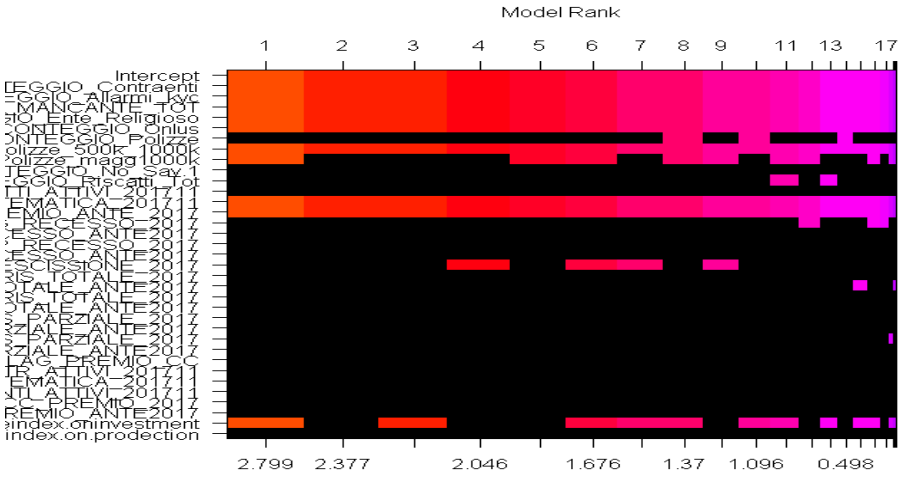
Clustering Repot

Variable importance was calculated from the combined dataset by Bayesian Adaptive Sampling for Bayesian Model Averaging and Variable Selection in Linear Models for indice & I1 & I2 after removing outliers which was based on Bayesian outlier detection. Calculation was based on the BIC parameter and we gave uniform weights for each 2^p models to find the important variables to explain dependent variables. After determining the variable importance for each variable outliers were added back to data set. Lastly base on these important variable we created a new dataset and did our calculations based on the new dataset we created. Where new data set is consist of 1353 obs. Of 36 variables where 12 of them non numeric,

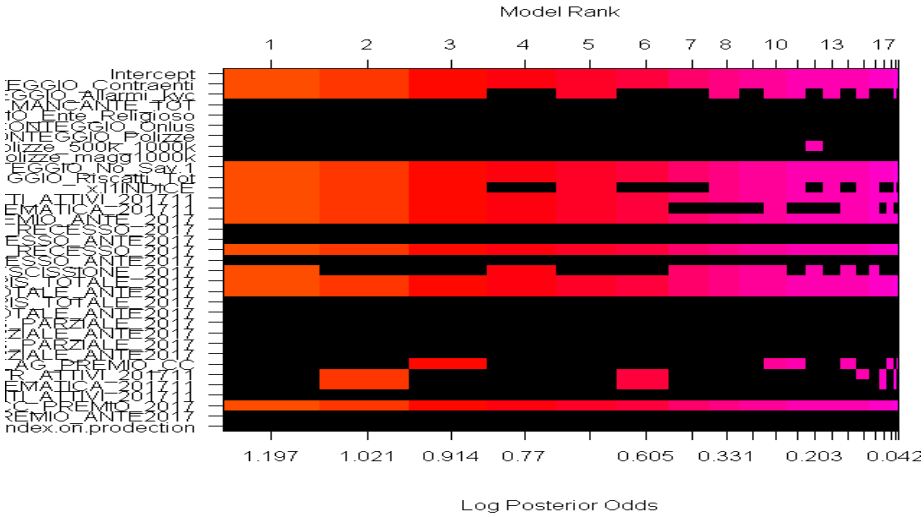
Variable importance for Indice:

indiceCONTEGGIO_Contraenti
indiceCONTEGGIO_Allarmi_kyc
indiceCONTEGGIO_DOC_MANCANTE_TOT
indiceCONTEGGIO_Ente_Religioso
indiceCONTEGGIO_Onlus
indiceCONTEGGIO_Polizze_500k_1000k
indiceCONTEGGIO_Polizze_magg1000k
indiceRISERVA_MATEMATICA_201711
indicePREMIO_ANTE_2017
indiceFLAG_RESCISSIONE_2017
indiceindex.oninvestment



