CS910: Coursework
BANK MARKETING ANALYSIS

INTRODUCTION

Marketing is a tool of attracting potential clients to a product, applying a variety of approaches and schemes. It significantly facilitates the buying of goods or services and helps in identifying the need of the product and persuading customers to buy it. The main goal of marketing is to improve sales for businesses and financial institutions.

Telemarketing is a form of marketing where salespersons contact the customer and provoke them to make a purchase. Nowadays, developing technologies are used very effectively in bank marketing campaigns as in many other fields of life. It is cost effective and keeps the customers up to date. In the banking sector, advertising and marketing is mostly based on an intensive knowledge of objective information about the market and the actual client needs. By digging into this kind of information it is possible to come up with a solution to build an effective marketing campaign that would help to target potential customers.

The contribution of this paper comes in two main dimensions. First is to analyse different characteristics provided by some telemarketing campaigns, and how they relate to the decision of a customer to subscribe to a term deposit or not. Data analysis techniques will be used in order to detect trends, such as whether certain attributes can affect the target result. A secondary aim of the paper is analysing the application of different feature selection and classification methods.

DATASET

This paper employed the bank marketing dataset that is publicly available from the University of California at Irvine Machine Learning Repository. The bank marketing dataset is introduced by Moro et al [1]. The marketing campaigns are based on phone calls and related to 17 campaigns, which occurred from May 2008 to November 2010.

The dataset contains information about clients that can be categorised into distinct categories: the client's personal information, the client's financial information, items related to a current marketing campaign as well as items related to a prior marketing campaign.

The dataset contains records about direct marketing campaigns carried out by the Portuguese bank with the aim of getting customers to subscribe to a term deposit. Term deposit refers to a deposit held by a financial institution for a certain period of time. This period is known to the customer. During this period, the customer is not able to get the deposit. Requesting for the deposit before the agreed maturity period may attract some of penalty.

The ability to identify potential customers for subscribing to a term deposit is important because more targeted marketing campaigns would reduce time and resources wasted. Collection of customer information seems to be important for development of the marketing strategies. Customer data is stored electronically and the size of this data is so immense that to analyse it manually would be impossible.

DATASET DESCRIPTION

Table A (appendix) shows the description of the bank telemarketing dataset.

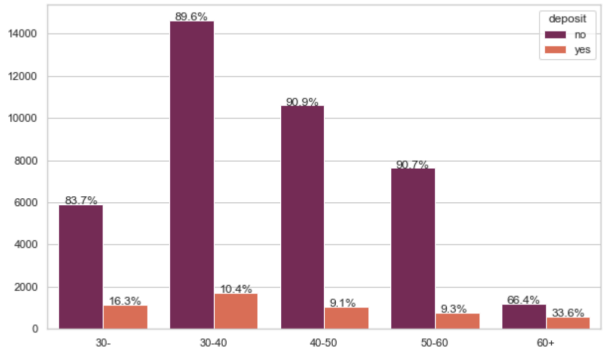
EXPLANATORY DATA ANALYSIS

The bank direct marketing data set contains 45211 number of samples with 17 attributes. The dataset is highly imbalanced as only 11.7% of the instances indicate a positive label (a client has subscribed for a term deposit).

To obtain a better understanding of the dataset, plots are used for an analysis of the relationship between attribute pairs as well as the relationship between attributes and the target variable. Visualization methods are used to illustrate hidden patterns inside the dataset.

