PREDICT THE SUCCESS OF TELEMARKETING DIRECT MARKETING – CASE 4

TEAM NUMBER: BADM_045
SUBMISSION NUMBER: BADM_1049

HARSH GUPTA
RIDHHI MEHTA
DOPPALAPUDI ABHISHEK

B.TECH IN DATA SCIENCE, 3RD YEAR, NMIMS MPSTME

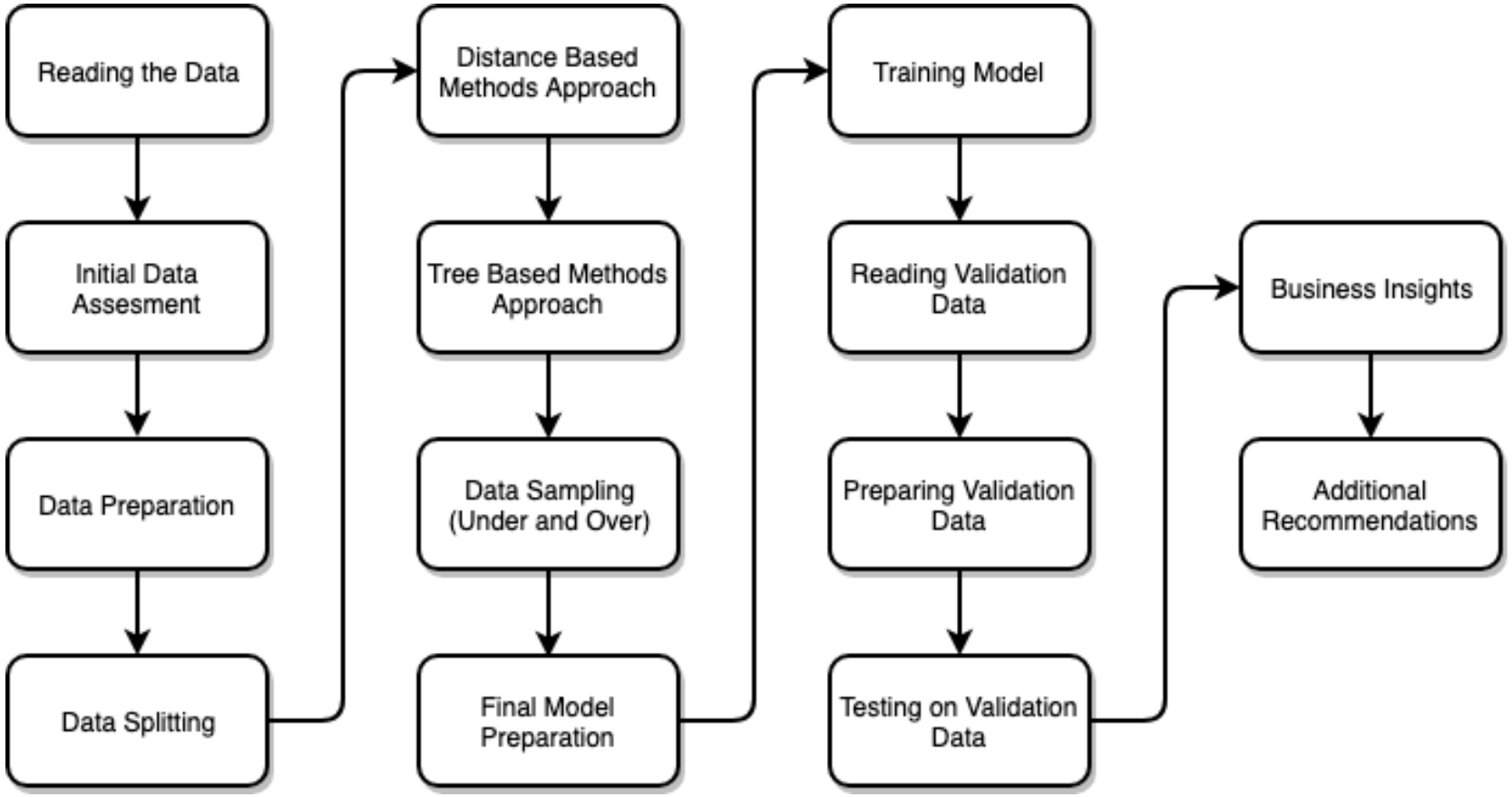
INTRODUCTION

- This following solution contains the results of our analysis on the dataset of a telecom company which used voice calls from call centers to pitch data packs to old customers and fresh targets.
- The goal of this report is to **EXPLAIN THE FACTORS THAT CAUSE A CHANGE IN CUSTOMERS BUYING DATA PACK and to PREDICT whether or not a customer will BUY the data pack**.
- We want to give you an unequivocal answer as to what the company should do in the future so in that spirit, we have paraphrased the goal and the meaning of the target variable as we have inferred it.
- We have curated this solution to try to explain or solve the problem for the telemarketing company.