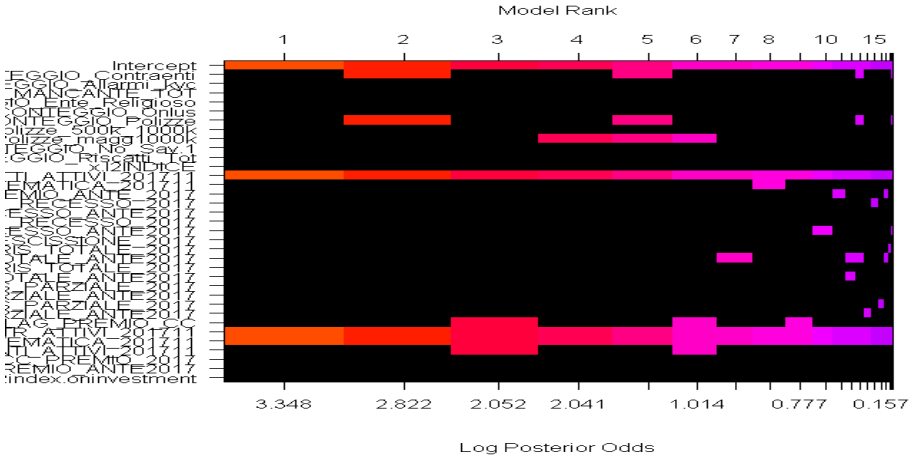
Variable importance for I1:

I1CONTEGGIO_Contraenti
I1CONTEGGIO_Allarmi_kyc
I1CONTEGGIO_DOC_MANCANTE_TOT
I1CONTEGGIO_No_Sav.1
I1CONTEGGIO_Riscatti_Tot
I1INDICE
I1CONTRATTI_ATTIVI_201711
I1RISERVA_MATEMATICA_201711
I1PREMIO_ANTE_2017
I1IMP_RECESSO_2017
I1FLAG_RESCISSIONE_2017
I1FLAG_RIS_TOTALE_2017
I1FLAG_RIS_TOTALE_ANTE2017
I1FLAG_PREMIO_CC
I1TCM_NUM_CONTR_ATTIVI_201711
I1TCM_RIS_MATEMATICA_201711
I1CC_PREMIO_2017



Variable importance for I2:

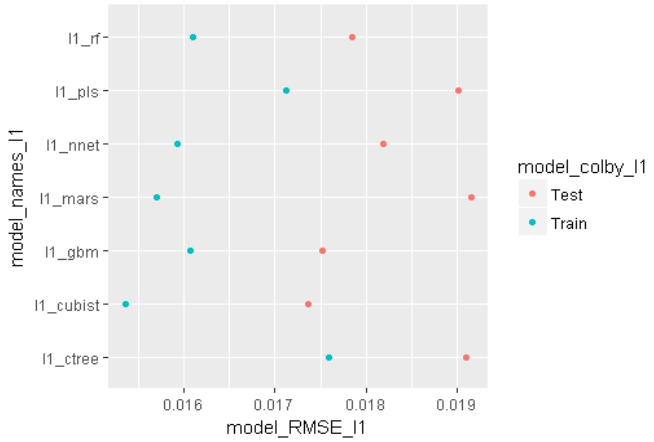
I2CONTEGGIO_Contraenti
I2CONTRATTI_ATTIVI_201711
I2FLAG_PREMIO_CC
I2TCM_NUM_CONTR_ATTIVI_201711
I2TCM_RIS_MATEMATICA_201711
I2CC_NUM_CLIENTI_ATTIVI_201711

