Bank Product's Subscriber Prediction Analysis

Swapnil Sahu | Tamilselvan Gurunathan | Mounika Cheera | Smriti Khilnani | Suman Pogul
University of Maryland Baltimore County

Introduction

A Financial institution invests millions of dollars yearly in marketing campaigns to expand its customer base for the new products that it introduces each year. Based on data gathered by their marketing team, this expense is typically used for random customer marketing. The financial institution could save millions of dollars from random marketing and utilize the same for targeted consumer marketing if it had an effective data model that could forecast the precise clients subscribing to their new products. We have built an efficient classification model to predict the precise customer who would subscribe to the banks new products.

Methodology

The data set for this task has been gathered from UCI data repository and data cleaning and pre-processing had been done on the dataset using linear interpolation. Each data model's performance was calculated using two methods - training set and cross validation. Below are the steps followed to achieve the task using WEKA.

01 Data Acqusition

The data is related with direct marketing campaign of a corporate banking institution.The marketing campaigns were based on phone calls.The bank data is acquired from UCI repository

02 Data Preprocess

Data imputation has been performed on the data set using linear interpolation method. Duplicate values has been removed using conditional formatting.

03 Data Modelling

Multiple classification models was built using KNN,J-48,Naive bayes Algorithms. The classification accuracy of each model has been determined using Held out data set method.

04 Data Visualization

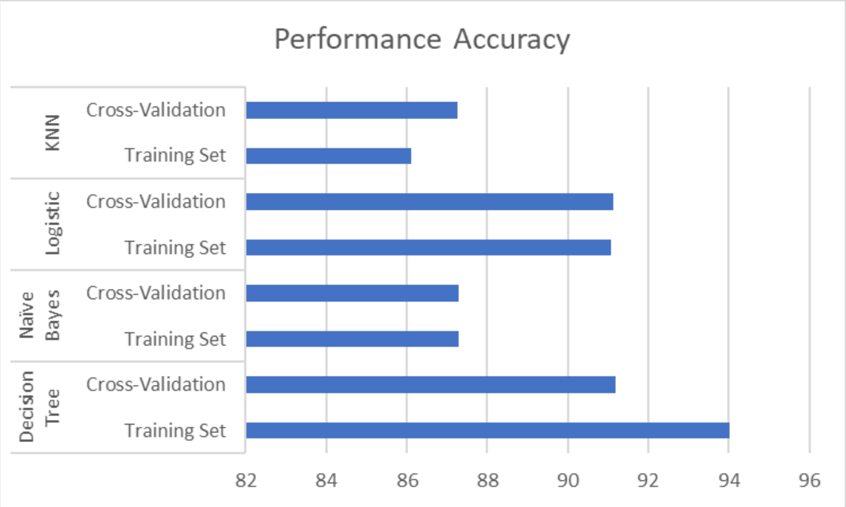
The performance of each model is visualized using ROC to identify the best model fit for this approach. This is done using WEKA tool.

