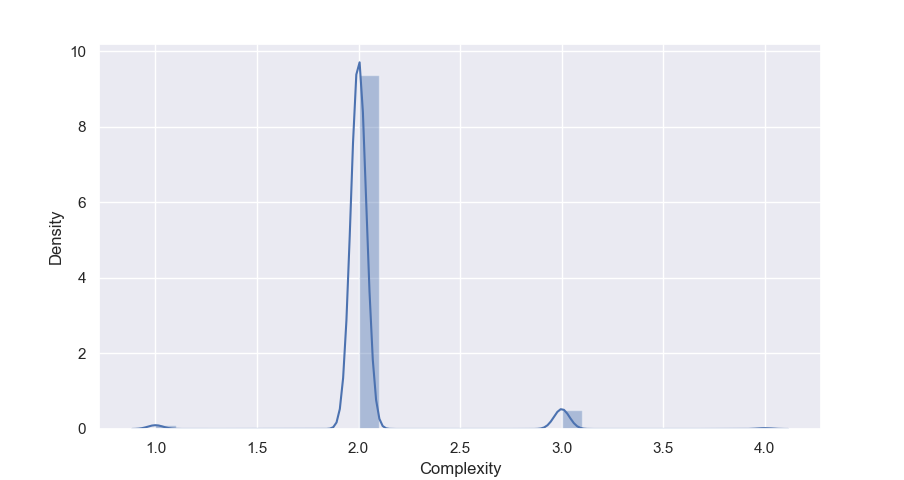
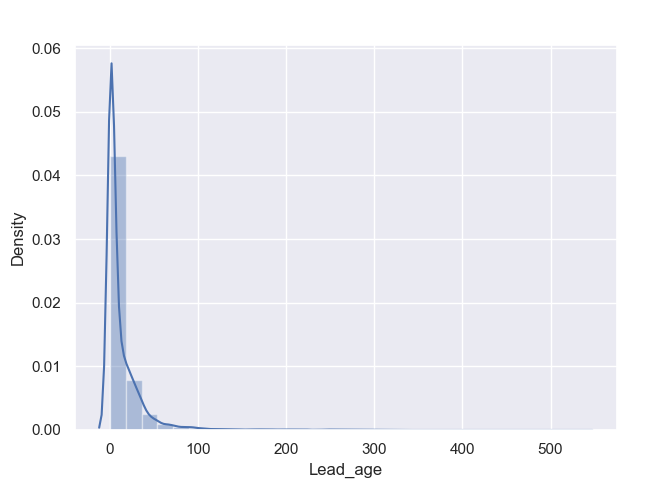
I use one-hot and label methods to encode our data and find out the classification result.

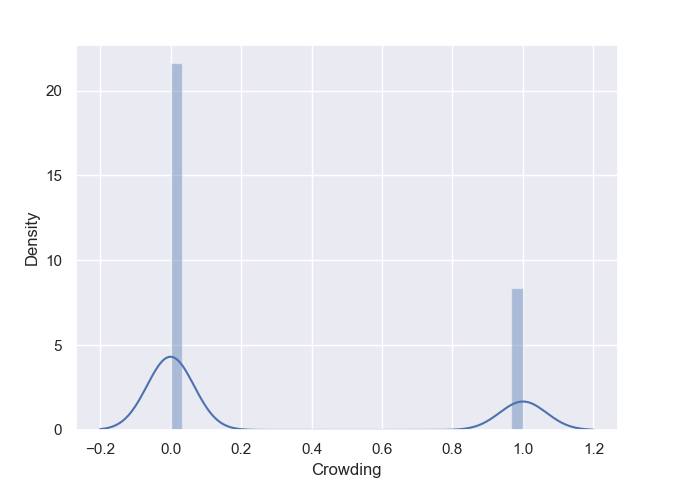
one-hot

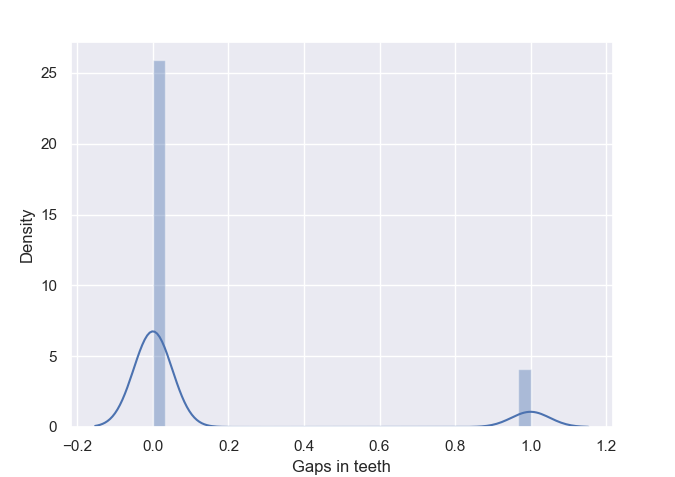
We use one-hot method to code our data, the value 0 and 1 represent whether this value is chosen by the row or not. And we divide columns 'Unhappy\_About', 'Complexity', 'Source\_Type', 'Stage','Lead\_age', 'Customer\_age', into columns 'Complexity','Lead\_age', 'Crowding','Gaps in teeth','Straighter teeth','Minor Crooked teeth', 'Protruding teeth', '1\_KOL\_Direct', '2\_KOL\_Ads', '3\_Promotion','4\_Content','5\_Direct','6\_Marketing Agency', '7\_Offline', '8\_Unknown', '9\_Referral', 'Closed Won', '15-25', '18-25','26-35','36-45','46-55','Above 55','Under 18' and use 0 and 1 to represent the value of each column rather than the String value.

