Machine Learning mini project

Katayoon Sadeghi mehr - March 2020

Project: Income (Classification)

Abstract:

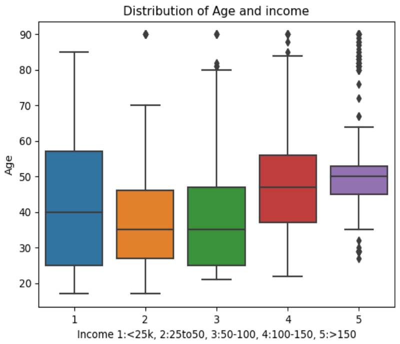
Different studies suggested that there are some features affected the amount of earnings people can make, whether they are employed or self-employed. The purpose of this study is to find the best model, using Logistic regression, KNN, Adaboost, Random Forest and SVM, to find out what are the features affecting the income.

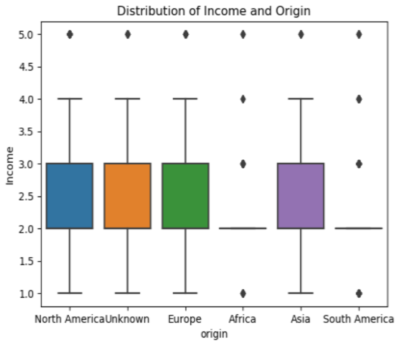
Data Explanation:

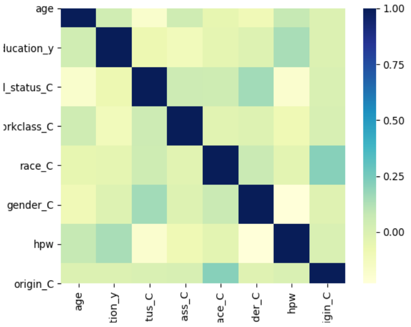
The data contains 48842 rows and the explanatory variables are age, education_y (years of formal education), marital status (coded), work class (coded), race (coded), gender (coded), hpw (hours of work her week), origin (coded from countries of origin. The dependent variable is income with 5 categories as less than 25k, 25-50k, 50-100k, 100-150k and more than 150k.

Findings:

The data shows that older people have more income and people from Africa and South America seems to earn less compared to other part of the world.







There no significant correlation in the features, so the model will be tested on all the Xs.

Logistic Regression: Applying logistic regression, on 75% of the data as training dataset and the rest of 25% of test dataset, p-values indicate the all the 8 features are significant. With R2(Train)=0.56 and R2(Test)=0.55.

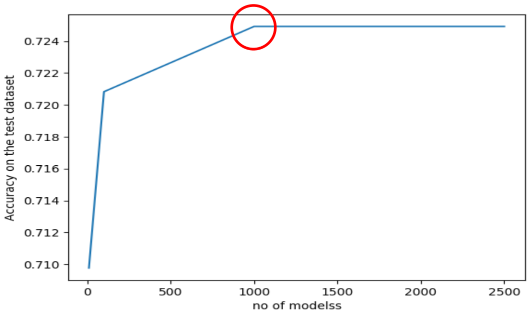
However the confusion matrix doesn't locate the values well. With accuracy of 0.55 this as a good model to fit the data.

	<25K	25-50K	50-100K	100-150K	>150K	Recall
<25K	0	1154	27	46	0	0
25-50K	0	6121	134	180	0	0.95
50-100K	0	2026	226	157	0	0.09
100-150K	0	1138	63	440	0	0.27
>150K	0	355	7	137	0	0
Precision	0	0.57	0.49	0.46	0	

Random Forest: Applying RF, using different number of estimators we see that after n_estimators= 1000 there is no gain in the accuracy.

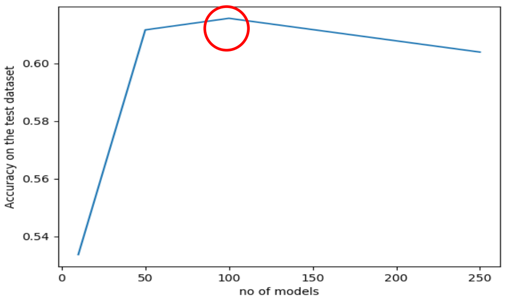
The highest possible accuracy is about 0.72.

(Here the data is not scaled.)



	<25K	25-50K	50-100K	100-150K	>150K	Recall
<25K	715	288	149	57	18	0.58
25-50K	66	5651	311	332	75	0.88
50-100K	109	520	1575	165	40	0.65
100-150K	74	545	148	747	127	0.46
>150K	9	158	41	127	164	0.33
Precision	0.73	0.79	0.71	0.52	0.39	

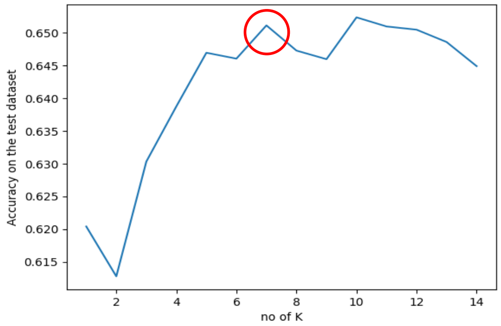
Adaboost: Applying adaboost, using 100 as the number of estimators gives the highest accuracy of 0.62.



	<25K	25-50K	50-100K	100-150K	>150K	Recall
<25K	519	439	135	133	1	0.42
25-50K	292	5669	233	235	6	0.88
50-100K	152	1402	658	181	16	0.27
100-150K	87	707	142	637	68	0.39
>150K	4	245	41	137	72	0.14
Precision	0.49	0.67	0.54	0.48	0.44	

KN Neighbors: Applying KNN Classifier and looking at the accuracy for different Ks, K=7 gives the highest accuracy of 0.65.

	<25K	25-50K	50-100K	100-150K	>150K	Recall
<25K	427	625	125	43	7	0.35
25-50K	198	5480	399	315	43	0.85
50-100K	134	809	1288	149	29	0.53
100-150K	56	681	130	665	109	0.41
>150K	11	195	44	158	91	0.18
Precision	0.52	0.70	0.65	0.50	0.33	



SVM: Applying SVM, RBF, Gamma=10 and C=10 The accuracy is 0.64.

	<25K	25-50K	50-100K	100-150K	>150K	Recall
<25K	391	643	129	56	8	0.32
25-50K	224	5519	320	324	48	0.86
50-100K	115	904	1192	180	18	0.49
100-150K	67	770	108	627	69	0.38
>150K	8	277	42	123	49	0.10
Precision	0.49	0.68	0.67	0.48	0.26	

Conclusion: Logistic Regression and SVM failed to estimate any values for first and last class (<25K and >150K). Random forest with Accuracy of 0.72 for 1000 estimator as the optimal number of estimators is slightly better than KNN and Adaboost. The precisions for the first 3 classes are above 70%.