Doordash Delivery Time Prediction

For Dr. Shaonan Tian
BUS 235C: Data Mining

By Paris Patel

Why study this dataset?

- World is moving to online retail and delivery services owing to the pandemic
- Businesses are going big on improving customer satisfaction
- Key element: the ‘Joy of Receiving’
- Hence accurate delivery time prediction becomes essential to delivering happiness to the customer.

About the data: Doordash Delivery Time Prediction

- This data was collected in early 2015 and has 197,428 records.
- This data has 16 variables representing date-time features, store and market features, order features and time estimate features of the orders placed during this time period.
- More features were derived from these variables, such as actual_delivery_duration, busy_dasher_ratio and estimated_non_prep_duration.

```
##  [1] "market_id"
##  [2] "created_at"
##  [3] "actual_delivery_time"
##  [4] "store_id"
##  [5] "store_primary_category"
##  [6] "order_protocol"
##  [7] "total_items"
##  [8] "subtotal"
##  [9] "num_distinct_items"
## [10] "min_item_price"
## [11] "max_item_price"
## [12] "total_onshift_dashers"
## [13] "total_busy_dashers"
## [14] "total_outstanding_orders"
## [15] "estimated_order_place_duration"
## [16] "estimated_store_to_consumer_driving_duration"
```

Snippet of Dataset

market_id	created_at	actual_delivery_time	store_id	store_primai	order_protoc	total_items	subtotal	num_distinct	min_item_pi	max_item_p	total_onshift	total_busy_d	total_outsta	estimated_o	estimated_store_to_consumer_driving_duration
1	2/6/15 22:24	2/6/15 23:27	1845	american	1	4	3441	4	557	1239	33	14	21	446	861
2	2/10/15 21:49	2/10/15 22:56	5477	mexican	2	1	1900	1	1400	1400	1	2	2	446	690
3	1/22/15 20:39	1/22/15 21:09	5477	NA	1	1	1900	1	1900	1900	1	0	0	446	690
3	2/3/15 21:21	2/3/15 22:13	5477	NA	1	6	6900	5	600	1800	1	1	2	446	289
3	2/15/15 2:40	2/15/15 3:20	5477	NA	1	3	3900	3	1100	1600	6	6	9	446	650
3	1/28/15 20:30	1/28/15 21:08	5477	NA	1	3	5000	3	1500	1900	2	2	2	446	338
3	1/31/15 2:16	1/31/15 2:43	5477	NA	1	2	3900	2	1200	2700	10	9	9	446	638
3	2/12/15 3:03	2/12/15 3:36	5477	NA	1	4	4850	4	750	1800	7	8	7	446	626
2	2/16/15 0:11	2/16/15 0:38	5477	indian	3	4	4771	3	820	1604	8	6	18	446	289