PROPOSED METHODOLOGY

- Although the campaign is the same, there are **two different groups of customers the telemarketing company is trying to target**. It is evident from the data that it is easy to separate out customers who have been contacted for the first time and people who have already been exposed to a previous marketing campaign:
- Our motivation in doing so is because we have identified that the company has two different objectives when approaching these two different customer types –
  - **Customer retention, Customer churning and unresponsive targets**. (Already been exposed to previous campaigns)
  - **Customer acquisition**. (fresh targets)
- Customers have been identified as **fresh targets** and those who have **already been contacted** and have been **split the data** on the **poutcome** column into **two different groups:**
  - Group 1 – Category – unknown – because they have never been contacted before – fresh targets
  - Group 2 – Category – success, failure and other – repeated targets

SOLUTION MODEL

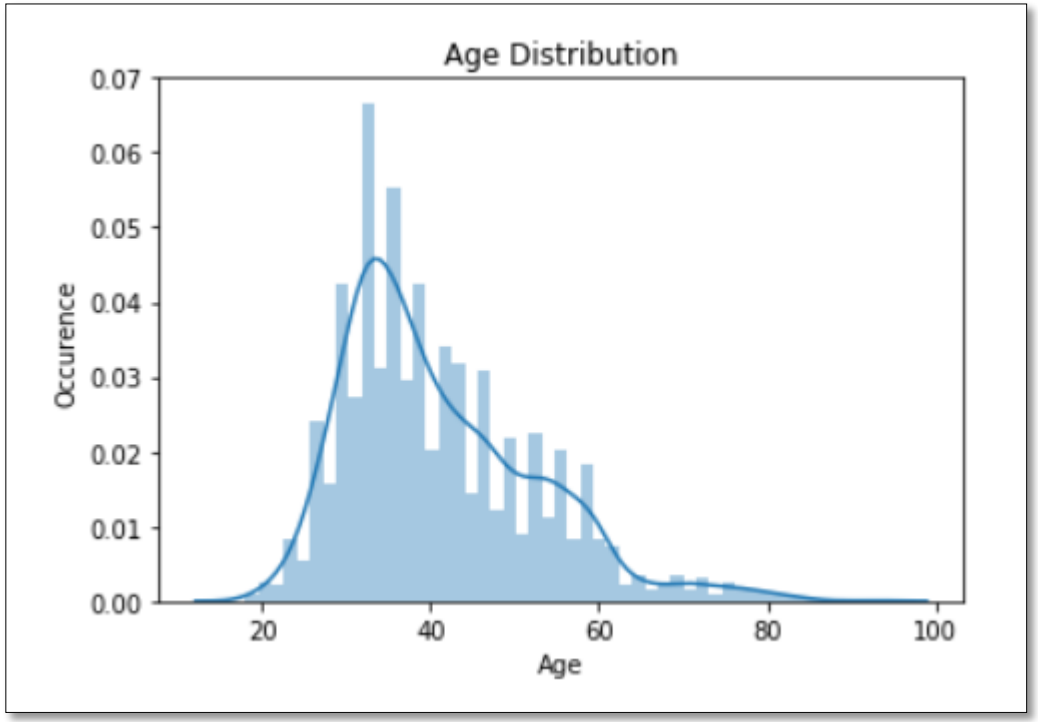


DATA CLEANING AND CONSOLIDATION

We have performed **three** very important transformations on the dataset before splitting it.

