Capital Bikeshare

Jake Berberian

Overview

The Company

Capital Bikeshare is a bikeshare system that supports the DMV-area. It has around 5000 bikes system-wide, with almost 600 stations throughout. They charge $2 for a 30-minute trip, $8 for the day, or $85 for a year-long membership, which gives access to unlimited 30-minute rides.

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

More on the variables

Let's take a closer look at some of the variables:

- `dteday` is the date of the observations
- `season` is the season, 1 = winter, 2 = spring, 3 = summer, 4 = fall
- `holiday` is decided by the District's official holiday calendar' 0 = no holiday, 1 = holiday
- `weathersit` describes the weather:
  - 1 = clear, few clouds, or partly cloudy
  - 2 = mist and/or cloudy
  - 3 = light snow, light rain, thunderstorm
  - 4 = heavy rain, ice pallets, heavy thunderstorm, snow + fog
- `temp` is a normalized temperature statistic in Celsius.
- `atemp` is a normalized "real feel" temperature statistic in Celsius.
- `casual`, `registed`, and `cnt` are count statistics counting the number of non-registered users, registered users, and total users, respectively.

The Plan

1. Explore the data
2. Multiple Linear Regression to predict number of riders
3. LDA/QDA to predict the binned number of riders
4. Random Forests to predict the number of riders

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.

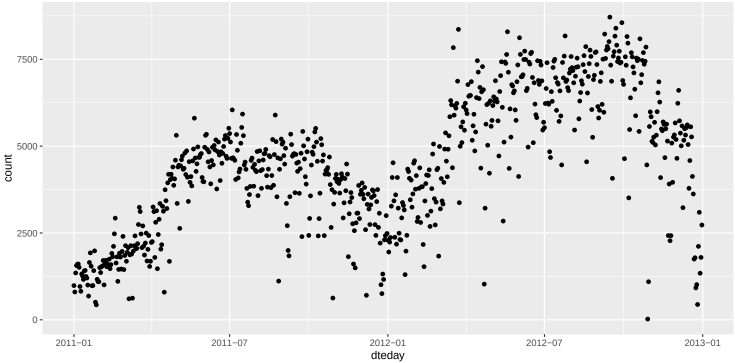
Problems

- Predicting number of users will be difficult, as Capital Bikeshare was gaining notoriety during these years.
- Year-to-year data has likely now stabilized, but we don't expect to see any definitive patterns.
- The data takes place over a two-year period, so it's hard to gauge a ton when each date has only two data points.
- While Capital Bikeshare has year-by-year data, it does not include all the same variables. As a result, we'll split our data into testing and training sets.

Data Exploration

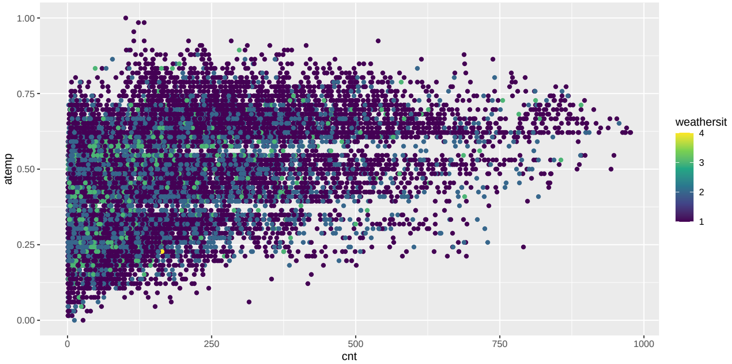
Riders per day

As mentioned, we can see that the number of users increased greatly during 2012. Furthermore, we see evidence of a cyclical shape, which seems to indicate that the date/season does have an effect: winter months see lower usage, while the summer months see some of the highest usage. This contradicts our original hypothesis that suggested that July and August would see slightly lower counts.



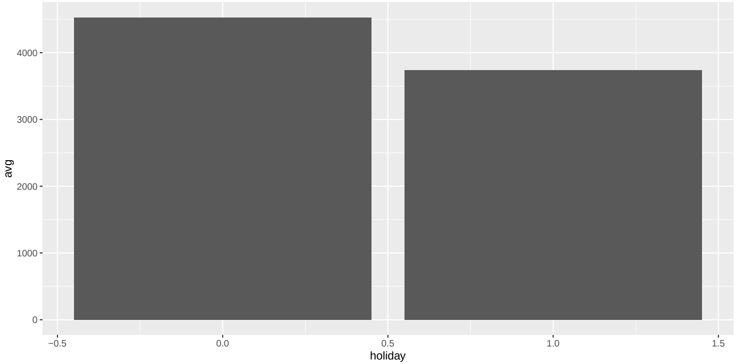
Temperature & weather

The first thing to notice is that there seems to be some sort of funnel shape to the plot. This would suggest that there's some "optimal" real-feel temperature for bikshares. Furthermore, we see very few observations of extreme weather. In fact, there are only three days over the course of the two years: 26 Jan 2011, 9 Jan 2012, and 21 Jan 2012.



