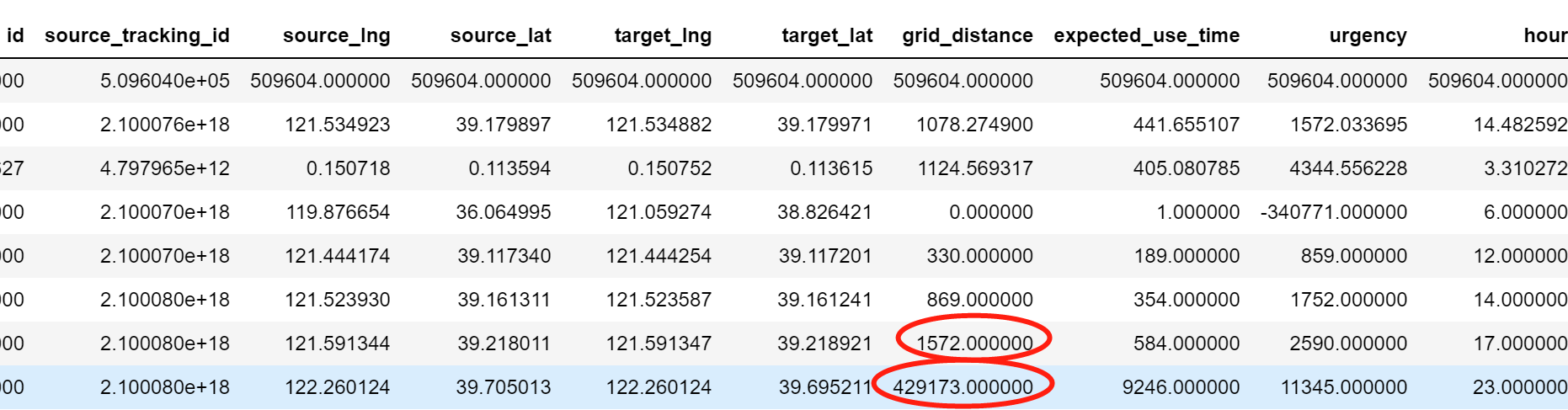
**Classification**

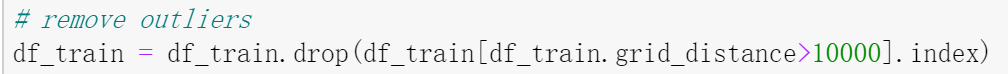
**Question(a)**

**Data Pre-Processing**

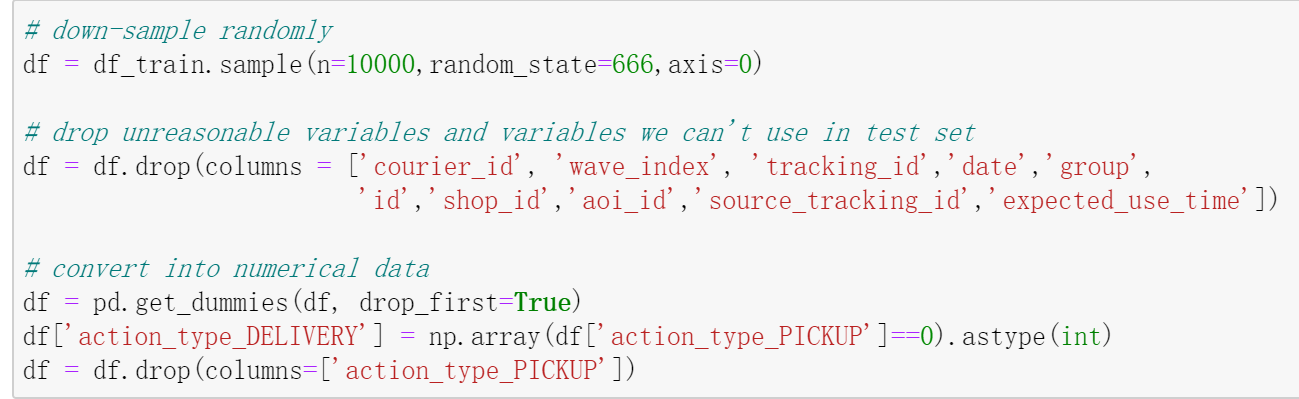
First, we described the data to see low variance variables and whether there exist some outliers:



