XGBoost

Code

- Installing Packages
- Loading Data
- XGBoost
- Model Evaluation
- Model Results
  - Importance Matrix
  - 实际值与预测值比较
  - Residual Analysis

Installing Packages

Hide

```
rm(list=ls(all=TRUE))
setwd('~/GitHub/IBD-EDA/aes/')
```

Loading Data

Hide

```
data <- read.csv('./data_processed/data.csv') %>%
  select(-X)
```

Hide

```
set.seed(123)
train_ratio = 0.8

train_indices <- sample(1:nrow(data), size = floor(train_ratio * nrow(data)))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]
```

Hide

```
dtrain <- xgb.DMatrix(
  data = as.matrix(train_data[,-1]), label = train_data[,1]
)

dtest <- xgb.DMatrix(
  data = as.matrix(test_data[,-1]), label = test_data[,1]
)
```

XGBoost

Hide

```
nrounds <- 50
params <- list(
  objective = "reg:squarederror",
  max_depth = 3,
  min_child_weight = 2,
  eta = 0.05,
  gamma = 0.1,
  subsample = 0.7,
  colsample_bytree = 0.8
)

final_model <- xgboost(
  data = dtrain,
  params = params,
  nrounds = nrounds,
  print_every_n = 10,
  early_stopping_rounds = 10,
  eval_metric = "rmse",
  evals = list(validation = dtest)
)
```

```
[16:24:36] WARNING: src/learner.cc:767:
Parameters: { "evals" } are not used.

[1] train-rmse:5.056297
Will train until train_rmse hasn't improved in 10 rounds.

[11]    train-rmse:4.277666
[21]    train-rmse:3.847145
[31]    train-rmse:3.542795
[41]    train-rmse:3.376484
[50]    train-rmse:3.243871
```

Model Evaluation

Hide

```
actuals_test <- test_data[,1]
preds_test <- predict(final_model, newdata = as.matrix(test_data[,-1]))

results <- postResample(pred = preds_test, obs = actuals_test)

print(paste("RMSE on Test Set: ", results[1]))
```

```
[1] "RMSE on Test Set:  4.40649594630852"
```

Hide

```
print(paste("MAE on Test Set: ", results[2]))
```

```
[1] "MAE on Test Set:  0.00422809586513368"
```

Hide

```
print(paste("R2 on Test Set: ", results[3]))
```

```
[1] "R2 on Test Set:  2.31902301856005"
```

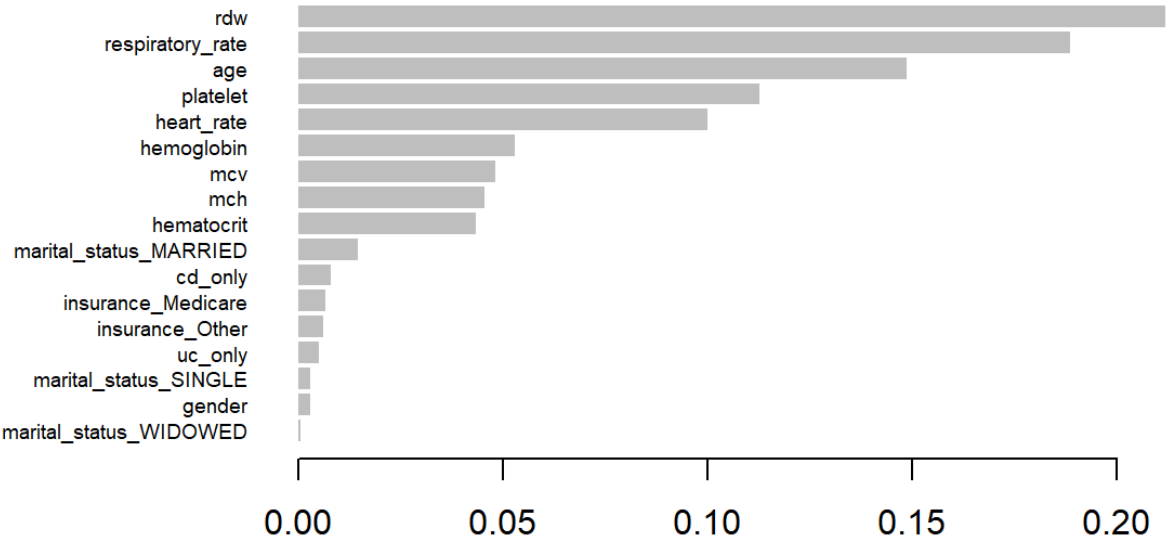
Model Results

Importance Matrix

Hide

```
importance_matrix <- xgb.importance(
  feature_names = colnames(train_data[, -1]),
  model = final_model
)

xgb.plot.importance(importance_matrix)
```

