MAR & Related Analysis

2019/06~2019/09

Part 1
MAR Related Data Exploratory

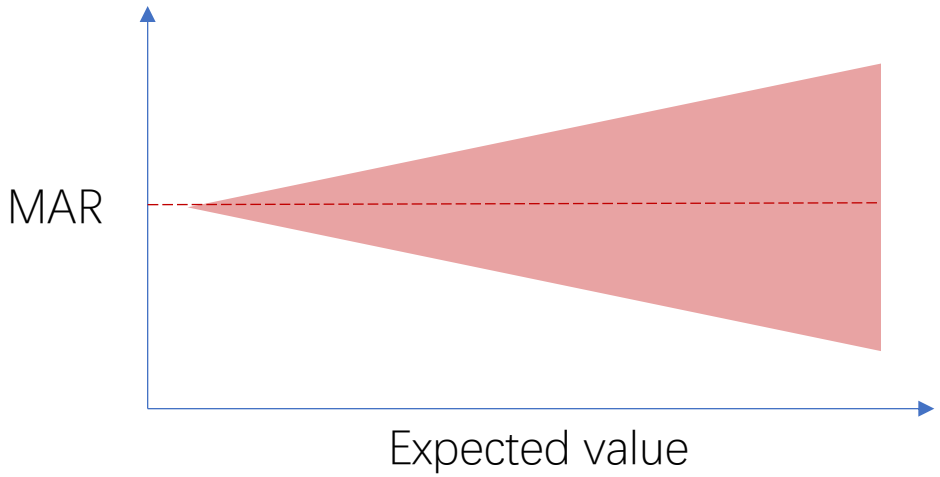
Part 1-MAR Introduction

1. MAR Definition:

MAR = (Actual Headcount – Expected Headcount) / Number of Classes

2. How Do We Evaluate Mar?

1. MAR plot (Non standardized Residual Plot)



2. MAR distribution of every Classes

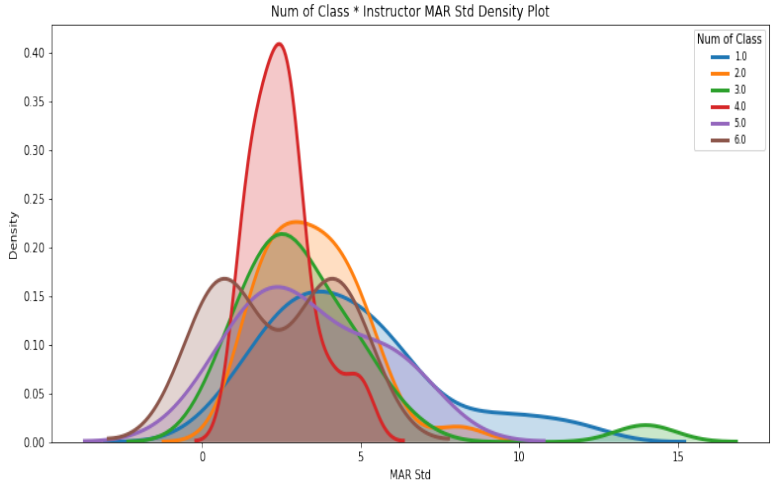
- Expect abnormal prediction ratio<10%
- Abnormal prediction ratio = overestimation + underestimation
- Overestimation: 90% of MAR score of every instructor(n>=2) is above zero;
- Underestimation: less than 10% of MAR score of every instructor(n>=2) is above zero

Part 1-Number of Class Recommendation

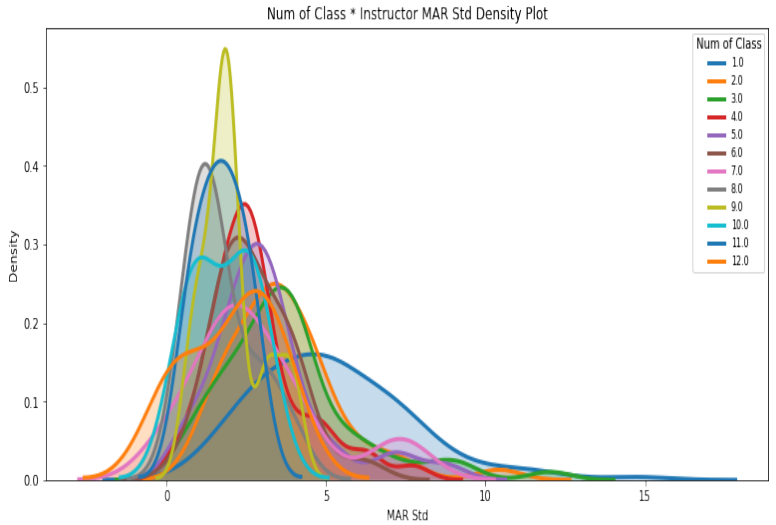
Based on baseline model, find out how many classes should an instructor take in order to have a more stable and reliable MAR scores

- Method:
  - First calculate instructor-wise standard deviation under same size of classes,
  - Second calculate class size-wise standard deviation from different instructors. This is to see the variation for different instructor under same size of classes.
- Results:
  - For small to medium size of gyms for example the max number of classes per instructor is less than 6, 3-4 classes per weeks, the standard deviation reaches its lowest level.
  - For large size of gyms for example the max number of classes per instructor teach is greater than 10, the difference of variance would not be that obvious when instructor teaches more than 2 classes per week.

