Intro2DA

April 19, 2021

Introduction to Data Analytics

Kevin Kam Fung Yuen
kevinkf.yuen@gmail.com
19 April 2021

1 Background and Objectives

Data includes the profiles of the clients and the charge how much insurance company to insure them. A data analyst needs to find the insights from the data. He/she needs to performance a KDD process for the activities below.
- Descriptive Analytics: distribution of data.
- Diagnostic Analytics: factors contributing to the charges.
- Predictive Analytics: predict the insurance charges for the new people based on the information we will get from them.
- Prescriptive Analytics: suggest deployments methods based on the previous steps.

2 Get Data from Internet

[1]:
```
path = "https://raw.githubusercontent.com/stedy/
↪Machine-Learning-with-R-datasets/master/insurance.csv"
```

[2]:
```
data <-read.csv(path,stringsAsFactors = TRUE );
head(data)
```

A data.frame: 6 × 7

	age <int>	sex <fct>	bmi <dbl>	children <int>	smoker <fct>	region <fct>	charges <dbl>
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622

3 Data cleansing and preparation

[3]:
```
# check missing values
sum(is.na(data))
```

0

[4]:
```
# check structure of data
str(data)
```

```
'data.frame':	1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 …
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 …
 $ bmi     : num  27.9 33.8 33 22.7 28.9 …
 $ children: int  0 1 3 0 0 0 1 3 2 0 …
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 …
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2
…
 $ charges : num  16885 1726 4449 21984 3867 …
```

[5]:
```
# it looks the factor number for smoker is not appropiate.
head(data$smoker)
```

1. yes 2. no 3. no 4. no 5. no 6. no

Levels: 1. 'no' 2. 'yes'

[6]:
```
# fix the data type
data$smoker = as.logical(as.numeric(as.character(factor(data$smoker,levels␣
↪=c("no","yes"), labels = c(0,1)))))
str(data)
```

```
'data.frame':	1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 …
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 …
 $ bmi     : num  27.9 33.8 33 22.7 28.9 …
 $ children: int  0 1 3 0 0 0 1 3 2 0 …
 $ smoker  : logi  TRUE FALSE FALSE FALSE FALSE FALSE …
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2
…
 $ charges : num  16885 1726 4449 21984 3867 …
```

[7]:
```
head(data$smoker)
```

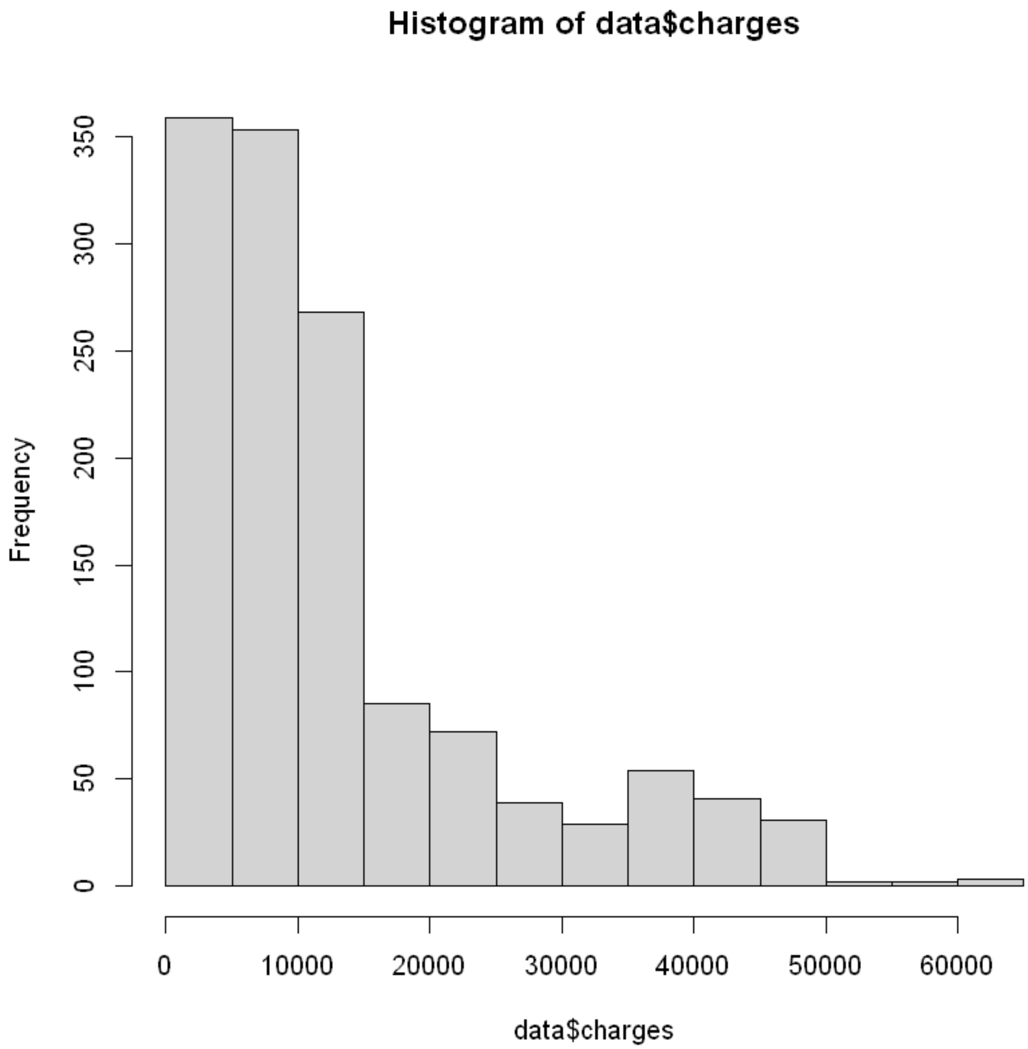
1. TRUE 2. FALSE 3. FALSE 4. FALSE 5. FALSE 6. FALSE

