The Analysis of Absenteeism At Work

Amy Ding

The Goal:

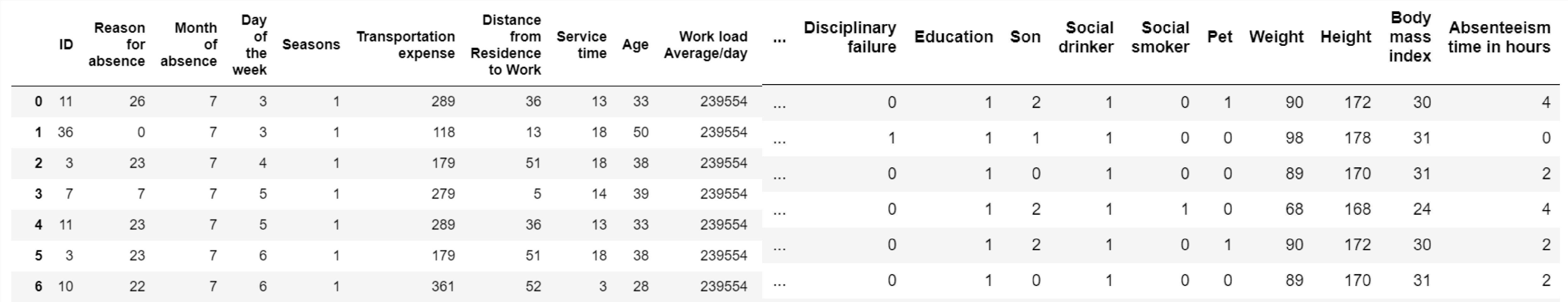
To predict the total hours people are absent at work by applying regression models to better understand underlying variables.

Background:

- The data set is from UCI (UC Irvine Machine Learning Repository)
- The data set was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.
- The data set (Absenteeism at work - Part I) was used in academic research at the Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management and a paper (Application of a neuro fuzzy network in prediction of absenteeism at work) was published based on the study of this data set.

Details about the data set:

- **Number of Instances:** 740
- **Number of Attributes:** 21
- **Attribute Characteristics:** Integer



	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
0	11	26	7	3	1	289	36	13	33	239554	...	0	1	2	1	0	1	90	172	30	4
1	36	0	7	3	1	118	13	18	50	239554	...	1	1	1	1	0	0	98	178	31	0
2	3	23	7	4	1	179	51	18	38	239554	...	0	1	0	1	0	0	89	170	31	2
3	7	7	7	5	1	279	5	14	39	239554	...	0	1	2	1	1	0	68	168	24	4
4	11	23	7	5	1	289	36	13	33	239554	...	0	1	2	1	0	1	90	172	30	2
5	3	23	7	6	1	179	51	18	38	239554	...	0	1	0	1	0	0	89	170	31	2
6	10	22	7	6	1	361	52	3	28	239554	...	0	1	0	1	0	0	89	170	31	2

Cleaning the data set:

- **In general, this data set is clean:**
- All values are integer
- No missing data
- **Work needed:**
- Rename some of the columns (make sure all the columns' names are one word).
- Drop some rows (Reason ==0 & Absent_time ==0).
- Drop some columns that are less interesting (I'm less interested in) or highly correlated.

Cleaning the data set (continue):

14 attributes left after cleaning, they are:

ID, Reason, Season, Trans_expense, Distance, Service_time, Age, Work_load, Son (number of children), Social_drinker, Social_smoker, Body_mass, ***Absent_time***.

Note:

ID, Reason and Season: Although they have numeric values, they are categorical attributes.