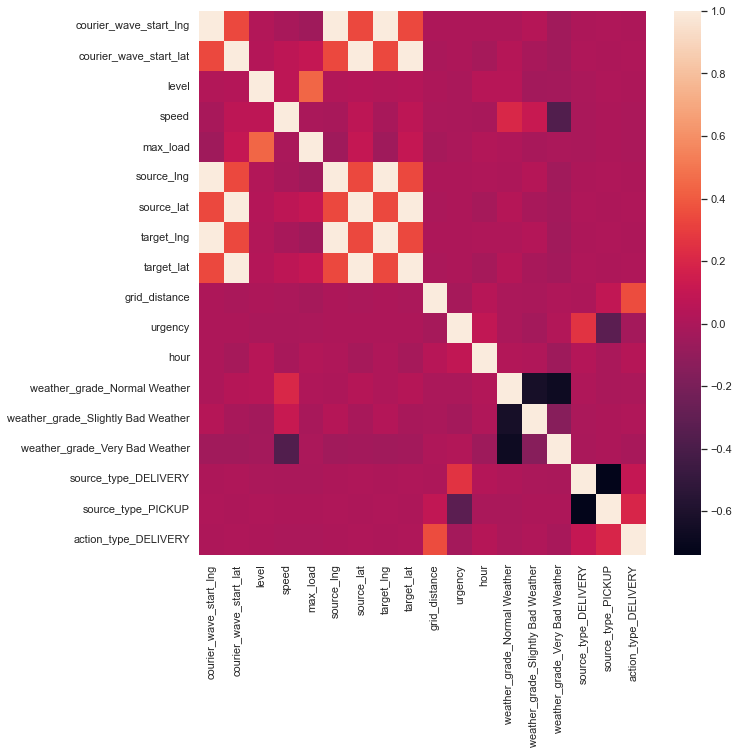
We found that grid\_distance has some outliers because 75% quantile is 1572 whereas its max value is 429173, which is approximate 300 times of 75% quantile. Therefore, we removed those observations by counting the number of values which more than 10000 meters, 20000 meters and decided deleting observations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

(courier\_wave\_start\_lng,source\_lng,target\_lng), (courier\_wave\_start\_lat, source\_lat, target\_lat)

