# On the Statistical Comparison of Feature Selection Methods and the Role of Experts. The case of Las Vegas strip

Nestor R. Barraza[1⋆] and Antonio A. Moreno[2]

[1] Universidad Nacional de Tres de Febrero
[2] Universidad Nacional de Moreno
nbarraza@untref.edu.ar

**Abstract.** A statistical comparison of feature selection methods is performed. Feature selection is an important issue in Data Mining and Data Science, and a comparison of the results obtained from different methods is hard to be performed. Then, the evaluation of metrics and ways of comparisons is an important matter of study. Our study is performed on a real dataset previously analyzed in the literature containing a small number of records, drawing the attention on the conclusions to be applied where poor statistical confidence levels of significance can be obtained because of a relative low number of samples are present. The use of inter rater agreement coefficients is introduced as a novel approach extending a previous study. Boruta and tree-based methodologies perform rather well even in small data as it is shown. Our metrics can be used to guide the expert opinion in order to take the final decision. This work extends the results obtained in a previous analysis performed on the mentioned dataset.

**Keywords:** Big Data · Feature selection · Wrapper · Filtered · Lasso · Expert role

## 1 Introduction

Feature selection is a quite important issue in Data Science for the sake of dimensionality reduction in order to improve the machine learning algorithms on one hand, and of correlation between attributes discovering on the other. Despite several methods has been developed as it will be explained below, there is a lack of metrics in order to help what method to choose in a particular problem. Due to the diversity of problems and datasets, a general list of metrics for comparison is also hard to be found. There have been some attempts in the literature in order to find metrics for comparison without a conclusive result, see for example [2]. Despite of that, there are important conclusions obtained from feature selection studies, like those presented in [8]. The last issue is quite important to be highlighted, since discovering the main subset of features and their correlations raises

⋆ N. Barraza is also with Facultad de Ingeniería, UBA. N. Barraza and A. Moreno are also with TicData *www.ticdata.com.ar*

to improve the knowledge domain. We specially want to draw attention to that, since feature selection is usually used for dimensionality reduction, as discussed in [8]. The knowledge domain is a quite important issue in Data Science, the opinion of experts is pretty relevant as to guide the algorithms and to help the interpretation of results, as discussed in [7] and [6]. Then, finding some metrics that helps to compare and evaluate the performance of different methods appears to be successful as to help the experts to achieve conclusions. The research in this area is mainly performed on synthetic and controlled datasets and little has been performed on real datasets where managers have to take important decisions. Our study is focused then on a real dataset which has been analyzed in [7] where the authors arrive to an important conclusion. Despite our study supports the conclusions achieved there, the detailed analysis presented in this work allows to think on other possibilities, leaving the final decision to experts, though with important information they can be taken into account. Despite the predictive validity, which is usually used to measure the performance of a subset of features, we introduce the analysis through inter rater agreement coefficients to compare the order of precedence of the features selected by each method. Those coefficients were also introduced in [2] though on a synthetic dataset. The necessary amount of data in order to achieve a good level of confidence is also a relevant issue in feature selection. Since the dataset analyzed here is composed of 500 records, our results show that a conclusive analysis of comparison and performance can be made even in a small data problem. Since the output variable is a numeric integer, the inter rater agreement coefficients are used both, to measure the performance on one hand and to compare the order of precedence of features on the other. The aim of this work is to draw the attention that despite there is not a unique feature selection method that performs the best, there exists however some important tools useful in order to guide the expert opinion. Despite we are not able to extrapolate our results to other cases, the analysis may help to similar studies, and this work can be considered as an extension of the analysis presented in [7]. All of our calculations have been performed in R language, the specific functions used in this work are listed in the following sections.

This work is organized as follows: An introduction to feature selection methods is presented in section 2, the dataset we use is detailed in section 3, the methods analyzed in this work are listed in section 4, the experiments performed and the results are shown in section 5, the inter rater agreement coefficients are explained in section 6, results obtained from cross validation are presented in 7, a statistical comparison of the methods analyzed are presented in section 8, a discussion on the results obtained are developed in section 9, finally, conclusions are presented in section 10.

