CCR-Logistic Based Variable Importance Visualization: Differentiating Prime and Suppressor Variables in Logit Models

Ana Perišić[1,2][a] and Ivan Sever[3][b]
[1]*Faculty of Science, University of Split, Split, Croatia*
[2]*Sibenik University of Applied Sciences, Sibenik, Croatia*
[3]*Institute for Tourism, Zagreb, Croatia*

Keywords: Variable Importance Visualisation, Logistic Regression, Correlated Component Regression, Suppression.

Abstract: Logistic regression typically involves assessing variable importance. This task becomes considerably more challenging in the presence of correlated variables (predictors) and suppression. We present a procedure for determining variable importance in multiple logistic regression models that can distinguish between suppressor variables and prime predictors. We propose a simple visualization tool for representing variable importance that can help practitioners to determine important prime and suppressor variables when building the multiple logistic regression model. The methodology relies on the extension of the Correlated Component Regression approach to logistic regression (CCR-Logit), which utilizes linear combinations of predictors instead of original predictors and can easily be generalized to various regression models. CCR-logistic methodology can handle a large number of predictors and is especially useful when dealing with correlated predictors. The variable importance is quantified by observing standardized regression coefficients from univariate models and higher-order component models, where univariate models capture the direct effect on the outcome, while the higher-order component models capture the suppressor effects. The proposed methodology is presented on a real-world dataset within the field of tourism.

1 INTRODUCTION

When building a regression model it can be more efficient to select a subset of relevant predictors, than to build a regression model on a large set of all possible variables. There are several reasons for this, from the theoretical and practical side. From the practical side, simple models are more effective than complex models, more cost-efficient and time-efficient, are easier to interpret, and often are more stable on out-of-sample data. On the theory side, good theories are parsimonious, containing only those constructs essential for understanding a certain phenomenon of interest (Braun and Oswald, 2011). Thus, assessing variable importance is essential when building regression models. A detailed review of various variable importance metrics developed for linear models together with several important properties that variable importance metrics should satisfy can be found in Grömping (2015).

Various metrics and approaches for variable importance assessment in logistic regression have been developed. The most widely adopted approaches include standardized regression coefficients which often rely on different approaches to standardization (see for instance Menard (2004)). Also, a popular method for evaluating predictor importance is dominance analysis where one predictor is considered as more important than another if it contributes more to the prediction of the criterion than does its competitor at a given level of analysis (Azen and Traxel, 2009). Moreover, the analyses often include calculating test values, information, and prediction performance measures for nested models (such as performing the LR test or comparing AIC, BIC, AUC for a model that includes and model that does not include a variable of interest). Some model-building strategies can be found in Hosmer and Lemeshow (2000). As in the general case, when building prediction models and assessing feature importance, there is no definitive or unambiguous method for establishing predictor importance (Braun and Oswald, 2011).

Assessing variable importance in logistic regres-

[a] https://orcid.org/0000-0001-9180-0270
[b] https://orcid.org/0000-0002-7043-4862

Perišić, A., Sever and I.
CCR-Logistic Based Variable Importance Visualization: Differentiating Prime and Suppressor Variables in Logit Models.
DOI: 10.5220/0013461700003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 43-52
ISBN: 978-989-758-758-0; ISSN: 2184-285X

sion with a large set of potential predictors is not straightforward. Similar to multiple linear regression, the relative importance of a predictor variable in logistic regression can vary depending on the subset of predictor variables included in the model (Azen and Traxel, 2009). Moreover, assessing predictor importance becomes more challenging in the presence of suppression. A suppressor variable shares no variance directly with the dependent variable and thus contributes to the regression model through removing irrelevant variance from the other independent variables (Nathans et al., 2012). There are different approaches to defining a suppressor variable, and thus different approaches for identifying a suppressor variable in the regression model (see for instance Friedman and Wall (2005); Ludlow and Klein (2014); Shieh (2006); Velicer (1978)). Some of the common approaches include observing regression coefficients and corresponding t-statistics. For instance, some approaches suggest that suppression exists if the squared multiple regression coefficient for a particular predictor is higher than the squared univariate regression coefficient for the same predictor. Instead of multiple regression coefficient and squared univariate regression coefficient, we can also evaluate the t-statistics of the estimated coefficient. Also, some approaches suggested observing the changes in the estimated regression coefficients when adding new predictors: suppression is present if the change in the estimated regression coefficient of a predictor is significant when adding a new predictor into the model. Other approaches suggest that variable X_j is a suppressor when the squared multiple correlation coefficient of Y with all predictors $X_1, X_2, \ldots, X_P$ is larger than the sum of the squared multiple correlation coefficient of Y with all predictors except X_j, and the squared correlation coefficient of Y and X_j.