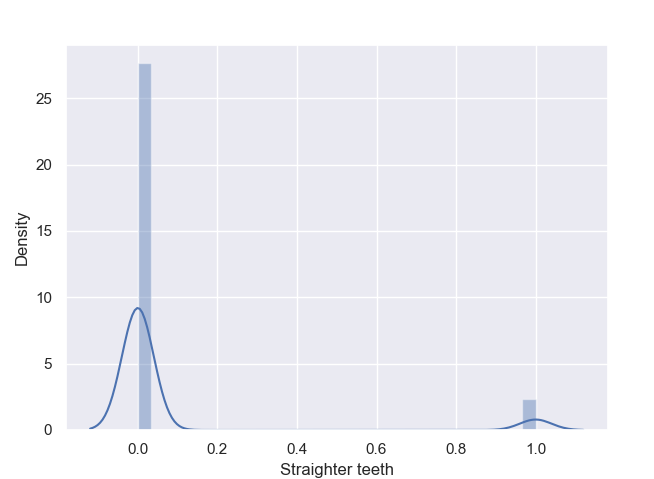
The pictures below are the distrubution of every attribute after one-hot encoding.

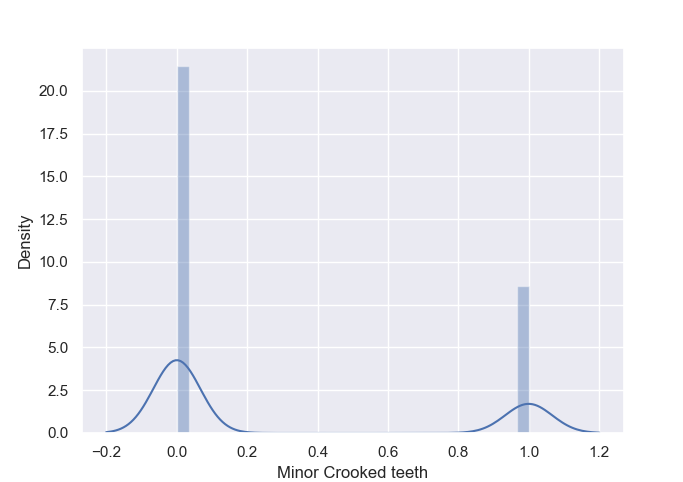


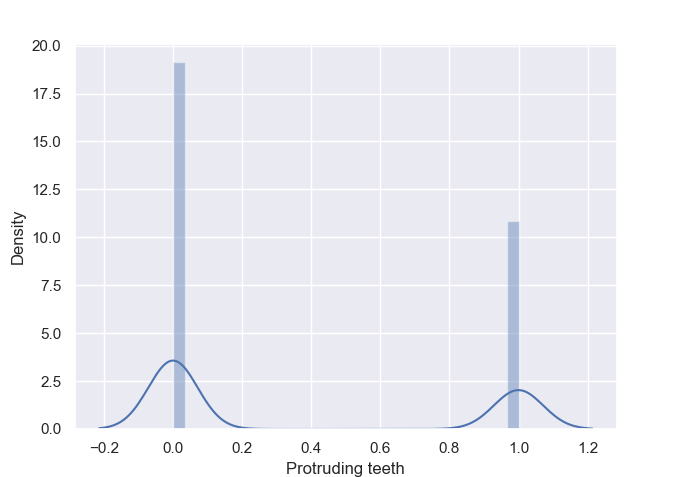


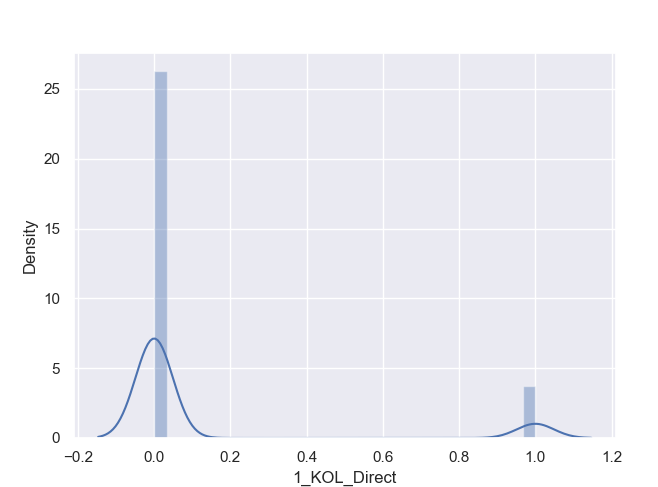


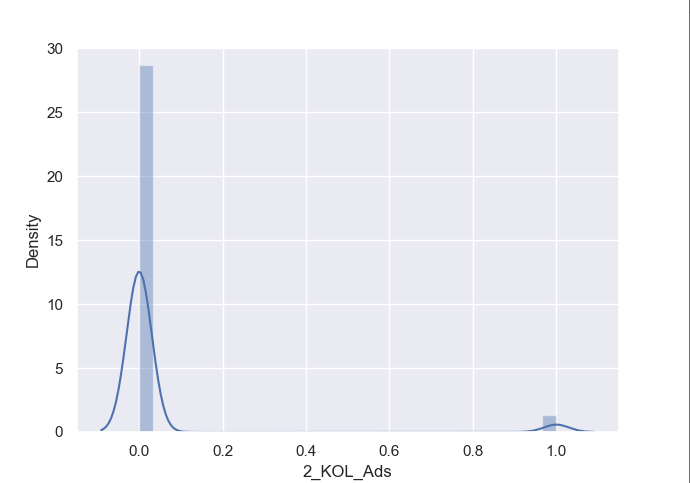