4 Descriptive Statistics

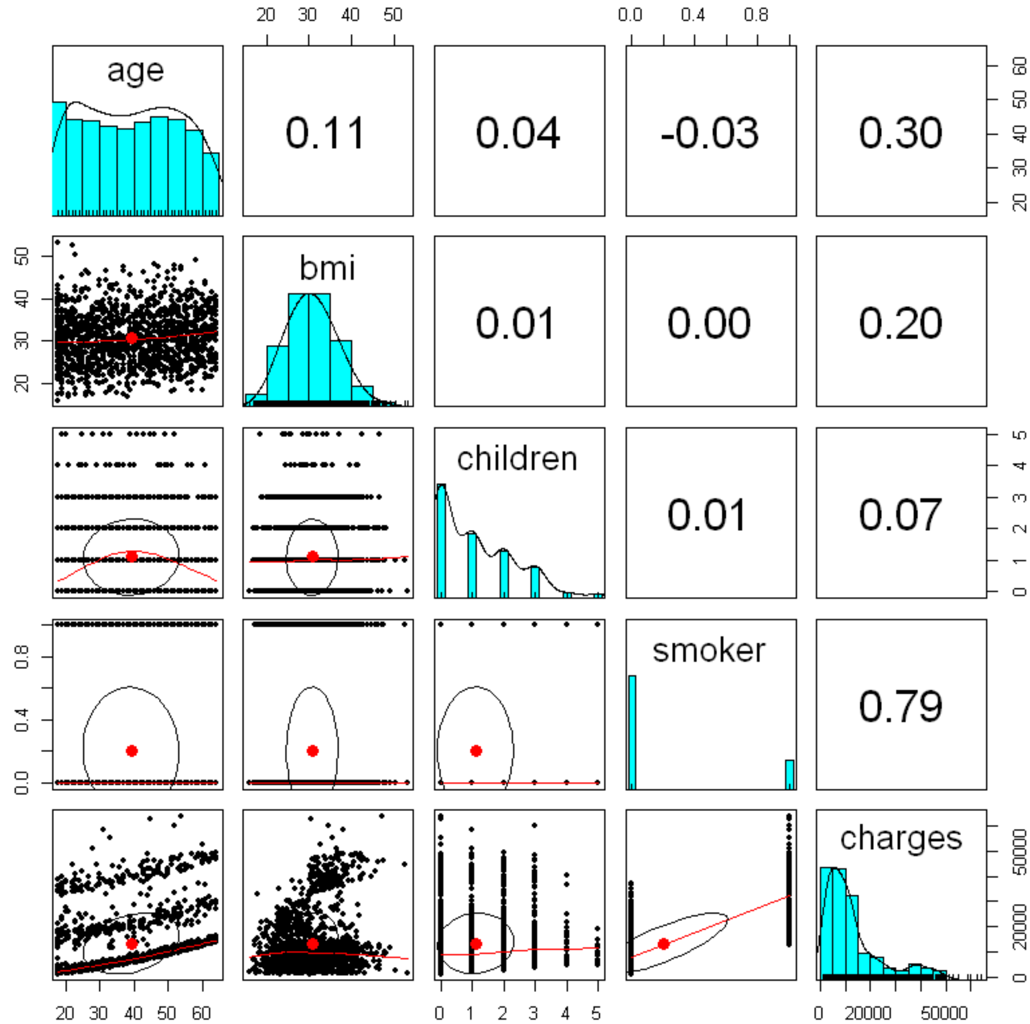
[8]:
```r
summary(data)
```

```
      age            sex           bmi           children      smoker
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   Mode :logical
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   FALSE:1064
 Median :39.00                Median :30.40   Median :1.000   TRUE :274
 Mean   :39.21                Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                Max.   :53.13   Max.   :5.000
       region       charges
 northeast:324   Min.   : 1122
 northwest:325   1st Qu.: 4740
 southeast:364   Median : 9382
 southwest:325   Mean   :13270
                 3rd Qu.:16640
                 Max.   :63770
```

