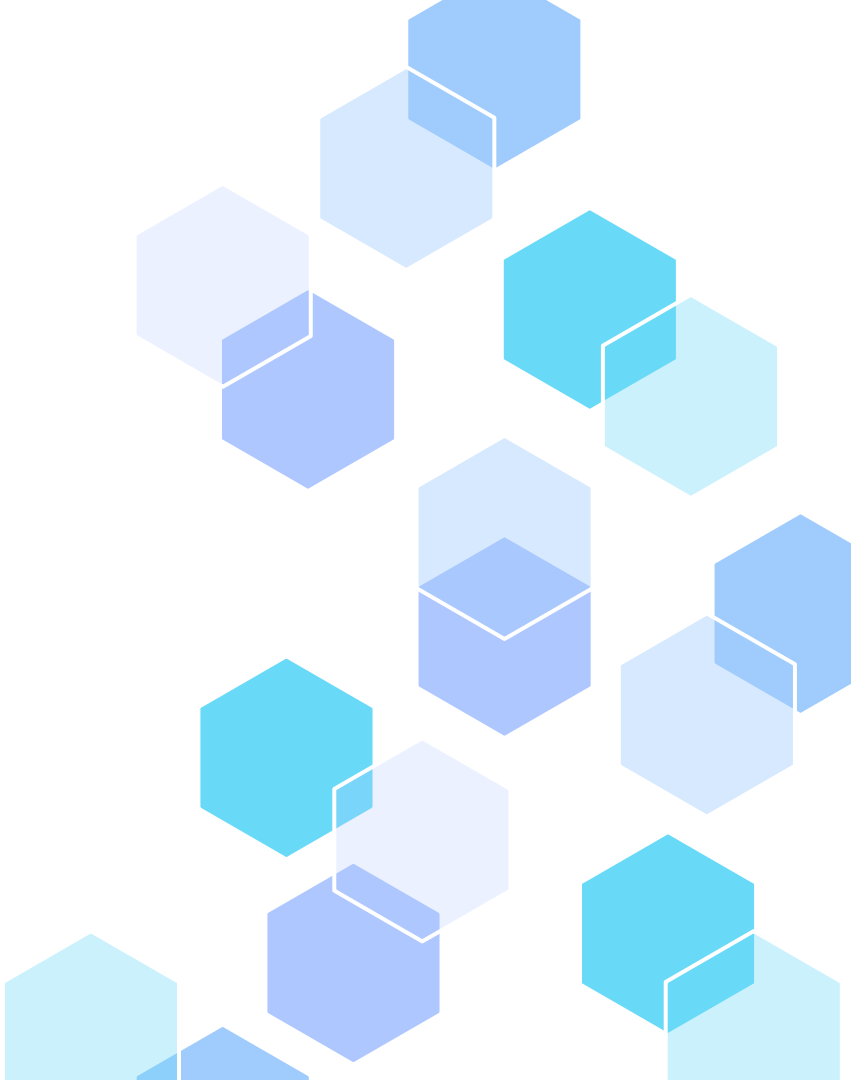
Top 10 Features for LGG

10. thick_length (0.12)
9. Unnamed: 0 (7.16)
8. blockSizes (-0.05)
7. mutation_length (-0.23)
6. Mut_Len_19-27 (-0.35)
5. Mut_Len_10-18 (-1.10)
4. Mut_Len_1-9 (-3.19)
3. autosome (-4.40)
2. blockCount (-4.46)
1. sex_chromosome (-4.76)

04

What’s Next?

Further steps



Next Steps

Although we have an adequate model, there are additional steps to take in the future.

- Create a Streamlit page
- Additional Modelling
  - XG Boost & RandomForest
- Further feature engineering (advanced)
- Additional cancer types

Thank you!