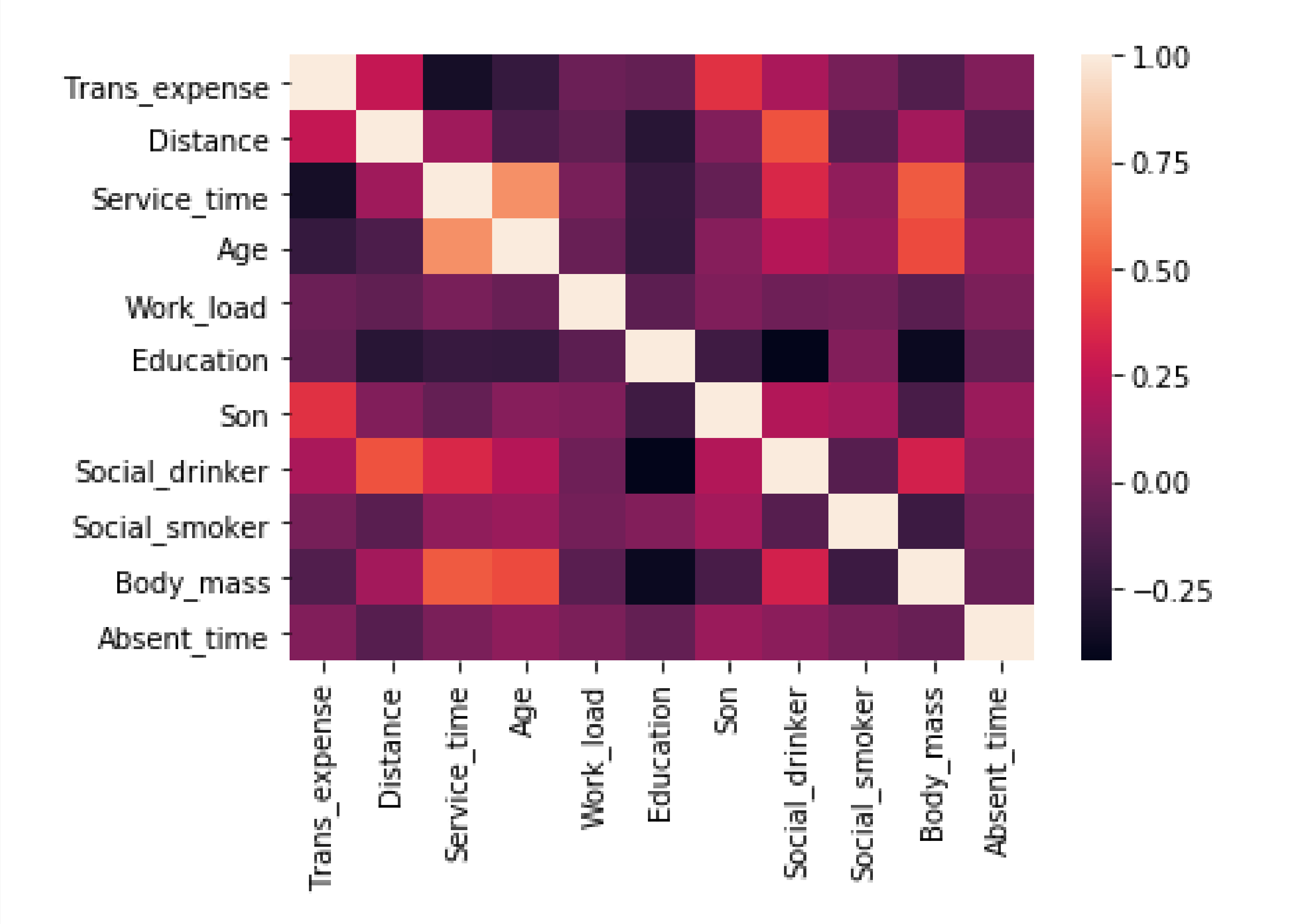
Reason: 28 different categories and are divided into two big categories (with International Code of Diseases (ICD) or without ICD). For example, blood donation, laboratory examination belong to category II and other common diseases fall into category I.

Analyzing the data:

All the numeric attributes have very limited correlation with Absent_time (both positive and negative).

- **Son (corr: 0.128)**
- **Distance (corr: -0988)**
- **Age (corr: 0.086)**

Linear regression may not be the best choice but let's still try!



Modeling – Linear Regression

Null RMSE (the baseline): 8.9

Model 1: Includes the three features with the largest corr with Absent_time (Son, Distance, Age)

RMSE: 9.0

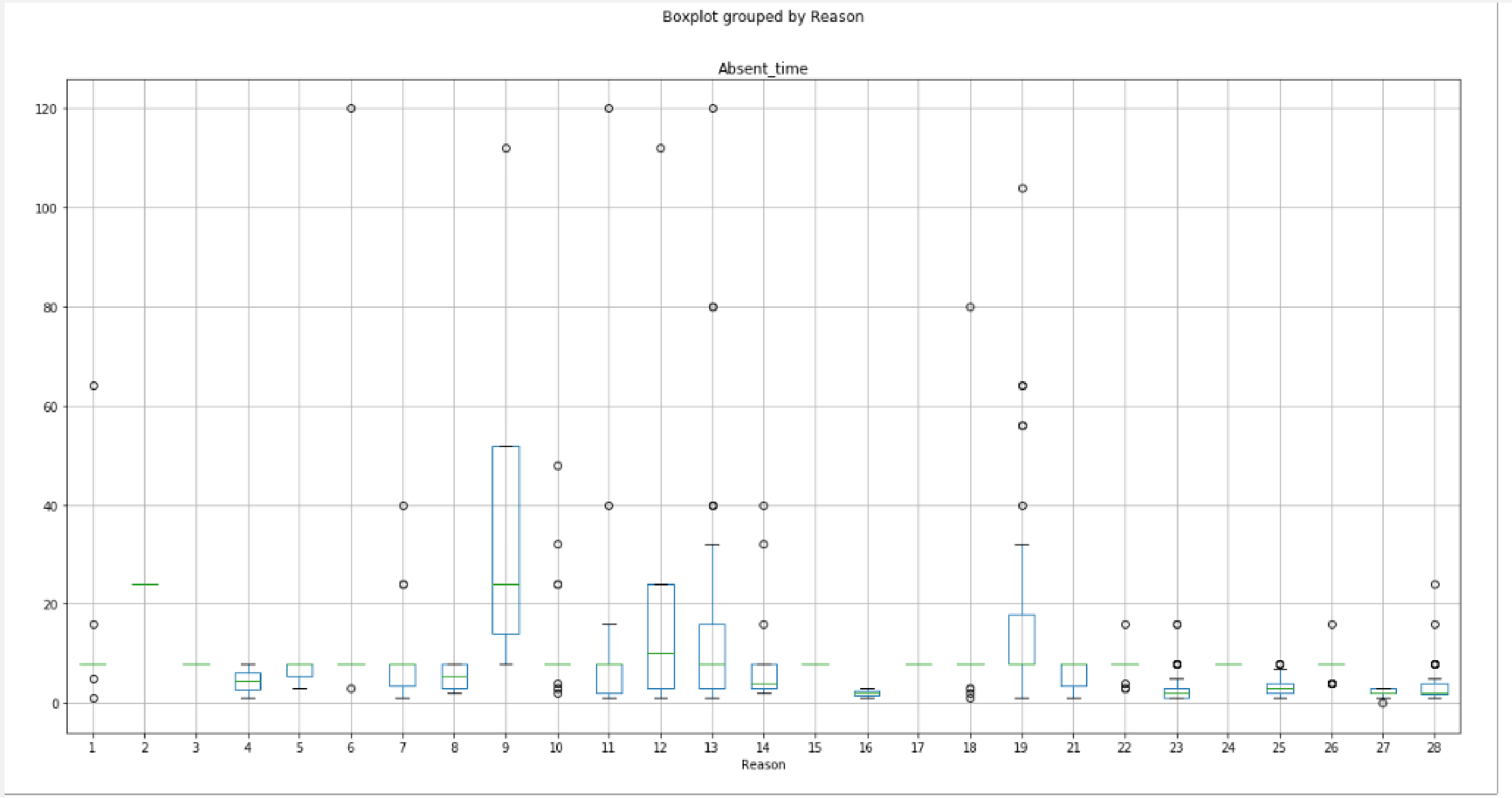
Model 2: Only includes the feature with the largest corr with Absent time.

