Name: Meet Patel

DSC 441: Fundamentals of Data Science
Project

Problem Statement

Lot of resources are invested in marketing campaigns in every industry. To keep up with current trends in fintech which leads people to invest money in stocks, bonds rather than deposits lot of resources are allocated for marketing deposits offered by the banks.

Lot of information and customer segmentation can be found from the dataset. Using the dataset to classify whether the customers will subscribe to term deposit or not, the bank can,

- Pre-classify the customers based on the demographics and current financial data if they will be interested in term deposit.
- Reduce marketing cost by finding target customers.
- Whether the marketing be carried out after a certain number of contacts.
- Make changes in the mode of contact.

Data

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. Using classification, the goal is to predict whether the customer will subscribe to a term deposit (variable y).

Link: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing[1]

The dataset consists of 45211 rows and 17 columns. The 17 columns are broken down into 10 discrete columns and 7 continuous columns. The data is in Comma-Separated

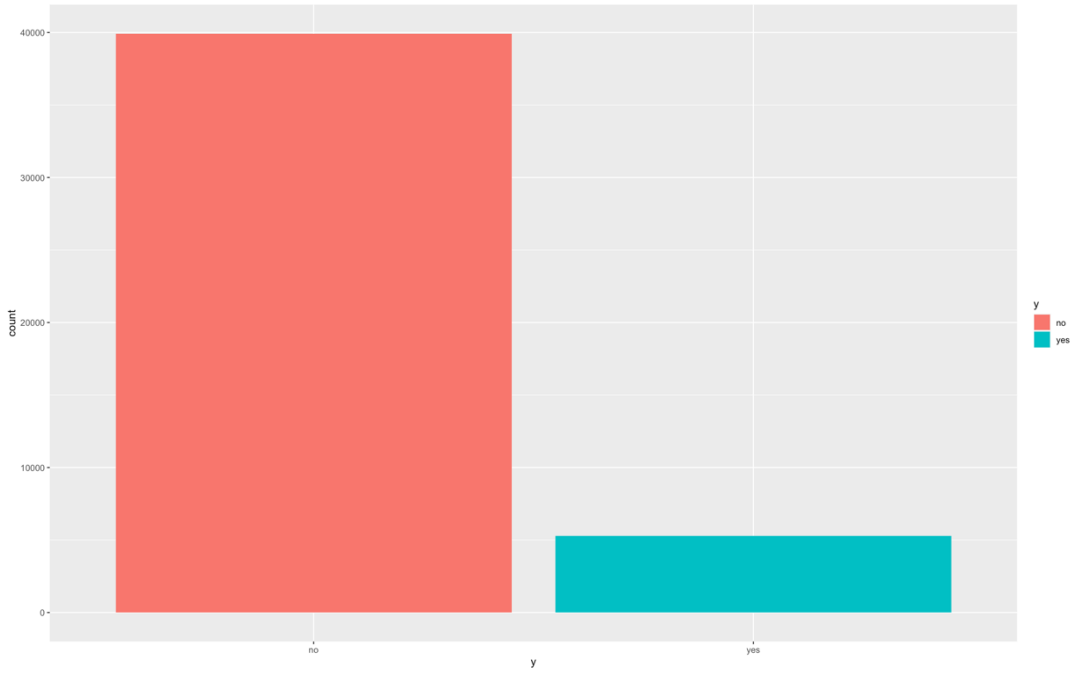
values file with a memory size of 4.7 MB. The data contains information that could be divided into several categories like customer demographics(age, job, marital status, education), financial data(credit default, balance, housing, loan) and campaign information(time, type and responses).

The dataset does not contain any missing values.

