Georgia State University CSC4780/6780 Fundamentals of Data Science

Final Project Presentation

Spring 2022
[AJKFDS]
Aryan Shrestha, Jagan Gedela, Sridutt Kalidindi, Caleb Tsai, Christopher Yoon

Introduction

- Banks are forced to expend time and resources to contact all of their clients to offer term plans that might be rejected
- Dataset was gathered through a phone call campaign for term deposits
- Our model aims to streamline the bank's operations by predicting which clients are more likely to accept term payments, and by extension, which clients should be called
- We followed the CRISP-Data Mining Technique to determine the most efficient model for predicting which clients to call.

Data Sources and Data Exploration

Dataset was gathered through a phone call campaign for term deposits.
This dataset includes 21 attributes and 41,189 instances.

- Exploratory Data Analysis (EDA) was used to summarize/preprocess our data
- Allows for a clean view for us to determine if our models are set up properly
- Normalization took place for all our continuous features on a range of [0,1]
- Transformation used one-hot encoding to change our data from text to ensure that our models can use the (initially) categorical features
- Pd.get_dummies was used to prevent a range of [0,infinity] from occurring and messing up the models' accuracy

Model Selection

We used Naive Bayes, Logistic Regression, ADABoost, SVM and Random Forest to make our models. HyperParameter FineTuning was also used to potentially improve our results on Logistic Regression and Random Forest.

Each of these models are implemented through training and testing to test the accuracy of the data, and for our implementation, we just tested several common models to find the best accuracy for our dataset.

- Feature selection procedures included both the filter and wrapper methods, using the Chi-squared test and Recursive Feature Elimination respectively
- Wrapper was only used on random forest due to model specificity

Evaluation

- Confusion Matrix and Accuracy Ratings were used as evaluation metrics.
- Models Used
    - Naive Bayes algorithm
    - Logistic Regression
    - ADABoost
    - SVM
    - Random Forest

Evaluation

Model	Accuracy	Recall		Precision		F1 Score	
		Yes	No	Yes	No	Yes	No
Naive Bayes	0.86	0.49	0.91	0.41	0.93	0.45	0.92
Logistic Regression	0.9086	0.39	0.97	0.67	0.92	0.49	0.95
Logistic Regression (Hyperparameter Finetuning)	0.9083	0.42	0.97	0.65	0.92	0.51	0.94
ADABoost	0.8987	0.65	0.93	0.55	0.95	0.59	0.94
SVM	0.8984	0.65	0.90	0.25	0.98	0.36	0.94
Random Forest	0.9080	0.46	0.96	0.63	0.93	0.53	0.94
Random Forest (Hyperparameter Finetuning)	0.9045	0.28	0.98	0.71	0.91	0.4	0.94

Conclusion

- Logistic Regression provides a model that is understandable and easy to use, which is important as real life datasets would be much more complex and larger sized
- Since it is an easy model to implement, the company would benefit from cost cuts and an increase in prediction accuracy when dealing with future customers
- A specific action the company could take in response to the model’s predictions is to concentrate their efforts on appealing to customers who don’t accept term deposits in order to make the most of advertising, promoting, and other campaigns