3	2/18/15 1:15	2/18/15 2:08	5477	NA	1	2	2100	2	700	1200	2	2	2	446	715
3	2/2/15 19:22	2/2/15 20:09	5477	NA	4	4	4300	4	1200	1500	1	1	1	446	453
3	2/16/15 4:19	2/16/15 6:34	5477	NA	1	2	2200	2	600	1600	3	3	4	446	642
3	2/7/15 1:34	2/7/15 2:17	5477	NA	1	1	1900	1	1900	1900	6	3	3	446	690
3	1/25/15 1:50	1/25/15 2:28	5477	NA	4	4	4986	4	699	2362	16	6	9	446	445
1	2/12/15 3:36	2/12/15 4:14	2841	italian	1	1	1525	1	1525	1525	5	6	8	446	795
1	1/27/15 2:12	1/27/15 3:02	2841	italian	1	2	3620	2	1425	2195	5	5	7	446	205
1	2/6/15 0:42	2/6/15 2:10	2841	italian	1	3	4475	3	925	1825	4	1	1	446	542
1	2/8/15 2:04	2/8/15 3:27	2841	italian	1	3	4375	3	1325	1625	6	4	3	446	789
1	1/31/15 4:35	1/31/15 5:47	2841	italian	1	2	3150	2	1425	1725	4	9	12	446	548
1	1/31/15 2:21	1/31/15 3:11	4139	mexican	1	2	950	2	150	700	24	24	26	446	212
1	1/31/15 23:45	2/1/15 0:14	4139	mexican	1	5	1285	3	150	400	12	13	11	446	424
1	2/17/15 3:13	2/17/15 4:00	5058	italian	1	4	5800	4	700	2000	19	19	30	446	344
1	2/15/15 1:26	2/15/15 2:16	5058	italian	1	2	2800	2	1200	1600	21	17	16	446	421
1	2/2/15 5:27	2/2/15 7:05	5058	italian	1	7	14900	5	1200	3900	8	11	11	446	901
1	2/16/15 2:21	2/16/15 3:54	5058	italian	1	3	3400	3	1400	2100	22	21	39	446	501
1	2/13/15 1:01	2/13/15 1:40	5058	italian	1	3	3800	3	700	1600	16	9	7	446	344
1	1/24/15 2:01	1/24/15 2:46	5058	italian	1	1	1800	1	1800	1800	27	25	24	446	424
1	2/10/15 4:36	2/10/15 5:20	5058	italian	1	4	5800	4	700	2000	13	16	27	446	344
1	1/26/15 2:09	1/26/15 2:47	5058	italian	1	3	3800	3	800	1800	21	18	20	446	530
1	2/7/15 1:44	2/7/15 2:48	5058	italian	1	1	1800	1	1800	1800	22	21	20	446	434
1	1/30/15 1:43	1/30/15 2:20	5058	italian	1	2	2700	2	1200	1500	16	16	13	446	456
3	1/30/15 19:49	1/30/15 20:28	4149	sandwich	2	4	3490	3	175	850	16	16	21	251	298
1	1/28/15 20:33	1/28/15 21:04	4149	NA	NA	3	1765	3	275	895	22	21	24	251	490
1	2/12/15 20:27	2/12/15 21:18	4149	NA	2	1	975	1	975	975	24	25	30	251	487
1	1/28/15 21:37	1/28/15 22:11	4149	NA	2	3	1425	3	175	900	13	12	7	251	704
2	2/9/15 20:29	2/9/15 21:17	4149	thai	5	1	1220	1	825	825	21	21	20	251	835
1	1/21/15 20:35	1/21/15 21:01	4149	NA	2	2	1750	2	825	825	18	18	20	251	259
1	2/16/15 19:19	2/16/15 20:16	4149	NA	2	8	7870	6	825	895	12	11	9	251	98