After plotting graphs of some variables for exploratory data analysis, several assumptions about the dataset can be derived,

1. The dataset is imbalanced. The target class has 88% values for majority class (No : Term Deposit) and remaining 12% values for minority class(Yes: Term Deposit).

```
> table(bank_factors$y)

   no   yes
39922  5289
```

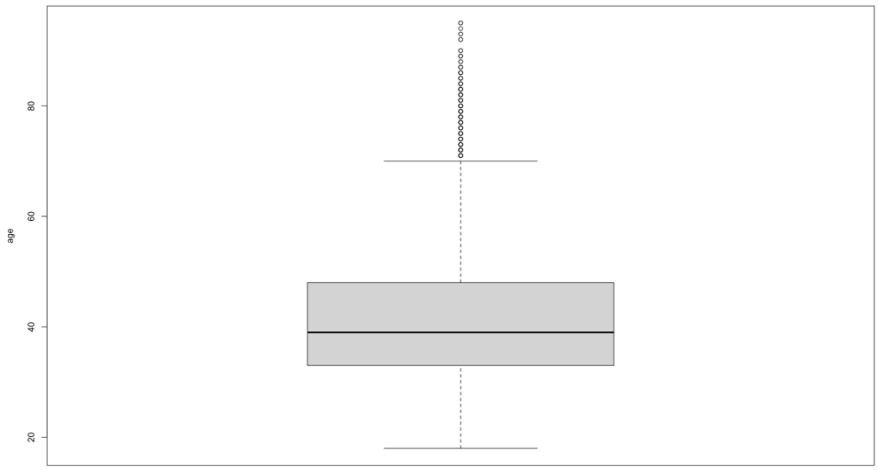


Bar graph summarizing number of customers subscribed and not subscribed

2. The two most interesting things that were discovered during Exploratory data analysis was about the outliers found for age and balance variables.

From the boxplots for age and balance, we may decide that the variables consist of outliers. However, when we look at the range of both the variables, we find that the outliers for age were people older than 65 years and for balance the values are as the average yearly balance could differ individually due to which some of the customers have negative balance and some of them have a very high balance.

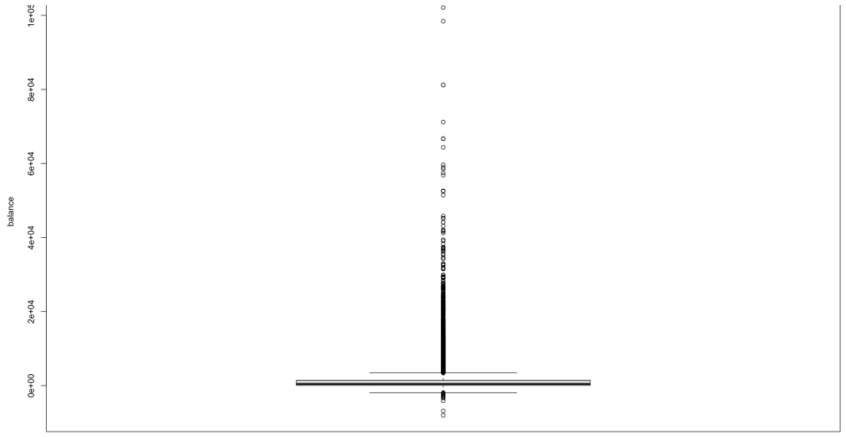
```
> summary(bank_factors$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   33.00   39.00   40.94   48.00   95.00
```