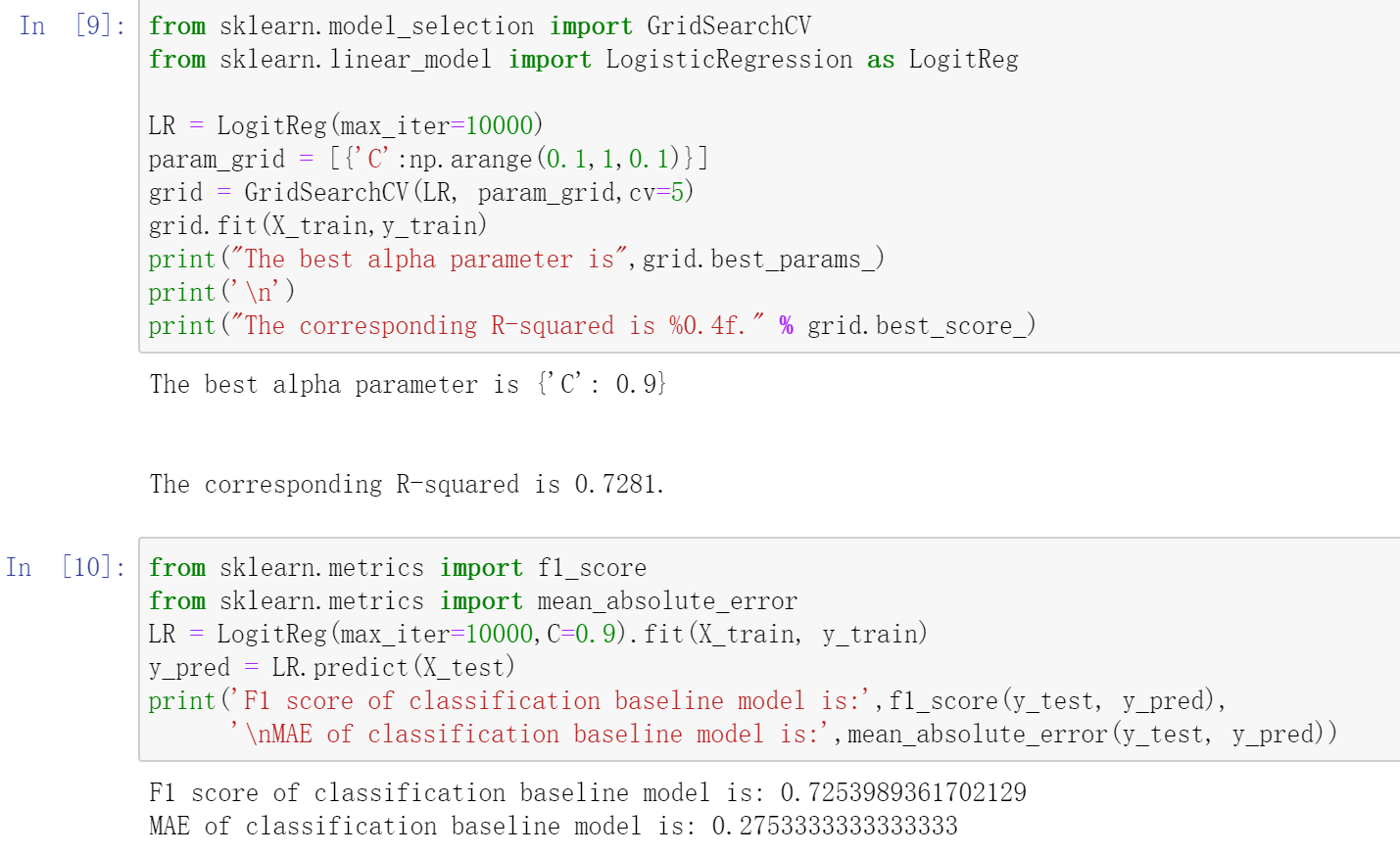
Additionally, the following variables are related to action\_type\_PICKUP more than 1%:

(urgency, grid\_distance, source\_type\_PICKUP, source\_type\_DELIVERY, hour, level)

Therefore, we chose them tentatively as our independent variables in our baseline model and in exploring models.

