Supervised Statistical Classification-Multinomial

Suraj Kumar(2601477K)

03/12/2021

Contents

1 Introduction

1.1 Background

Supervised multinomial classification is an essential aspect of Machine Learning. Sectors such as Business analysts, Data scientists, Financial analysts, and medicine are required to predict the data category based on feature instance vectors. The simple multinomial regression model is the most fundamental statistical method to serve the purpose; however, several advanced classification models such as randomForest, and Neural Network have evolved over the years, which have brought a revolution in prediction accuracy. randomForest has a very versatile application based on a decision tree hierarchy. On the other side, Neural Network performance has improved over a few decades as more and more data are available for training. Performance comparison is a crucial aspect of choosing the best statistical model for data. Overall classification accuracy and group-specific classification accuracy are the two most prominent used performance measures for supervised multinomial classification.

1.2 Aim of Analysis

The research project aims to evaluate the best multinomial classification models based on some performance measure criteria. Specifically, the main objectives are to design a study to access performance in classification, measure the performance of a simple multinomial regression model, and compare the performance to that of other statistical methods.

1.3 Workflow

The study goes through exploratory and formal data analysis of 4 datasets; Abalone, Car evaluation, Nursery school application, and Contraceptive method choice. All of the datasets have multi-class categorical output and a set of covariates. This report focuses on box plots, summaries, scatterplots, barplots, bar plots, and proportional tables to develop initial impressions about the data in section 3. While the study deepens on designing a performance measure, fitting multinomial regression, Random forest, and Neural Network to each dataset, and comparing the outcomes in section 4. 2nd section explains the background and motivation for the methods. Eventually, 5th section concludes the remark extracted from the overall analysis, and 6th section discusses the possible future extension in the work.

1.4 Data Descriptions

The aim of the analysis is specific for each dataset; predicting the age of abalone from physical measurements for Abalone dataset; predicting the car acceptability for given features of cars for Car dataset; predicting the contraceptive method choice of women based on several characteristics about women for Contraceptive method choice dataset; predicting success or failure of nursery class application provided socio-demographic information of the parents for Nursery dataset. The datasets have been sampled at different locations and sectors through a randomized method. The abalone dataset has been sampled from the Abalone population in Tasmania. I. Blacklip and Abalone found in North Coast and Islands of Bass Strait; the car evaluation dataset has been obtained from a simple hierarchical decision model previously developed for the demonstration of DEX; the contraceptive method choice dataset has been collected from a part of the 1987 National Indonesia Contraceptive Prevalence Survey; the nursery dataset has been obtained from the applications of nursery schools in Ljubljana, Slovenia.

2 Description of the Methods

The cornerstone of the formal analysis is based upon required statistical classification methods such as multinomial regression, randomForest, and Neural Networks. Additionally, principal components analysis and variable selection have been used as supplementary to the multinomial process.

2.1 Multinomial Regression

Nominal logistic regression models for more than two categories can be used as a classification tool for observation to a class based on the highest predicted probability among groups. The statistical method considers one of the groups as a baseline and fits logistic regression to the ratio of each group member to the baseline. Eventually, the probability of observation falling into each class is evaluated, and a class is assigned based on the highest value.

2.2 Variable Selection

Wrapper methods are the greedy search approach that stepwise removes variables based on an evaluation criterion. A bi-directional elimination method works similar to forwarding selection but does extra backward elimination at each iteration of adding a variable. The process reaches optimal features until no new feature can be added or removed. Due to its greedy nature, different wrapper methods can give different results. Additionally, they are prone to over-fitting and have high computation time.

2.3 Principal Component Analysis

PCA is a robust unsupervised machine learning algorithm for feature extraction. The method is useful when a significant correlation is present among continuous covariates. Moreover, it holds no assumption for the distribution of data. PCA considers both covariance and correlation matrix depending upon the evenness of the order of data variance. The algorithm is badly affected by outliers; therefore, it's necessary to remove outliers firstly. Eventually, discretionary components are retained based on the proportion of variance(proportion desired), Cattell's scree plot method, or Kaiser's method.

2.4 RandomForest

RandomForest, likewise bagging, bootstrap samples from the training data with replacement and average across all out-of-bag errors to give overall miss-classification error rate. Additionally, the former method attempts to decorrelate decision trees by randomly selecting a subset of features, which helps reduce the variance. randomForest beautifully handles missing data, mixed datasets, and multicollinearity.

2.5 Neural Network

The Neural Network has garnered the attention of the whole world in every sector as more and more data are available for training. The method is very effective in modeling the complex non-linear relationships among data features. The hyperparameter for the hidden layer needs to be carefully adjusted for every dataset to avoid overfitting and to model appropriately. Feedforward Neural Networks are most common in applications. It attempts to find a locally optimal solution based on initial reference using gradient descent and backpropagation.

3 Exploratory Analysis

3.1 Abalone

The dataset contains 4177 observations with no missing values, two categorical variables, sex and rings, and seven numerical variables length, diameter, height, whole weight, sucked weight, viscera weight, and shell weight.The number of rings is the response variable that gives the age in years on the addition of 1.5. The data is in the tidy format as each column corresponds to one feature of abalone, each row corresponds to a different abalone entity, and the characteristics of abalone are the single observational unit. Table 1 & 2 show the general information of the data set, including data type,number of factors, missing values, mean, quantile, and standard deviation.

Table 1: Summary statistics of Abalone: Factors

skim_variable	n_missing	factor.top_counts	factor.n_unique
sex	0	M: 1528, I: 1342, F: 1307	3
rings	0	9: 689, 10: 634, 8: 568, 11: 487	28

Table 2: Summary statistics of Abalone: Numerical

skim_variable	n_missing	numeric.mean	numeric.sd	numeric.p25	numeric.p50	numeric.p75	numeric.p100
length	0	0.5239921	0.1200929	0.4500	0.5450	0.615	0.8150
diameter	0	0.4078813	0.0992399	0.3500	0.4250	0.480	0.6500
height	0	0.1395164	0.0418271	0.1150	0.1400	0.165	1.1300
whole.weight	0	0.8287422	0.4903890	0.4415	0.7995	1.153	2.8255
shucked.weight	0	0.3593675	0.2219629	0.1860	0.3360	0.502	1.4880
viscera.weight	0	0.1805936	0.1096143	0.0935	0.1710	0.253	0.7600
shell.weight	0	0.2388309	0.1392027	0.1300	0.2340	0.329	1.0050

The response variable rings has 28 groups ranging from 1 to 29 except 28. All sex groups have a fairly similar number of observations, while the ring group 9 has the maximum number of observations as table 1 depicts. Moreover, the variances of continuous covariates differ in order notably. A barplot of the response variable can give deep insight into the frequency distribution. Figure 1 displays the barplot and sex-wise distribution of the ring.