## 2 Feature selection

Feature selection methods attempt to select features that carry most of the information of the target variable, and the features in the selected set should be

independent the more possible in order to avoid redundancy of information in the given set. The metrics used in order to measure the connection between each feature and the target determines the method. Feature selection algorithms can be grouped in three different categories:

### 2.1 Filter Methods

Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithm. Instead the features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Some common filter methods are Correlation metrics (Pearson, Spearman, Distance), Chi-Squared test, Anova, Fisher's Score etc.

### 2.2 Wrapper Methods

In wrapper methods, a model is trained using a given subset of features. Based on the inferences drawn from the previous model, features are added or removed from the subset. Forward Selection and Backward elimination are some of the examples for wrapper methods.

### 2.3 Embedded Methods

These are the algorithms that have their own built-in feature selection methods. LASSO regression is one such example.

## 3 Dataset

We use the Las Vegas dataset previously analyzed in [8]. The analysis on this dataset appears to be quite interesting since it has just 504 records, what can be considered as a small data problem. The output variable is the *Score* assigned to hotels. Data were collected from the Authors of [8] from *TripAdvisor.com*. Features are listed in table 1, see [8] for details.

In the mentioned reference, using a DSA (Data sensitivity analysis) with a Support Vector Machine learning procedure, what is itself a wrapper method, the authors conclude that qualifications in TripAdvisor obtained for the hotels are determining to choose the hotels by the customers due to the reviewing variables are two of the most relevant.

The order of importance of features selected according to [8] is shown in Fig. 1.

## 4 Methods Analyzed

### 4.1 Filtered

**ji-square test of independence** The ji-square test of independence can be used as a first approach in order to evaluate feature correlation with the output variable.

**Table 1.** List of features

| Feature name | Data type | Description |
|---|---|---|
| Username | Categorical | Username as registered in TripAdvisor |
| User country | Categorical | User's nationality |
| Nr. reviews | Numerical | Number of reviews |
| Nr. hotel reviews | Numerical | Total hotel reviews |
| Helpful votes | Numerical | Helpful votes regarding review's info |
| Score | Numerical | Review score 1 2 3 4 5 |
| Review date | Date | Date when the review was written |
| Review text | Text | Textual content of the review |
| Review language | Categorical | Language of the review |
| Period of stay | Categorical | Period of stay: Dec–Feb Mar–May Jun–Aug Sep–Nov |
| Traveler type | Categorical | Business couples families friends solo |
| Member registered year | Date (year) | Year the user has registered in TripAdvisor |
| Pool | Categorical | If the hotel has outside pool |
| Gym | Categorical | If the hotel has gym |
| Tennis court | Categorical | If the hotel has tennis court |
| Spa | Categorical | If the hotel has spa |
| Casino | Categorical | If the hotel has a casino inside |
| Free internet | Categorical | If the hotel provides free internet |
| Hotel name | Categorical | Hotel's name |
| Hotel stars | Categorical | Hotel's number of stars |
| Nr. rooms | Numerical | Hotel's number of rooms |
| User continent | Categorical | Continent where the user's country is located |
| Member years | Numerical | Number of years the user is member of TripAdvisor |
| Review month | Categorical | Month when the review was written (from review date) |
| Review weekday | Categorical | Day of the week the review was written (from review date) |

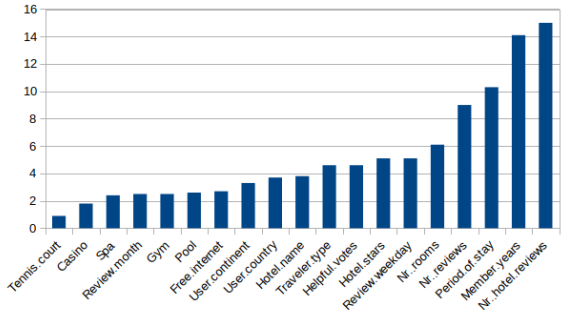



**Fig. 1.** Features selected in [8]

**OneR** The One rule feature selection method is one of the simplest criterion. It assigns the most frequent class of the output variable for each value of the predictor and order the predictors according to the root mean square error (RMSE). Despite there is a classification technique involved in this method, since a single performance metric is considered for each attribute separately, we can straightly compare this method with the *ji-square* test of independence. Then, we consider this method as a filtered one.