[9]:
```r
hist(data$charges)
```



[10]:
```
# install.packages("psych")  # if it is not ready
library(psych)
pairs.panels(data[c("age","bmi","children","smoker","charges")])
```



5 Modelling – Regression

5.1 Train-Test Splitting for Data

[11]:
```
# Total observations
m = nrow(data)
m
```

1338

[12]:
```
# randomize the data
set.seed(1000)
rand = sample(m)
```

```
rData = data[rand,]
```

[13]:
```
# size of training data
(nTrain = round(m * 0.9))
# size of testing data
(nTest = m - nTrain)
```

1204

134

[14]:
```
# Assign training and testing data from randomized data
trainData =  rData[1:nTrain, ]
testData =  rData[(nTrain+1):m, ]
```

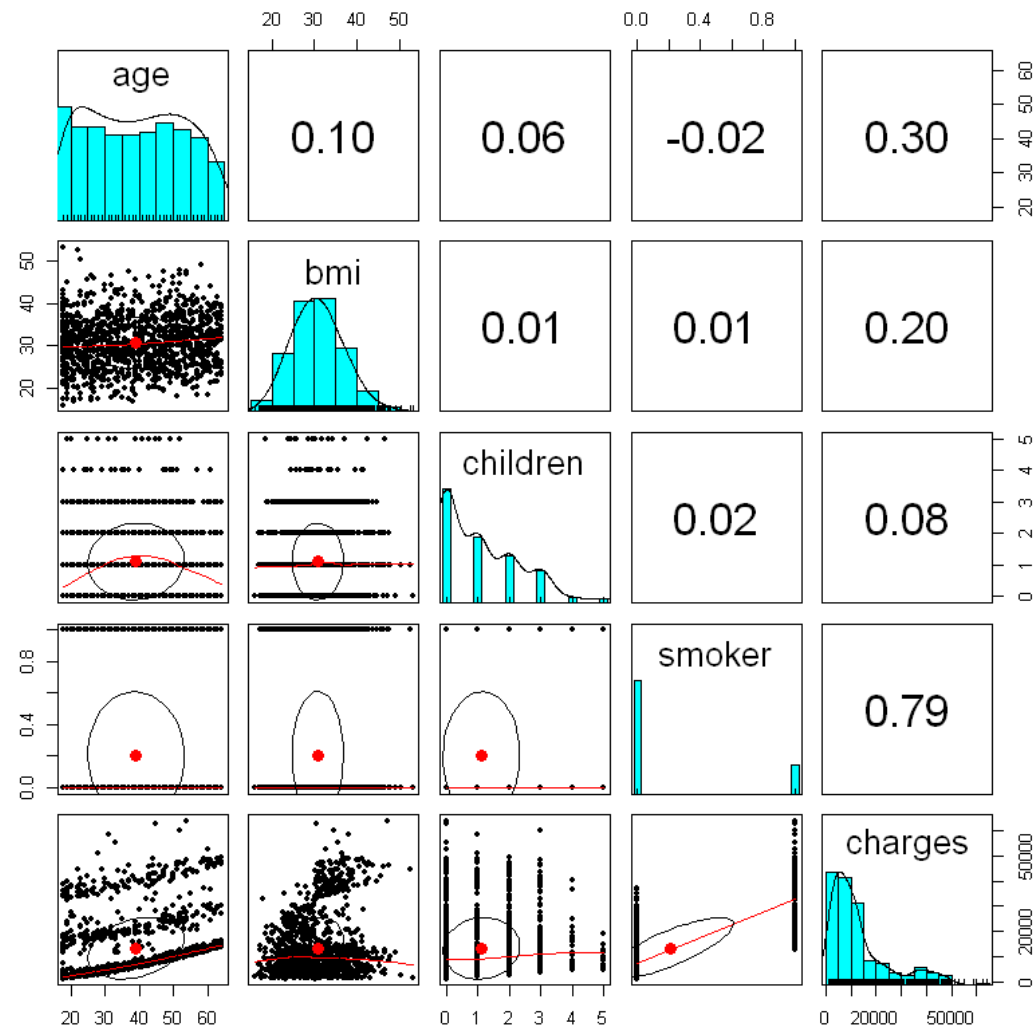
[15]:
```
str(trainData)
```

```
'data.frame':	1204 obs. of  7 variables:
 $ age     : int  47 25 62 34 60 63 25 47 52 47 …
 $ sex     : Factor w/ 2 levels "female","male": 2 1 2 1 1 2 1 2 1 2 …
 $ bmi     : num  28.2 23.5 38.8 26.7 18.3 …
 $ children: int  3 0 0 1 0 0 0 1 2 3 …
 $ smoker  : logi  TRUE FALSE FALSE FALSE FALSE FALSE …
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 2 1 3 3 1 1 4 3 1 2
…
 $ charges : num  24915 3206 12981 5003 13204 …
```

