**Prediction of Bank Client’s Term-deposit Subscription Behavior utilizing Supervised Learning Models**

Hetian Bai (hb1500), Jieyu Wang (jw4937), Zhiming Guo (zg758), Fu Shang (fs1520)

1. **Introduction**

This is the report for *Introduction to Data Science* course team project. In this machine learning project, we found the data from [UCI](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) data pool and utilized the techniques introduced in class to process data and build predictive models. At the end of the project, we will make predictions of the target variable with the best supervised learning model. We present a complete data mining and modeling workflow with details in this report.

**Business Scenario and Motivation**

A Portuguese banking institution attempts to get more of its clients to subscribe for a term deposit. A higher amount of term deposit subscription creates more opportunities for the bank to increase profit, which allows the bank to invest in higher gain financial products and to pay higher interest to its customers. Therefore, we are solving a classification problem and the target variable is to make binary predictions on whether the client will make subscription of term deposit with the bank or not. To solve the classification problem, we will build several predictive models then compare their performances to obtain the best model.

**Problem Formulation**

* Proactively reach out to bank clients whom with higher likelihood of making term deposit

subscriptions

* Study the features to gain insights into key drivers to promote clients to make subscriptions
* Create an efficient marketing strategy based on data mining and analysis

The predictive model will contribute to the bank’s marketing team to better targeting their potential clients who are more likely to make subscriptions for term deposit. Given a client’s information, this model will make a predictive result of whether the customer will subscribe for a term deposit with the bank. Then, the marketing team can focus on advertising the bank’s term deposit products to those clients whose predictive results are positive. Also, the dataset includes some data about when and how often the client was contacted. By dealing with those data, we might be able to give suggestions on how to conduct campaigns in a more effective way.

1. **Data Exploration**

**Predictor Variables**

The historical data contains 41,188 instances. Each data instance includes 20 features before data processing, which are categorized into three groups: personal information, previous market campaign result, and current economic indicators. Personal information includes age, job, marital status, education, default, housing, loan, and contact methods; previous market campaign result includes days and month that last contact was made, duration of last contact in seconds, number of days since the client was last contacted in a previous campaign, number of contacts performed during this campaign for this client and outcome of the previous marketing campaign; current economic indicators includes employment variation rate, consumer price index, consumer confidence index, euribor 3-month rate and number of employees.

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Type |
| age | Age of the client | Numeric |
| job | Client's occupation | Categorical |
| marital | Marital status | Categorical |
| education | Client's education level | Categorical |
| default | Indicates whether the client has credit in default | Categorical |
| housing | Indicates whether the client has a housing loan | Categorical |
| loan | Indicates whether the client as a personal loan | Categorical |
| contact | Type of contact communication | Categorical |
| month | Month that last contact was made | Categorical |
| day\_of\_week | Day that last contact was made | Categorical |
| duration | Duration of last contact in seconds | Numeric |
| campaign | Number of contacts performed during this campaign for this client (including last contact) | Numeric |
| pdays | Number of days since the client was last contacted in a previous campaign | Numeric |
| previous | Number of contacts performed before this campaign for this client | Numeric |
| poutcome | Outcome of the previous marketing campaign | Categorical |
| empvarrate | Employment variation rate (quarterly indicator) | Numeric |
| conspriceidx | Consumer price index (monthly indicator) | Numeric |
| consconfidx | Consumer confidence index (monthly indicator) | Numeric |
| euribor3m | Euribor 3-month rate (daily indicator) | Numeric |
| nremployed | Number of employees (quarterly indicator) | Numeric |

**Target Variable**

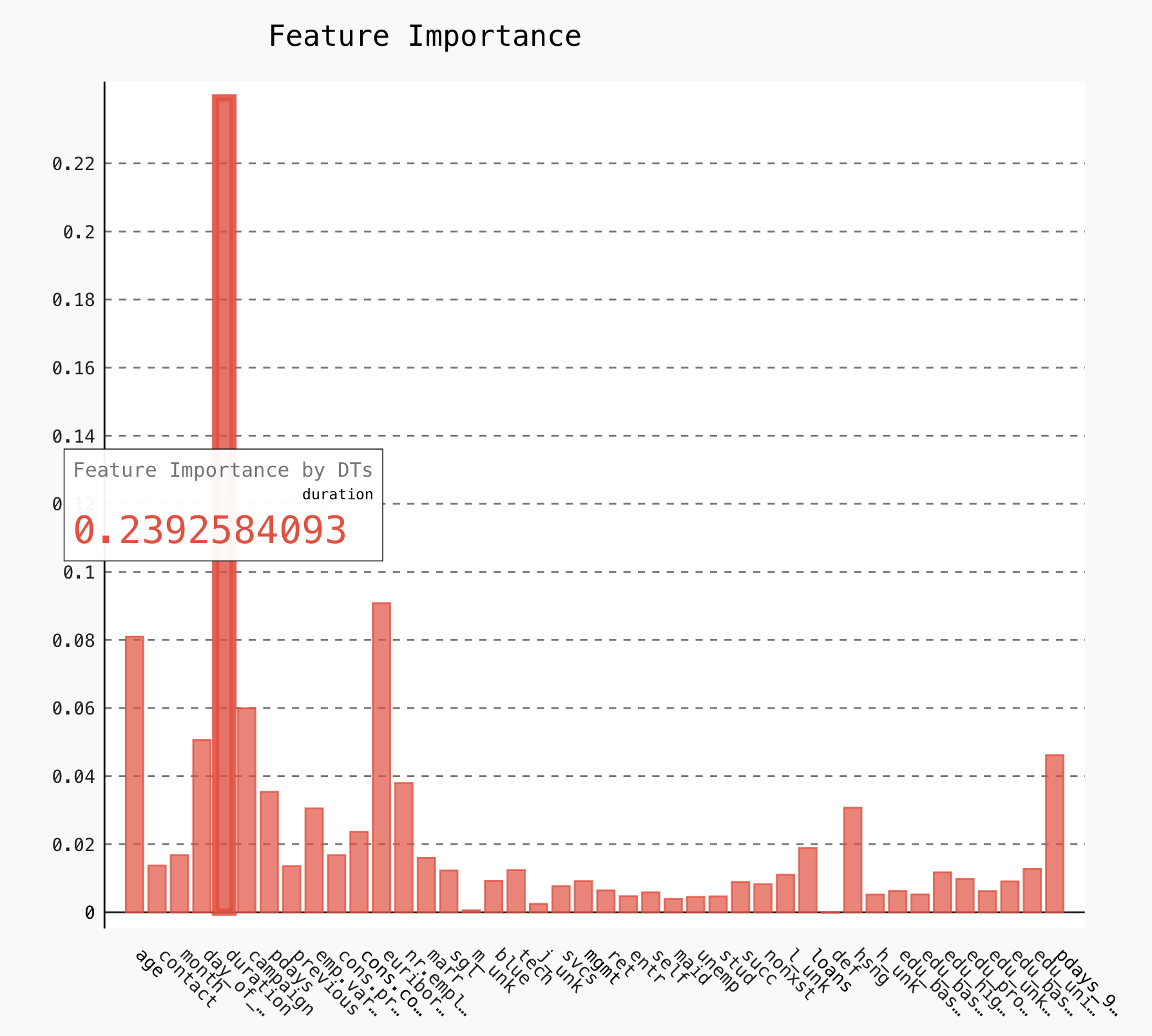
The target variable, named "Y" in the dataset, is binary and indicates whether the client has subscribed for a term deposit or not. Among the 41,188 customer data, 4,640 instances made term deposits with bank and 36,549 customers did not.