CincinnatiSports



MidTown

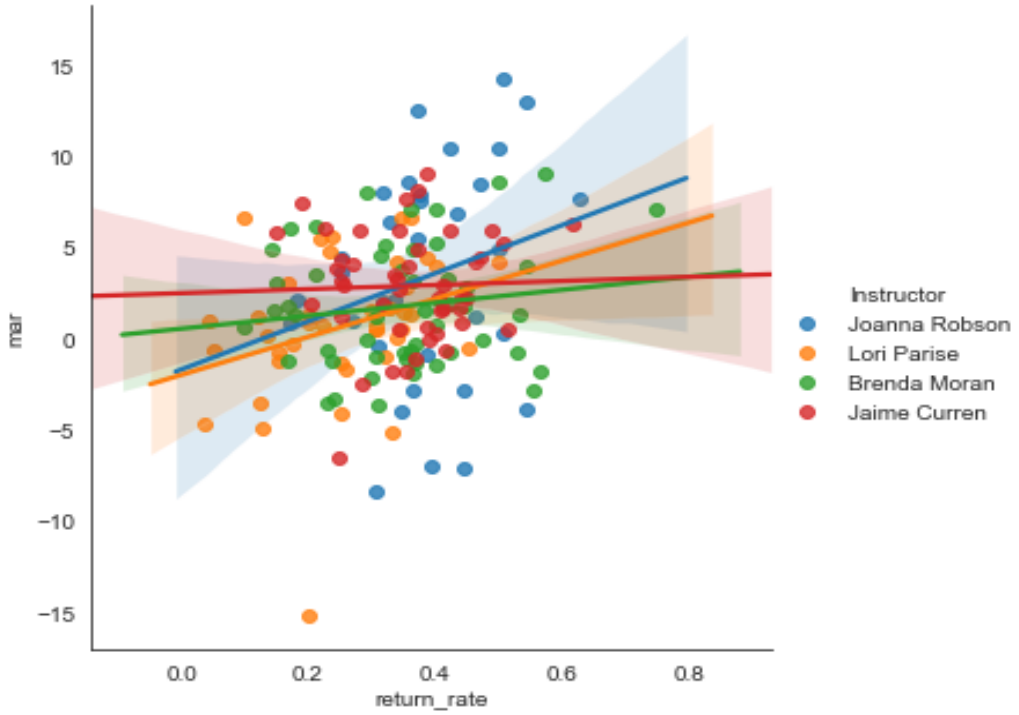


Part 1-Return Rate Correlation Analysis

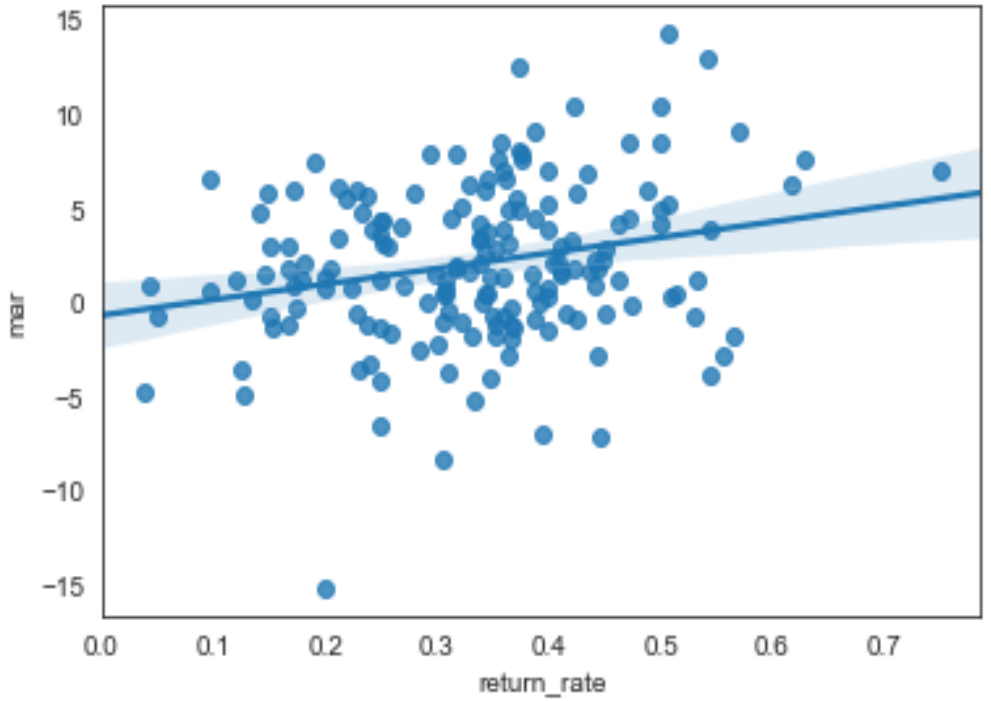
Below is a sample results of MAR(Instructor-wised) vs Return Rate correlation (Inshape).
This result below deliberately chose Instructors with positive MAR mean score with stable class teaching records for every week.

MAR Vs Return Rate (by Instructor)
Avg(MAR) >0
Correlation rate = 0.23, p-value <0.005

Distribution of MAR for each Instructor



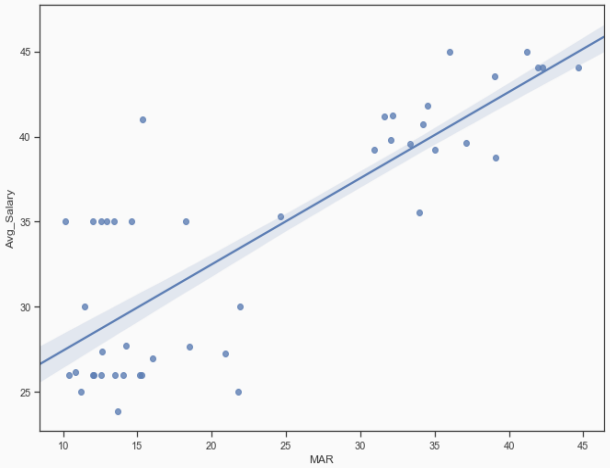
Combined results for all Instructors(condition)



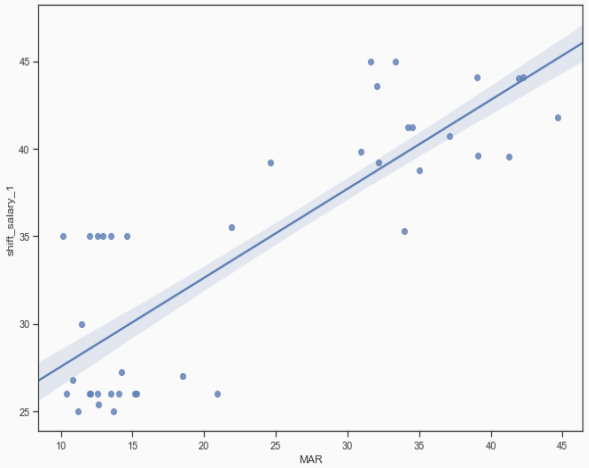
Part 1-MAR & Management Decision (NAC vs 24Hour)