**Question(b)**

**Baseline Model**

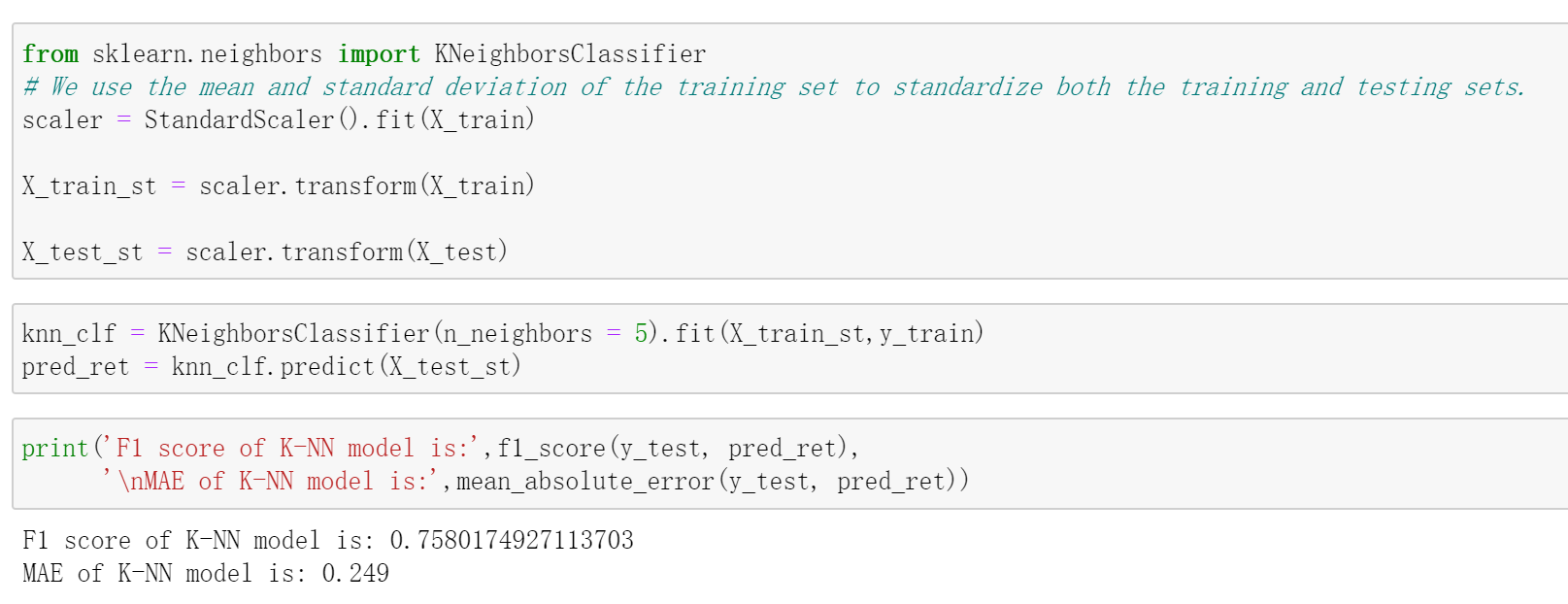


We used GridSearchCV to cross validate the logistic regression model which is our baseline model. We found that when C parameter equals 0.9, the corresponding R-squared is the highest, 0.7281. Then, we refitted the model with C=0.9 and got the F1 score of 0.725. This is not bad but still not good enough. As a result, we need to try more models and feature engineering.