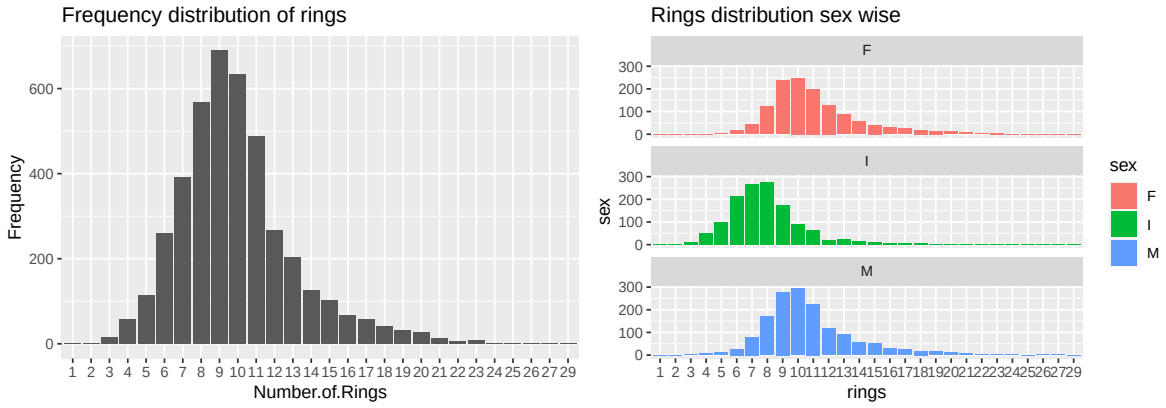

Figure 1: Distribution of rings

From the figure 1 barplot, the ring seems to have a right-skewed bell shape curve and a long tail extending in the right direction. The distribution is very sparse, mainly concentrated around 9, and very little data are available for rings below 5 and above 20. The uneven nature may adversely affect

the classification accuracy. Additionally, from the figure 1 sex-wise distribution, the ring has the same bell shape for all sex groups. The female sex group has slightly more rings than other groups. Before exploring the relationship of rings with continuous covariates, let's check for relation within continuous covariates themselves. A multicollinearity matrix can help better to identify any correlation.

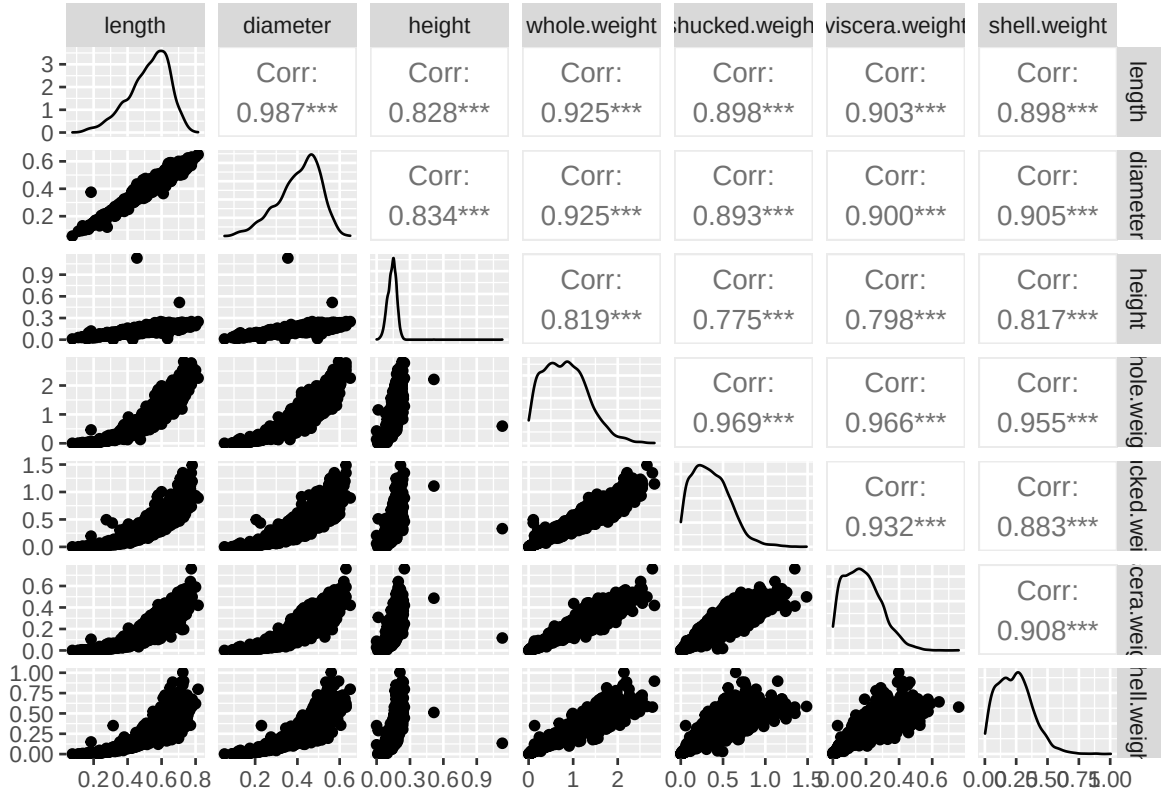

Figure 2: Multicollinearity check among continuous covariates

There exist a very high correlation among continuous covariates, so Principal Component Analysis can effectively reduce the dimension. Besides it, the squeezed scatterplots between length against height and height against whole-weight suggest the presence of two potential outliers. Since outliers can severely distort the performance of the majority of statistical methods, so they should be identified and removed. The figure 3 demonstrates the change in the scatterplot for length against height as outliers are removed.

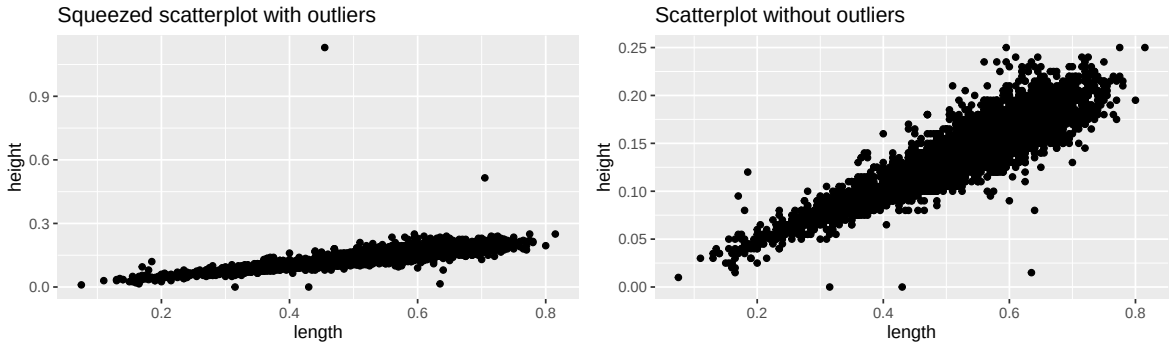

Figure 3: Outliers identification and removal

Boxplots in figure 4 of the ring against other continuous covariates are stuffy because of the large

number of the response groups in rings.