NAC
MAR(>=10) vs Hourly Rate

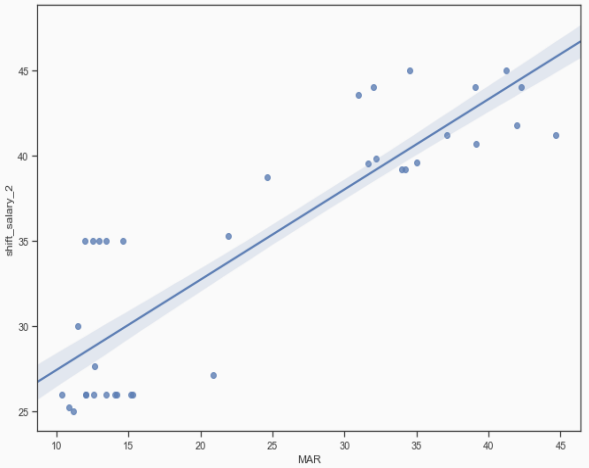
No Time Lag -0.82



Time Lag 1 -0.83

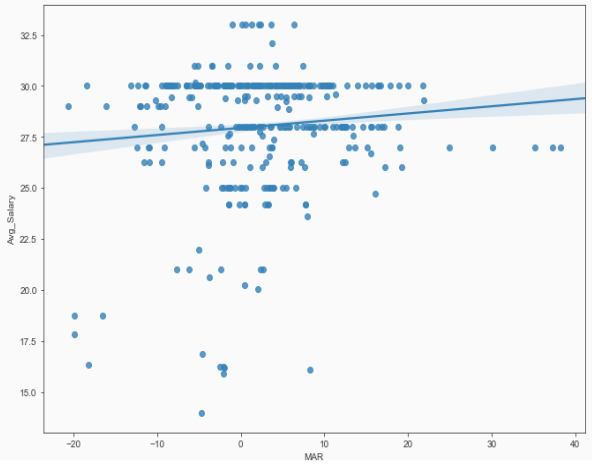


Time Lag 2 -0.87

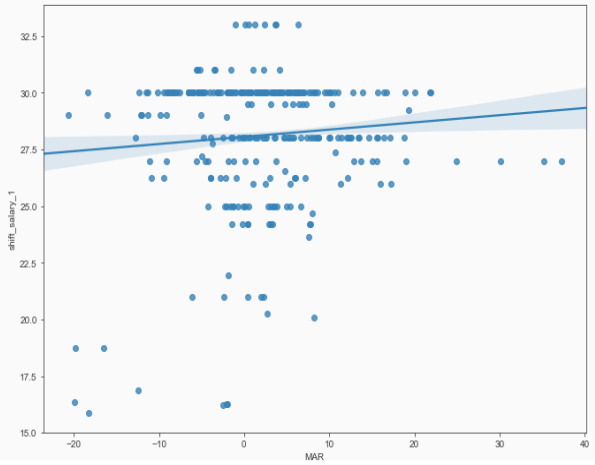


24Hour
MAR vs Hourly Rate

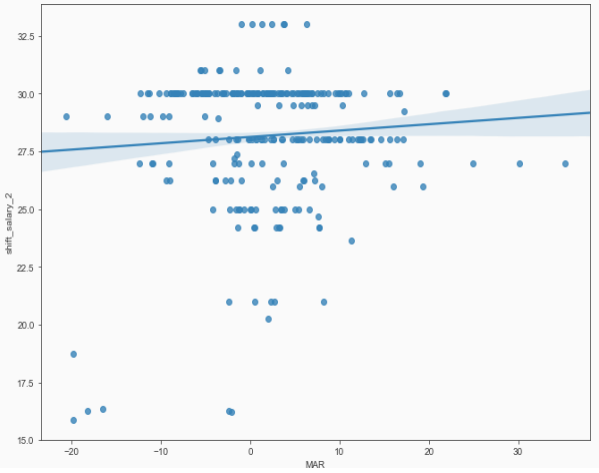
No Time Lag -0.1



Time Lag 1-0.09



Time Lag 2-0.07



Part 2
Model Modification

Part 2-Data Cleaning – Drop Invalid Data

01

Virtual Classes

- No actual Instructor presence

- Unrelated to instructor evaluation
- Bring bias to prediction especially on smaller size of classroom

02

Outdoor Classes

- Pool Classes/Field Classes related outdoor classes

- Not suitable for prediction
- Would focus more for indoor classes

03

Teen Classes

- Classes for kids for example Teen KongFu class etc.

- Legal issue
- Attendance pattern is hard to predict
- Would focus on adult prediction

Part 2-Data Cleaning – Analysis of Irregular Classes

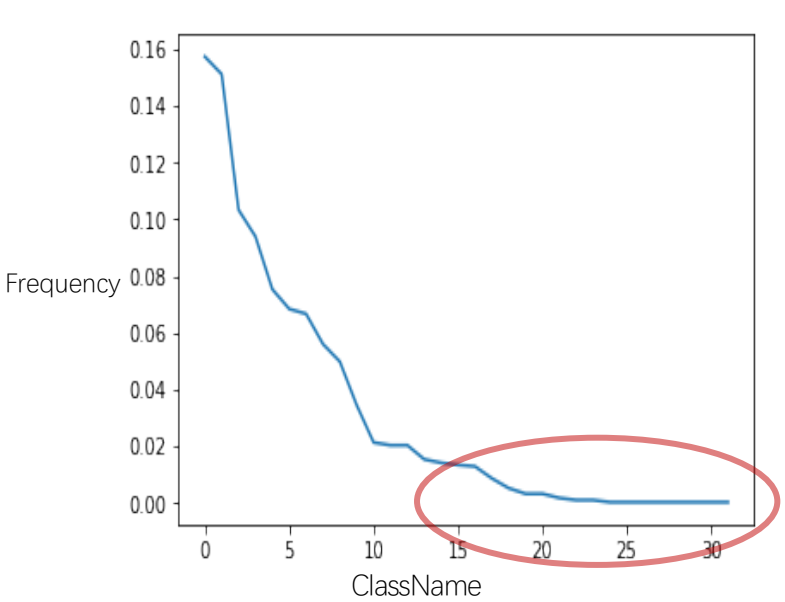
ClassName	Cum_Freq	Freq
freestyle ride	0.157334	0.157334
bodypump	0.308474	0.151140
yoga	0.411794	0.103320
gentle yoga	0.505699	0.093905
zumba	0.581021	0.075322
cxworx	0.649405	0.068385
bodycombat	0.716056	0.066650
barre	0.772052	0.055996
power yoga	0.821853	0.049802
u-jam fitness	0.856293	0.034440
beatz ride	0.877602	0.021308
yoga flow	0.897919	0.020317
turbo kick	0.918236	0.020317
ultimate conditioning	0.933598	0.015362
circuit xtreme	0.947721	0.014123
easy does it	0.961100	0.013380