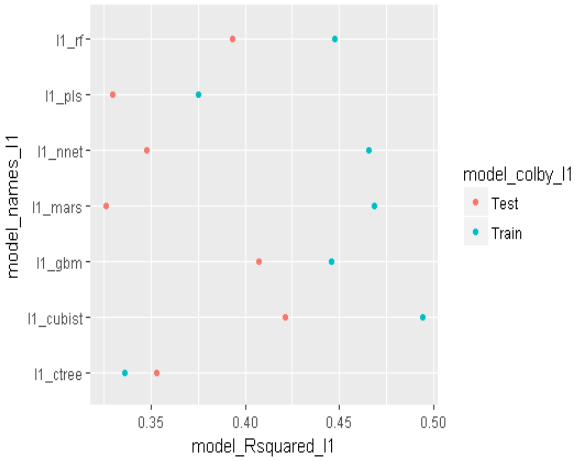


```
> cor.test(bysn.combined.aml$index.on.prodection, bysn.combined.aml$index.oninvestment)

        Pearson's product-moment correlation

data:  bysn.combined.aml$index.on.prodection and bysn.combined.aml$index.oninvestment
t = -1.7365, df = 1351, p-value = 0.0827
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.100232198  0.006116954
sample estimates:
        cor
-0.04719135

> cor.test(bysn.combined.aml$index.on.prodection, bysn.combined.aml$INDICE)

        Pearson's product-moment correlation

data:  bysn.combined.aml$index.on.prodection and bysn.combined.aml$INDICE
t = 0.5494, df = 1351, p-value = 0.5828
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03837779  0.06818432
sample estimates:
      cor
0.0149457

> cor.test(bysn.combined.aml$INDICE, bysn.combined.aml$index.oninvestment)

        Pearson's product-moment correlation

data:  bysn.combined.aml$INDICE and bysn.combined.aml$index.oninvestment
t = 4.4276, df = 1351, p-value = 1.03e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06672768 0.17179329
sample estimates:
      cor
0.1195953
```

→The following comparison was graph based on the means,

Estimating I1 ->index.on.investment





RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I1	0.01542602	0.01641275	0.01751741	0.01759150	0.01919424	0.01949445	0
rf_I1	0.01284787	0.01502519	0.01568235	0.01609822	0.01683771	0.01951014	0
gbm_I1	0.01306033	0.01408800	0.01628811	0.01607586	0.01798763	0.01904597	0
pls_I1	0.01437217	0.01518400	0.01701572	0.01712603	0.01854567	0.02037379	0
mars_I1	0.01229912	0.01454462	0.01504471	0.01569839	0.01688562	0.02009571	0
nnet_I1	0.01341288	0.01471561	0.01571494	0.01592296	0.01661212	0.01884627	0
cbst_I1	0.01236540	0.01442805	0.01528627	0.01536186	0.01637243	0.01952784	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I1	0.2859043	0.2969305	0.3161927	0.3362339	0.3690729	0.4188746	0
rf_I1	0.3523008	0.4171862	0.4580275	0.4476496	0.4932004	0.4970251	0
gbm_I1	0.3832324	0.4248606	0.4332964	0.4455592	0.4624253	0.5376617	0
pls_I1	0.2342569	0.3329217	0.3955806	0.3753043	0.4118315	0.5029045	0
mars_I1	0.3419738	0.3918476	0.4745209	0.4685313	0.5378474	0.6090214	0
nnet_I1	0.3108030	0.4627664	0.4836370	0.4657853	0.5040439	0.5314476	0
cbst_I1	0.4022848	0.4228213	0.5006847	0.4939933	0.5602743	0.5813861	0