**Near zero variance** This simple criterion is implemented in order to eliminate constant and almost constant predictors across samples. It is based in a logical common sense that a near zero variance predictor has poor discriminant power.

### 4.2 Information Gain

The metric used by this method in order to measure the connection between a predictor and the target variable is the Mutual Information, an extensive explanation and application of this method can be found in [1].

### 4.3 Wrapper Methods

Wrapper methods considered here are based on decision trees and their variants, see for example [3] and [5].

**ctree** This is a tree based method that uses the entropy as a measure of impurity.

**CART** A tree method based on the Gini impurity.

**Random Forest** This is also a tree based algorithm that extends a previous concept of bagging of trees by randomly selecting a subset of features (feature bagging). Features highly correlated with the output variable will be selected in many of the B trees.

**Boruta** The Boruta algorithm is based on random forest. It adds another order of randomness by creating shuffled copies of all features (shadow features), then it chooses features having more importance than the best of the shadow features.

### 4.4 Embedded Methods

**Lasso** The *lasso* ((least absolute shrinkage and selection operator) method is a penalized version of the least sum of squares method, it adds the penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ to the RSS. This term allows the coefficients $\beta_j$ to become zero, selecting this way a given subset of features.

**Table 2.** R functions

| Method | *function* |
|---|---|
| ji-square | *chisq.test* |
| OneR | *OneR* |
| Near zero Var. | *nearZeroVar* |
| Information Gain | *information.gain* |
| ctree | *train(..., method="ctree",...)* |
| CART | *rpart* |
| Random Forest | *random.forest.importance* |
| Boruta | *Boruta* |
| lasso | *model.matrix* |
| | *data.matrix* |
| | *cv.glmnet* |

### 4.5 R functions

We list in table 2 the R functions we have used to implement each of the mentioned feature selection methods.

The implementation of *lasso* deserves an explanation. Since the dataset is composed by a few number of records, *lasso* gives an all zero result because a poor level of confidence. Then, we apply the *model.matrix* function on the factors attributes in order to divide each factor according to group of values, as shown also in Fig. 4 below.

## 5 Experiments

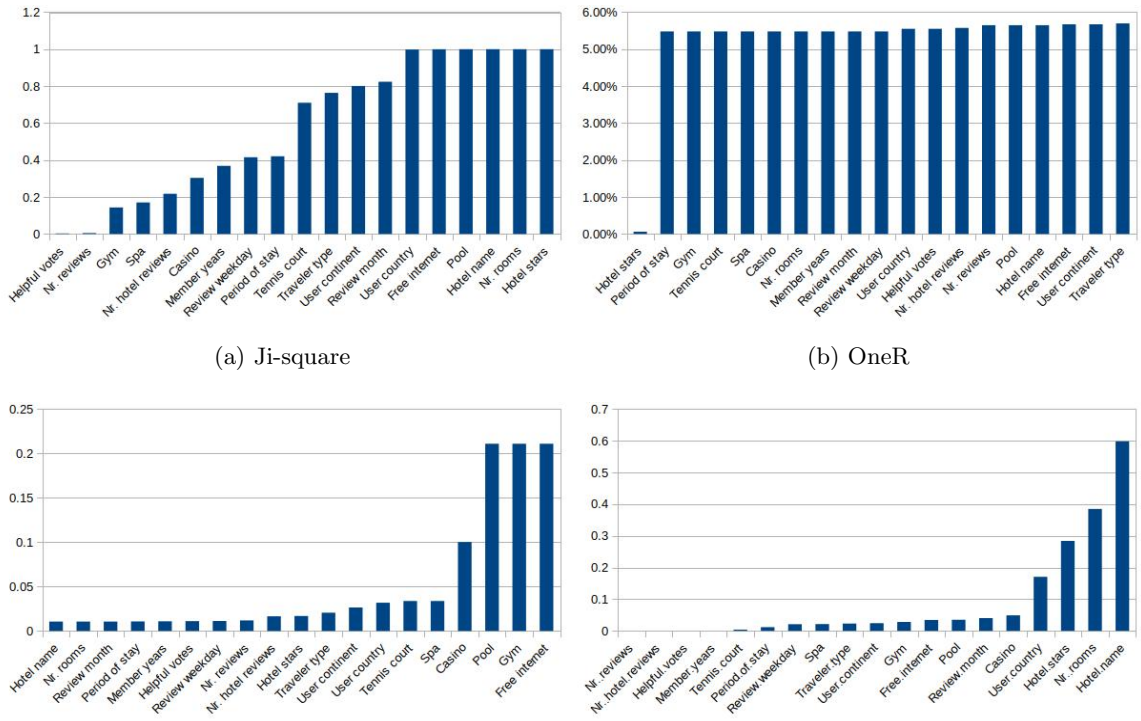
We show in this section all the results obtained by implementing the mentioned methods, all the calculations were performed in the R language. The figures show the relative order of importance of the features. Filter methods are show in Fig. 2, wrapper methods are shown in Fig. 5 and lasso coefficients are shown in Fig. 4.