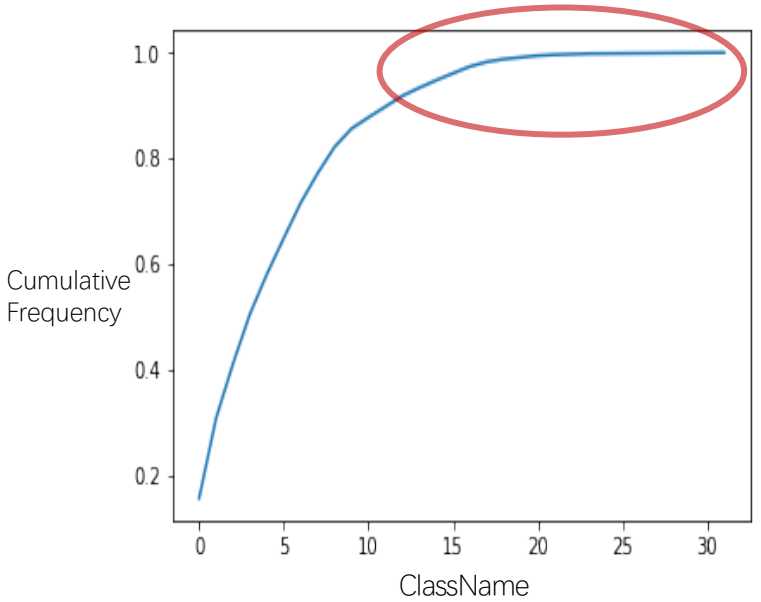
... ...

balletone	0.998761	0.000248
les mills sprint™	0.999009	0.000248
hot dance nights - zumba	0.999257	0.000248
free team training boot camp	0.999504	0.000248
beatz ride - 1 year anniversary party	0.999752	0.000248
pop pilates	1.000000	0.000248

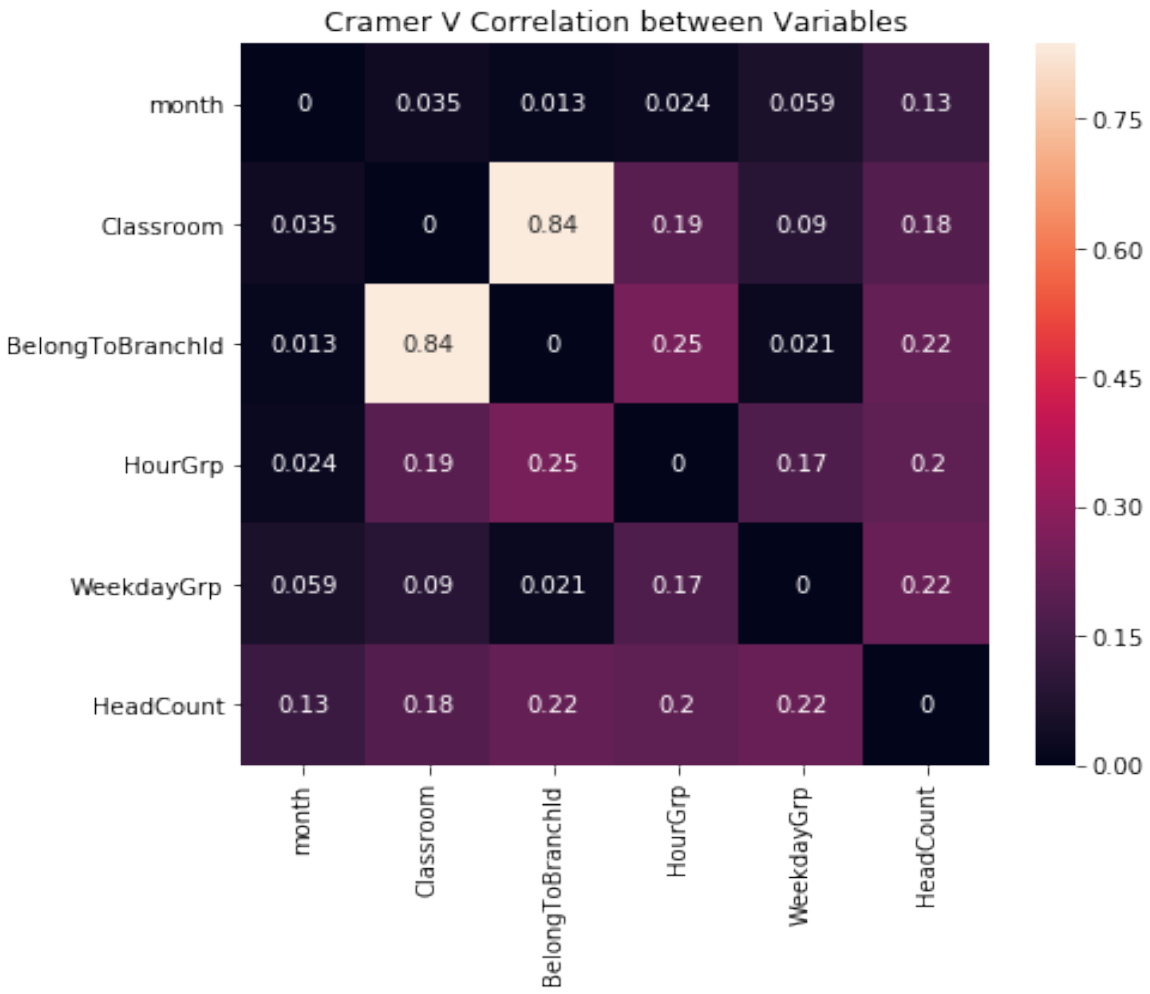
Irregular Classes

- The definition of irregular Classes here means that aside from regularly happened classes, there are a lot of classes that is happening seasonally. For example some classes are ended with boot camp, festival or party. The characteristics of this kind of classes are low frequency, which may bring bias when we try to use it to predict the expected attendance.
- Therefore before add those classes into our prediction, we need to test the actual attendance of those classes to see if it has similar pattern with the other regular classes by fulfilling various conditions and decide if it is suitable to add those classes





Part 2- Feature Engineering(1)



01 Check Variable Dependency

- Our Goal
  - Check the dependency of independent variables.
  - A strong correlation on independent variables may affect model stability when apply to different dataset or slightly change of data
- Our Findings
  - Except branch and classroom, the correlation between the other variables are low. For all gyms if there is more than 1 branch, the correlation between brand and classroom is very strong.
- Our Solutions
  - Combine branched and classroom into one variable. Some of the gym may not have 2 or more branch, but for long term consideration, combine those two variables would be a practical solution.