RMSE: 8.9

We are not getting good results, maybe because we haven't included a very important attribute ***Reason***. Since it is an unordered category, we have to dummify this column.

Modeling – Linear Regression

First, let's take a look of the relationship with the column *Reason* and *Absent_time* by drawing a box plot.



Modeling – Linear Regression

Model 3: Include all the dummy variables from column ***Reason.***

RMSE: 10.2

Model 4: Includes all the dummy variables from column ***Reason*** and also the three features with the largest corr with Absent time.

RMSE: 10.4

We are getting even worse results when including the very important attribute ***Reason.***

*** Linear regression is not a good fit, even the best model we got so far has the same RMSE with the Null RMSE!**

Modeling – Decision Tree

Model 1: Includes all the dummy variables from column ***Reason***

Model 2: Includes the three features with the largest corr with Absent_time

Model 3: Includes all the attributes

Model 4: Includes top three attributes that have the largest importance rate

- Reason_12 (0.23)
- Reason_29 (0.17)
- Age (0.17)

*** For all of the models above, I used the for loop to get the best RMSE among a set range for the Tree depth. But unfortunately I'm not getting better RMSE (The best RMSE I got here is 9.4).**

Modeling – Random Forest

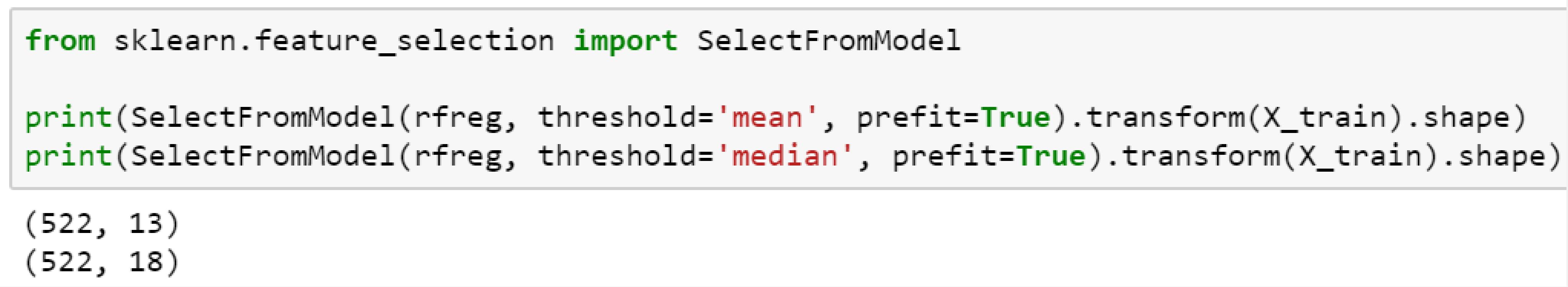
Original Model:

1. Includes all the attributes 2. n_estimators=150, max_features=5

RMSE: 12.4

Improve the original model:

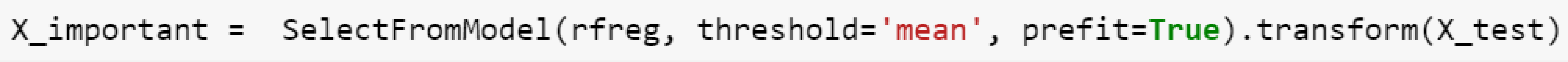
1. Reducing X to its Most Important Features



```
from sklearn.feature_selection import SelectFromModel

print(SelectFromModel(rfreg, threshold='mean', prefit=True).transform(X_train).shape)
print(SelectFromModel(rfreg, threshold='median', prefit=True).transform(X_train).shape)
```

```
(522, 13)
(522, 18)
```