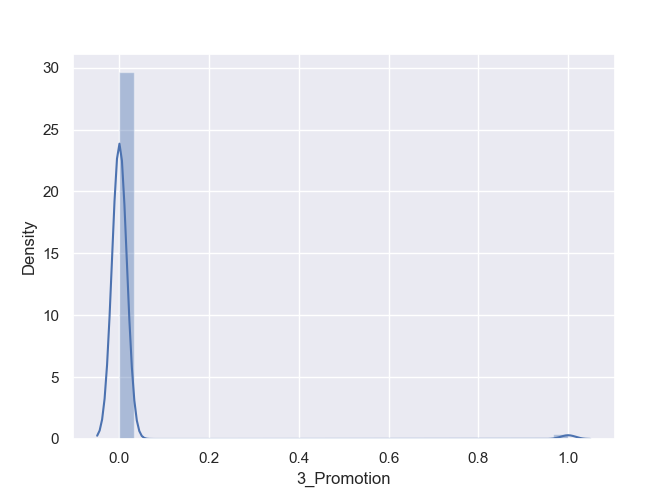


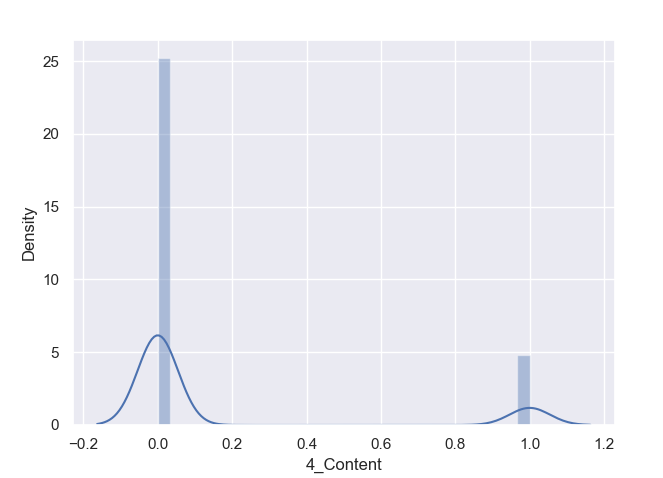


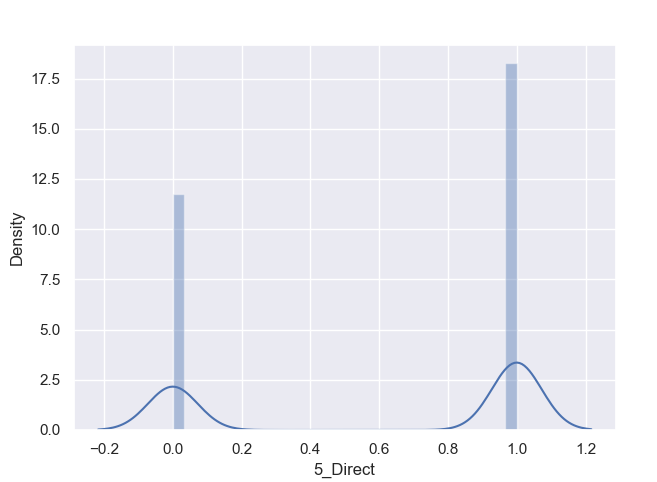


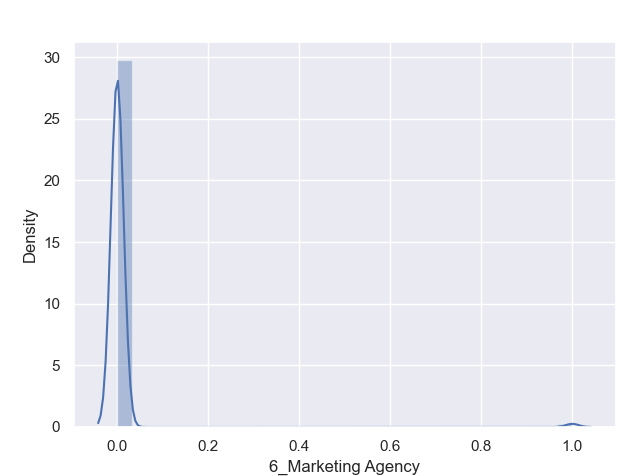


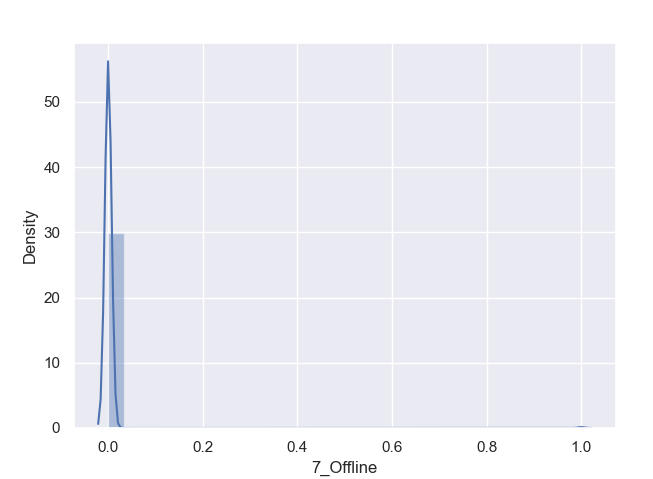


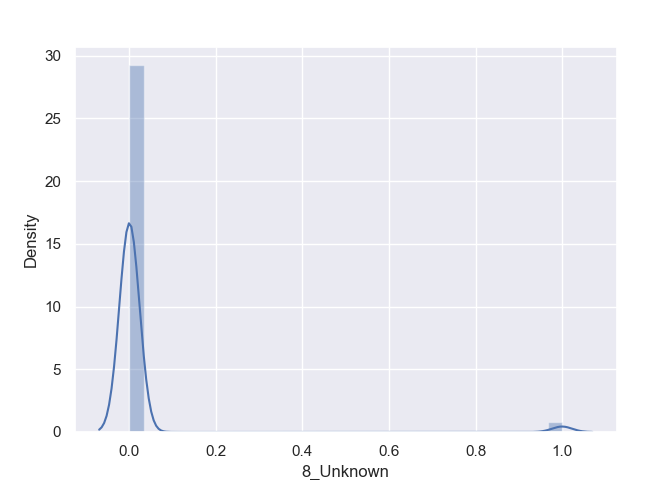


