Predicting Transport Expenditures

Alberto Rodriguez

4/12/2019

QUESTION

Can we use information on Household Characteristics, Income and Expenditure to predict their transport costs?

BUILDING A DATABASE

- National Survey of Household Income and Expenses by the National Institute of Statistics and Geography of Mexico which includes information on more than 74,000 observations of households in the country
- Inflation Reports from the National Bank
- Gas Prices from the Energy Regulatory Commission

(and merging it all together with geographic and month data)

Splitting the Data

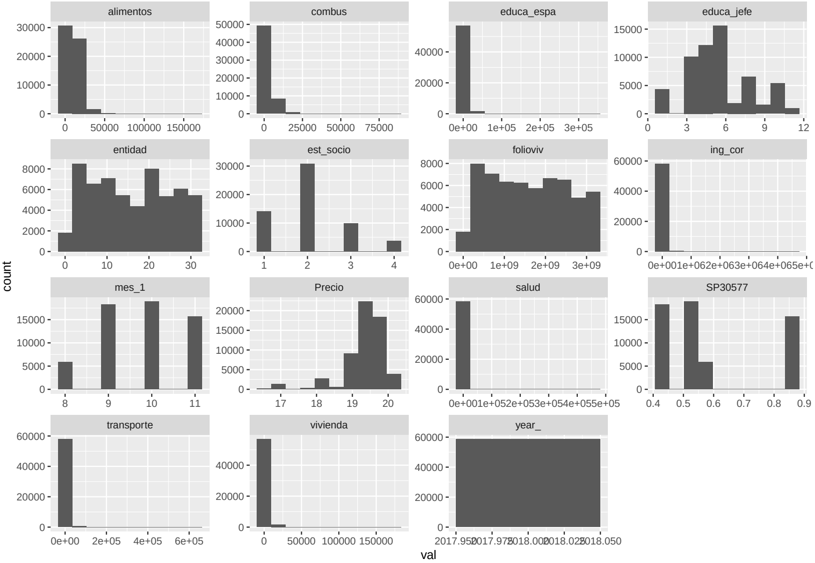
```r
set.seed(123)
index = createDataPartition(enigh2018_clean$transporte,p=.8
train_data = enigh2018_clean[index,] # Use 80% of the data
test_data = enigh2018_clean[-index,] # holdout 20% as test
dim(train_data)
```

```
## [1] 58725    16
```

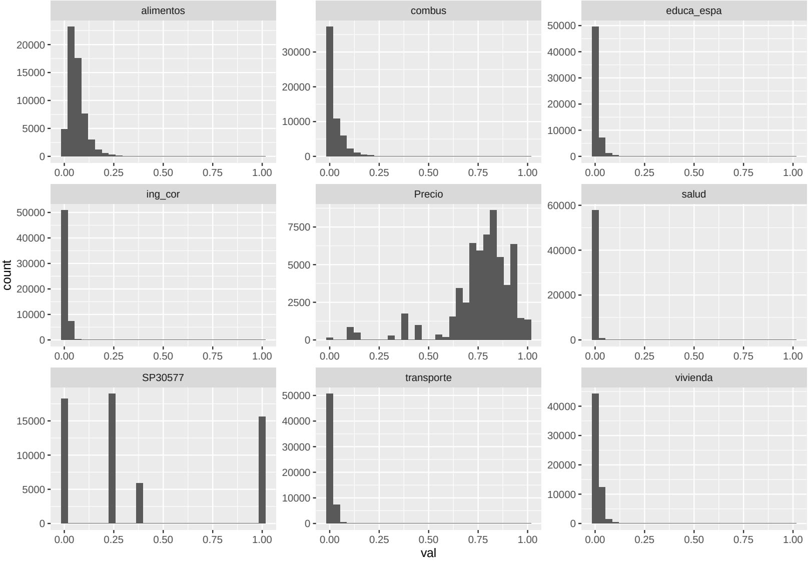
```r
dim(test_data)
```

```
## [1] 14680    16
```

Looking at Data



Baked Goods!



