COMP9321 Assignment3 Report

Name: Qiyao Zhou

zId: z5379852

1.Data explore and preprocess

The first step to take is to get a general idea of the data set. The information collected on the dataset shows that there are no null values in the dataset, so there is no need for default value processing, while the columns in the dataset are of type int64 or object, and the data of type object needs to be uniquely coded, which retains the information of the categorical variables while avoiding the size relationship between the numerical variables and also improves the generalization ability of the model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150211 entries, 0 to 150210
Data columns (total 13 columns):
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   Number_of_Shops_Around_ATM                150211 non-null  int64
 1   ATM_Zone                                  150211 non-null  object
 2   No_of_Other_ATMs_in_1_KM_radius           150211 non-null  int64
 3   Estimated_Number_of_Houses_in_1_KM_Radius 150211 non-null  int64
 4   ATM_Placement                             150211 non-null  object
 5   ATM_TYPE                                  150211 non-null  object
 6   ATM_Location_TYPE                         150211 non-null  object
 7   ATM_looks                                 150211 non-null  object
 8   ATM_Attached_to                           150211 non-null  object
 9   Average_Wait_Time                         150211 non-null  int64
 10  Day_Type                                  150211 non-null  object
 11  rating                                    150211 non-null  int64
 12  revenue                                   150211 non-null  int64
dtypes: int64(6), object(7)
```

After the unique heat coding, the dataset has 34 columns, excluding "revenue" for regression and "rating" for classification, there are still 32 columns, based on common sense analysis it is clear that these columns are correlated with both revenue and rating. Therefore, I decided to use feature selection to rank the relevance of each feature to the predicted attributes and select the features with the highest relevance for model building. In the regression task, I used f_regression for correlation analysis, while in the classification task, I chose f_classif. As for the number of features, in the subsequent model tuning tests, the number of features for the two parts was finally set at 16 and 25 respectively, taking into account the running time of the program and the accuracy of the model.

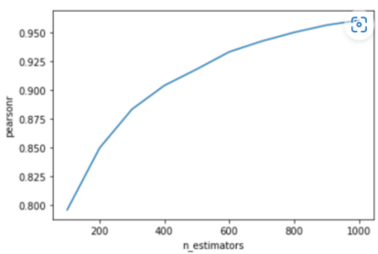
Finally, for regression problems, it is also important to normalize the features to improve both the accuracy and stability of the model, as well as to speed up the convergence of the model, while the target values can be reduced by using the logarithm method to obtain better model results. These pre-processes are not needed and cannot be performed in classification problems.

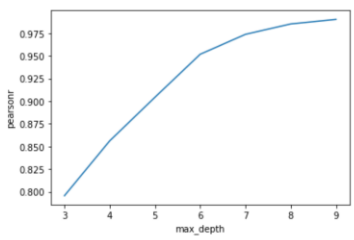
2. Model selection and modulation

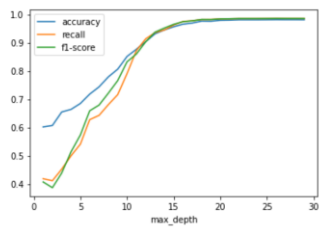
Before formally starting the model construction, the original training data also needs to be divided in order to facilitate the testing of the resulting model. I decided to divide the data in train.tsv into a training set and a test set in a 4:1 ratio.

Firstly, for the regression problem, I tried SGD and AdaBoost, but the results were not very satisfactory, so I finally used GradientBoost, while for the classification problem, the decision tree algorithm model I used for the first time achieved satisfactory results, so I chose the decision tree to complete the classification model.

In terms of the choice of model parameters, after several tests (below), the GradientBoostingRegressor parameters were finally chosen as n_estimators =1000, max_depth=7 and the decision tree model was only set to criterion='entropy'.







3. Training results and analysis

After the above analysis and model construction, we finally obtained the required regression and classification models, with the Pearson coefficient of the former reaching 0.99 in the divided test set, while the accuracy of the latter, recall and f1-score, both exceeded 0.98.

	feature	importance
0	Estimated_Number_of_Houses_in_1_KM_Radius	0.466546
1	No_of_Other_ATMs_in_1_KM_radius	0.201806
2	Average_Wait_Time	0.079700
3	ATM_Zone_RM	0.060932
4	ATM_Attached_to_Petrol Bunk	0.034209
5	ATM_Zone_C	0.029840
6	ATM_TYPE_Urban	0.026322
7	ATM_Zone_FV	0.020826
8	ATM_Location_TYPE_Passbook Printing and Withdraw	0.018584
9	Day_Type_Working	0.012562
10	ATM_TYPE_Town	0.012350
11	ATM_Attached_to_Building	0.009864
12	Day_Type_Festival	0.009272
13	ATM_Zone_RL	0.008943
14	ATM_Location_TYPE_Deposit and Withdraw	0.006229
15	ATM_TYPE_Semi Urban	0.002014

	feature	importance
0	Estimated_Number_of_Houses_in_1_KM_Radius	0.406050
1	No_of_Other_ATMs_in_1_KM_radius	0.224377
2	Number_of_Shops_Around_ATM	0.117402
3	Average_Wait_Time	0.070201
4	ATM_Attached_to_Petrol Bunk	0.032484
5	ATM_Zone_FV	0.026503
6	ATM_Zone_RL	0.019096
7	Day_Type_Working	0.016967
8	ATM_TYPE_Urban	0.014482
9	ATM_TYPE_Town	0.012030
10	ATM_looks_Normal	0.009955
11	ATM_looks_New	0.008450
12	ATM_Attached_to_Building	0.008076
13	ATM_Location_TYPE_Only WIthdraw	0.006992
14	ATM_Location_TYPE_Deposit and Withdraw	0.006656
15	ATM_Location_TYPE_Checkdrop and Withdraw	0.005598
16	ATM_Location_TYPE_Passbook Printing and Withdraw	0.004654
17	ATM_TYPE_Semi Urban	0.003874
18	ATM_Zone_RM	0.002860
19	ATM_Zone_RH	0.001414
20	ATM_Zone_C	0.001284
21	ATM_Placement_Facing Road	0.000596