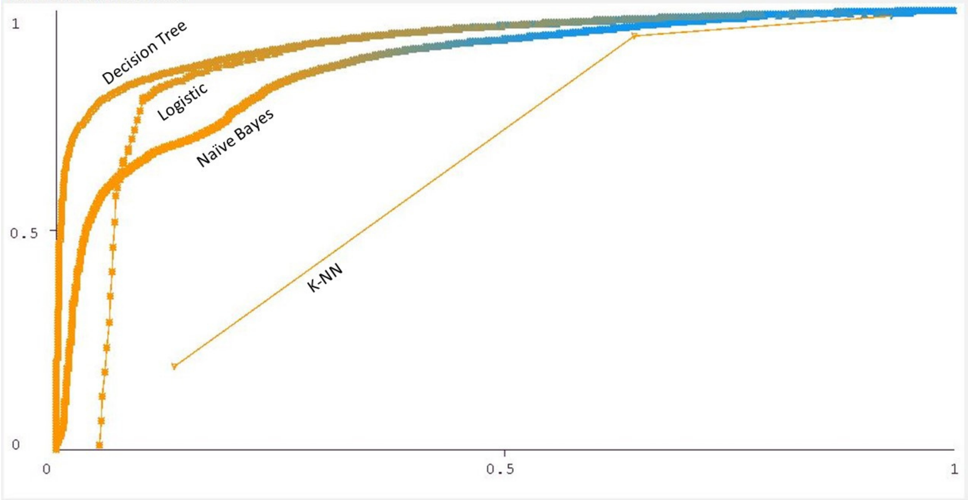
05 Model Finalization

The best model fit for our objective will be finalized after analyzing the performance of each model that will be suggested for deployment.

Analysis

The downloaded dataset consists of variables like age, job status, marital status, annual income, etc. We implemented J48 algorithm on the training dataset. The model predicted the target class variable (Subscribe) as "Yes" or "No" (whether the customer will subscribe to the product or not) based on the data provided. The accuracy is 94% which is quite good performance for training set. The subsequent model deployment will further enhance the performance and understanding of the model.

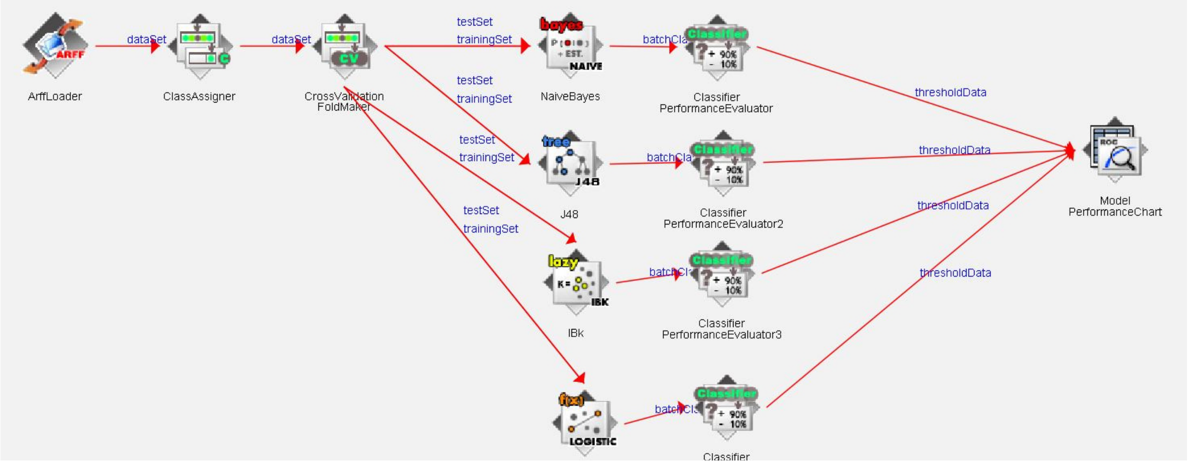




Multiple classification model was built using KNN, Naive bayes and logistic regression to compare the performance and identify the efficient model.   Receiver Operating Characteristic (ROC) curve is used for the graphical representation of the model developed. We have chosen ROC curve as it is independent on the cost benefit matrix and class priors. By varying the threshold value for the classifier creates a curve. As we can see the curve depicted illustrates that area under ROC curve which means perfect performance accuracy.

Knowledge Flow

Knowledge Flow represents a data-flow interface in WEKA.A data-flow was created using component nodes for processing and analyzing data. We derived model performance chart using the training data set and running different algorithms.



Conclusion

Based on the analysis of the performance accuracy out of the four data models (Decision Tree, Naive Bayes, Logistic, KNN) used, we conclude that Decision Tree has the highest accuracy percentage of 94% which would be more efficient model to predict whether a customer will subscribe the bank's product or not.