MACHINE LEARNING ENGINEER NANODEGREE

Using Supervised learning to determine whether the customer subscribes for a bank marketing campaign or not

Capstone Proposal :

Deepika Kothapalli

June 26, 2018.

Domain Background :

Banks contain huge information about the customers. They can use this data for several useful purposes. In order to have a good relationship with the customers and to market several other schemes they introduced they will select a means of communication like personal contact, through telephone etc.

Sometimes Banks can use TV and other means to inform or market their offers and schemes. But customers may not pay a good interest in that. So, they use telephone as a medium where they can directly speak with the customer about the offers and clarify about certain things and know whether they are interested or not.

In this way they can get to know about the genuine feedback what they think of the certain offer. Through direct contact with the customer it is also possible to convince the customer about their ideas. This type of marketing product or service is called direct marketing which came into existence in 1960’s.

Problem Statement :

In this I want to determine whether the customer subscribe to the campaign or not. I want to classify the customers subscribed and unsubscribed to the campaign

There are certain features based on which I want to classify the data points like age, type of job, marital status, education, loan, housing number of days before the bank contacted the customer etc.

I decided to use several supervised learning classification algorithms like Decision trees, logistic regression etc. I will find the best model among those using many performance metrics through which I can get good results.

This project consists of several phases like Data Exploration, Data preprocessing, application of various classification algorithms, finding the best model etc.

Datasets and inputs :

I have selected the data set from the kaggle.com.

It contains nearly 4000 rows and 20 columns.

Features are :

- age (numeric)
- typeofjob('admin.','bluecollar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- education ('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- default: has credit in default? (categorical: 'no','yes','unknown')
- housing: has housing loan? (categorical: 'no','yes','unknown')
- loan: has personal loan? (categorical: 'no','yes','unknown')
- contact: contact communication type (categorical: 'cellular','telephone')
- month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri'
- duration: last contact duration, in seconds (numeric)
- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric)
- previous: number of contacts performed before this campaign and for this client (numeric)

- poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
  # social and economic context attributes
- emp.var.rate: employment variation rate - quarterly indicator (numeric)
- cons.price.idx: consumer price index - monthly indicator (numeric)
- cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- euribor3m: euribor 3 month rate - daily indicator (numeric)
- nr.employed: number of employees - quarterly indicator (numeric)

Among these features most important features during the course of the project which are useful for producing the accurate results

Solution Statement :

The solution I will provide is :

Selecting the best optimal features necessary for accurate classification

It is a supervised learning classification problem which requires a binary classification

So, I will use several classification algorithms like logistic regression, decision trees, SVM etc

Among them I will find the best model using the performance metrics.

I will use GridSearch optimization technique for providing more accurate results.

Benchmark model :

The output variable for this problem is binary. So, the bench mark model I choose for this problem is logistic regression.

F-Beta score and accuracy score can be used as performance metrics. If any other model has good F-Beta score than logistic regression that can be considered as good model.

Evaluation metrics :

Accuracy : It determines the proportion of correct predicts among all the predictions

Accuracy= TP+TN/(TP+TN+FP+FN)

Where True positives and True Negatives are the correct predictions. False positives and False negatives are wrong predictions.

Precision :

Precision=TP/(TP+FP)

It determines among all the customers that are predicted as subscribed to the campaign who are actually subscribed.

Recall :

Recall determines the proportion of subscribed customers correctly predicted among the actual subscribed customers.

Recall=TP/(TP+FN)

F-Beta score :

F-beta score is the weighted harmonic mean of precision and recall. β

F-beta=(1+ β2) *precision *recall/(β2*precision+recall)

Project Design :

These are the steps I would like to perform

Data Acquisition : I have got the data set from kaggle.

Data preprocessing: In this step I will clean and remove the unnecessary data and also apply one hot encoding to convert categorical data to numerical data.I will remove the data points with missing values.

Data Exploration and visualization : I will plot the data to determine the skewness and normalize the data using min_max_scaler if any skewness is present.

Model Evaluation and Validation :

In model evaluation I will use performance metrics like Accuracy and F-beta score. If the F-beta score is high the model is considered to be a good model. Otherwise it is a bad model.

Optimization: I will optimize the model using one of the most popular optimization technique Grid Search.