Feature Testing - State and Region

Ryan Rogers

2022-10-17

Library Imports and General Setup

Data Ingestion and Processing

Data Ingestion

```
data <- read.csv("~/Documents/GitHub/Prediction-of-commercial-insurance-payments-for-surgical-procedure
hospital_data <- read.csv("~/Documents/GitHub/Prediction-of-commercial-insurance-payments-for-surgical-
```

Data Processing

```
# Working / Predict Split - Function courtesy of Shruti
split_dataset <- data %>% data_split(count_thresh = 50)
working_set <- split_dataset[[1]]
predict_set <- split_dataset[[2]]
rm(data)
rm(split_dataset)
```

```
# Region isolation
state_reg_mapping <- hospital_data %>% select(MSA_CD, prov_region) %>% distinct() %>% rename(msa = MSA_

# Hospital Dataset Prep - Taken from Baseline Model
hospitals_msa <- hospital_data %>%
  group_by(MSA_CD) %>%
  summarise(Hospitals = n(),
            PctTeaching = sum(teaching == "YES")/n(),
            PctLargeHospital = sum(beds_grp == "500+")/n(),
            Urban = ifelse(sum(urban_rural == "URBAN")/n() == 1, "Urban","Rural"),
            PctPrivate = sum(ownership == "PRIVATE (NOT FOR PROFIT)" | ownership == "PRIVATE (FOR PROFI
  rename(msa = MSA_CD)

rm(hospital_data)
```

```
# Merge working data with hospital data - Taken from Baseline Model
working_set_with_hosp <- left_join(working_set, hospitals_msa, by = "msa") %>%
  select(-priv_pay_mean, -priv_pay_iqr, -mcare_pay_mean, -mcare_pay_sd)

rm(working_set)

working_set_with_reg <- left_join(working_set_with_hosp, state_reg_mapping, by = "msa")
```

```
rm(working_set_with_hosp)
```

Train/Test Split

```
# Dev/Test Split - Taken from Baseline Model
dt = sort(sample(nrow(working_set_with_reg), nrow(working_set_with_reg)*.8)) #Split data
dev_set <-working_set_with_reg[dt,] #80% training data
test_set <-working_set_with_reg[-dt,] #20% test data

#rm(working_set_with_reg)
```

Baseline Model

Initialization

```
# Random Forest model - Taken from Baseline Model
set.seed(123) #Set seed for reproducibility
# Fit Random Forest Model on training data
Random_Forest <- randomForest(
  formula = priv_pay_median ~ .,
  data    = dev_set,
  num.trees = 500,
  mtry = 7,
  nodesize = 20,
  na.action = na.omit
)
```

Prediction on dev_set

```
# Prediction - Taken from Baseline Model
train_predict <- dev_set %>%
  mutate(pred_priv_pay_median = predict(Random_Forest, dev_set)) %>%
  filter(!is.na(pred_priv_pay_median))
```

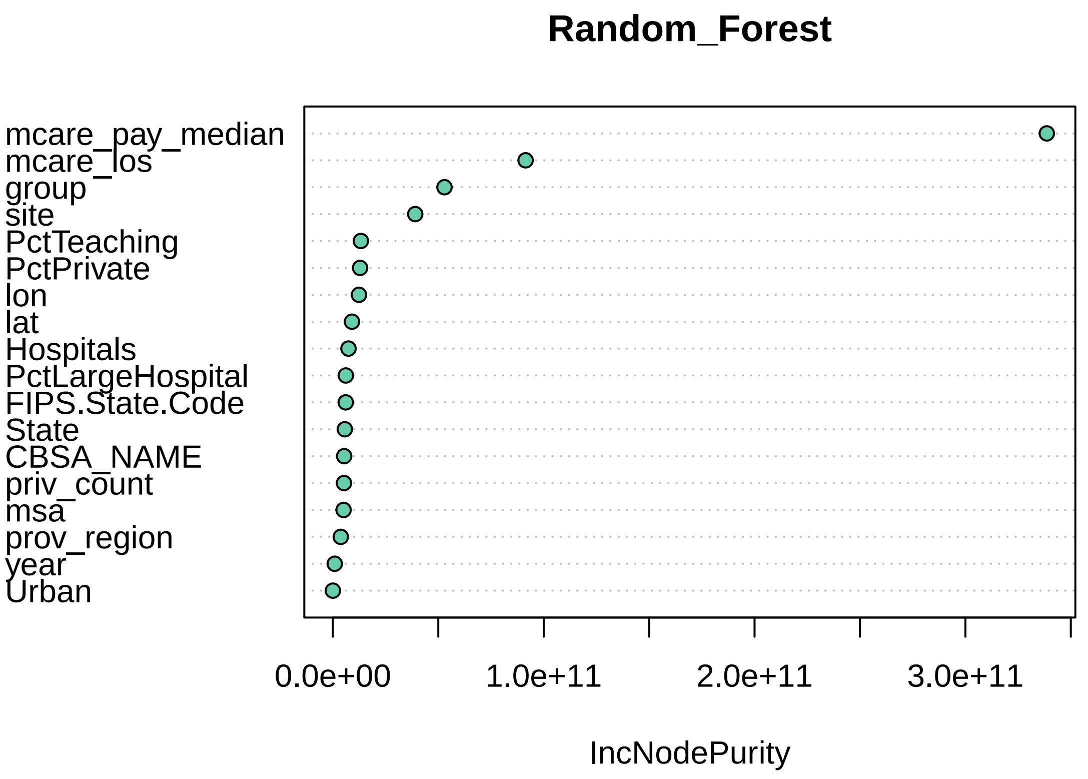
Model Evaluation

```
# Evaluation - Taken from Baseline Model
trn_m = MAPE(train_predict$pred_priv_pay_median, train_predict$priv_pay_median)

train_mape_percent = mean(abs((train_predict$priv_pay_median - train_predict$pred_priv_pay_median)/trai
```

