Bike Rental Demands

By
Emmanuel Cocom
Kristen Marenco
Imelda Flores

Understanding The Problem

Background Info:

Bike Rentals in Washington, D.C.

Our Challenge:

Predict the demand of bicycle rentals.

The Data

1 Label:

- Count

11 Features

- Datetime
- Holiday
- Weather
- Season
- Holiday
- Humidity
- Temperature
- Atemp
- Working Day
- Windspeed
- Casual
- Registered

Leakage Variables – Bad Data

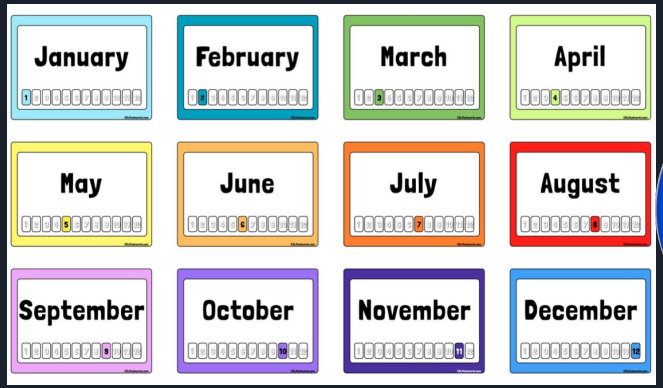
- Casual + Registered == Count (our label!)

Team Decision: NOT Features, MUST be DROPPED

DATA EXTRACTED - From Datetime Feature

- Hour
- Months
- Day of the Year
- Weekday
- Year

DAY	CODE
Monday	1
Tuesday	2
Wednesday	3
Thursday	4
Friday	5
Saturday	6
Sunday	0



Categorical Features - One Hot Encoded

Weather -> Clear, Cloudy, Raining, Ice

Season -> Fall, Spring, Summer, Winter

Hour -> 4 Bins

Bin 1: 1 - 6 AM

Bin 2: 7AM - 12 PM

Bin 3: 1 - 6 PM

Bin 4: 7 PM - 12 AM

	spring,	summer	fall	winter
0	1.0	0.0	0.0	0.0
50	1.0	0.0	0.0	0.0
100	1.0	0.0	0.0	0.0
150	1.0	0.0	0.0	0.0
200	1.0	0.0	0.0	0.0
250	1.0	0.0	0.0	0.0

Normalizing Data

SKLEARN Preprocessing

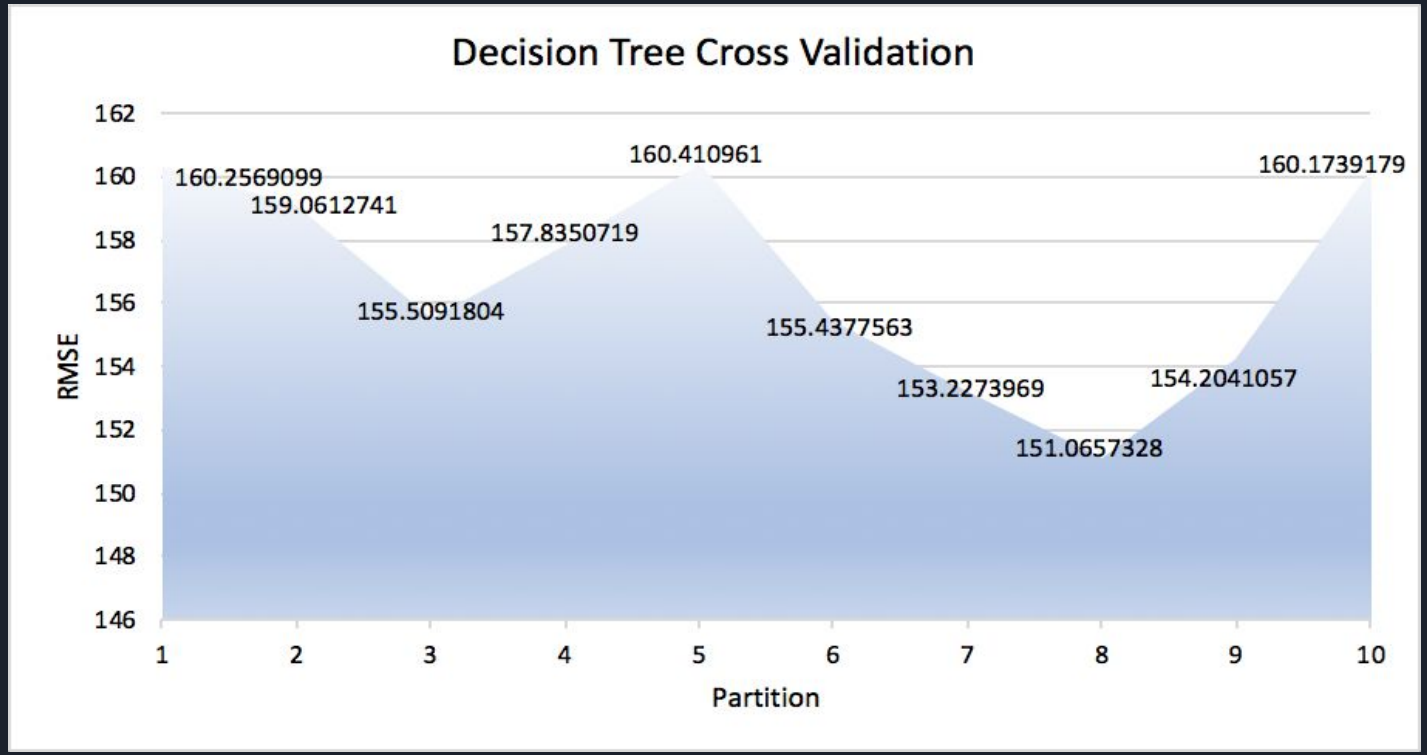
```
from sklearn import preprocessing

#normalize data
for x in range(len(df_splits)):
    #scale it -> d type changes to numpy array
    scaled_feature_matrix_month_numpyarray = preprocessing.scale(df_splits[x][0])

    #change back to df
    df_month_scaled = pd.DataFrame(scaled_feature_matrix_month_numpyarray, columns = df_splits[x

    #store back the scaled data back into list.
    df_splits[x][0] = df_month_scaled
```

Algorithms & Methods

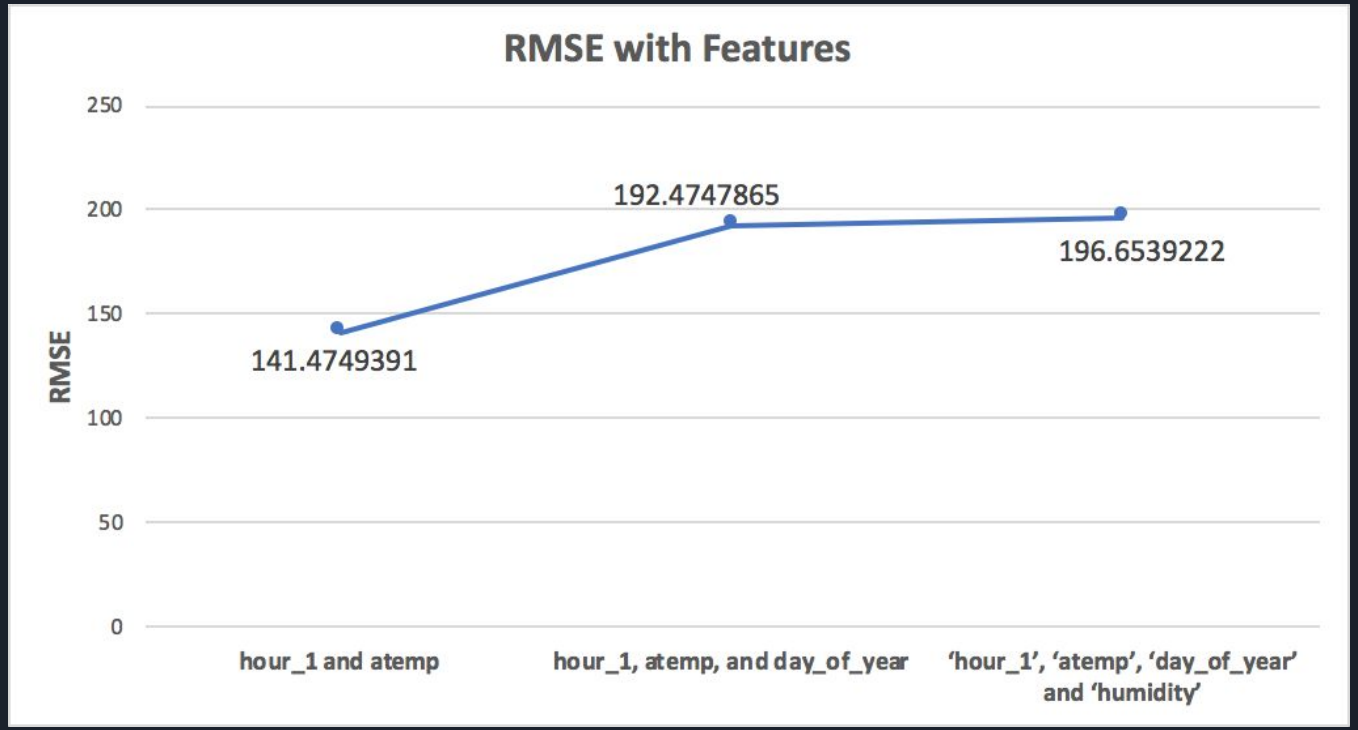
- KNN Regression
- Decision Tree Regression
- Random Forest Regression
- XGBoost Regression
- ADA Boost Regression