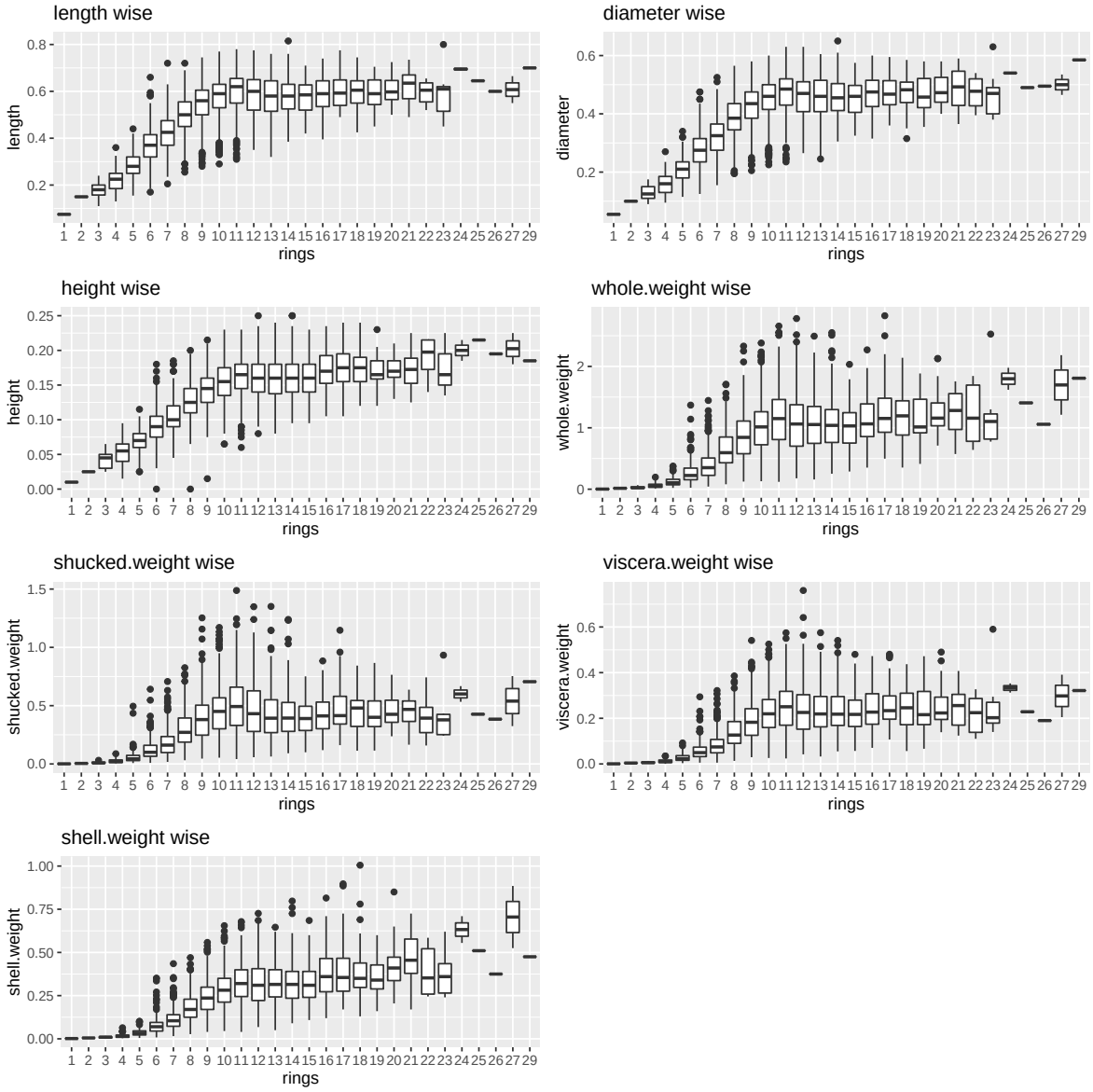

Figure 4: Rings vs continuous covariates: Ablone

The rings have a very uneven number of observations for every covariate; very few for lower and upper rings groups, many for middle ring groups. As the number of rings increases, variables such as length, diameter, height, and whole weight increase, whereas shucked weight, viscera weight, and shell weight remain the same approximately. There is evidence of many observations out of the Inter Quartile Range in any of the plots. The abalone dataset is now ready to jump into formal analysis.

3.2 Car Evaluation

The dataset consists of 1728 observations with no missing values and all categorical covariates such as buying, maint, doors, persons, lug_boot, safety. The car attributes cover three main components; luxury, comfort, and safety. The data is tidy as each column is a unique car feature, each row is an observation of a specific car, and the only observational unit is car characteristics. Table 3 shows the summary of the data set, including data type, number of factors, and missing values.

Table 3: Summary statistics of car

skim_type	skim_variable	n_missing	factor.top_counts	factor.n_unique
factor	Buying	0	hig: 432, low: 432, med: 432, vhi: 432	4
factor	maint	0	hig: 432, low: 432, med: 432, vhi: 432	4
factor	doors	0	2: 432, 3: 432, 4: 432, 5mo: 432	4
factor	persons	0	2: 576, 4: 576, mor: 576	3
factor	lug.boot	0	big: 576, med: 576, sma: 576	3
factor	safety	0	hig: 576, low: 576, med: 576	3
factor	class	0	una: 1210, acc: 384, goo: 69, vgo: 65	4

The response variable class has four groups; unacc(unacceptable), acc(fairly acceptable), good(good acceptable), vgood(very good acceptable). The class groups have uneven proportional data distribution as unacc accounts for 1210 observations, whereas vgood has only 65. Other categorical covariates have a fairly uniform group-wise distribution. Figure 5 barplot can give better visualization of the class response variable.

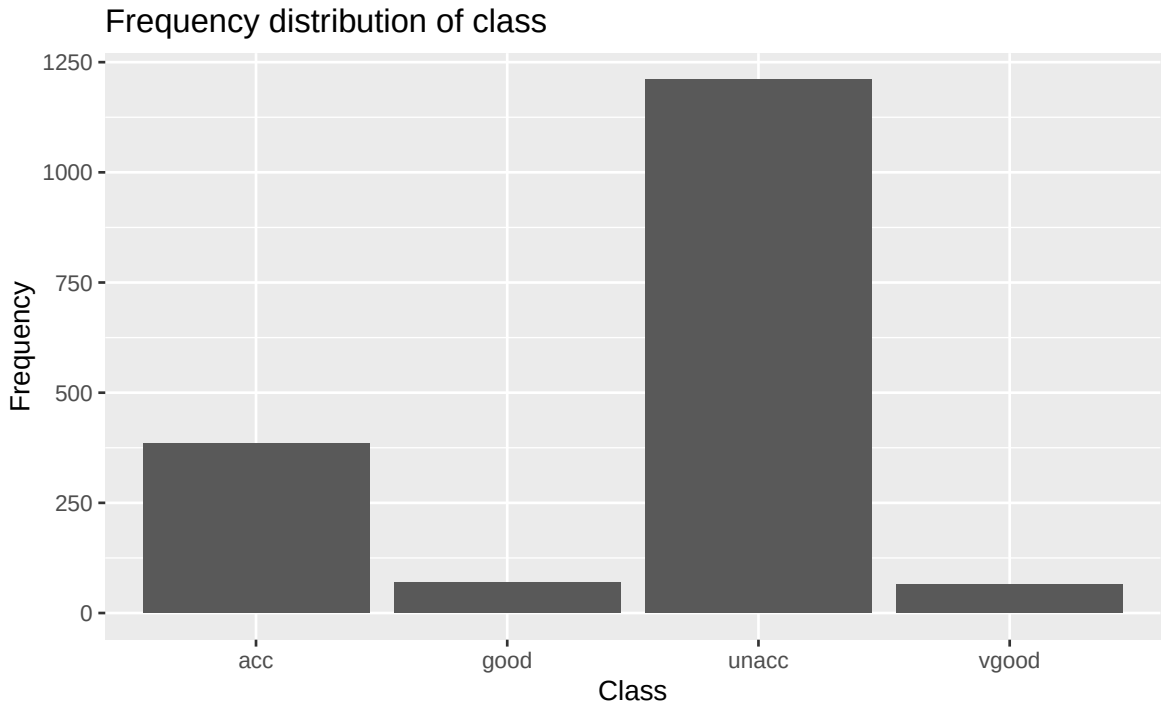

Figure 5: Distribution of class

The proportion of good quality cars are much lower than bad cars. As shown in Table 4, around 70% of total vehicles are unacceptable, while 3.8% are outstanding.

Now, it's right time to understand the relationship between the class and other categorical covariates. Figure 6 plots dodged barplots for class vs other variables.

Table 4: Propotion of groups in the class

class	n	percent
acc	0.2 (384)	22.2% (0.222222222222222)
good	0.0 (69)	4.0% (0.039930555555556)
unacc	0.7 (1210)	70.0% (0.700231481481482)
vgood	0.0 (65)	3.8% (0.0376157407407407)

Figure 6: Class vs covariates: Car

the above figure 6 shows that car acceptability decreases with high buying and maintenance prices, whereas increases with safety. Moreover, a lower space for luggage and lesser person accountability increase the car's unacceptability. People tend to reject low safety cars outright.Great cars have medium buying and maintenance costs, have above four doors, accommodate more than two persons, have plenty of luggage space, and have high safety measures. People outright discard low safety and two persons cars. In all, a customer looks for high safety cars with good comfort available at lower prices. There is no outlier in the case of car data; therefore, it is ready for formal analysis.