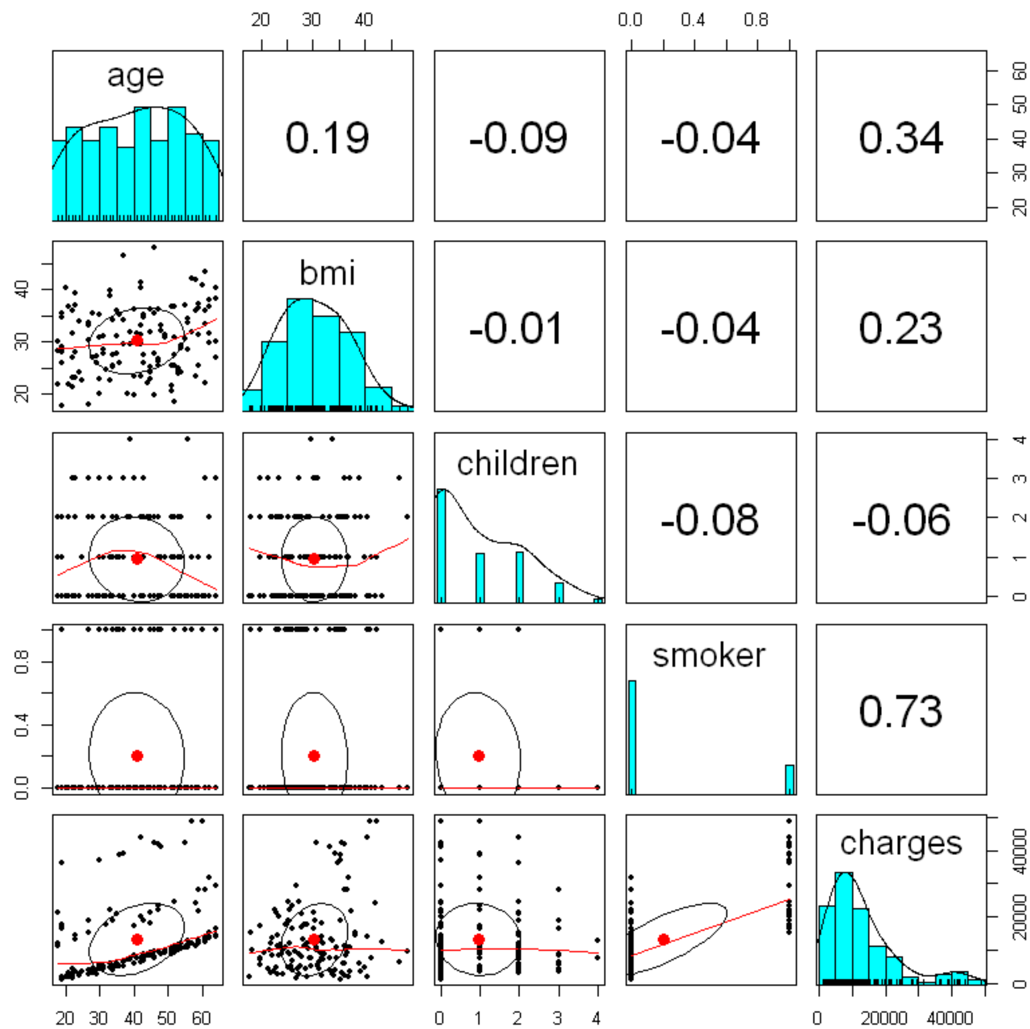
[16]:
```
str(testData)
```

```
'data.frame':	134 obs. of  7 variables:
 $ age     : int  62 37 45 47 42 46 21 53 64 37 …
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 1 2 1 2 2 1 2 …
 $ bmi     : num  36.9 29.6 24 26.6 31.3 …
 $ children: int  1 0 2 2 0 0 0 0 3 1 …
 $ smoker  : logi  FALSE FALSE FALSE FALSE FALSE FALSE …
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 1 2 1 1 2 2 2 2 2 1
…
 $ charges : num  31620 5028 8604 9716 6359 …
```

