Capstone Project_1 – In-depth Analysis

Prediction of telemarketing conversion probability using bank user dataset

Problem Statement/ Objective:

In this project the main goal is to predict if a Tele-marketing campaign for term deposit will lead to a conversion. The dataset provided consists of close to twenty features consisting of demographic details of the potential clients, any previous banking details of the client and details on the marketing campaign of a Portuguese Banking Institution . The target variable is to predict the probability of a potential client subscribing to the term deposit, so that clients with higher predicted probability of subscribing could be chosen for the campaign.

As part of this capstone project I plan to execute the following steps -

- Data cleaning and Preprocessing
- A thorough EDA of the provided data set
- Experimentation with different classification modules learned so far
- Analysis of the results and conclusion

Potential Client

Potential client of this project will be banking institution.

Exploratory Analysis- Data Wrangling

Data exploration —

- Columns : bank_data.head()
- Unique values in a column: bank_data.info()
- Describe: bank_data.describe()
  There are 4521 different rows in the dataset
- Duplicates: len(bank_data[bank_data.duplicated()])

Dealing with missing values:

— Quantifying missing values per column, filling & dropping missing values.

bank_data.isnull().sum()

bank_data.isna().sum()

No missing values in the bank_dataset

To find Outliers:

```
sns.boxplot(x=bank_data['age'])
sns.boxplot(x=bank_data['balance'])
sns.boxplot(x=bank_data['day'])
sns.boxplot(x=bank_data['duration'])
sns.boxplot(x=bank_data['campaign'])
sns.boxplot(x=bank_data['pdays'])
sns.boxplot(x=bank_data['previous'])
```

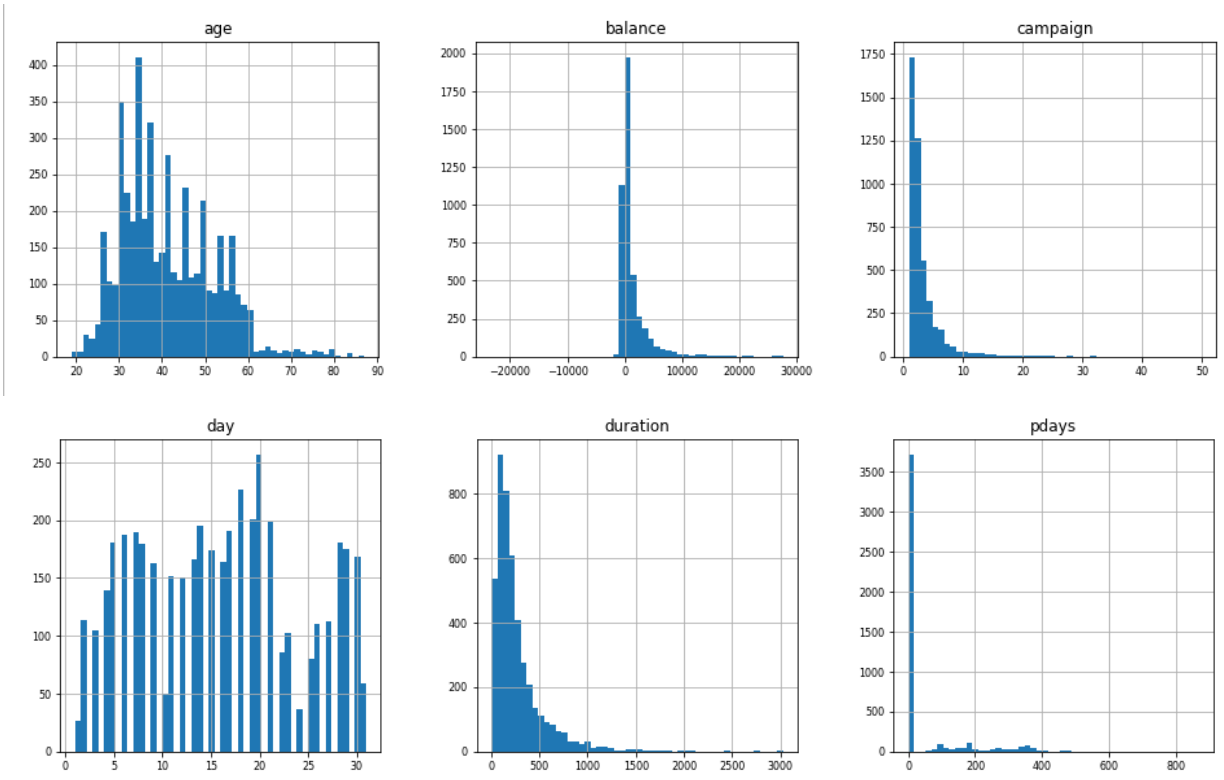
To remove Outlier:

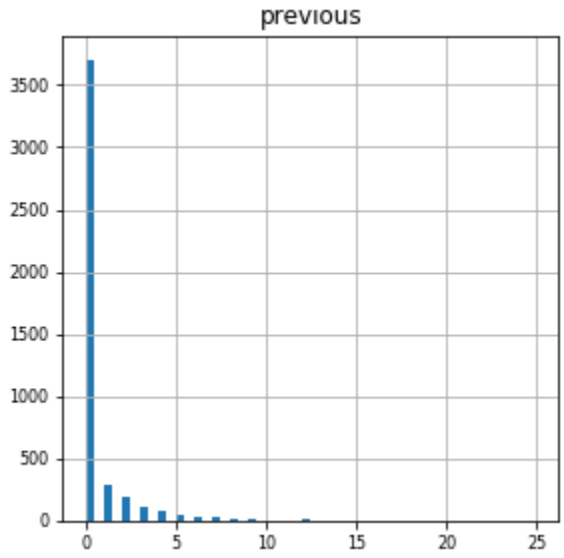
Outliers found in all the numerical attribute (age, balance, duration, campaign, Pdays,previous), **except day attribute.**

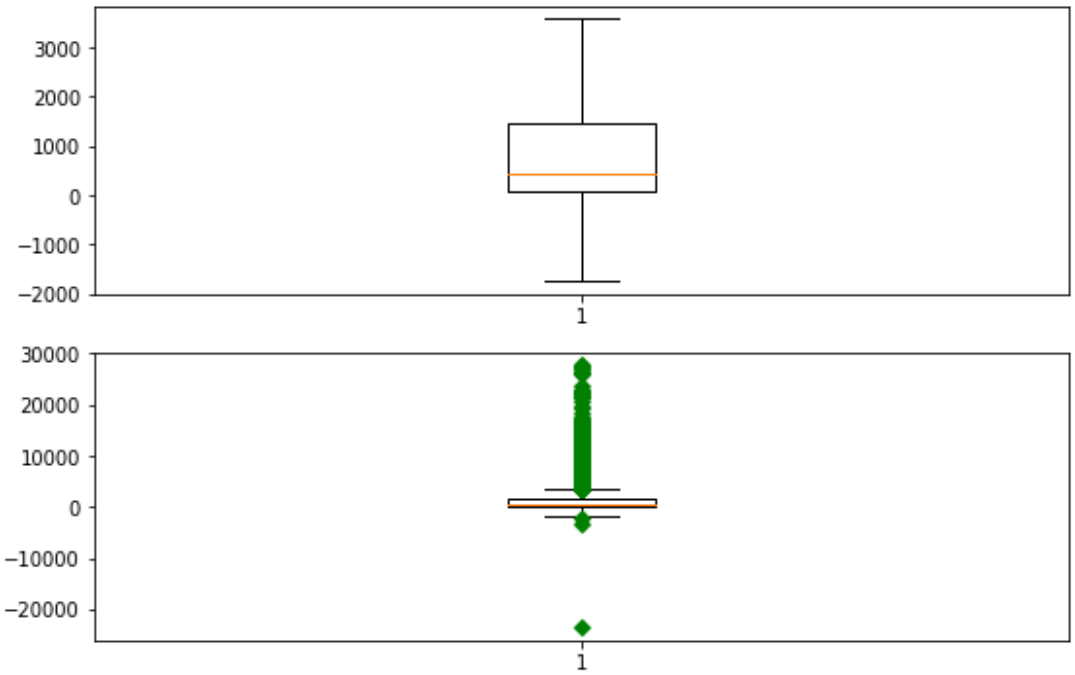
```
sns.boxplot(x=bank_data['age'])
q3 =bank_data['age'].quantile(0.75) q3
q1 =bank_data["age"].quantile(0.25) q1
iqr = q3-q1
iqr
upper_limit= q3+(1.5*iqr)
upper_limit
lower_limit= q1-(1.5*iqr)
lower_limit
bank_data.loc[bank_data['age'] < (q1 - 1.5 * iqr),['age']] = q1 - 1.5 * iqr
bank_data.loc[bank_data['age'] <= (q1 - 1.5 * iqr),['age']]
bank_data.loc[bank_data['age'] > (q3 + 1.5 * iqr),['age']] = q3 + 1.5 * iqr
bank_data.loc[bank_data['age'] >= (q3 + 1.5 * iqr),['age']]
After removed outliers:
sns.boxplot(x=bank_data['age'])
```