3.3 Contraceptive Method Choice

The dataset consists of 1473 observations with no missing values, seven categorical variables, and two continuous variables; Wife's age, Wife's education, Husband's education, Number of children ever born, Wife's religion, Wife's now working, Husband's occupation, Standard of living index, and Media exposure.All women were not pregnant during the time of the survey. The data attributes cover the demographic and socio-economic characteristics of women in Indonessia. The dataset is tidy as each column is a characteristic of women, each row is a unique observation, and the only observational unit is women characteristics. Table 5 & 6 show the general information of the data set, including data type, number of factors, and missing values.

Table 5: Summary statistics of contraceptive method choice:Factors

skim_variable	n_missing	factor.top_counts	factor.n_unique
w.education	0	4: 577, 3: 410, 2: 334, 1: 152	4
h.education	0	4: 899, 3: 352, 2: 178, 1: 44	4
religion	0	1: 1253, 0: 220	2
working	0	1: 1104, 0: 369	2
occupation	0	3: 585, 1: 436, 2: 425, 4: 27	4
standard	0	4: 684, 3: 431, 2: 229, 1: 129	4
media	0	0: 1364, 1: 109	2
contraceptive	0	1: 629, 3: 511, 2: 333	3

Table 6: Summary statistics of contraceptive method choice: Numerical

skim_variable	n_missing	numeric.mean	numeric.sd	numeric.p25	numeric.p50	numeric.p75	numeric.p100
age	0	32.538357	8.227245	26	32	39	49
children	0	3.261371	2.358549	1	3	4	16

Contraceptive is consists of three groups; 1=No-use, 2=Long-term use, 3=Short-term use. Most of the surveyed women population in Indonesia are educated, non-working, have good living standards, follow Islamism, and stay away from media. The median age is 32, and the median number of children is 3. Surprisingly, numeric.p100 is 16, whereas numeric.p75 is only 4. That suggests to do some further analysis through boxplots.

Table 7: Propotion of groups in the contraceptive

contraceptive	n	percent
1	0.4 (629)	42.7% (0.427019687712152)
2	0.2 (333)	22.6% (0.226069246435845)
3	0.3 (511)	34.7% (0.346911065852003)

From the table 7, group 1 has the highest proportion of observations around 43%, while group 2 has the lowest around 22.6%. That means majority of wives don't intend to use contraceptives due to adverse health issues or religious belief. Boxplots and barplots can help to deepen the impression further through figures 7 & 8.

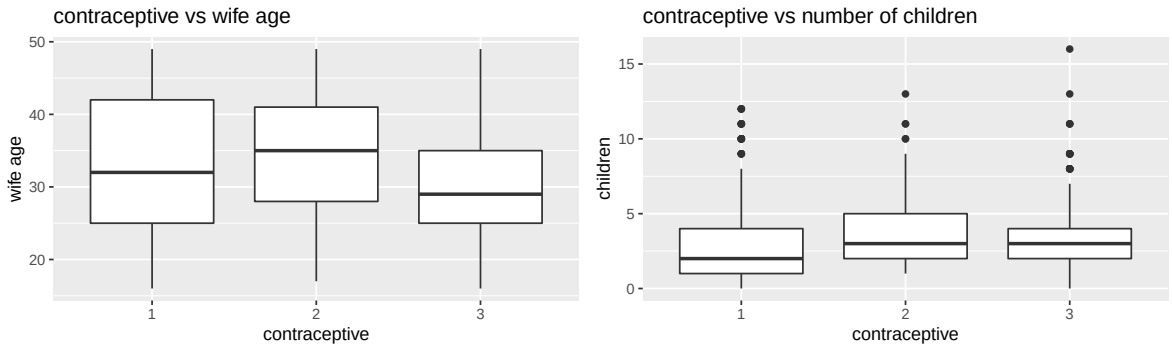

Figure 7: Boxplots of Contraceptive vs Covariates:Numerical

From figure 7, wives with short-term contraceptive usage have lower median age, while wives with long-term contraceptive usage have higher median age than any other. Each boxplot in wife age against contraceptive indicates no outliers, has uniform spread and slightly misses symmetry. Whereas, boxplots in children against contraceptives are unsymmetric, indicate the presence of outliers mostly of age above 12, and have unsymmetric spread. Outliers can distort the classification performance severely, so they are removed.

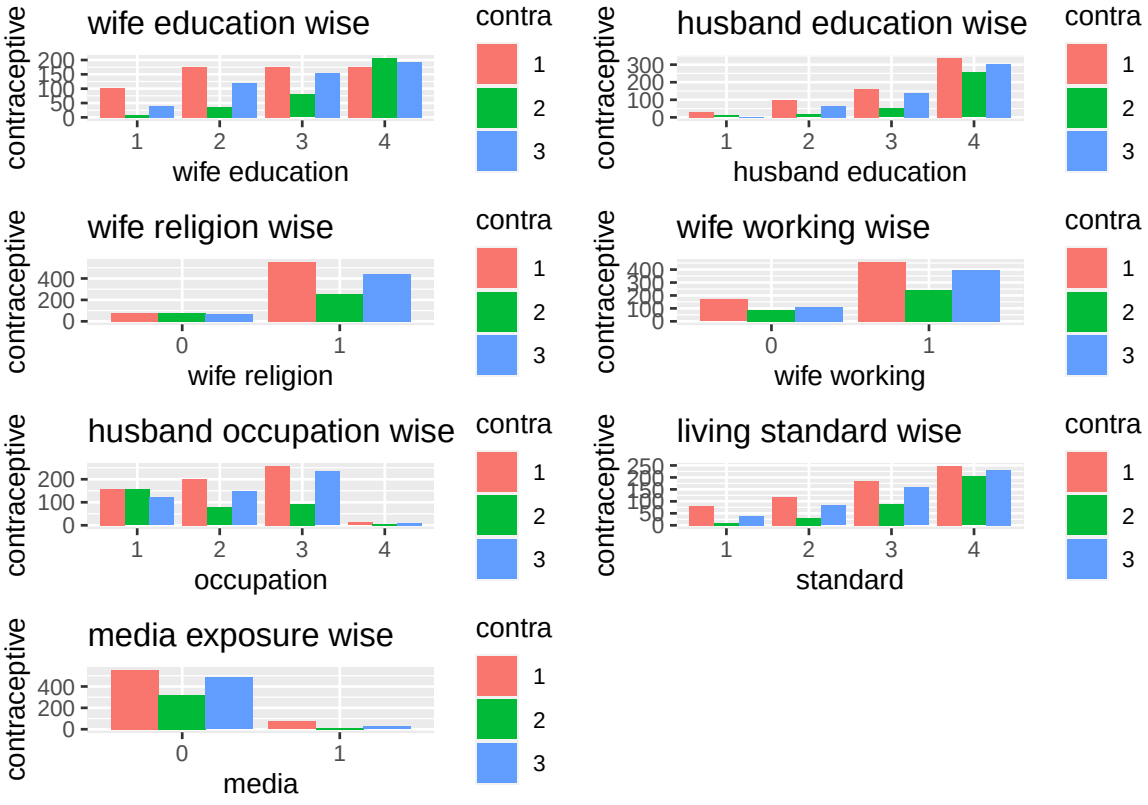

Figure 8: Plots of Contraceptive vs Covariates:Categorical

Uneducated wives have a higher proportion of no_use, while educated ones have a higher proportion of long_term_use. Likewise, wives, husbands have the same trend in the education arena. Working and not working wives have a very similar pattern regarding the usage of contraceptives. The majority of Islamic wives are reluctant to use contraceptives. No usage mindset is prevalent in all classes of living standards and media exposures. The contraceptive dataset is now ready for formal analysis.