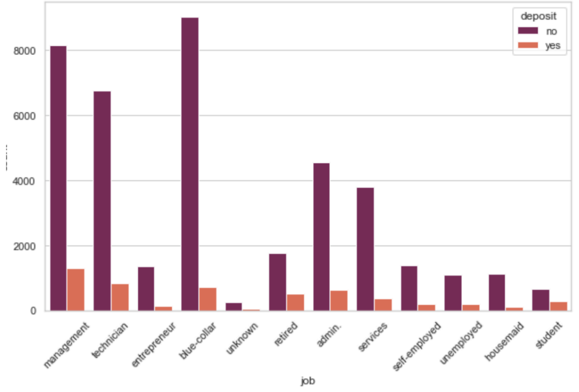
Age

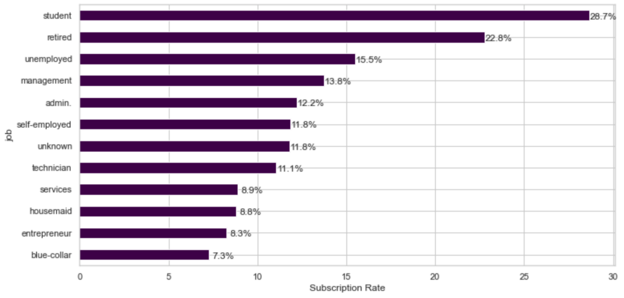
There is an extensive age range from 18 to 95 amongst clients called by the bank.



The graphic demonstrates that main target of the bank are people from thirties to forties. However, this group presents lower subscription rates compared to the younger and older groups. People who are above 60 has the largest subscription rate(33.6%). The second largest subscription rate is obtained by people who are below 30 years old(16.3%). This pattern is predictable since older people are saving money for retirement, and using term deposits as the least risky investment tool.

Job

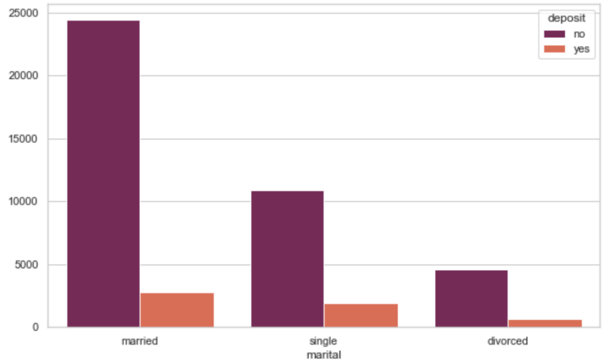




Looking closer into 'job' data we can notice that blue-collars and management are the most targeted. However, the ratios between 'yes' and 'no' are not the largest ones. Retired people and student take more than 50% of subscription despite the fact that they are less contacted. This result is consistent with the previous finding of higher subscription rates among the young and old people.

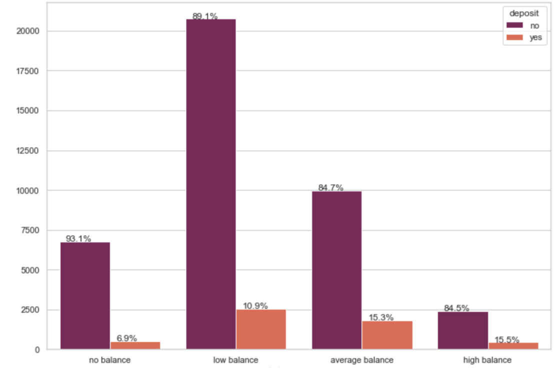
Marital

The graphic illustrates that married people are most contacted. It is not surprising since the main target of the bank are people in mid thirties. However, the subscription rate shows that single and divorced people are more likely to subscribe for a term deposit compared to married people.



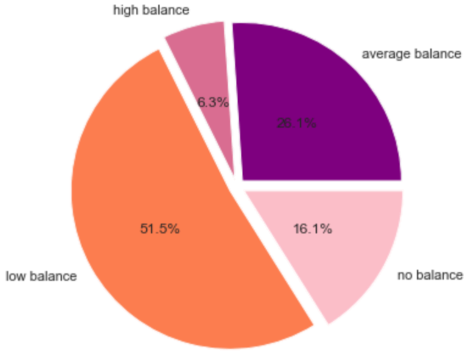
Balance

In order to identify the patterns, customers are classified into four categories based on their balances. Thus, clients with a negative value of balance are referred to 'no balance' group, clients with a balance between 0 and 1000 euros are in low balance group, clients with a balance between 1000 and 5000 euros are in average balance group and high balance group contains clients with a balance greater than 5000 euros.