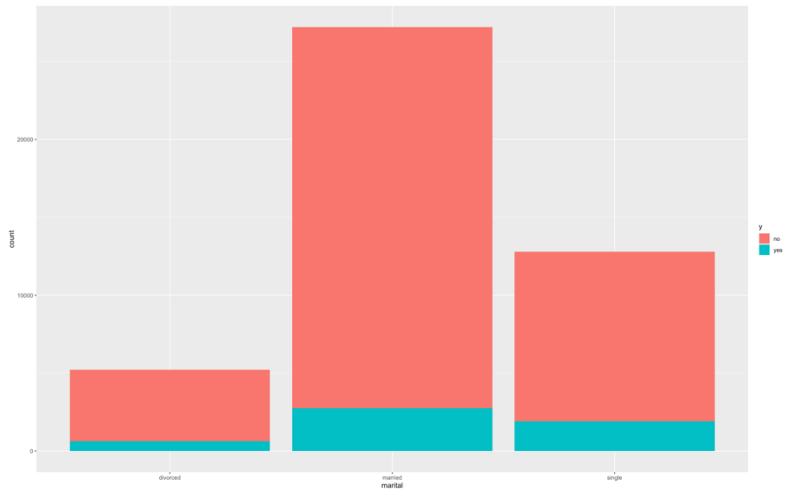
Age Box Plot

```
> summary(bank_factors$balance)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -8019      72     448    1362    1428  102127
```

