Project – Propensify

Submitted by – Gayathri Keerthivasagam

Problem Statement

Are you aware of what, when, and why your customers will make a purchase? Many businesses undertake an intense pursuit to discover these answers, dedicating valuable resources to data-driven campaigns and high-cost strategies - yet the actual outcomes often remain elusive and disappointing.

Customer information is considered to be a valuable asset. Propensity modeling is a method that aims to forecast the chance that individuals, leads, and customers will engage in specific actions.

Suppose you are working for a company as a Data Scientist. Your company is commissioned by an insurance company to develop a tool to optimize their marketing efforts.

This project is aimed at building a **propensity model to identify potential customers**.

Data:

The insurance company has provided with a historical dataset (train.csv).

The company has also provided with a list of potential customers to whom to market (test.csv). From this list of potential customers, you need to determine yes/no whether you wish to market to them. (Note: Ignore additional columns such as 'profit' and 'id').

The dataset for this project can be accessed from Propensify.zip

Success Metrics :

Below are the metrics for the successful submission of this case study.

- The accuracy of the model on the test data set should be > 85% (Subjective in nature)
- Add methods for Hyperparameter tuning.
- Perform model validation.

Let's delve deeper into the project's details and objectives.

Process Overview: Machine Learning Model Implementation

1. Data Collection and Exploration
2. Data Cleaning and Preprocessing
3. Feature Engineering and Selection
4. Dealing with Imbalanced Data
5. Model Selection
6. Model Training and Evaluation
7. Hyperparameter Tuning
8. Model Validation and Deployment
9. Predictions performed for test file
10. Reporting and Documentation

Let's explore each of the outlined steps in detail.

1.Data collection and Exploration

- The dataset was sourced from the provided link, and the data description is mentioned below in the table.
- A comprehensive exploration of the dataset was conducted, meticulously examining each feature in detail.
- Different visualization techniques such as histograms, box plots, pair plots, count plots are used to understand relationships between features and target variable.

Insurance client data:

Column	Description	Sample data
custAge	Age of the customer	18 to 94
profession	type of job	‘admin.','bluecollar', 'technician', 'services', 'management', 'retired','entrepreneur','housemaid', 'self employed' ,'unemployed', 'student','unknown'
marital	marital status	'divorced','married','single','unknown'
schooling	Education level of the customer	'basic.4y','basic.6y','basic.9y','high.school','illiterate', 'professional.course','university.degree',' unknown'
default	has credit in default?	‘yes’,’no’,’unknown’
housing	has housing loan?	‘yes’,’no’,’unknown’
loan	has personal loan?	yes’,’no’,’unknown’

Related with the last contact of the current campaign:

Column	Description	Sample data
contact	contact communication type	'cellular','telephone'
month	last contact month of year	‘mar’,’apr’….’dec’
day_of_week	last contact day of the week	'mon','tue','wed','thu','fri'

Other attributes:

Column	Description	Sample data
campaign	number of contacts performed during this campaign and for this client	1 to 40
pdays	number of days that passed by after the client was last contacted from a previous campaign	0 to 25,999(Not contacted)
previous	number of contacts performed before this campaign and for this client	0 to 6
poutcome	outcome of the previous marketing campaign	'failure','nonexistent','success'
pmonths	number of months that passed by after the client was last contacted from a previous campaign	0 to 0.7,999(Not contacted).
pastEmail	Number of previous Emails sent to the customer	0 to 25

Social and economic context attributes:

Column	Description	Sample data
emp.var.rate	employment variation rate - quarterly indicator	-3.4 to 1.4
cons.price.idx	consumer price index - monthly indicator	92.2 to 94.7
cons.conf.idx	consumer confidence index - monthly indicator	-50.8 to -26.9
euribor3m	euribor 3 month rate - daily indicator	0.6 to 5
nr.employed	number of employees - quarterly indicator	4963.6 to 5228.1

Output variable (desired target):

Column	Description	Sample data
responded	Did the customer respond to the marketing campaign and purchase the policy	‘yes’,’no’

2.Data Cleaning and Preprocessing

- Missing values in the *'custAge'* column have been imputed with the mean value based on the profession. Considering that professions encompass a diverse range including students and retirees, this approach is deemed suitable for accurately estimating missing ages.
- For the *'day_of_week'* feature, missing values have been filled with the mode, ensuring consistency and preserving the distribution of weekdays.
- Regarding the *'schooling'* attribute, missing values have been imputed with the mode corresponding to each profession. Given the strong correlation between profession and schooling, this strategy ensures that the imputed values align closely with the characteristics of each profession.
- Duplicates are successfully identified and removed from the training dataset.

3.Feature Engineering and Selection

- Feature engineering has been conducted on several features including *'pdays'*, *'pastEmail'*, and *'custAge'*. Labels have been assigned to delineate ranges of values, facilitating better comprehension and analysis of these features. The dataset was analyzed to identify numerical, categorical, and binary features.

