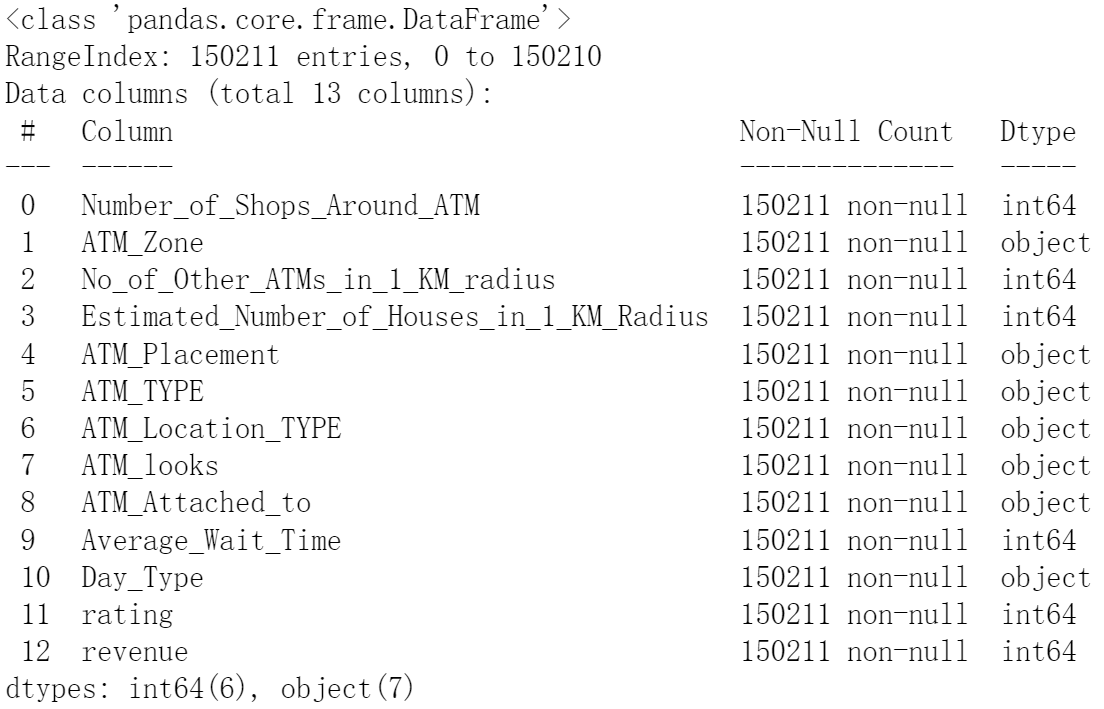
**COMP9321 Assignment3 Report**

**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.



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.