A powerful tool in understanding regression models is visualization. Variable importance in regression models is mostly visualized through bar plots and line plots that present the variable importance metric. Visualizations that exceed the one-dimensional aspect of presenting variable importance have also been developed. For instance, Inglis et al. (2022) constructed heatmap and graph-based displays showing variable importance and interaction jointly.

In this work, we present a visualization tool that presents variable importance in a logistic regression model and distinguishes between the direct and indirect variable effects. The proposed methodology is capable of handling a large set of correlated predictors. Along with distinguishing between the direct and indirect effects, the proposed visualization covers three dimensions of interest when evaluating a predictor: statistical significance, its total effect and direction of the relationship. We introduce the methodology in the second section and present the application to a real-world problem in section 3.

2 PROPOSED METHODOLOGY

The methodology for visualizing variable importance presented in this work relies on the Correlated Component Regression (CCR) method. The CCR method, introduced by Magidson J. (Magidson, 2010, 2013), is a dimension reduction method developed for multiple regression models that utilizes $K < P$ correlated linear combinations of the predictors instead of the original P predictors, to predict an outcome variable. The first component captures the effects of predictors that have a direct effect on the outcome, while the higher-order components capture indirect effects, i.e. the effects of suppressor variables that improve prediction by removing extraneous variation from one or more of the predictors that have direct effects. This approach identifies prime predictors as those having substantial loadings on the first component, and suppressor variables as those having substantial loadings on higher-order components, and relatively small loadings on the first component (Magidson, 2013). For instance, in the case of two components, pure suppressor variables have zero loadings on the first but highly significant loadings on the second CCR component (Magidson, 2010).

The CCR algorithm is developed for multiple regression models and has different variants depending on the scale type of the outcome variable. For instance, when the outcome variable is dichotomous, we can apply the CCR-logistic regression (CCR-Logit) approach. The easiest way of adapting the CCR methodology to the logistic regression case is performing the logit transformation of the outcome variable and then evaluating the model as the multiple linear regression model. We first present the CCR algorithm extended to logistic regression (CCR-Logistic) introduced in Magidson (2013). Also, we present the approach for identifying prime and suppressor variables.

Assume we have a collection of $X_1, X_2, \ldots, X_P$ predictor variables and we are building a logistic regression model where the dichotomous outcome variable is denoted by Y. For ease of understanding, we denote the logit transformation $Logit(Y)$ simply by Y. The algorithm is executed through the following steps, denoted as S_1 to S_3:

S1. Univariate Models

Step 1.1. Estimate P univariate models
For each predictor $X_i, i = 1,2,\ldots,P$, estimate the univariate model

$$Y = \beta_{0i} + \lambda_i^{(1)} X_i + \varepsilon_i.$$

Here, β_{0i} represents the intercept, ε_i is the error term, $\lambda_i^{(1)}$ is the univariate regression coefficient of interest that captures the direct effect of the predictor variable X_i on the outcome. For each predictor $X_i, i = 1,2,\ldots,P$, check the associated p-value and denote it by $\text{pv}_i^{(1)}$. The associated p-values are measures of significant direct effects. Predictors that have significant coefficients are considered as prime predictors (here we take $\text{pv}_i^{(1)} < 0.1$, but this bound can be changed).

Step 1.2. Univariate regression coefficient standardization
For each predictor $X_i, i = 1,2,\ldots,P$, standardize the univariate regression coefficients by calculating:

$$\lambda_i^{*(1)} = \lambda_i^{(1)} \sigma_{X_i}, \quad i = 1,2,\ldots,P$$

where $\lambda_i^{(1)}$ is the regression coefficient estimated in the univariate regression model, and σ_{X_i} is the standard deviation of the predictor X_i.

S2. Higher Order Components

Step 2.1. Estimate the first component
The first component S_1 is defined as the weighted linear combination of P predictors, with weights being proportional to estimated coefficients $\lambda_i^{(1)}$:

$$S_1 = \frac{1}{P} \sum_{i=1}^{P} \lambda_i^{(1)} X_i.$$

The first component captures the total direct effect of all predictors.

Step 2.2. Estimate the higher-order components
For $k = 2,\ldots,K < P$, define the k-th component S_k as the weighted average of all 1-predictor partial effects:

$$S_k = \frac{1}{P} \sum_{i=1}^{P} \lambda_i^{(k)} X_i$$

where weights $\lambda_i^{(k)}$ are estimated from the regression models:

$$Y = \alpha_i + \gamma_{1.i}^{(k)} S_1 + \cdots + \gamma_{(k-1).i}^{(k)} S_{k-1} + \lambda_i^{(k)} X_i + \varepsilon_i,$$

$i = 1,2,\ldots,P$. Higher-order components capture the effect of suppressor variables that improve predictions by removing extraneous variation from prime predictors. For each $\lambda_i^{(k)}$, $i = 1,2,\ldots,P$, check the associated p-value and denote it by $\text{pv}_i^{(k)}$. The associated p-values are measures of significant suppressor effect. Predictors that have at least one significant coefficient $\lambda_i^{(k)}$, for $k = 2,\ldots,K$ ($\text{pv}_i^{(k)} < 0.1$) are considered suppressor predictors.