3.4 Nursery

The dataset consists of 12960 observations with no missing values and all categorical covariates; parents, has_nurs, form, children, housing, finance, social, health. The attributes cover concept structure such as parents' employment, family financial structure, family structure, and social health conditions. The data is in a tidy format. Each column is a unique attribute, each row is a unique observation of a nursery application, and the application characteristics are the only observation unit. Table 8 shows the general information of the data set, including data type, number of factors, and missing values.

Table 8: Summary statistics of of Nursery Applications

skim_variable	n_missing	factor.top_counts	factor.n_unique
parents	0	gre: 4320, pre: 4320, usu: 4320	3
has_nurs	0	cri: 2592, imp: 2592, les: 2592, pro: 2592	5
form	0	com: 3240, com: 3240, fos: 3240, inc: 3240	4
children	0	1: 3240, 2: 3240, 3: 3240, mor: 3240	4
housing	0	con: 4320, cri: 4320, les: 4320	3
finance	0	con: 6480, inc: 6480	2
social	0	non: 4320, pro: 4320, sli: 4320	3
health	0	not: 4320, pri: 4320, rec: 4320	3
class	0	not: 4320, pri: 4266, spe: 4044, ver: 328	5

The class response variable has four categories; not:not_recomended, priority, special priority, recommended, and very recommended. The group recommended has very few data, so may need further diagnostics. Every other variable has a very even distribution. A proportional table can help to get a clear picture.

Table 9: Propotion of groups in the Class

class	n	percent
not_recom	0.3 (4320)	33.3% (0.333333333333333)
priority	0.3 (4266)	32.9% (0.329166666666667)
recommend	0.0 (2)	0.0% (0.000154320987654321)
spec_prior	0.3 (4044)	31.2% (0.312037037037037)
very_recom	0.0 (328)	2.5% (0.0253086419753086)

The table 9 confirms that there are just two observations for the recommendation group, while other groups have a significant amount of data. The prediction may not go well with minimal data, so these are removed. Only 25.3% of the applications were very recommended, while 33.33% were not recommended. Figures 9 & 10 display the class distribution and relationship with other covariates.

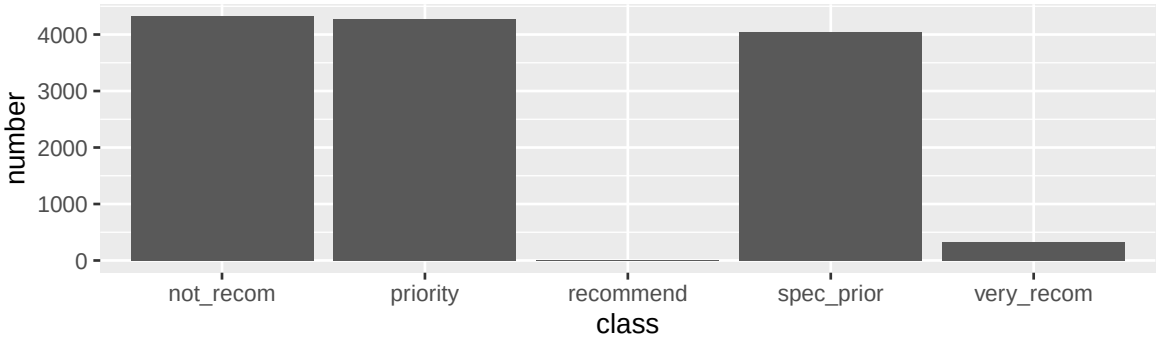

Figure 9: Nursery class barplot

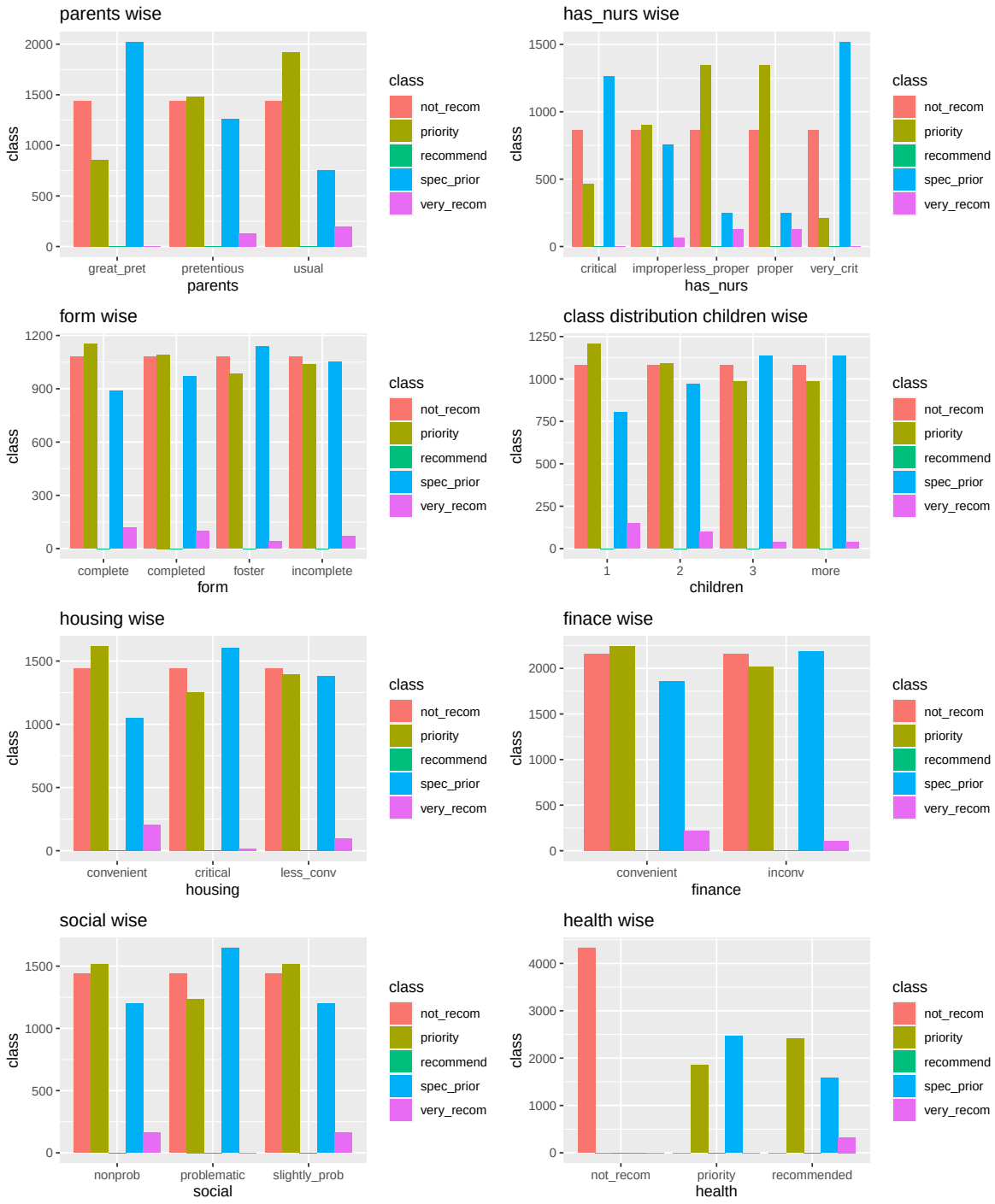

Figure 10: Nursery application class vs Covariates

The barplot in figure 9 shows that few applications were recommended due to the high volume of enrollment.Other groups such as not_recom, priority, and special priority have roughly the same number of observations. Additionally, figure 10 depicts that very good recommendation chance increases with usual parents, having a proper nursery at home, complete form of the family, lesser number of children, convenient home, convenient finance, and high health recommendation. Health has been a major cause of rejection as all applications were rejected with the least health priority. A good social image and convenient housing conditions can also boost the probability of recommendation.

