Logistic Regression Model for Predicting Brain Tumor Malignance

Statistical Analysis Report

May 2024

Breck Emert

Stat 835 – Categorical Analysis

Abstract

This statistical analysis report finds a logistic regression model aimed to predict the malignancy of brain tumors between either glioblastoma (GBM) or lower-grade glioma (LGG). Using data from 857 patients diagnosed with glioma, gathered from the UC Irvine Machine Learning Repository, this study examines the relationship between these glioma grades and the mutation status of 20 commonly mutated genes. The analysis involved preprocessing to remove irrelevant features, variable selection using Hosmer et al. (2013)'s purposeful selection method, and 10-fold cross validated stepwise selection based on Bayesian Information Criterion (BIC). This methodology was designed to significantly minimize model complexity while maintaining predictive accuracy. The final model incorporates four gene mutations (IDH1, IDH2, TP53, PIK3R1) and shows significant predictive power for tumor grading. The model validation shows an AUC of 0.90 which shows great discrimination between LGG and GBM based on genetic profiles. This study highlights the potential of logistic regression in medical diagnostics and suggests areas for further research which may allow genetic profiling to be used in aide of diagnosis of the degree of malignancy in gliomas.

Contents

Title

Logistic Regression Model for Predicting Brain Tumor Malignance

Introduction

Glial cells perform many supporting operations in the brain, including regulation of the chemical environment, acting as myelination to improve signal speed, and physical structure (Gazzaniga & Mangun, 2013). They make up half of the cells in the brain, roughly equal to the number of neurons (Gazzaniga & Mangun, 2013). Glioma, a term for a group of cancers that involve the glial cells of the brain, primarily occurs in the cerebrum and makes up 33 percent of all brain tumors (John Hopkins Medicine, 2022). The need for diagnostic prediction and classification of glioma is evident, as glioma is the most malignant and rapid-onset type of brain cancer (Schwartzbaum et al., 2006).

In this exploratory analysis we intend to uncover the relationship between specific genetic mutations and the mutation status of 20 commonly mutated genes. We hypothesize that the expression of one or more gene mutations can predict whether the cancer is graded as glioblastoma (GBM) or lower-grade glioma (LGG). The dataset spans 857 patients diagnosed with glioma and contains data on the cancer grade, gender, age, and the mutation status of 20 commonly mutated genes. The variable for the specific type of glioma is not included due to the scope of the research question.

We considered the following predictor variables:

1. **Grade**: The grade of glioma, either glioblastoma (GBM) or lower-grade glioma (LGG).
2. **Gender**: The gender of the patient, either male or female.
3. **Gene Expression (20 variables):** Binary variables indicating the expression of 20 commonly mutated genes, 1 for mutated and 0 for a regular gene.

Materials and Methods

Data Collection

The dataset used for analysis was downloaded from the UC Irvine Machine Learning Repository (Tasci et al., 2022). It contains data from 857 patients diagnosed with glioma and covers their diagnosed grade (the degree of malignancy), gender, age and 20 mutation features. The gene mutations are not specific or meaningful necessarily, they are just the 20 most commonly mutated genes according to The Cancer Genome Atlas, the provider of the dataset (Tasci et al., 2022). Each row of the data represents the data from one patient, recorded at the time of diagnosis.

Data Preprocessing

The dataset contained two features unrelated to our analysis: project and case ID. These variables contain administrative data irrelevant to the cohort or specific diagnosis, and were removed. In addition, we eliminated the primary diagnosis column as it is not related to the research question of this paper - it contained the information on the specific subtype of glial cells affected by the cancer. 4 observations were removed due to missingness.

Statistical Analysis

All calculations were performed with the statistical language R. Variables were chosen under the guidance of Hosmer et al. (2013)'s purposeful selection method. Model training was performed using the 'caret' package for 10-fold training and validation splits and 'MASS' for stepwise BIC selection. All code output is visible in attached PDFs and a Word Document.

Final Dataset

Variable Name	Data Type	Data Format	Description	Example
Grade	Binary	{LGG, GBM}	Glioma grading of Lower-Grade (LGG) or Glioblastoma Multiforme (GBM)	LGG
Gender	Binary	{Female, Male}	Binary sex of the individual	Female
Age At Diagnosis	Continuous	Numeric	Age when the patient received diagnosis	51.3
Gene Name 1	Binary	{0, 1}	Expression of Gene 1 in the patient	1
...	...	...	...	...
Gene Name 20	Binary	{0, 1}	Expression of Gene 20 in the patient	0

Results

Exploratory Analysis

The mean age of patients in the study is 50.9 with a standard deviation of 15.7. The descriptive statistics in Appendix A1 show the contingency tables for each binary variable which highlight their counts. We can see that 359 of the patients are female and 498 are male, with a higher prevalence of LGG among both genders. Appendix A2 shows the mutation heatmap of the dataset, ordered by age for both grades. From the figure we can see the patients diagnosed with GBM are much younger on average. Some variables have very low prevalence which may warrant scrutiny for their inclusion in the final model, highlighted in figure 1 below.

