Capital Bikeshare

Jake Berberian

Overview

The Data

The dataset was found using UCI's Machine Learning Repository. It contains data spanning from 1 January 2011 to 31 December 2012 from Capital Bikeshare's official website joined with weather data from I-weather and the district's official holiday schedule. It contains the following variables:

Variables		
instant	holiday	hum
dteday	weekday	windspeed
season	workingday	casual
yr	weathersit	registered
mnth	temp	cnt
hr	atemp	hum

Hypotheses

Our initial hypotheses are the following:

- Workdays/holidays and days with lower temperatures/worse weather will result in lower usage.
- We will see a decrease in users in the high summer months (specifically July and August).
- We can expect to see `holiday` and `weekday` play the largest role in the number of casual users.

Discriminant Analysis

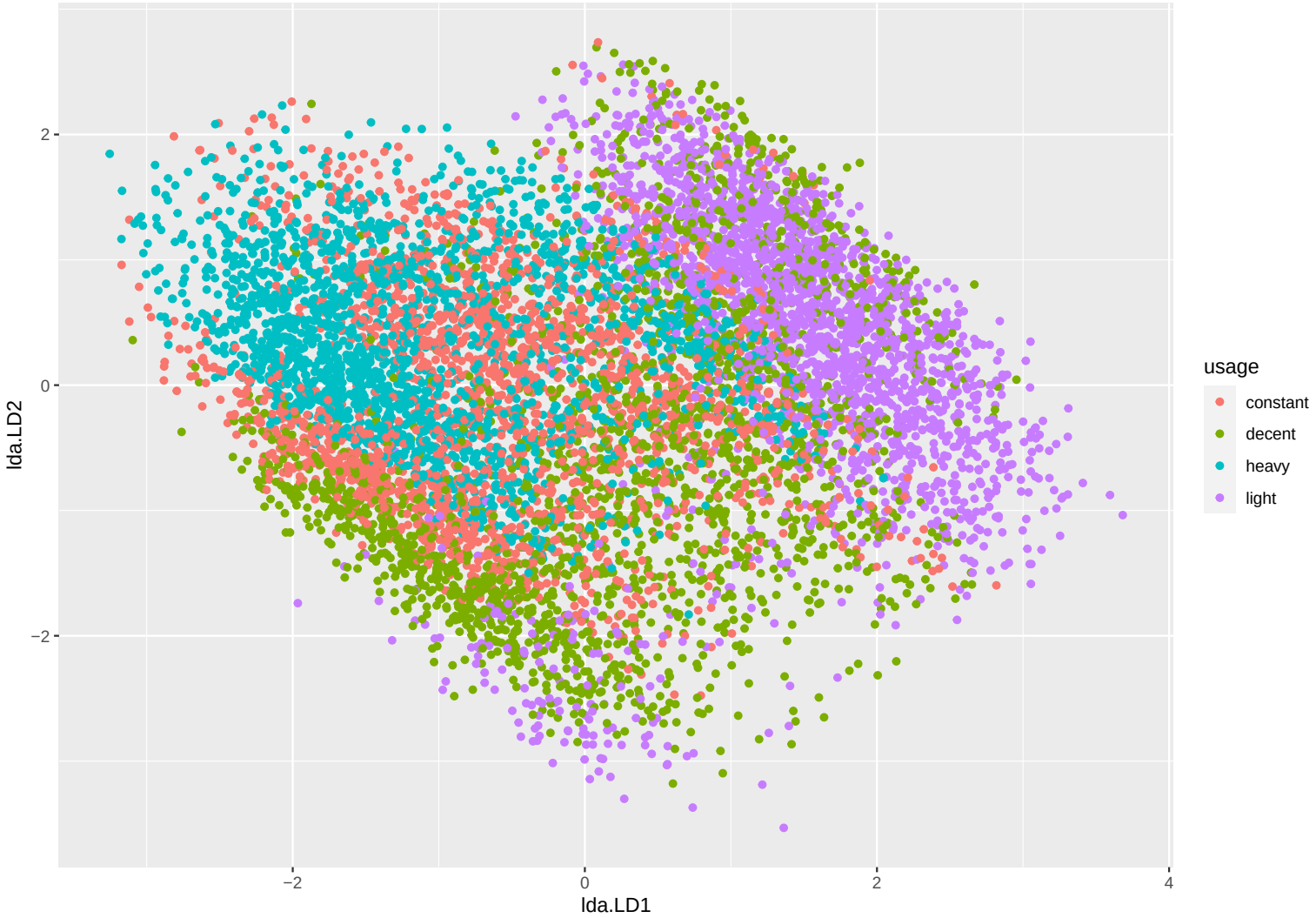
The LDA Model

- First, we'll want to bin the hourly data into four categories: heavy usage, constant usage, moderate usage and light usage.
- Run calculations twice
  (1) Using `CV = TRUE` to get prediction of class membership from LOOCV.
  (2) Using `CV = FALSE` to allow us to use `predict()` on our test set and get a classification rate.
- Our classification rate indicates that we've correctly classified a little over half of the counts (~56.25%).

	constant	decent	heavy	light
constant	1548	899	1486	421
decent	823	1740	368	1421
heavy	879	317	2706	452
light	55	385	98	3781

	constant	decent	heavy	light
constant	734	444	393	26
decent	441	854	163	184
heavy	761	195	1381	53
light	229	723	211	1898

class_rate
0.560069


lda.LD2
2
0
−2
−2
0
2
4
lda.LD1
usage
constant
decent
heavy
light

The QDA Model

	constant	decent	heavy	light
constant	803	590	355	39
decent	308	805	105	301
heavy	840	137	1567	28
light	214	684	121	1793

Our classification rate is 0.572, which a little better than LDA (56.01%).