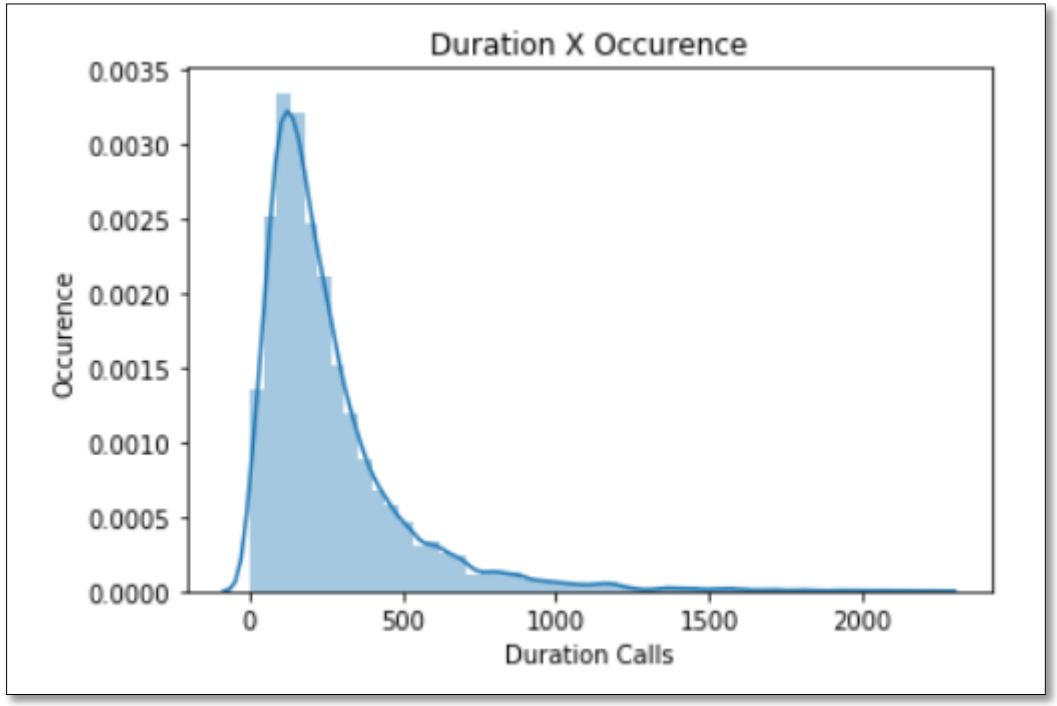
- **Campaign column**: The column includes the last call made to the customer. So, it is imperative that we subtract 1 to find out the number of times they have been contacted before the last call.
- **Duration column**: Our intention here is to compare two realistic models in predicting if a customer will subscribe to the data pack. The duration is not known before a call is performed. After the last call, the result is obviously known. Thus, "duration" could not be used as an independent variable and was excluded after the data was read in.
- **Age column**: The age column has been converted from a continuous variable to a categorical variable, so as to make it easier for interpreting it from a business perspective. Customers from the same age category are expected to display similar data pack usage characteristics

EXPLORATORY DATA ANALYSIS



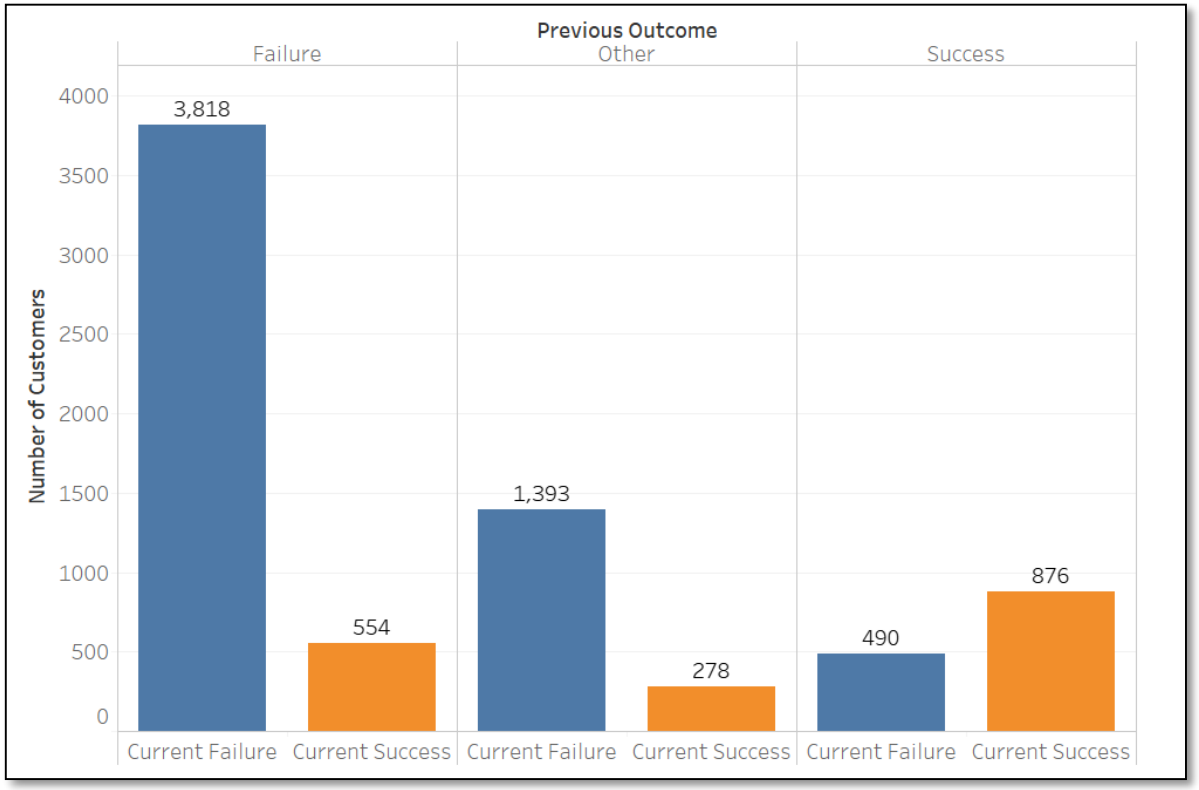
This graph helped us to decide the bins of the different clusters of age. We have focused more on the middle density of the distribution, whereas the data points after 4th quartile have been altogether condensed into one cluster. Although, it was a trial and error method because we wanted to be sure about the different combinations of the bins that could be made and used.