Correlation Analysis

- After conducting the correlation analysis, it was observed that features such as *employment rate*, *number of employees*, and *euribor 3 month rate* exhibit a high correlation with the target variable 'responded'.

Feature Selection

- Additionally, there was a notable correlation between the features *'pmonths'* and *'pdays'*. Upon further investigation, it was discovered that *'pmonths'* essentially represents the duration indicated by *'pdays'* in terms of months.
- For instance, if *'pdays'* equals 6, then *'pmonths'* equals 6/30, which equals 0.2. As both features convey the same temporal information, the decision was made to drop the *'pmonths'* column.

Skewness

- Skewness was calculated for numerical features, and log transformations were applied to positively skewed features. Outliers were visualized, revealing outliers in the *'campaign'* feature.

Encoding and Scaling

- The target variable '*responded'* is encoded as class 0 and class 1. Categorical features were encoded using both one-hot encoding and label encoding techniques, depending on the nature of the categorical values.
- Numerical features were scaled to maintain data normalization. Scaling numerical features helps maintain data normalization, ensuring that all features contribute equally to the model's learning process.

4.Dealing with Imbalanced Data

- Techniques used for balancing the data:
  i. SMOTE
  ii. Random Undersampling
  iii. Random Oversampling
- A detailed comparative analysis was conducted for different resampling techniques, revealing that *SMOTE* outperforms others in terms of both accuracy and precision.

5.Model Selection

- ML models used are
  i. Logistic regression
  ii. K Nearest Neighbors
  iii. Support Vector Machine
  iv. Random Forest Classifier
  v. Gradient Boosting.
- A comprehensive comparison of various classifier models was undertaken, indicating that *Gradient Boosting* emerged as the most suitable machine learning model for this marketing dataset.

6.Model Training and Evaluation

- All selected models were trained and metrics used are
  i. Accuracy
  ii. Precision
  iii. Recall
  iv. F1 Score
  v. Area Under the ROC Curve (AUC).
- Among them, Gradient Boosting with SMOTE technique exhibited higher precision and achieved an accuracy greater than 0.85, meeting the project's success criterion.
- Given the highly imbalanced nature of the data, further analysis will proceed with Gradient Boosting + SMOTE.

7.Hyperparameter Tuning

- A grid search is conducted to find optimal parameters:
  i. n Estimators
  ii. Learning Rate
  iii. Max Depth

- Following the search, the best estimator and model are identified.

- Finally, the code prints the best parameters, score, estimator, and model, facilitating analysis and interpretation.

8.Model Validation and Deployment

- Thus, the trained model is rigorously evaluated using various validation techniques such as cross-validation, train-test splitting, and performance metrics assessment. This ensures that the model generalizes well to unseen data.
- Feature importance is visualized which reflects the overall contribution of the feature to the predictive performance of the model. Features with higher importance scores are considered more influential in making predictions.
- The trained model is saved in a pickle file for future use and applied to the preprocessed test data, following the same preprocessing steps as the training data.
- This allowed to identify potential customers for positive responses in the test dataset efficiently.

9.Predictions performed for test file

- After employing the deployed model, predictions were made for the target variable *'responded'* in the test file.
- Additionally, a column was appended and saved as '*test_predictions.xlsx',* indicating for each observation whether marketing to that customer is preferred, denoted by 1 (yes) or 0 (no).

10.Reporting and Documentation

- All Python notebooks covering data preprocessing, sampling techniques, model selection, hyperparameter tuning, models, visuals and a detailed README.md have been thoroughly documented.
- Packaged the solution in a zip file included with a README that explains the installation and execution of the end-to-end pipeline.
- Additionally, a comprehensive PDF titled 'Documentation on Propensify' encapsulating the entire machine learning end-to-end pipeline has been prepared and is attached herewith.

Future work

- Perform customer segmentation analysis to identify distinct customer groups based on their characteristics and behavior.
- Explore how different customer segments respond to marketing campaigns and tailor the propensity model for each segment.
- Develop an interactive dashboard that allows the insurance company to visualize and explore the propensity model's results.

Instructions for running the code:

Important Note: Please ensure to update the file paths in both of the below mentioned python notebook for the train and test datasets to reflect the local directory structure on your machine before running the code.

In this project, two Python notebooks were utilized.

1.The notebook titled *'Machine_learning_model_implementation.ipynb'* encompasses preprocessing, exploratory data analysis, model selection and the identification of the optimal machine learning model for the dataset.

2. The notebook named *'Marketing_source_code_pipeline.ipynb'* illustrates the end-to-end pipeline for best identified model training, including saving the model as a pickle file. Furthermore, predictions were made for the target variable 'responded' in the test file, thereby identifying potential customers for the company.

Please run the *'Marketing_source_code_pipeline.ipynb'* file to get the predictions.

Thanks for reading!!!

##