4 Formal Analysis

Formal analysis for each dataset starts with creating standard and consistent performance measure criteria. Every four datasets are split into multiple training and test pairs. Then, All suitable statistical methods for a dataset are run on each pair. In the end, all training and test errors for each statistical method are averaged to give final classification errors. Eventually, all performances specific to a dataset are compared based on classification error rate. The lesser is classification error, the better is the performance of a method. Furthermore, each continuous covariate is scaled before feeding to Neural Network. In order to choose the best hyperparameter for each dataset, the training set is run on exhaustive set of hyperparamters from 1 to 5,5,5; the model having least classification error is choosen for further comparison with other statistical methods.

4.1 Abalone

Section 3.1 has discussed the preliminary analysis of the dataset. The formal analysis commences with Principal Component Analysis for dimension reduction of the dataset as whole. The numerical variables have a very high correlation, so PCA can decrease the model complexity while maintaining the performance. Outliers have already been removed. Firstly, we checked for variances of the numerical covariates.

Table 10: Variance of numerical covariates of abalone

length_v	diameter_v	height_v	whole.weight_v	shucked.weight_v	viscera.weight_v	shell.weight_v
0.0144202	0.0098467	0.0014814	0.2401263	0.0491569	0.0119976	0.0193661

The highest and lowest value of variance is 0.0098467 for diameter and 0.2401263 for whole_weight. Due to the high difference in the order of variance, a correlation matrix is more suitable for the PCA.

Table 11: Cumulative proportion of variance

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
sd	2.5432458	0.4634750	0.3731185	0.3074646	0.2528881	0.1127751	0.0816665
Prop_of_var	0.9240141	0.0306870	0.0198882	0.0135049	0.0091361	0.0018169	0.0009528
Cum_prop	0.9240141	0.9547012	0.9745894	0.9880943	0.9972303	0.9990472	1.0000000

As the table 11 states, the very first component itself captures 92.4% of the total variance. Combined component 1 and component 2 can cover more than 95% of the total variance. Henceforth, PCA has been successful in reducing the dimension from seven to two. Dataset is split into three training and test pair subsets in ratios; 50:50, 60:40, and 70:30.

Firstly, simple multinomial regression model is fitted on all the training sets, and the classification error rate is evaluated on the corresponding test sets.

Table 12: Multinomial regression output for different splits: Abalone

split	training.error	test.error
50:50	0.7019645	0.7332375
60:40	0.7077844	0.7311377
70:30	0.7111567	0.7174781
Overall	0.7069686	0.7272844

Table 12 presents the classification performance of multinomial regression on all the splits. The overall test error is 72.7284%, which is worst. However, the test error slightly improves with training set having more data. A confusion matrix can help to understand the class-specific errors.

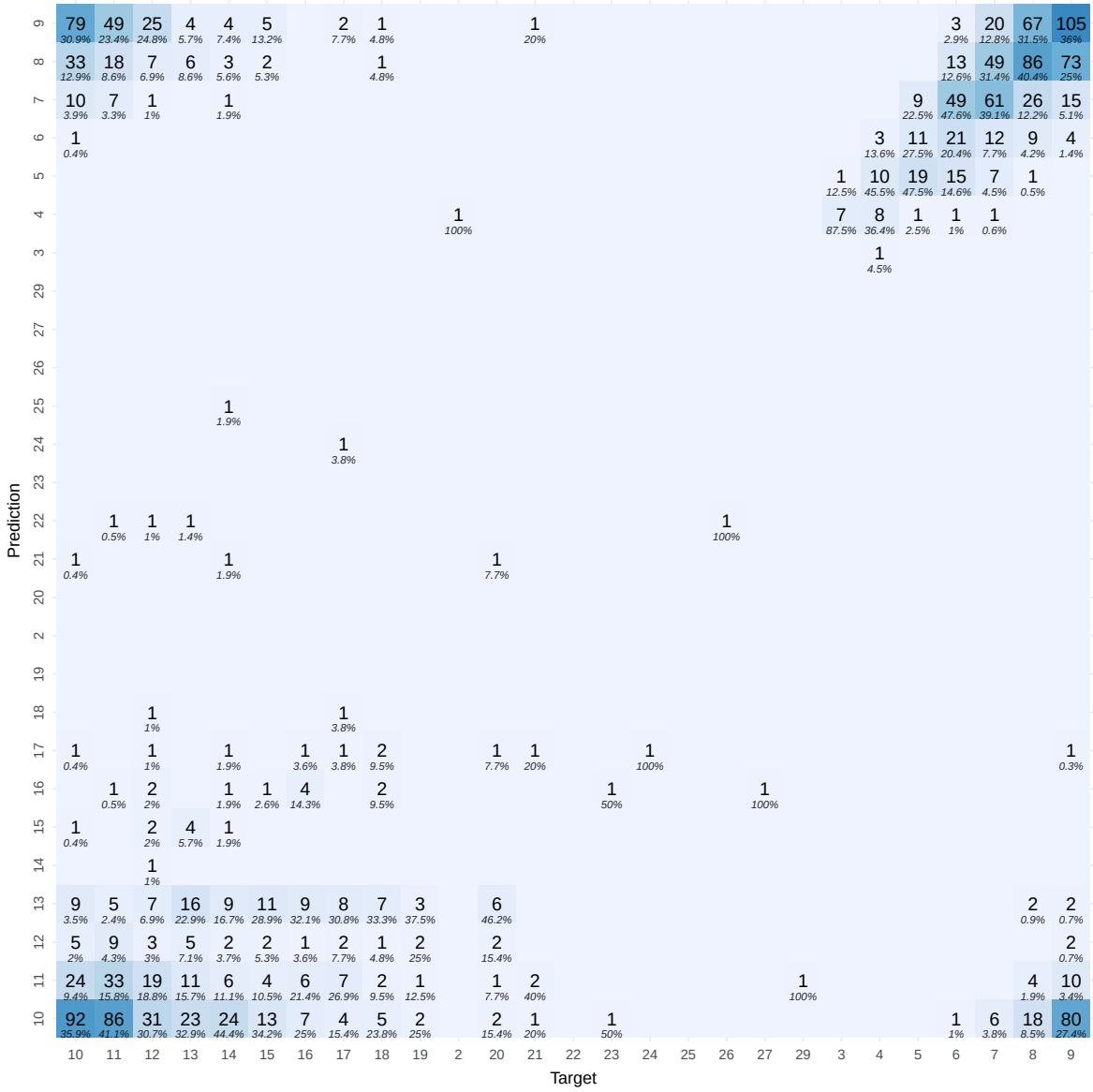

Figure 11: Confusion matrix for multinomial regression: Abalone

The confusion matrix in figure 11 clearly states that the prediction has missed the target very closely. The diagonal pattern in the density of blue color confirms the previous statement. Now we proceed with variable selection on all the three last multinomial models.

Table 13: Variable reduction output for different splits: Abalone

split	training.error	test.error
50:50	0.7163392	0.7241379
60:40	0.7085828	0.7305389
70:30	0.7145791	0.7190742
Overall	0.7131670	0.7245837

The bi-direction wrapper method leads to variable reduction to 4, 5,8 for the corresponding multinomial models first, second, and third, each having 9 input variables. Comparing table 13 with the table 12, a marginal improvement in the overall test error rate can be noticed. The advantage is of having lesser complexity of the reduced model for the same output.