[17]:
```
pairs.panels(trainData[c("age","bmi","children","smoker","charges")])
```



[18]:
```
pairs.panels(testData[c("age","bmi","children","smoker","charges")])
```



5.2 Features Selection

[19]:
```
fullModel =  lm(formula = charges ~ ., data = trainData)
summary(fullModel)
```

```
Call:
lm(formula = charges ~ ., data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max
-11614.8  -2747.7   -927.3   1397.8  29753.4
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11872.52    1046.38 -11.346  < 2e-16 ***
age                255.99      12.56  20.378  < 2e-16 ***
sexmale           -124.62     351.77  -0.354 0.723210    
bmi                338.19      30.30  11.163  < 2e-16 ***
children           477.09     144.27   3.307 0.000971 ***
smokerTRUE       24213.76     435.97  55.540  < 2e-16 ***
regionnorthwest   -343.62     507.84  -0.677 0.498768    
regionsoutheast  -1164.02     509.37  -2.285 0.022476 *  
regionsouthwest  -1035.77     506.16  -2.046 0.040943 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6074 on 1195 degrees of freedom
Multiple R-squared:  0.7556,	Adjusted R-squared:  0.754 
F-statistic: 461.9 on 8 and 1195 DF,  p-value: < 2.2e-16
```

Select those features with significant level, p <= 0.01 only. We may consider the dummy variables in region. to simply the introduction, we do not discuss at this moment.

[20]:
```
reduceModel =  lm(formula = charges ~ age + bmi + children + smoker , data =
 ↪trainData)
summary(reduceModel)
```

```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max 
-12246.6  -2820.3   -951.9   1366.5  29202.5 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12025.20    1000.53 -12.019  < 2e-16 ***
age            256.88      12.56  20.456  < 2e-16 ***
bmi            318.70      29.05  10.969  < 2e-16 ***
children       476.15     144.22   3.302  0.00099 ***
smokerTRUE   24180.23     434.31  55.675  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6082 on 1199 degrees of freedom
Multiple R-squared:  0.7542,	Adjusted R-squared:  0.7534 
F-statistic: 919.6 on 4 and 1199 DF,  p-value: < 2.2e-16
```

Now we can confirm all features are statistical significant.

[21]:
```
anova(fullModel,reduceModel)
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1195	44092467572	NA	NA	NA	NA
2	1199	44358476853	-4	-266009281	1.802355	0.1259959

If P < 0.05, we reject H0: reduced model = full model, or assert Ha: reduced model != full model. In this case, P >0.05. we do not reject H0. If H0: reduced model = full model is favored, we choose the less variable one.

6 Prediction / test /evaluation

[22]:
```
str(testData)
```