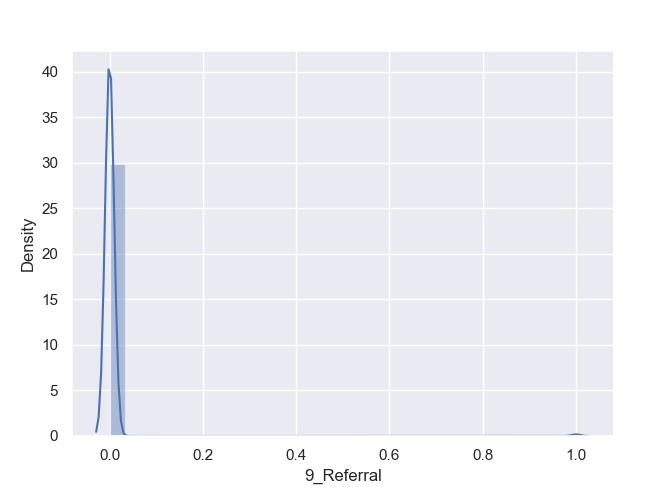


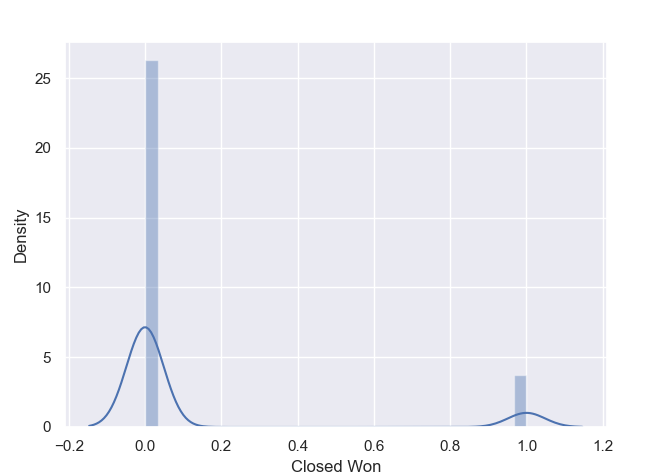


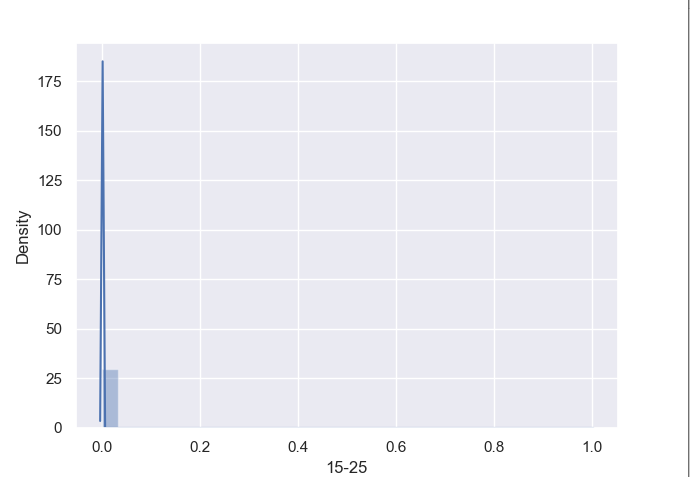


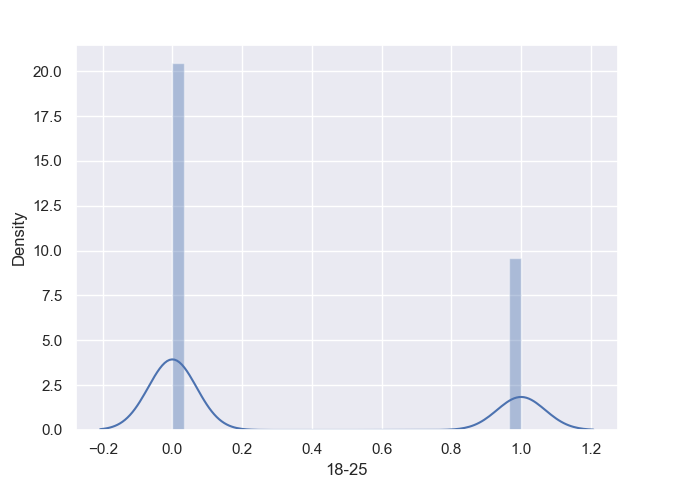


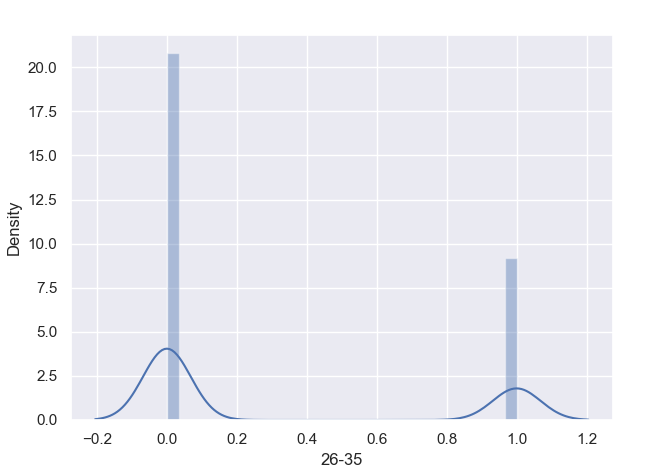


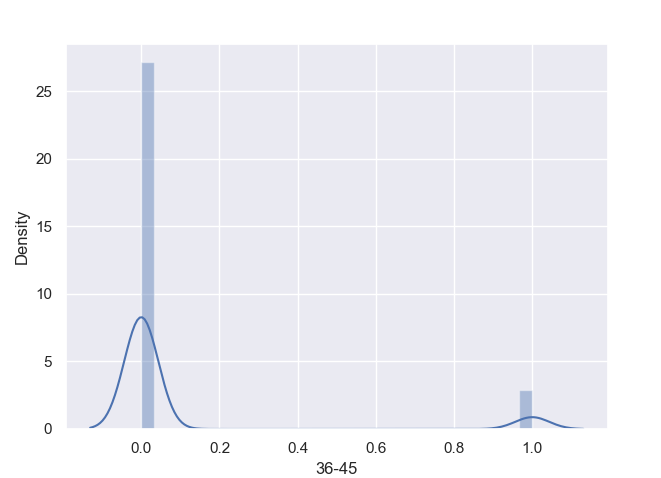


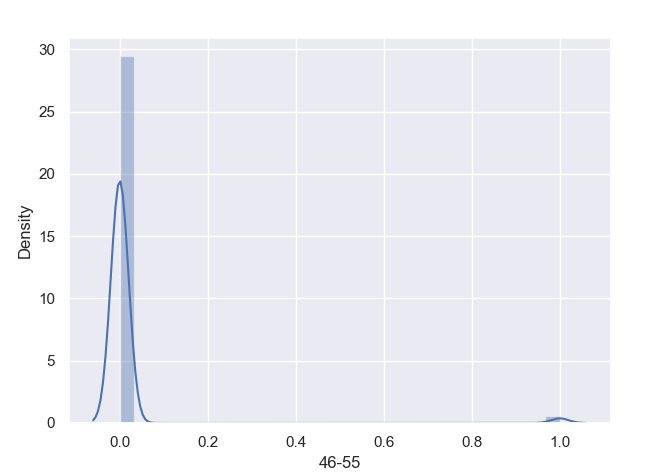


