Machine Learning Engineer Nanodegree

Capstone Proposal

Arthur Vetori
December 03rd, 2018

Proposal

Domain Background

With the development of the computer processing power and machine learning techniques, data science become more popular and demanded in the all business sectors. One of the promising application areas is Banking and Financial Industry. The large amount of data that banks have, such as personal education, income and spending habits makes predictive models a strong tool for direct marketing. The wide spread use of internet turns decision processes more based in data than it ever was. In bank marketing, for example, large amount of data with different techniques are used for classifying potential costumers for a specific product.

Problem Statement

The goal of this project is to classify if the client will subscribe for a term deposit based on multiple client features such as personal information, last contact of the current marketing campaign and economic context attributes. Different techniques will be used and compared in accuracy terms.

Datasets and Inputs

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls.

The dataset used in this project is available at:

> https://archive.ics.uci.edu/ml/datasets/bank+marketing

The file downloaded from the link contains two datasets, one with 20 features (**bank-additional-full.csv**) and another with 17 features (**bank-full.csv**). I will be using the dataset with 20 features for this work. This dataset, consists of 41188 registers and are ordered by date (from May 2008 to November 2010).

Inputs are:

- **age** (numeric)

- **job** : type of job (categorical: 'admin.','blue-collar','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** : (categorical:'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). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

- **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; 999 means client was not previously contacted)

- **previous**: number of contacts performed before this campaign and for this client (numeric)

- **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

- **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)

Solution Statement

For this project, different classification methods (AdaBoost, Random Forest, Naive Bayes and SVM) will be used to see which of them will perform better in terms of accuracy and F1-score. Other techniques such as feature selection, cross validation and grid search will be used to improve classifier performance.

Benchmark Model

I will be using Decision Trees as the benchmark model, which will be compared with other models described in the solution statement.

Evaluation Metrics

Accuracy and F1-score will be used as evaluation metrics to compare described techniques.

Project Design

The structure followed in project will be:

- Exploratory data analysis
- Cleaning data if needed.
- Feature selection
- Classifier training/testing with cross validation
- Compare all methods using the described evaluation metrics

- Conclusion

References

https://archive.ics.uci.edu/ml/datasets/bank+marketing