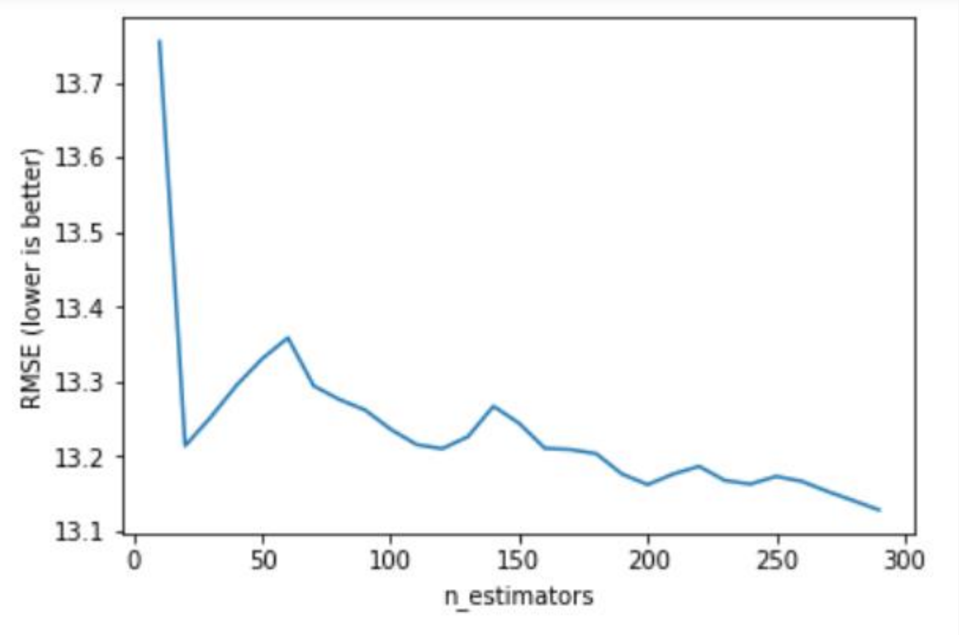


```
X_important =  SelectFromModel(rfreg, threshold='mean', prefit=True).transform(X_test)
```

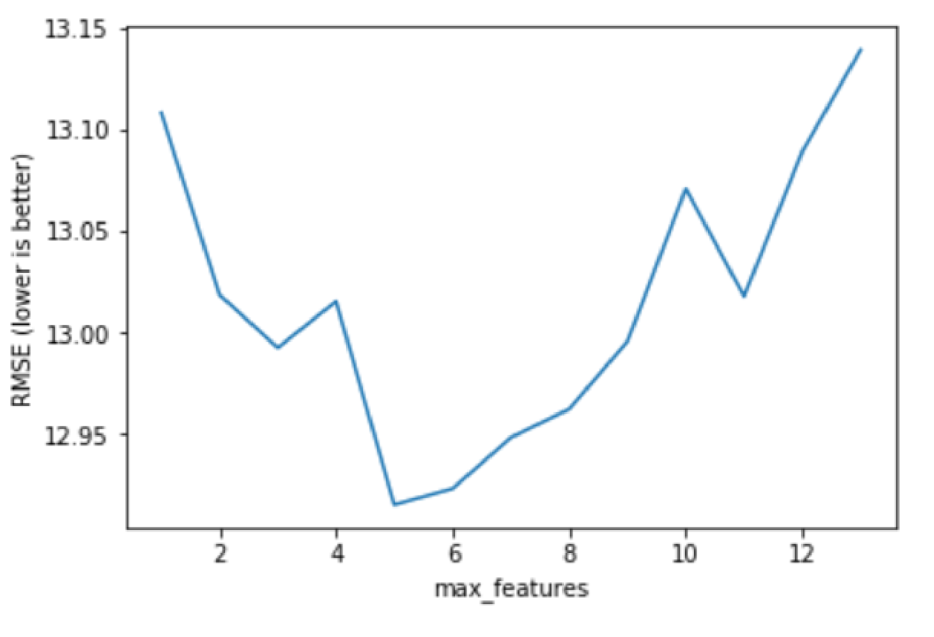
Modeling – Random Forest

Improve the original model:

2. Tuning n_estimators



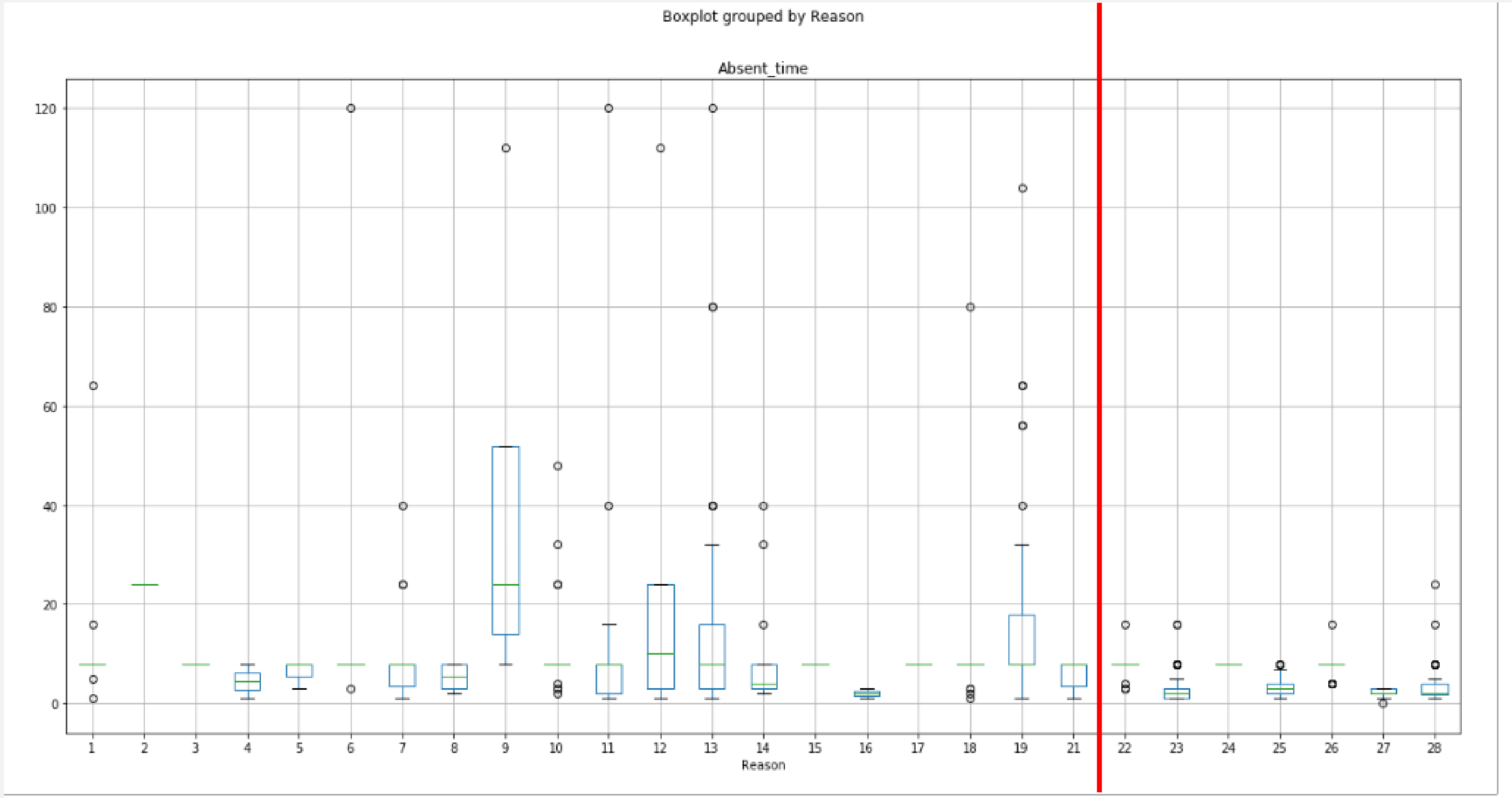
3. Tuning max_features



After re-train our data by using the optimized n_estimators and max_features values, the RMSE decreased to 8.9 but it's still not good compare to Null RMSE.

Modeling – Random Forest

What can I do next to improve the prediction?



Modeling – Random Forest

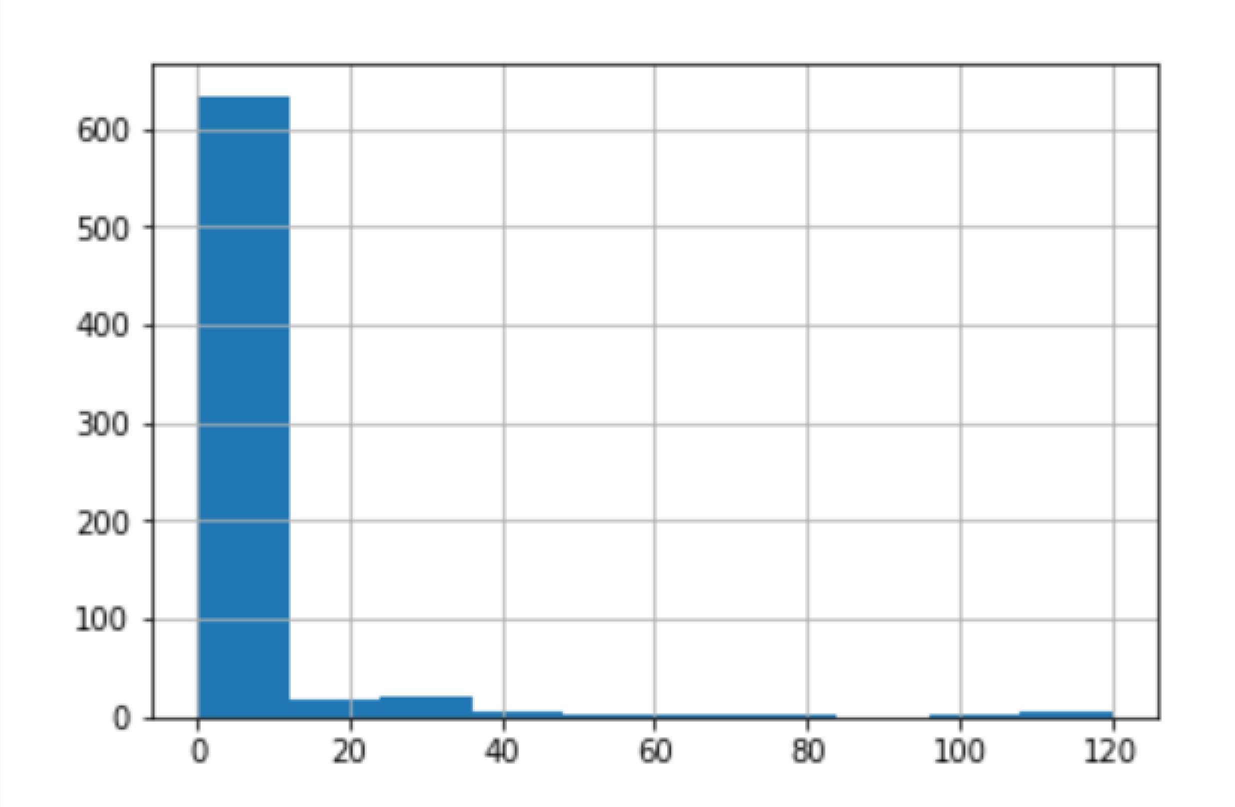
- Map the values from column ***Reason*** to 1 or 0.
- Set X to its Most Important Features (Top 6 out of 12 features)