## 6 Inter rater agreement coefficients

Inter rater agreement coefficients are intended to evaluate the agreement between rankers who assign subjects to categories. Then, they can be used in order to compare the ordered subset of features selected from each method. We explain below the coefficients used in this work.

### 6.1 Cohen-Kappa

Cohen's kappa coefficient compares the observed probability of disagreement of two raters to the probability of disagreement expected by chance. Let $p_{i,j}$ be the



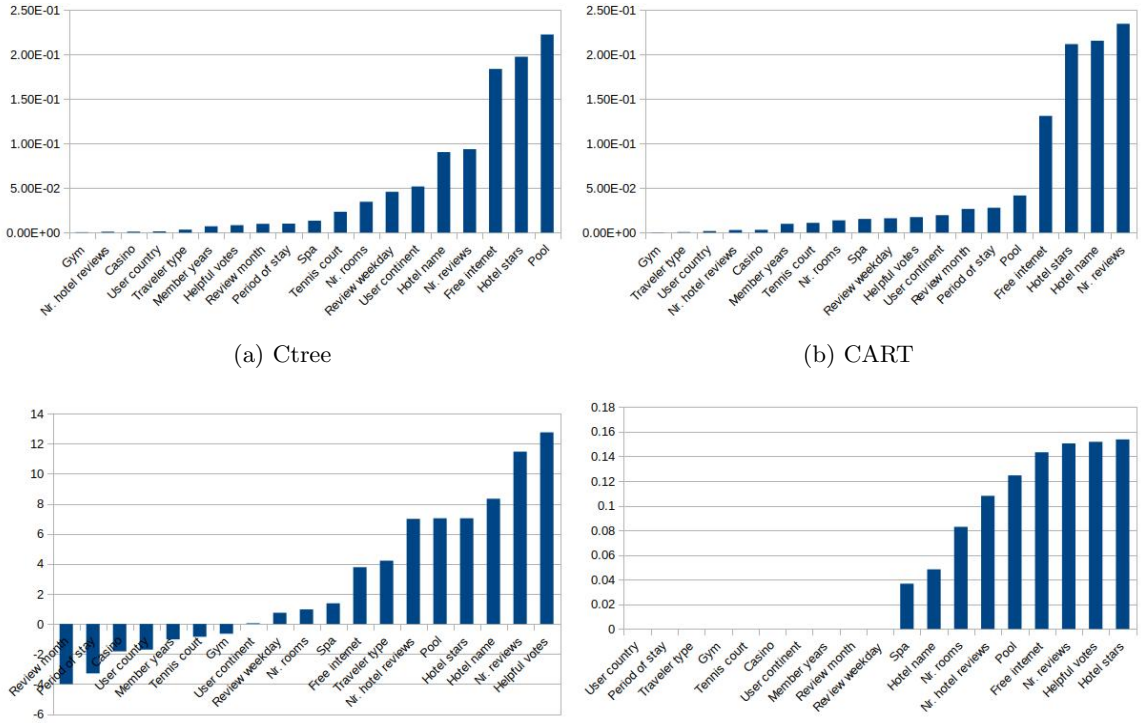

(a) Ji-square

(b) OneR

(c) Near Zero Variance

(d) Information Gain

**Fig. 2.** Filter methods.




(a) Ctree (b) CART

(c) Random forest (d) Boruta

**Fig. 3.** Wrapper methods.

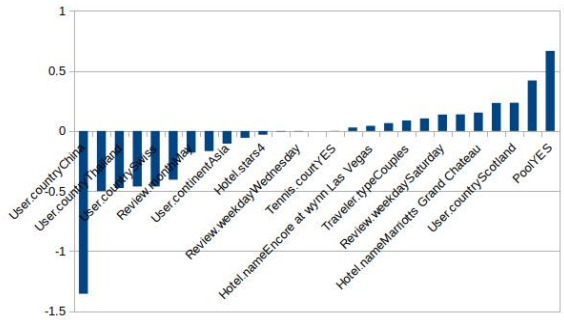



**Fig. 4.** Lasso coefficients

proportion of subjects that were assigned to the $i$th category by the first rater and to the $j$th category by the second rater, and $p_i = \sum_{j=1}^{m} p_{i,j}$, $p_j = \sum_{i=1}^{m} p_{i,j}$ the corresponding marginals. The weighted Cohen-Kappa statistics is defined as:

$$\hat{\kappa}_w = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}p_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}p_i\, p_j}{1 - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}p_{ij}} \tag{1}$$

The unweighted kappa is obtained as a special case of $\hat{\kappa}_w$ with $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ for $i \neq j$. In case the $m$ categories form an ordinal scale with numerical values $1, 2, \cdots, m$, weights can be set by: $w_{ij} = 1 - (i-j)^2/(m-1)^2$, and $\hat{\kappa}_w$ can be interpreted as an interclass correlation coefficient.