Data Clean Up

Missing and infinite values:

- `store_primary_category`: The missing values here were replaced by the most frequent primary category associated to the store_id.
- `busy_dasher_ratio`: There were erratic values in the values of total available values and busy dashers, which resulted in infinite values. The records with these values were removed.

Data Preparation for Model Inputs

- Redundancy Removal:
  - Redundant variables were removed based on the correlation index.
  - Correlation index > 0.5, high value

	Var1 <fctr>	Var2 <fctr>	value <dbl>
502	total_busy_dashers	total_onshift_dashers	0.9417415
503	total_outstanding_orders	total_onshift_dashers	0.9346389
602	total_outstanding_orders	total_busy_dashers	0.9312948
904	estimated_non_prep_duration	estimated_store_to_consumer_driving_duration	0.9230859
806	order_protocol1	estimated_order_place_duration	0.8976453
3	num_distinct_items	total_items	0.7581458
102	num_distinct_items	subtotal	0.6828903
2	subtotal	total_items	0.5571750
302	max_item_price	min_item_price	0.5412406
104	max_item_price	subtotal	0.5079474

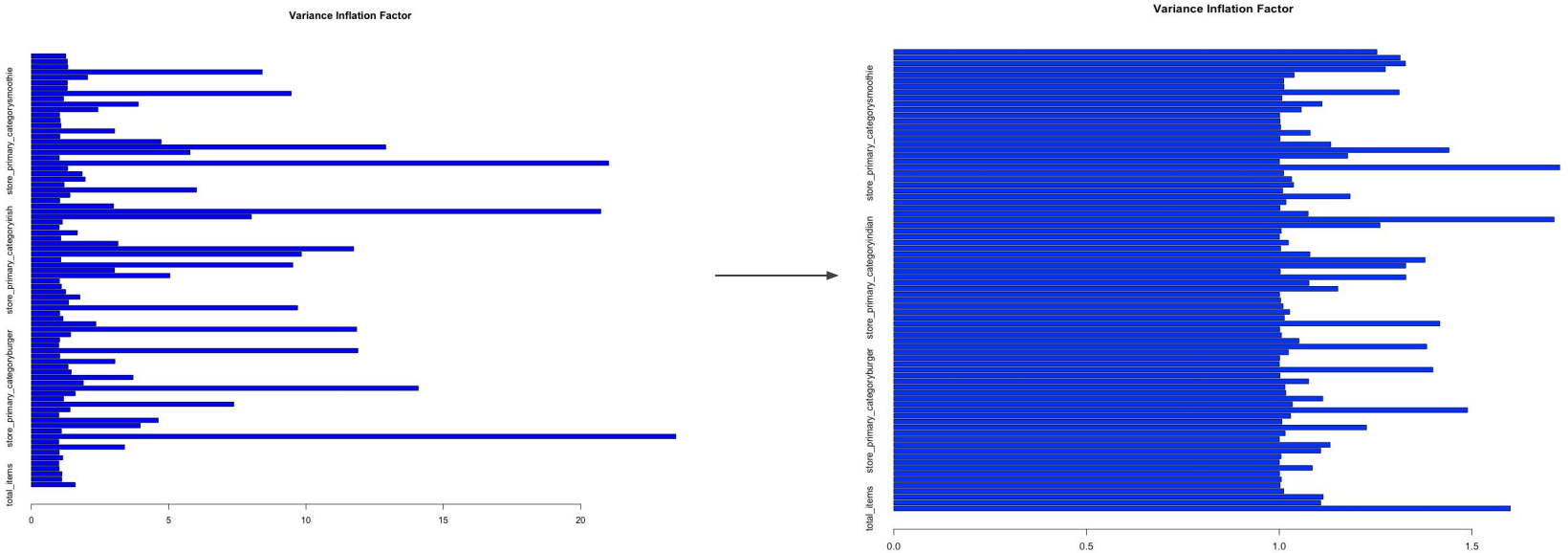
1–10 of 20 rows Previo

	Var1 <fctr>	Var2 <fctr>	value <dbl>
80	percent_distinct_item_total	total_items	0.44575119
82	price_range_of_items	total_items	0.33330429
81	avg_price_per_item	total_items	0.31075877
6559	avg_price_per_item	percent_distinct_item_total	0.22671088
5083	avg_price_per_item	store_primary_categorypizza	0.21256456
251	actual_delivery_duration	estimated_store_to_consumer_driving_duration	0.18841753
85	estimated_order_place_duration	total_outstanding_orders	0.17104873
2869	avg_price_per_item	store_primary_categoryfast	0.16726071
35	store_primary_categoryfast	total_items	0.16394007
87	actual_delivery_duration	total_outstanding_orders	0.15760431

1–10 of 20 rows Previou

Data Preparation for Model Inputs

- Multicollinearity Check
  - VIF > 15 considered high

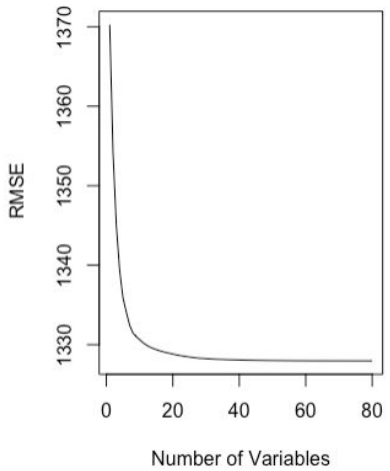


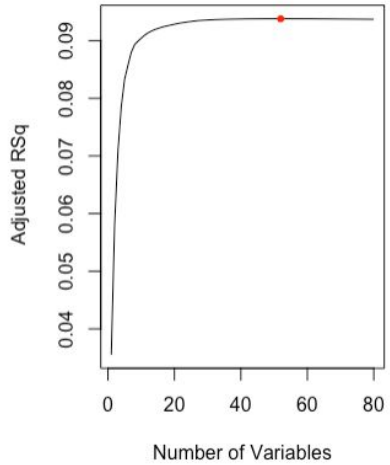
Data Preparation for Model Inputs

- Feature Selection
  - Forward step variable selection
  - Criteria: Median RMSE (~1328 seconds)