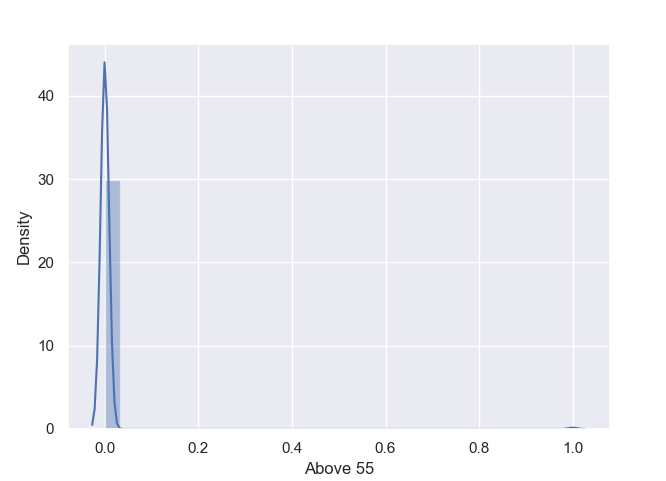


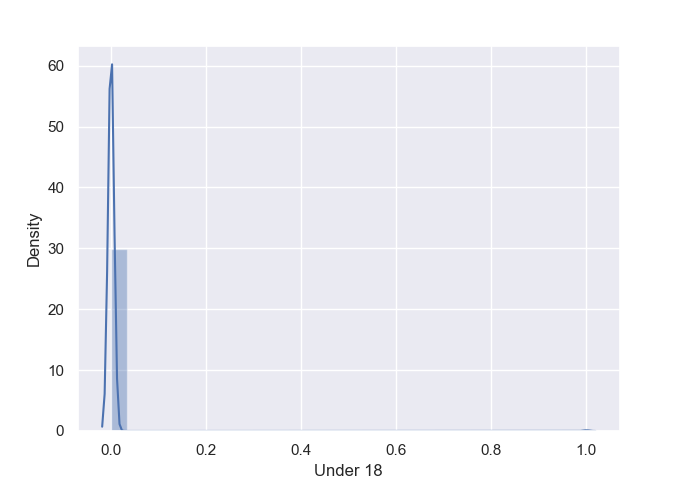




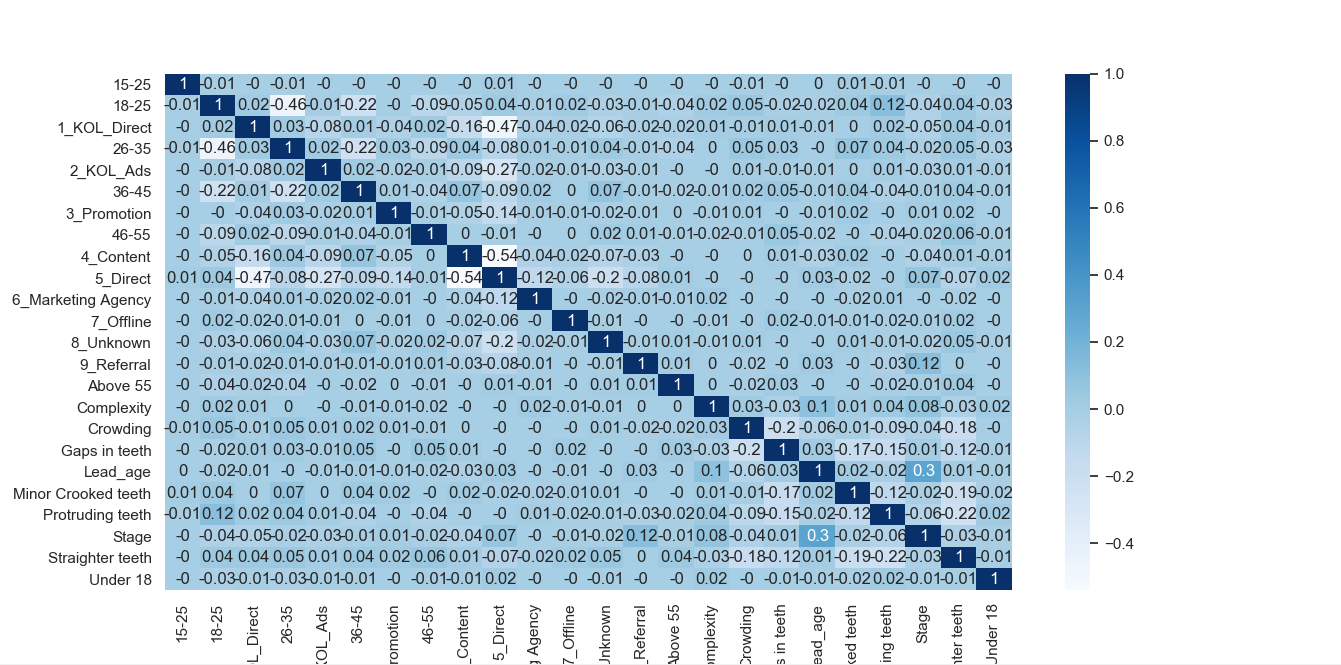








The picture below is heatmap, we can find that lead\_age is most relavant to Stage which determines whether people win or lost.