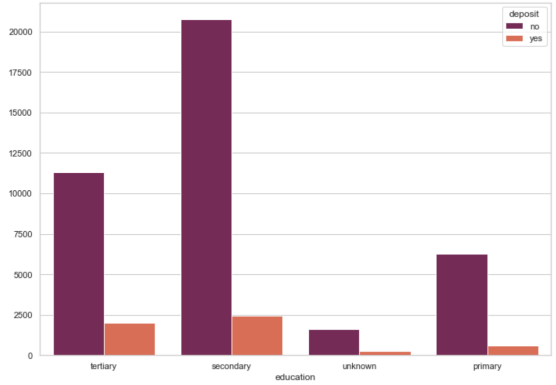


This plot shows a positive correlation between balance levels and subscription rate. Customers with negative balances only have the lowest subscription rate which is equal to 6.9% while these rates for clients with average and high balances are more than 15%.

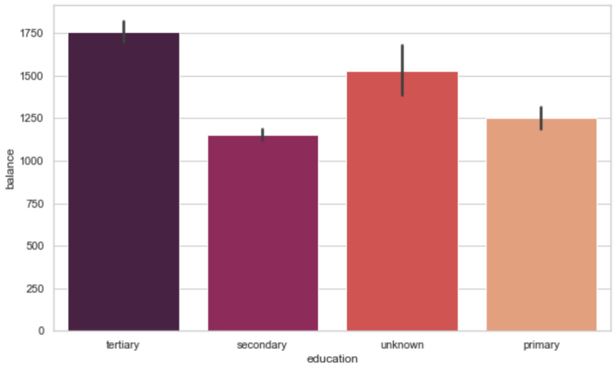
Nevertheless, 51.5% of contacted clients have a low balance. The bank should pay more attention to customers with average and high balances in order to increase its effectiveness.



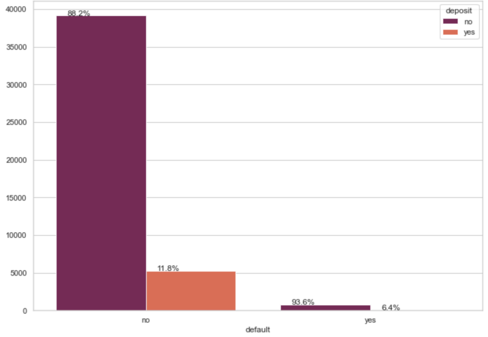
Education



The graphic indicates that clients who have a secondary education are most contacted by telemarketing campaign, but clients with tertiary education are more likely to subscribe to a term deposit. What is more, by comparing education with the balance it can be concluded that there is a significant impact on the balance of people with tertiary education. This confirms the previous conclusion: clients with high balance are more likely to subscribe to a term deposit.