Holidays

As we see, the average number of users on holidays (3735) is sizabley smaller than the average number on non-holidays (4527). However, some of these holidays have less bikers than expected. For example, New Year's Eve and Day in both 2012 and 2013 had around 2200 bikers (sans New Years Day 2011). Average for any given day is 4504 bikers.



Multiple Linear Regression

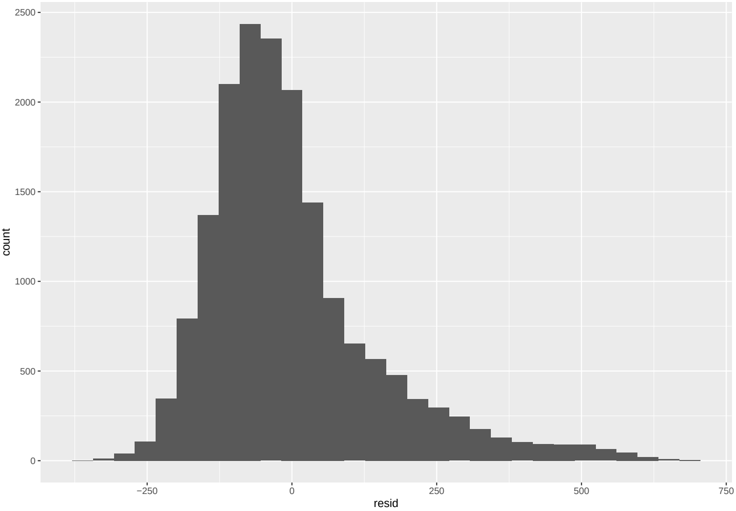
Model

We'll run a regression on our data. We removed redundant variables and the variable `yr`, as that is years after 2011 which isn't necessary. The resulting coefficients are below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.6571542	10.0588514	1.8547997	0.0636587
season	18.7825275	2.7384276	6.8588732	0.0000000
mnth	0.0294694	0.8532431	0.0345380	0.9724489
hr	7.4451260	0.2443268	30.4720036	0.0000000
holiday	-15.3813096	9.7965393	-1.5700758	0.1164339
weekday	1.8987827	0.8053713	2.3576488	0.0184132
workingday	10.3343411	3.5684198	2.8960553	0.0037883
weathersit	1.0685877	2.8036207	0.3811456	0.7031045
atemp	332.1139884	10.0096283	33.1794526	0.0000000
hum	-224.9190547	10.2061765	-22.0375431	0.0000000
windspeed	41.7130077	14.0929410	2.9598512	0.0030862

Residual Analysis

- We'll first check to see if our residuals are normally distributed. The plot shows a decent right skew, so we'll proceed with caution.



Multicollinearity

	vif
season	3.5537
mnth	3.3532
hr	1.1242
holiday	1.0855
weekday	1.0155
workingday	1.0772
weathersit	1.2931
atemp	1.1687
hum	1.5125
windspeed	1.1390

As expected, `temp` and `atemp` have extremely high variance inflation factors. Furthermore, `season` and `mnth` have higher VIFs, which also would make sense. We'll create a reduced model without `temp` or `season`. This is because their

Model 2

We'll try out this model with `temp` and `season` removed.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.2253423	10.0246343	2.6160897	0.0089097
mnth	4.8352030	0.4882193	9.9037520	0.0000000
hr	7.3991103	0.2448816	30.2150471	0.0000000
holiday	-17.0673875	9.8193986	-1.7381296	0.0822234
weekday	1.7911213	0.8073514	2.2185150	0.0265455
workingday	10.7512667	3.5773538	3.0053686	0.0026602
weathersit	0.4910302	2.8097796	0.1747575	0.8612742
atemp	349.7268637	9.7002644	36.0533329	0.0000000
hum	-222.4138532	10.2266588	-21.7484379	0.0000000
windspeed	39.3025143	14.1258804	2.7823055	0.0054091

Model Comparison: General Linear F-test

With an F-stat of 47.0441, we have a corresponding p-value of less than 0.00001. Thus, we can conclude with strong statistical certainty that our full model is favored. However, we need to remember that there was strong multicollinearity in our full model.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
8679	193106141	NA	NA	NA	NA
8678	192064942	1	1041200	47.04414	0

Model Comparison: Cross-validation

- Our mean square error, calculated through cross-validation is very high.
- Linear regression may not perform well as a predictive power.
  - Look at year-to-year inconsistent

PMSE
21536.48

Conclusions

- Overall, neither linear model did a great job of explaining the variance in `cnt`. Their respective R^2 values:
  - Full model: 0.3379
  - Reduced model: 0.3343
- It seems that there is a lot of correlation between variables and that with so many variables, we could perhaps try some dimensionality-reduction techniques in the future (PCA, etc.)
  - Majority of variables are important, so look at better variable selection methods.