Balance Box Plot

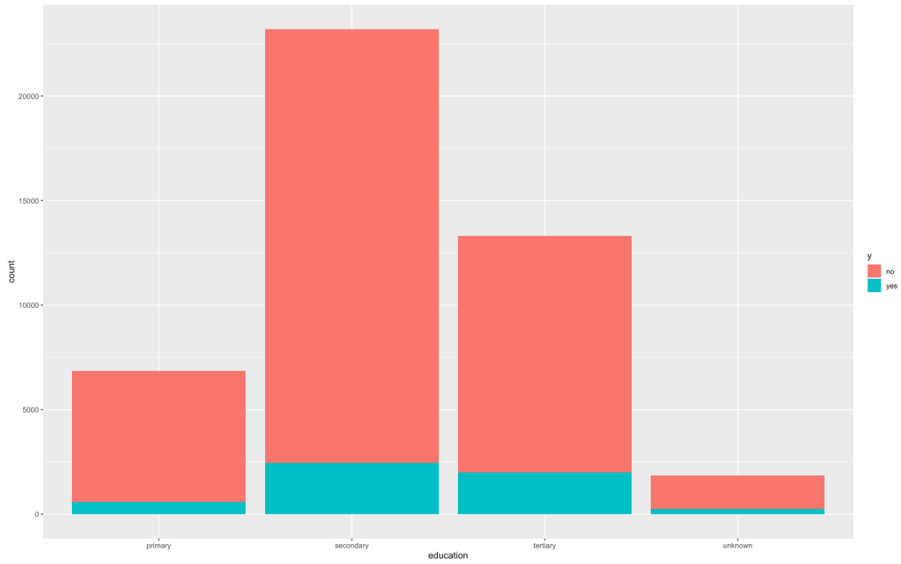


3. In terms of count married people were the most who did and did not subscribed to deposits.



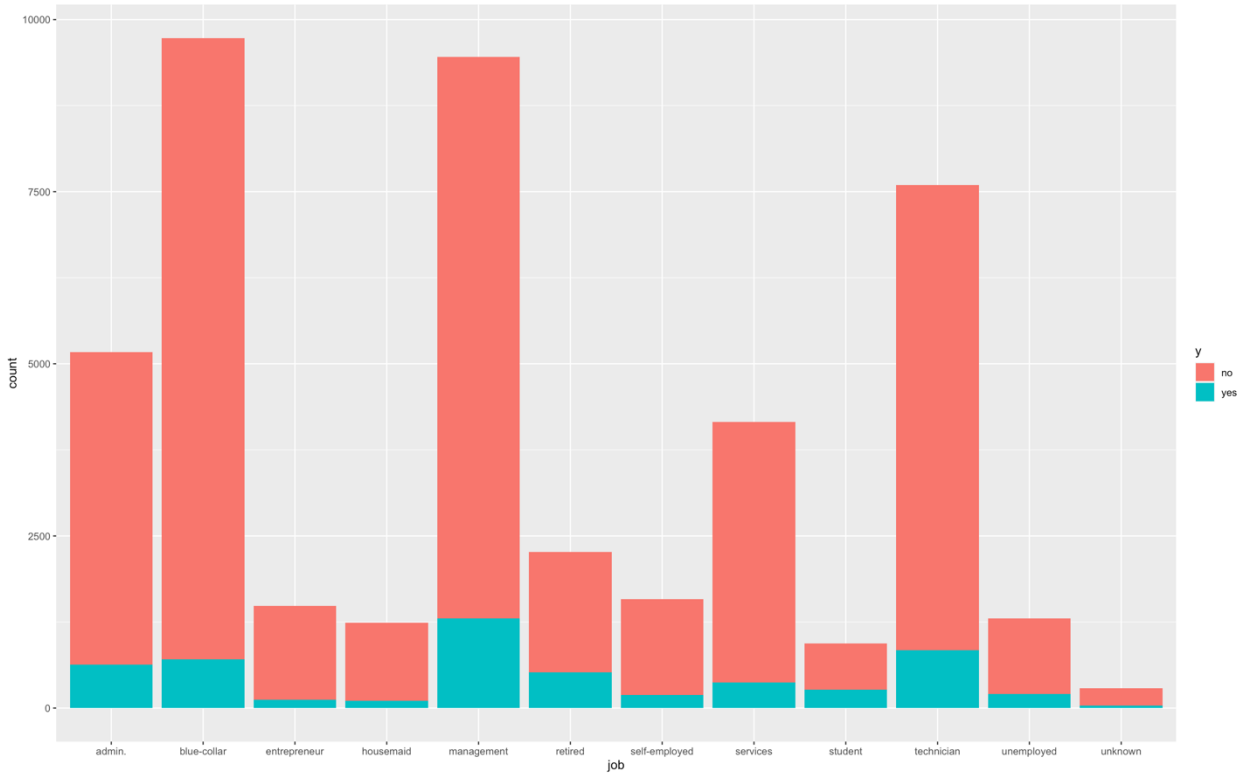
Deposit subscription based on Marital Status

4. Based on education customers with secondary and tertiary education subscribed to deposits.



Deposit Subscription based on education

5. When looking at the plot for job type/role for customer, an interesting finding was that people in blue-color jobs, management jobs, retiree, and technician subscribed most to the deposits, whereas entrepreneur, housemaid and self-employed customers did not subscribe as much. It can be assumed that entrepreneurs and self-employed people might be getting higher rate of returns from their business or people who are employed would like to have their savings invested in simple deposits that yield them yearly interest along with salaries from their jobs.



Deposit subscription based on job type

Data Preparation

After understanding the problem statement and results from exploratory analysis, the following steps were taken,

1. The categorical data were converted to factors. Although Decision tree and Naïve Bayes work with numerical and categorical data, to use the SMOTE function the categorical data was converted to factors using as.factor() function.

2. The age and balance variable showed outliers, but they convey important information, so no action was required.

3. The data was split into training and testing sets using createDataPartition function from caret package. The training set consisted of 70% of the data and the remaining 30% was allocated to testing set.

4. The dataset was imbalanced, so using SMOTE function the minority class was oversampled for training dataset.

Modeling

After understanding the problem faced by business and the results from exploratory analysis, it was decided that Decision tree and Naïve Bayes algorithms were the best to work with for the dataset. The main reason was the flexibility that both models offer, both the model work well with mixed data i.e., categorical and numerical data.

Distinguished reasons for choosing the two algorithms were,

- Decision Tree
    - Decision tree results are easy to interpret, and they could be express more efficiently in business terms to the business stakeholders. We can provide a reason for the occurrence of the outcome. For instance, a customer with X balance, Y job and Z age would opt for yearly deposit subscription.
    - Decision tree allows us to set business rules that help to determine the outcome based on specific needs.
    - Decision tree are not distance based so no scaling and normalization would be required.
    - Outliers have negligible effect on the algorithm.
    - They work with smaller number of classes.
    - One of the main problems of Decision tree is that of over-fitting and complexity of tree.