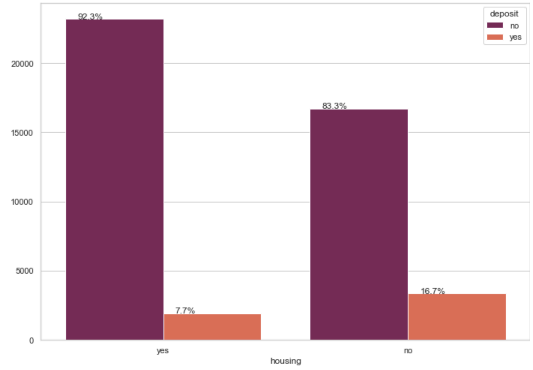


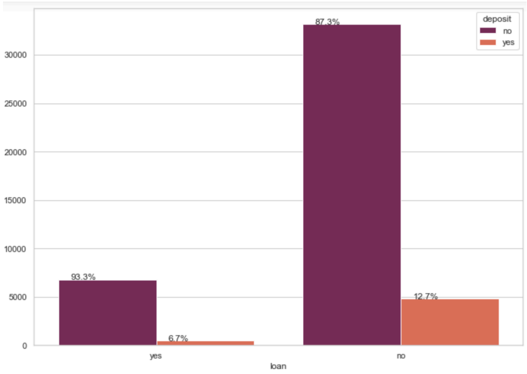
Default



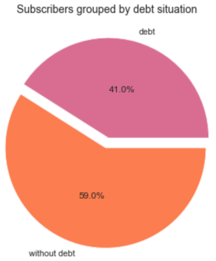
The graphic shows that customers without any credit default are more likely to subscribe to term deposit than customers with credit default.

Housing and Loan

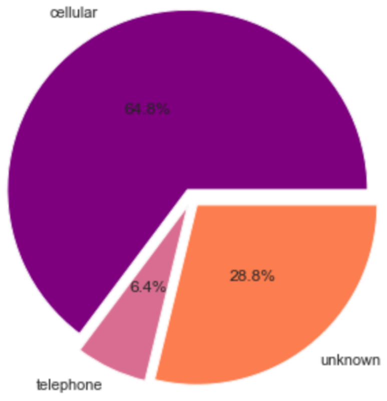




It could be concluded from the graphs that clients with any type of debts are less likely to subscribe to term deposit. However, personal loans have less influence on subscribing than housing loans. Furthermore, almost 60% of subscribers do not have any types of debts(loan, housing or default).

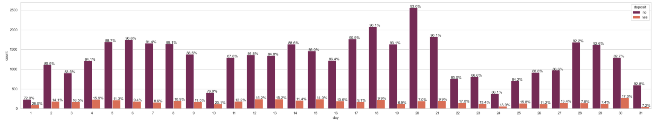


Contact



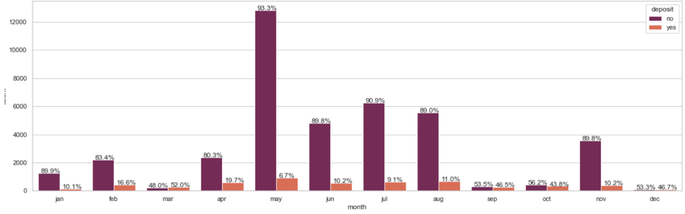
The pie chart presents the subscribers grouped by the type of communication. It indicates that 64.8% of subscribers are contacted by cellular and only 6.4% are contacted by telephone.

Day



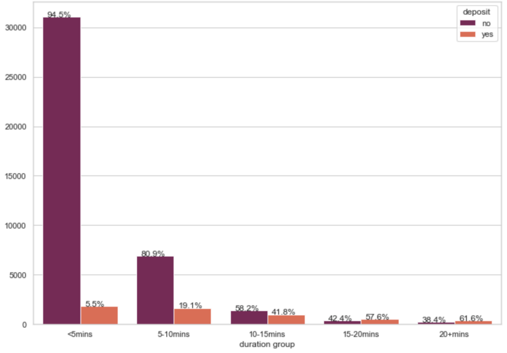
It can be concluded from the graph that the bank should initiate the telemarketing campaign at the first day of the month. What is more, the subscription rate tends to be the lowest at the last day of the month.

Month



Analysing the month infers that the bank contacted most clients from May to August. The highest contact rate is achieved in May and equal to 30%. These rates are about 0% in March, September, October, and December. Nevertheless, the subscription rate presents a completely different tendency. The highest subscription rate (52%) is obtained in March, and all subscription rates in September, October, and December are more than 43%. Thus, this analysis identifies the completely wrong time of bank telemarketing campaigns. The bank should increase the telemarketing campaign in the fall and the beginning of the spring.

