BANKING TERM DEPOSIT ANALYSIS

Machine Learning CS 697AB

Instructor: Kaushik Sinha

Done by:

Caleb Rettig (undergraduate student) & Sai Rohan Reddy Gaddam (graduate student)

1. Problem Definition

We are working on the marketing campaigns of a Portuguese banking institution. The goal is to predict if the client will subscribe a term deposit or not. Further, we need to be reducing the number of features (from the initial 17) for data storage efficiency, finding out significant features with an increase/decrease in classification accuracy.

So, our goal would be to apply and algorithm and test the data while removing certain features and checking accuracy on each case. Finally, we would be determining the feature that has affected the dataset the most and also how the features are different or same from each other. We would be reaching an ideal point where our accuracy is high with ideal number of features (neither less nor more).

2. Solution Techniques:

Initially, the data consists of inputs in the character formal (for example marital status). So we first converted all such types of data into a labeled numerical format for MATLAB convenience.

Secondly, we have tested 2 different algorithms on the datasets and decided to go by the Naive Bayes theorem. We would be training a small portion of the data and then test the remaining to find out the accuracy obtained.

Then, we have written a script that performs feature reduction to eliminate one feature at a time and remove the one which has the least impact on accuracy. The idea is to keep on going till we have one feature left and then conclude the stage where we had the highest accuracy. Such a method would determine the characteristics of the features.

3. Datasets used:

There are Four datasets:
1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010
2) bank-additional.csv with 10% of the examples (4119), randomly selected from set 1, and 20 inputs.
3) bank-full.csv with all examples and 17 inputs, ordered by date.
4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 .

Dataset 1 and 2 which contain an additional 4 features. We only calculated the accuracy for these as working on the already reduced datasets seemed more reasonable to us.

Dataset 4 has been used to train our classifier for prediction and then dataset 3 has been tested on.

The inputs from set 4 are selected randomly from set 3, so we could not separate them out from the original. Hence this can result in a slight false improvement in the final accuracy.

4. Software used:

MATLAB R2016a which is a great tool for numerical integrations.

5. Results:

Initial accuracy = 81.9120 without classifying features showing lot of variance (e.g age & balance)

Feature Reduction:

Each stage removes the specified features and the highest accuracy specified when removing the feature at that particular stage.

Initial accuracy adjusting age and balance - 83.5571

1st reduction: (pdays) - 87.0474

2: (previous) - 88.4364

3: (campaign) - 88.8390

4: (loan) - 88.9341

5: (month) - 88.9562

6: (balance) - 89.0137

7: (default) - 89.0182

8: (day) - 89.0137

9: (education) - 89.0115

10: (age) - 89.0093

11: (job) - 88.9983

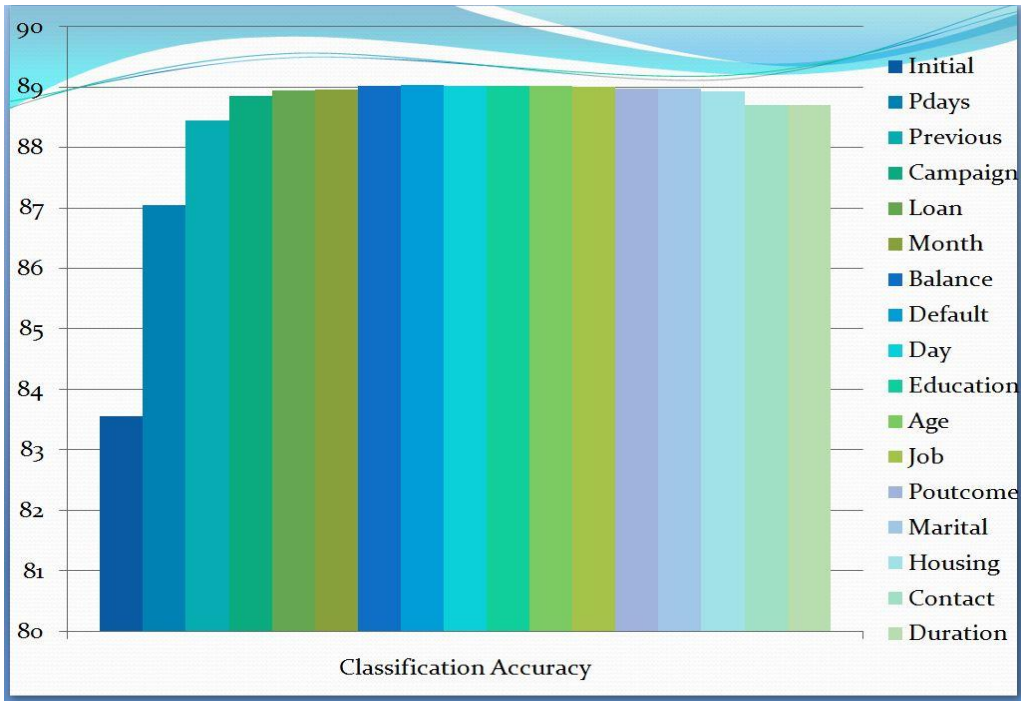
12: (poutcome) - 88.9651

13: (marital) - 88.9651

14: (housing) - 88.9142

15: (contact) - 88.6908

Duration is the final feature with 88.6908% accuracy. Our best accuracy hit after we removed 7 features i.e 89.0182.



6. Discussion

Firstly, the results were particularly not showing much variation after removal of certain number of features (As we can see a fractional change in % of accuracy). We expected each feature to be distinct from one another and the difference be shown in the form on numbers.

We could have used more techniques like KNN classification and Decision trees to check the accuracy and how data reacts to such algorithms. Initially, we have tried feature reduction for our datasets. The accuracies did not go as expected and it failed. So, we switched to feature reduction.

Given the short period of a month, we would say that we are not really satisfied by the final results. Given more time, we could have performed more classification technique, studied and understood the data behavior with different methods. Lastly, we would say that we gained a lot of experience with this project and our failures in this project taught us a lot. Our results are just another attempt which turned out right.

7. Division of labor

Although the division initially was 60% - 40%, we would say that division was pretty much equal.

Caleb - Worked on the conversion script that coverts character inputs to numerical. Worked on the Decision tree technique as well which could not be presented but was completed.

Rohan - Structure of scripts and implementations, ideas. Documentation including abstract, report and the presentation.

The decision of techniques and the main body of the project was done by both of us. Although the division for Caleb was only 40%, he has put up a tremendous amount of effort and time into the project. Concluding to the point that we cannot really think on how much of the whole each of us contributed. We can say it goes around equally.

8. Reference

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]