Comparing predictions

- Paul: **submit36.csv** / LB: 0.21034
- Gino: **avg__Nov2__15.csv** / LB: 0.21465

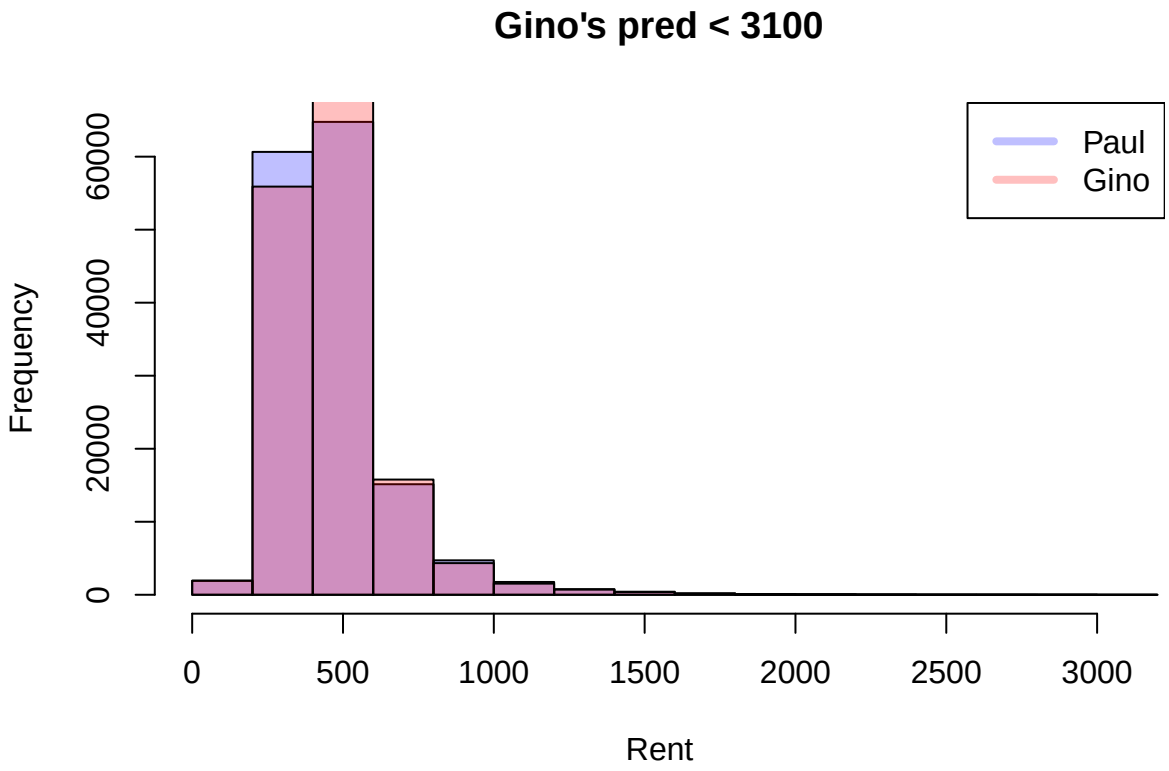
Comparing predictions' distribution

```
## pred_Paul$REN_BASE_RENT
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##  150508       0  142777       1   464.5   260.2   289.8   350.5   424.2
##     .75     .90     .95
##   523.3   674.5   813.3
##
## lowest :   11.98   13.63   29.04   31.80   32.65
## highest: 2693.61 2806.75 2815.76 2925.03 3089.38
```

