Feature Testing - SAIPE

Ryan Rogers

2022-10-17

Library Imports and General Setup

Data Ingestion and Processing

Data Ingestion

General Data

```
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-p
```

Feature Data

```
saipe_2018 <- read_excel("~/Documents/GitHub/Prediction-of-commercial-insurance-payments-for-surgical-p
saipe_2019 <- read_excel("~/Documents/GitHub/Prediction-of-commercial-insurance-payments-for-surgical-p
saipe_2020 <- read_excel("~/Documents/GitHub/Prediction-of-commercial-insurance-payments-for-surgical-p

saipe_data <- rbind(saipe_2018, saipe_2019)
saipe_data <- rbind(saipe_data, saipe_2020)

saipe_data <- saipe_data %>%
  select(c(Name, `Poverty Percent, All Ages`, `Median Household Income`, year)) %>%
  rename(`State_Poverty_Percent_All_Ages` = `Poverty Percent, All Ages`) %>%
  rename(`State_Median_Household_Income` = `Median Household Income`)

rm(saipe_2018)
rm(saipe_2019)
rm(saipe_2020)
```

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)
```

```
# 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)
```

```
# Merge working data with SAIPE data
working_set_with_saipe <- left_join(working_set_with_hosp, saipe_data, by = c('State' = 'Name','year'))

rm(saipe_data)
rm(working_set_with_hosp)
```

Train/Test Split

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

#rm(working_set_with_saipe)
```

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))
```

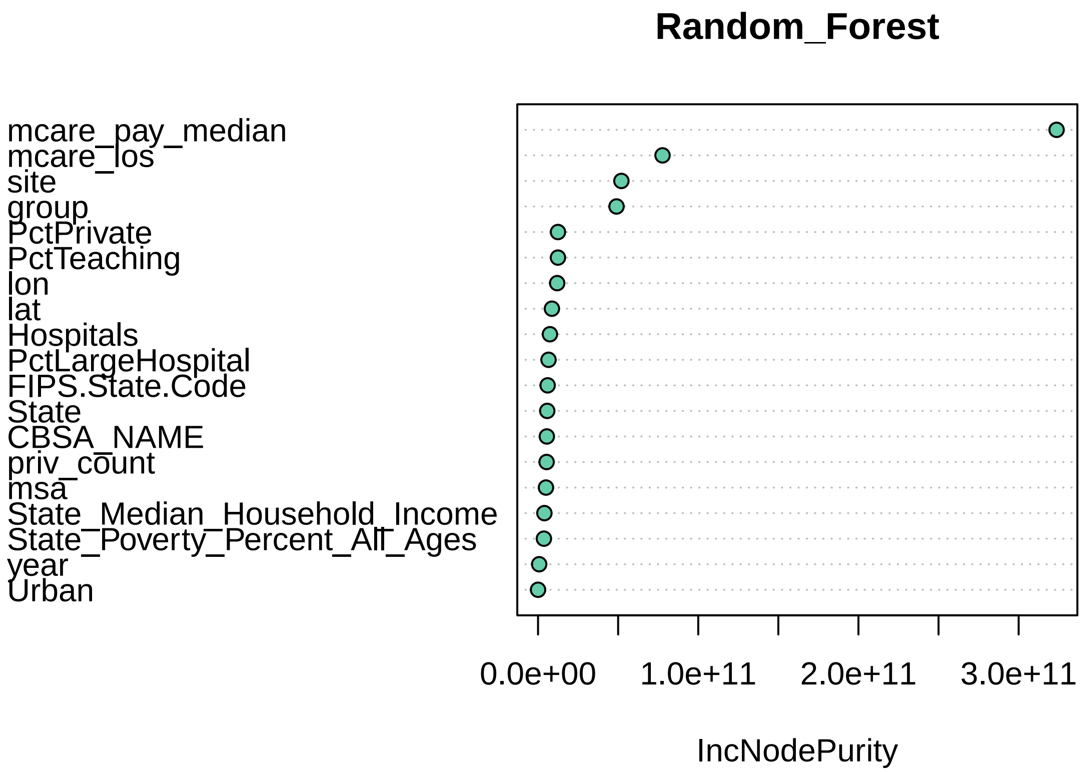
Model Evaluation

```r
# 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

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




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

```
##                                  IncNodePurity
## mcare_pay_median                  323693014955
## mcare_los                          77713361746
## site                               52025405767
## group                              49003692292
## PctPrivate                         12439323810
## PctTeaching                        12435264329
## lon                                11945030444
## lat                                 8659541723
## Hospitals                           7420241813
## PctLargeHospital                    6553211647
## FIPS.State.Code                     6001861457
## State                               5747585302
## CBSA_NAME                           5490612710
```

```
## priv_count                        5344420131
## msa                               4953786691
## State_Median_Household_Income     3975563055
## State_Poverty_Percent_All_Ages    3676547083
## year                               720689665
## Urban                                      0
```

```
rm(feat_imps)
```

Correlations at Group Level

```
##                        group_priv_pay_median group_mcare_pay_median     poverty
## group_priv_pay_median             1.00000000              0.9110905 -0.05882791
## group_mcare_pay_median            0.91109049              1.0000000 -0.09426150
## poverty                          -0.05882791             -0.0942615  1.00000000
## income                            0.09981004              0.1603948 -0.84826563
##                             income
## group_priv_pay_median   0.09981004
## group_mcare_pay_median  0.16039480
## poverty                -0.84826563
## income                  1.00000000
```

- Not super strong correlations of poverty and income with payments, but could potentially be helpful after clustering