Project Report: Attribute Importance Analysis of Bank Marketing Data

Meghana Gaopande
meghanag@vt.edu

Pranavi Rambhakta
pranavi7@vt.edu

1 Problem Statement

We have two client datasets of telephonic direct marketing campaigns, which are disjoint in time, and share a majority of features (with some differences). Given this, for each dataset we want to analyze the importance of each attribute in terms of its contribution to classifying whether a given client will subscribe to the term deposit at this bank. We would like to study whether generalization holds true; i.e. the same set of common attributes is important for both datasets, or whether specialization holds true; i.e. the important attribute set is unique to the datasets, or has little overlap.

Keywords

Bank Marketing Campaign; Classification; Ablation; Attribute Importance

2 Data Description

The data is from direct marketing campaigns of a Portuguese banking institution, and is available on the UCI Machine learning repository. The marketing campaigns involved telephonic contact with clients to offer subscription for a term deposit with the bank. There can be multiple contacts to the same client in the dataset.

There are two datasets:

1 bank-additional-full.csv: There are 41188 records, 20 attributes and the target attribute (y), ordered by date (from May 2008 to November 2010).

2 bank-full.csv: 45211 records, 16 attributes, and the target attribute (y), ordered by date (older version of the first dataset with less attributes)

Table 1: Target Attribute

Attribute	Type	Description
Y	Binary Categorical	Subscribed to term deposit – Yes/No balance

The data set 'bank-additional-full.csv' has 88.7% records belonging to the 'no' class label and 11.3% records belonging to the 'yes' class label while 'bank-full.csv' has 88.3% records belonging to the 'no' class label and 11.7% records belonging to the 'yes' class label.

Table 2: Attributes common to both data sets

Attribute	Type	Description
Age	Numeric	Age of customer
Job	Categorical	Type of job
Marital status	Categorical	Single/Divorced/Married
Education	Categorical	Education Level
Default	Categorical	Credit in default – Yes/No
Housing	Categorical	Housing loan – Yes/No
Loan	Categorical	Personal loan – Yes/No
Contact	Categorical	Telephonic/Cellular
Month	Categorical	Month of last contact with customer
Day	Categorical	Day of last contact with the customer
Duration	Numeric	Duration of last contact (seconds)
Campaign	Numeric	Number of contacts with customer
P-days	Numeric	Number of days passed after last contact
Previous	Numeric	Number of contacts in previous campaign
Poutcome	Categorical	Outcome of previous campaign

Table 3: Attributes included in bank-additional-full.csv only

Attribute	Type	Description
Employment variation rate	Numeric	Quarterly indicator
Consumer price index	Numeric	Monthly indicator
Consumer confidence index	Numeric	Monthly indicator
Euribor	Numeric	Euro Interbank Offered Rate – three month rate – daily indicator
Number employed	Numeric	Quarterly indicator

Attributes in Table 3 are social and economic attributes for the country.

Table 4: Attributes included in bank-full.csv only

Attribute	Type	Description
Balance	Numeric	Average yearly bank balance

The demographic used in both datasets is similar in some respects. In both datasets the mean age is 40, about 65% are married, the most common jobs are 'technician', 'blue collar' and 'admin' and almost 82% have no personal loan.

3 Data Preprocessing

The datasets contain both numeric and categorical attributes. We have preprocessed the data as described below, to remove data skewness as well as to handle missing values:

1 Numerical attributes are standardized i.e. mean centered to zero and variance scaled to one.

2 The categorical attributes have been converted to binary attributes by using Integer Encoding for ordinal categorical data, and One Hot Encoding for unordered categorical data.

3 Though the dataset does not have any explicit missing values, it contains some nominal attributes which has a category called 'Unknown'. The total number of unknown values in bank-full dataset and bank-additional-full dataset are 12718 and 52124 respectively. The percentage of unknown values in both datasets are shown in Figure 1. Due to the high percentage of unknown values, we have neither estimated the values nor discarded them. Since One Hot Encoding transforms the nominal attributes into multiple binary categorical attributes, the columns corresponding to the unknown attribute have been removed from the dataset.