- Tests corrupt/missing data
- Normalization
- Cross Validation
- ADA Boost
- Dimensional Reductionality (PCA)

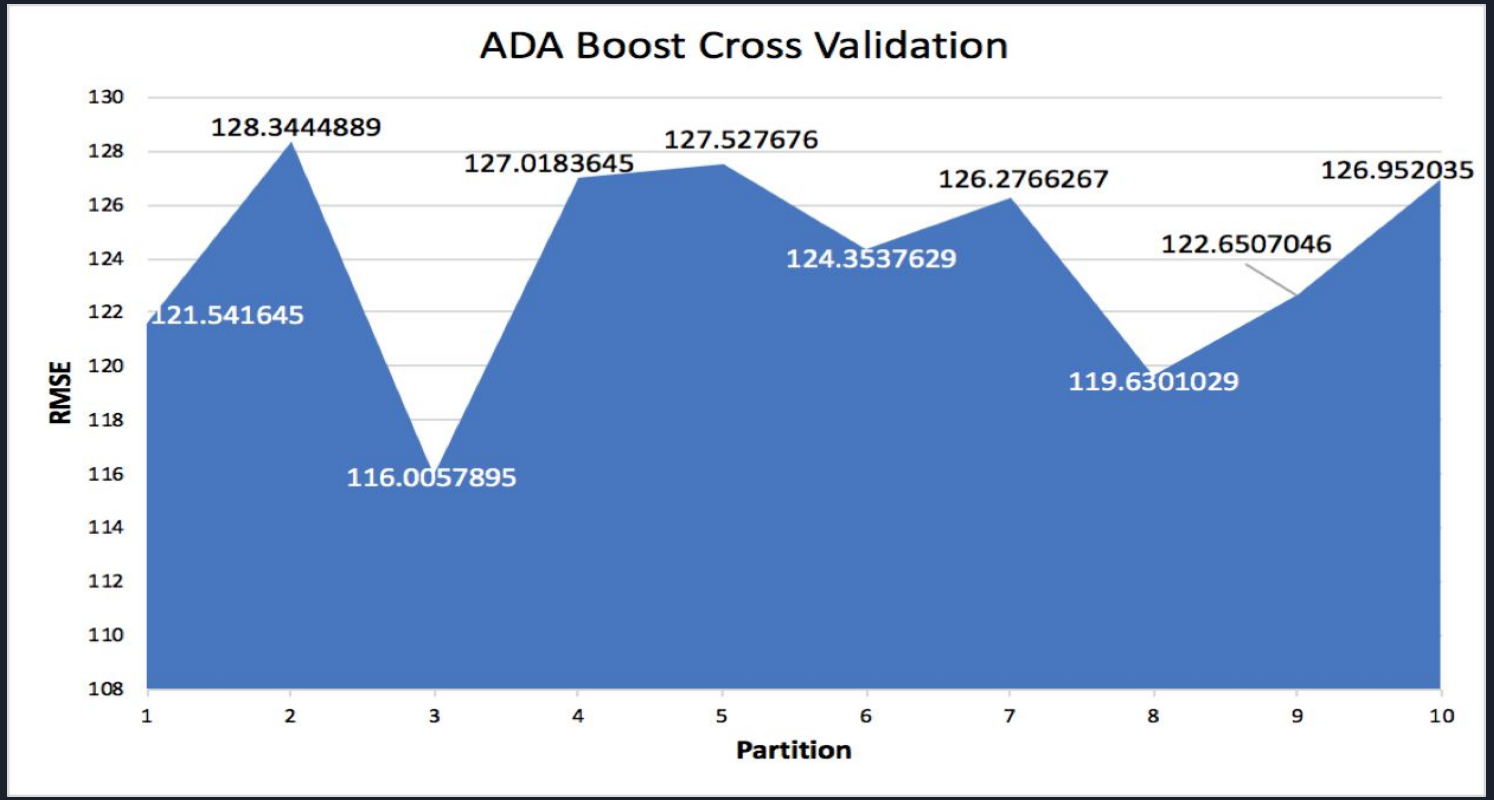
Decision Tree Regression - Cross Validation



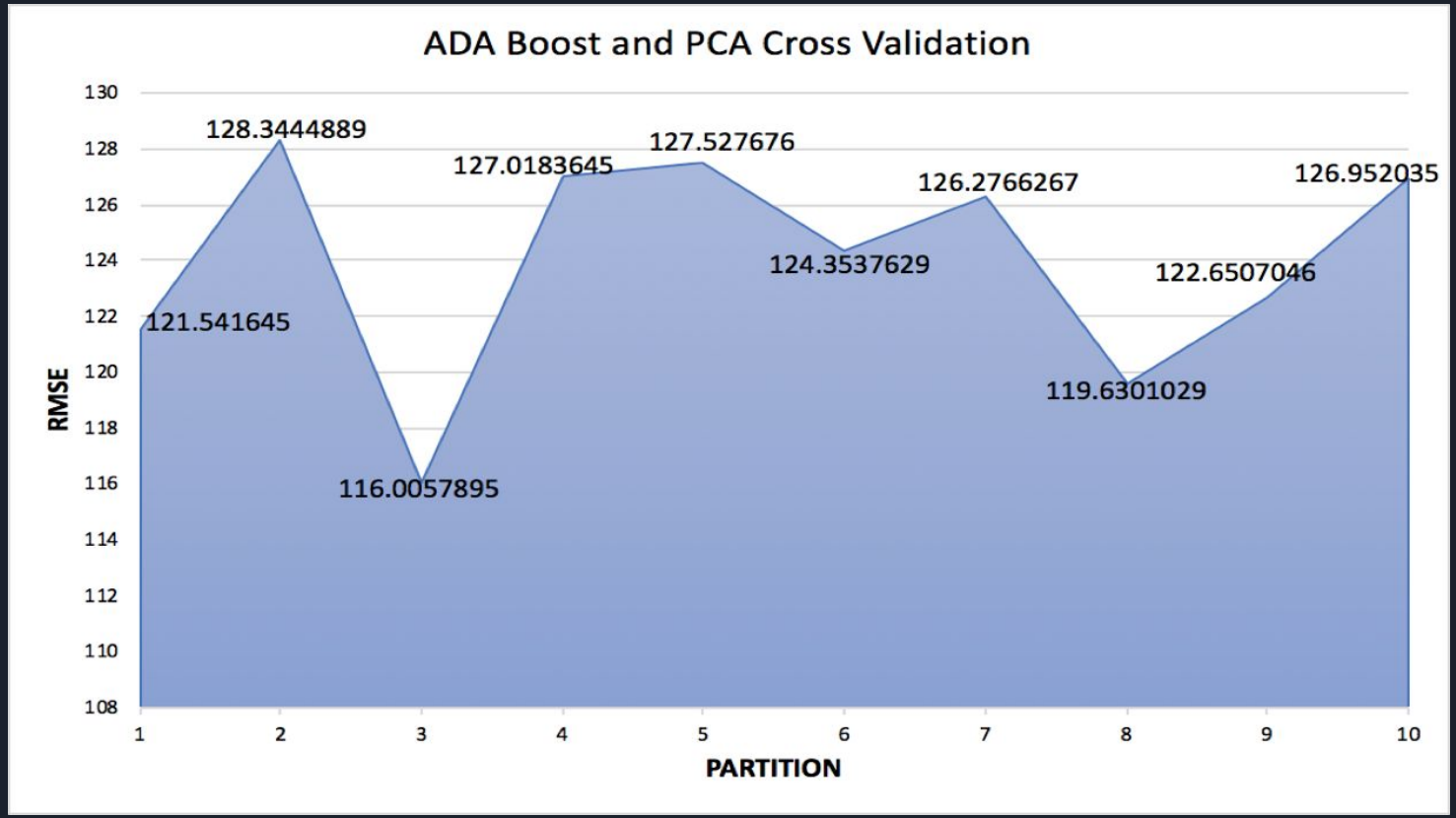
Decision Tree - Feature Reduction - 3 best features