Discriminant Analysis

The Setup

- First, we'll want to bin the hourly data into four categories: heavy usage, constant usage, moderate usage and light usage.
- We'll then run LDA and QDA on our data and cross-validate using our testing set
- Finally, we'll discuss if LDA or QDA provides a better clustering method.

'# Discriminant Analysis

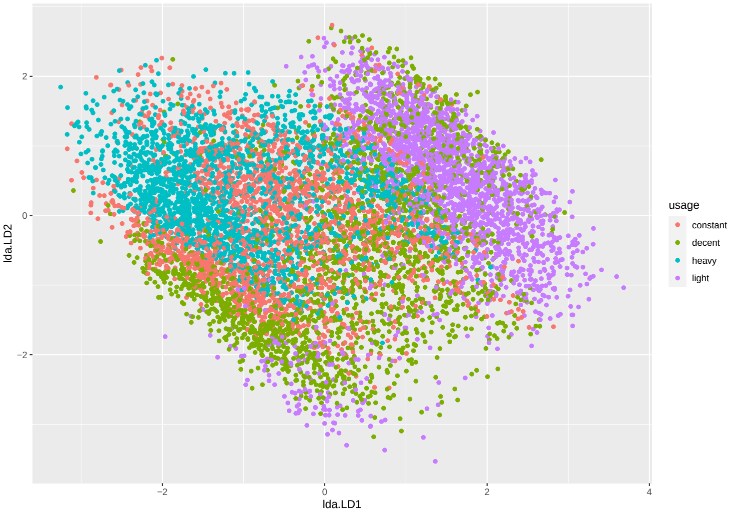
The LDA Model

- 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 (~56.01%).

	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



2
0
−2
lda.LD2
−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.

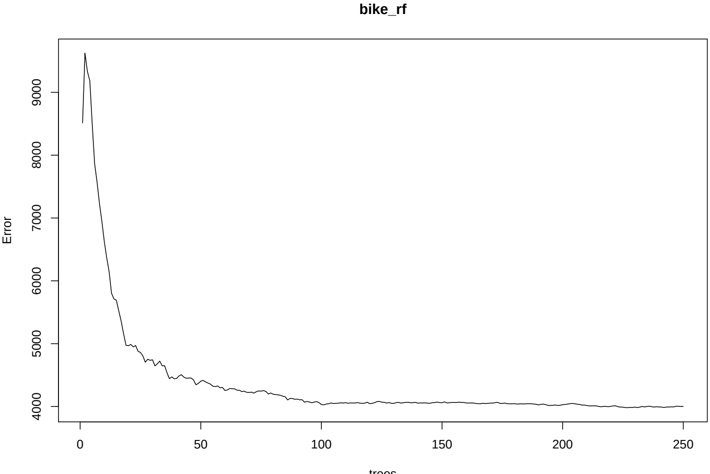
Random Forest

The Setup

- We can use either a classification approach (with the binned data) or regression approach (with `cnt` variable)
  - Ultimately, this is relatively important to Capital Bikeshare, the binned data provides too much variance between groups (heavy takes the range of 281 to 977 bikers. That's a difference of three-fold).
- So, we'll use the `cnt` variable and run a random forest regression.
- We'll try to find the optimal number of trees and number of variables that are randomly sampled at each split.

The Model

- 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.

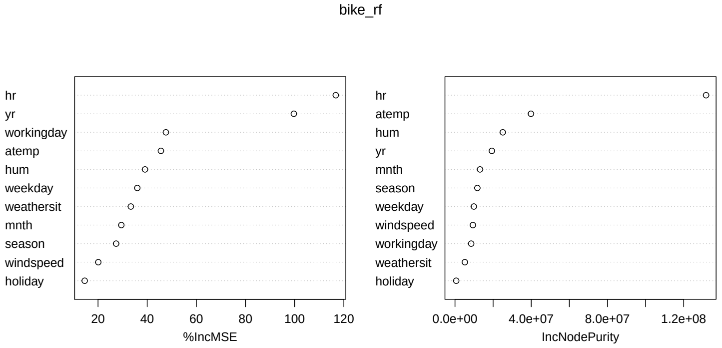


Importance of variables

	%IncMSE	IncNodePurity
season	27.36379	11739072.7
yr	99.65486	19333599.4
mnth	29.47383	13115835.9
hr	116.71486	131750937.0
holiday	14.56832	651400.7
weekday	35.97770	9895061.2
workingday	47.60292	8501793.3
weathersit	33.33767	5164409.5
atemp	45.57385	39847498.5
hum	39.10847	25067128.7
windspeed	20.10389	9385486.9

Importance of variables

- We can see from both the table and plot 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 211. 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 3261.23.
- 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` = 211 and `mtry` = 3.

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

Conclusions

- Our PMSE is very large, but much smaller than that of linear regression (by over 6x). So while the % of variance explained it 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.