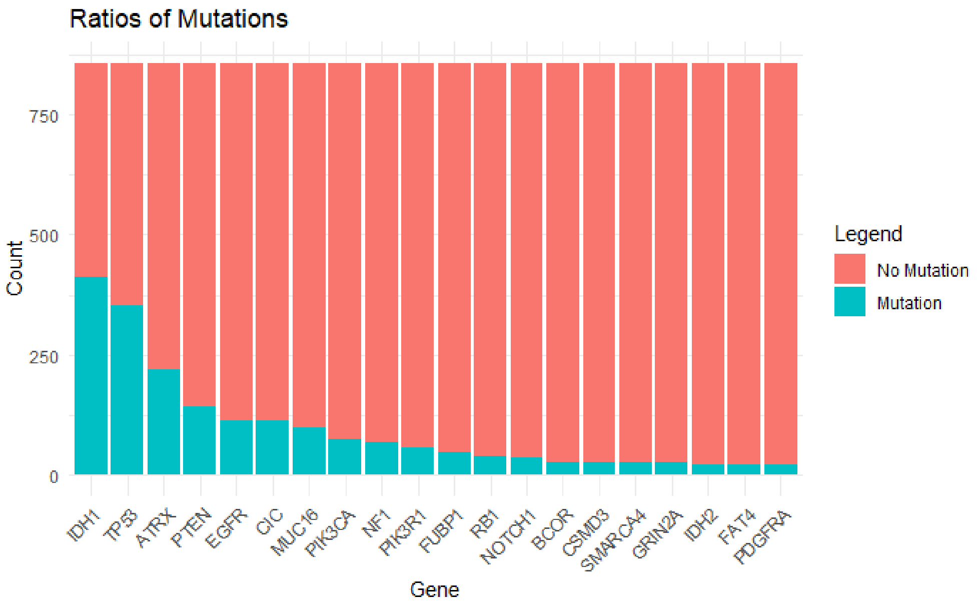


Figure 1: Ratios of Mutations

A Pearson correlation matrix in Figure 2 below visually highlights the linear relationships among every variable. This information is useful as logistic regression models suffer from multicollinearity, as it inflates the standard deviation of coefficients and lowers the significance of individual variables. We highlight that only one relationship shows a correlation coefficient greater than 0.5, a common threshold for high covariance, which is the genes ATRX and TP53 with a coefficient of 0.54. All of the exploratory information uncovered in this section will be considered in the variable selection phase.

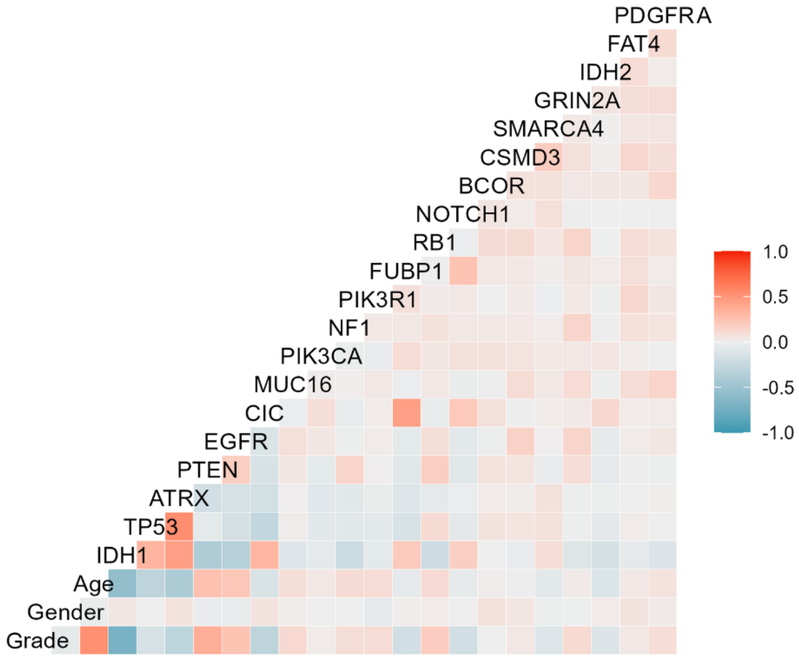


Figure 2: Correlation Matrix

Variable Selection

To select variables for the model in light of the relatively limited amount of data (relative to the 22 available variables), purposeful selection under the guidance of Hosmer et al. (2013) was employed. Using the contingency tables generated in Appendix A1.1, a Pearson chi-square test was used to calculate the respective score statistics. After selecting only variables where the associated p-value was less than 0.2, 5 variables were removed as potential explanatory variables (Age, PIK3CA, BCOR, CSMD3, FAT4). The full results of these tests can be found in Appendix A3.

Model Selection

With the strongest candidate variables selected, we performed backwards selection with 10-fold cross validation. To aide in selecting a parsimonious model, BIC was used with k = log(857) = 2.93. The model yielded by this method has four variables and a BIC of 560. However, all models included NOTCH1, but this variable introduces extreme instability in the coefficients across these models; in the stepwise selected model the standard error was over 32 times the estimate itself. Because of the uncertainty, stepwise selection was run without this variable. Removal of influential datapoints and artificial adjustments for the low-EPV did not resolve the uncertainty issue so the exact cause is unknown. The models found without NOTCH1 are all stable, yielding a lowest BIC of 607, with the next best 3 models having 631 BIC, a moderately distinct result.

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Grade ~ Gender + IDH1 + TP53 + ATRX + PTEN + EGFR + CIC + MUC16 +
NF1 + PIK3R1 + FUBP1 + RB1 + SMARCA4 + GRIN2A + IDH2 + PDGFRA

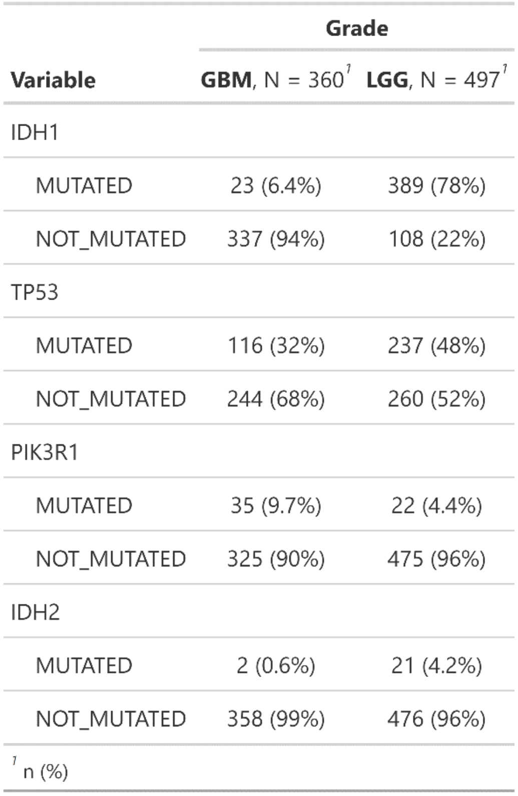
Final Model:
Grade ~ IDH1 + TP53 + PIK3R1 + IDH2

	Step	Df	Deviance	Resid. Df	Resid. Dev	BIC
1				840	565.5	680.4
2	-ATRX	1	0.08	841	565.6	673.7
3	-PDGFRA	1	0.73	842	566.4	667.7
4	-Gender	1	0.87	843	567.2	661.8
5	-RB1	1	0.96	844	568.2	656.0
6	-FUBP1	1	0.86	845	569.1	650.1
7	-SMARCA4	1	2.23	846	571.3	645.6
8	-MUC16	1	2.84	847	574.1	641.7
9	-CIC	1	2.81	848	576.9	637.7
10	-EGFR	1	3.45	849	580.4	634.4
11	-GRIN2A	1	4.17	850	584.6	631.8
12	-PTEN	1	6.25	851	590.8	631.3
13	-NF1	1	6.69	852	597.5	631.3

Research was done on potential gene interaction effects, and none were found among the candidate variables. Because of the potential 231 two-way interactions and near infinite combinations of models that can be made with them, there is no statistically reliable way to automatically search for these effects, and they will not be included in the model.