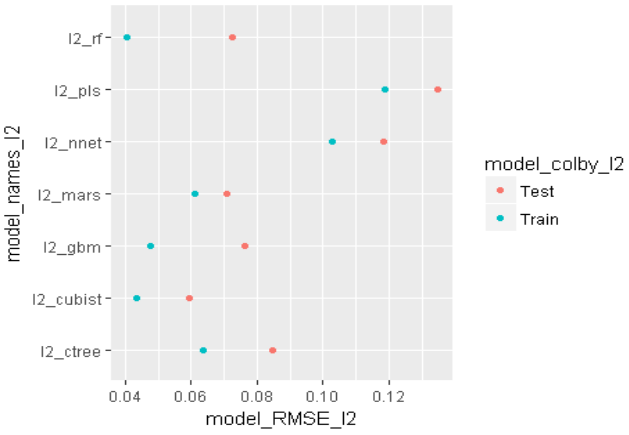
RMSE

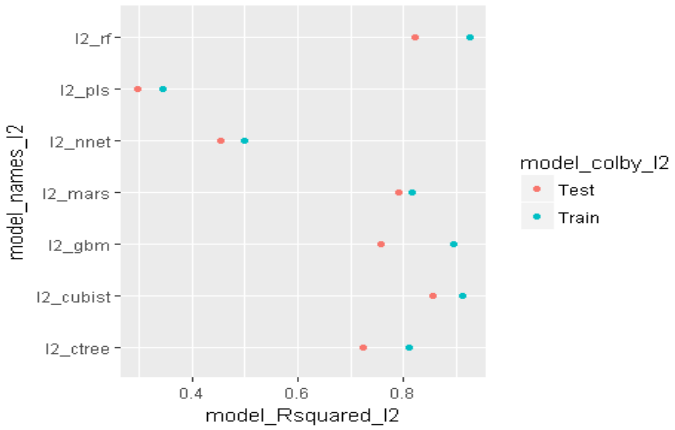
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I1_test	0.01186228	0.01498069	0.01772579	0.01908971	0.02155228	0.03316876	0
rf_I1_test	0.01226679	0.01496519	0.01749706	0.01783970	0.02075557	0.02527725	0
gbm_I1_test	0.01150516	0.01461999	0.01661548	0.01751595	0.01892843	0.02786480	0
pls_I1_test	0.01423982	0.01501247	0.01754712	0.01900199	0.02208771	0.02883784	0
mars_I1_test	0.01343995	0.01575383	0.01830090	0.01914568	0.02348802	0.02601192	0
nnet_I1_test	0.01381960	0.01568037	0.01704357	0.01818624	0.01813087	0.02763419	0
cbst_I1_test	0.01324759	0.01425276	0.01512506	0.01735688	0.01931261	0.02937716	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I1_test	0.01101860	0.2854048	0.3895197	0.3528524	0.4659670	0.6100093	0
rf_I1_test	0.18307675	0.3206591	0.3828809	0.3933717	0.4398956	0.7207495	0
gbm_I1_test	0.15766924	0.2313889	0.4521077	0.4074834	0.5553492	0.6500910	0
pls_I1_test	0.07407383	0.3013056	0.3508209	0.3296193	0.4059688	0.5187042	0
mars_I1_test	0.05529647	0.2386155	0.2792297	0.3261955	0.4683707	0.5372340	0
nnet_I1_test	0.17028488	0.2921902	0.3457919	0.3475166	0.3943688	0.5402541	0
cbst_I1_test	0.25495345	0.3194055	0.3772459	0.4214019	0.5044294	0.6912741	0