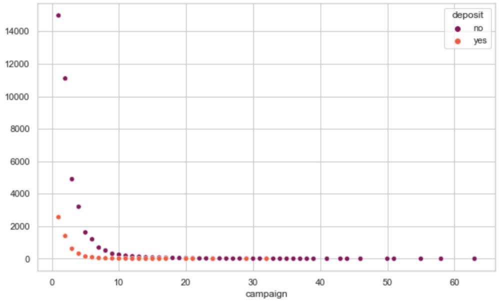
Duration



According to the dataset information, duration of the campaign plays a significant role and has a good correlation with deposit subscription. The more the client is engaged with the conversation, the more likely they are to subscribe to a term deposit.

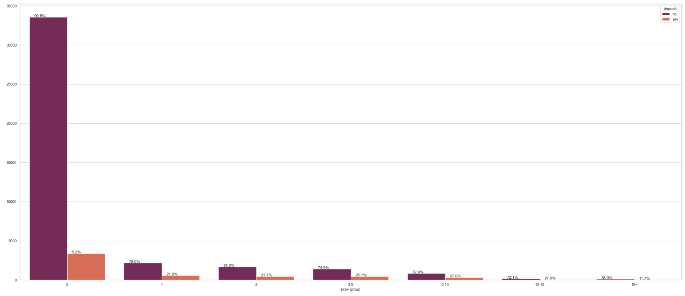
Campaign

The attribute represents the number of contacts performed during this marketing campaign per client.



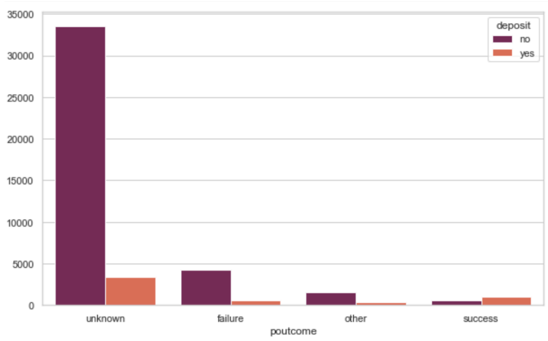
The scatter plot infers that more contacts, the less likely the client will decide to make a term deposit.

Previous



Number of contacts with the customers matters. Too many contacts with the customer could make him decline the offer.

Poutcome
This column has many unknown data. This is because most of these people involved in this campaign now have not been involved before. However, there is a tendency that people who have a successful previous outcome are likely to subscribe for a term deposit again.



DATA PREPROCESSING

There are missing values in four categorical attributes: ’job’, ‘education’, ‘contact’, ‘poutcome’ coded with ‘unknown’ label. These missing values are treated as separate values due to the fact that they are not random and may themselves be information.

There are categorical nominal values, categorical ordinal values, and numeric values presented in the data set. Attribute ‘education’ could be put in to the categorical ordinal values as the education level could be ordered from lowest to highest.

The data preprocessing steps have been taken in order to remove skewness. Firstly, numerical attributes are standardized by removing the mean and scaling the variance to one using StandardScaler. Secondly, attribute ‘education’ is converted to its numerical representation using OrdinalEncoder as this attribute can be considered as ordinal. Thirdly, all other categorical attributes are transformed into numeric values using LabelEncoder.

FEATURE SELECTION

Feature selection is an important process of reducing the number of input variables when developing a predictive model. Decreasing the size of the features helps to increase the efficiency of the training process as well as deleting redundant features improving the accuracy of the classifiers. Feature selection algorithms are based on statistical measures that assign a scoring value to each feature and then the features are ranked by this value. In this paper, Information Gain [2] and classification and regression trees (CART) algorithm implemented in scikit-learn are used as feature selection tools.

FEATURE SELECTION RESULTS

Two rankings are compared to see patterns in importance amongst attributes and to derive some insights. The ranks for features are presented in Table 1.