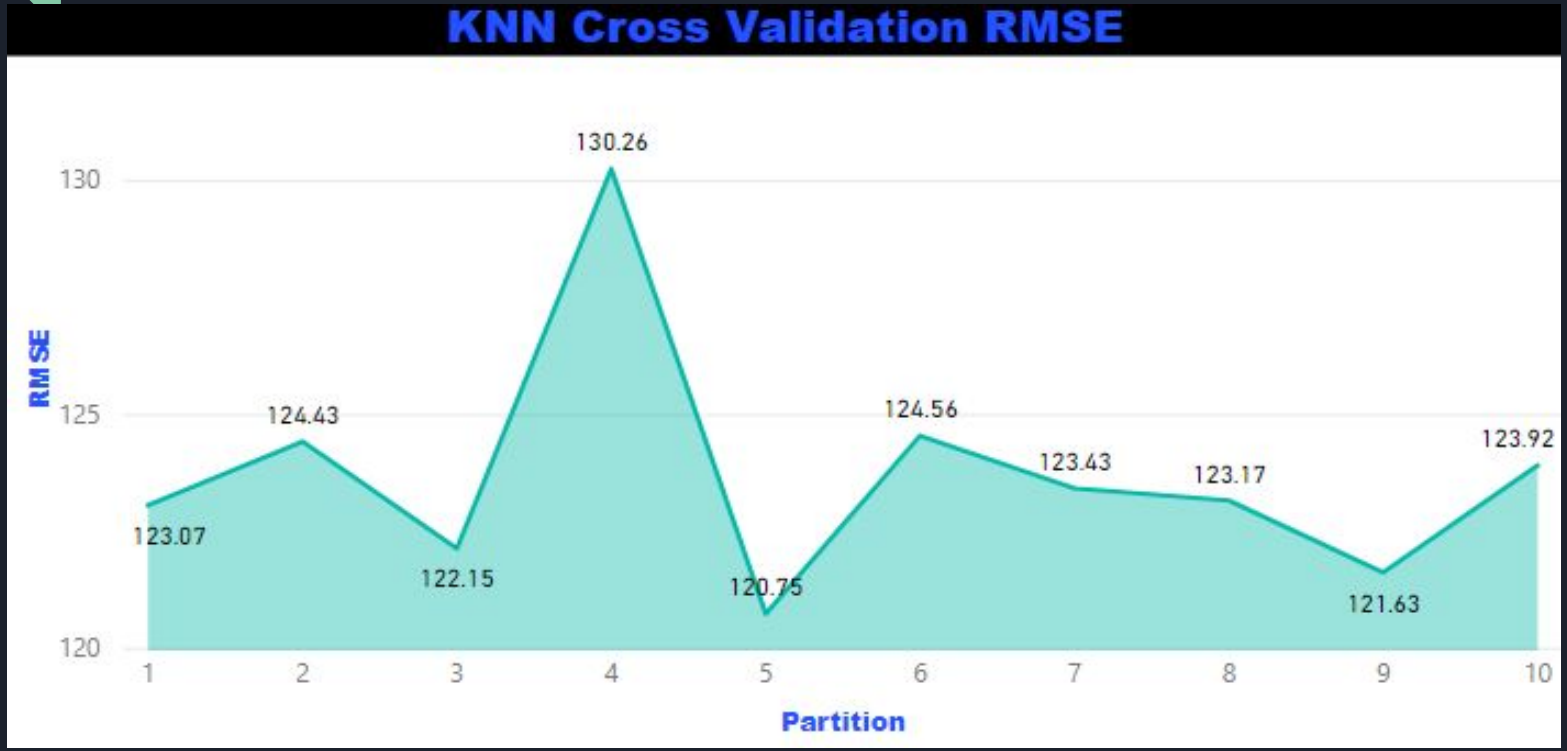
Decision Tree - Ada Boost - Cross Validation