Finally we use three models to do classification work, and target is the attribute Stage. For every model, the test error is below.

DecisionTree

The test error of DecisionTree is 0.12200435729847492

SVM

The test error of SVM is 0.1328976034858388

QDA

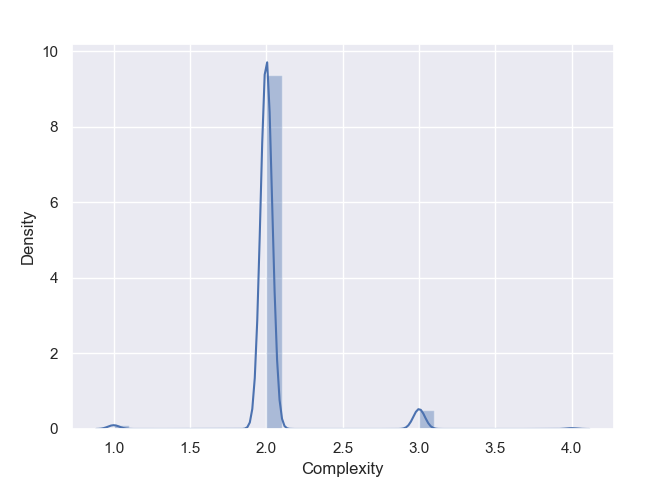