EXPLORATORY DATA ANALYSIS



Seeing this graph, it makes sense to eliminate the column of duration. It is highly skewed and there is one more important thing to note that the call duration for the most part is always going to be in a particular range. Moreover, this column is dependent on the last call, and it can get very subjective depending on the situation of the caller and the client in the last call. Hence we are planning to drop it.

EXPLORATORY DATA ANALYSIS



- High percentage of users who have given positive responses in the past are more likely to give positive responses in the current campaign
- Very high number of users tend to reject the offer of the current campaign if they have not subscribed to the deal in the previous campaigns
- We see that 490 users out of a total of 1366 that had given successful responses in the last campaign have not shown positive results in the current campaign. The churn percentage is 35.8%.

DATA PROCESSING

- We have not used any scaling on the dataset since:
  - All but two columns of the dataset are categorical, which are unaffected by scaling
  - We're building a tree based model whose performance is scaling independent.
- All categorical columns have been Label Encoded
- The dataset is split into two categories on the basis of the **poutcome** column as mentioned before

OBJECTIVES

- The crux of our problem statement in order to **predict the target is NOT JUST ACCURACY**.
- **Our primary objective is two-fold:**
  - **To identify important features in our model so that we can gain insights which can be implemented in business.**
  - **TO TACKLE THE VERY IMPORTANT PROBLEM OF CLASS IMBALANCE. The fundamental problem with the dataset is that 90% of the data is with target 0 and only 10% with target 1.**
- We have two types of wrong values:
  - False Positive – means the client HAS NOT SUBSCRIBED to the data pack, but the model thinks he did
  - False Negative – means the client SUBSCRIBED to the data pack, but the model said he hasn't.

METHODS USED – DISTANCE BASED

- The first model used is the Logistic Regression Model, the standard model to start with while building a supervised classification problem.
- Logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model.
- Since the dataset has a huge class imbalance, we have used Synthetic Minority Oversampling Technique (SMOTE) to increase the number of cases in data set in a balanced way
- After Logistic Regression, with the aim of still improving recall and accuracy we move on to tree-based methods.

METHODS USED - TREE BASED

- We have used Decision Trees as they provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options.
- Random Forest has been used to create the final model because:
  - It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
  - It runs efficiently on large databases.
  - It can handle thousands of input variables without variable deletion.
  - It gives estimates of what variables are important in the classification.

METHODS USED - TREE BASED

- Grid Search Cross Validation: The model has been improvised using Grid Search CV as we needed to evaluate the model using all the different and essential combinations of the hyper parameters. We also observe that, the recall as well as the precision increases when we select the best estimator given by grid search CV.
- Using **under-sampling** and **over-sampling** methods simultaneously to solve class imbalance problems:
  - Simply under sampling the overrepresented sample is leading to a massive information loss as there are only 10% of the small sample leading to loss of over 25,000 rows
  - Simply using smote to oversample the data in this case is CREATING 40% Synthetic data which is not ideal for the final model that we want to learn.
- **Hence, we have MANUALLY IMPLEMENTED THIS METHOD WHICH COULD HAVE BEEN DONE USING SMOTETOMEK or SMOTEENN.**

BUSINESS INSIGHTS BASED ON MODEL

- INTERPRETATION
- PROFILE TARGETING

ADDITIONAL RECOMMENDATIONS

- **Additional steps the telemarketing company should take in order to predict the target variable with better accuracy in the future.**
  - Service provider used by other family members – This data will be very useful because if there are family members who use the same service then the company can pitch different family pack schemes to them.
  - Location of user – This is one of the most important aspects that has been missed out in this dataset. With this information we can find out if the company has a leading position or trailing position with regards to the number of towers.
  - Time when the call is made to the customer/potential customer – This is very important because we can infer from the JOB COLUMN about the rough schedule of each person and avoid calling them when he or she is busy. We can also analyse the times when customers are the most responsive.
  - **This is for customers only** – Number of calls made to complain about the service. With this data the company can then analyse how to retain its customers better.

THANK YOU