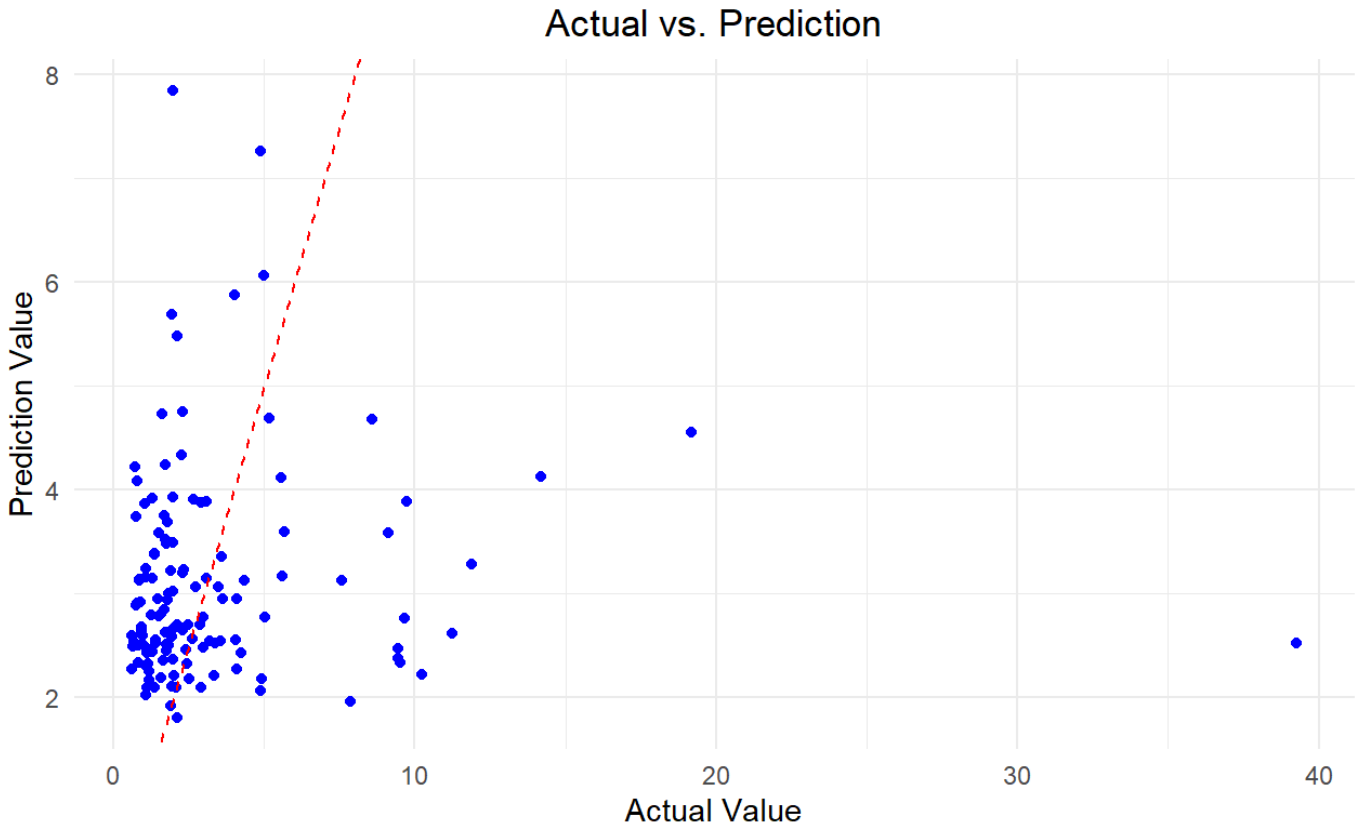


实际值与预测值比较

Hide

```
comparison_df <- data.frame(
  Actual = actuals_test,
  Prediction = preds_test
)

ggplot(comparison_df, aes(x = Actual, y = Prediction)) +
  geom_point(colour = "blue") +  # 绘制散点
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +  # 添加等值线
  theme_minimal() +  # 使用简洁主题
  labs(title = "Actual vs. Prediction", x = "Actual Value", y = "Prediction Value") +  # 添加图标题和轴标题
  theme(plot.title = element_text(hjust = 0.5))  # 居中标题
```