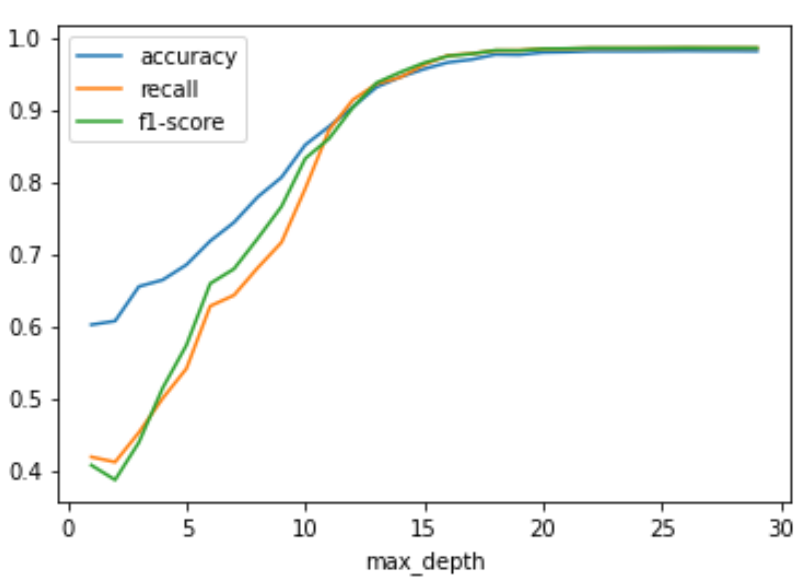
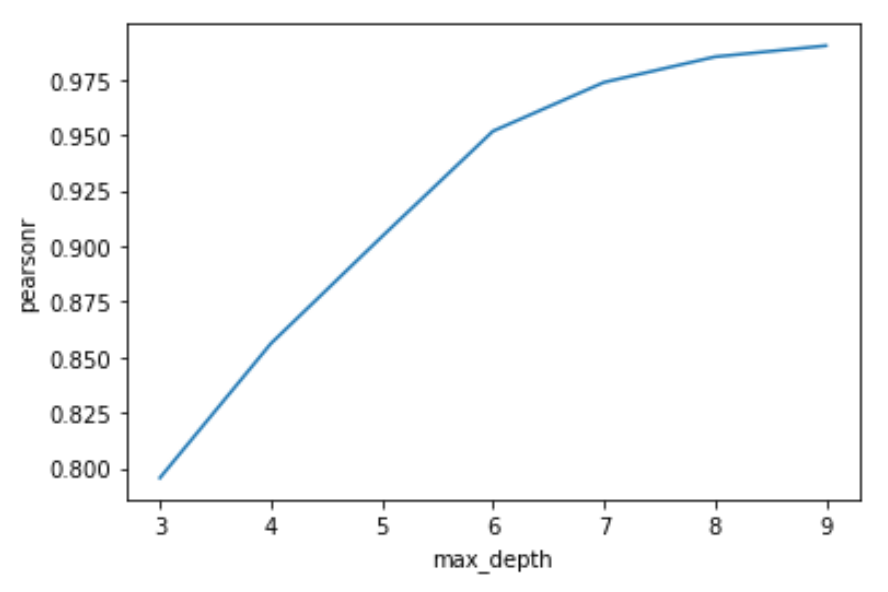
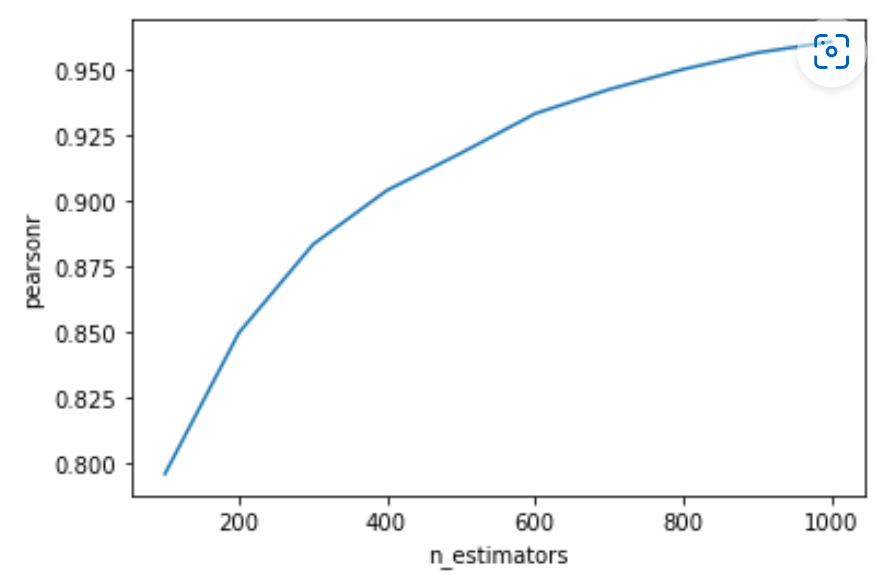
Finally, for regression problems, it is also important to normalize the features to improve both the accuracy and stability of the model, as well as to speed up the convergence of the model, while the target values can be reduced by using the logarithm method to obtain better model results. These pre-processes 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.