The test error of QDA is 0.2848583877995643

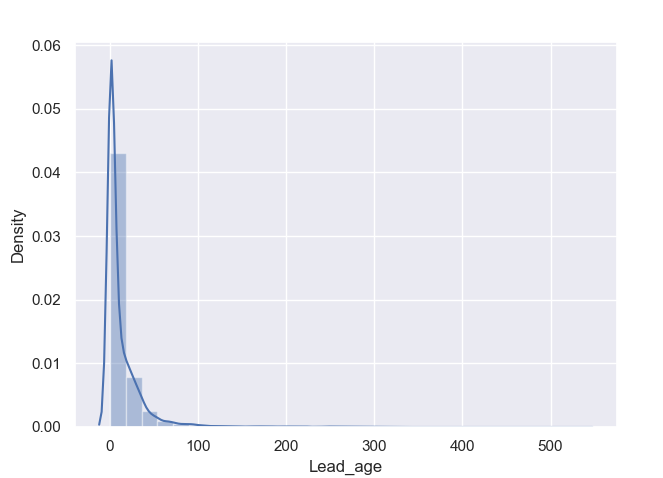
label

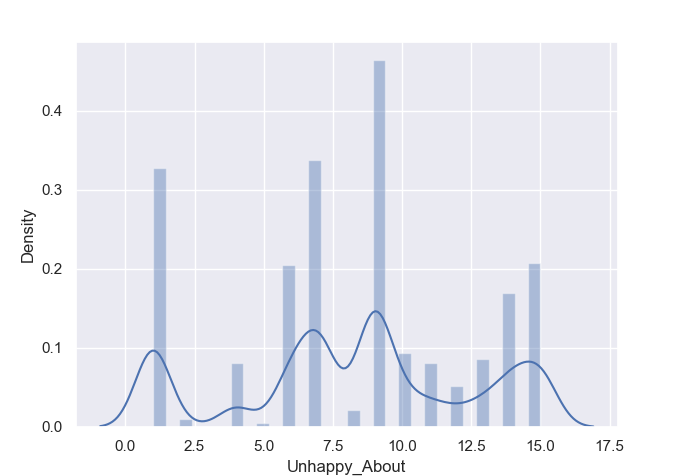
The second method we use to encode our data is label, we don’t divide columns and just use different value to represent different string.