Part 2- Feature Engineering(2)

02 Interaction Term Analysis

- Our Goal
  - Apply different combination of model to dataset
  - Find a simple and accurate model in predicting MAR

- How to Decide Which Interaction Terms to Add ?
  - Check deviance of all interaction terms.
  - If the deviance drop a lot when we add one interaction term, it means that one is significant.

★ Baseline Model:

HeadCount ~ month + BranchId*HourGrp*WeekdayGrp + Classroom

Test Model 1:

HeadCount ~ Month + BranchId_Classroom*HourGrp*WeekdayGrp

Test Model 2:

HeadCount ~ Month + HourGrp*WeekdayGrp + BranchId_Classroom + BranchId_Classroom:HourGrp:WeekdayGrp

Test Model 3:

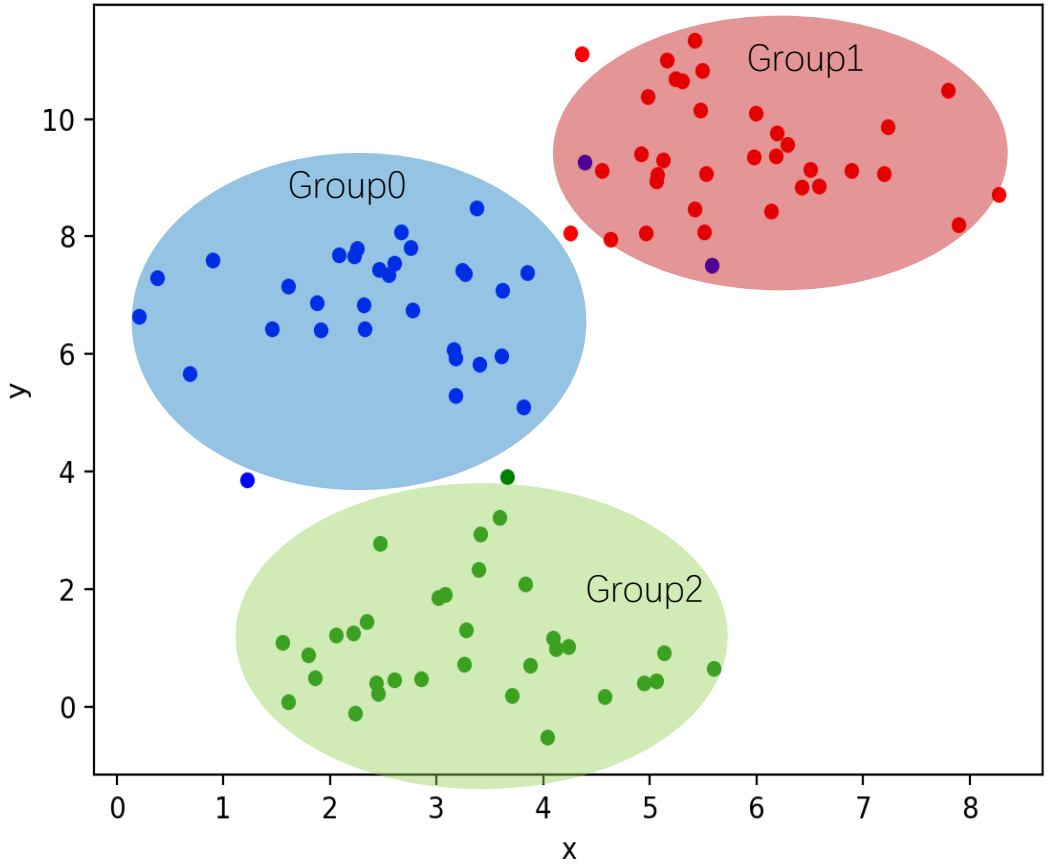
HeadCount ~ month + HourGrp:WeekdayGrp + BranchId_Classroom + BranchId_Classroom:HourGrp:WeekdayGrp

Part 2- Feature Engineering(3)

03 Add Adjustment (Conditional)

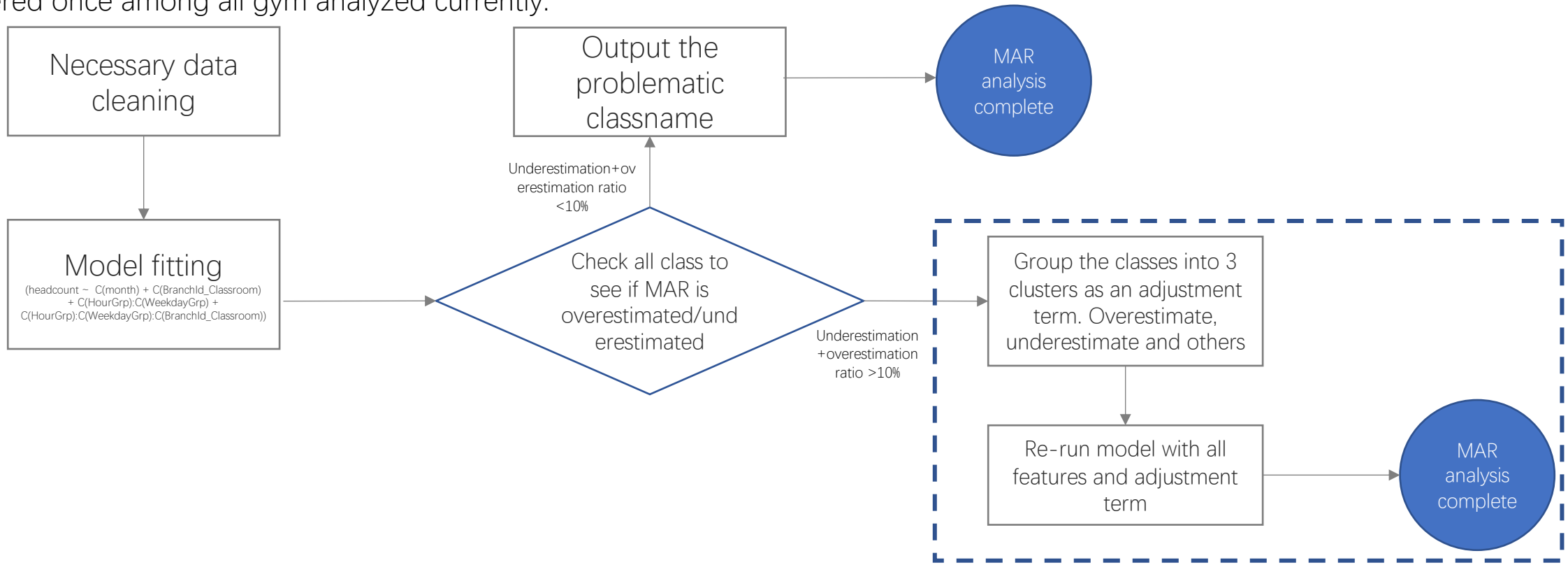
- Our Goal
  - Make sure that MAR distribution for all classes are reasonable to assess instructors 'performance
  - If underestimation and overestimation ratio in our prediction is high we need to consider a solution
- Our Solutions
  - Cluster classes into different clusters based on conditions
  - Mark underestimated classes as group1, overestimated classes as group 2 and the other classes are group 0.
  - Re run model if (group1 + group2)/(group0+group1+group2) >0.1

