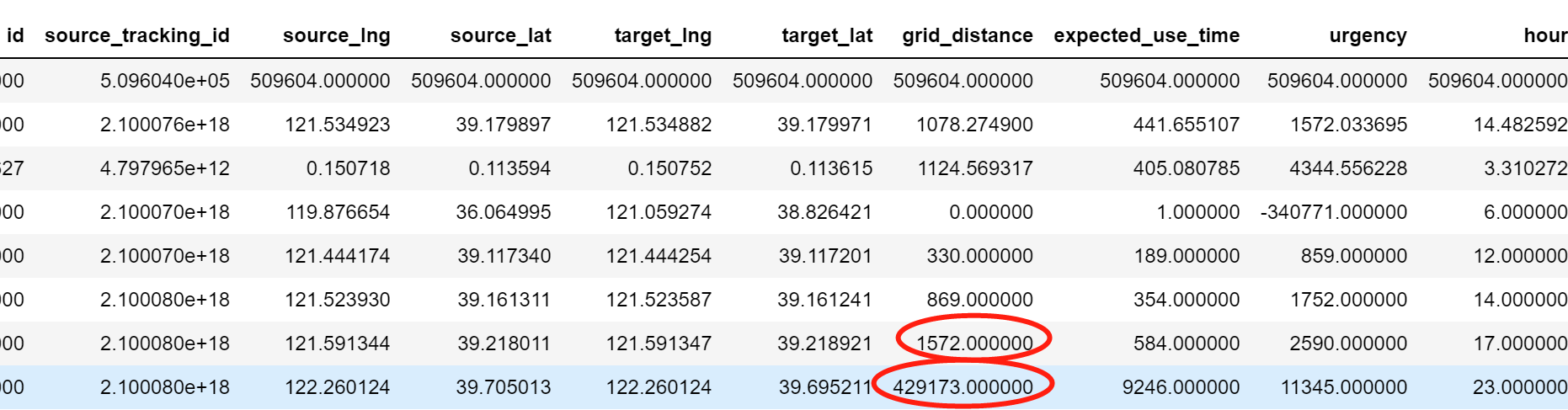
**Regression**

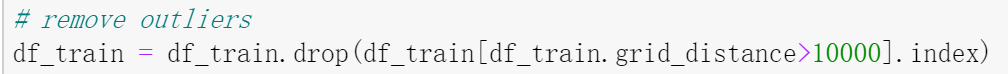
**Question(a)**

**Data Pre-Processing**

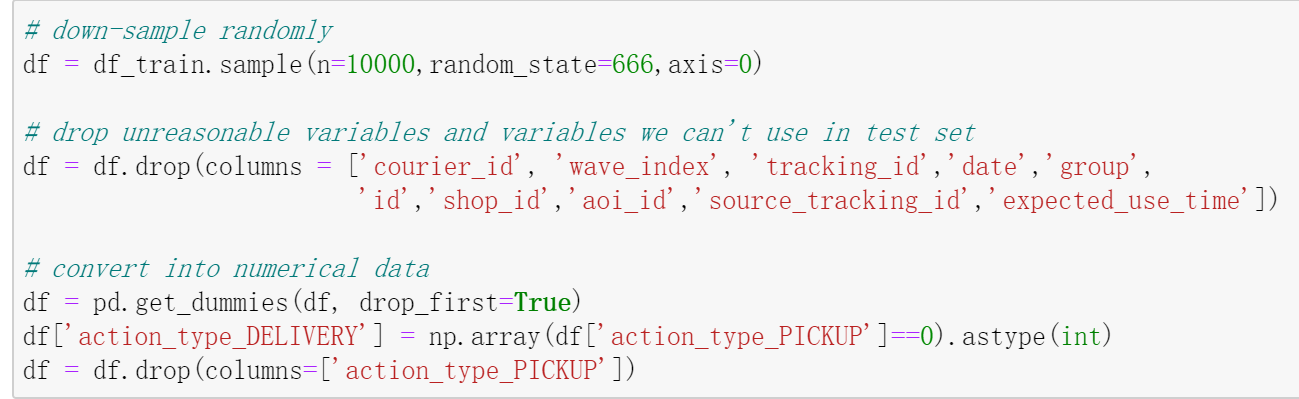
First, we described the data to see low variance variables and whether there exist some outliers:



