Regression Analysis

- At first I handled object columns. I did one hot encoding for each, except day(0, …, 6)(we don't have Mon, Tue and Wed=> it is not good to do one hot on this column). We don't have any nan values. I also plotted tips distribution and correlation matrix.
- The data is too small. So expecting much from this is not real.
- I'll use the Mean squared error in most of the models as it is standard for regression model evaluation, also there might be additional losses and metrics for each model.
- First I do a simple linear regression. As we look at coefficients and the correlation matrix we can see a clear relation. Here I also calculated the R2 coefficient of determination(`0.4429399687489898`) and as it is not closer to 1 this means that linear regression is not a good estimator. MSE score for Test was `0.6963090766605349` but the training MSE is `1.1041697586041952` and training R2 is `0.45653112545009544` which means that in MSE term we are underfitting but in R2 term the model is ok.
- Regression tree was done using gridsearch in order to get the best hyperparameters(here loss was taken negative MSE as the gridsearch function maximizes the score but we need minimal MSE). The MSE score for Test was `1.0689216741175624` worse than for linear regression.
- Random forest regression tree was done using gridsearch in order to get the best hyperparameters(here loss was taken negative MSE as the gridsearch function maximizes the score but we need minimal MSE). The MSE score for Test was `0.8857390625671349` better than the regression tree, but still worse than the linear regression.
- Gradient boosting regression tree was done using gridsearch in order to get the best hyperparameters(here loss was taken negative MSE as the gridsearch function maximizes the score but we need minimal MSE). The MSE score for Test was `0.7344606040110865`
- I think that these 4 models are enough to see that we don't need more complicated algorithms as the data has a poor quality and it's time to play with some metrics and losses.
- RMSE is the square root of MSE so it won't give any additional info, that is why I won't use it.

- As MSE sums the squares of the difference of t_true and y_pred => it can be sensitive to outliers, so it is worth checking MAE as well. After training with negative_mae(gridsearch maximizes but we need minimal MAE). The results are the same linear regression is the best on test data but undefits, other models don’t underfit but give poorer results on test data then linear regression.
- Considering the results I would choose Gradient boosting as it doesn’t underfit and give the best results among non underfitting models.