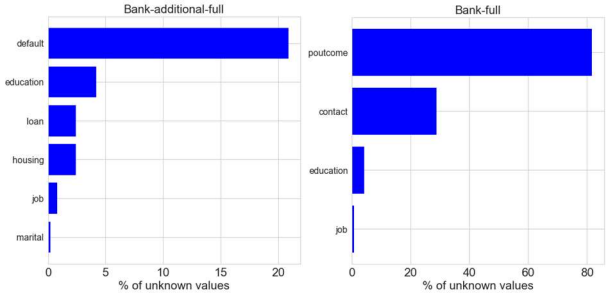

Figure 1: Percentage of unknown values

4 Data Exploration

Scatter plots have been used for preliminary analysis of the relationship between attribute pairs as well as the relationship between attributes and the target variable. In addition, correlation plots are used to visualize the correlations between the numeric attributes. The correlation matrices of both datasets are shown in Figure 2 and Figure 3.

We observe that there are some attribute pairs which have a significant correlation. In the dataset bank-full.csv, attributes 'pdays' and 'previous' have a correlation coefficient of 0.45. In the dataset bank-additional-full.csv, 'Euribor' and 'Number Employed' have a high correlation coefficient of 0.95. Other negatively correlated pairs are: 'P-days' and 'previous', 'Number Employed' and 'Previous', 'Euribor' and 'Previous'. The attribute pair 'Euribor' and 'Consumer price index' is positively correlated.

We have used Chi-squared test to establish the relationships between the categorical attributes. In bank-full.csv, categorical attributes 'housing' and 'loan' are associated. In the dataset bank-full-additional.csv, 'loan' is a categorical attribute, associated with attributes 'marital status', 'education' and 'poutcome'.

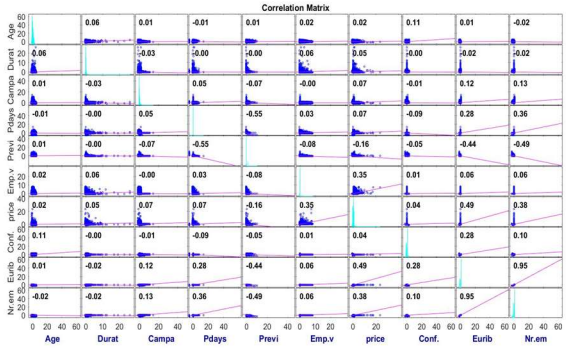

Figure 2: Correlation matrix of bank-additional dataset

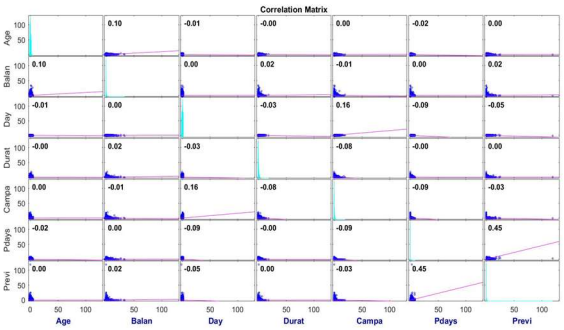

Figure 3: Correlation matrix of bank-full dataset

Boxplots are used to capture the spread of the data. We observe that attributes that though some attributes are common between the two datasets, they have different ranges. For this reason, data standardization has been included in preprocessing.

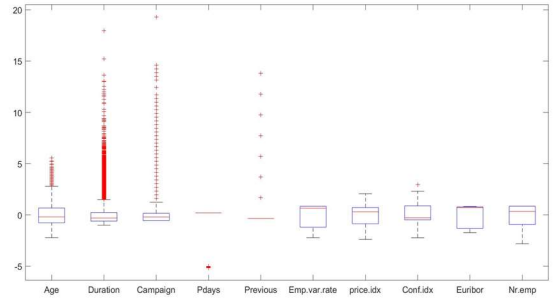

Figure 4: Box plot for bank-additional-full dataset

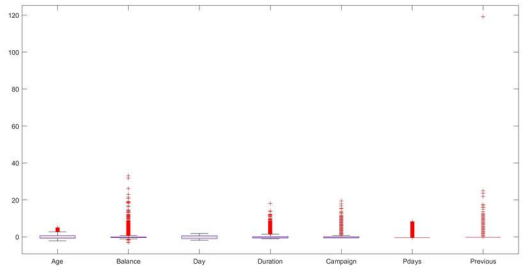

Figure 5: Box plot for bank-additional-full dataset

5 Model Building

We have conducted an ablation study on the two data sets independently, using the following classifier models:

