Department of Management Sciences - IIT Kanpur

DMS 672 – Data Mining

Project Report

Group 3 - Bias in Advertising Data

Submitted to:
Dr. Faiz Hamid
Associate Professor
DOMS IIT Kanpur

Name	**Roll Number**
Akash Tripathi	231250015
Anshi Srivastav	241140602
Devaguptapu Rama Krishna Sandilya	231250043
Varun Pratap Singh Chauhan	231250161

1. Introduction

Overview of the problem statement

Classification of users/customers conversion based on an advertisement shown.

Brief description of the dataset and objectives:

Dataset contains synthetic generated data for users who were shown a certain advertisement (ad). Each instance of the dataset is specific to a user and has feature attributes such as gender, age, income, political/religious affiliation, parental status, home ownership, area (rural/urban), and education status.

In addition to the features, it also contains information on whether users actually clicked on or were predicted to click on the ad. Clicking on the ad is known as conversion, and the three outcome variables included are: (1) The predicted probability of conversion, (2) Predicted conversion (binary 0/1) which is obtained by thresholding the predicted probability, (3) True conversion (binary 0/1) that indicates whether the user actually clicked on the ad.

2. Exploratory Data Analysis (EDA) and Data Preprocessing

EDA / Preprocessing Steps:

- Extracting Unique Values in each of the Features of the Dataset and finding the number of Unknown / Missing values in each column.
- Finding Occurrences / Data Points where the Conversion differs despite being the same on all other features and comparing True vs. Predicted Conversion from the Dataset.
- Finding the threshold for *predicted_probability* where *predicted_conversion* is 1 happens to be 0.3658223221090007
- We dropped the column of predicted_probability and also dropped the duplicate rows in the original dataset. (1443140 rows reduced to 2591 rows)
- As the missing data is very high, instead of dropping the column, we are replacing the values of "Unknown" in the dataset based on the proportion of other unique values in that specific column / feature of the dataset

Bivariate Analysis

- The highest conversion rate is observed among college-educated homeowners, while the lowest conversion rate is among non-college-educated non-homeowners.

- College-educated individuals in both areas have notably higher conversion rates, suggesting that education level significantly influences conversions regardless of the area.
- While urban areas have higher representation, the likelihood of conversion does not differ significantly across gender or area
- Regardless of Income (<100k or >100k), Younger age groups (18-24) have significantly lower conversion rates)
- Political affiliation and religion both impact conversion rates, with "Others" generally showing higher average conversion rates than "Christianity" across various political categories, especially in "Liberalism." This suggests that individuals with certain religious and political affiliations may respond differently to conversion opportunities.

3. Data Mining Algorithms and Result Analysis

The Summary of Data Mining Algorithms used as part of the Project are:

- Decision Tree Model (Pre-Pruned & Post-Pruned)
- Bagging
- Random Forest
- Naive Bayes Classifier
- Artificial Neural Network
- XGBoost

The metrics used to evaluate the performance of these classification models which are also indicated as part of the report are

- Accuracy of the Model
- Receiver Operating Characteristic Curve
- Confusion Matrix
  - True Positive Rate / Sensitivity / Recall - Predicting correctly that a customer will be converted
  - True Negative / Specificity - Predicting correctly that a customer will not be converted
  - False Positive - Predicting incorrectly that a customer will be converted
  - False Negative - Predicting incorrectly that a customer will not be converted
  - Precision & F-Score

From the perspective of the importance, the most important metrics to assess the model performance shall be

- Lower values of False Negative Rate (to not lose out on potential customers)
- High values of True Positive Rate (if required can give strategic discounts or design products as appropriate)
- Several strategies can be thought of and built around each group of customers to enhance the conversion rate.

Comparison of Performance of Models

Performance Metric	**Original Dataset (Data Imbalance)**	**Balanced Dataset (Post SMOTE)**	**Balanced Dataset (Post ADASYN)**
Decision Tree Classification Model			
ROC	0.55	0.66	0.86
Accuracy	89%	62.23%	80.96%
Error Rate	11%	37.77%	19.04%
Sensitivity / Recall	0%	62.86%	91.71%
Specificity	97%	61.6%	70.46%
Precision	0	59.79%	75.04%
F-Score	0	0.6128	0.8254
Pre-Pruned Decision Tree			
ROC	0.58	0.66	0.86
Accuracy	91.8%	62.23%	80.96%
Error Rate	8.2%	37.77%	19.04%
Sensitivity / Recall	0%	62.86%	91.71%
Specificity	100%	61.6%	70.46%
Precision	-	59.79%	75.04%
F-Score	0	0.6128	0.8254
Post-Pruned Decision Tree			
ROC	0.58	0.66	0.86
Accuracy	91.8%	62.66%	81.17%
Error Rate	8.2%	37.34%	18.83%