Step 2.3. The standardized coefficient
For each predictor $X_i, i = 1,2,\ldots,P$, calculate the standardized coefficient

$$\lambda_i^{*(k)} = \lambda_i^{(k)} \sigma_{X_i}, \quad i = 1,2,\ldots,P, \quad k = 2,\ldots,K.$$

S3. The final K-component model

Step 3.1. The Final K-Component Model
Estimate the final K-component model, which is defined as a regression model with outcome Y and predictors $S_1, S_2, \ldots, S_K$:

$$Y = \alpha^{(K)} + \sum_{k=1}^{K} b_k^{(K)} S_k + \varepsilon$$

Step 3.2. Regression coefficients for the predictors
The predicted values of the outcome variables are then:

$$\hat{Y} = \alpha^{(K)} + \sum_{k=1}^{K} b_k^{(K)} S_k$$

and can then be easily re-expressed to obtain regression coefficients for the predictors by substituting as follows:

$$\hat{Y} = \alpha^{(K)} + \sum_{k=1}^{K} b_k^{(K)} \sum_{i=1}^{P} \lambda_i^{(k)} X_i = \alpha^{(K)} + \sum_{i=1}^{P} \beta_i X_i.$$

The coefficient β_i for predictor X_i is the weighted sum of the loadings, where the weights are the regression coefficients of the components in the K-component model:

$$\beta_i = \sum_{k=1}^{K} b_k^{(K)} \lambda_i^{(k)}.$$

Step 3.3. Standardized final CCR coefficients
Calculate the associated standardized coefficient as:

$$\beta_i^* = \beta_i \sigma_{X_i}, \quad i = 1,2,\ldots,P.$$

The optimal number of components and predictors involved can be found by performing cross-validation on the training dataset. Results from simulations and applications with real high-dimensional data suggest that CCR models rarely require more than 10 components regardless of the number of predictors and usually perform well with 3 or 4 components, while the estimation is fast (Magidson (2010)).

Having the results of the CCR-Logit algorithm, we establish the visualization in the Cartesian coordinate system by covering 5 dimensions of interest:

Figure 1: Proposed visualization.

(D1) (Prime/Direct Effect)
We observe the direct impact of each variable on the outcome by presenting the absolute value of the standardized univariate regression coefficient $\lambda_i^{*(1)}$ on the y-axis. Variables that have significant univariate regression coefficients are considered as significant prime predictors.

(D2) (Suppressor Effect)
We observe the indirect impact of each variable on the outcome by presenting the largest absolute value of the standardized regression coefficient $\lambda_i^{*(k)}, k > 1$ on the x-axis, i.e. we present $\lambda_i^{*(Amax)} = \max_{k>1} |\lambda_i^{*(k)}|$. Variables that have at least one significant $\lambda_i^{*(k)}$ coefficient are considered as significant suppressor predictors.

(D3) (Statistical Significance)
In the proposed visualization, each variable is presented by a data point $(\lambda_i^{*(Amax)}, |\lambda_i^{*(1)}|)$. The significance of each variable is captured by associated p-values $(\min_{k>1} pv_i^{(k)}, pv_i^{(1)})$. Following this simple visualization strategy, we consider a categorization of a variable into 4 cases: a predictor can be a (I) a significant prime predictor, (II) a significant suppressor predictor, (III) both a significant prime and a significant suppressor predictor, and (IV) a nonsignificant prime and a nonsignificant suppressor. Thus, we divide the visualization area into four quadrants according to the significance of the predictors in the univariate and higher-order models. The vertical line is placed at $x = \frac{1}{2}(\lambda_m + \lambda_M)$ where λ_M is the highest value of the standardized coefficients $\lambda_i^{*(Amax)}$ for the variables that had no significant coefficients in the higher-order components, i.e. we take $\lambda_M = \max_{X_i \text{not suppressor}} \lambda_i^{*(Amax)}$, while λ_m is the lowest value of standardized coefficients $\lambda_i^{*(Amax)}$ for the predictors that had significant coefficients in the higher-order components, i.e. $\lambda_m = \min_{X_i \text{ suppressor}} \lambda_i^{*(Amax)}$. The horizontal line is placed at $y = \frac{1}{2}(\lambda_{um} + \lambda_{UM})$, where λ_{um} is the lowest absolute value of the univariate standardized coefficient $|\lambda_i^{*(1)}|$ of the univariate significant predictors, i.e. $\lambda_{um} = \min_{X_i \text{ prime}} |\lambda_i^{*(1)}|$, while λ_{UM} is the highest absolute value of the univariate standardized coefficients $|\lambda_i^{*(1)}|$ of the univariate non-

significant predictors, i.e. $\lambda_{UM} = \max_{X_i \text{ not prime}} |\lambda_i^{*(1)}|$.

