BANKING TELEMARKETING CAMPAIGN - A CASE STUDY

INTRODUCTION

Today's world revolves around data and the valuable insights it provides. Making wise use of data helps a company manage its resources and improve its performance in various areas. Marketing is one such area where companies invest a lot of resources into. The ultimate goal of marketing is to influence customers to buy your product or use your service. Making use of data driven technology can be a game changer in this area.

The finance sector is one of the sectors that has been most affected by recent developments in machine learning. If it's forecasting market prices or, in this case, determining whether a customer will sign up for a term deposit, Machine learning has the potential to be a very helpful tool for increasing profitability.

In our project, we studied, analyzed data and made a classification model using existing Machine Learning algorithms.

PROBLEM STATEMENT

AB Bank is a large public sector bank which has branches across the cities in North America. It provides various services like savings accounts, current account, term deposits, personal loans, home loans etc. to customers. Whenever the bank conducts marketing on its new schemes, it will keep track of data related to customers' personal, social and economic details. Also, it maintains the detailing on efforts made to achieve success in the campaign.

Recently, the bank has conducted a campaign to market their term-deposit scheme. Campaigns were conducted based mostly on direct phone calls, soliciting the bank's customers to place a term deposit. After all the marketing efforts, if the client had agreed to place a deposit, then the campaign is successful, otherwise not (Target variable marked 'yes', or 'no').

ABOUT DATA SET

The data gives specifics regarding the bank's client information, data pertaining to the most recent campaign contact, data on social and economic context characteristics

Bank client data:

1. Customer id : Unique customer id
2. custAge: Age of the customer.
3. profession: type of job

4. marital: marital status
5. schooling: Educational qualification
6. default: has credit in default?
7. housing: has a housing loan?
8. loan: has a personal loan?
9. State_Code: Code representing unique state name
10. Region_Code: Code representing unique Region name
11. City_Code: Code representing City of the customer
12. Postal_Code: Postal code of the area to which the customer belongs to.

Data related with the last contact of the current campaign:

1. contact: contact communication type
2. month: last contact month of year
3. day_of_week: last contact day of the week
4. campaign: number of contacts performed during this campaign and for this client (includes last contact)
5. pdays: number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
6. previous: number of contacts performed before this campaign and for this client
7. poutcome: outcome of the previous marketing campaign
8. duration: duration of the last call

Data related to social and economic context attributes:

1. emp.var.rate: employment variation rate - quarterly indicator
2. cons.price.idx: consumer price index - monthly indicator
3. cons.conf.idx: consumer confidence index - monthly indicator
4. euribor3m: euribor 3 month rate - daily indicator
5. nr.employed: number of employees - quarterly indicator

INSIGHTS

1. The dataset contained few anomalies that include
    a. A city being present in multiple states
    b. A state being present in multiple regions

    Using correct geographical data helped us get rid of the anomalies

2. The dataset we used in this project also had several null values. Details of customers like their job, education, loan details and default credit in their accounts were left unfilled. In our project we made use of KNN Imputer to fill these details.

3. Our findings from exploratory data analysis are

a. The target variable having class imbalance. The ratio of yes to no is 1:8
b. All the categorical variables except “poutcome”, i.e success of previous campaign had very little impact on the target variable.
c. Economic factors like euribor rate, consumer confidence index, showed significant influence on the target variable including contact information like duration of previous call, number of days that passed by after the client was last contacted, number of contacts performed before this campaign.

TECHNIQUES USED IN PROJECT

1. Encoding Techniques: Ordinal Encoder, Dummies
   Categorical columns in the dataset are encoded using Ordinal Encoder and Dummy variables.
   The columns with categories more than 5 like “City_Name”, “State_Name” and “job” are encoded using Ordinal Encoder.
   Columns like “day_of_week”, “month”, “education” are manually mapped to numeric values.
   Other columns like “marital”, “loan”, ”default”, “housing”, “contact”, “Region_Name”, “poutcome” and “y”.

2. Imputation Techniques: KNN Imputer
   Non-existent values in the dataset are imputed using KNN Imputer. KNN Imputer unlike other imputers, imputes missing values with the mean on n nearest neighbors and not just the mean of the attribute.

3. Standardization: RobustScaler, StandardScaler
   RobustScaler is used to scale features since it is robust to outliers. StandardScaler is used to standardize features for individual models and models with PCA.

4. Feature Selection Techniques: Recursive Feature Elimination, Principal Component Analysis
   It is important to go for Feature Selection Techniques if most of the attributes do not have strong influence on the target variable. In our project, we used both Recursive Feature Elimination and Principal Component Analysis in combination with Classification models. Both of these techniques gave better results than the models alone.

5. Oversampling Techniques: Synthetic Minority Oversampling Technique
   The dataset we worked with suffered from class imbalance. This problem was solved using Oversampling the data and making the classes balanced.
   Oversampling the training data helped models learn better and produced more True Positives than models trained with original data.

MODEL OVERVIEW

The objective of this project is to predict customers who will subscribe to the bank's term deposit, i.e, give a positive response for the campaign. Here, generating maximum True Positives is the goal of the predictive model. A bank can spend more resources on customers that might eventually after contacting may say no, but cannot afford to miss out customers that might respond positively. Hence, the performance metric used to compare models here is Recall.

Models			
MODEL	ACCURACY	PRECISION	RECALL
Logistic Regression	91.05	70.68	40.46
Dtree	90.77	60.88	59.77
XGBoost	90.94	63.71	52.87
KNN	84.77	26.49	16.78
SVM	89.69	68.02	22.98
Random Forest	90.398	75.07	27.01
RFE with Logistic Regression	90.29	66.81	34.25

Models with SMOTE			
MODEL	ACCURACY	PRECISION	RECALL
Logistic Regression	84.69	42.49	86.2
Dtree	80.62	37.07	93.44
XGBoost	90.86	62.94	53.67
KNN	83.12	40.04	87.12
SVM	83.86	41.09	86.78
Random Forest with smote	84.29	39.9	67.01
Logistic Regression with RFE	84.27	41.76	86.32

Models with PCA and SMOTE			
MODEL	ACCURACY	PRECISION	RECALL
Logistic Regression	84.47	42.02	85.29
Dtree	79.37	33.45	81.66
XGBoost	88.32	50.17	69.88
KNN	83.82	40.23	76.55
SVM	82.44	38.97	87.81
Random Forest with pca and smote	88.89	51.78	76.89

RESULTS

Models trained on oversampled data performed better than models trained on data suffering from class imbalance. It is observed that Logistic Regression, KNN Classifier and Decision Tree models produced maximum True Positives when trained on oversampled data.
Decision Tree showed the best recall score of 93 percent and is hence deployed.

col_0	0.0	1.0
target		
0.0	5340	1207
1.0	72	798