Bike Sharing Demand

Max Llewellyn

https://www.kaggle.com/c/bike-sharing-demand

Agenda

- Problem and Data Overview
  - Who, What, Why, Where, How
- Exploratory Data Analysis
- Other Solutions
  - EDA & Ensemble Model (Top 10 Percentile)
  - Comprehensive EDA with XGBoost (Top 10 percentile)
  - bikes
- Modeling

Problem and Data Overview

What is Bike Sharing?

- Stations with bikes to rent
- Users ride the bike from station to station
- Popular in most metro areas

- CDPHP *Cycle!* bike Sharing is expanding in the capital district!

Where/Who is this data from?

- Located in Washington DC
- Capital Bikeshare provided data on each individual ride and Hadi Fanaee Tork condensed it into day to day data for Kaggle.

How is the Data Formatted

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
1	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0000	3	13	16
2	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0000	8	32	40
3	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0000	5	27	32
4	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0000	3	10	13
5	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0000	0	1	1
6	2011-01-01 05:00:00	1	0	0	2	9.84	12.880	75	6.0032	0	1	1
7	2011-01-01 06:00:00	1	0	0	1	9.02	13.635	80	0.0000	2	0	2
8	2011-01-01 07:00:00	1	0	0	1	8.20	12.880	86	0.0000	1	2	3
9	2011-01-01 08:00:00	1	0	0	1	9.84	14.395	75	0.0000	1	7	8
10	2011-01-01 09:00:00	1	0	0	1	13.12	17.425	76	0.0000	8	6	14

How is the Data Formatted

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
1	2011-01-01 00:00:00	spring	FALSE	FALSE	Good	9.84	14.395	81	0.0000	3	13	16
2	2011-01-01 01:00:00	spring	FALSE	FALSE	Good	9.02	13.635	80	0.0000	8	32	40
3	2011-01-01 02:00:00	spring	FALSE	FALSE	Good	9.02	13.635	80	0.0000	5	27	32
4	2011-01-01 03:00:00	spring	FALSE	FALSE	Good	9.84	14.395	75	0.0000	3	10	13
5	2011-01-01 04:00:00	spring	FALSE	FALSE	Good	9.84	14.395	75	0.0000	0	1	1
6	2011-01-01 05:00:00	spring	FALSE	FALSE	Fair	9.84	12.880	75	6.0032	0	1	1
7	2011-01-01 06:00:00	spring	FALSE	FALSE	Good	9.02	13.635	80	0.0000	2	0	2
8	2011-01-01 07:00:00	spring	FALSE	FALSE	Good	8.20	12.880	86	0.0000	1	2	3
9	2011-01-01 08:00:00	spring	FALSE	FALSE	Good	9.84	14.395	75	0.0000	1	7	8
10	2011-01-01 09:00:00	spring	FALSE	FALSE	Good	13.12	17.425	76	0.0000	8	6	14

Why are we analyzing it?