For Example, this is how we label classes into different groups.



Part 2- Overall Model Pipeline Summary

Below is the summary of model pipeline for all gyms, the only change compared to last report is the add on terms on the adjustment for the model after find out the ratio of overestimate and underestimate MAR score for all classes. Currently the threshold for underestimate and overestimate condition is set really high: over 90% positive of MAR score for a class is considered overestimate and less than 10% positive MAR score for a class is considered underestimate. And the adjustment process would not be trigger until the sum of the underestimate + overestimate ratio above 10%. Under this standard the adjustment is only triggered once among all gym analyzed currently.

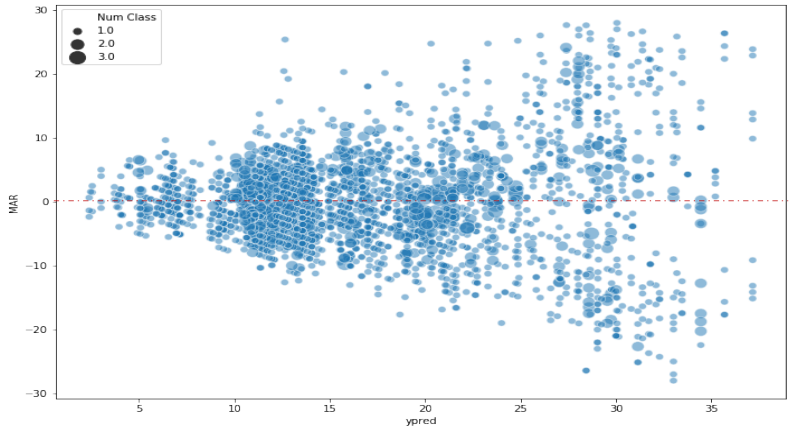


Part 3
Model Generalization & Comparison

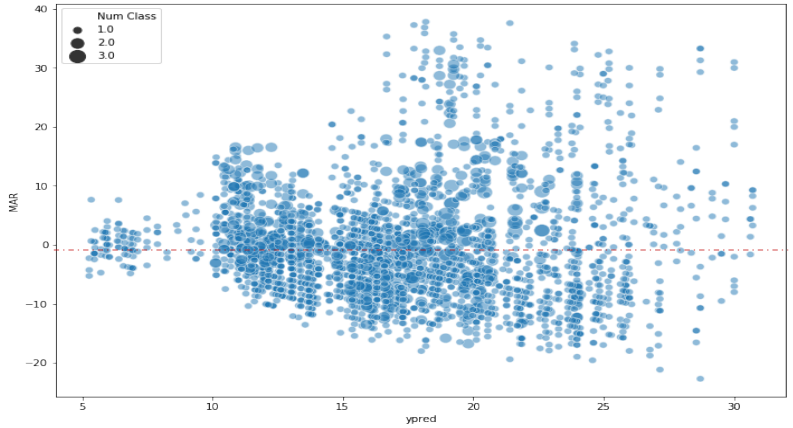
Part 3 – Inshape

Visualization of points distributions for Inshape revised model, baseline model and comparison

Revised Model



Baseline Model

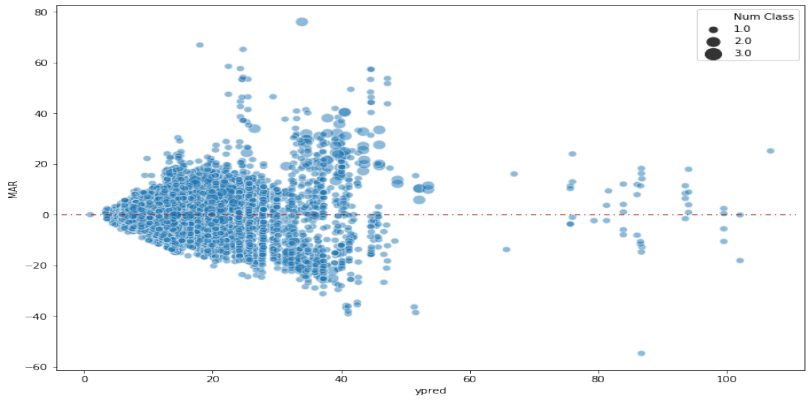


Comparison

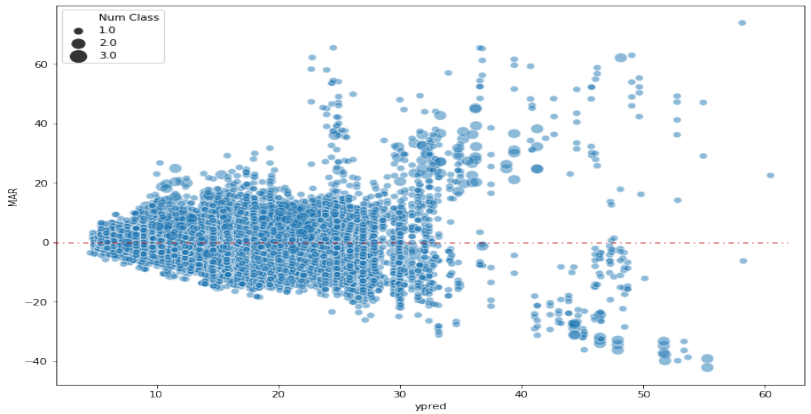
Part 3 – NAC

Visualization of points distributions for NAC revised model, baseline model and comparison