Estimating I2 ->index.on.investment





RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I2	0.04585467	0.05783657	0.06731025	0.06372817	0.06936552	0.07816629	0
rf_I2	0.02845453	0.03353768	0.03928597	0.04040607	0.04823362	0.05171032	0
gbm_I2	0.03286040	0.04060742	0.05009036	0.04766317	0.05414199	0.06005584	0
pls_I2	0.08370220	0.10723381	0.12545963	0.11882468	0.12840907	0.13814785	0
mars_I2	0.03676346	0.04385694	0.04664393	0.06130615	0.08173271	0.10636585	0
nnet_I2	0.07332708	0.08084510	0.09801012	0.10294320	0.11607041	0.15991970	0
cbst_I2	0.02087343	0.02624006	0.04334785	0.04329685	0.05101579	0.08019352	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I2	0.7530929	0.7833580	0.8126783	0.8105757	0.8412591	0.8694705	0
rf_I2	0.8526138	0.9154663	0.9364723	0.9248587	0.9446754	0.9630564	0
gbm_I2	0.8440580	0.8686723	0.8949590	0.8955251	0.9162521	0.9477657	0
pls_I2	0.2307107	0.2651897	0.3379017	0.3447531	0.3799344	0.5624962	0
mars_I2	0.6120466	0.7361073	0.8570340	0.8172974	0.9066151	0.9231193	0
nnet_I2	0.1997073	0.4266706	0.5053577	0.4989399	0.6162053	0.7062162	0
cbst_I2	0.8315106	0.8518614	0.9359500	0.9127353	0.9613200	0.9771620	0

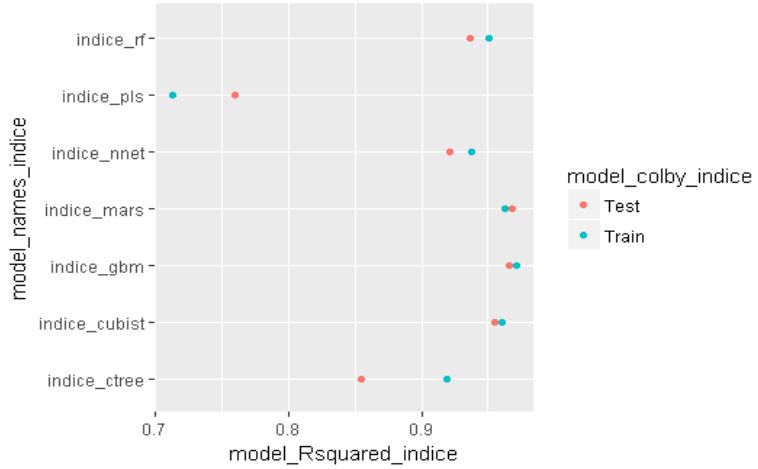
RMSE

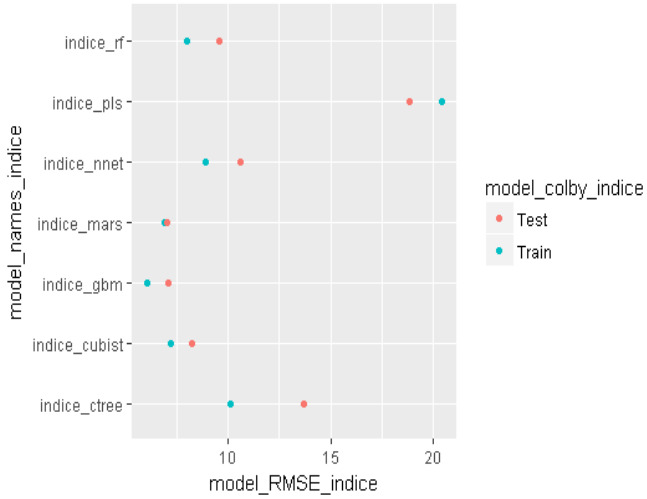
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I1_test	0.01186228	0.01498069	0.01772579	0.01908971	0.02155228	0.03316876	0
rf_I1_test	0.01226679	0.01496519	0.01749706	0.01783970	0.02075557	0.02527725	0
gbm_I1_test	0.01150516	0.01461999	0.01661548	0.01751595	0.01892843	0.02786480	0
pls_I1_test	0.01423982	0.01501247	0.01754712	0.01900199	0.02208771	0.02883784	0
mars_I1_test	0.01343995	0.01575383	0.01830090	0.01914568	0.02348802	0.02601192	0
nnet_I1_test	0.01381960	0.01568037	0.01704357	0.01818624	0.01813087	0.02763419	0
cbst_I1_test	0.01324759	0.01425276	0.01512506	0.01735688	0.01931261	0.02937716	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_I1_test	0.01101860	0.2854048	0.3895197	0.3528524	0.4659670	0.6100093	0
rf_I1_test	0.18307675	0.3206591	0.3828809	0.3933717	0.4398956	0.7207495	0
gbm_I1_test	0.15766924	0.2313889	0.4521077	0.4074834	0.5553492	0.6500910	0
pls_I1_test	0.07407383	0.3013056	0.3508209	0.3296193	0.4059688	0.5187042	0
mars_I1_test	0.05529647	0.2386155	0.2792297	0.3261955	0.4683707	0.5372340	0
nnet_I1_test	0.17028488	0.2921902	0.3457919	0.3475166	0.3943688	0.5402541	0
cbst_I1_test	0.25495345	0.3194055	0.3772459	0.4214019	0.5044294	0.6912741	0