Model Validation

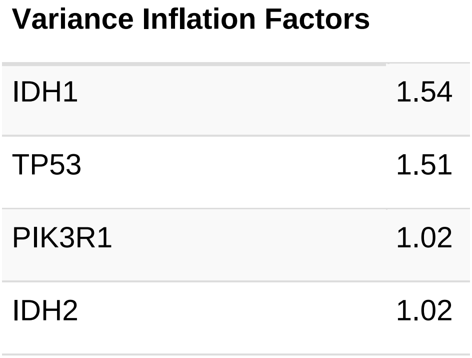
ML estimates are subject to drastic bias with a small number of events per variable (EPV), and backwards selection further compounds this issue (Van Smeden et al., 2018). As such, it is important to have a reasonable of EPV and low VIF among the predictors. For the candidate model, these numbers are shown in the figures below. The IDH2 gene has the lowest count of 23 mutations, and while this does not meet a stringent threshold of 10 variables per predicted included, this estimate is conservative. van Smeden further highlights that few studies actually support the requirement of 10 EPV, noting that most problems occur within 2-4 EPV. The EPV for IDH2 is 23/5 = 4.6 and the p-value for the coefficient of <0.0001, the variable was kept in the candidate model.



	Grade	
Variable	**GBM**, N = 360[1]	**LGG**, N = 497[1]
IDH1		
MUTATED	23 (6.4%)	389 (78%)
NOT_MUTATED	337 (94%)	108 (22%)
TP53		
MUTATED	116 (32%)	237 (48%)
NOT_MUTATED	244 (68%)	260 (52%)
PIK3R1		
MUTATED	35 (9.7%)	22 (4.4%)
NOT_MUTATED	325 (90%)	475 (96%)
IDH2		
MUTATED	2 (0.6%)	21 (4.2%)
NOT_MUTATED	358 (99%)	476 (96%)

[1] n (%)

Figure 3: Final Contingency Table



Variance Inflation Factors

IDH1	1.54
TP53	1.51
PIK3R1	1.02
IDH2	1.02

Table 2: Variance Inflation Factors

Next, Cook's distance was calculated for every observation in the dataset, which is a measure of the standardized residuals. For logistic regression, this is performed by measuring the overall change in the fitted logits after deleting the ith observation. The only large values were around 0.1, which is less than our cutoff of 0.5. For a dataset of this size and from the strongly varying effects of genes, a more stringent cutoff of 4/n cuts off many observations. The full plot is available in Appendix A4. In addition, removal of strong observations in logistic regression can cause a drastic decrease in sensitivity (and the opposite).

Finally, the Type III sum of squares was computed for each predictor variable in the candidate model. This assesses each predictor's contribution while controlling for the other predictors in the model. The results show strong significance for each predictor, which is interpreted as the model fit being significantly worse if it is not included.

Analysis of Deviance Table (Type III tests)				
	LR	Df	Pr(>Chisq)	
Age	24.01	1	9.57E-07	***
IDH1	276.71	1	2.20E-16	***
TP53	20.59	1	5.68E-06	***
NOTCH1	15.13	1	0.000100	***
IDH2	37.23	1	1.05E-09	***

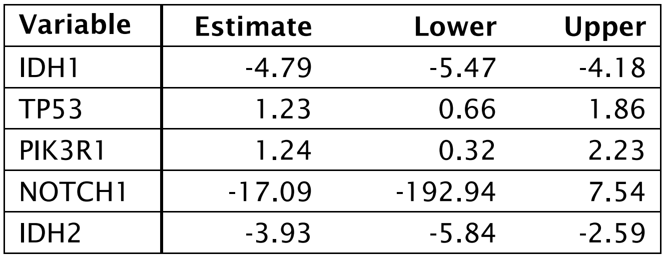
Final Model

The final model equation is described as:

logit[P(Grade=GBM)] = 1.07 - 4.71 * IDH1 + 1.07 * TP53 + 1.18 * PIK3R1 - 3.87 * IDH2

The estimates are visible in table 4 and figure 6 below:

Term	Estimate	Std Error	Z-Value	P-Value
(Intercept)	1.07	0.13	8.14	0.0000
IDH1	-4.71	0.31	-15.33	0.0000
TP53	1.07	0.28	3.78	0.0002
PIK3R1	1.18	0.47	2.54	0.0112
IDH2	-3.87	0.78	-4.96	0.0000



Variable	Estimate	Lower	Upper
IDH1	-4.79	-5.47	-4.18
TP53	1.23	0.66	1.86
PIK3R1	1.24	0.32	2.23
NOTCH1	-17.09	-192.94	7.54
IDH2	-3.93	-5.84	-2.59

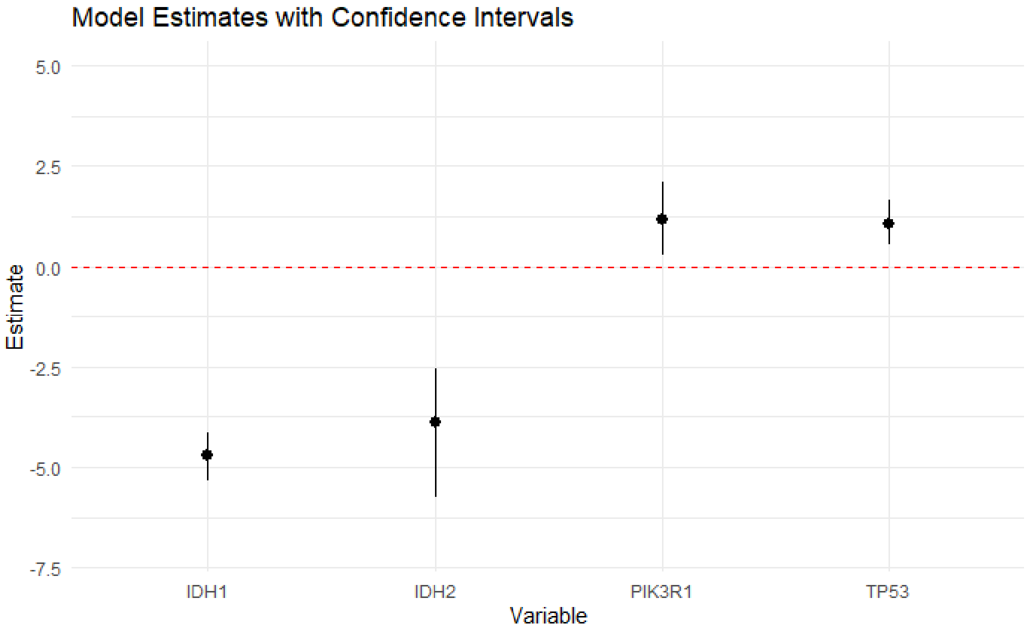
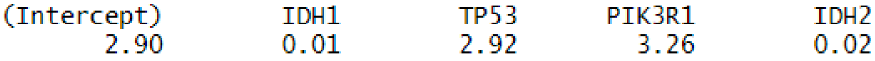