- Final total number of variables:

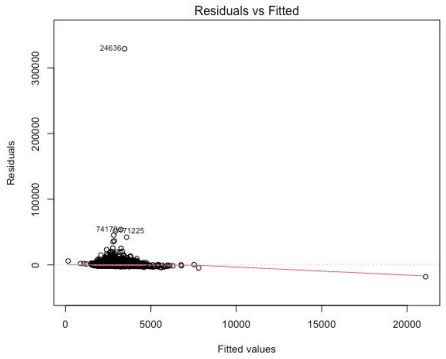
  34 predictors + 1 response variable

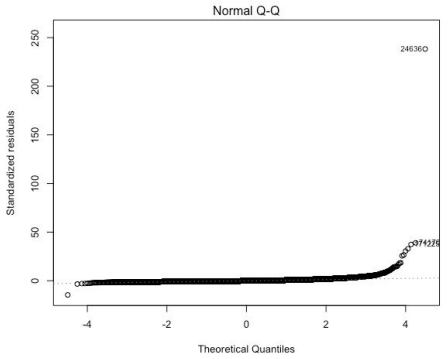


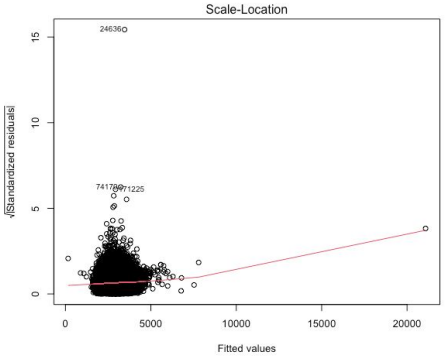


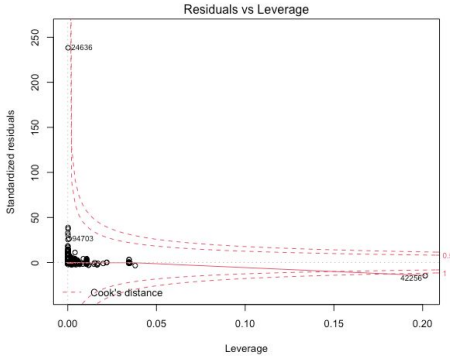
Linear Regression

```
Call:
lm(formula = actual_delivery_duration ~ ., data = delivery_train)

Residuals:
   Min     1Q Median     3Q    Max
-18031   -673   -203    432 329024

Coefficients:
                                              Estimate Std. Error t value             Pr(>|t|)
(Intercept)                                  2854.399      3.669 777.920 < 0.0000000000000002 ***
total_items                                   122.919      4.119  29.842 < 0.0000000000000002 ***
total_outstanding_orders                      219.151      3.812  57.497 < 0.0000000000000002 ***
estimated_order_place_duration                158.400      3.791  41.787 < 0.0000000000000002 ***
estimated_store_to_consumer_driving_duration  255.023      3.681  69.284 < 0.0000000000000002 ***
busy_dasher_ratio                              72.006      3.672  19.611 < 0.0000000000000002 ***
store_primary_categoryalcohol                 -15.727      3.696  -4.255 0.00002092437238406 ***
store_primary_categorybarbecue                -21.835      3.727  -5.859 0.00000000466265203 ***
store_primary_categorybrazilian                24.055      3.675   6.545 0.00000000005972405 ***
store_primary_categorybreakfast               -14.184      3.752  -3.781            0.000157 ***
store_primary_categorybubble.tea                6.188      3.684   1.679            0.093066 .
store_primary_categoryburger                    9.143      3.828   2.388            0.016929 *
store_primary_categorycafe                     17.688      3.723   4.751 0.00000202892236683 ***
store_primary_categoryconvenience.store        12.563      3.685   3.409            0.000653 ***
store_primary_categorydim.sum                  -7.685      3.688  -2.084            0.037148 *
store_primary_categoryfrench                  -12.170      3.679  -3.308            0.000941 ***
store_primary_categorygreek                    -9.104      3.721  -2.447            0.014407 *
store_primary_categoryhawaiian                -13.239      3.697  -3.581            0.000342 ***
store_primary_categoryindian                   23.659      3.782   6.256 0.00000000039716272 ***
store_primary_categoryjapanese                 27.842      3.793   7.341 0.00000000000021381 ***
store_primary_categorykorean                  -14.966      3.701  -4.043 0.00005275538782700 ***
store_primary_categorymediterranean             8.477      3.755   2.258            0.023963 *
store_primary_categorymexican                 -48.867      3.880 -12.595 < 0.0000000000000002 ***
store_primary_categorypasta                   -11.329      3.683  -3.076            0.002097 **
store_primary_categorypersian                 -16.978      3.682  -4.611 0.00000401737130795 ***
store_primary_categorypizza                    33.905      3.937   8.612 < 0.0000000000000002 ***
store_primary_categorysalad                    29.896      3.726   8.024 0.00000000000000103 ***
store_primary_categorysandwich                -19.597      3.818  -5.132 0.00000028672175021 ***
store_primary_categoryseafood                 -17.355      3.715  -4.671 0.00000299917030925 ***
store_primary_categorysouthern                  7.652      3.670   2.085            0.037085 *
store_primary_categorytapas                    18.231      3.672   4.965 0.00000068949501536 ***
store_primary_categoryturkish                  -6.897      3.674  -1.877            0.060473 .
store_primary_categoryvietnamese              -29.301      3.763  -7.787 0.00000000000000693 ***
avg_price_per_item                             86.875      4.075  21.317 < 0.0000000000000002 ***
price_range_of_items                           73.872      3.977  18.577 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1381 on 141621 degrees of freedom
Multiple R-squared:  0.08852,   Adjusted R-squared:  0.0883
F-statistic: 404.5 on 34 and 141621 DF,  p-value: < 0.00000000000000022
```