Table 1: Features rankings

Rank	Information Gain	CART
1	duration	duration
2	poutcome	balance
3	pdays	month
4	month	age
5	balance	day
6	previous	poutcome
7	contact	pdays
8	housing	job
9	age	campaign
10	marital	education
11	day	housing
12	job	marital
13	campaign	contact
14	education	previous
15	loan	loan
16	default	default

As it may be concluded, attributes ‘default’ and ‘loan’ are ranked lowest using both feature selection methods. This indicates that the fact of having debts does not impact client’s decision to subscribe to term deposit. The analysis also shows that the personal information related attributes which include ‘education’, ‘marital’, ‘job’ are not important in this decision.

In terms of importance, ‘duration’ is the most significant across both rankings. ‘Month’ is also indicated as a significant feature by both presented methods of feature selection. This suggests that term subscriptions are seasonal. ‘Poutcome’ which takes high positions in both rankings signifies repetition in subscription. Moreover, customer’s account balance seems to contribute to the classifier according to both rankings.

MODEL BUILDING

To further investigate if there is a relationship between clients' information and deposit subscription, the following classifiers were modelled to select the best classifier:

1. *LogisticRegression*
2. *k-Nearest Neighbours using euclidean distance*
3. *k-Nearest Neighbours using manhattan distance*
4. *Decision Tree Classifier*
5. *Gaussian Naive Bayes*
6. *Random Forest Classifier*

PERFORMANCE EVALUATION

As mentioned before, the dataset is significantly imbalanced. Since the issue of class imbalance has the potential of producing extremely high or low prediction accuracies (i. e. overfitting) f1 weighted score is used along with the accuracy score. weighted means that the f1 score is calculated for each class independently and then added together using a weight that depends on the number of true labels of each class.

The classifiers have been evaluated on the preprocessed data with the entire set of attributes. This is done with the purpose of obtaining baseline results to compare the further analysis.

RESULTS

Table 2 presents performance of baseline settings in terms of accuracy and f1 measure of the classifiers.

Table 2: Baseline results dataset

Classifiers	Accuracy	F1 weighted
LR	0.88	0.86
kNN eucl	0.76	0.77
kNN manh	0.77	0.78
DTC	0.62	0.67
NB	0.80	0.82
RFC	0.76	0.77

From the baseline results, it is observed that Logistic Regression Classifier produces the best accuracy of 88% as well as the best f1 weighted measure. The worst results are obtained with Decision Tree Classifier with 62% of the instances correctly classified.

Modelling of the various classifies is done using the k-fold (10-folds) cross-validation. That means the whole dataset is divided into ten equal sized sets and classifiers are trained on nine train sets and tested on one test set. This process is repeated ten times and then an average of all folds is taken as a result. Tables 3,4,5 and 6 present performances of Information Gain and CART feature selection methods in terms of accuracy and F measure values using the whole set of classifiers when trained and tested on the reduced the size of features. Four feature sizes (5, 8, 10, 12) are tried for the bank marketing dataset.

Table 3: results for 5 features

	Information Gain		CART	
Classifier	Accuracy	F1	Accuracy	F1
LR	0.88	0.86	0.89	0.85
kNN eucl	0.77	0.78	0.69	0.72
kNN manh	0.77	0.78	0.69	0.73
DTC	0.75	0.77	0.66	0.70
NB	0.83	0.83	0.88	0.86
RFC	0.78	0.79	0.70	0.73

Table 4: results for 8 features

	Information Gain		CART	
Classifier	Accuracy	F1	Accuracy	F1
LR	0.88	0.86	0.88	0.85
kNN eucl	0.77	0.78	0.75	0.77
kNN manh	0.77	0.78	0.76	0.77
DTC	0.75	0.77	0.66	0.70
NB	0.82	0.82	0.83	0.83
RFC	0.78	0.78	0.76	0.77

