Principles of Data Science Assignment Stage 1

This report is written by: Mimi Mukherjee

Student ID: mmuk3840

Subject Name: Principles of Data Science

Subject Code: COMP5310

Year: 2019

Semester: 1

Problem

Telemarketing is an effective strategy of selling goods and services by using phone calls. With the ever-growing competition rates in the business world, tele-sales need innovation. With the help of machine learning, salesperson's approach to the potential buyers can be more effective.

This project data is related to direct telemarketing campaigns of a Portuguese banking institution. This dataset is retrieved from MCI Machine Learning Repository(link). This is a classification problem, where the goal is to predict if the client will purchase a term deposit. This data-driven approach using social and economic attributes can help us to target potential buyers with better efficiency.

Data

Dataset Characteristics:	Multivariate	Number of Instances:	41188	Number of Numerical Attributes:	10
Associated Tasks:	Classification	Number of Attributes:	20	Number of Categorical Attributes:	10

In order to clean and visualize the data below are the steps performed, in order:

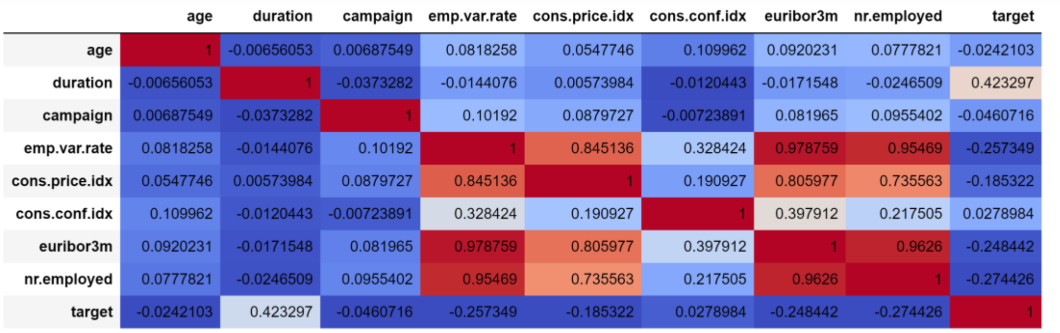
1. Deal with missing values: There are several missing values in some categorical attributes, all coded with the "unknown" label. Instead of using deletion or imputation techniques, they have been treated as a possible class label. **Justification**: In real scenario, if some data of a person is unknown to the salesperson, they can still approach the person based on the model which takes 'unknown' data in consideration.

2. Convert two numeric columns ('pdays' and 'previous') into categorical columns: 'Pdays' represents number of days that passed by after the client was last contacted - 999 means client was not previously contacted. 'Previous' represents number of contacts performed -0 means never contacted. **Justification**: Here, 999 and 0 signifies not their numeric value but some different category. So, these attributes have been changed to categorical columns.

3. Remove outliers from numeric data: I have removed the values outside the range (mean +/- 2.5 * S.D.) from numerical columns. **Justification**: Outliers can worsen performance of the model.

4. Change target variable type from categorical to numeric: I have changed the target variable which signifies if the customer will buy term deposit from categorical values yes/no to numeric values 1/0 respectively. **Justification**: Numeric data can give us a probability distribution indicating the chances of buying term deposit from 0 to 1.

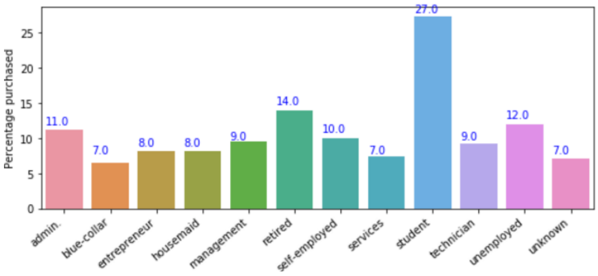
5. Identify correlation between columns of the dataset: I visualized the correlations of numeric columns in matrix format.