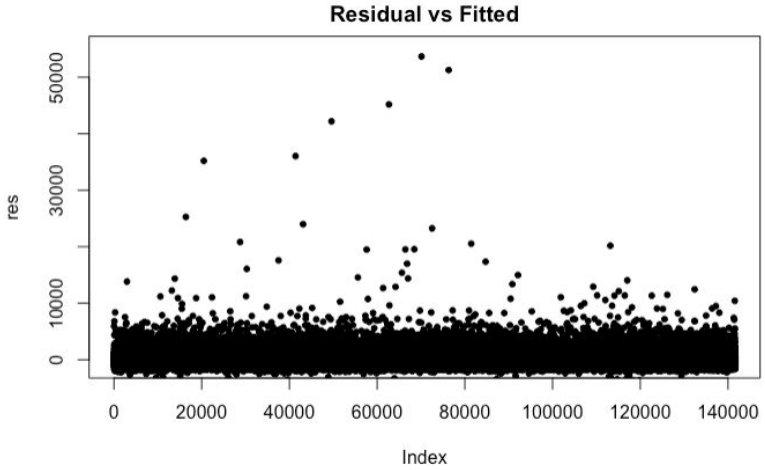


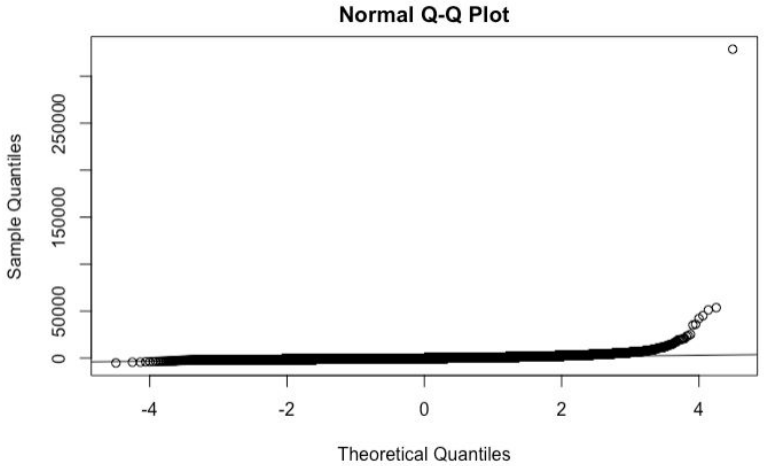
Random Forest Regressor

```{r}
rf_model
```