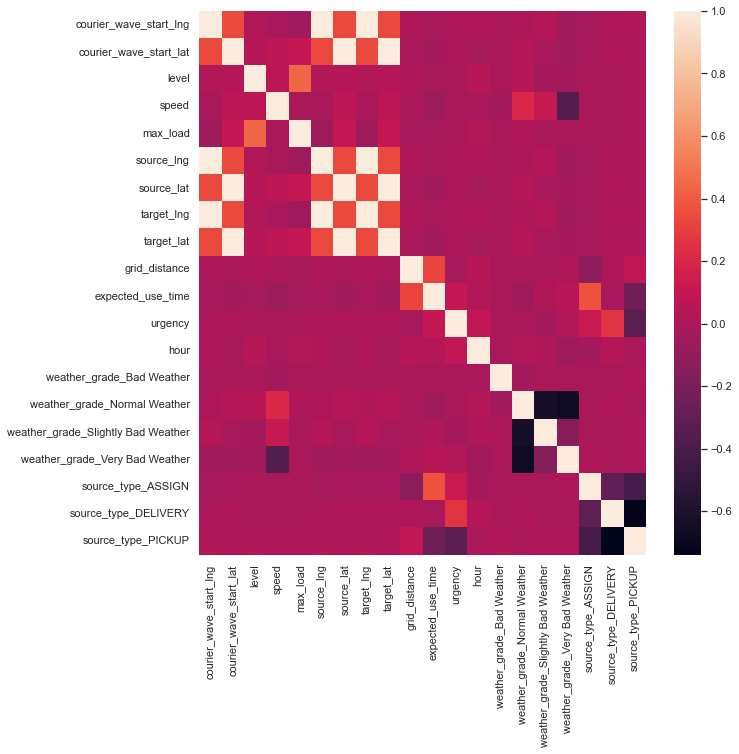
We found that grid\_distance has some outliers because 75% quantile is 1572 whereas its max value is 429173, which is approximate 300 times of 75% quantile. Therefore, we removed those observations by counting the number of values which more than 10000 meters, 20000 meters and decided deleting observations whose grid\_distance is more than 10000 meters.



Second, we down-sampled the training data set into 10000 and dropped some unreasonable variables like id and some variables we cannot use to predict the testing data set. Then, we converted some discrete variables into numerical data by getting dummies. Incidentally, we generated our labels which represents action type is pick-up when label equals 1.



Third, we standardized the data sample and ran a heat map to reveal the correlations between all the variables to see whether independent variables have strong correlations and which independent variables have correlations with label.

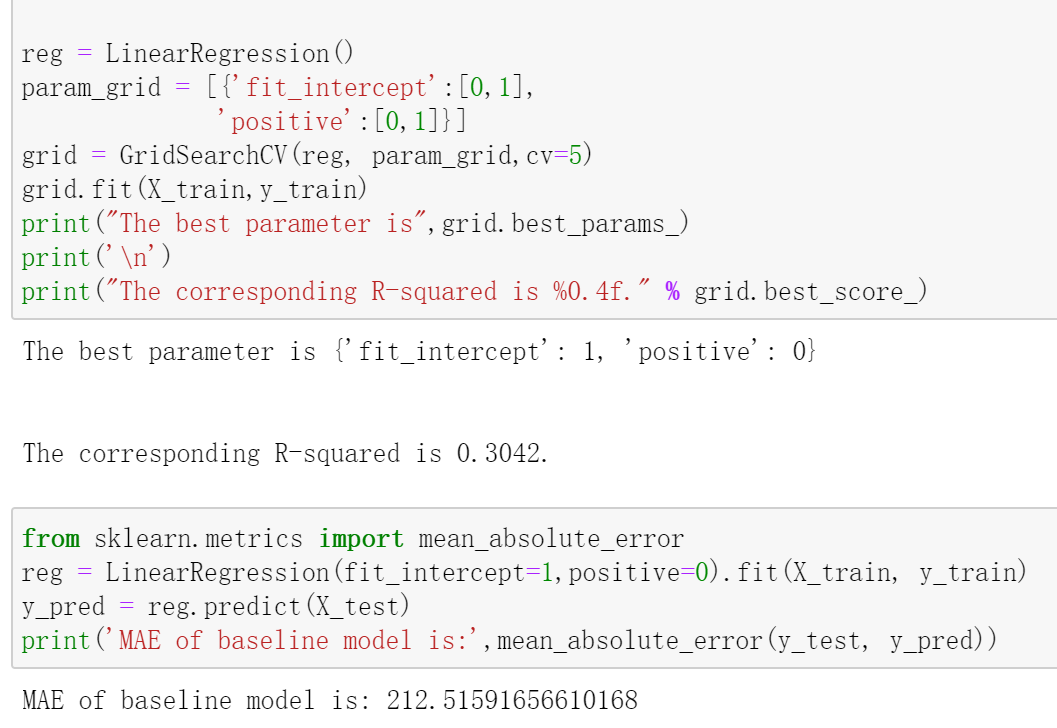


According to the figure above, we chose the following variables as our features for now:

speed', 'grid\_distance','urgency','weather\_grade\_Normal Weather', 'weather\_grade\_Very Bad Weather', 'source\_type\_ASSIGN','source\_type\_PICKUP'