- Naïve Bayes
    - It assumes that every pair of features are independent to each other.
    - Requires small training datasets.
    - Naïve Bayes algorithm is pretty fast and robust as compared to other classifier algorithms

All the variables were selected for both the algorithms and no tuning of hyperparameters were used for both the algorithms.

Evaluation

1. Decision Tree:

The decision tree was built without tuning the hyperparameters. The decision tree has 4 levels and consists of duration, pdays, poutcome, month and day as features used for classification at nodes.

From confusion matrix we can evaluate the model by observing the Accuracy, Sensitivity, and Specificity as the dataset was balanced using SMOTE function. The accuracy for the model was 0.83, i.e., the model was able to guess 83% of items in each class. The sensitivity was .8433, which means that 84% of all the customers who did not opt for deposits were identified. The specificity was 0.7648 which means 76% of the data was correctly identified the customers that did not opted for deposits.

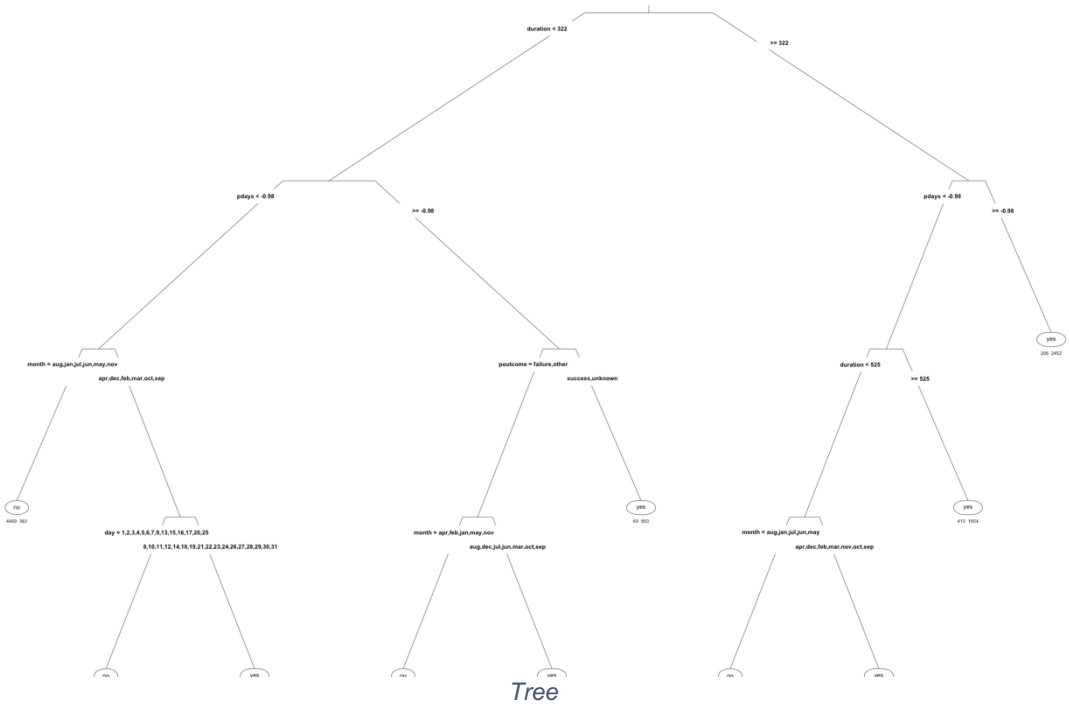
Tree pruning was not possible as it was found that the least error requires 10 splits which made the pruned tree same as the main decision tree.