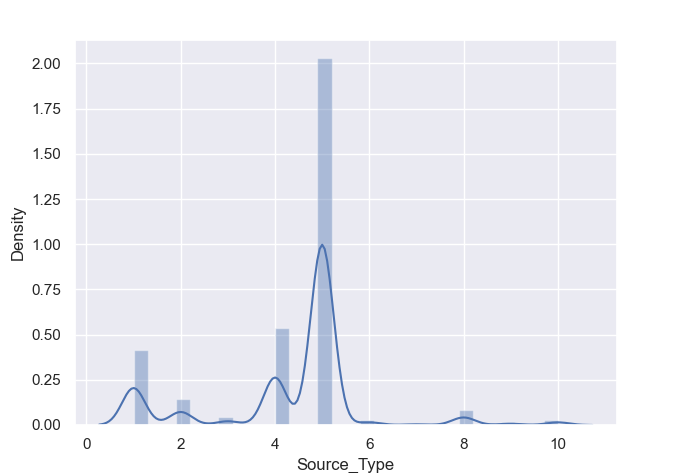
The columns of the data are 'Complexity', 'Lead\_age', 'Unhappy\_About', 'Source\_Type', 'Stage', 'Customer\_age'

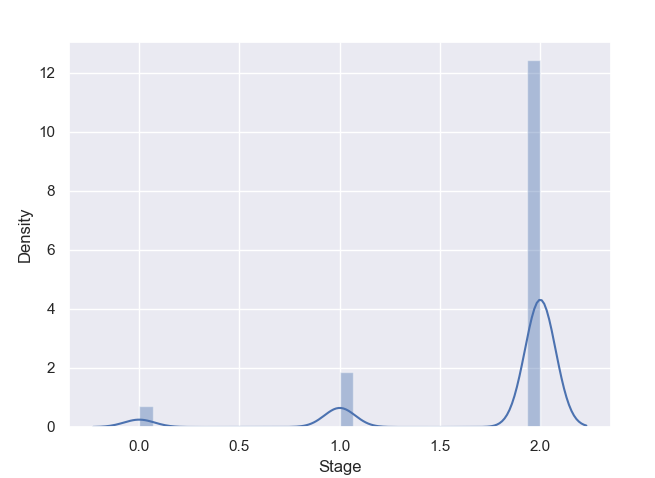
The pictures below are the distrubution of every attribute after label encoding.

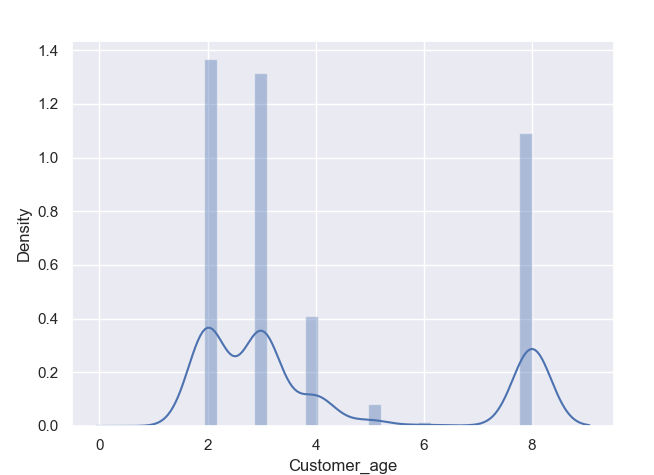




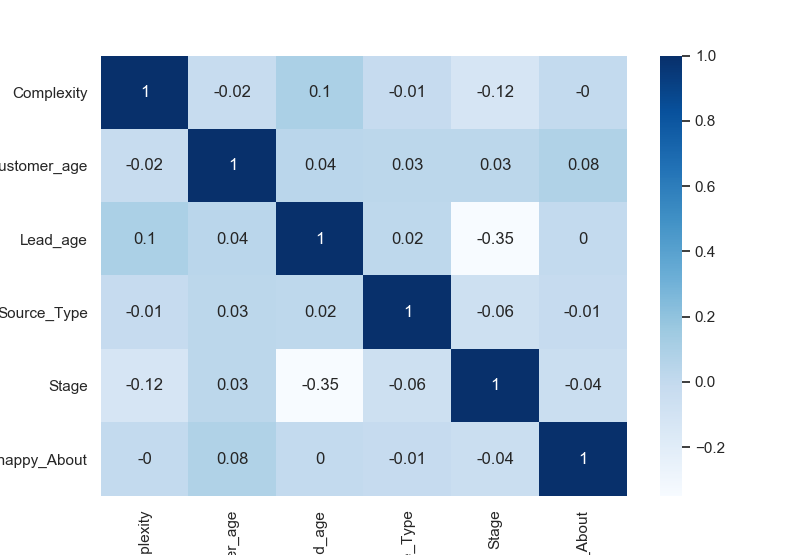








The picture below is heatmap, we can also find that lead\_age is most relavant to Stage which determines whether people win or lost.



Finally we use three models to do classification work, and target is the attribute Stage. For every model, the test error is below.

DecisionTree

The test error of DecisionTree is 0.1748366013071896

SVM

The test error of SVM is 0.1753812636165577

QDA

The test error of QDA is 0.20043572984749458

After the total work, we can fnd out that no matter what type of encoding method we use, lead\_age is the most important part to determine whether one people will win or get lost. Also, Decision tree have the best effect for our classification work.