Estimating Indice





RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_indice	9.029230	9.828524	10.161198	10.096756	10.503884	11.176221	0
rf_indice	6.796377	7.739795	8.061188	7.983610	8.482594	8.756092	0
gbm_indice	5.427057	5.636639	5.941744	6.016301	6.228426	6.994148	0
pls_indice	16.179238	17.953358	18.773233	20.424679	20.025395	36.968546	0
mars_indice	5.980166	6.368818	6.967507	6.915122	7.275391	8.310679	0
nnet_indice	7.785479	8.314171	8.757670	8.874570	9.393729	10.288779	0
cbst_indice	6.068120	6.508730	7.223263	7.182089	7.656383	8.357050	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_indice	0.8857935	0.9101766	0.9230331	0.9184522	0.9259165	0.9426996	0
rf_indice	0.9413274	0.9473544	0.9501459	0.9506102	0.9533528	0.9609760	0
gbm_indice	0.9619747	0.9668890	0.9722686	0.9713866	0.9755456	0.9791575	0
pls_indice	0.4587045	0.7122840	0.7272693	0.7122850	0.7544259	0.7895978	0
mars_indice	0.9509894	0.9576298	0.9649955	0.9626640	0.9676132	0.9703773	0
nnet_indice	0.9125504	0.9250454	0.9425518	0.9372015	0.9455476	0.9582405	0
cbst_indice	0.9397953	0.9527484	0.9646043	0.9601957	0.9675202	0.9755200	0

RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_indice_test	8.787414	12.185092	13.722466	13.665028	16.259373	17.091905	0
rf_indice_test	7.330678	7.818214	8.655143	9.543438	10.577960	13.444280	0
gbm_indice_test	6.177040	6.756068	6.975983	7.048507	7.237112	8.274875	0
pls_indice_test	15.255240	17.385341	19.492933	18.821389	20.137287	21.395492	0
mars_indice_test	4.916345	6.790318	7.312377	7.019557	7.470368	7.852858	0
nnet_indice_test	7.727798	8.857104	9.463006	10.590604	12.881061	13.660112	0
cbst_indice_test	6.044474	7.507581	8.081364	8.217886	8.505008	11.935196	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ctree_indice_test	0.7636808	0.7827571	0.8676797	0.8545207	0.9163721	0.9422405	0
rf_indice_test	0.8858765	0.9257331	0.9398799	0.9364682	0.9558248	0.9625534	0
gbm_indice_test	0.9535288	0.9595288	0.9664424	0.9662817	0.9724500	0.9816326	0
pls_indice_test	0.6807267	0.7458886	0.7485130	0.7599599	0.7973212	0.8185136	0
mars_indice_test	0.9568033	0.9633877	0.9641683	0.9679698	0.9725227	0.9845424	0
nnet_indice_test	0.8698511	0.9032943	0.9320379	0.9211965	0.9376427	0.9614333	0
cbst_indice_test	0.9319900	0.9477834	0.9553141	0.9553093	0.9636814	0.9796752	0

After modelling indice & I1 & I2, we found rf, gbm and cubist are the best models which can be used for the prediction. In this study we group indice & I1 & I2 and clustered them with using KNN, PAM, Clara and Hierarchical clustering method with the same distance method ("euclidean") and we model again the clustered columns one by one and found clustering method can be explain better by rf and gbm (note that cubist is only for the regression models.)

