LUT
Lappeenranta
University of Technology

Bank Marketing Campaign

A Machine Learning algorithm comparison

Members of Research Group

Mikael Sommarberg, BSc
Kimmo Flykt, BSc
Sami Kabir, BSc
Ajesh Sharma, BSc

Introduction and Background

In today's world more and more tasks are moved to digital world from real world. Long gone are days when reports where in paper and copied to everyone. At the same time the amount of data that exists and is collected has exploded in our hands. The more we move towards digital the more data we create. This creates a unique opportunity to use computers to analyse data. We know that human brain is great calculator and its recognition capabilities are better than any other solution but with large amounts of data our brains start to struggle. If we can model dataset in a way that computer understands it we can use it's great calculation speeds to run through the data and get results out faster than before. This would be great help when we have to run through hundrest or event hundred thousand rows of data with multiple attributes.

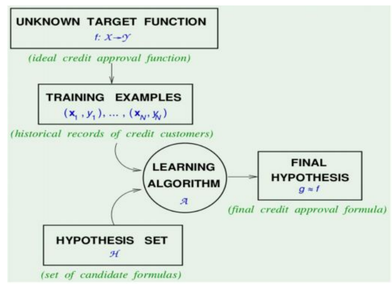
The main goal for this project is to compare prediction accuracy of different machine learning models with the given historical bank marketing campaign data. If classifier has a very high accuracy, it can help the manager to filter clients and use available resources more efficiently to achieve the campaign goal. Proper strategy would reduce cost and improve long term relations with the clients. We have chosen to compare five different machine learning algorithms: K-Nearest Neighbours, Logistic Regression, Linear Discriminant Analysis, Decision Tree and Gaussian Naive Bayes.

Data

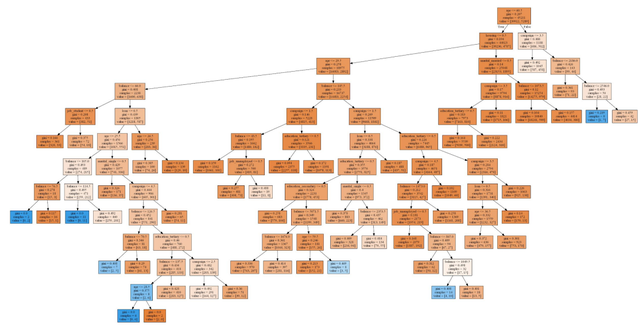
Sl. no.	Variable Name	Data Type	Expected Values	Significance
01	**Age**	**Numeric**	**18, 30, 60, 90 ...**	**Describes the age of the customer.**
02	**Job**	**Character**	**'Student', 'unemployed' 'manager' etc.**	**This is a categorical information describing the employment scenario of a customer.**
03	**Marital**	**Character**	**'Married', 'Single' etc.**	**This describes the marital status of the customer.**
04	**Education**	**Character**	**'Unknown', 'secondary', etc**	**This describes the education qualification of the customer.**
05	**Default**	**Binary**	**'Yes' or 'No'**	**This describes if the customer has credit in default or not.**
06	**Balance**	**Numeric**	**10000, 1000, 999 ...**	**This describes the average yearly balance in euros of a customer.**
07	**Housing**	**Binary**	**'Yes' or 'No'**	**This describes if the customer has housing loan or not.**
08	**Loan**	**Binary**	**'Yes' or 'No'**	**This describes if the customer has any personal loan or not.**
09	Contact	Character	'Unknown', 'Cellular' etc.	This describes the categorical information about the contact communication type of a customer.
10	Day	Numeric	01, 15, 20, 30 ...	This describes the day of the month when the last contact was made with the customer.
11	Month	Character	'Jan', 'Feb', 'Mar', etc.	This describes the month of the year when the last contact was made with the customer.
12	Duration	Numeric	10, 19, 300, ...	This describes the last contact duration in seconds.
13	**Campaign**	**Numeric**	**1, 3, 5 ….**	**This describes the number of times contacts were made for this campaign and for a customer.**
14	Pdays	Numeric	-1, 10, 15 ...	This describes the number of days that passed by after the client was last contacted from a previous campaign.
15	Previous	Numeric	1, 10, 14 ...	This describes the number of contacts performed before this campaign and for this client.
16	Poutcome	Character	'Unknown', 'Success', Failure etc.	This describes the outcome of the previous marketing campaign.
17	**Y**	**Binary**	**'Yes' or 'No'**	**This describes if a client has subscribed a term deposit or not**