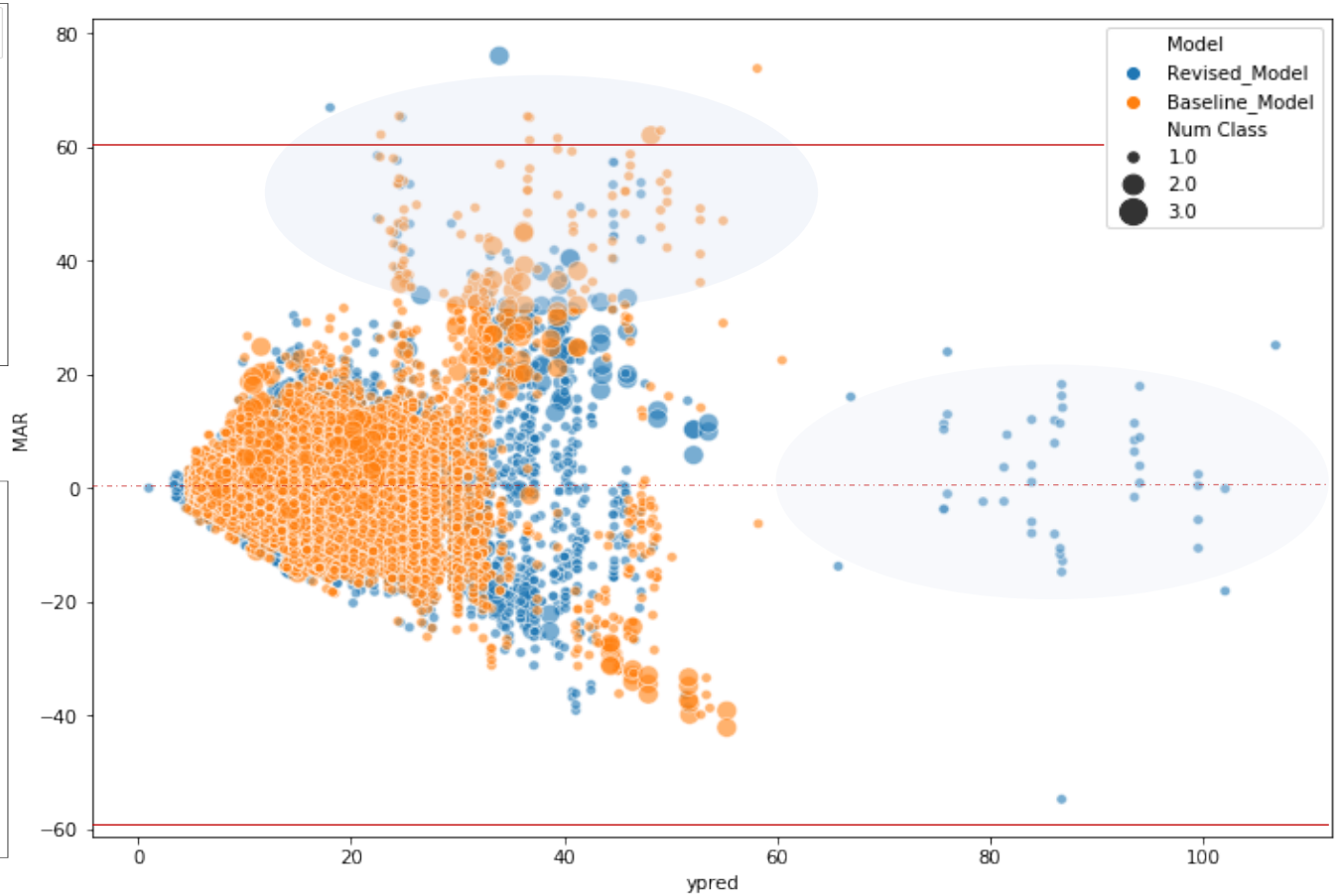
Revised Model



Baseline Model



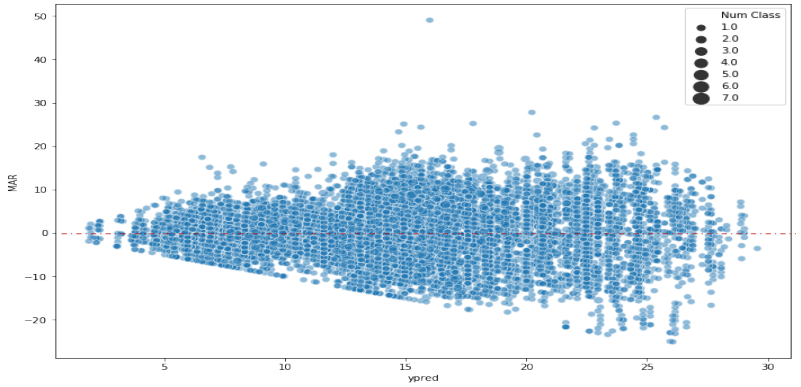
Comparison



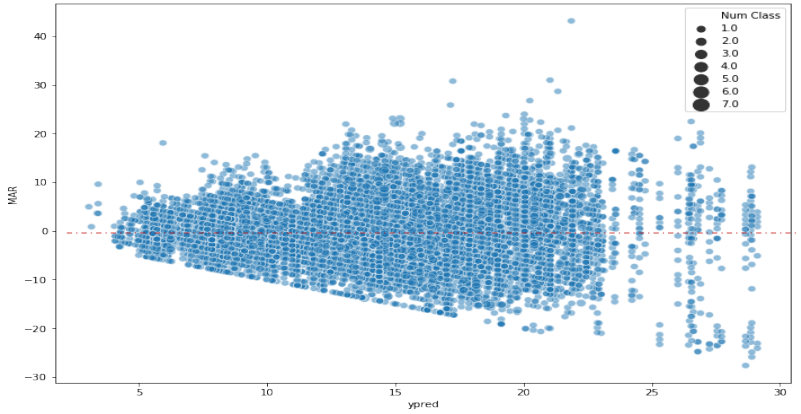
Part 3 – Sawmill

Visualization of points distributions for sawmill revised model, baseline model and comparison