Conclusions

- LDA vs. QDA trade-off (Bias-variance trade-off)
  - LDA is less flexible than QDA, with fewer parameters.
  - LDA can suffer from high bias when when the classes have different covariance matrices.
- Since our training set is fairly large (8689 observations), the variance of the classifier is not a major concern.

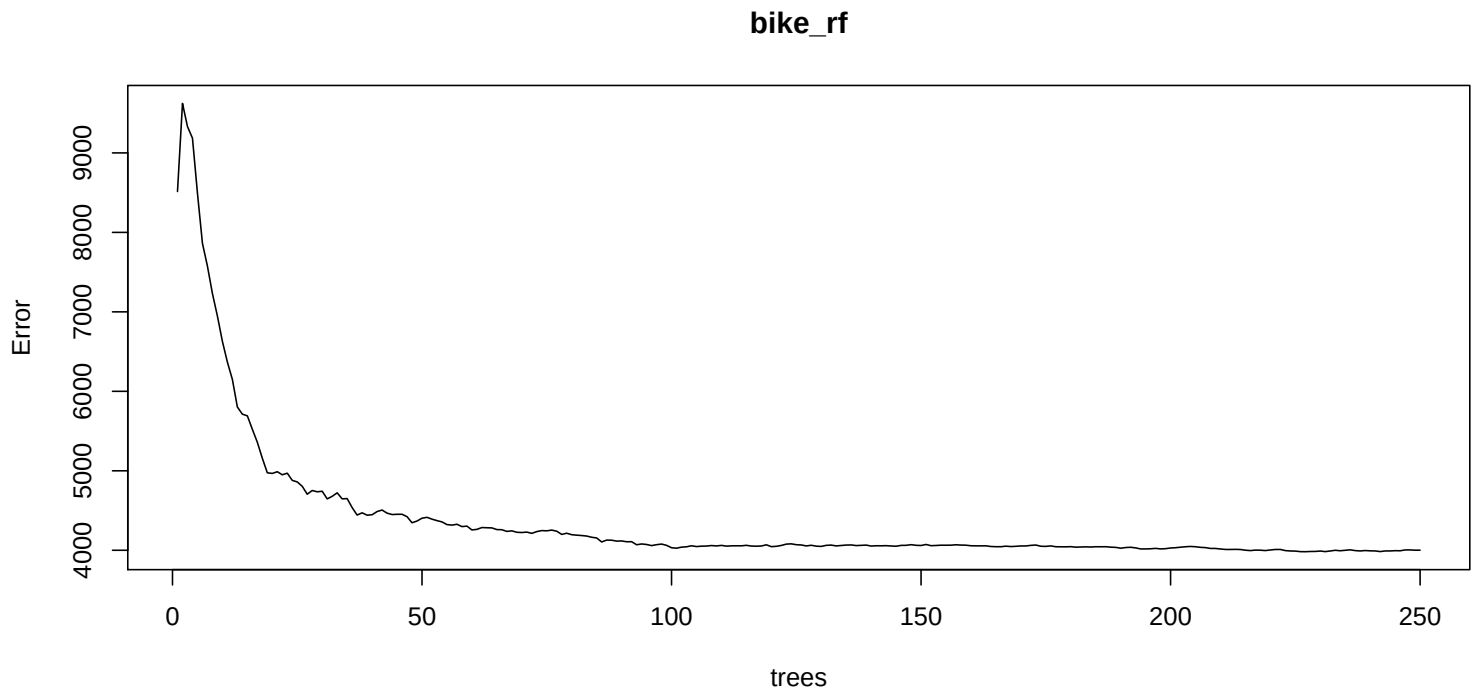
Conclusions

- A possible issue here is that much of the "heavy"-classified data comes from year 2, which skews the data. Since the weather is evenly distributed between years, it makes it difficult for the model to correctly classify observations.
- Since we'll trade some bias for variance, we'll go with our QDA model. It better explains the data and since n in the training set is fairly large, the effects of variance are mitigated.