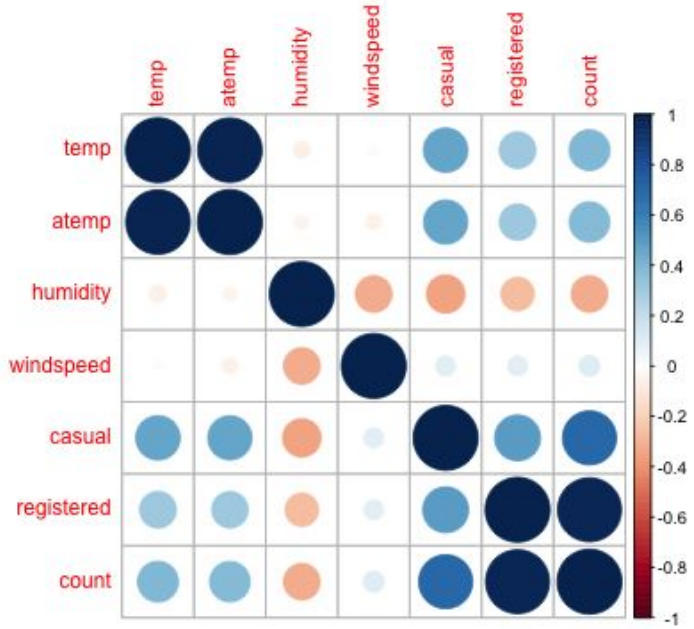
kaggle

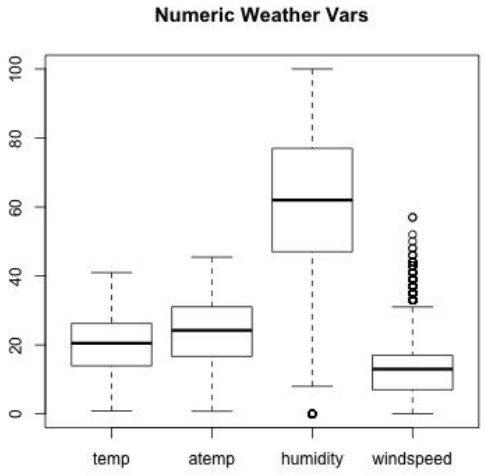
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

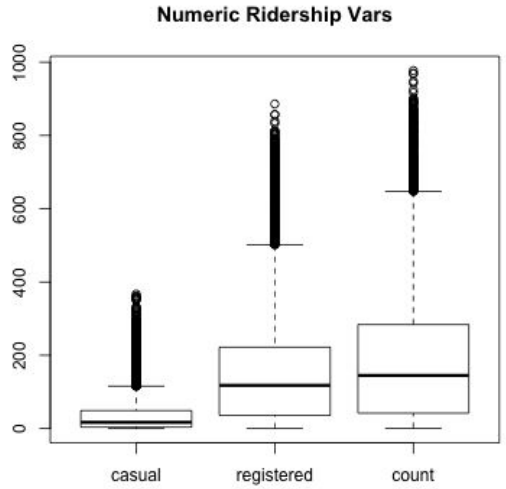
- TO WIN
- Goal: To predict the total number of rides on a future day based on the temp, windspeed, weather, date, type of day (eg: weekday, holiday), humidity
- Room Mean Squared Log Error (RMSLE) is the evaluation metric