**Question(b)**

**Baseline Model**

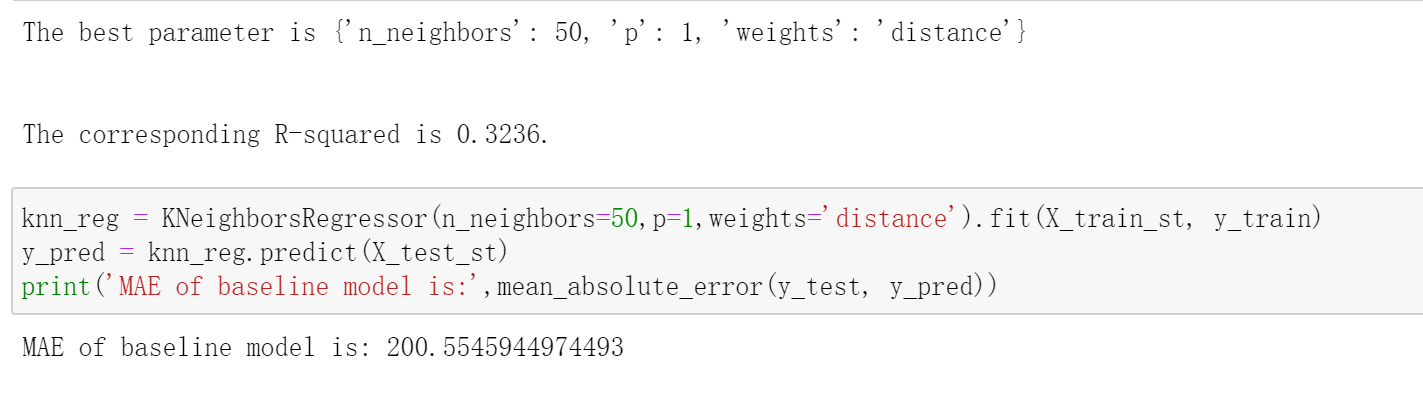


We used GridSearchCV to cross validate the linear regression model which is our baseline model. The MAE is 212.5, which is a relatively high error.