### 6.2 Fleiss-Kappa

Let $N$ be the total number of subjects, $n$ the number of ratings per subject and $k$ the number of categories. Let $n_{ij}$ represents the number of raters who assigned the $i$th subject to the $j$th category. The proportions of assignments to the $j$ category is given by:

$$p_j = \frac{1}{N\, m}\sum_{i=1}^{n} n_{ij} \tag{2}$$

Let $P_i$ the proportion of agreement for the $i$th subject computed as the proportion of the rater-rater pairs in agreement:

$$P_i = \frac{1}{n\,(n-1)}\sum_{j=1}^{k} n_{ij}\,(n_{ij}-1) \tag{3}$$

and the mean:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{4}$$

The probability of coincidence by chance is computed as the proportion (2):

$$P(coincidence|j) = p_j \tag{5}$$

then, the total probability of coincidence by chance results:

$$\bar{P_e} = \sum_{j=1}^{k} p_j^2 \tag{6}$$

The Fleiss-Kappa coefficient is then defined as:

$$\kappa = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}} \tag{7}$$

### 6.3 Kendall's $\tau_b$

The Kendall's tau_b coefficient for two sets of ordered pairs with ties is given by:

$$\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_x)\,(N_c + N_d + T_y)}} \tag{8}$$

where $N_c$ and $N_d$ account for concordant and disconcordant pairs, $T_x$ denotes the number of pairs tied for the first response variable only and $T_y$ denotes the number of pairs tied for the second variable only.

**Table 3.** R functions

| Coefficient | *function* |
|---|---|
| Cohen-Kappa | *cohen.kappa* |
| Fleiss-Kappa | *kappam.fleiss* |
| Kendall's $\tau_b$ | *tau_b* |

The interest in using this coefficient comes from the fact that it takes into account the order in the sets of pairs. It is specially useful in order to compare the order of precedence assigned to the features by each method.

Functions implementing the mentioned coefficients are listed in table 3.