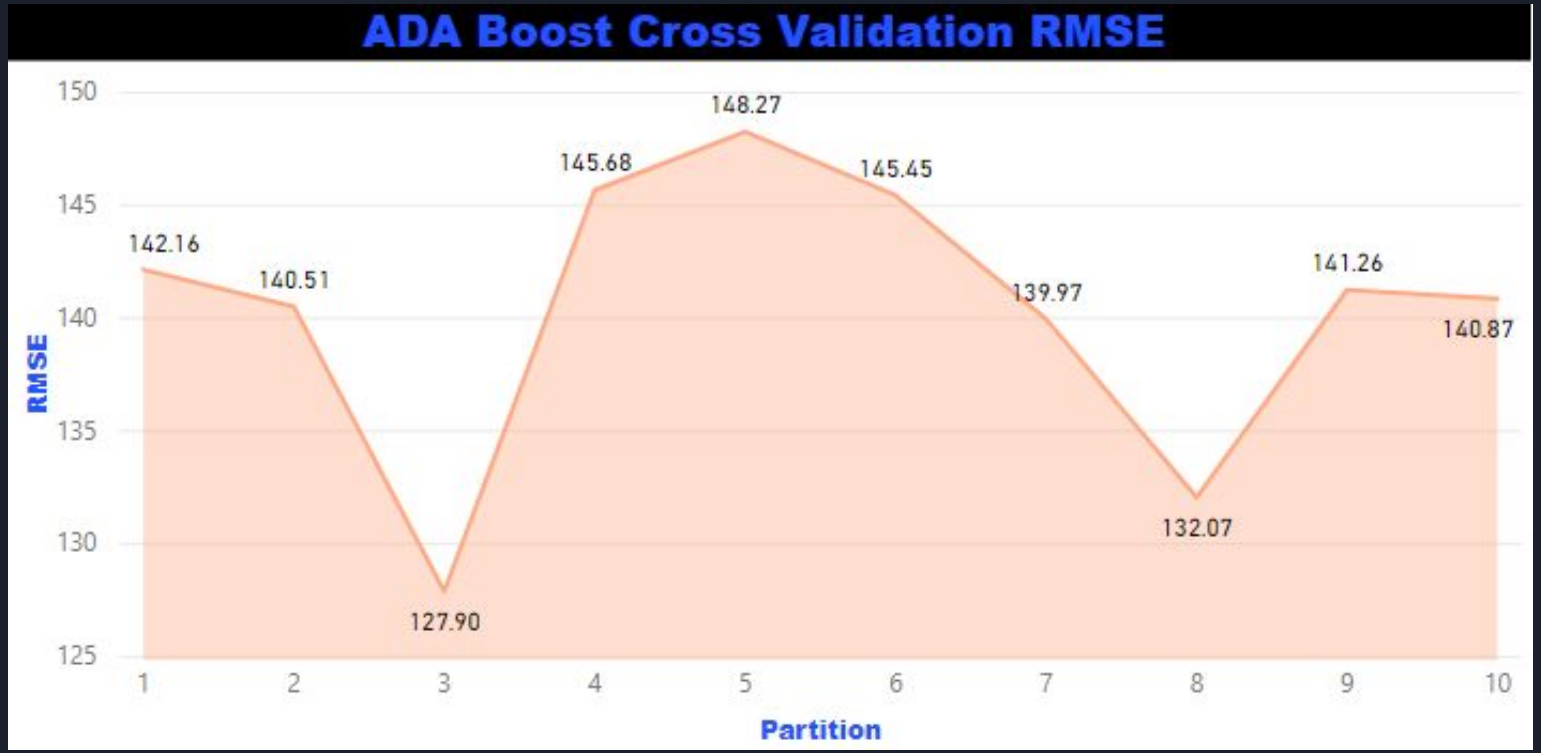
ADA Boost - PCA - Cross Validation



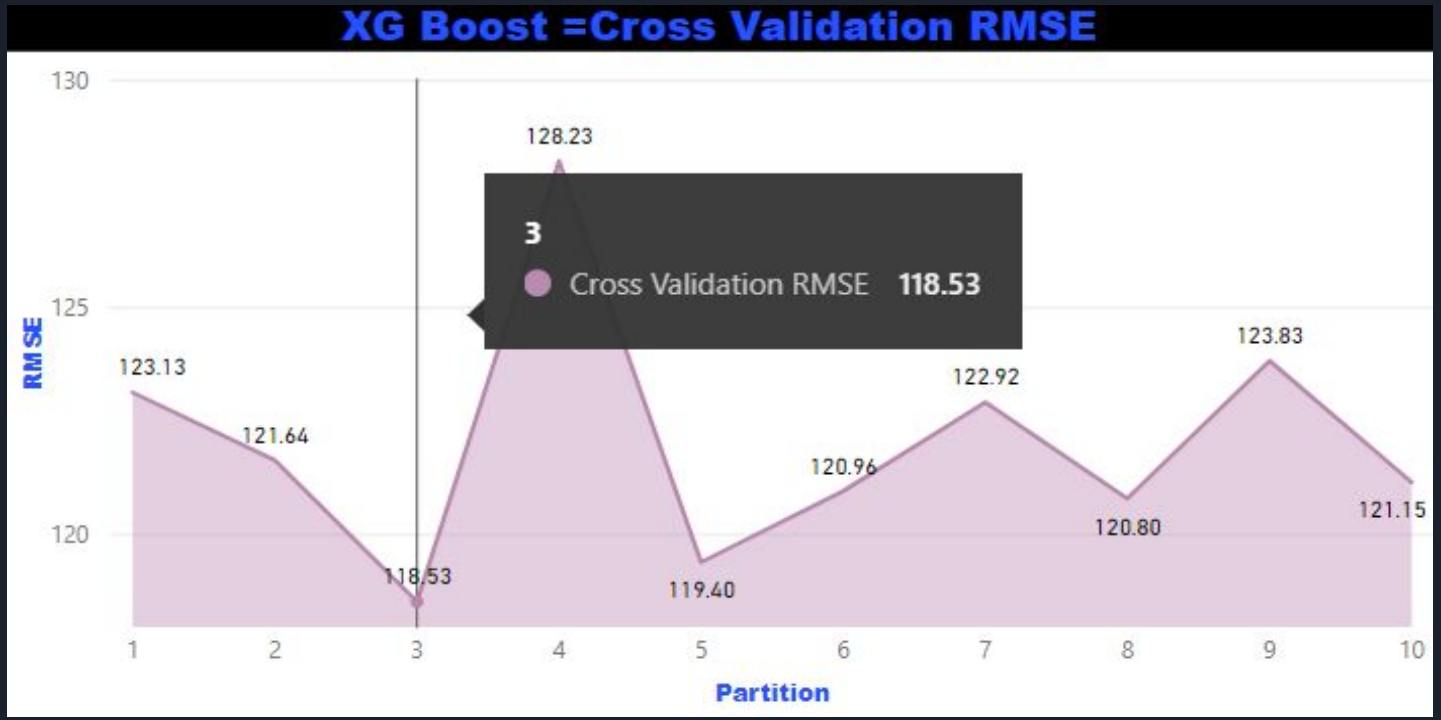
KNN Regression - Cross Validation



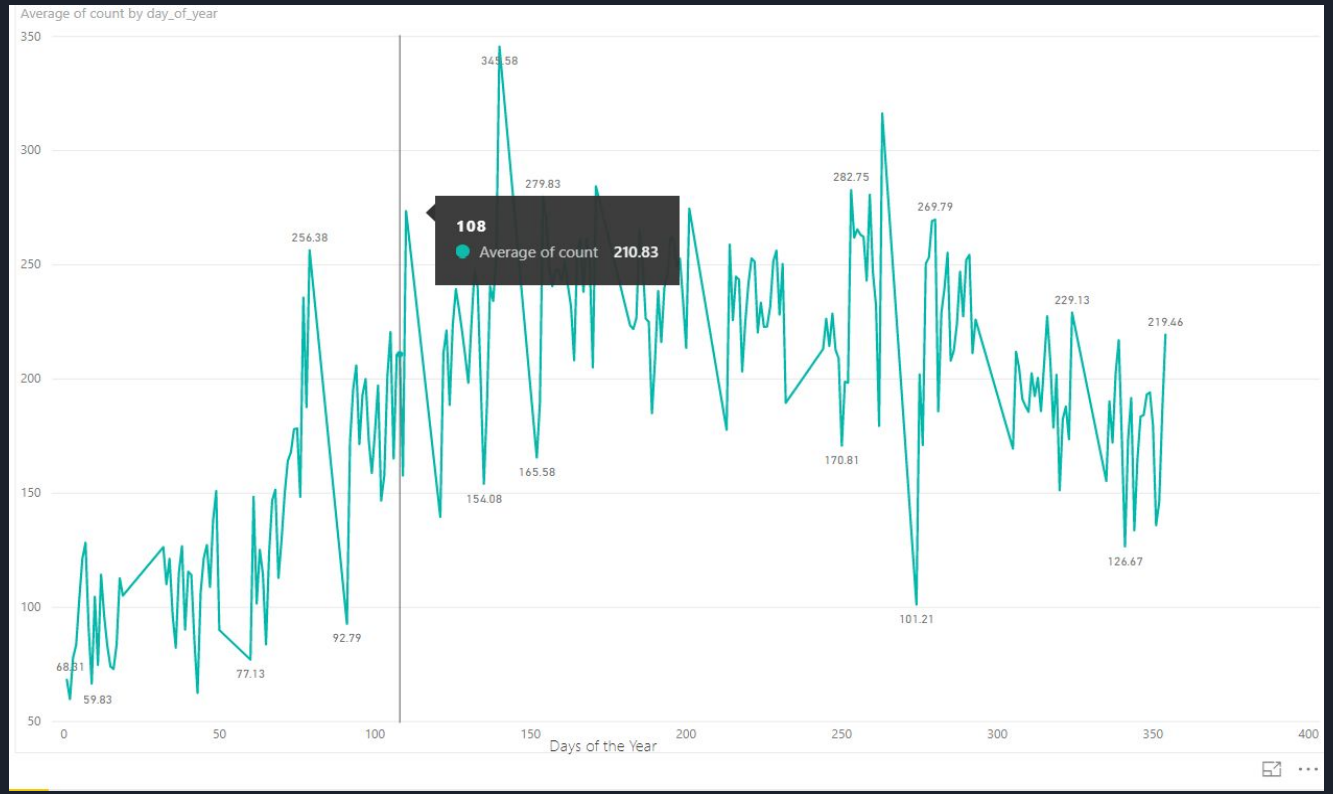