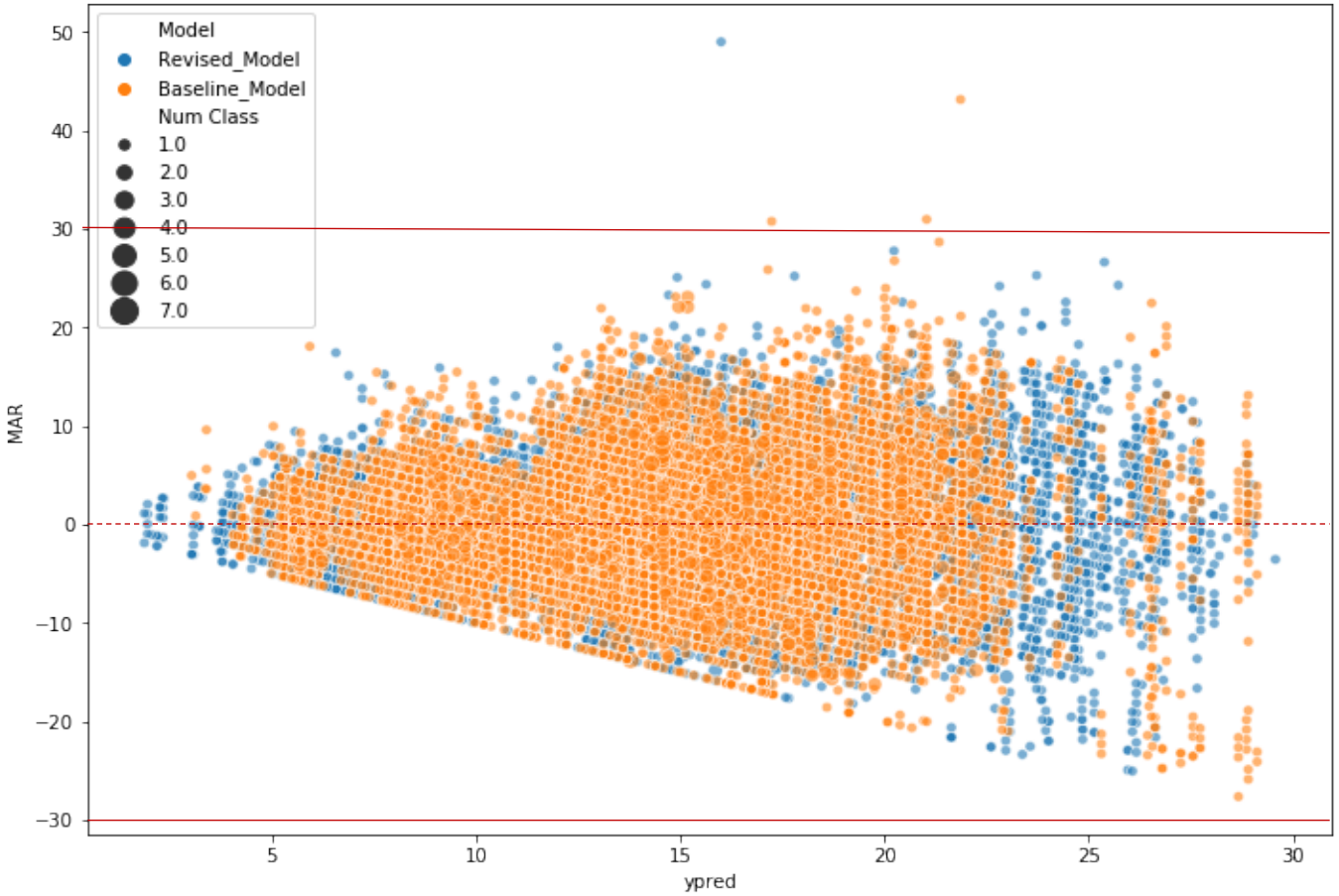
Revised Model



Baseline Model



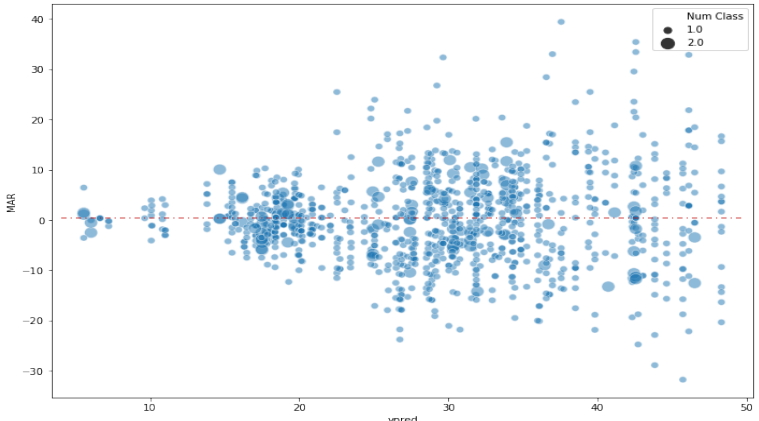
Comparison



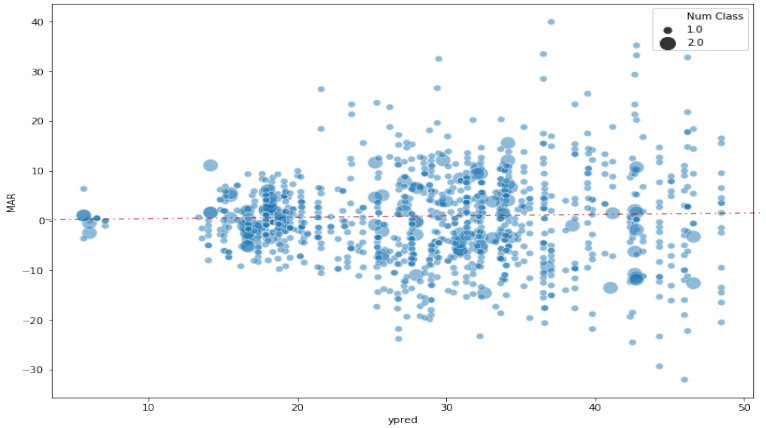
Part 3 – 24Hour

Visualization of points distributions for sawmill revised model, baseline model and comparison

Revised Model



Baseline Model



Comparison