Table 14: randomForest output for different splits: Ablone

split	training.error	test.error
50:50	0.7388596	0.7485632
60:40	0.7469062	0.7586826
70:30	0.7412731	0.7621708
Overall	0.7423463	0.7564722

RamdomForest has even worst performance than the simple multinomial as the overall training and test errors have jumped up by 2-3% in table 14. Now we go for the Neural Network. The numerical data is scaled beforehand.

Table 15: Neural Network output for different splits: Ablone

split	training.error	test.error
50:50	0.7982750	0.7404215
60:40	0.7936128	0.7359281
70:30	NA	NA
Overall	0.7959439	0.7381748

The Neural Network having hyperparameter 3 fails to calculate weight for 70:30 split; henceforth, the overall error is calculated with the rest two splits. In terms of performance, the method is not giving any improved results. The overall training error shifts up to 0.7959 because of sparse and inadequate data for good training.

Figure 12: Comparision of performances :Ablone

The overall classification error rates are pretty high for all the statistical methods for the Abalone dataset as shown in figure 12. The reasons could be attributed to very sparse data distribution and inadequate observations. Moreover, observations outside of IQR ranges of height and whole_weight covariates could have distorted the performance. Random forest and PCA have performed more poorly than simple multinomial. However, variable selection helped to reduce the number of inputs keeping the classification performance close to the multinomial regression.

4.2 Car Evaluation

From exploratory analysis about car evaluation dataset in section 3.2, we are ready to apply simple multinomial regression, variable reduction by a wrapper method, randomForest, and Neural Network.

Likewise abalone dataset, the car data is split into three training and test pairs in different proportions. We start with applying simple multinomial to all training sets:-

Table 16: Multinomial regression output for different splits: Car

split	training.error	test.error
50:50	0.0439815	0.0856481
60:40	0.0472973	0.0895954
70:30	0.0488007	0.0751445
Overall	0.0466931	0.0834627

The multinomial regression for the car data has given reasonably good performance with an overall training error of 4.6693% and a test error of 8.3463%. Moreover, the test error decreases significantly in split 70:30. Furthermore, we can see confusion matrix 13 for class specific errors.

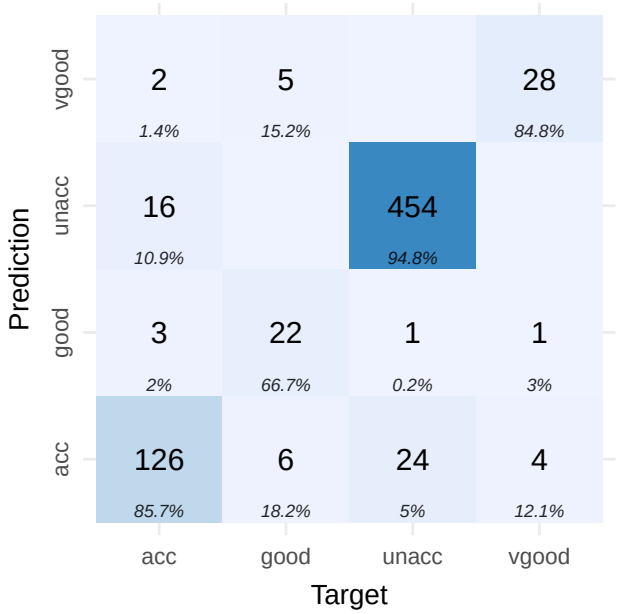

Figure 13: Confusion matrix for multinomial regression: Car

The group unacc has the lowest classification error, while good has the highest. Groups good and vgood have fewer observations than unacc and acc, which cause higher classification errors.

Table 17: Variable selection output for different splits: Car

split	training.error	test.error
50:50	0.0439815	0.0856481
60:40	0.0472973	0.0895954
70:30	0.0488007	0.0751445
Overall	0.0466931	0.0834627

The wrapper method doesn't reduce complexity of initial models as all input variables remain intact for all cases. The classification performance is the same as in the previous table 16.

From table 18, randomForest performance is good as the comprehensive test and training errors are 4.9666% and 4.5727% respectively. Moreover, the error rate declines with an increase in the training set size; both test and training errors are lowest for the 70:30 split. The performance is also better than the previous two models.

Table 18: randomForest output for different split: Car

split	training.error	test.error
50:50	0.0497685	0.0497685
60:40	0.0444015	0.0606936
70:30	0.0430108	0.0385356
Overall	0.0457269	0.0496659

Table 19: Neural Network output for different splits: Car

split	training.error	test.error
50:50	0	0.0231481
60:40	0	0.0072254
70:30	0	0.0346821
Overall	0	0.0216852

The Neural Network with hyperparamter (3,5,3) has given extraordinary performance for the car dataset with an overall 0% training error and 2.1685% test error. The performance is much better than all previous methods. The test error has improved by around 6%, which is a significant achievement. The figure 14 presents all the performances method wise in a barplot.

Figure 14: Comparision of performances :Car

As displayed in figure 14, every statistical method has performed well. However, advanced statistical methods have even better performance than simple multinomial regression. A Neural Network emerges to be the clear winner. The excellent performance attribute goes to a reasonably uniform distribution of categorical covariates and absence of numerical covariate. Multinomial performance lagged because of two groups good and vgood which were fewer in numbers. Randomforest has around 3% more classification accuracy than multinomial.

4.3 Contraceptive Method Choice

In section 3.3, we had performed the exploratory analysis of contraceptive method choice, removed the outliers of children having age greater than 12, and analyzed boxplots, barplots. From here, we delve deeper by doing formal analysis. Likewise, for abalone and car, the steps are the same. So, let's start with the performance checking of multinomial regression.

Table 20 presents the performance report of multinomial regression. The method is performing slightly better on the 50:50 split. The classification error rate is very high as the overall test error rate crosses 50%. A confusion matrix in figure 15 can help to understand the core of the problem.

Table 20: Multinomial regression output for different splits: Contraceptive

split	training.error	test.error
50:50	0.4258503	0.5074830
60:40	0.4444444	0.5119048
70:30	0.4450923	0.5238095
Overall	0.4384624	0.5143991

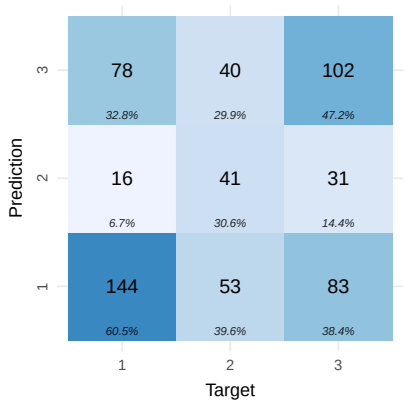

Figure 15: Confusion matrix for multinomial regression: Contra

Group 1 has the highest true classification rate of 60.5%, whereas group 2 has the lowest, around 30.6%. 39.6% of the predictions of group 2 have been wrongly done to group 1. Moreover, 38.4% of group 3 predictions have been wrongly attributed to group 1. Conclusively, the more is data , the better is prediction.

Table 21: Variable reduction output for different splits: Contraceptive

split	training.error	test.error
50:50	0.4285714	0.5061224
60:40	0.4489796	0.5153061
70:30	0.4392614	0.5124717
Overall	0.4389375	0.5113001

Applying bi-directional elimination methods leads to 16, 11, and 12 to the respective multinomial models with split 50,60, and 70. The original model has 19 variables, so the wrapper methods have reduced the complexity, but the overall test and training errors remain the close to the former ones.