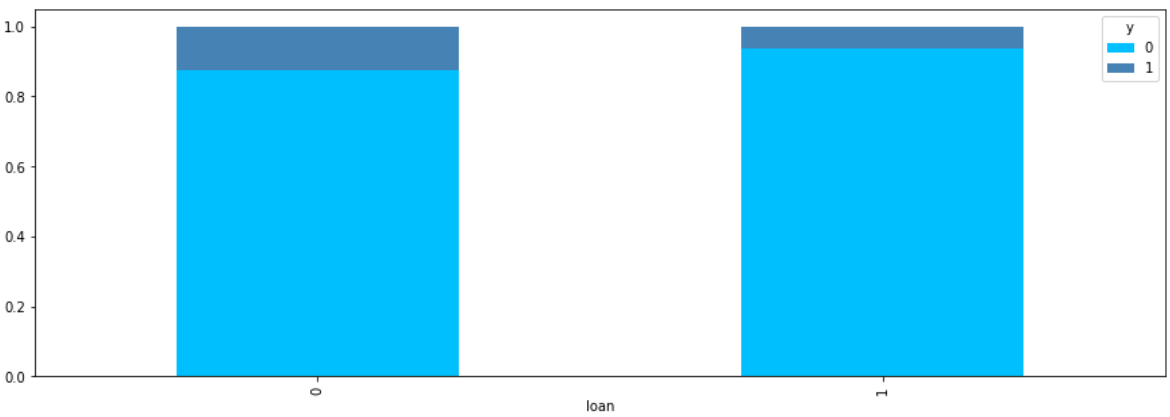
Similarly for all the attributes except "**days"** which is not having outliers.

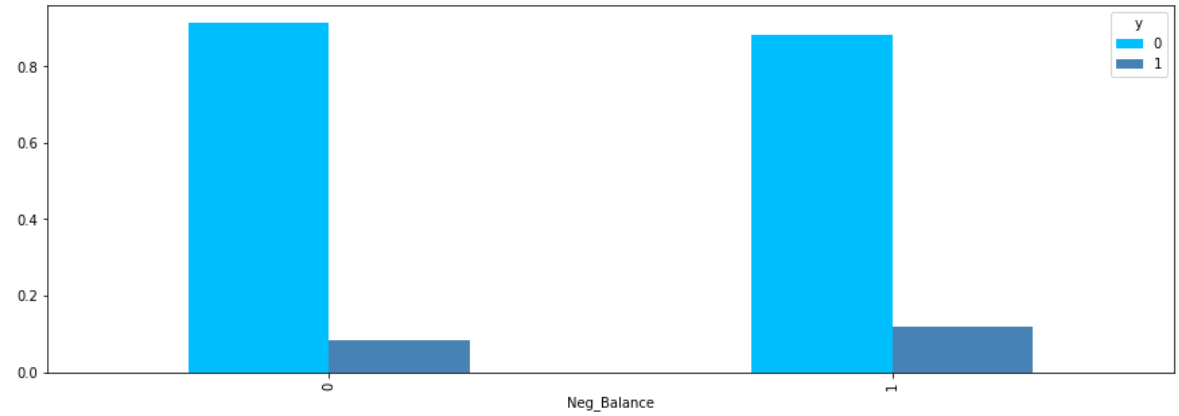
Visualization:

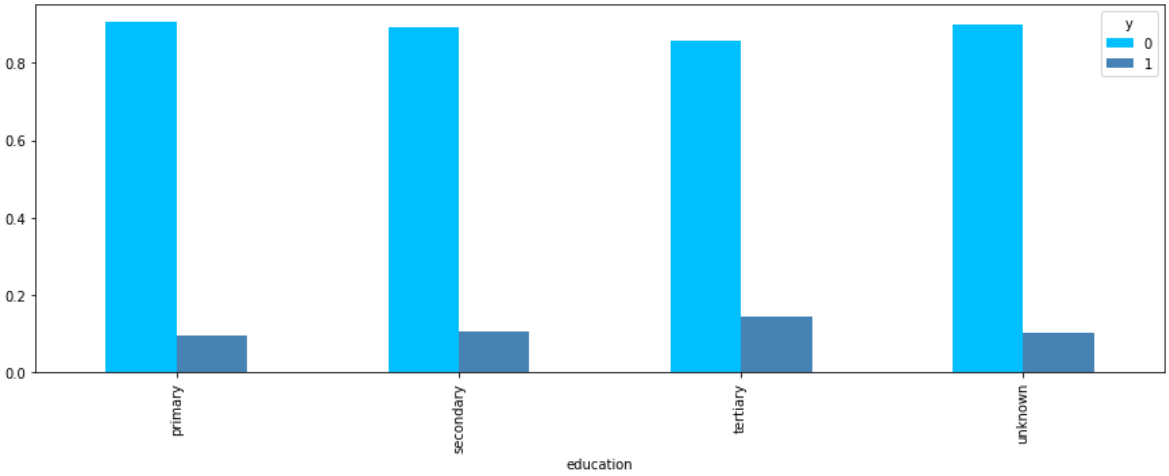



previous
3500
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
25


3000
2000
1000
0
−1000
−2000
1
30000
20000
10000
0
−10000
−20000
1


1.0
0.8
0.6
0.4
0.2
0.0
0
1
loan
y
0
1





Inferential Statistics

Dependent Variable

Deposit: y - has the client subscribed a term deposit? (binary: 'yes','no')

Independent variable

1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has a housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

'Observation is that 4521 rows and 17 numerical features after transformation. Target variable shape is (4521, 0) as expected

Primary analysis of several categorical features reveals:

- Administrative staff and technical specialists opened the deposit most of all. In relative terms, a high proportion of pensioners and students might be mentioned as well.
- Although in absolute terms married consumers more often agreed to the service, in relative terms the single was responded better. Best communication channel is secular.
- The difference is evident between consumers who already use the services of banks and received a loan.
- Home ownership does not greatly affect marketing company performance.

Observation from correlation matrix:

- Most correlated with target feature is call duration. So we need to transform it to reduce the influence.

- Highly correlated features (Previous day and Previous campaign) may describe clients state from last contact of current campaign. Their variance might support model capacity for generalization.

- Since categorical variables dominate in the dataset and the number of weakly correlated numeric variables is not more than 4, we need to transform categorical variables to increase the model's ability to generalize data. (we cannot drop them)
- Particular attention should be paid to the Duration Feature and categories that can be treated as binary. It suggests using binning and simple transformation accordingly (0 and 1).
- For categories of more than 3 types of possible option (marital and education) it is proposed to use the encode targeting - it will allow correctly relate the values to the target variable and use indicated categories in numerical form.