KNN Regression - ADA Boost



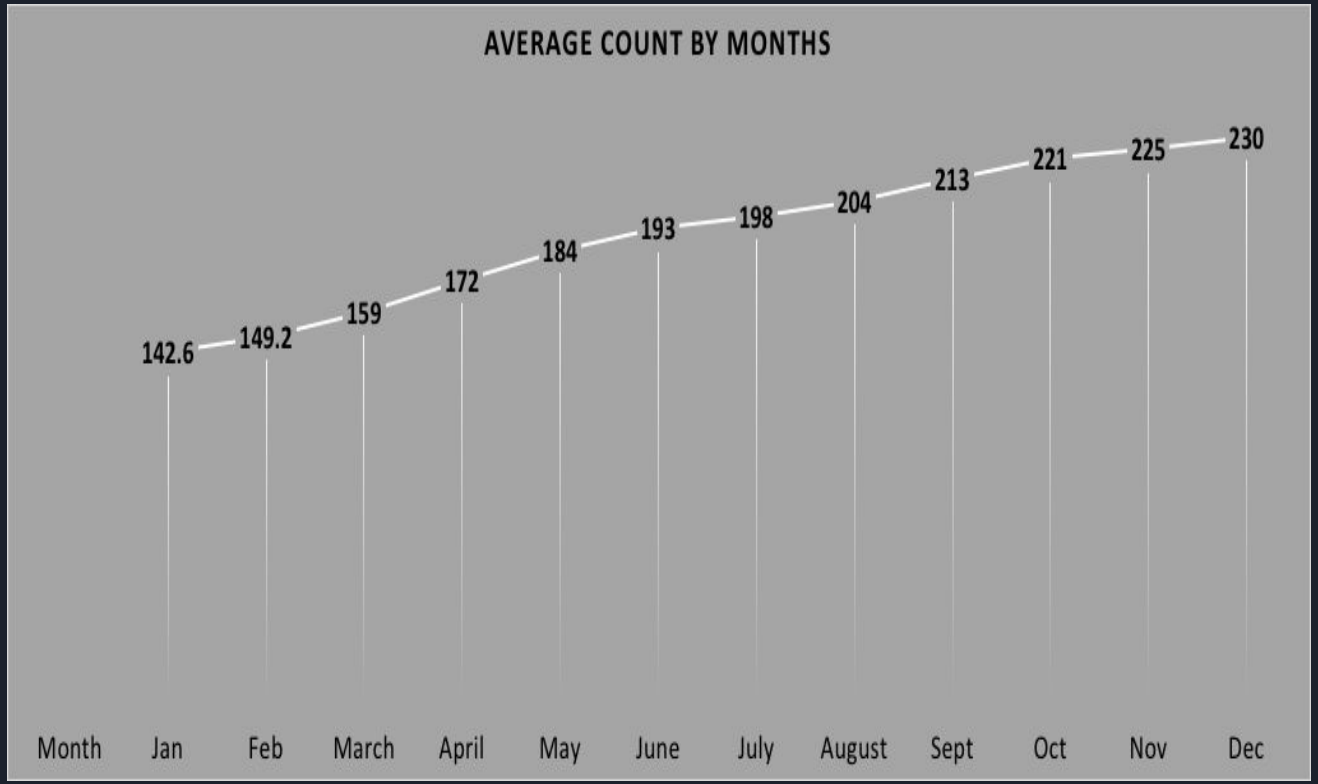
XGBoost Regression - Cross Validation



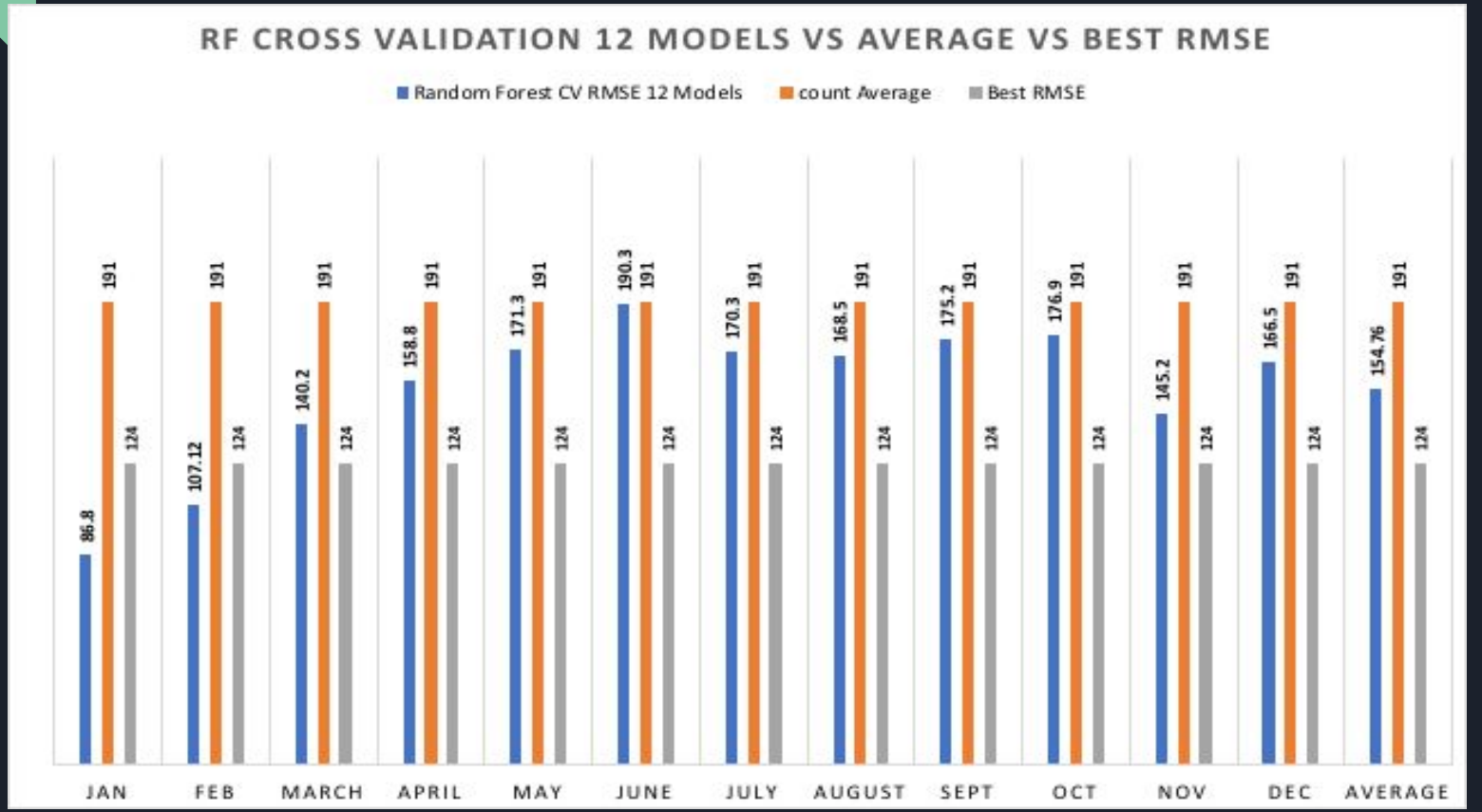
Now with 12 Models! But why?