We get the best RMSE so far: 8.6 (It’s slightly better than the Null RMSE: 8.9).

There are more ways to tune this model, but I want to spend time on other thoughts.

Back to analyzing the data:

Absent_time distribution

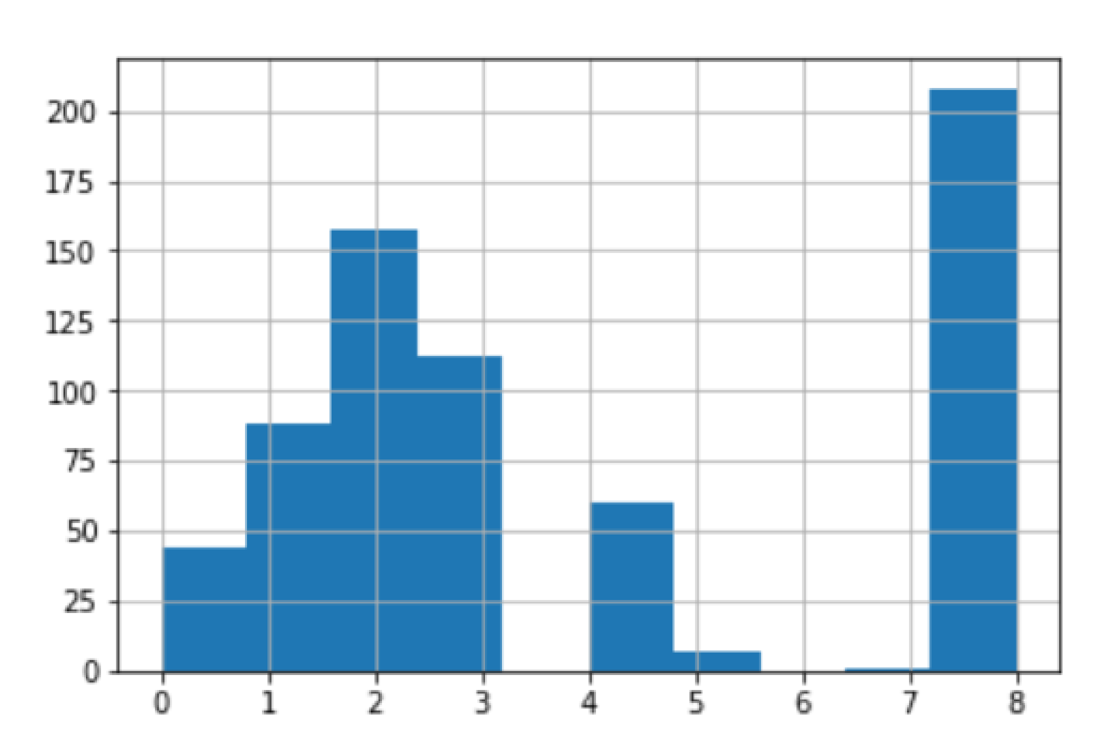


Super skewed!!
Why didn’t I
check this first???

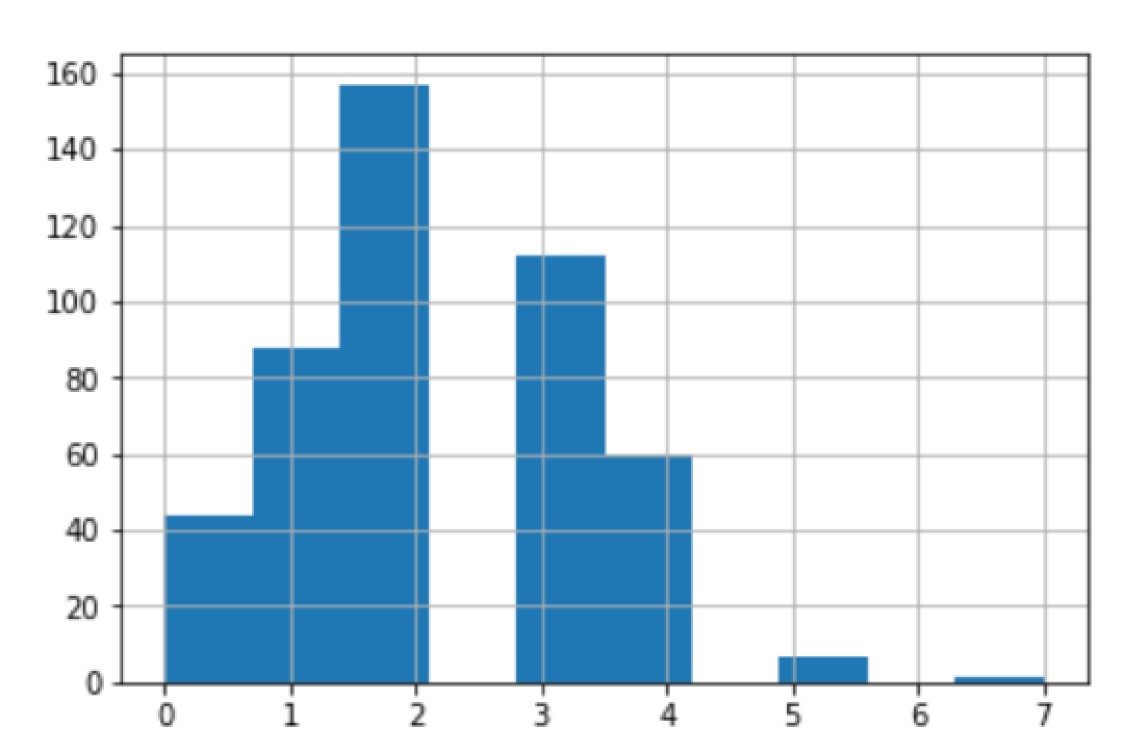
Back to analyzing the data:

Trying to narrow down the range of Absent_time to get a better distribution.

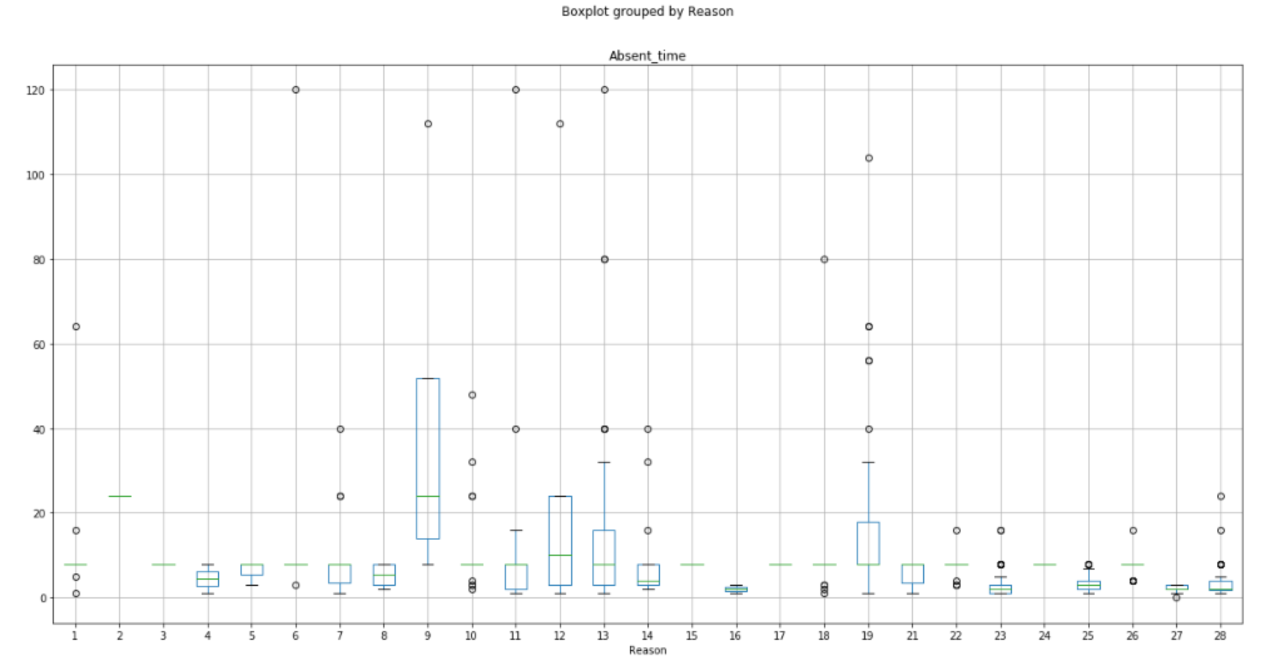
Absent_time<10 **91% of the whole data set**