1 Naïve Bayes Classifier
2 C4.5 Decision tree
3 Random Forest

We ran the classification models on the preprocessed dataset with all attributes to get baseline models. Then we used the ablation technique i.e. ran classifiers by removing one attribute at a time and compared the new model with the baseline model to understand whether the removal of the attribute improved or degraded the classifier model. The improvement in performance of the classifier on removal of the attribute indicates that the attribute removed is not importance whereas degradation in performance of the classifier implies that the attribute is significantly contributing to the classifier.

Since our datasets have unequal distribution of the class labels, we have used 10-fold stratified cross validation for building and testing our models.

6 Model Evaluation

We used the Kappa statistic as a measure of the performance of classifiers.

We ranked attributes based on the difference between the baseline Kappa statistic (with all attributes) and the Kappa statistic after removing the attribute. A positive difference indicates that the removal of the attribute degraded the Kappa statistic, which in turn means that the attribute is important and contributes in the classification. A negative or zero difference indicates that the Kappa statistic after removing the attribute, increases or stays the same which in turn means that the attribute was not important in building the classifier or was misleading the classifier. For each classifier, we ranked attributes in decreasing order of:

$$\text{(Kappa statistic)}_{Baseline} - \text{(Kappa statistic)}_{-attribute} \quad (1)$$

These rankings are referred to as local ranks in our description. We obtained a consolidated ranking using Mean Reciprocal Ranks, which is referred to as the global rank in this description.

$$\text{Mean Reciprocal Rank (MRR)} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (2)$$

The local and global ranks for both datasets are detailed in Table 6 and Table 6.

To select a subset of important attributes and to identify the attributes that are not important for building a classifier for the particular dataset, we iteratively removed the lowest ranked attribute and ran the classifier. Removing the lowest ranked attribute in the first iteration and running the classifier, in the next iteration we removed the two lowest ranked attributes and continued with iterations till we had removed all but the highest ranked attribute.

We observed that as we go on removing attributes starting with the lowest ranked, our classifier continues to improve (based on Kappa statistic) till a point after which it starts to degrade. This is the point where we have removed an attribute that is important to the classifier and can be thought of as the point of importance in this process.

The results for the first dataset Bank-full.csv are shown in Figures 6 to 8, and results for Bank-additional-full.csv are shown in Figures 9 to 11. The baseline Kappa is indicated in blue. The plot of the Kappa statistic, removing attributes based on local ranks is indicated in orange, and the plot based on global ranks is indicated in grey.

For each classifier, we have a local maximum (based on local ranks) and a global maximum (based on the global ranks). We observe that for Naïve Bayes, the local maximum is higher than the global maximum. For C4.5 trees (J48 in Weka), as well as for Random Forest, the global maximum is higher than the local maximum which indicates that the mean reciprocal ranking finds a smaller subset of attributes which is a better model based on the Kappa statistic. This indicates that Naïve Bayes improves the mean reciprocal ranks. One of the reasons for this could be that Naïve Bayes considers all attribute probabilities when predicting the class label, so it might be able to better distinguish between less important attributes.

For each classifier, we derived two subsets of attributes: the first subset which gives the local maximum, and the second which gives the global maximum. For each of our datasets, we have 6 such subsets, two per classifier. These are detailed in Tables 8 and 9. We compared these to see patterns of importance amongst our attributes and to derive some insights detailed in Section 8.

We have summarized the final global ranks (based on MRR ranks) for both our datasets in Table 5.