Count Average by Month

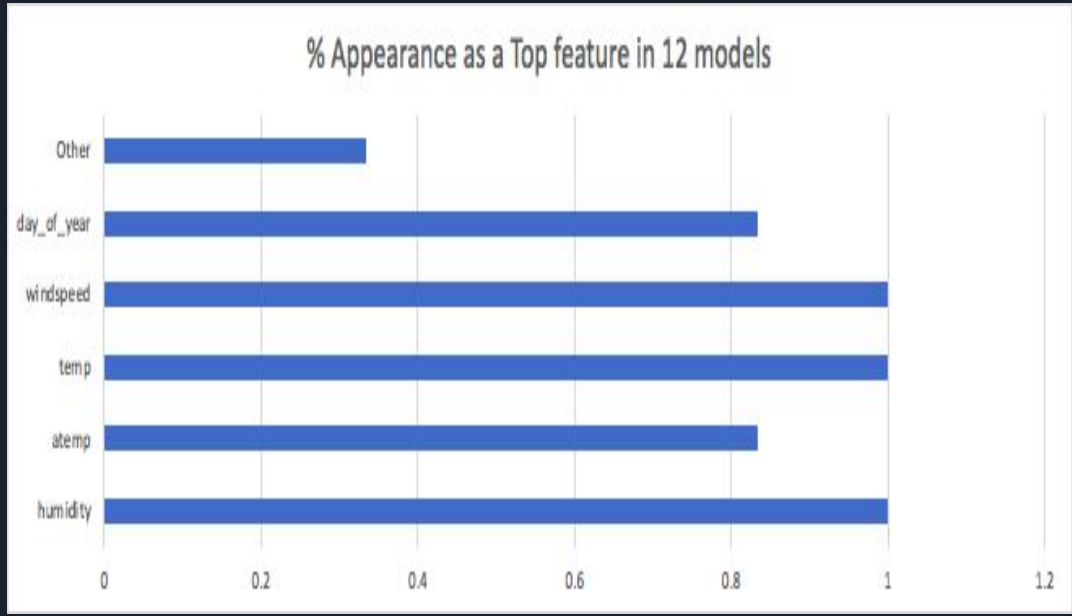


Random Forest Regression (RF) - CV

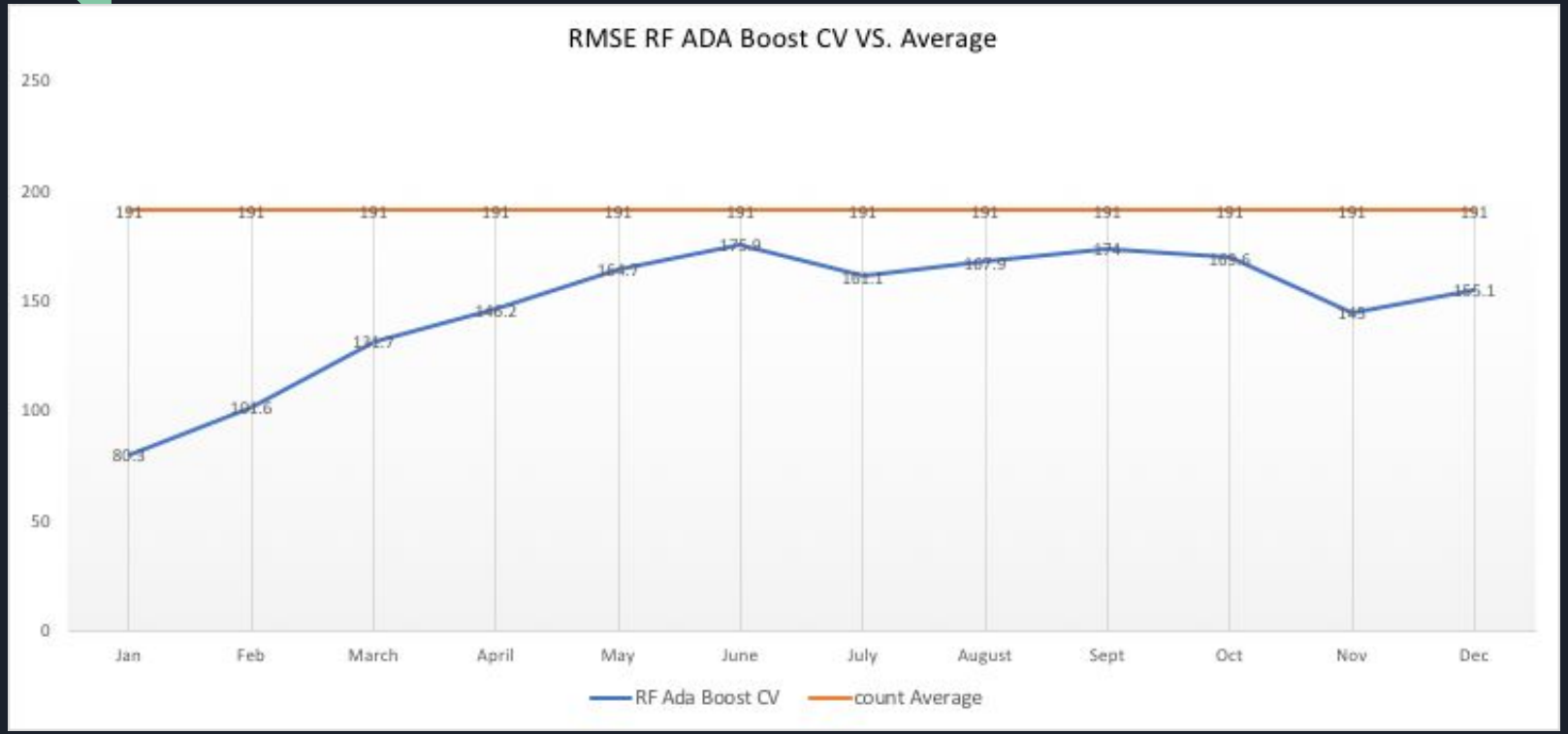


Different RMSE but same Features contributing!

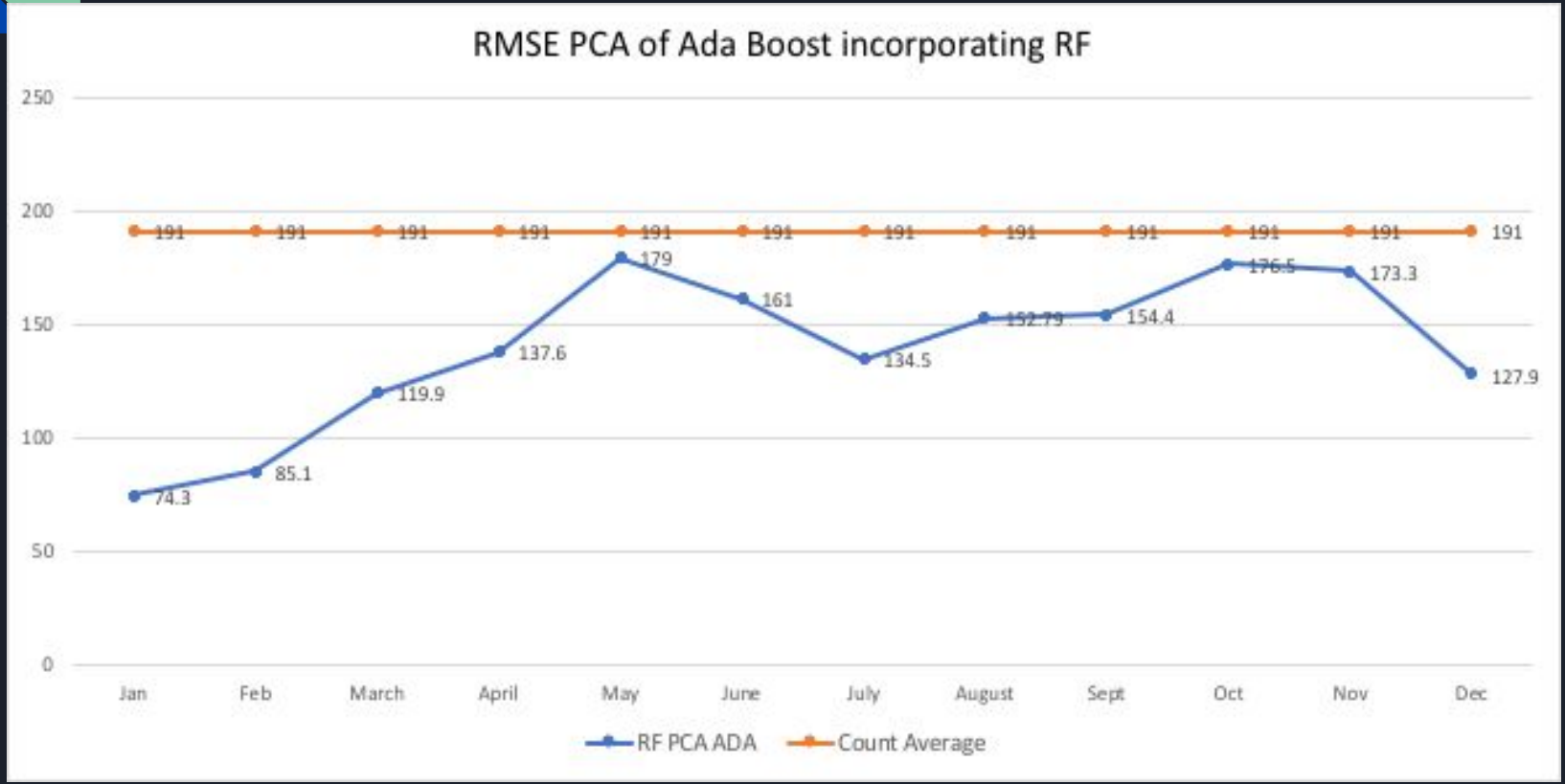
Month	#1 Feature	#2 Feature	#3 Feature	#4 Feature	#5 Feature
Jan	humidity	atemp	temp	windspeed	year
Feb	humidity	temp	windspeed	atemp	day_of_year
March	temp	atemp	humidity	windspeed	year
April	humidity	windspeed	temp	atemp	day_of_year
May	atempt	humidity	windspeed	day_of_year	temp
June	humidity	windspeed	year	day_of_year	temp
July	temp	humidity	day_of_year	windspeed	atemp
August	temp	humidity	windspeed	day_of_year	year
Sept	humidity	atemp	windspeed	temp	day_of_year
Oct	humidity	windspeed	atemp	temp	day_of_year
Nov	humidity	temp	windspeed	day_of_year	temp
Dec	humidity	temp	windspeed	day_of_year	atemp



RF - ADA Boost


RMSE RF ADA Boost CV VS. Average
250
200
150
100
50
0
191
80.3
101.6
131.7
146.2
164.7
175.9
161.1
167.9
174
165.6
145
155.1
Jan
Feb
March
April
May
June
July
August
Sept
Oct
Nov
Dec
RF Ada Boost CV
count Average

