Group Project Second Deliverable

Team Details

- Group Name: Solo Banking
- Name: Ian Paulino Velez
- Email: ianpvelez@gmail.com
- Country: United States
- College: Fordham University
- Specialization: Data Science

Problem Description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model to help them understand whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Data Understanding

We have a dataset containing 17 columns: 16 features and the target column ‘y’. 7 of these features are numerical, while the rest are categorical (either multilabel or binary). This is a list of the features and their definition:

- Bank client data:

1. age (numeric)
2. job: type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed')
3. marital: marital status (categorical: 'divorced','married','single'; note: 'divorced' means divorced or widowed)
4. education (categorical: ‘primary’, ‘secondary’, ‘tertiary’)
5. default: has credit in default? (categorical: 'no','yes')
6. housing: has housing loan? (categorical: 'no','yes')
7. balance: average yearly balance (numeric)
8. loan: has personal loan? (categorical: 'no','yes')

- Related with the last contact of the current campaign:

9. contact: contact communication type (categorical: 'cellular','telephone')
10. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
11. day_of_week: last contact day (numeric: 1 to 31)
12. 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:

13. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric)

15. previous: number of contacts performed before this campaign and for this client (numeric)
16. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
17. y: has the client subscribed a term deposit? (binary: 'yes','no')

The numerical variables, other than ‘age’ and ‘day_of_week’ were all suspected to have outliers. However, they were also mostly skewed, which made IQR not a good decision maker to drop all of these values, which made up a substantial amount. Thus, very extreme values were manually removed.

For missing values, a very large amount was found on ‘poutcome’ and ‘contact’, which may have to be removed entirely. A more manageable number of missing values was found on ‘job’ and ‘education’. Both being categorical variables, an attempt was initially made to impute these values with a knn classifier and a random forests classifier, but both models performed very poorly on the validation set. Because of this, the decision was made to fill in values with proper intentionality with the most common category (e. g. relationship between variables, such as more than 80% of the people in management having a tertiary education) and see how models perform for those sensitive to missing values. For other machine learning models, such as random forests, it may be valuable to explore the performance both in the presence and absence of missing values in our dataset.

Github Repo link:

https://github.com/ianpv04/Data-Glacier-Project.git