Table 5: Global Ranks

Rank	Attributes:Bank-additional-full	Attributes:Bank-full
1	Duration	Duration
2	Pdays	Poutcome
3	Euribor3m	Month
4	Month	Housing
5	Cons.conf.idx	Age
6	Poutcome	Contact
7	Marital	Day of week
8	Education	Job
9	Nr.Employed	Default
10	Default	Previous
11	Age	Education
12	Day of week	Pdays
13	Emp.var.rate	Marital
14	Contact	Loan
15	Job	Campaign
16	Loan	Balance
17	Cons.price.idx	
18	Housing	
19	Previous	

Table 6: Ranking for Bank- full.csv

Attribute removed	Naïve Baye's			C4.5 Decision tree			Random forest			Ranking based on Mean Reciprocal Rank
	Kappa	$(Kappa)_{Baseline}$ - $(Kappa)_{-Attribute}$	Rank	Kappa	$(Kappa)_{Baseline}$ - $(Kappa)_{-Attribute}$	Rank	Kappa	$(Kappa)_{Baseline}$ - $(Kappa)_{-Attribute}$	Rank	
None (Baseline)	0.4271			0.4685			0.4394			
Age	0.4088	0.0183	5	0.4491	0.0194	5	0.4406	-0.0012	12	5
Job	0.4085	0.0186	4	0.4699	-0.0014	16	0.4526	-0.0132	15	8
Marital	0.4147	0.0124	7	0.4647	0.0038	15	0.4346	0.0048	7	13
Education	0.4112	0.0159	6	0.4572	0.0113	14	0.435	0.0044	8	11
Default	0.4147	0.0124	7	0.4512	0.0173	7	0.4385	0.0009	11	9
Balance	0.4187	0.0084	10	0.4551	0.0134	13	0.4491	-0.0097	14	16
Housing	0.4061	0.021	3	0.445	0.0235	4	0.4279	0.0115	5	4
Loan	0.4156	0.0115	8	0.4521	0.0164	10	0.4379	0.0015	10	14
Contact	0.4178	0.0093	9	0.4537	0.0148	12	0.4109	0.0285	4	6
Day	0.4147	0.0124	7	0.4494	0.0191	6	0.4355	0.0039	9	7
Month	0.4232	0.0039	13	0.393	0.0755	2	0.4049	0.0345	2	3
Duration	0.2727	0.1544	1	0.2812	0.1873	1	0.2753	0.1641	1	1
Campaign	0.4211	0.006	12	0.4515	0.017	8	0.4437	-0.0043	13	15
Pdays	0.4343	-0.0072	14	0.4524	0.0161	11	0.4279	0.0115	5	12
Previous	0.4203	0.0068	11	0.4519	0.0166	9	0.4295	0.0099	6	10
Poutcome	0.3706	0.0565	2	0.4284	0.0401	3	0.4088	0.0306	3	2

Table 7: Ranking for Bank-additional-full.csv

Attribute removed	Naïve Baye's			C4.5 Decision tree			Random forest			Ranking based on Mean Reciprocal Rank
	Kappa	$(Kappa)_{Baseline}$ - $(Kappa)_{-Attribute}$	Rank	Kappa	$(Kappa)_{Baseline}$ - $(Kappa)_{-Attribute}$	Rank	Kappa	$(Kappa)_{Baseline}$ - $(Kappa)_{-Attribute}$	Rank	
None (Baseline)	0.3917			0.509			0.4972			
Age	0.3909	0.008	7	0.5052	0.0038	6	0.5032	-0.0060	13	11
Job	0.3895	0.0022	5	0.5091	-0.0001	14	0.5192	-0.0220	19	15
Marital	0.3935	-0.0018	12	0.5012	0.0078	4	0.5029	-0.0057	12	7
Education	0.3885	0.0032	4	0.5076	0.0014	10	0.5180	-0.0208	18	8
Default	0.3897	0.0020	6	0.5065	0.0025	7	0.5021	-0.0049	11	10
Housing	0.3913	0.0004	9	0.5091	-0.0001	14	0.5066	-0.0094	16	18
Loan	0.3912	0.0005	8	0.5118	-0.0028	18	0.4997	-0.0025	8	16
Contact	0.3949	-0.0032	13	0.5072	0.0018	9	0.4996	-0.0024	7	14
Month	0.3882	0.0035	3	0.5093	-0.0003	15	0.4961	0.0011	4	4
Day of week	0.3914	0.0003	10	0.5086	0.0004	11	0.4974	-0.0002	6	12
Duration	0.3270	0.0647	1	0.3216	0.1874	1	0.3254	0.1718	1	1
Campaign	0.3966	-0.0049	14	0.5093	-0.0003	15	0.5054	-0.0082	15	20
Pdays	0.3807	0.0110	2	0.5090	0	13	0.4941	0.0031	2	2
Previous	0.4193	-0.0276	18	0.5099	-0.0009	17	0.5010	-0.0038	10	19
Poutcome	0.4300	-0.0383	19	0.5006	0.0084	3	0.5113	-0.0141	17	6
Emp.var. rate	0.4116	-0.0199	17	0.5089	0.0001	12	0.4971	0.0001	5	13
Cons.price.idx	0.3978	-0.0061	15	0.5070	0.0020	8	0.5047	-0.0075	14	17
Cons.conf. idx	0.3912	0.0005	8	0.5094	-0.0004	16	0.4957	0.0015	3	5
Euribor3m	0.4111	-0.0194	16	0.4986	0.0104	2	0.4941	0.0031	2	3
Nr.employed	0.3916	0.0001	11	0.5023	0.0067	5	0.5003	-0.0031	9	9