Furthermore, we select KNN clustering method and found the means for each cluster;

Row Labels	Average of INDICE	Average of index.oninvestment	Average of index.on.prodection
1	179.8181818	0.037727273	0.249545455
2	116.1	0.03075	0.24275
3	103.6631579	0.027368421	0.281578947
4	153.5245902	0.032786885	0.230327869
5	133.7307692	0.027211538	0.220384615
6	53.6090535	0.025226337	0.233045267
7	65.05820106	0.02978836	0.22994709
8	37.51769912	0.02199115	0.230530973
9	77.44805195	0.03025974	0.254805195
10	91.91366906	0.030071942	0.223381295
Grand Total	**80.94161123**	**0.027730968**	**0.237117517**

The table above shows;

- ➔ Mean Indice = 80.94
- ➔ Mean I1 = 0.02773 (index.on.investment)
- ➔ Mean I2 =0.2371 (index.on.prodection)

Since cluster 1 is exceeding our calculations, we can conclude that cluster 1 is consisted of the most dangerous branches and Cluster 2,3,4,5 also may be investigated in the near future.

Note that: Variable selection should be calculated also with AIC, MCMC methods and one most compare how test and train datasets are close to each other for the most accurate data selection. (train and test sets are based on the stratified sampling) For the clustering analysis further studying may indicate the use of more complicated unsupervised technics.

The location of each cluster

Count of c_knn_I1I2Indice	Column Labels										
Row Labels	1	2	3	4	5	6	7	8	9	10	Grand Total
Abruzzo		1	2	4	1	3	5	8	3	1	28
Basilicata			1	1	1		4	1		1	9
Calabria	5		1	3	1	1	2	1	2	4	20
Campania	8	15	16	15	20	15	18	11	16	13	147
Lazio	1	28	30	10	29	70	49	61	47	36	361
Marche		6	2	3	4	13	7	9	5	9	58
Molise		1	1	1	2	4	1	5	1	1	17
Puglia		9	4	2	13	20	18	17	23	13	119
Sardegna		2	3		3	10	6	14	1	3	42
Sicilia	8	50	25	14	23	46	38	24	37	38	303
Toscana		1	4	5	4	34	16	30	9	8	111
Trentino-Alto Adige		1	1	2		3	2	16	3	3	31
Umbria		5	3	1	3	16	14	13	5	7	67
Veneto		1	2			8	9	16	2	2	40
Grand Total	**22**	**120**	**95**	**61**	**104**	**243**	**189**	**226**	**154**	**139**	**1353**