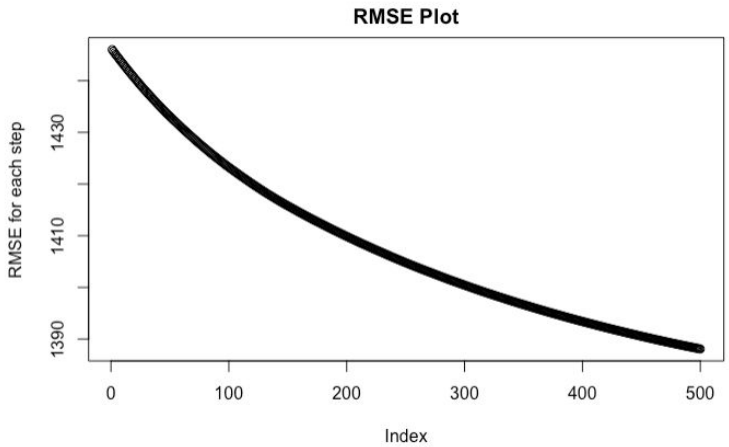
```
Ranger result

Call:
 ranger(x = delivery_train[-c(ncol(delivery_train))], y = delivery_train$actual_delivery_duration,      seed = 123, num.tree = 500,
write.forest = TRUE, replace = TRUE,      importance = "impurity")

Type:                             Regression
Number of trees:                  500
Sample size:                      141656
Number of independent variables:  34
Mtry:                             5
Target node size:                 5
Variable importance mode:         impurity
Splitrule:                        variance
OOB prediction error (MSE):       1849498
R squared (OOB):                  0.1158833
```

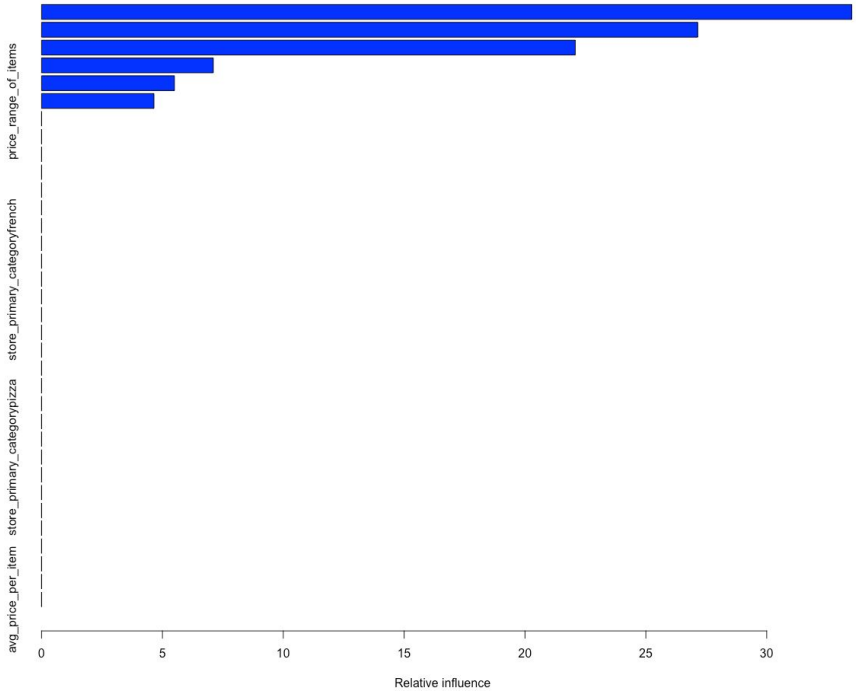
Random Forest Regressor





Gradient Boosting Regressor





Findings

In-Sample RMSE

Linear.RMSE <dbl>	Random_forest.RMSE <dbl>	Gradient_boost.RMSE <dbl>
1380.84	1046.452	1388.084

Out-of-sample RMSE

Linear.RMSE <dbl>	Random_forest.RMSE <dbl>	Gradient_boost.RMSE <dbl>
1092.528	1066.675	1100.065

How to make it better?

- Try other feature selection methods and different combination of features
- Add loss function to minimize underestimation of time
- Should also try single Decision Tree, XGBoost models
- Modify the problem to estimate cooking time

Thank you!