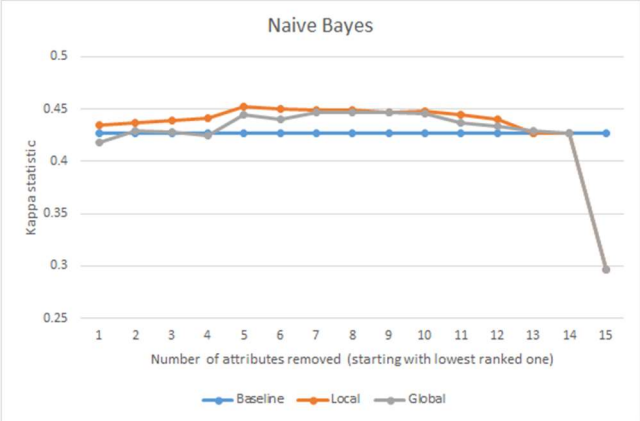

Figure 6: Bank-full.csv: Naive Bayes: Plot of Kappa statistic vs number of lowest ranked attributes removed

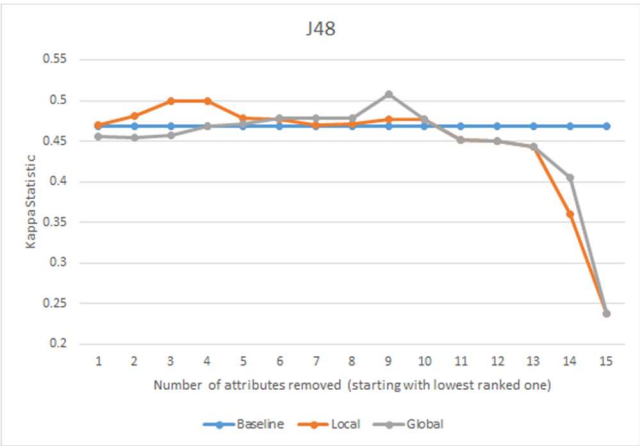

Figure 7: Bank-full.csv: J48: Plot of Kappa statistic vs number of lowest ranked attributes removed

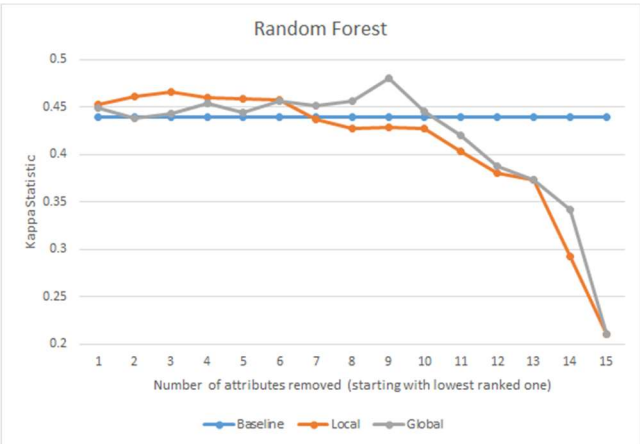

Figure 8: Bank-full.csv: Random Forest: Plot of Kappa statistic vs number of lowest ranked attributes removed