Table 5: results for 10 features

	Information Gain		CART	
Classifier	Accuracy	F1	Accuracy	F1
LR	0.88	0.86	0.88	0.86
kNN eucl	0.80	0.80	0.77	0.78
kNN manh	0.80	0.80	0.78	0.78
DTC	0.75	0.77	0.67	0.72
NB	0.82	0.83	0.83	0.83
RFC	0.79	0.79	0.76	0.77

Table 6: results for 12 features

	Information Gain		CART	
Classifier	Accuracy	F1	Accuracy	F1
LR	0.88	0.86	0.88	0.86
kNN eucl	0.74	0.76	0.79	0.79
kNN manh	0.75	0.76	0.79	0.79
DTC	0.62	0.67	0.66	0.70
NB	0.81	0.82	0.82	0.83
RFC	0.74	0.75	0.76	0.76

As seen in tables, the results of classification when applying Information Gain and CART methods of feature selection are quite similar despite the fact that they have differences for the first five highest ranked features. The performance of classification is improved with reduced size of features and the best results for all classifiers are obtained when we take into account 10 features. These highest ranked ten features are 'duration', 'poutcome', 'pdays', 'month', 'balance', 'contact', 'housing', 'previous', 'age', ‘job’, 'day' for Information Gain and 'duration', 'balance', 'month', 'age', 'day', 'pdays', 'poutcome', 'job', 'campaign', ‘education' for CART feature selection. It is also worth mentioning that Logistic Regression presents the best results across all classifiers and for both feature selection methods.

CONCLUSION

The aim of this paper has been to find any tendencies amongst dataset attributes.Taking into account previous data analysis, it is possible to create some advice that could help to increase telemarketing campaigns’ effectiveness:

1. Most potential clients tend to be under 30 or older than 60 years old, students or retired with a balance of more than 1000 euros.

2. Banks should initiate campaigns in March, September, October and December paying more attention to the first day of the month.

3. Banks could benefit by creating a method of increasing the duration of call.

What the paper has also found is that the Logistic Regression model has the best performance in prediction of the potential subscribers for a term deposit among all other classification methods. Applying feature selection methods leads to increasing performances of classifiers. Thus, the best results are obtained when the number of features is equal to 10.

POSSIBLE EXTENSIONS

Possible extensions to this research are as follows:

1. Combining this dataset with other bank marketing datasets with the aim to compare the results of predicting with the same algorithms and metrics but with more attributes. Adding more information may allow categories to be created for grouping the attributes (i.e. personal information, demographic, social and economic situation) with subsequent analysis of the relationship between each category and the target class.

2. Use of other feature selection and classification methods to find the best performance. The goal is to make the feature vector smaller and dense which will make the classifier learn more easily.

REFERENCE

[1] Moro S., Laureano R., Cortez P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. – 2011.

[2] Yang Y., Pederson J. O. A comparative study on feature selection in text categorization in: Proc. of the 14th International Conference on Machine Learning. – 1997.

Appendix

Table A

Attributes	Kind	Attributes illustration
Age	Numeric	Customer's age [18; 95]
Job	Categorical	Type of job
Marital	Categorical	Customer’s marital status
Education	Categorical	Customer's educational status
Default	Binary	Credit debt situation? ‘yes', 'no'
Balance	Numeric	Average annual balance, in euros [-8019; 102127]
Housing	Binary	Real estate debt situation? ‘yes’, 'no'
Loan	Binary	Personal debt situation? ‘yes', 'no'
Contact	Categorical	Contact communication type
Day	Numeric	Last interview day [1; 31]
Month	Categorical	Last interview month
Duration	Numeric	Last call duration, in seconds [0; 4918]
Campaign	Numeric	Number of contacts performed during this campaign and for this client [1; 63]
Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign -1 means client was not previously contacted
Previous	Numeric	Number of contacts performed before this campaign and for this client [0; 275]
Poutcome	Categorical	Outcome of the previous marketing campaign
Deposit	Binary	Has the client subscribed a term deposit? ‘yes’, 'no'