```
'data.frame':   134 obs. of  7 variables:
 $ age     : int  62 37 45 47 42 46 21 53 64 37 …
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 1 2 1 2 2 1 2 …
 $ bmi     : num  36.9 29.6 24 26.6 31.3 …
 $ children: int  1 0 2 2 0 0 0 0 3 1 …
 $ smoker  : logi  FALSE FALSE FALSE FALSE FALSE FALSE …
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 1 2 1 1 2 2 2 2 2 1
…
 $ charges : num  31620 5028 8604 9716 6359 …
```

[23]:
```
X = testData[,-7]
```

[24]:
```
Y = testData[,7]
```

[25]:
```
Yp = predict(reduceModel, X)
```

[26]:
```
#rmse: root mean square error
RMSE = sqrt(mean((Y-Yp)^2))
RMSE
```

5954.53393416666

[27]:
```
data.frame(Y, Yp, abs(Y-Yp))
```

A data.frame: 134 × 3

	Y	Yp	abs.Y...Yp.
	<dbl>	<dbl>	<dbl>
574	31620.001	16124.748	15495.25350
165	5028.147	6925.631	1897.48391
741	8604.484	8146.614	457.86955
276	9715.841	9477.844	237.99716
949	6358.776	8724.722	2365.94530
89	8026.667	8631.984	605.31732
788	1917.318	5116.645	3199.32710
684	9863.472	9340.153	523.31840
379	16455.708	15441.148	1014.56028
948	39047.285	33035.308	6011.97728
1181	7650.774	12410.225	4759.45164
857	40974.165	35037.412	5936.75272
537	5972.378	10277.782	4305.40441
456	21797.000	15050.064	6746.93595
1091	41676.081	35762.144	5913.93697
645	18806.145	11226.250	7579.89577
700	3500.612	7350.782	3850.16964
486	4347.023	5838.523	1491.49972
381	15006.579	25765.344	10758.76430
714	1984.453	6010.291	4025.83734
766	11842.624	12759.535	916.91148
1316	11272.331	2097.248	9175.08383
832	5266.366	5457.678	191.31236
893	10422.917	9506.200	916.71712
228	24227.337	16230.543	7996.79391
626	3736.465	3720.094	16.37061
188	5325.651	6957.519	1631.86784
765	9095.068	8509.937	585.13143
795	7209.492	9118.335	1908.84362
865	8782.469	9170.600	388.13126
...	...	...	...
661	6435.624	13737.002	7301.37856
444	28287.898	15245.758	13042.13994
817	2842.761	1860.449	982.31168
1120	5693.431	3467.708	2225.72203
1226	4795.657	9618.684	4823.02758
92	10942.132	9967.691	974.44126
535	13831.115	17316.057	3484.94205
636	14410.932	16586.225	2175.29266
818	3597.596	7135.347	3537.75087
61	8606.217	9168.705	562.48728
558	3935.180	7611.478	3676.29812
17	10797.336	11618.258	820.92208
321	4894.753	5238.416	343.66260
382	42303.692	36062.692	6240.99986
547	3268.847	6460.630	3191.78310
1326	13143.337	14332.027	1188.68990
840	12622.180	13121.904	499.72488
430	18804.752	6027.536	12777.21606
1291	7133.903	5046.570	2087.33249
1027	16450.895	25807.712	9356.81762

7 Deloyment

Normally, we compare different ML models and choose the best one with settting the best hyperparameters. To simply the illustration, we assume that the model is accepted and is going to deploy for further use.

8 Reviews

8.1 Please give the orders for the steps in the advanced analytics!.

- ( )Prescriptive Analytics
- ( )Diagnostic Analytics
- ( )Descriptive Analytics
- ( )Predictive Analytics

8.2 Please give the order according to pyramid.

- ( ) Information
- ( ) Data
- ( ) wisdom
- ( ) knowledge

8.3 Please rank the size of the scope: smaller number means smaller scope.

- ( ) Deep Learning
- ( ) Artificial Intelligence
- ( ) Machine Learning

8.4 Please give the orders according to KDD process.

- ( ) selection
- ( ) Preprocessing
- ( ) Transformation
- ( ) Data Mining
- ( ) Interpretation / Evaluation

9 Solutions

9.1 Please give the orders for the steps in the advanced analytics.

- (4)Prescriptive Analytics
- (2)Diagnostic Analytics
- (1)Descriptive Analytics
- (3)Predictive Analytics

9.2 Please give the orders according to pyramid.

- (2) Information
- (1) Data

- (4) wisdom
- (3) knowledge

9.3 Please rank the sizes of the scopes: smaller number means smaller scope.

- (1) Deep Learning
- (3) Artificial Intelligence
- (2) Machine Learning

9.4 Please give the orders according to the KDD process.

- (1) selection
- (2) Preprocessing
- (3) Transformation
- (4) Data Mining
- (5) Interpretation / Evaluation

[ ]:

```

```