Exploratory Data Analysis

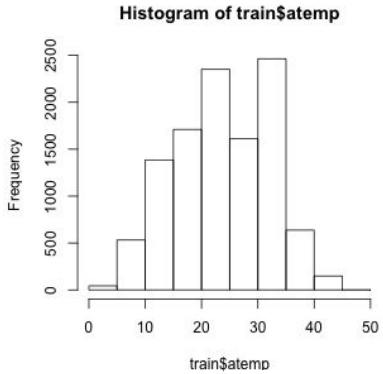
Exploratory Data Analysis

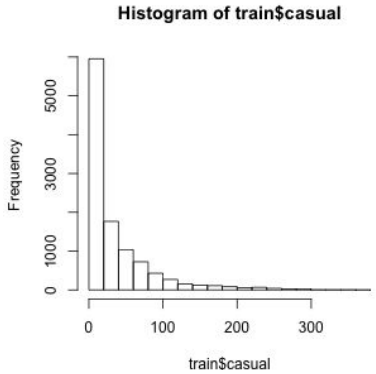


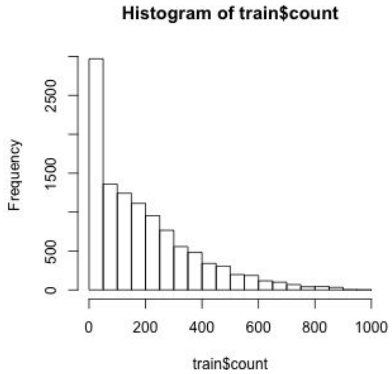


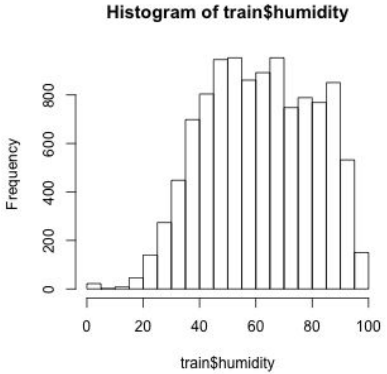


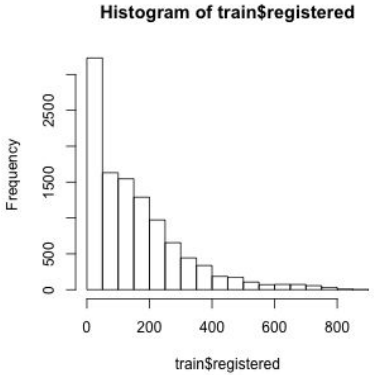
Exploratory Data Analysis

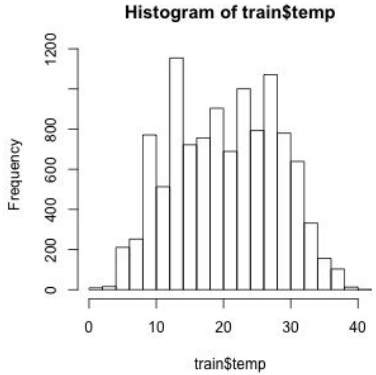


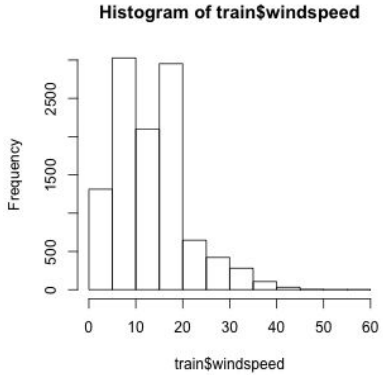












Other Solutions

Kaggle Kernels

Kernel	Features	Modeling Approach	Performance (RSMLE)
EDA & Ensemble Model (Top 10 Percentile)	-New cols from Datetime -Coercing of columns to “categorical” data type	Linear Regression Ridge Regression Lasso Regression Random Forest Gradient Boost	0.977996 0.977996 0.978133 0.102804 0.189973
Comprehensive EDA with XGBoost (Top 10 percentile)	-Transforming count to log(count) -Creation of dummy vars for categorical data	XGBoost Random Forest	Did not put in notebook :-(
bikes	- Creation of dayofweek, hour, month and year columns from the datetime column	Random Forest	0.106703