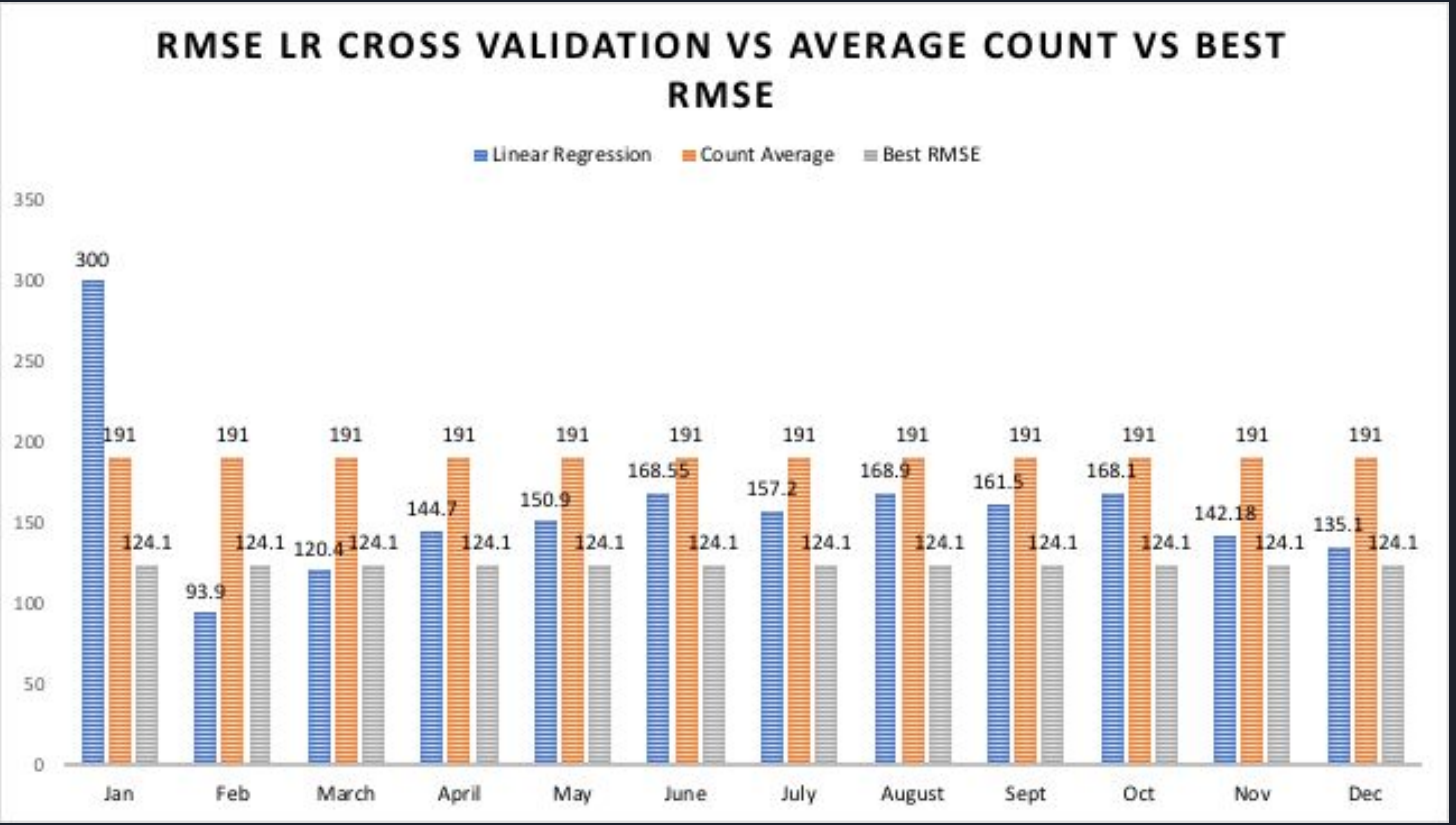
RF - PCA


RMSE PCA of Ada Boost incorporating RF
250
200
150
100
50
0
191
74.3
85.1
119.9
137.6
179
161
134.5
152.79
154.4
176.5
173.3
127.9
Jan
Feb
March
April
May
June
July
August
Sept
Oct
Nov
Dec
RF PCA ADA
Count Average

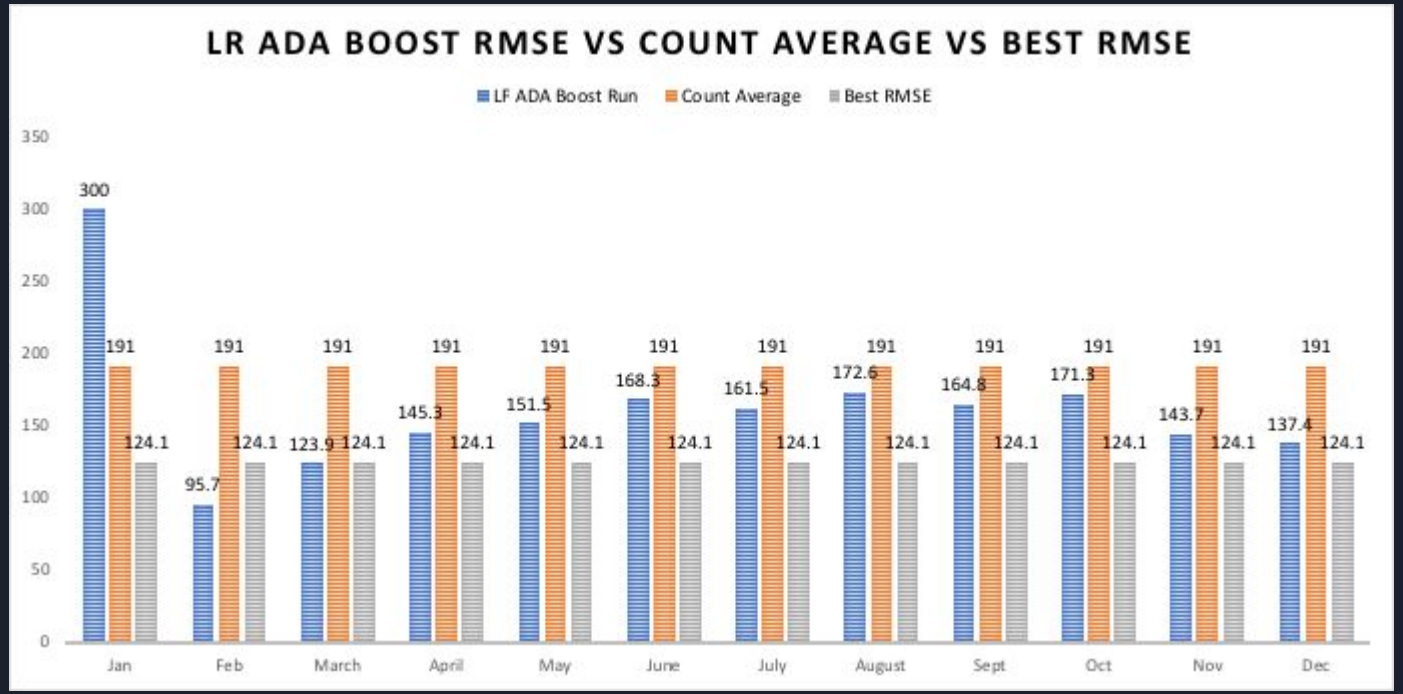
RF - Let's look at the numbers!

Months	RF Individual Run	RF RMSE CV Run	RF Feature Reduction	RF ADA CV2	RF PCA ADA	COUNT AVERAGE
Jan	74.8	86.8	91.5	80.3	74.3	191
Feb	85.4	107.12	107.2	101.6	85.1	191
March	120	140.2	137.8	131.7	119.9	191
April	138.1	158.8	154.5	146.2	137.6	191
May	132.4	171.3	179	164.7	179	191
June	162.2	190.3	188.5	175.9	161	191
July	135.2	170.3	178.5	161.1	134.5	191
August	152.5	168.5	172.8	167.9	152.79	191
Sept	154.4	175.2	182.9	174	154.4	191
Oct	172.4	176.9	176.5	169.6	176.5	191
Nov	119.6	145.2	159	145	173.3	191
Dec	129.3	166.5	166.6	155.1	127.9	191
Average	131.3583333	154.76	157.9	147.7583333	139.6908333	191

