Example R Markdown

0.1 Data

In this report, we use the `breastcancer` dataset from the `RforHDSdata` package, which comes from the Original Wisconsin Diagnostic Breast Cancer Database and was obtained from the UC Irvine Machine Learning Repository. This data contains information on various features of cell nuclei present in images of 212 malignant and 357 benign breast masses.

0.2 Exploratory Analysis and Model Fitting

We first created a subset of the `breastcancer` dataset that contains only the 10 mean measures in addition to the diagnosis variable. Then, we fit a logistic regression with diagnosis as the outcome (we set the diagnosis value equal to 1 if the tumor is malignant and 0 otherwise) and include the main effects of all 10 mean measures. In this model, only texture, area, smoothness, and concave points were significantly associated with the diagnosis, so we fit a new model containing just these independent variables and then conducted a forward stepwise selection procedure based on AIC to find possible interactions that could improve the model's fit. Our final model included an interaction between texture and all three other measures as well as an interaction between smoothness and concave points, but we removed this last interaction due to a resulting infinite odds ratio. The coefficients from this model can be seen in Table 1 below, and the resulting ROC curve is shown in Figure 1.

Table 1: Final Model Intercepts and Odds Ratios

Term	Estimate	Std. Error	P Value
(Intercept)	6.869068e+11	17.6414961	0.1223551
texture	9.486210e-02	0.9416523	0.0123747
area	9.765352e-01	0.0119297	0.0465505
smoothness	0.000000e+00	161.5610625	0.0738243
concave_points	2.877549e+64	93.1948212	0.1112492
texture:area	1.001836e+00	0.0006544	0.0050702
texture:smoothness	1.131759e+08	8.4031024	0.0273240
texture:concave_points	2.680370e-02	4.8668672	0.4570920

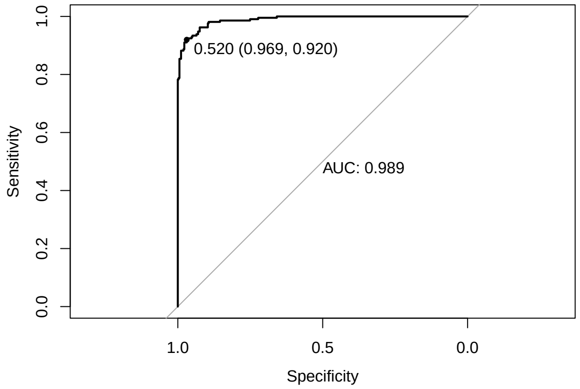

Figure 1: ROC Curve for Final Model

We can further examine the relationship between the four predictors in our model and our outcome of interest by drawing our attention to the plots in Figure 2 below.

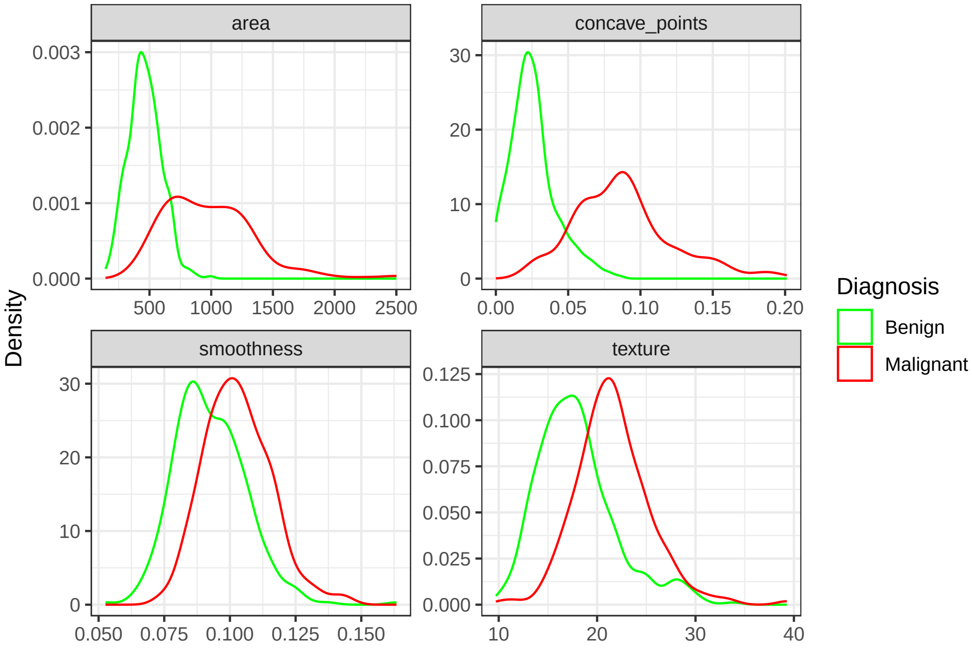

Figure 2: Predictor Variable Distributions by Diagnosis

1 Code Appendix

```r
knitr::opts_chunk$set(message=FALSE, warning=FALSE, error=FALSE, echo = FALSE, fig.pos = "H", out.extra
library(tidyverse)
library(knitr)
library(kableExtra)
library(broom)
library(HDSinRdata)
library(pROC)
#load in data
data(breastcancer)
#subset the data and simplify the column names
bc_sub <- breastcancer[2:12]
names(bc_sub) <- c("diagnosis", "radius", "texture", "perimeter", "area",
                   "smoothness", "compactness", "concavity", "concave_points",
                   "symmetry", "fractal_dimension")

#change the levels of the diagnosis variable
bc_sub <- bc_sub %>% mutate(diagnosis = case_when(diagnosis == "M" ~ 1,
                              diagnosis == "B" ~ 0))
original_mod <- glm(diagnosis ~ ., data = bc_sub, family = binomial)

#include only the significant main effects from the previous model
simple_mod <- glm(diagnosis ~ texture + area + smoothness + concave_points,
                  data = bc_sub, family = binomial)

#specify our upper bound for the model scope
fullmod <- glm(diagnosis ~ (texture + area + smoothness + concave_points)^2,
               data = bc_sub, family = binomial)

#perform forward stepwise selection
step_mod <- step(simple_mod, scope = list(upper = fullmod),
                 direction = "forward", trace = 0)

#view the summary
final_mod <- glm(diagnosis ~ texture + area + smoothness + concave_points +
                   texture:area + texture:smoothness + texture:concave_points,
                 data = bc_sub, family = binomial)

tidy(final_mod, exponentiate=TRUE) %>%
  select(-statistic) %>%
  kable(caption = "Final Model Intercepts and Odds Ratios",
        col.names = c("Term", "Estimate", "Std. Error", "P Value")) %>%
  kable_styling(latex_options = c("HOLD_position"),
                font_size=8)
roccurve <- roc(predictor=predict(final_mod, type="response"),
                    response=as.factor(final_mod$y),
                    levels = c(0,1), direction = "<")
plot(roccurve, print.auc=TRUE, print.thres = TRUE)
bc_long <- bc_sub %>% select(diagnosis, texture, area, smoothness, concave_points) %>%
  pivot_longer(cols = texture:concave_points) %>%
  mutate(diagnosis = case_when(diagnosis == 1 ~ "Malignant",
                               diagnosis == 0 ~ "Benign"))
```

```r
ggplot(data = bc_long) +
  geom_density(aes(x = value, y = ..density.., color = diagnosis)) +
  facet_wrap(~ name, scales = "free") +
  labs(y = "Density", x = "") +
  scale_color_manual(name = "Diagnosis", values = c("green", "red")) +
  theme_bw()
```