Absent_time<8 **60% of the whole data set**

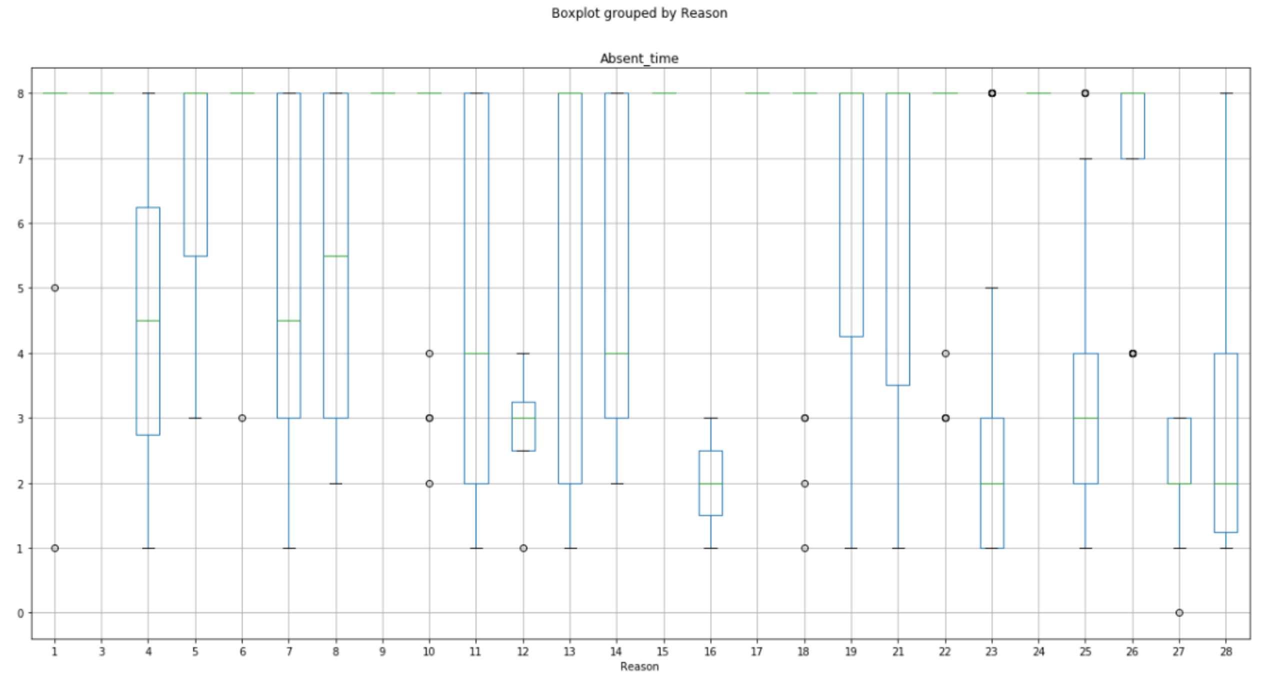


Back to analyzing the data:

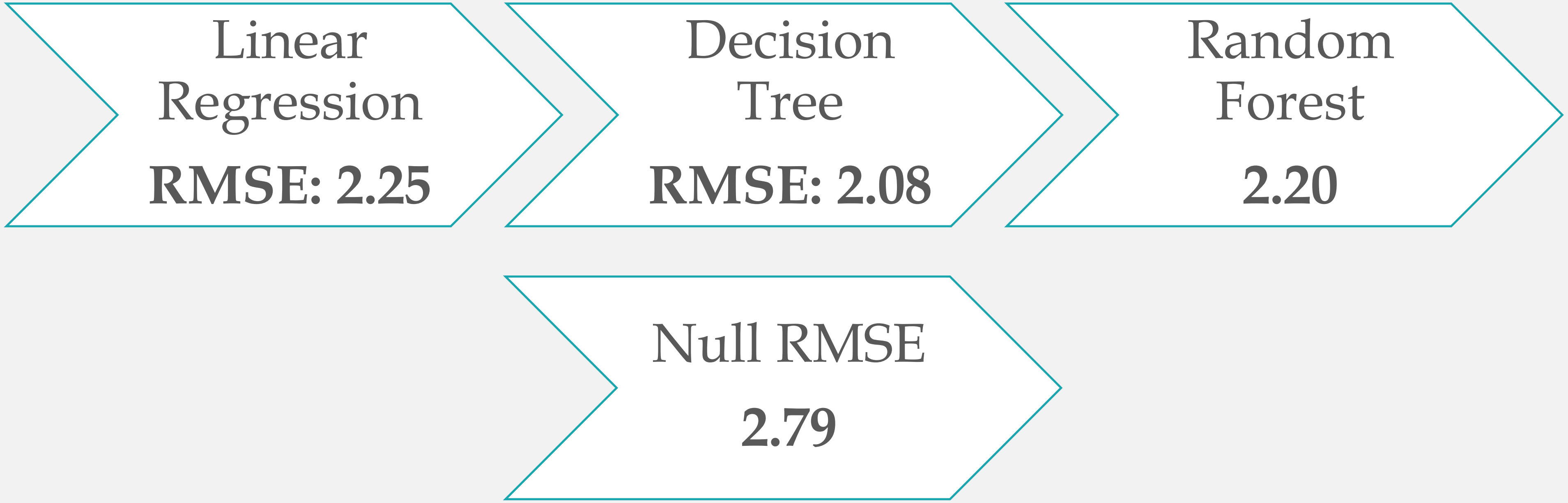


Original Absent_time distribution

After filtered out 9% of the outliers



Modeling –Linear Regression, Decision Tree, Random Forest



Linear Regression

RMSE: 2.25

Decision Tree

RMSE: 2.08

Random Forest

2.20

Null RMSE

2.79

Our models are getting better results compared to the baseline, even though they are not perfect.

Observation:

The original data set (with more outliers): In general, models have better RMSE value when using less features for prediction.

Filtered data set (with less outliers): In general, models have better RMSE value when using more features for prediction.

Why?!

Lesson learned and next steps:

- It’s always a good idea to take a look at the data’s distribution first. As a business owner or decision maker, a model that can predict well the majority of the time may be good enough.
- A small data set may not give the best prediction and the outliers have more impacts on the models.
- There are other interesting attributes we can analyze. For example, the relationship between Age, Body_mass, Social_drinker, Social_smoker, Season and Reason (Figure out possible causes for certain disease).
- Study more about how to better choose features for prediction and get more experience through practice.

Thank you!

It’s been a nice journey!