Figure 6: Model Coefficients

Discussion

Interpretation of results

This logistic regression model predicts glioma grading using a set of 4 binary gene mutation expression variables to estimate the probability that Y=1 (a diagnosed grade of GBM). The exponent of the coefficients tells us the multiplicative effect on the odds that a patient with glioma receives a diagnosis as GBM. Note that the exponentiated intercept term is interpreted as the odds when all genes are not mutated, which is not as useful. The values in the following table are calculated in this way. We see that the genes for IDH1 and IDH2 have little effect on the outcome when mutated, whereas for TP53 and PIK3R1 the odds of being diagnosed as GBM triple.



```
(Intercept)        IDH1        TP53      PIK3R1        IDH2
       2.90        0.01        2.92        3.26        0.02
```

Table 5: Exponentiated Coefficients

We now summarize the predictor power of the model as a whole. Receiver Operating Characteristic (ROC) curves summarize predictive power over all possible values of π_0, in which π_0 represents our decision boundary for deciding if a case is GBM or LGG. The area under the curve (AUC) of the final model's ROC curve is 0.90, visible in figure 7 below. An AUC of 0.90 means that our model has a 90% chance of correctly distinguishing between a randomly chosen positive and negative example; a value of .5 is random and 1 shows perfect discrimination. At the optimal threshold of this curve, the sensitivity and specificity are 0.82 and 0.93, respectively. These values mean that the model identifies 82% of the patients with GBM while correctly rejecting 93% of the patients without GBM. While the ROC curve shows a full visual of every value of π_0, it is still useful to visualize the classification table at the sample proportion of GBM which is 0.42. This classification table to help put the model's prediction to numbers is also visible below in table 6.

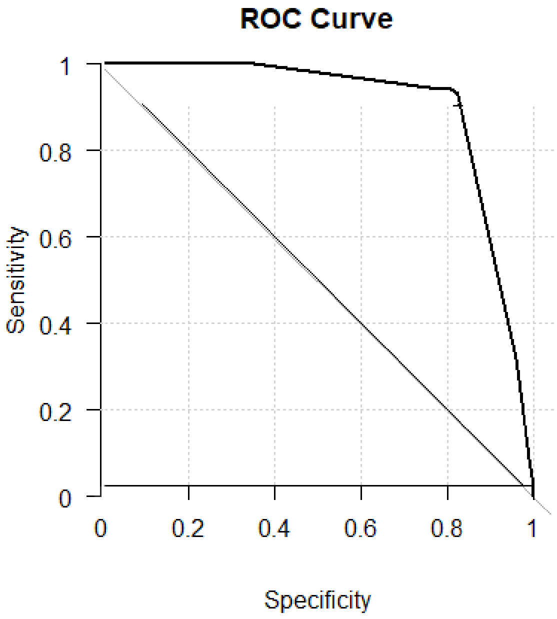


Figure 7: ROC Curve

Classification Table at $\pi_0 = 0.42$

		Predicted	
		LGG	**GBM**
Actual	**LGG**	410	87
	GBM	25	335

Table 6: Classification Table

Implications of findings

Knowing what genes are best predictive of malignant glioma is very useful. If a patient who has glioma shows specific combinations of genetic mutations, this model will help identify if they will be diagnosed as GBM. This could potentially improve the relevance of the treatment they receive. Most importantly, our model specifically finds that the presence of mutations in the TP53 or PIK3R1 genes is associated with 2.92 and 3.96 times the odds of their cancer being classified as GBM compared to those who do not. We again note that interaction effects were not considered for the study. These findings encourage further studies to explore potential causal mechanisms in these genes and to investigate how similar genetic profiles may be able to predict malignancy at earlier stages such as during diagnosis.

Limitations of the study

More data is needed to be able to consider a wider range of variables - the exclusion of the NOTCH1 variable from automatic selection highlights this limitation of the dataset. The logistic regression model inherently assumes a linear form for the logit of the probability of success, which was not tested compared to other potential relationships. The full details of the dataset used are not known, and the model assumption of independence of observations could be violated if there are clusters within the data (e.g. patients from the same location or hospital). More stringent cutoffs for EPV and Cook's distance can be used which would change the results of the model.

While the model is useful for genetically stratifying patients with a diagnosed grade of glioma, there are several reasons why the uncovered gene mutation patterns are not necessarily causal. The model looks at patients who were already diagnosed with a grade of glioma, so it does not necessarily mean that someone with these mutations is more likely to have cancer. Given the complex genetic and environmental interactions involved in glioma, the exclusion of existing knowledge in the field is limiting to a comprehensive understanding of what leads to high-grade glioma. Potential interaction effects and other domain knowledge limits the breadth that this model can accurately cover, which resulted in a very parsimonious model.

References

Cohen, A. L., Holmen, S. L., & Colman, H. (2013). IDH1 and IDH2 mutations in gliomas. Current neurology and neuroscience reports, 13(5), 345. https://doi.org/10.1007/s11910-013-0345-4

Gazzaniga, M. S., & Mangun, G. R. (Eds.). (2014). The cognitive neurosciences (5th ed.). Boston Review. https://doi.org/10.7551/mitpress/9504.001.0001

Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd Edition, John Wiley & Sons, Hoboken, NJ.