Statistical Test

The choice of metrics result

It is proposed to use ROC_AUC metrics for evaluating different models with additional monitoring of the accuracy metric dynamic.

RECALL:

Recall – Specificity: TN / (TN + FP) [MATRIX LINE 1]

- For all NEGATIVE(0) **REAL** VALUES how much we predict correct ?
- other way to understand, our real test set has 4521+20 = 4541 clients that didin't subscribe(0), and our model predict 98% correct or 4521 correct and 20 incorrect

  print(round(4521 /(4521 + 20),2))

PRECISION:

- Precision: TN / (TN + FN) [MATRIX COLUMN 1]

  (For all NEGATIVE(0) **PREDICTIONS** by our model, how much we predict correct ?)

- Precision: TN / (TN + FN) [MATRIX COLUMN 1]

  (For all POSITIVE(1) **PREDICTIONS** by our model, how much we predict correct ?)

F1-SCORE:

F1-SCORE: F1-Score is a "median" of Recall and Precision, consider this when we want a balance between this metrics

- F1 = 2(Precision(0) Recall(0)) / (Precision(0) + Recall(0))

```
F1_0 = 2*0.91*0.98/(0.91+0.98)
round(F1_0,2)
```

AVG/ TOTAL

Consider the weights of sum of REAL VALUES [line 1] [line2]

- AVG_precision = (0.91*(7279/8238))+ (0.69*(959/8238))
  round(AVG_precision,2)
- AVG_Recall = (0.98*(7279/8238))+ (0.26*(959/8238))
  round(AVG_Recall,2)
- AVG_f1 = (0.95*(7279/8238))+ (0.38*(959/8238))
  round(AVG_f1,2)

This approach will allow us to explore models from different angles.

Estimator is **Logistic regression**: F1 score is 0.505724

ROC_AUC is 0.74165 and accuracy rate is 0.886188

Building the Model

The choice of the most effective model, build learning curve rate

Estimator is **Logistic regression**: F1 score is 0.505724

ROC_AUC is 0.74165 and accuracy rate is 0.886188

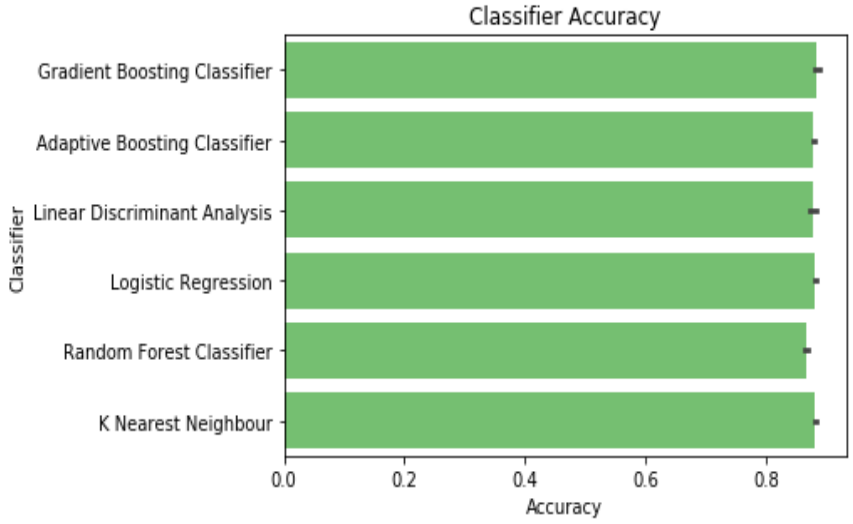
Estimator is **GradientBoostingClassifier**: F1 score is 0.583217

ROC_AUC is 0.81165 and accuracy rate is 0.892818

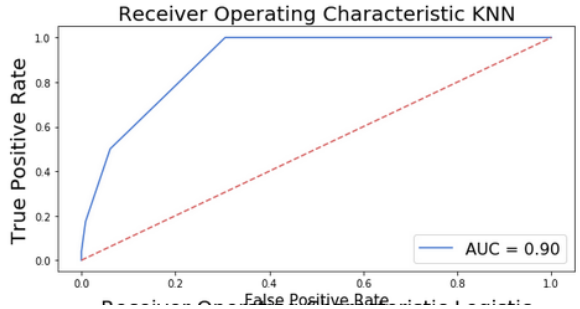
Estimator is **KNN** : F1 score is 0.496126