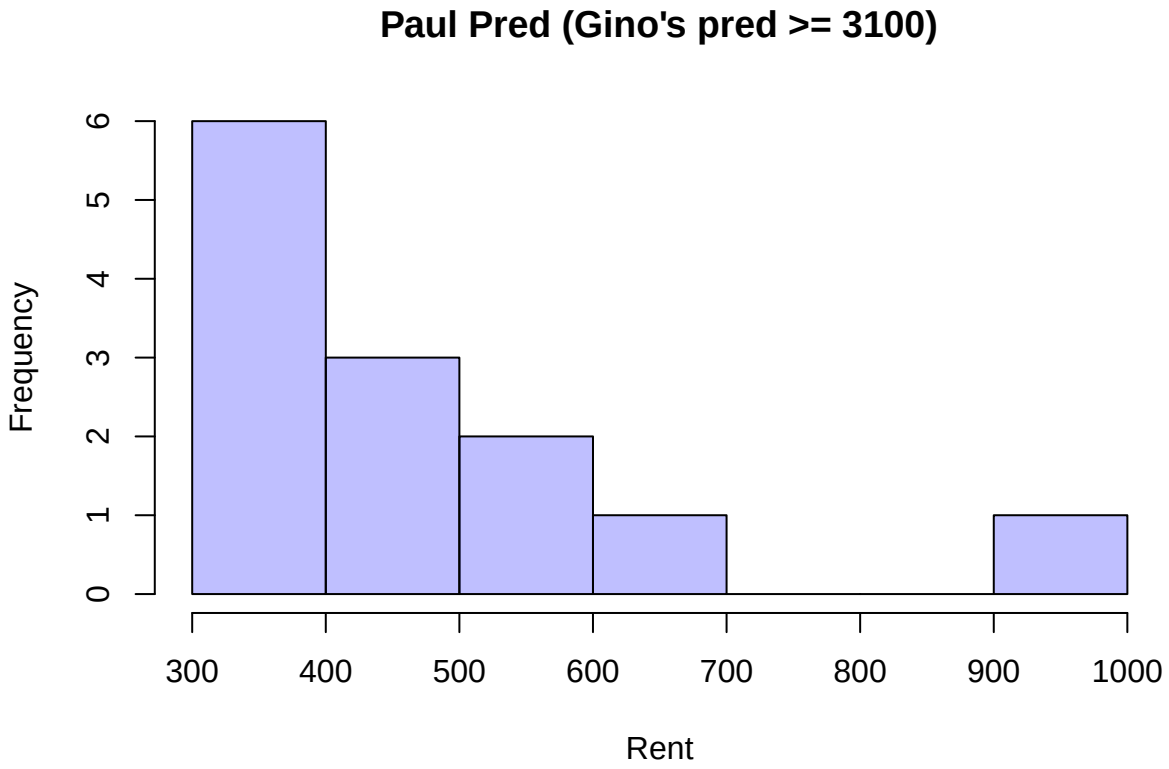
```
## pred_Gino$REN_BASE_RENT
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##  150508       0  149630       1   468.3   261.4   293.5   357.9   432.3
##     .75     .90     .95
##   528.0   669.8   793.9
##
## lowest :    59.65    64.94    68.92    69.82    70.35
## highest:  4245.56  7356.94  9277.14 27884.67 48413.29
```

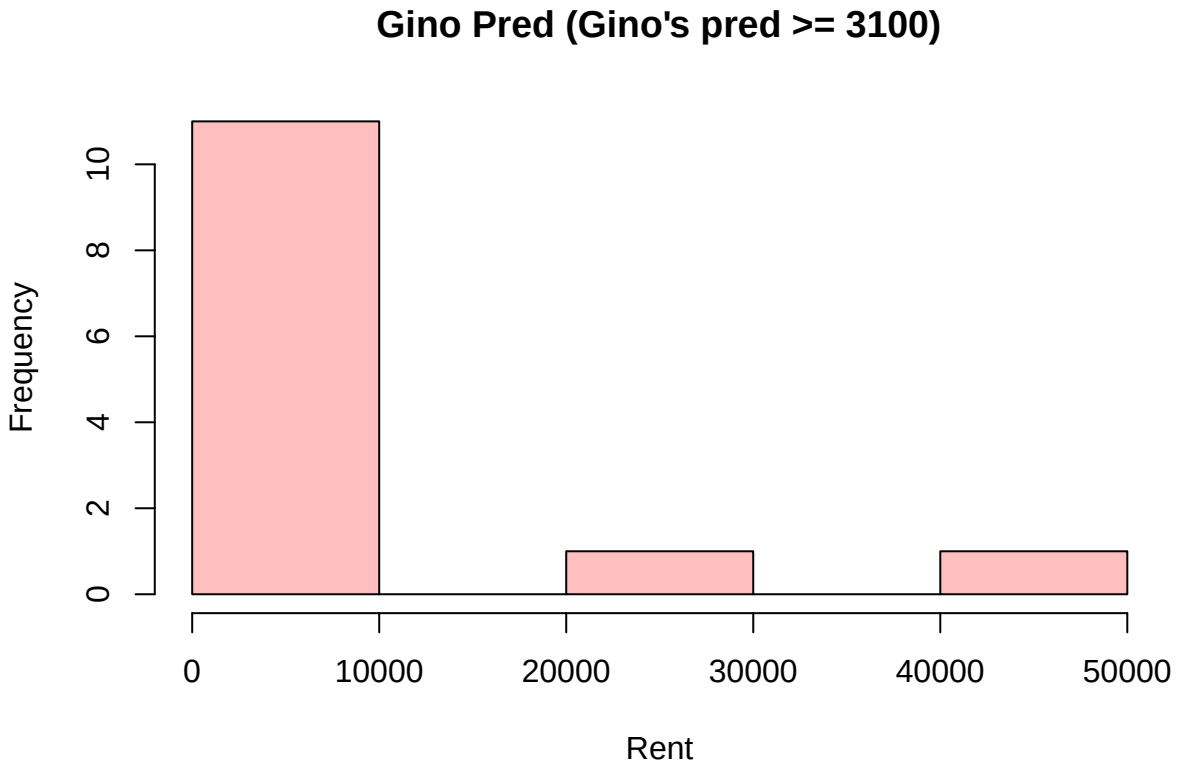
- similar mean (464.5 vs 468.3)
- different standard deviation (193.3795 vs.239.4788). Gino's prediction has more outliers >3089.38 (= highest value of Paul's prediction).