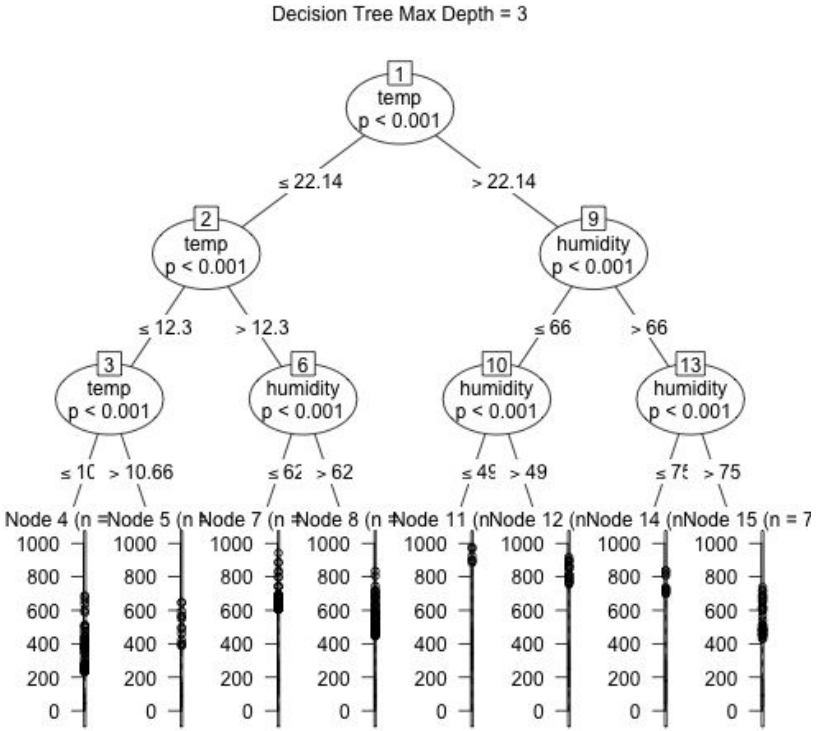
Modeling

Modeling: Linear Regression

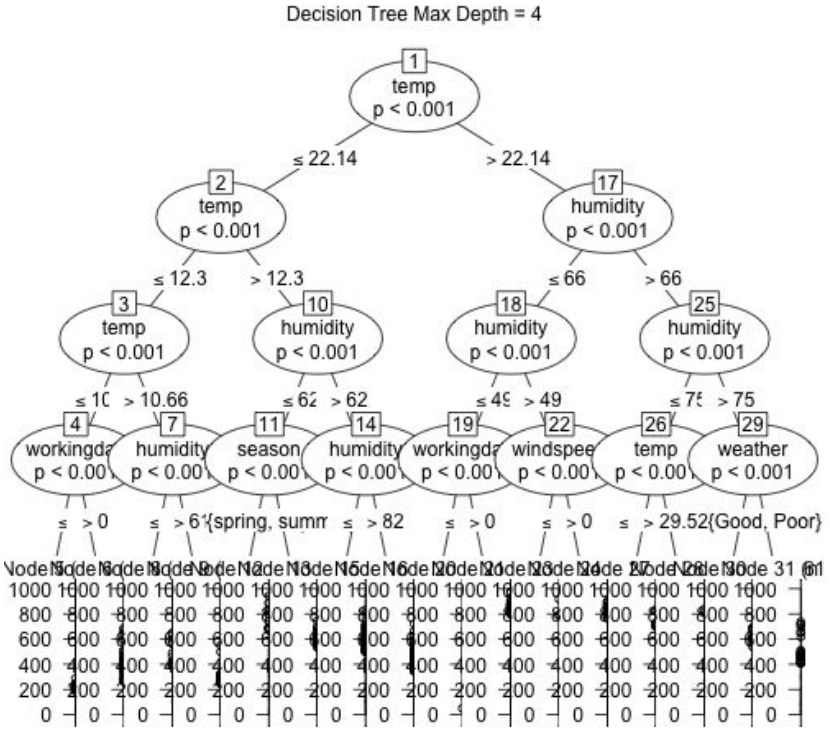
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    128.1877    10.8558  11.808  < 2e-16 ***
temp             9.4333     1.4031   6.723  1.9e-11 ***
atemp            1.2011     1.2275   0.979 0.327853
humidity        -2.8475     0.1117 -25.484  < 2e-16 ***
windspeed        0.5734     0.2379   2.410 0.015981 *
seasonsummer     5.1365     6.4424   0.797 0.425307
seasonfall     -30.0799     8.2776  -3.634 0.000281 ***
seasonwinter    68.2408     5.4584  12.502  < 2e-16 ***
holidayTRUE     -6.1375    11.1109  -0.552 0.580702
workingdayTRUE  -0.8556     3.9533  -0.216 0.828661
weatherFair     15.6759     4.3059   3.641 0.000274 ***
weatherPoor     -9.9765     7.3427  -1.359 0.174287
weatherBad     187.1072   155.3143   1.205 0.228357
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- “Default” Season was spring
- “Default” weather was good
- Error = 1.383519

Modeling Decision Tree



Error = 1.358835



Error = 1.335442

Modeling Random Forest

No pretty picture

But the best error of 1.287416!

Questions

Thank you for your time!

Sources

- https://www.flickr.com/photos/volvob12b/21562863658
- https://www.flickr.com/photos/taedc/14248535232
- https://commons.wikimedia.org/wiki/File:Kaggle_logo.png
- https://www.kaggle.com/c/bike-sharing-demand
-

TOO FAR

Thank you for your time!

EDA & Ensemble Model (Top 10 Percentile)

- https://www.kaggle.com/viveksrinivasan/eda-ensemble-model-top-10-percentile.
- By Vivek Srinivasan
-

Comprehensive EDA with XGBoost (Top 10 percentile)

- https://www.kaggle.com/miteshyadav/comprehensive-eda-with-xgboost-top-10-percentile
- By Mitesh Yadav

bikes

- https://www.kaggle.com/meenaj/bikes
- By meena
-