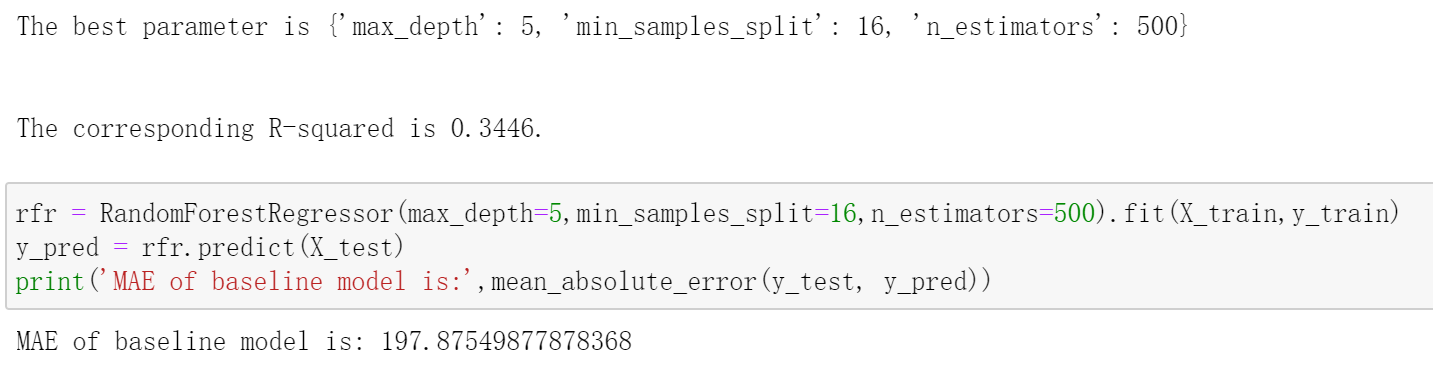
**Question(c)**

**More Complex Models**

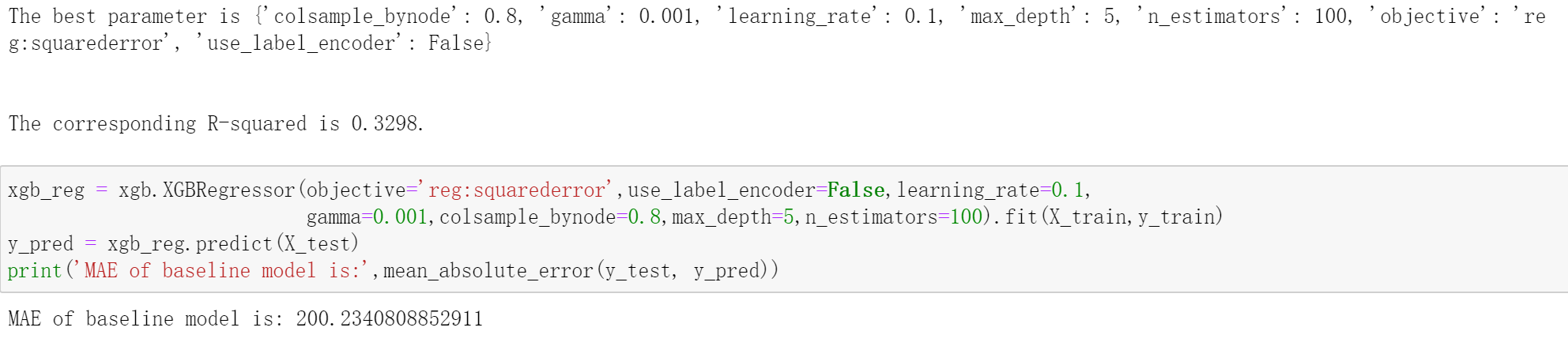
K-NN:



Random Forest:



XGBT:



After trying more complex models, we found that MAEs dropped to 200. And the three models above have approximately same results. So, we still chose XGB as our final model.

**Question(d)**

**Feature Engineering**

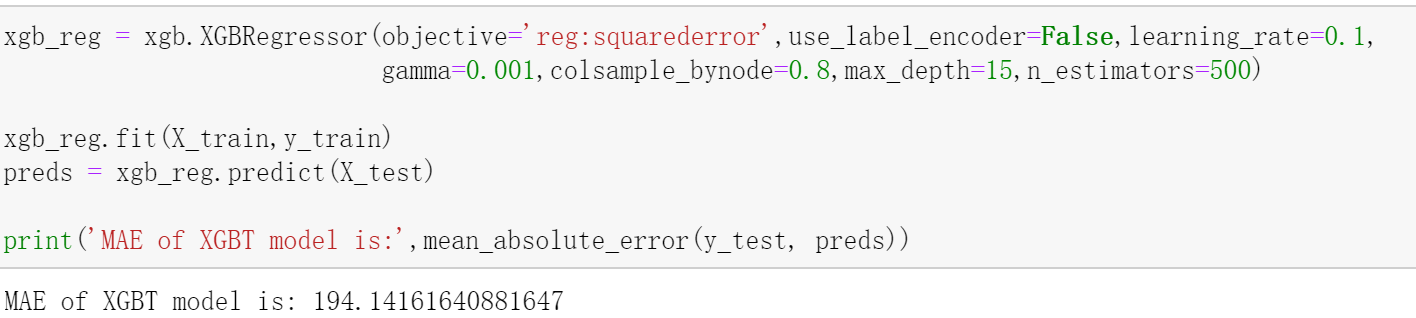
First, we polynomial features and found degree=1 is the best.



Second, we conducted feature selection by RFECV using all the variables except id. This do help us to find some useful features but I still added some other features to make the MAE smaller. As a result, we chose the following features:

'grid\_distance','urgency','source\_type\_DELIVERY','source\_type\_PICKUP','speed','target\_lng','target\_lat','source\_lng','source\_lat','courier\_wave\_start\_lng','courier\_wave\_start\_lat','hour','max\_load'

Third, we split the whole training data set into train set and test set and refitted the model using the recommended features.



Finally, the MAE dropped to 194. We believe this would drop below 190 after fitting all the 500000 observations.

**Question(e)**

We used the whole training set to fit our model and made the parameters of XGBT better and predicted our testing set. We found a mistake in test data set because the tracking id and source\_type were swapped.

According to our model, source\_type, grid\_distance and urgency are the most important features in predicting the expected use time. Weather condition seemed no effects on courier’s action type.