Johns Hopkins Medicine. (2022, February 11). Gliomas. Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/gliomas

National Cancer Institute - Center for Genomics. (2019, March 6). The Cancer Genome Atlas Program (TCGA). cancer.gov. https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history/ethics-policies

National Cancer Institute - Center for Genomics. (n.d.). The Cancer Genome Atlas Program (TCGA). cancer.gov. https://www.cancer.gov/ccg/research/genome-sequencing/tcga

National Cancer Institute - Genomic Data Commons. (n.d.). About the Data. gdc.cancer.gov. https://gdc.cancer.gov/about-data

Schwartzbaum, J. A., Fisher, J. L., Aldape, K. D., & Wrensch, M. (2006). Epidemiology and molecular pathology of glioma. Nature Clinical Practice Neurology, 2(9), 494–503. https://doi.org/10.1038/ncpneuro0289

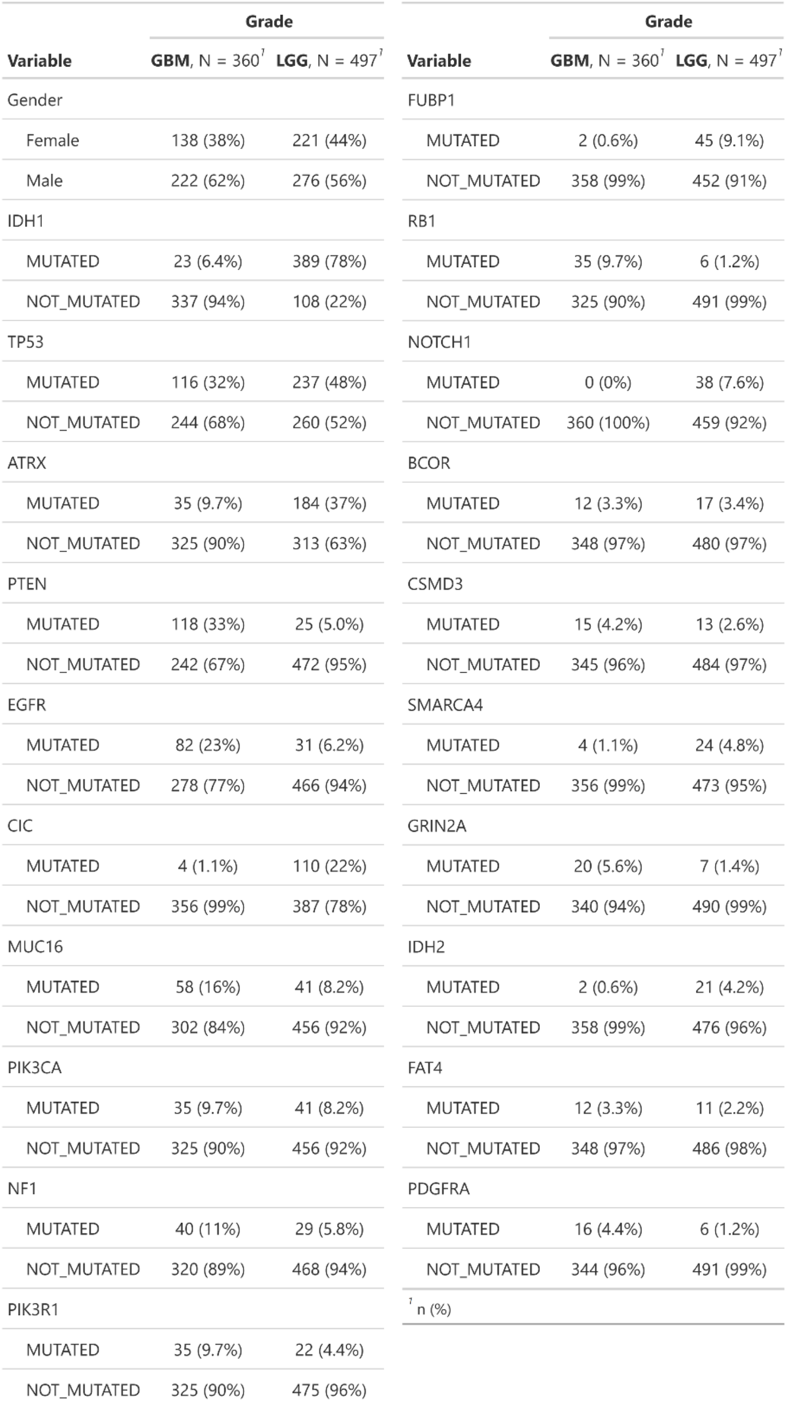
Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., & Krauze, A. V. (2022). Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. International Journal of Molecular Sciences, 23(22), 14155.

Tasci, E.; Zhuge, Y.; Kaur, H.; Camphausen, K.; Krauze, A.V. Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with

Clinical and Molecular Characteristics. Int. J. Mol. Sci. 2022, 23, 14155. https://doi.org/10.3390/ijms232214155

Van Smeden, M., Moons, K. G., De Groot, J. a. H., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research, 28(8), 2455–2474. https://doi.org/10.1177/0962280218784726

Appendix



	Grade	
Variable	**GBM**, N = 360[1]	**LGG**, N = 497[1]
Gender		
Female	138 (38%)	221 (44%)
Male	222 (62%)	276 (56%)
IDH1		
MUTATED	23 (6.4%)	389 (78%)
NOT_MUTATED	337 (94%)	108 (22%)
TP53		
MUTATED	116 (32%)	237 (48%)
NOT_MUTATED	244 (68%)	260 (52%)
ATRX		
MUTATED	35 (9.7%)	184 (37%)
NOT_MUTATED	325 (90%)	313 (63%)
PTEN		
MUTATED	118 (33%)	25 (5.0%)
NOT_MUTATED	242 (67%)	472 (95%)
EGFR		
MUTATED	82 (23%)	31 (6.2%)
NOT_MUTATED	278 (77%)	466 (94%)
CIC		
MUTATED	4 (1.1%)	110 (22%)
NOT_MUTATED	356 (99%)	387 (78%)
MUC16		
MUTATED	58 (16%)	41 (8.2%)
NOT_MUTATED	302 (84%)	456 (92%)
PIK3CA		
MUTATED	35 (9.7%)	41 (8.2%)
NOT_MUTATED	325 (90%)	456 (92%)
NF1		
MUTATED	40 (11%)	29 (5.8%)
NOT_MUTATED	320 (89%)	468 (94%)
PIK3R1		
MUTATED	35 (9.7%)	22 (4.4%)
NOT_MUTATED	325 (90%)	475 (96%)