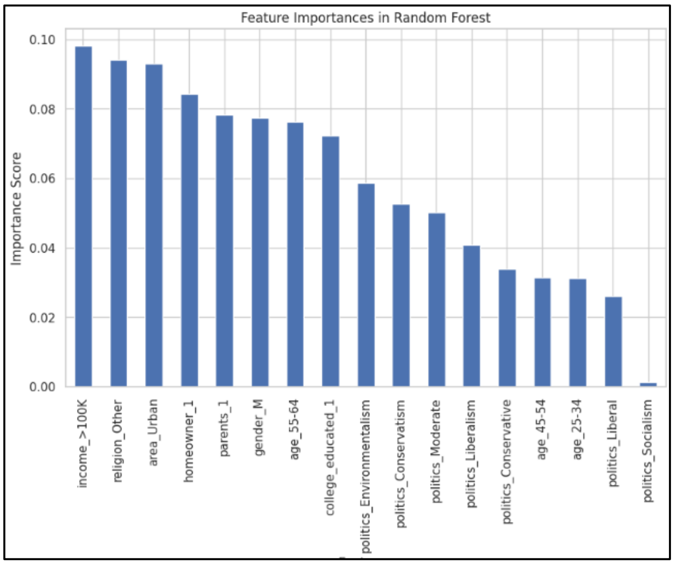
Sensitivity / Recall	0%	62.64%	93.07%
Specificity	100%	62.63%	69.62%
Precision	-	60.34%	74.78%
F-Score	0	0.6147	0.8293
Bagging Classifier			
ROC	0.65	0.66	0.89
Accuracy	91.8%	-	81.28%
Error Rate	8.2%	-	18.72%
Sensitivity / Recall	0%	-	92.21%
Specificity	100%	-	70.71%
Precision	-	-	75.26%
F-Score	0	-	0.8288
Random Forest Classifier			
ROC	0.64	0.66	0.89
Accuracy	91.8%	-	80.85%
Error Rate	8.2%	-	19.15%
Sensitivity / Recall	0%	-	92.21%
Specificity	100%	-	69.83%
Precision	-	-	74.74%
F-Score	0	-	0.8256
Naive Bayes Classifier			
ROC	0.64	-	0.65
Accuracy	10.65%	-	60.74%
Error Rate	89.35%	-	39.26%
Sensitivity / Recall	100%	-	65.8%
Specificity	2.79%	-	55.86%
Precision	8.3%	-	59.03%
F-Score	0.153	-	0.6223

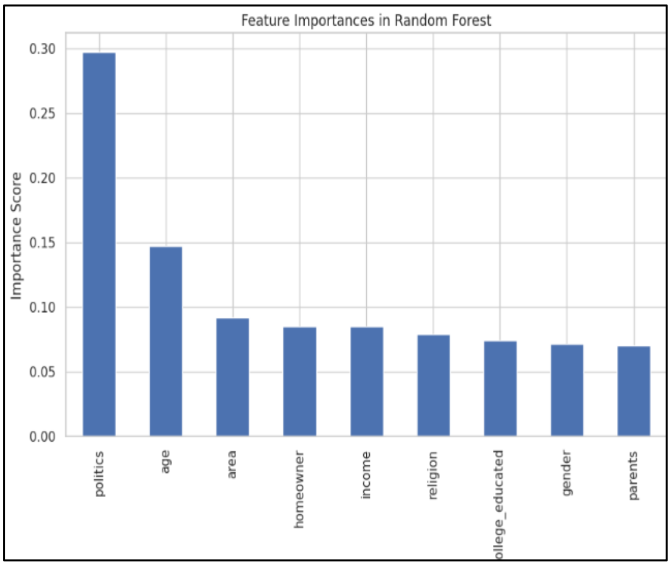
Artificial Neural Network			
ROC	0.49	0.66	0.87
Accuracy	92.1%	56.81%	77.34%
Error Rate	7.9%	43.19%	22.66%
Sensitivity / Recall	0%	90.67%	96.54%
Specificity	100%	24.03%	58.79%
Precision	-	53.58%	69.36%
F-Score	0	0.6736	0.8072
XG Boost			
ROC	0.64	0.66	0.89
Accuracy	92.1%	61.81%	82.13%
Error Rate	7.9%	38.19%	17.87%
Sensitivity / Recall	0%	64.72%	91.56%
Specificity	100%	58.91%	73.01%
Precision	-	60.3%	76.63%
F-Score	0	0.6243	0.8343

Inferences from Models Performance Comparison

- XGBoost model has given the best performance when trained with a balanced dataset with regards to all the metrics in consideration
- The performance of all the models is very poor on the original dataset, i.e., the unbalanced dataset. All the models are nearly performing to a random guess as it can be seen through the confusion matrices too.
- The performance of the models has not improved in the case of the dataset developed through the SMOTE algorithm despite establishing a balance in the target class data.
- The performance of the models has been significantly improved through the dataset generated by the ADASYN algorithm
- We can see that the feature importance has changed in predicting the true_conversion.
    - Initially the top 3 features are Income, Religion, and Area and also the score of importance is only around 10%

- However, in the synthetically developed balanced dataset, the features of importance have changed to Politics, Age, and Area and also the score of importance for politics is around 30%





- The Hyperparameters in each of the models are tuned and implemented to get the maximum performance in each of the cases