Location of Cluster1

ANNOMESE	COD_CAB_GE	DES_REGION	DES_PROVINCIA_ABI_G	FILIALE_ABI_GEST	INDIRIZZO	COD_ABI_GE	REGIONE	AREA_COMM	AREA_COMM.1	c_knn_I1I2Indice
201706	40023	Campania	Napoli	FIL. DI NOLA	VIA ROMA, 30	2008	6 - SUD	AREA RETAIL NAPOLI PR	AREA RETAIL NAPOLI PR	1
201706	81491	Calabria	Reggio Calabria	FIL. DI PALMI	VIA ROMA, 23	2008	6 - SUD	AREA RETAIL CALABRIA	AREA RETAIL CALABRIA	1
201706	83330	Sicilia	Caltanissetta	FIL. DI GELA	CORSO VITTORIO EMAN	2008	7 - SICILIA	AREA RETAIL CALTANISS	AREA RETAIL CALTANISS	1
201706	39892	Campania	Napoli	FIL. DI FRATTAMAGGIOF	CORSO DURANTE, 201 A	2008	6 - SUD	AREA RETAIL NAPOLI PR	AREA RETAIL NAPOLI PR	1
201706	16530	Sicilia	Messina	FIL. DI MESSINA	PIAZZA CAIROLI, 46	2008	7 - SICILIA	AREA RETAIL MESSINA	AREA RETAIL MESSINA	1
201706	42830	Calabria	Vibo-Valentia	FIL. DI VIBO VALENTIA	VIA ENRICO GAGLIARDI,	2008	6 - SUD	AREA RETAIL CALABRIA	AREA RETAIL CALABRIA	1
201706	25904	Sicilia	Trapani	FIL. DI MARSALA	VIA XI MAGGIO,91	2008	7 - SICILIA	AREA RETAIL TRAPANI	AREA RETAIL TRAPANI	1
201706	16511	Sicilia	Messina	FIL. DI MESSINA	VIA GARIBALDI 102 - CO	2008	7 - SICILIA	AREA RETAIL MESSINA	AREA RETAIL MESSINA	1
201706	16303	Calabria	Reggio Calabria	FIL. DI REGGIO CALABRI	CORSO GARIBALDI 331	2008	6 - SUD	AREA RETAIL CALABRIA	AREA RETAIL CALABRIA	1
201706	16918	Sicilia	Catania	FIL. DI CATANIA	CORSO SICILIA,8	2008	7 - SICILIA	AREA RETAIL CATANIA	AREA RETAIL CATANIA	1
201706	4404	Calabria	Catanzaro	FIL. DI CATANZARO	PIAZZA BASILICA DELL'I	2008	6 - SUD	AREA RETAIL CALABRIA	AREA RETAIL CALABRIA	1
201706	3443	Campania	Napoli	FIL. DI NAPOLI	VIA VERDI, 18,D	2008	6 - SUD	AREA RETAIL NAPOLI CE	AREA RETAIL NAPOLI CE	1
201706	4615	Sicilia	Palermo	FIL. DI PALERMO	VIA TERRASANTA,8	2008	7 - SICILIA	AREA RETAIL PALERMO (	AREA RETAIL PALERMO (	1
201706	5340	Lazio	Roma	FIL. DI ROMA	VIALE G. CESARE, 54-C	2008	5 - CENTRO	AREA RETAIL ROMA PRA	AREA RETAIL ROMA PRA	1
201706	14906	Campania	Caserta	FIL. DI CASERTA	CORSO TRIESTE ANG. P	2008	6 - SUD	AREA RETAIL CASERTA	AREA RETAIL CASERTA	1
201706	16600	Sicilia	Agrigento	FIL. DI AGRIGENTO	PIAZZALE ALDO MORO,1	2008	7 - SICILIA	AREA RETAIL AGRIGENT	AREA RETAIL AGRIGENT	1
201706	75042	Campania	Caserta	FIL. DI SANTA MARIA CA	CORSO ALDO MORO, 75	2008	6 - SUD	AREA RETAIL CASERTA	AREA RETAIL CASERTA	1
201706	16304	Calabria	Reggio Calabria	FIL. DI REGGIO CALABRI	VIA DEGLI ARCONTI, 6	2008	6 - SUD	AREA RETAIL CALABRIA	AREA RETAIL CALABRIA	1
201706	40132	Campania	Napoli	FIL. DI QUARTO	VIA CUCCARO, 1	2008	6 - SUD	AREA RETAIL NAPOLI CE	AREA RETAIL NAPOLI CE	1
201706	43440	Sicilia	Palermo	FIL. DI MISILMERI	PIAZZA COMITATO 1860	2008	7 - SICILIA	AREA RETAIL PALERMO I	AREA RETAIL PALERMO I	1
201706	14903	Campania	Caserta	FIL. DI CASERTA	PIAZZA VANVITELLI, 25	2008	6 - SUD	AREA RETAIL CASERTA	AREA RETAIL CASERTA	1
201706	89800	Campania	Caserta	FIL. DI CAPODRISE	S.S. SANNITICA 87 - KM.	2008	6 - SUD	AREA RETAIL CASERTA	AREA RETAIL CASERTA	1