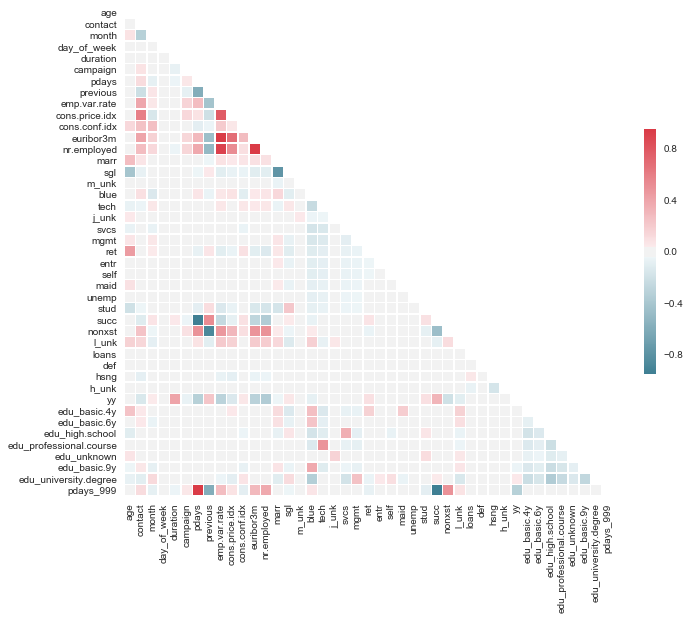
1. **Data Processing**

**Data cleaning**

1. Deal with textual data:
   * 1. Convert categorical variables into dummy variables: convert *marital, education, default, housing, loan, contact,* *poutcome,* and target variable *y* into dummy variables
     2. Convert time data into numeric: convert variable *month (jan/feb/…/dec)* and *day\_of\_week(mon/tue/.../fri)* into sequence of integers ordered by time.
2. Deal with missing values: In this dataset, the missing values only appears in the variable *pdays* as ‘999’.
3. Special Data Preparation:
   1. Data for SVM: On the one hand, when we are dealing with missing values, we created a new binary variable named ‘*pdays\_mv’* which value is ‘1’ if there is a missing value in ‘*pdays’,* instead of using ‘999’ to stand for missing values*.* Additionally, we take the mean for non-NaN to replace the missing value in order to normalize the cleaned data set to run SVM model efficiently. On the other hand, we normalized data in order to fit the SVM model efficiently.
   2. Data for Decision Trees model: No additional process on data. Fitted the model directly on the cleaned data set.
   3. Data for Logistic Regression model: Used the same method to deal with missing values. ?
4. Outliner and duplication check: Applied pandas duplicate function, no duplication detected. ………

**Feature exploration**

****



**Business Understanding**

* Identify and motivate the business problem that you are addressing.
* How (precisely) will a data mining solution address the business problem?

**Data Understanding**

* Identify and describe the data (and data sources) that will support data mining to address the business problem.  Include those aspects of the data that we routinely talk about in class and/or in the homework.

**Data Preparation**

* Specify how these data are integrated to produce the format required for data mining.
* Give a clear and precise definition of the target variable.
* Make a summary of any feature engineering that should be performed, which may include binning, non-linear transformations and domain knowledge based feature extraction.

**Modeling & Evaluation**

* Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?
* Identify an appropriate baseline model and report its performance.
* Describe an evaluation framework you will use to improve upon the baseline.
* Perform an analysis of possible algorithms and use the data science experimental framework to choose an

   optimal candidate.

* Demonstrate how you were able to improve upon the baseline and document the process of doing so.
* Discuss why and how this model should “solve” the business problem (i.e., improve along some dimension of interest to the firm).
* Discuss the type of evaluation metric that should be used to choose the best algorithm. How does this metric relate to the business problem?