Tree Model Performance

Gwen Rino

4/20/2018

Model Description

This decision tree model uses all the variables except `week_start_date`, which is the unique identifier. It includes the engineered feature `year.season`. I ran the model with missing values imputed two ways, first by median values and second by k nearest neighbor values. I used the `caret` package to cross validate these models using the bootstrapping method.

```
# With knn imputation
tree_2 <- train(total_cases ~ . -week_start_date,
                data = dengue.knn,
                method = "rpart")

# With median value imputation
tree_3 <- train(total_cases ~ . -week_start_date,
                data = dengue.med,
                method = "rpart")
```

Model Evaluation

Both the cross validations returned best models with the tuning parameter cp = 0.016. However, in both these cases the MAE was a bit greater than 20 – better than the Naive and Linear models, but not by much. There was little difference between the MAEs of the rpart models with different imputation methods, which makes sense because a tree model shouldn't be very sensitive to this kind of preprocessing.

Most of the nodes in the visualized decision trees concern elements of time rather than weather. Again, it is clear that time is critical to understanding this data set. I need to learn about time series!

```
print(tree_2)
```

```
## CART
##
## 936 samples
##  26 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 936, 936, 936, 936, 936, 936, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE      Rsquared   MAE
##   0.01609336  38.21654  0.4681548  21.39936
##   0.02550193  38.97228  0.4410404  22.01423
##   0.19421475  45.27368  0.4982142  25.51078
##
## RMSE was used to select the optimal model using the smallest value.
```

```
## The final value used for the model was cp = 0.01609336.
```

```
print(tree_3)
```

```
## CART
##
## 936 samples
##  26 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 936, 936, 936, 936, 936, 936, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE      Rsquared   MAE
##   0.01609336  36.03093  0.5213433  20.60488
##   0.02550193  36.81828  0.5029362  21.21105
##   0.19421475  44.70778  0.4863066  24.99904
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.01609336.
```