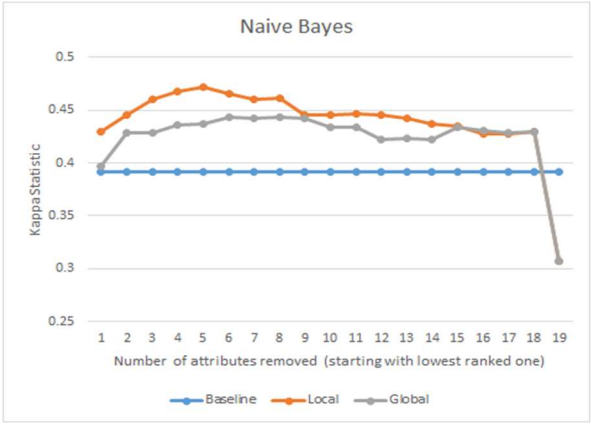

Figure 9: Bank-additional-full.csv: Naive Bayes: Plot of Kappa statistic vs number of lowest ranked attributes removed

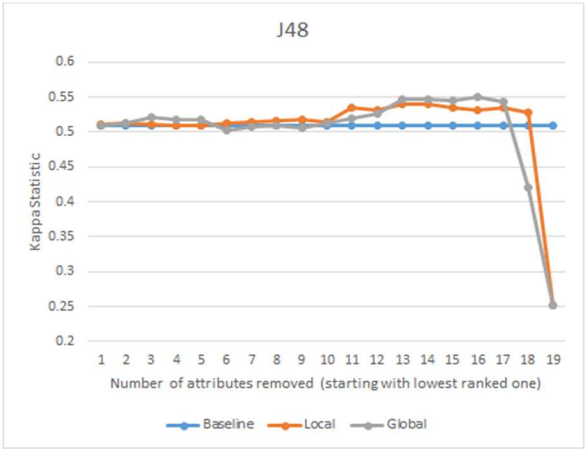

Figure 10: Bank-additional-full.csv: J48: Plot of Kappa statistic vs number of lowest ranked attributes

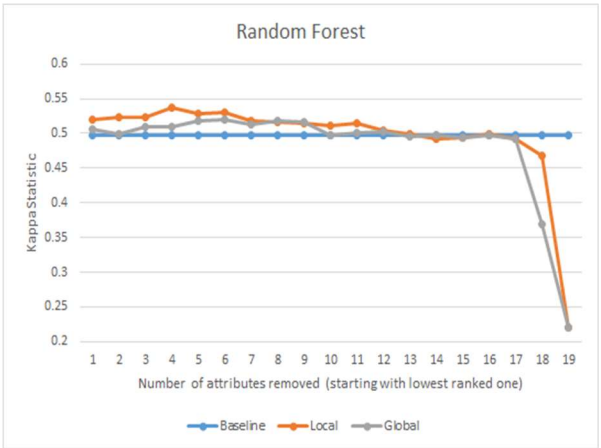

Figure 11: Bank-additional-full.csv: Random Forest: Plot of Kappa statistic vs number of lowest ranked attributes

Table 8: Subsets for bank-full.csv

Naïve Bayes Subsets	Global	Duration, Poutcome, Month, Housing, Age, Contact, Day, Job and Default
	Local	Duration, Poutcome, Housing, Job, Age, Education, Default, Day, Marital Status, Loan and Contact
Decision Tree Subsets	Global	Duration, Poutcome, Month, Housing, Age, Contact and Day
	Local	Duration, Month, Poutcome, Housing, Age, Day, Default, Campaign Previous, Loan, Pdays and Contact
Random Forest Subsets	Global	Duration, Poutcome, Month, Housing, Age, Contact and Day
	Local	Duration, Month, Poutcome, Contact, Housing, Pdays, Previous, Marital Status, Education, Day, Loan, Default and Age