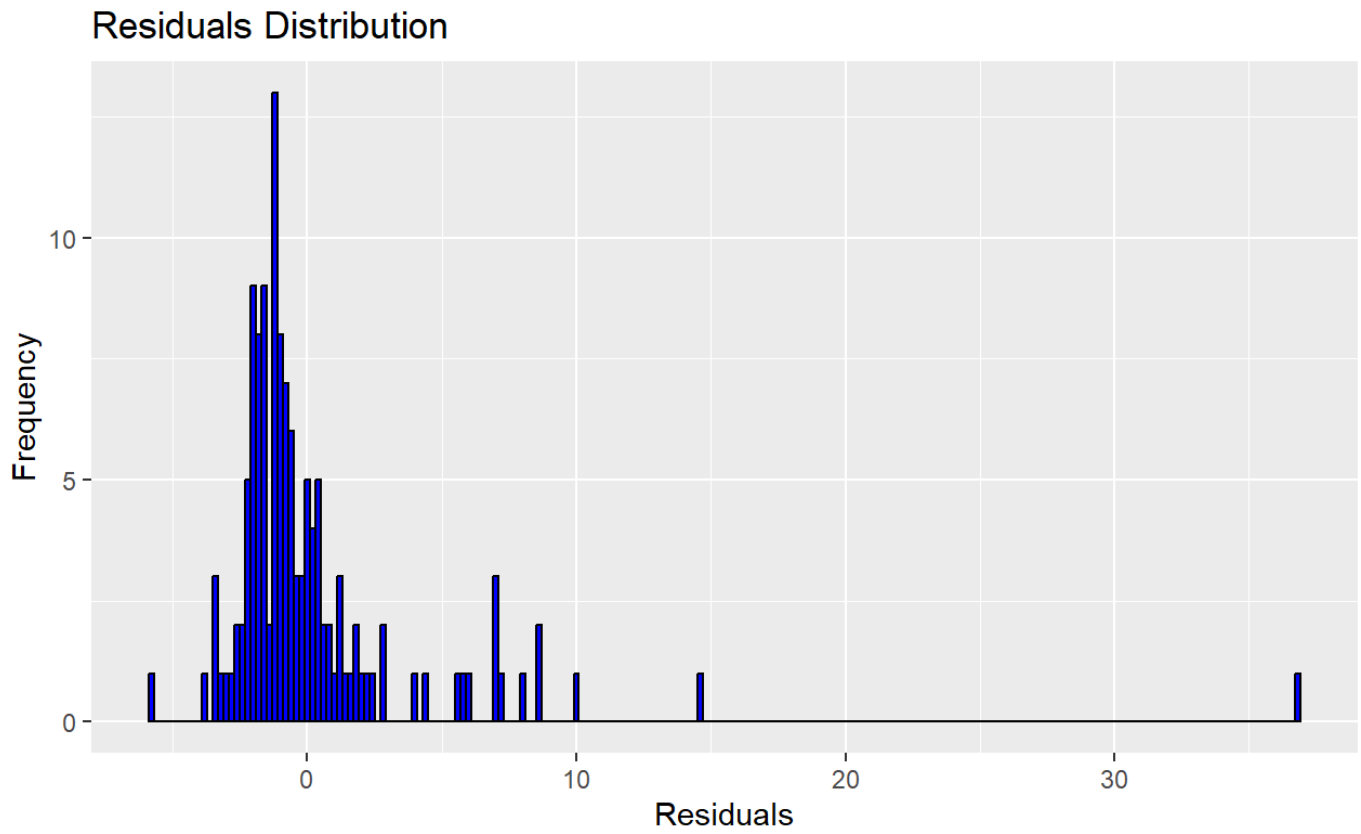


Residual Analysis

Hide

```
residuals <- actuals_test - preds_test

ggplot() +
  geom_histogram(aes(x=residuals), binwidth = 0.2, fill="blue", color="black") +
  ggtitle("Residuals Distribution") +
  xlab("Residuals") +
  ylab("Frequency")
```



Hide

```
ggplot() +
  geom_point(aes(x=preds_test, y=residuals), color="red") +
  ggtitle("Residuals vs. Predicted Values") +
  xlab("Predicted Values") +
  ylab("Residuals") +
  geom_hline(yintercept=0, linetype="dashed", color = "blue")
```