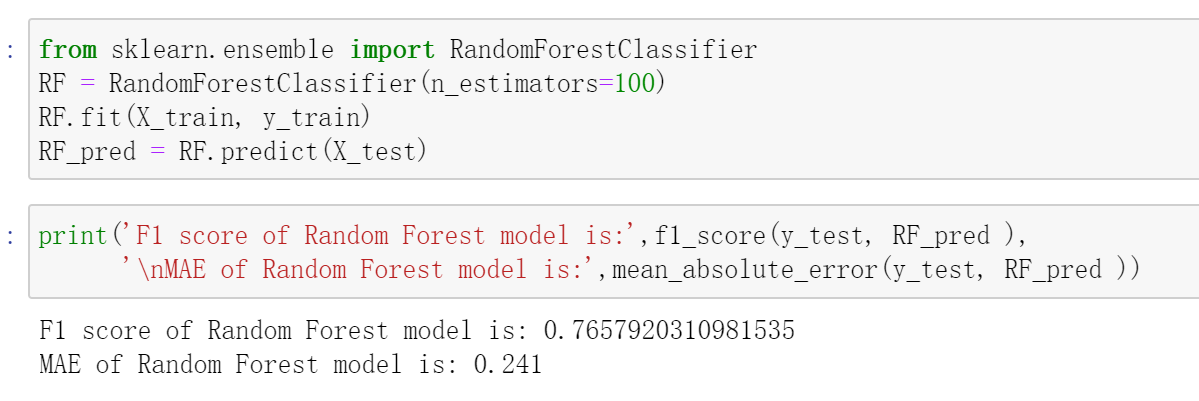
**Question(c)**

**More Complex Models**

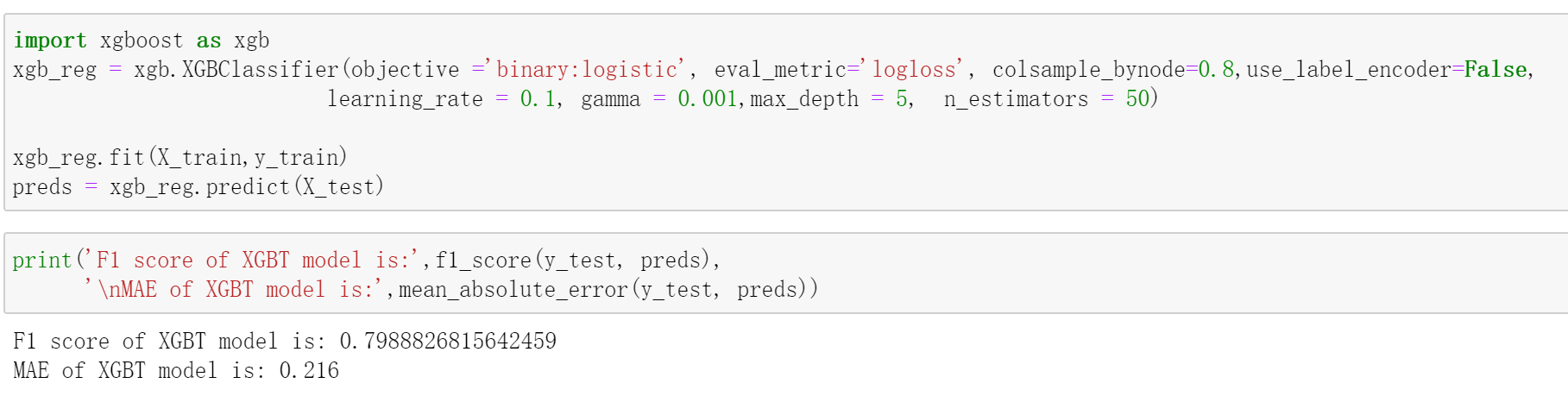
K-NN:



Random Forest:



XGBT:

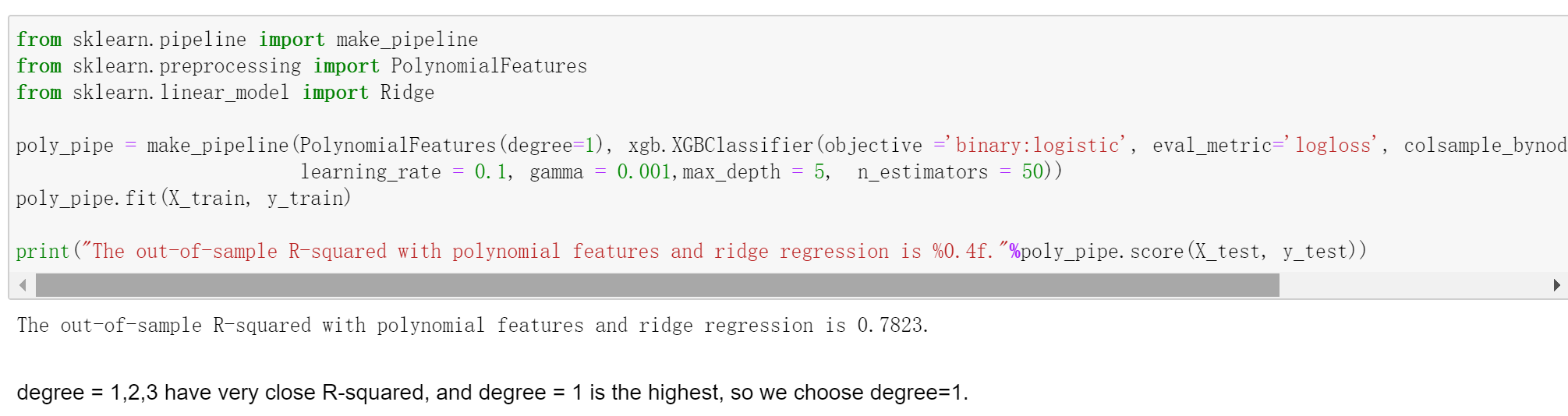


After trying different models, the F1 score improved significantly. Among these models, XGBT has a F1 score reaching 0.8. As a result, we chose XGBT as our final model in feature engineering and predicting our test data set.