	Grade	
Variable	**GBM**, N = 360[1]	**LGG**, N = 497[1]
FUBP1		
MUTATED	2 (0.6%)	45 (9.1%)
NOT_MUTATED	358 (99%)	452 (91%)
RB1		
MUTATED	35 (9.7%)	6 (1.2%)
NOT_MUTATED	325 (90%)	491 (99%)
NOTCH1		
MUTATED	0 (0%)	38 (7.6%)
NOT_MUTATED	360 (100%)	459 (92%)
BCOR		
MUTATED	12 (3.3%)	17 (3.4%)
NOT_MUTATED	348 (97%)	480 (97%)
CSMD3		
MUTATED	15 (4.2%)	13 (2.6%)
NOT_MUTATED	345 (96%)	484 (97%)
SMARCA4		
MUTATED	4 (1.1%)	24 (4.8%)
NOT_MUTATED	356 (99%)	473 (95%)
GRIN2A		
MUTATED	20 (5.6%)	7 (1.4%)
NOT_MUTATED	340 (94%)	490 (99%)
IDH2		
MUTATED	2 (0.6%)	21 (4.2%)
NOT_MUTATED	358 (99%)	476 (96%)
FAT4		
MUTATED	12 (3.3%)	11 (2.2%)
NOT_MUTATED	348 (97%)	486 (98%)
PDGFRA		
MUTATED	16 (4.4%)	6 (1.2%)
NOT_MUTATED	344 (96%)	491 (99%)

[1] n (%)

Figure A1: Contingency Tables

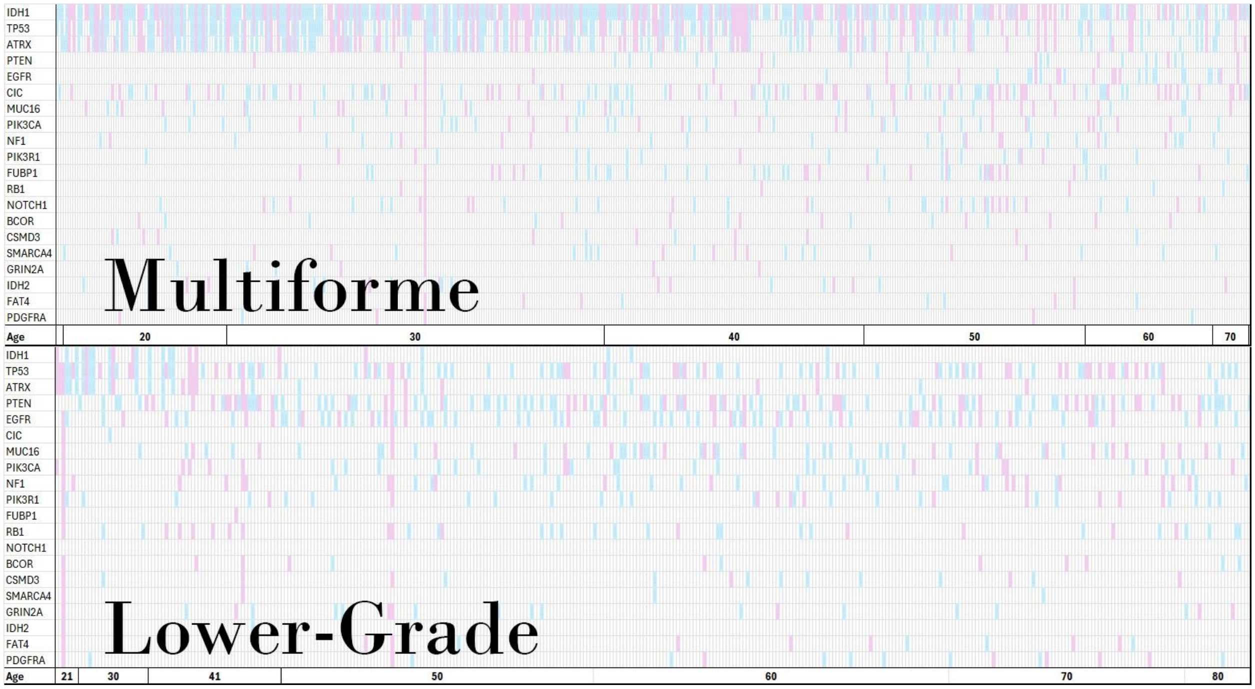
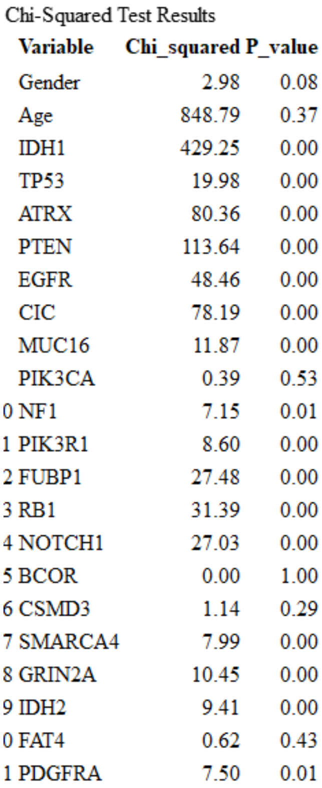


Figure A2: Gene Mutation Heatmap



Chi-Squared Test Results

	Variable	Chi_squared	P_value
	Gender	2.98	0.08
	Age	848.79	0.37
	IDH1	429.25	0.00
	TP53	19.98	0.00
	ATRX	80.36	0.00
	PTEN	113.64	0.00
	EGFR	48.46	0.00
	CIC	78.19	0.00
	MUC16	11.87	0.00
	PIK3CA	0.39	0.53
0	NF1	7.15	0.01
1	PIK3R1	8.60	0.00
2	FUBP1	27.48	0.00
3	RB1	31.39	0.00
4	NOTCH1	27.03	0.00
5	BCOR	0.00	1.00
6	CSMD3	1.14	0.29
7	SMARCA4	7.99	0.00
8	GRIN2A	10.45	0.00
9	IDH2	9.41	0.00
0	FAT4	0.62	0.43
1	PDGFRA	7.50	0.01