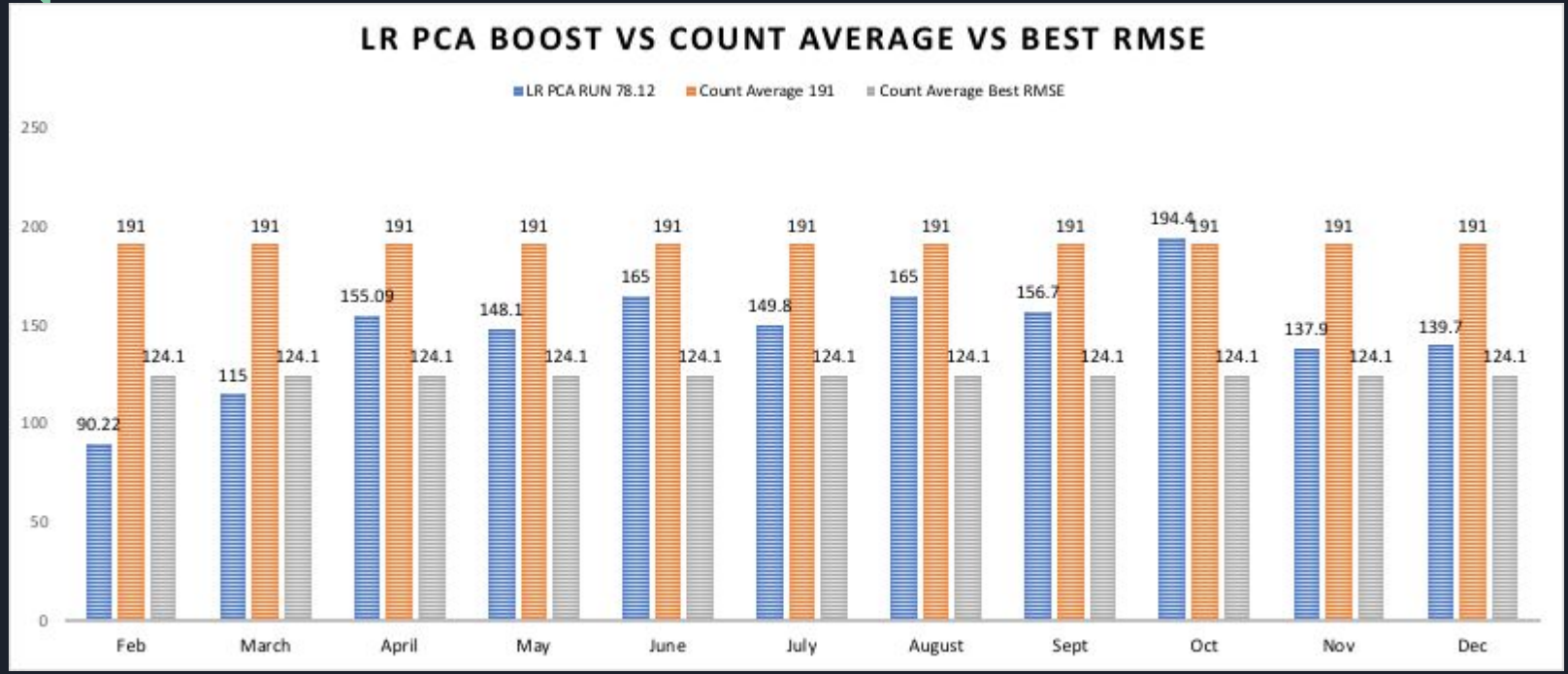
Linear Regression (LR)



LR - ADA Boost


LR ADA BOOST RMSE VS COUNT AVERAGE VS BEST RMSE
LF ADA Boost Run
Count Average
Best RMSE
350
300
250
200
150
100
50
0
Jan
Feb
March
April
May
June
July
August
Sept
Oct
Nov
Dec
300
191
124.1
95.7
191
124.1
123.9
191
124.1
145.3
191
124.1
151.5
191
124.1
168.3
191
124.1
161.5
191
124.1
172.6
191
124.1
164.8
191
124.1
171.3
191
124.1
143.7
191
124.1
137.4
191
124.1

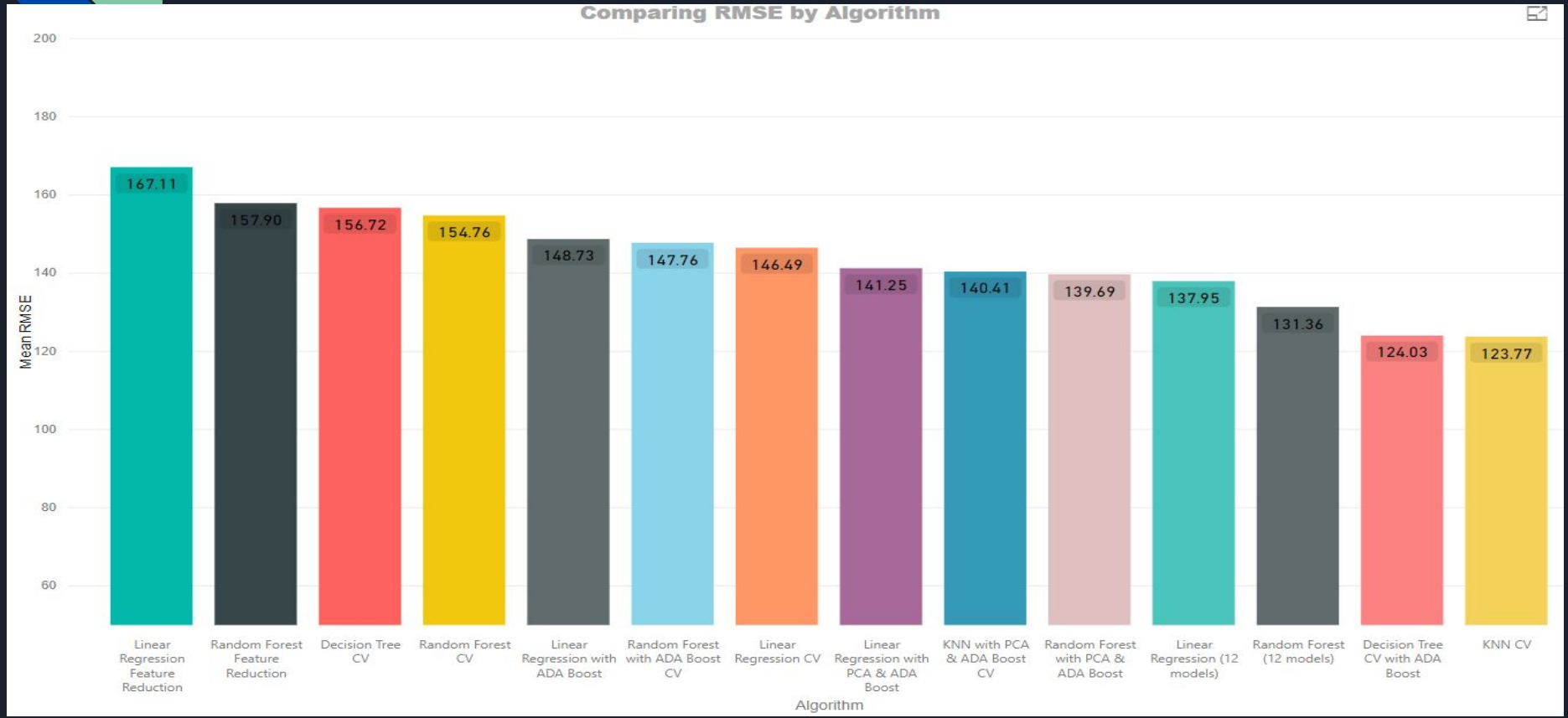
LR - PCA


LR PCA BOOST VS COUNT AVERAGE VS BEST RMSE
LR PCA RUN 78.12
Count Average 191
Count Average Best RMSE
250
200
150
100
50
0
Feb
March
April
May
June
July
August
Sept
Oct
Nov
Dec
90.22
115
155.09
148.1
165
149.8
165
156.7
194.4
137.9
139.7
191
124.1

LR - Let's Look at the numbers

Linear Regression	Linear Regression RMSE	Linear Regression CV RMSE	LF Feature Reduction	LF Ada boost RMSE	LF PCA RMSE	Count Average	BEST RMSE
Jan	78.1	1.31822E+12	90.5	6.88051E+13	78.12	191	141
Feb	90.2	93.9	106.9	95.7	90.22	191	141
March	115.1	120.4	148.5	123.9	115	191	141
April	115	144.7	173.7	145.3	155.09	191	141
May	148.1	150.9	185.5	151.5	148.1	191	141
June	165.338	168.55	197.6	168.3	165	191	141
July	149.82	157.2	183	161.5	149.8	191	141
August	165	168.9	195.7	172.6	165	191	141
Sept	156.7	161.5	204.5	164.8	156.7	191	141
Oct	194.4	168.1	200.6	171.3	194.4	191	141
Nov	137.9	142.18	164.2	143.7	137.9	191	141
Dec	139.7	135.1	154.6	137.4	139.7	191	141
Average	137.9465	1.09852E+11	167.1083333	5.73376E+12	141.2525	191	141

RMSE of All Algorithms



12 Models VS 1 Model? Clear Winner?

- Random Forest Regression and Linear Regression:

  12 models works for January and February

- KNN, Decision Tree, XGBoost:

1 model works better for rest of the year

Best Approach

Better Results:

Linear and Random Forest - PCA

Decision Tree - ADA Boost

KNN - Cross Validation

XGBoost - Cross Validation

Overall:

Jan & Feb - Random Forest & PCA

March - Linear Regression & PCA

The Rest: XGB Cross Validation and Decision Tree (ADA Boost and PCA)