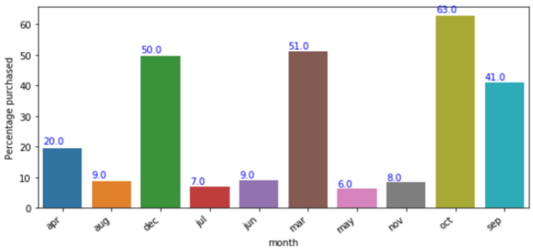


	age	duration	campaign	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	target
age	1	-0.00656053	0.00687549	0.0818258	0.0547746	0.109962	0.0920231	0.0777821	-0.0242103
duration	-0.00656053	1	-0.0373282	-0.0144076	0.00573984	-0.0120443	-0.0171548	-0.0246509	0.423297
campaign	0.00687549	-0.0373282	1	0.10192	0.0879727	-0.00723891	0.081965	0.0955402	-0.0460716
emp.var.rate	0.0818258	-0.0144076	0.10192	1	0.845136	0.328424	0.978759	0.95469	-0.257349
cons.price.idx	0.0547746	0.00573984	0.0879727	0.845136	1	0.190927	0.805977	0.735563	-0.185322
cons.conf.idx	0.109962	-0.0120443	-0.00723891	0.328424	0.190927	1	0.397912	0.217505	0.0278984
euribor3m	0.0920231	-0.0171548	0.081965	0.978759	0.805977	0.397912	1	0.9626	-0.248442
nr.employed	0.0777821	-0.0246509	0.0955402	0.95469	0.735563	0.217505	0.9626	1	-0.274426
target	-0.0242103	0.423297	-0.0460716	-0.257349	-0.185322	0.0278984	-0.248442	-0.274426	1

We can see duration column which represents last contact duration highly affects the target output. But the duration is not known before a call is performed and after the end of the call target 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. Also, we observed that euribor3m, emp.var.rate and nr.employed are very closely related. Any one of these three economic context attributes can be used to predict the other two attributes. So, we have dropped these three attributes from the dataset – duration, emp.var.rate and nr.employed.

6. We visualized the categorical variables to identify if any of the category affects the target variable more than others. Below is one such visualization which indicates students are more attracted to buying term deposit than people in other occupations.

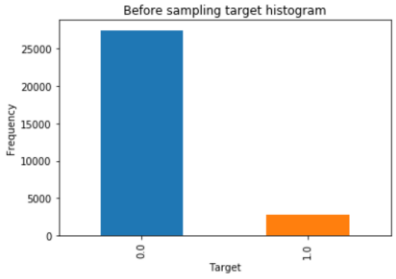


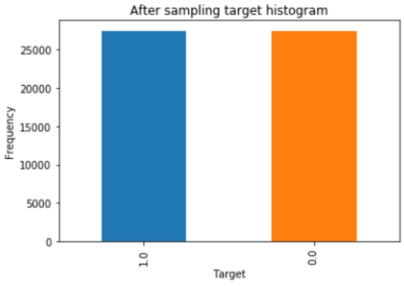


The visualization of ‘month’ attribute indicate there could be a seasonal impact on the term deposit purchase, because compared to other months, deposit purchase ratio is higher in October. However, these correlations do not guarantee any causality. These visualisations help only to get sense of the data.

7. One hot encoding of categorical columns: We have converted categorical attributes to dummy/indicator variables by using get_dummies() method.
Justification: Numerical value helps machine learning algorithms to do a better job in predication.

8. Dealing with class imbalance problem: Out of 40000+ instances only 4600 observations have purchased term deposit. The overwhelming ‘not purchased’ class has outnumbered ‘purchased’ class by 9:1 ratio and our target is to find out those customers who are likely to buy term deposit. This significantly under-represented ‘purchased’ data is difficult to train. To handle this problem, I have synthetically generated minority samples (SMOTE), i.e. oversampled this minority class.





Tools used

1. Data cleaning: pandas, numpy
2. Data split into test and train set: train_test_split from sklearn.model_selection
3. Balance target class: SMOTE from imblearn.over_sampling
4. Data visualisation: seaborn, matplotlib

Approach

In the second part of the assignment I will be using logistic regression, neural network (NN) and support vector machine and will compare the performance of the models. The ROC curve will be used to visualise the accuracy.

Dataset source:

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Citation:

This dataset is available to public for research. The details are described in [Moro et al., 2014].

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014