Project Objective: To predict if an existing customer will subscribe to a data pack or not through telemarketing activities

Technique Used: Logistic Regression for model implementation

About Dataset : In this data set there are 16 variables and 4521 tuples. There are 8 categorical variables in this data set. Following are the columns:

1. **Age:** age of customer (numeric)
2. **job:** type of job (categorical)
3. **marital:** marital status (categorical)
4. **education** (categorical)
5. **connect:** has more than one connection? (binary: "yes","no")
6. **balance:** voice credit balance (numeric)
7. **landline:** has landline? (binary: "yes","no")
8. **smart:** has smart phone? (binary: "yes","no")
9. **last_day:** last contact day of the month (numeric)
10. **last_month:** last contact month of year (categorical)
11. **duration:** last contact duration, in seconds (numeric)
12. **campaign:** number of contacts performed during this campaign includes last contact
13. **passdays:** number of days that passed by after the customer was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
14. **previous:** number of contacts performed before this campaign
15. **poutcome:** outcome of the previous marketing campaign (categorical)
16. **target(OUTPUT) :** has the customer subscribed a data pack? (binary: "1","0)

Methodology followed:

Step1: Data visualizations

Step2: Data preparation

Step 2.1: One-hot encoding
Step 2.2: SMOTE algorithm for up-sampling the output
Step 2.3: Recursive Feature Elimination (RFE)

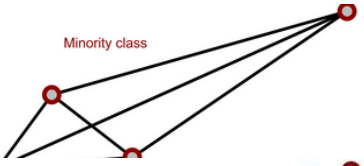
Step3: Model Implementation using Logistic Regression & its Interpretation

---

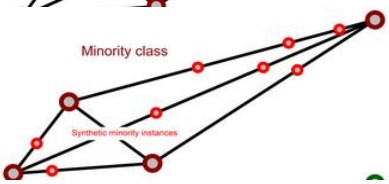
About the techniques(SMOTE, RFE and logistic regression):

1. SMOTE: SMOTE algorithm aka **S**ynthetic **M**inority **O**versampling **T**echnique is used for dealing with imbalanced data distribution (over sampling). In our case, the output column(target) has ~88% values as '0' and ~12% values as '1'. This is the case of oversampling because we want to increase the number of people buying the data pack ('1').

SMOTE synthesises new minority instances between existing (real) minority instances. Imagine that SMOTE draws lines between existing minority instances like this.



SMOTE then imagines new, synthetic minority instances somewhere on these lines.

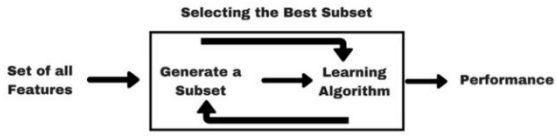



```
from imblearn.over_sampling import SMOTE
```

2. RFE: Recursive **F**eature **E**limination, or RFE for short, is a feature selection algorithm. It is a wrapper method i.e model is trained using the subset of features. Next on the basis of the outcome of previous model we add or remove some features.

This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

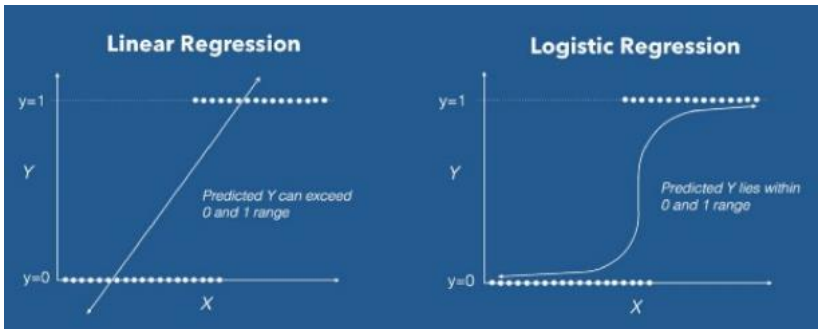
```
from sklearn.feature_selection import RFE
```



3. Logistic Regression: It is a classification technique and is used when the dependent variable (target) is categorical. For example, To predict whether an email is spam (1) or (0).

Assumptions

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little or no multicollinearity.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.



References:

https://rikunert.com/SMOTE_explained

https://medium.com/@vijain2010/beginners-guide-to-feature-selection-techniques-using-python-4d23fcb8951a

https://towardsdatascience.com/logistic-regression-b0af09cdb8ad

https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8