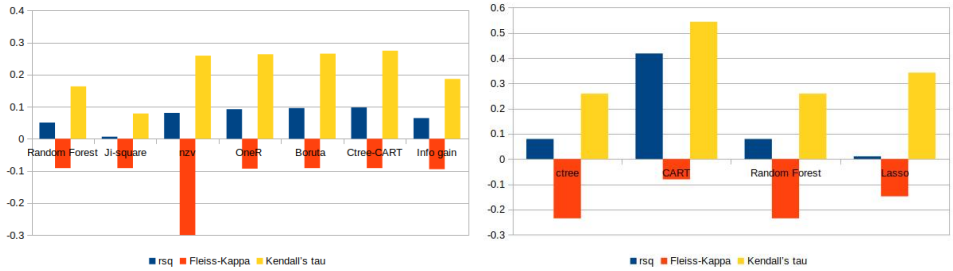
## 7 Cross-validation

Wrapper and embedded methods allow to measure the efficiency of the selected features through a cross-validation process. The same algorithm used to remove irrelevant features is evaluated by means of its predictive validity. Since the output variable *Score* is a numerical integer, these coefficients are not just used to compare the subsets of selected features but also to evaluate the performance.The accuracy is measured using different metrics: the R-square coefficient of determination, $rsq = 1 - \frac{rss}{tss}$, where $rss$ is the residual sum of squares and tss is the total sum of squares, the Fleiss-Kappa and Kendall's $\tau_b$ inter rater agreement coefficients.

As another way of comparison, a linear predictive model was built with the five first selected features ordered according to the importance established by all the considered methods. Results of applying those metrics are shown in Fig. 5.

## 8 Statistical Comparison

For the sake of comparison, we compare the agreement in the order of features assigned by each method using the inter rater agreement coefficients mentioned




(a) Linear regression (b) Algorithm

**Fig. 5.** Predictive validity.

before and the R-square coefficient of determination. The use of inter rater agreement coefficients on synthetic data has been previously analyzed in [2]. The application of the Cohem-Kappa coefficient in feature selection in order to measure the performance of a fuzzy criterion was presented in [9]. Results of comparison are shown in Fig. 6.

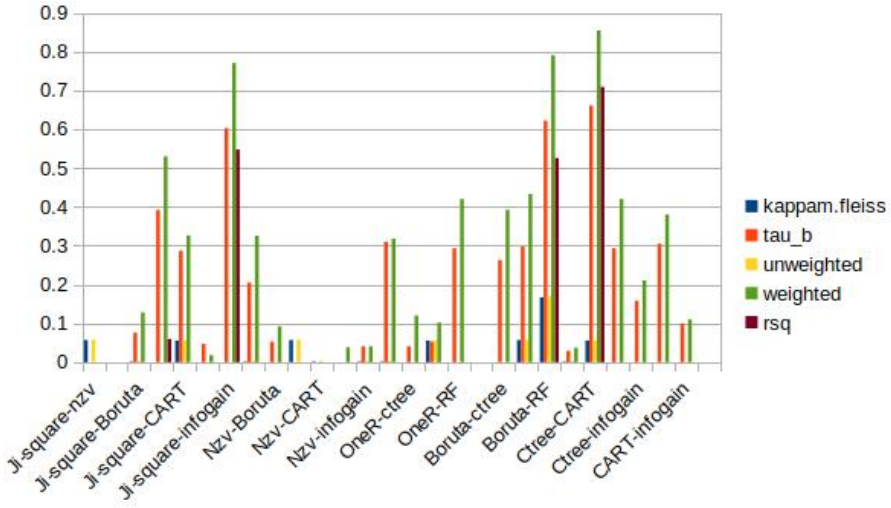



**Fig. 6.** Inter-rater agreement.

## 9 Discussion

From the results shown above, we can highlight the following remarks:

1. Boruta-Random.forest, ctree-CART and ji-square-info.gain are the methods that agree the most.
2. Features selected by Boruta, ctree and CART methods reinforce the study presented in [8] in the way that reviews from Trip Advisor members strongly influences decision of tourists in choosing the hotel.
3. Inter rater agreement statistics appears to be a good metrics in order to compare feature selection methodologies.
4. Inter rater agreement coefficients show poor predictive validity performance for all the methods. The Fleiss-Kappa coefficient indicates the poorest performance, being negative in most of the cases. This requires an extensive analysis regarding what is really considered as agreement due just to chance, as discussed in [4].
5. The predictive validity performance using a linear regression model using the first five features selected by each model is poor for all the models, being the poorest performance for the ji-square test of independence.
6. The weighted Kappa coefficient shows more agreement between methods.
7. Since it takes into account the selected order of features, the $\tau_b$ coefficient appears to be a good metric in order to compare feature selection methods.
8. Mutual information measure (Information Gain) agree with the ji-square test of independence.
9. We have worked on a real dataset.
10. Lasso method does not work well with a relative small quantity of data and its result does not agree with any of the other methods.
11. Though the simplicity of the method, the features selected by the near zero variance reflects that the most variability is achieved by attributes related to fun and recreation. A low variability shows an almost equally spreaded yes/no responses. Since those features are not considered as relevant by the other methods, variations in their values do not follow variation in the target variable. However, some expert may decide to take into account this set of features according to his/her experience.

Despite we are not able to extrapolate our conclusions to a more general case, our analysis allows to evaluate the behaviour of several tools of feature selection, including not just some methods but metrics for comparison and performance evaluation. We have then performed a more detailed study on this dataset than that presented in [8], though arriving to the same conclusions, reinforcing this way that previous analysis.

## 10 Conclusions

An extensive analysis of feature selection methods and metrics for comparison was presented. We have applied those tools to a dataset previously analyzed in the literature. Our study supports that previous conclusions about hotels in Las Vegas are chosen mainly based on information appeared in Trip Advisor regarding reviews by members. The analysis presented is also important in order

to evaluate the tools behaviouring in a small data problem. The use of inter rater agreement coefficients as a metric for comparison in a real dataset was introduced. Fleiss-Kappa, weighted and unweighted Cohen-Kappa and Kendall's $\tau_b$ were analyzed this way. Our study intends to show how to build a feature selection framework in order to guide the expert opinion or managers who have the final decision. We are developing a similar study on other datasets in order to extend the present analysis, results will be presented in future publications.

## Acknowledgment

N. Barraza wishes to thank Universidad Nacional de Tres de Febrero for support under grant no. 32/473 A. A. Moreno wishes to thank Universidad Nacional de Moreno for support.

## References

1. Barraza, N., Moro, S., Ferreyra, M., de la Peña, A.: Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. Journal of Information Science **45**(1), 0165551518770967 (2019). https://doi.org/10.1177/0165551518770967, https://doi.org/10.1177/0165551518770967
2. Bolón-Canedo, V., Rego-Fernández, D., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Sánchez-Maroño, N.: On the scalability of feature selection methods on high-dimensional data. Knowledge and Information Systems (12 2017). https://doi.org/10.1007/s10115-017-1140-3
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Ed). Wiley (2001)
4. Gwet, K.: Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters. Advanced Analytics, LLC (2014), https://books.google.com.ar/books?id=fac9BQAAQBAJ
5. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer (2013), https://faculty.marshall.usc.edu/gareth-james/ISL/
6. Moreno, A., Barraza, N., Daicich, O.: Big data, enfoques multidisciplinarios para la gestión del conocimiento. In: 48JAIIO STS 2019, SIMPOSIO ARGENTINO SOBRE TECNOLOGIA Y SOCIEDAD. pp. 79–92 (Sep 2019)
7. Moro, S., Cortez, P., Rita, P.: A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. Expert Systems **35**(3) (6 2018). https://doi.org/10.1111/exsy.12253, moro, S., Cortez, P., & Rita, P. (2018). A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. Expert Systems, 35(3), [e12253]. DOI: 10.1111/exsy.12253
8. Moro, S., Rita, P., Coelho, J.: Stripping customers' feedback on hotels through data mining: The case of las vegas strip. Tourism Management Perspectives **23**, 41–52 (7 2017). https://doi.org/10.1016/j.tmp.2017.04.003
9. Vieira, S.M., Kaymak, U., Sousa, J.M.C.: Cohen's kappa coefficient as a performance measure for feature selection. In: International Conference on Fuzzy Systems. pp. 1–8 (July 2010). https://doi.org/10.1109/FUZZY.2010.5584447