Cross Validation

```
set.seed(1004) # set a seed for replication purposes
```

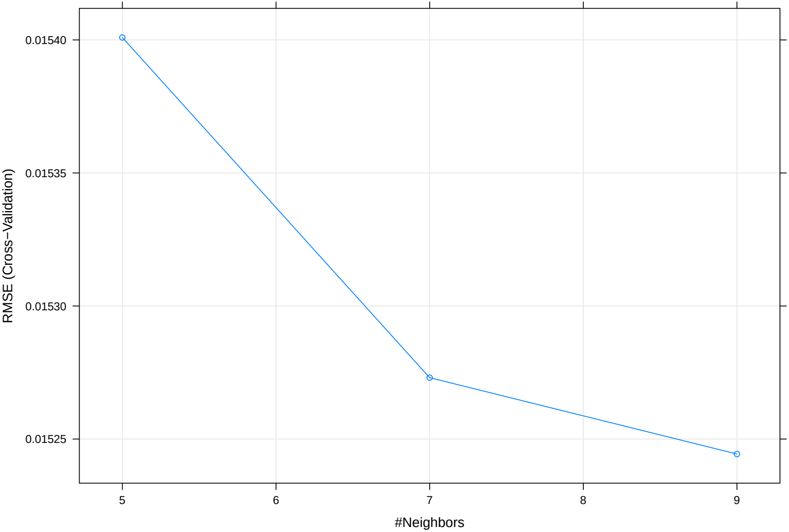
Linear Model

```
mod_lm
```

```
## Linear Regression
##
## 58725 samples
##    21 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 9787, 9787, 9788, 9788, 9788, 9
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.01351838  0.3824566  0.004729118
##
## Tuning parameter 'intercept' was held constant at a valu
```

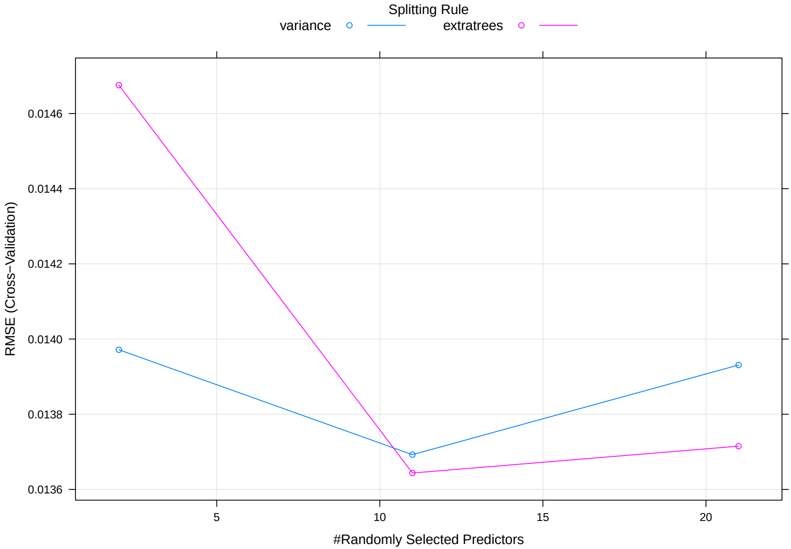
K-Regressors

```
plot(mod_knn)
```



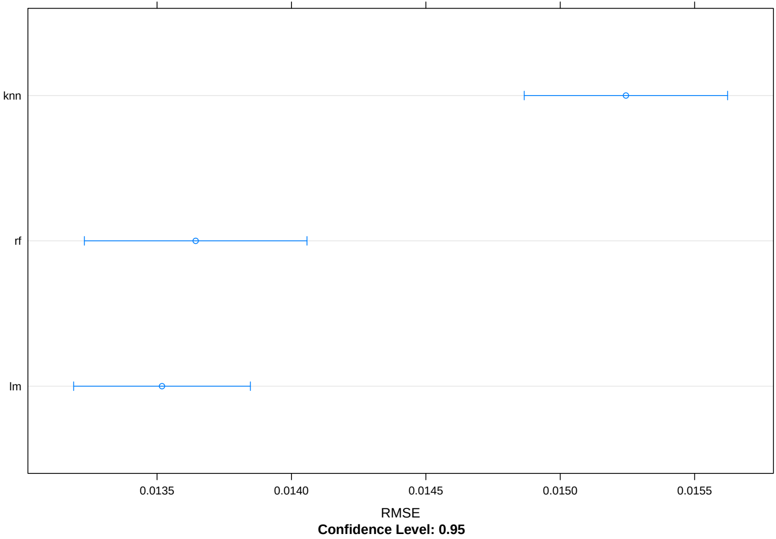
Random Forest

```r
plot(mod_rf)
```



Comparison

```
dotplot(resamples(mod_list),metric = "RMSE")
```



Test!

(not good at all)

```
pred <- predict(mod_knn,newdata = test_data2)
mse = sum(test_data2$transporte-pred^2)/nrow(test_data2)
mse
```

```
## [1] 0.009360972
```

Next Steps

- Change folds
- Build Visualizations
- Maybe Add Variables