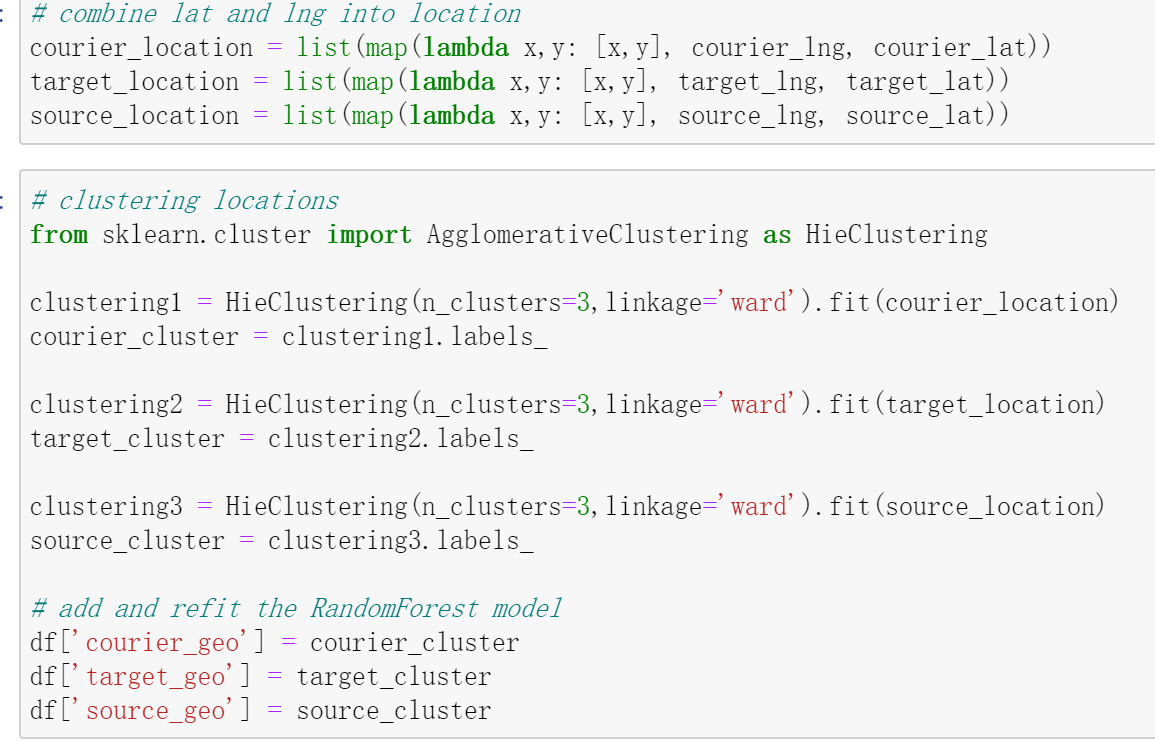
**Question(d)**

**Feature Engineering**

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

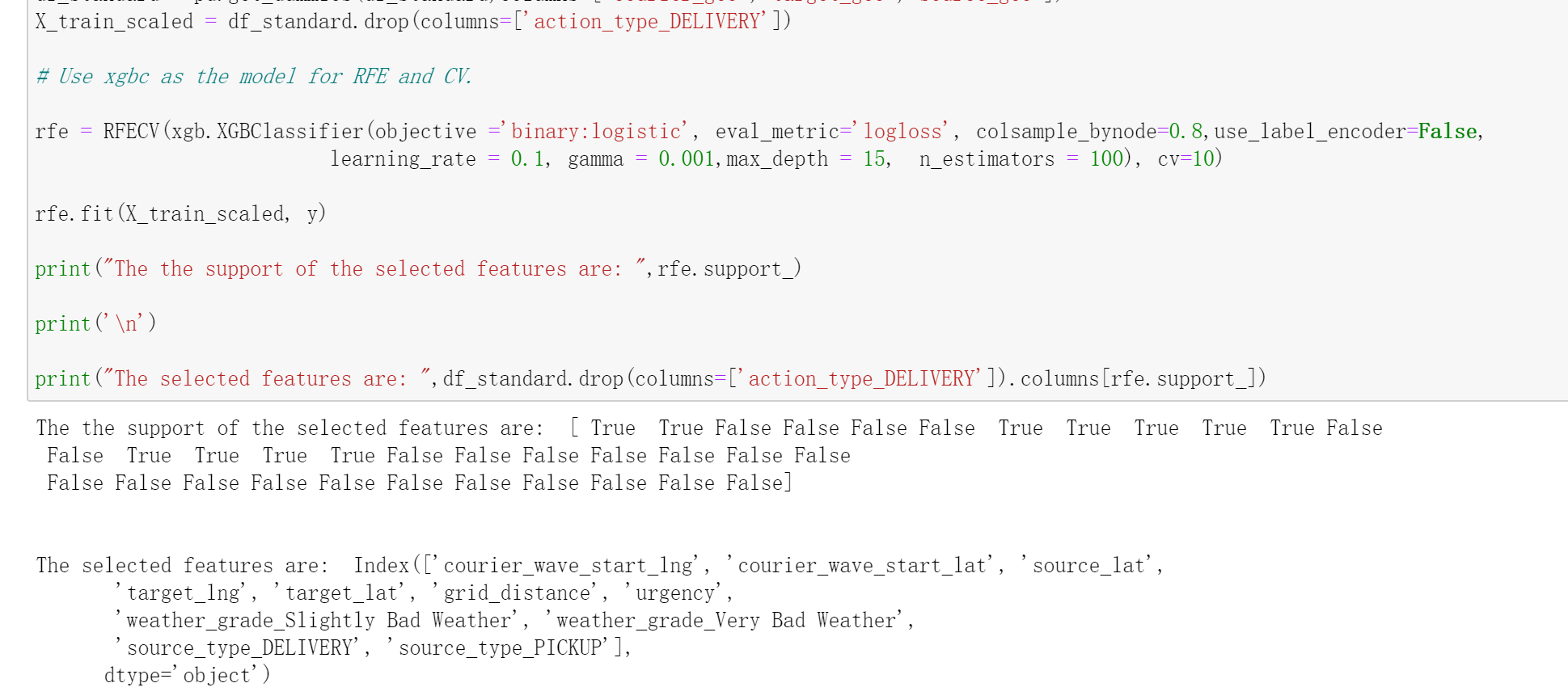


Second, we clustered the geographic information into 3 categories and added them as dummy variables into our model.

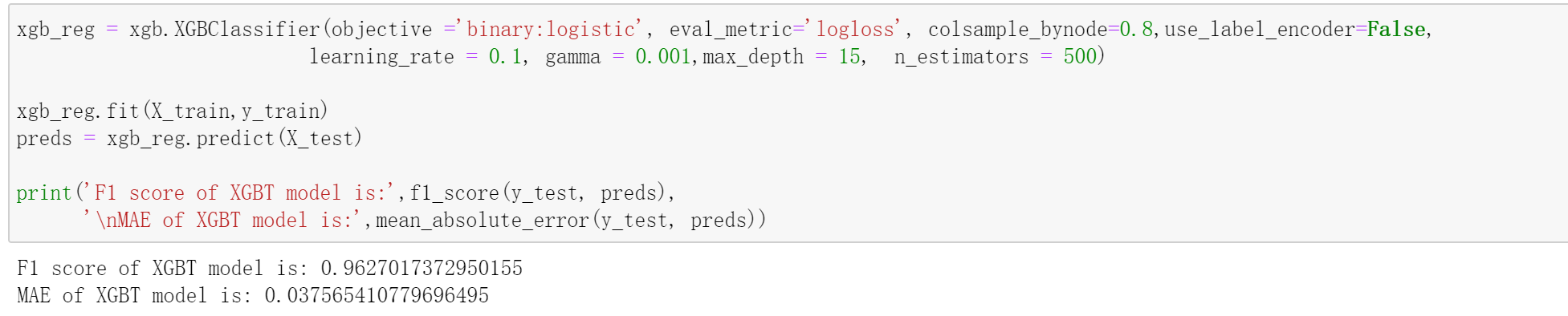


Actually, this did not improve our F1 score too much, so we conducted RFECV to select independent variables.

Third, we conducted feature selection by RFECV using all the variables except id.



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



Finally, the F1 score improve to 0.96.

**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 courier’s action type. Weather condition and geographic information can help predict courier’s next move. But level, max load of courier, speed, hour seemed no effects on courier’s action type.