Figure A3: Purposeful Selection Chisq Tests

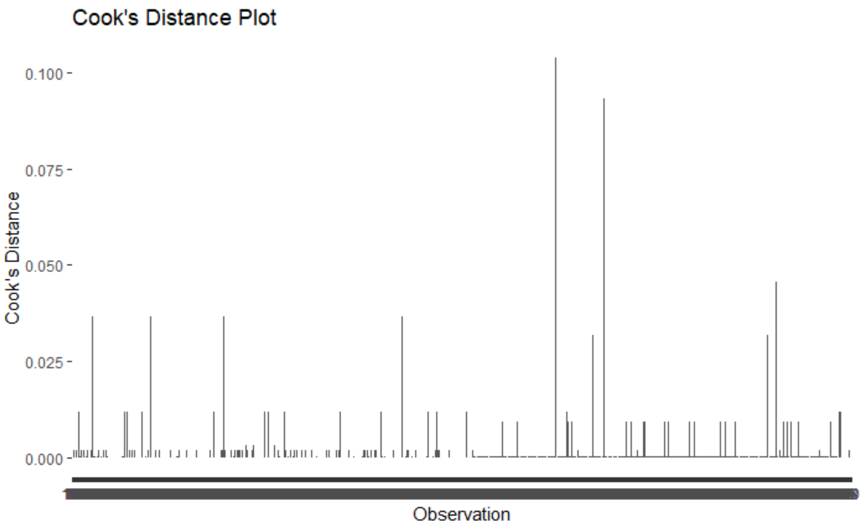


Figure A4: Cook's Distances

Appendix: R Code

These are rendered via three knit attachments as pdf. There are three unique files, clearly labeled as ‘Overview.Rmd’, ‘Model Experimenting.Rmd’, and ‘Final Model.Rmd’ containing their respective code. The copyable code is pasted in the following pages.

~~~
---
title: "Overview"
author: "Breck Emert"
date: '`r Sys.Date()`'
output: pdf_document
---


```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)


library(broom)
library(car)
library(dplyr)
library(GGally)
library(ggpubr)
library(gtsummary)
library(mvnormalTest)
library(reshape2)
library(webshot)
~~~

```{r load}
glioma <- read.csv("../Data/glioma.csv")


glioma_num <- read.csv("../Data/glioma_num.csv")
glioma_num$Age <- as.double(glioma_num$Age)
```

~~~
glioma_factor <- as.data.frame(lapply(glioma, factor))
glioma_factor$Age <- glioma$Age

head(glioma_num)
```

```{r ggpairs}
png <- ggcorr(glioma_num)
png
#cor(glioma_num$ATRX, glioma_num$TP53)
#ggsave("correlation_plot.png", plot=png)

ggpairs(
  glioma_num[, c("Grade", "Gender", "Age")],
  aes(alpha = 0.5),
  lower = list(
    continuous = wrap(
      "points",
      shape = 16,
      position = position_jitter(width = 0.1, height = 0.1)
    )
  ),
  progress = FALSE
)
```
~~~

~~~
```{r summary table}
summary_table <-
  subset(glioma, select = -Age) %>%
  tbl_summary(by=Grade) %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Grade**")

summary_table

#gt::gtsave(as_gt(summary_table), zoom=5, "summary table.png")
~~~

```{r descriptive statistics}
# Age
age_stats <- glioma_num %>%
  summarise(
    Mean = mean(Age, na.rm = TRUE),
    Median = median(Age, na.rm = TRUE),
    SD = sd(Age, na.rm = TRUE),
    Min = min(Age, na.rm = TRUE),
    Max = max(Age, na.rm = TRUE)
  )
print(age_stats)

# Counts
```

~~~~
binary_counts <- glioma_num[, 4:23] %>%
  summarise(across(everything(), list(zeroes = ~sum(. == 0), ones = ~sum(. == 1))))
print(binary_counts)
```


```{r mutation ratios graph}
# Exclude genes
genes_data <- glioma_num[, -(which(names(glioma_num) %in% c("Grade", "Gender", "Age")))]

# Melt and factor
data_melted <- melt(genes_data, id.vars = NULL)
data_melted$value <- factor(data_melted$value, levels = c(0, 1), labels = c("No Mutation", "Mutation"))

# Plot
plot <- ggplot(data_melted, aes(x = variable, fill = value)) +
  geom_bar(position = "stack") +
  labs(x = "Gene", y = "Count", title = "Ratios of Mutations", fill = "Legend") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readability
print(plot)
```


```{r Cancer Grades hist}
ggplot(glioma, aes(x = Grade)) +
~~~~

~~~
  geom_bar() +
  ggtitle("Distribution of Cancer Grades") +
  xlab("Grade") +
  ylab("Count")
```


```{r individual chisq tests}
library(knitr)


# Chi-sq test for each variable
variables <- setdiff(names(glioma_num), "Grade")


chi_sq_results_df <- data.frame(Variable = character(),
                                Chi_squared = numeric(),
                                P_value = numeric(),
                                stringsAsFactors = FALSE)


for (var in variables) {
  # Contingency table between Grade and the current variable
  table <- xtabs(~ get(var) + Grade, data = glioma_num)


  # Append chi-sq test
  test_result <- chisq.test(table)
  chi_sq_results_df <- rbind(chi_sq_results_df, data.frame(Variable = var,
                                                           Chi_squared = round(test_result$statistic, 2),
                                                           P_value = round(test_result$p.value, 2)))
~~~

~~~
}


# Print
chisq_kable <- kable(chi_sq_results_df, format = "html", caption = "Chi-Squared Test Results", table.attr = "style='background-color: white;'")


# Save to PNG
# html_file <- tempfile(fileext = ".html")
# writeLines(as.character(chisq_kable), html_file)
# png_file <- tempfile(fileext = ".png")
# webshot(html_file, png_file, delay = 0.2)
# file.rename(png_file, "chisq_kable.png")
```
~~~

~~~
---
title: "Model Experimenting"
author: "Breck Emert"
date: '`r Sys.Date()`'
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)

library(caret)
library(MASS)
library(stringr)
library(lmtest)
```

```{r load}
glioma <- read.csv("../Data/glioma.csv")

glioma_num <- read.csv("../Data/glioma_num.csv")
glioma_num$Age <- as.double(glioma_num$Age)

glioma_factor <- as.data.frame(lapply(glioma, factor))
glioma_factor$Age <- glioma$Age

glioma_rem <- glioma_num[-c(579, 626), ]
~~~

~~~
head(glioma_num)
```


```{r Purposeful Selection similar to the one in Overview RMD}
data <- glioma_num
predictor <- 'Grade'
variables <- setdiff(names(data), predictor)

# Chi-squared test for each variable
glioma_purposeful <- data.frame(Grade = data[[predictor]])
for (var in variables) {
  # Contingency table between Grade and the current variable
  table <- xtabs(~ get(var) + get(predictor), data = data)

  # Add if chi-sq p > 0.2
  test_result <- chisq.test(table)
  if (test_result$p.value < 0.2) {
    glioma_purposeful[[var]] <- data[[var]]
  } else {
    print(var)
  }
}

# Print
glioma_purposeful <- as.data.frame(lapply(glioma_purposeful, factor))
~~~

~~~~
glioma_purposeful$Grade <- factor(glioma_purposeful$Grade, levels = c(0, 1), labels = c("LGG", "GBM"))

glioma_purposeful <- subset(glioma_purposeful, select = -NOTCH1)

glioma_purposeful_rem <- glioma_purposeful[-c(579, 626), ]

head(glioma_purposeful)