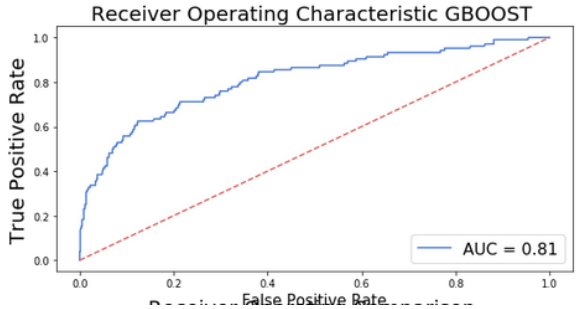
ROC_AUC is 0.90198 and accuracy rate is 0.883978

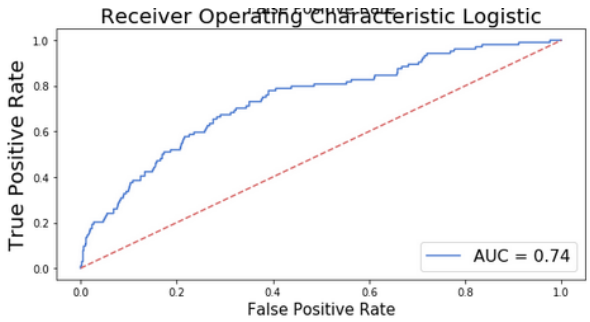
The Accuracy and Classifier Accuracy of various models

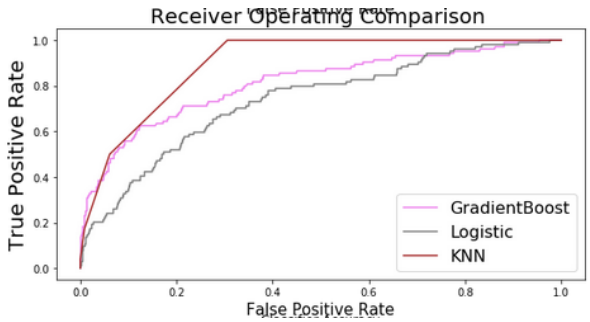


ROC_AUC Curve for KNN, Gradient Boost, and Logistic Regression Model and then the comparison of both









ANALYZING THE RESULTS

So now we have to decide which one is the best model:

- False Positive, means the client do NOT SUBSCRIBED to term deposit, but the model thinks he did.
- False Negative, means the client SUBSCRIBED to term deposit, but the model said he don't.

- The first one its most harmful, because we think that we already have that client but we don't and maybe we lost him in other future campaigns.
- The second its not good but its ok, we have that client and in the future we'll discovery that in truth he's already our client

So, our objective here, is to find the best model with the lowest False Positive as possible.

Our best performed model with the ROC_AUC (0.90198) metric is **KNN** . This classifier could achieve accuracy rate 0.88 that is average accuracy among all classifiers (0.90).

Conclusions and Future Scope:

This analysis can be carried out at the level of individual bank branches as does not require much resources and special knowledge (the model itself can be launched automatically with a certain periodicity). Potentially similar micro-targeting will increase the overall effectiveness of the entire marketing company.

1. Take into account the time of the company (May is the most effective)
2. Increase the time of contact with customers (perhaps in a different way formulating the goal of the company). It is possible to use other means of communication.
3. Focus on specific categories. The model shows that students and senior citizens respond better to proposal.
4. Age, income level (not always high), profession can accurately determine the marketing profile of a potential client.

Given these factors, it is recommended to **concentrate on those consumer groups** that are potentially more promising. The concentration of the bank's efforts will effectively distribute the company's resources to the main factor - the bank's contact time with the client - it affects most of all on conversion.