Table 22: randomForest output for different split

split	training.error	test.error
50:50	0.4435374	0.4829932
60:40	0.4410431	0.4914966
70:30	0.4606414	0.4965986
Overall	0.4484073	0.4903628

RandomForest's outcomes in table 22 are slightly better than the former two methods. The 50:50 split seems to be more suitable for the technique with the lowest test rates 48.2993%. The overall test and training classification error rates have marginally gone below 50.

The training errors in table 23 have reduced significantly, whereas test errors are still high. The big gap between training and test errors for all splits could be due to over-fitting. However, hyperparameters

Table 23: Neural Network output for different splits: Contraceptive

split	training.error	test.error
50:50	0.2680272	0.5564626
60:40	0.3174603	0.5051020
70:30	0.3168124	0.5079365
Overall	0.3007667	0.5231670

other than 6 for the Neural Network have given even the worst results. Now we compare the overall performances of all statistical methods.

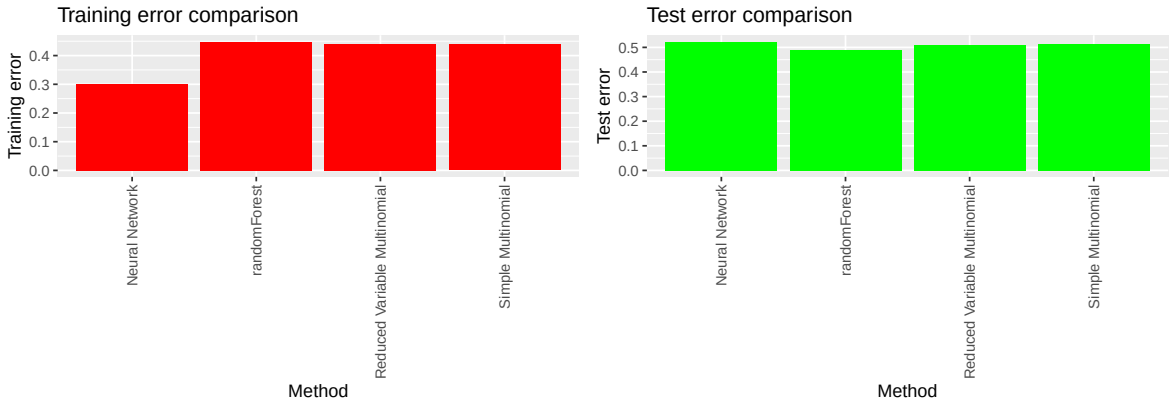

Figure 16: Comparision of performances :Contraceptive

All statistical methods have failed to give a good result. Among these, randomForest has slightly better test performance than any other method. The Neural Network has the lowest training error and seems to be overfitted. The variable reduction has successfully reduced the multinomial model complexity while producing the same classification error rate. The poor result could be due to the presence of numerical covariates.

4.4 Nursery

Section 3.4 has already discussed the initial exploration of the dataset, removed recommended group from the response variable, and understood the relationship of the class variable with other covariates. Likewise, the previous procedure, we split the data into three subset pairs in different proportions and fit a multinomial regression.

Table 24: Multinomial regression output for different split

split	training.error	test.error
50:50	0.0760920	0.0711530
60:40	0.0747363	0.0682870
70:30	0.0751929	0.0704733
Overall	0.0753404	0.0699711

The multinomial regression on the nursery dataset gives a fairly good performance with an overall training error 7.534% and test error 6.9971%. The error rate is consistent across all splits; however, the 60:40 break gives the minimum classification errors. We can analyze the confusion matrix in figure 17

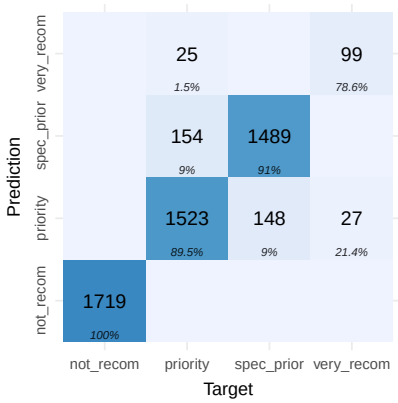

Figure 17: Confusion matrix for multinomial regression: Nursery

There is an almost 100% true classification rate for the not_recom group. Conversely, the group very_recom has suffered the most, with the lowest true classification rate of 78.6%. The performance difference could be due to data availability, as the very_recom group has much lesser data than not_recomm. Most of the misprediction for the very_recom group has been done as a priority.

Table 25: Variable selection output for different split

split	training.error	test.error
50:50	0.0760920	0.0711530
60:40	0.0747363	0.0682870
70:30	0.0751929	0.0704733
Overall	0.0753404	0.0699711

Variable reduction retains the same number of variables as in the former multinomial model for all the splits. Henceforth, the performance outcomes are also the same as original multinomial models. Now, we go for randomForest.

Table 26: RandomForest output for different split

split	training.error	test.error
50:50	0.0325668	0.0288625
60:40	0.0284281	0.0225694
70:30	0.0273429	0.0244342
Overall	0.0294459	0.0252887

RandomForest has performed much better than the multinomial regression as the overall test error is 0.0252887, which is around 4% less than the later one. Also, the 50:50 split gives the lowest training and test errors.

Table 27: Neural Network output for different splits: Nursery

split	training.error	test.error
50:50	0	0.0018521
60:40	0	0.0017361
70:30	0	0.0000000
Overall	0	0.0011961

Table 27 depicts the extraordinary performances of the Neural Network for the nursery dataset. The overall training and test errors have reduced to 0 and 0.0012, which are way less than the former

models. The performance improves as more and more data are available for training, and the best comes out for the 70:30 split with 100% true classification performances. Now, compare the overall performances of all statistical models.

Figure 18: Comparision of performances : Nursery

Performance comparison in figure **??** shows that the Neural Network is the foremost runner, followed by randomForest. The variable reduction has not helped either way. The minimum overall test error is around 0.1%. The multinomial regression performance is fair, but other advanced statistical models have overshadowed it. The attribute for good performances is the absence of numerical variables, outliers, and sufficient observations for groups.

5 Conclusion

A performance measure criteria can be designed by splitting data into multiple training and test of different proportions. Each statistical method is trained on every training set for each dataset, and classification error is evaluated on the corresponding test set. Eventually, all the test errors for a statistical method are averaged to give a final classification error. In the end, all classification errors of respective statistical methods are compared to evaluate the best one for the dataset. The research project used the same performance criteria to calculate multinomial regression performance.

The overall classification error rate can be used to compare the performances of multinomial regression with advanced statistical methods such as randomForest and Neural Network. The specific class error can help to evaluate group-wise performance. As more data is available for the group, the classification accuracy is better, as shown in the confusion matrix of different datasets.

The presence of numerical covariates and sparse data can severely distort the classification performance. The dataset abalone and contra suffered because of the same reasons.

The neural network can give high accuracy in classification for only categorical features datasets. The method has more than 99% classification accuracy for car and nursery datasets. Furthermore, the performance improves as more and more data are available for training. For instance, the classification accuracy was slightly better for the nursery dataset than the car dataset.

RandomForest performs better than the multinomial regression but poorer than the Neural Network in the case of categorical feature dataset. Besides it, both PCA and variable selection can help reduce dimensionality, keeping the performance the same as multinomial regression, as shown in the abalone and contraceptive dataset.

6 Future Work

1. Use of LASSO, Chi-square, Mutual information of variable selection instead of greedy wrapper methods.
2. Introduce more hyper-parameters in Neural Network beyond 3 hidden layers.
3. Performance comparison on datasets having size greater than 50,000.

7 References