Bolded values were used in the study

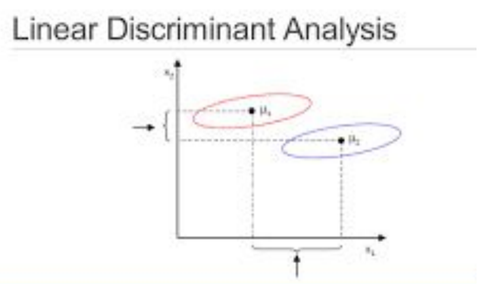
Methodology



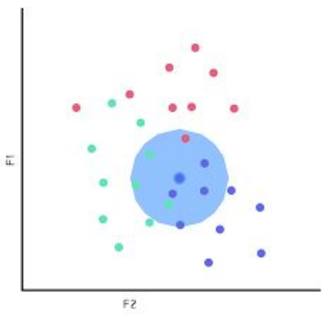
Linear regression



Decision tree

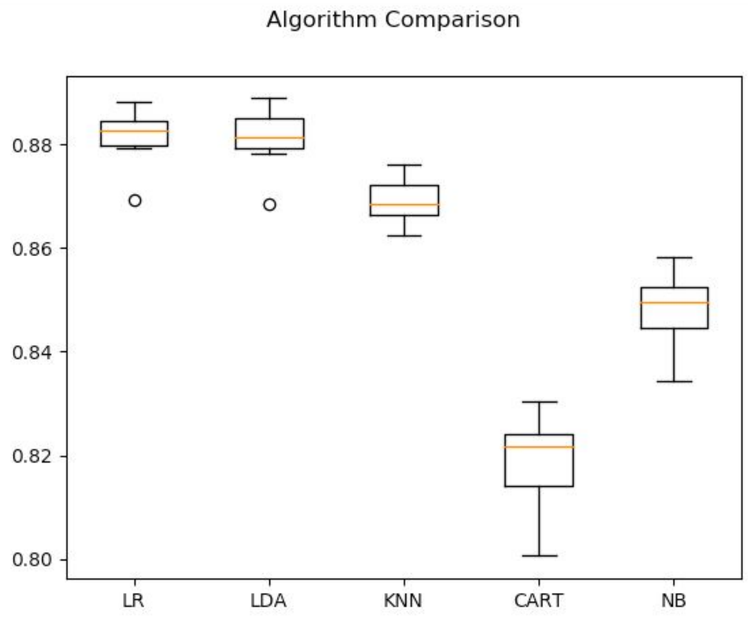


Linear Discriminant Analysis



K-nearest neighbor

Results



Machine Learning Model	Prediction Accuracy
Logistic Regression (LR)	88.02%
Linear Discriminant Analysis (LDA)	88.11%
K-Nearest Neighbour (KNN)	86.90%
Decision Tree Classifier (CART)	81.88%
Gaussian gaussian Naive Bayes (NB)	84.85%

The results of this study indicate that the best performing algorithm with the used data is the **Linear Discriminant Analysis.** A close second place goes for **Logistic Regression** algorithm.

RESEARCH GROUP CONTACT INFO

Mikael Sommarberg, BSc, Lappeenranta University of Technology
Mikael.Sommarberg@student.lut.fi

Kimmo Flykt, BSc, Lappeenranta University of Technology
Kimmo.Flykt@student.lut.fi

Sami Kabir, BSc, Lappeenranta University of Technology
Sami.Kabir@student.lut.fi

Ajesh Sharma, BSc, Lappeenranta University of Technology
Ajesh.Sharma@student.lut.fi