Comparing predictions - Gino's pred $<$ 3100



Comparing predictions - Gino's pred $>=$ 3100





Simple Model averaging

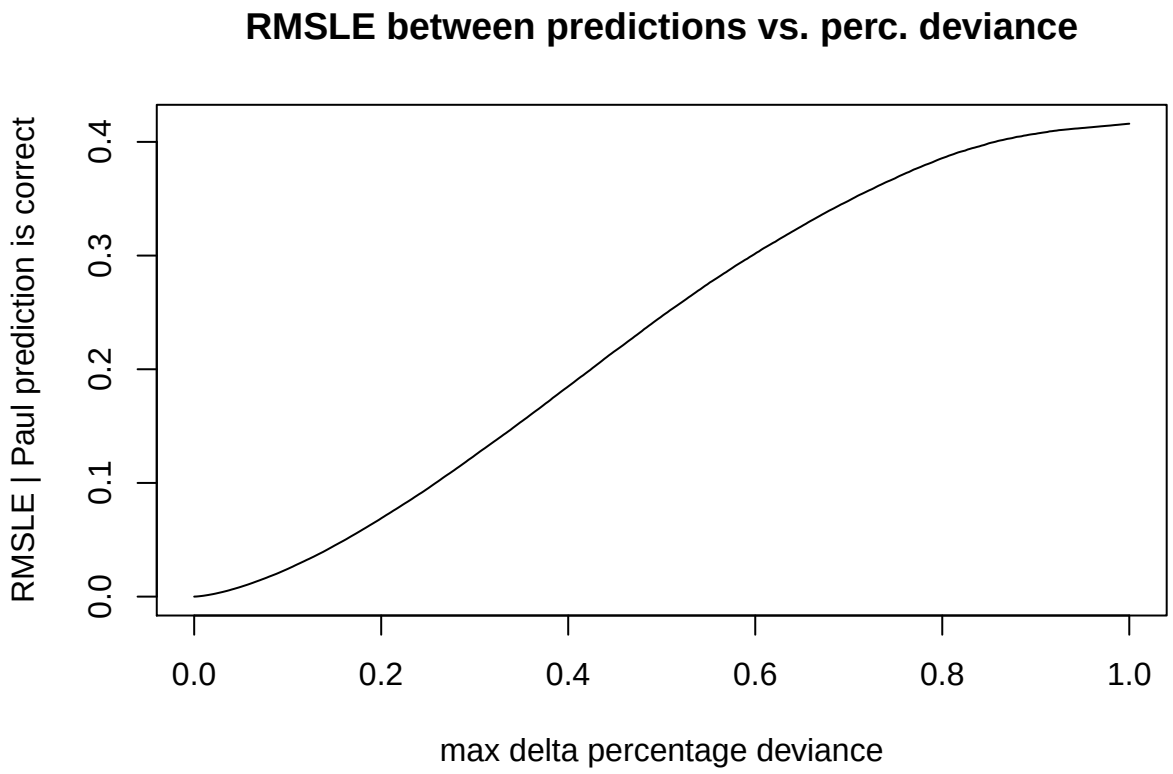
- weight Paul: 0.6
- weigth Gino: 0.4
- ~0.28 LB

Concordance Model averaging

- Assumed Paul prediction as better, averaged models (same weigths above) only if the max delta percentage deviance between predictions < 0.29 (covering ~50% of th test set).
- LB ~ 0.214

```
###
delta_distribution_perc = data.frame(max_delta =
                                       seq(from = 0,to = 1,length.out = 400) , err_perc = NA)

for (i in seq_along(delta_distribution_perc$max_delta) ) {
  MAX_DELTA = delta_distribution_perc$max_delta[i]
  delta = abs((pred_Paul$REN_BASE_RENT-pred_Gino$REN_BASE_RENT)/pred_Paul$REN_BASE_RENT)

  delta_idx = which(delta<=MAX_DELTA)

  pred_1_overlap = pred_Paul$REN_BASE_RENT[delta_idx]
  pred_2_overlap = pred_Gino$REN_BASE_RENT[delta_idx]

  pred_test = pred_Paul$REN_BASE_RENT
  pred_test[delta_idx] = pred_Gino$REN_BASE_RENT[delta_idx]
```

```
  rmsle_overlap = RMSLE(pred=pred_test, obs=pred_Paul$REN_BASE_RENT)

  delta_distribution_perc[delta_distribution_perc$max_delta==MAX_DELTA,]$err_perc = rmsle_overlap
}

plot(x = delta_distribution_perc$max_delta,y = delta_distribution_perc$err_perc , type = "l" ,
     xlab="max delta percentage deviance" , ylab="RMSLE | Paul prediction is correct" ,
     main = "RMSLE between predictions vs. perc. deviance")
```



Some findinds

- Averaging predictions with simple linear coefficients doesn't work
- A % of Gino's outliers should be correct, otherwise LB~0.214 isn't explicable
- If it was possible to identify a significant part of Gino's good outliers and then merging them with Paul base prediction our score should boost

Mapping Gino's outliers vs. Paul base in the feature space

TODO