Table 9: Subsets for bank-full-additional.csv

Naïve Bayes Subsets	Global	Duration, Pdays, Euribor3m, Month, Cons.conf.idx, Poutcome, Marital Status, Education, Nr.Employed, Default, Age and Day of week
	Local	Duration, Pdays, Month, Education, Job, Default, Age, Cons.conf.idx, Loan, Housing, Day of week, Nr.Employed, Marital Status, Contact and Campaign
Decision Tree Subsets	Global	Duration, Pdays, Euribor3m, Month, Cons.conf.idx and Poutcome
	Local	Duration, Euribor3m, Poutcome, Marital Status, Nr.Employed, Age, Default, Cons.price.idx and Contact
Random Forest Subsets	Global	Duration, Pdays, Euribor3m, Month, Cons.conf.idx, Poutcome, Marital Status, Education, Nr.Employed, Default, Age, Day of week, Emp.var.rate and Contact
	Local	Duration, Pdays, Euribor3m, Cons.conf.idx, Month, Emp.var.rate, Day of week, Contact, Loan, Nr.Employed, Previous, Default, Marital Status, Age, Cons.price.idx and Campaign

7 Contributions

The preprocessing, model building, and model evaluation work steps are performed by Meghana for 'bank-full.csv' and Pranavi for 'bank-full-additional.csv', for all the three classifiers.

Comparison of data and the analysis on importance of attributes were done jointly.

8 Real World Insights

The purpose of our analysis is to determine which of the input attributes out of the client data highly impact client subscription to term deposits and which attributes are not adding to the classifier. Based on the subsets obtained, we listed the attributes missing on most subsets and categorized them as unimportant. For Bank-full.csv dataset, attribute 'balance', which is exclusive to this dataset, is ranked lowest. This indicates that the bank balance of the customer does not impact the decision to subscribe to term deposits. The attributes 'campaign', 'pdays' and 'previous' are related to number of contacts and time elapsed after last contact in the current and previous campaigns. Our analysis shows that these are not important. The personal information related attributes which include 'education', 'marital status' and 'job' are also not important.

In Bank-additional-full.csv, 'job' is missing in 5 of the 6 subsets. This indicates that a person's job isn't important in the decision of subscribing to a term deposit. Similarly, whether a person has a personal loan or housing loan, as indicated by attributes 'loan' and 'housing' are also unimportant in this decision. Similar to bank-full.csv, attributes related to number of contacts, 'previous' and 'campaign' are unimportant.

Bank-additional-full.csv has five extra attributes which are social and economic attributes for the country. Our analysis shows that out of these, employment variation rate and consumer price index do not seem to be important. However, the Euro Interbank Offered Rate, which influences interest rates, has a great impact on a client's decision to subscribe to a term deposit. The Consumer Confidence Index, which also indicates the health of the economy based on trends in savings and spending, contributes as an important attribute. The attribute 'number employed' which indicates the employment rate is also significant. These three attributes add value to the newer dataset bank-additional-full.csv.

We see that some attributes' importance has changed over the two datasets. In the older data set bank-full.csv, contact related attributes are not very important. The attribute 'pdays', ie. number of days that passed by after the client was last contacted from a previous campaign grows in importance in the new dataset, but did not contribute in the older dataset. Education is important in the newer dataset, as this has more detailed categories compared to the older dataset. Whether or not a client has a housing loan, did not seem to contribute to the classifier in the older dataset, but in the new dataset, this grows in importance. Marital status was not important in the older dataset but is ranked high in the new dataset.

In terms of importance, 'duration' is the most significant across datasets. This means that the duration of the last contact with the client is an indicator of subscription, even though number of contacts is not. 'Month' is important across both datasets, which indicates that term subscriptions are seasonal. 'Default' which describes whether or not a person has defaulted their credit, is equally important across both the datasets. 'Poutcome', which is the outcome of the previous marketing campaign signifies repetition in subscription.

Based on our analysis, customer databases for term deposit subscription information can comprise of fewer attributes. The attributes categorized as unimportant can be left out. This should help narrow down the customer base most likely to subscribe, so that the direct marketing campaigning resources can be spent on these, instead of targeting all clients as potential term deposit buyers.

REFERENCES

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

[2] [Moro et al., 2014] 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

[3] https://www.bportugal.pt/estatisticasweb