```



```{r stepwise selection}

#glioma_purposeful <- subset(glioma_purposeful, select=-fitted_values)


initial_model <- glm(Grade ~ ., data=glioma_purposeful, family = binomial)

step_model <- stepAIC(initial_model, direction = "backward", k=log(857))


step_model$anova

```
~~~~

~~~
---
title: "Final Model"
author: "Breck Emert"
date: '`r Sys.Date()`'
output: pdf_document
always_allow_html: true
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)

library(broom)
library(car)
library(dplyr)
library(ggplot2)
library(gtsummary)
library(kableExtra)
library(knitr)
library(pROC)
library(webshot)
```

```{r loading}
glioma <- read.csv("../Data/glioma.csv")

glioma_num <- read.csv("../Data/glioma_num.csv")
~~~

~~~
glioma_num$Age <- as.double(glioma_num$Age)


glioma_factor <- as.data.frame(lapply(glioma, factor))
glioma_factor$Age <- glioma$Age


head(glioma_num)
```



```{r final model}
final_model <- glm(Grade ~ IDH1 + TP53 + PIK3R1 + IDH2, data=glioma_num, family=binomial)
summary(final_model)
```



```{r summary table}
summary_table <-
  subset(glioma, select = c(Grade, IDH1, TP53, PIK3R1, IDH2)) %>%
  tbl_summary(by=Grade) %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Grade**")


summary_table
#gt::gtsave(as_gt(summary_table), zoom=5, "summary table final.png")
```



```{r Influence Measures}
# Influence Measures
~~~

```
influence_measures <- influence.measures(final_model)
influence_df <- as.data.frame(influence_measures$infmat)
influence_df$observation <- row.names(influence_df)

high_influence <- influence_df %>%
  filter(hat > (2 * mean(hat)), cook.d > 0.05)
high_influence

# Plot
ggplot(influence_df, aes(x = observation, y = cook.d)) +
  geom_bar(stat = "identity") +
  labs(title = "Cook's Distance Plot",
       x = "Observation",
       y = "Cook's Distance")


# VIF
vif_results <- vif(final_model)
vif_table <- data.frame(Gene = names(vif_results), VIF = round(vif_results, 2), row.names = NULL)

vif_styled <- kable(vif_table,
                    format = "html",
                    col.names = c("Gene", "VIF"),
                    caption = "Variance Inflation Factors") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
```

~~~~
                full_width = F,
                font_size = 12) %>%
  column_spec(1, bold = TRUE, width = "10em") %>%
  column_spec(2, width = "5em")

vif_styled

```


```{r removal of outliers}
final_model <- glm(Grade ~ IDH1 + TP53 + PIK3R1 + IDH2, data=glioma_num, family=binomial)
summary(final_model)

glioma_rem <- glioma_num[-c(579, 626), ]
removed_model <- glm(Grade ~ IDH1 + TP53 + PIK3R1 + IDH2, data=glioma_rem, family=binomial)
summary(removed_model)
```


```{r Type 3 SS}
anova_results <- Anova(final_model, type="III")
print(anova_results)
```


```{r equation}
kable(tidy(final_model))
~~~~

~~~
paste(round(coef(final_model), 2), names(coef(final_model)), sep = ' * ', collapse = ' + ')
```


```{r visualization}
# Tidy model
d0 <- tidy(final_model, conf.int = TRUE) %>%
  select(term, estimate, conf.low, conf.high) %>%
  filter(term != "(Intercept)")

# Plot
ggplot(data = d0, mapping = aes(x = term, y = estimate, ymin = conf.low, ymax = conf.high)) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  geom_pointrange() +
  theme_minimal() +
  labs(x = "Variable", y = "Estimate", title = "Model Estimates with Confidence Intervals") +
  theme(legend.position = "none") +
  coord_flip() +
  coord_cartesian(ylim = c(-7, 5))

```


```{r interpretation}
# Cross Tab
print(round(exp(coef(final_model)), 2))
mean(glioma_num$Grade)
~~~

```
predicted <- as.numeric(fitted(final_model) > mean(glioma_num$Grade))
xtabs(~ glioma_num$Grade + predicted)


# ROC
response <- factor(glioma_num$Grade)
predictor <- predict(final_model, type = "response")


rocobj <- roc(response, predictor)


plot(rocobj,
    main = "ROC Curve",
    xlim = c(0, 1), ylim = c(0, 1),  # Set exact limits to 1 for both axes
    xlab = "Specificity", ylab = "Sensitivity",
    axes = FALSE
)


axis(1, at = seq(0, 1, by = 0.2), labels = as.character(seq(0, 1, by = 0.2)), las = 1)
axis(2, at = seq(0, 1, by = 0.2), labels = as.character(seq(0, 1, by = 0.2)), las = 1)


abline(h = seq(0, 1, by = 0.2), v = seq(0, 1, by = 0.2), col = "lightgray", lty = "dotted")


ciobj <- ci.se(rocobj,                    # CI of sensitivity
            specificities = seq(0, 1, 10)) # over a select set of specificities
plot(ciobj, type = "shape", col = "pink")     # plot as a blue shape
plot(ci(rocobj, of = "thresholds", thresholds = "best")) # add one threshold
```

~~~
# AUC
auc_value <- auc(rocobj)
print(paste("The AUC of the ROC curve is:", auc_value))


# Optimal Threshold
coords <- coords(rocobj, "best", ret=c("threshold", "specificity", "sensitivity"))
print(paste("Best threshold:", coords$threshold))
print(paste("Specificity at best threshold:", coords$specificity))
print(paste("Sensitivity at best threshold:", coords$sensitivity))
```
~~~