From the decision tree, following variables were important,

- Duration: last contact duration
- Poutcome : outcome of previous marketing campaign
- Pdays: number of days passed after client was last contacted
- Previous: number of contacts performed before this campaign
- Month: last contact month of year
- Day: last contact day of month

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  10105   373
       yes  1871  1213

               Accuracy : 0.8345
                 95% CI : (0.8282, 0.8408)
    No Information Rate : 0.8831
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4317

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8438
            Specificity : 0.7648
         Pos Pred Value : 0.9644
         Neg Pred Value : 0.3933
             Prevalence : 0.8831
         Detection Rate : 0.7451
   Detection Prevalence : 0.7726
      Balanced Accuracy : 0.8043

       'Positive' Class : no
```

Decision Tree Confusion Matrix



Tree

2. Naïve Bayes:

The Naïve Bayes model was built tuning two main hyperparameters, one was *trainControl* which applied cross validation with 3 folds to resample the data for training and second set parameters were *fL*(control whether we add 1 to all cases to prevent a 0 probability), *usekernel*(a normal density is estimated) and *adjust*(1 smoothing).

Naive bayes is based on Bayes Theorem that assumes that each feature is independent and equal. The accuracy for model was .7965 i.e., the model was able to guess 79 % of items is each class. The sensitivity was 0.8020, which indicates that 80% of all the customers who said NO to deposits were identified. The specificity value was 0.7547 which indicates 75% of data that was correctly identified the customers that did not opted for deposits.

From the Naive Bayes model, following variables were found important,

- Duration: last contact duration
- Previous: number of contacts performed before this campaign
- Pdays: number of days passed after client was last contacted
- Poutcome : outcome of previous marketing campaign
- Balance: average yearly balance
- Housing: whether customer has housing loan

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  9605  389
       yes 2371 1197

               Accuracy : 0.7965
                 95% CI : (0.7896, 0.8032)
    No Information Rate : 0.8831
    P-Value [Acc > NIR] : 1

                  Kappa : 0.361

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8020
            Specificity : 0.7547
         Pos Pred Value : 0.9611
         Neg Pred Value : 0.3355
             Prevalence : 0.8831
         Detection Rate : 0.7082
   Detection Prevalence : 0.7369
      Balanced Accuracy : 0.7784

       'Positive' Class : no
```

Naive Bayes Confusion Matrix

Discussion and Conclusions

The models meet the original goal of the project to discover important features which decides whether customer opts for yearly deposit subscription or denies it based on the marketing campaign. There can be several assumptions drawn based on both the models that were created for the problem,

- Duration plays important role for subscription, when the marketing agent invests more time for each customer, they are likely to opt for deposit subscription. So banks can train the marketing agent to explain customers the deposit features in more details with pros and cons.
- Banks should focus more on previous customers who opted for deposit subscription. They are more likely to opt in for new subscriptions.
- The customer will likely opt in for deposit subscription, if they were contacted previously.

For approaching the problem in the future, one of the interesting variables to add would be the rate of interests on deposits. Also, I would like to set rules for decision tree by applying domain knowledge to the process. Furthermore, applying clustering to the dataset would help use to group the customers and decide the important variables in terms of categories(like demographics, finance, campaign data) . To improve the model efficiency and accuracy PCA can be applied to reduce the total predictor features in the dataset.

One of the most important things that I learned is that models are not build based upon the pros and cons of the algorithm, but they should be selected based on the business understanding. Several businesses required algorithms and data processing techniques based upon the problems. For this particular dataset, the age and balance features seemed outliers, but they help us divide our customers into age or income groups and helps us to invest time and resources to generate revenues.

Appendix:

[1] Citation for Database use:

[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

Input variables:

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

   # related with the last contact of the current campaign:
9. **contact:** contact communication type (categorical: "unknown","telephone","cellular")
10. **day:** last contact day of the month (numeric)
11. **month:** last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. **duration:** last contact duration, in seconds (numeric)

    # 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, -1 means client was not previously contacted)
15. **previous:** number of contacts performed before this campaign and for this client (numeric)
16. **poutcome:** outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

17. **y** - has the client subscribed a term deposit? (binary: "yes","no")