Model Feature Importances

```
# Feature Importances Plot - Taken from Baseline Model
varImpPlot(Random_Forest, bg = "aquamarine3")
```

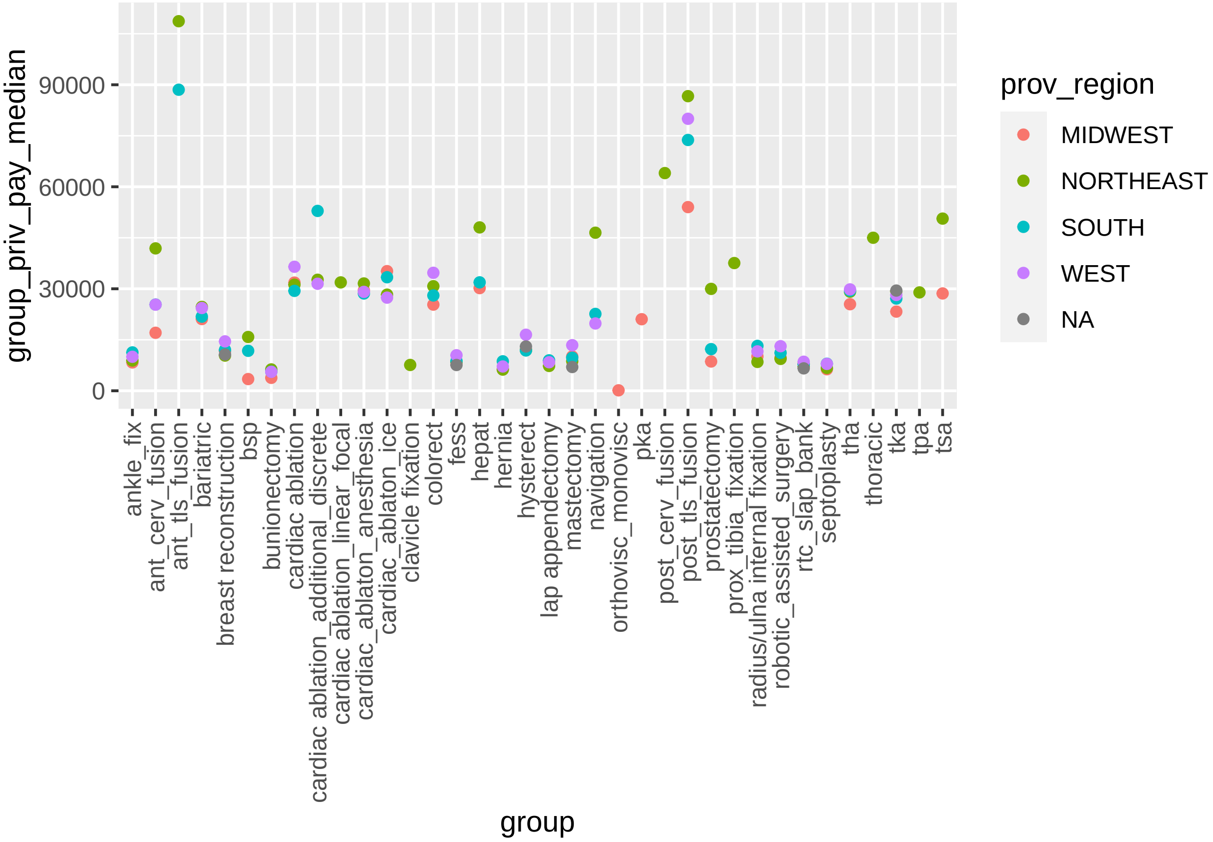


```
# Feature Importances - Tabulated
feat_imps <- data.frame(Random_Forest$importance)
show(feat_imps %>% arrange(desc(IncNodePurity)))
```

```
##                    IncNodePurity
## mcare_pay_median  338603166668
## mcare_los          91406707684
## group              52936650828
## site               39070100385
## PctTeaching        13276172374
## PctPrivate         12910913232
## lon                12377927738
## lat                 9056959123
## Hospitals           7403559784
## PctLargeHospital    6157787714
## FIPS.State.Code     6126455602
## State               5672062422
## CBSA_NAME           5354260554
## priv_count          5271776854
## msa                 5074966570
## prov_region         3747919896
## year                 894707987
## Urban                        0
```

```
rm(feat_imps)
```

Correlation by Group and Region



- Not a ton of variation in private payment median based on region for most procedures
- Some procedures like ant_cerv_fusion, ant_tls_fusion do see some variation by region