Random Forest

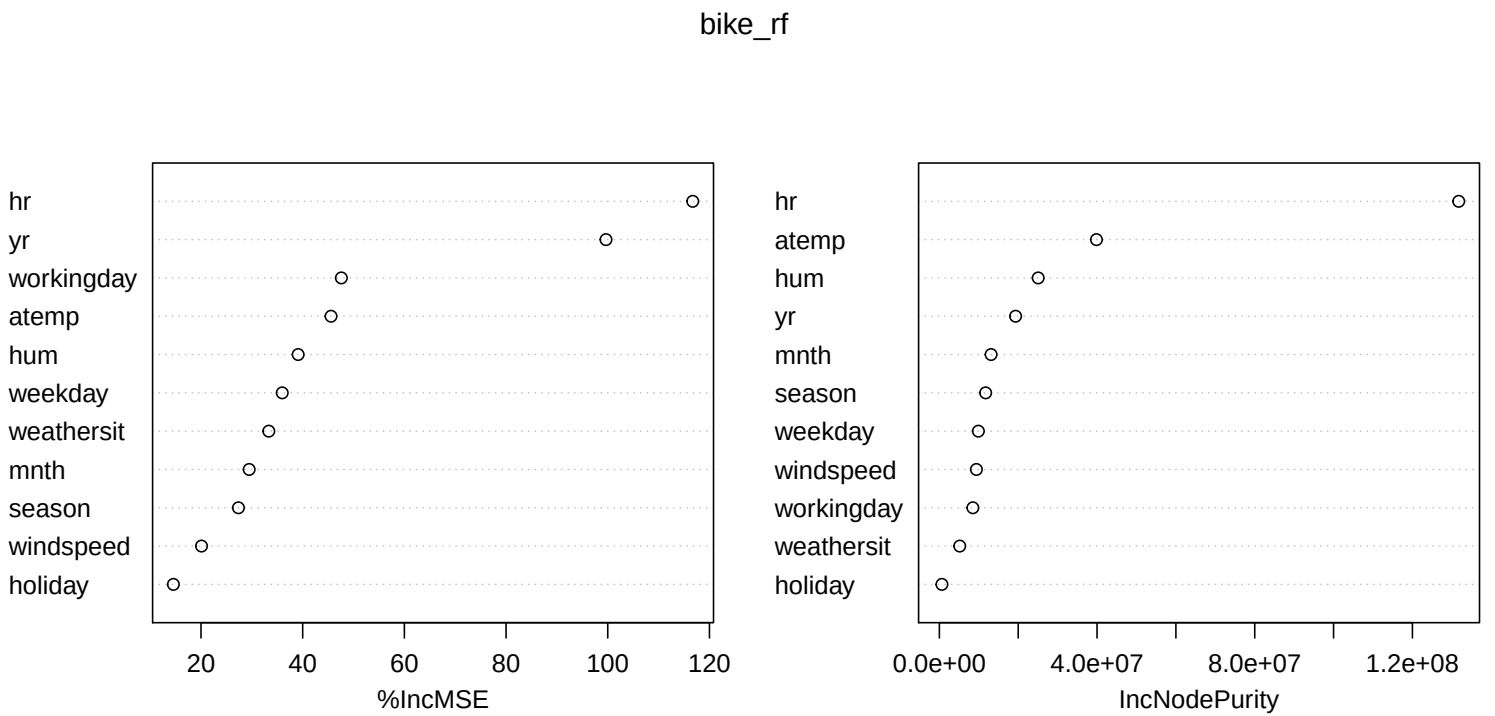
The Model

- We'll use a regression approach, as the binned data provides too much variance between groups.
- We'll first try with 250 trees, as to get a good baseline and to not use too much computational power.
- Judging from our plot, it seems the error levels off around 100 trees, but we'll explore further.



Importance of variables

- We can see that the hour of day has a large influence in the model. Meanwhile, `holiday` has the lowest impact, which disproves our original hypothesis that holiday vs. non-holiday would have a big influence.



Cross-validation

- The optimal number of trees seem to be 227. We'll optimize our model to to follow this.
- From our new, optimized model, the mean square error of prediction, given by the test set cross-validation, is 3770.262.
- Furthermore, we can optimize the number of variables tried at each split, but this would take a good amount of computational power, so we'll be content with `ntree` $= 227$ and `mtry` $= 3$.

Min. MSE	Max. R Sq.	PMSE
227	227	3770.262

Conclusions

- Our PMSE is somewhat large. So while the % of variance explained is above 90%, our model doesn't seem to actually perform that well on the testing data.
- Random forests take a lot of computational power compared to running a regular regression using `lm()`.

Discussion

Discussion

- Overall, it's difficult to predict the number of bikers. Perhaps binning would've made it easier, but would not have provided actionable insights.
- If the data has converged more to the norm (aka the company still wasn't growing), perhaps our models would have more predictive power.
- Outcome of hypotheses:
  - Holidays nor weather played a large part in any of our models. However, workdays did.
  - There's no decrease in users in the hotter summer months.
  - Again, `holiday` didn't play a large role, but `weekday` did.
- Further studies
  - Logistic regression
  - More years of data
  - Testing w/o weather (seemed to not play that large of an impact)

Citations

Citations

Fanaee-T, Hadi, and Gama, Joao, *Event labeling combining ensemble detectors and background knowledge*, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R.* 7th ed., Springer, 2017.