(D4) (Overall Effect)
The overall effect of each predictor on the outcome is visualized by the size of each data point associated with the predictor. The size of each data point is proportional to the normalized value of the associated absolute value of the final standardized coefficient β_i^* from the final CCR model. The normalized value is calculated as

$$\text{NORM}\beta_i^* = \frac{1}{\sum_{i=1}^{P} |\beta_i^*|} |\beta_i^*|.$$

(D5) (Direction)
The visualization is enriched by adding the information on the direction of the (overall) relationship between the predictor and the outcome variable. This is achieved by presenting positive final standardized coefficients in one color, and negative final standardized coefficients in another color.

The example of such a visualization is presented in Figure 1. The Figure is divided into four areas distinguishing between the (significant) pure prime, (significant) pure suppressor, (significant) prime and suppressor, and nonsignificant variables. Predictors having a positive overall effect on the outcome are presented in blue, while predictors having a negative overall effect are presented in red. The size of each dot is proportional to the absolute value of the overall effect. In this theoretical example, we have a collection of 8 variables included in the regression analysis. Three variables, P5, P4 and P6 have a significant direct effect on the outcome. Predictors P4 and P5 have the largest overall effect on the outcome and are positively related to the outcome. Predictor P6 is negatively related to the outcome. Predictor P2 is a (pure) suppressor variable, positively related with the outcome. Predictor P8 has both direct and indirect effect on the outcome. The overall effect of the predictor P8 on the outcome is positive. This hypothetical example classifies three variables as both nonsignificant prime and nonsignificant suppressor variables, meaning that these variables should be excluded from the regression analysis. Note that this example is theoretical and that in practice we expect that the number of both not prime and not suppressor variables should be low. In fact, when dealing with carefully planned analyses (this means that the variables (predictor candidates) included in the regression analysis are carefully selected) we expect that the selected variables will have direct, indirect or both direct and indirect effect on the outcome.

3 APPLICATION

We present the application of the proposed methodology on a real-world dataset from the survey on residents' perceptions of tourism impacts and their attitudes toward tourism in the city of Split, Croatia. Split is the second-largest city in Croatia and the largest Croatian city on the Adriatic coast, with approximately 160,000 inhabitants. As a Mediterranean city with exceptional cultural-historical heritage and natural beauty, Split is a highly attractive tourist destination. In 2022, 2.6 million overnight stays were realized in its commercial accommodation facilities. The intensive growth of tourism over the past decade has put a lot of pressure on residents' well-being and their living environment (Matečić et al., 2022). The survey of local residents in the city of Split, which was conducted in June 2022 on a sample of 385 respondents, was designed to identify the key drivers of adverse tourism impacts in the city and thus support effective monitoring, management, and mitigation of risks associated with overtourism. The sample was representative at the city level by gender and age group of residents. Computer Assisted Telephone Interview (CATI) was used as a data collection method.

The dataset comprises eleven variables related to residents' perceptions of tourism impacts in the city of Split. A detailed description of included variables (i.e., impact indicators) can be found in the Appendix. Six numerical variables are used in their original form where Appearance, Apartmentization, Authenticity, Space, and Services are responses to a 5-point rating scale, while Displacement is a binary variable. Other four numerical variables F1:Social crowding, F2:Waste and cleanliness, F3:Current expenses, and F4: Housing affordability are constructed through exploratory factor analysis. Factors F1:Social crowding and F2:Waste and cleanliness were established by performing factor analysis on the set of crowding-related variables: Noise, Traffic, Crowding, Transport, Littering, Smell, Tourist behavior and Parking. Factors F3:Current expenses and F4: Housing affordability are constructed through exploratory factor analysis applied on a set of price-related tourism impact items: Housing affordability, Realestate prices, Rent, Utility prices, Grocery prices, and Restaurant prices. These ten variables are (theoretically) assumed to affect the outcome variable. The outcome variable Perception is a binary variable that presents the perception of overall tourism impacts. It is formed by categorizing the overall attitude toward tourism impacts, measured on a 5-point Likert scale anchored by very negative and very positive, as either positive or neutral/negative.

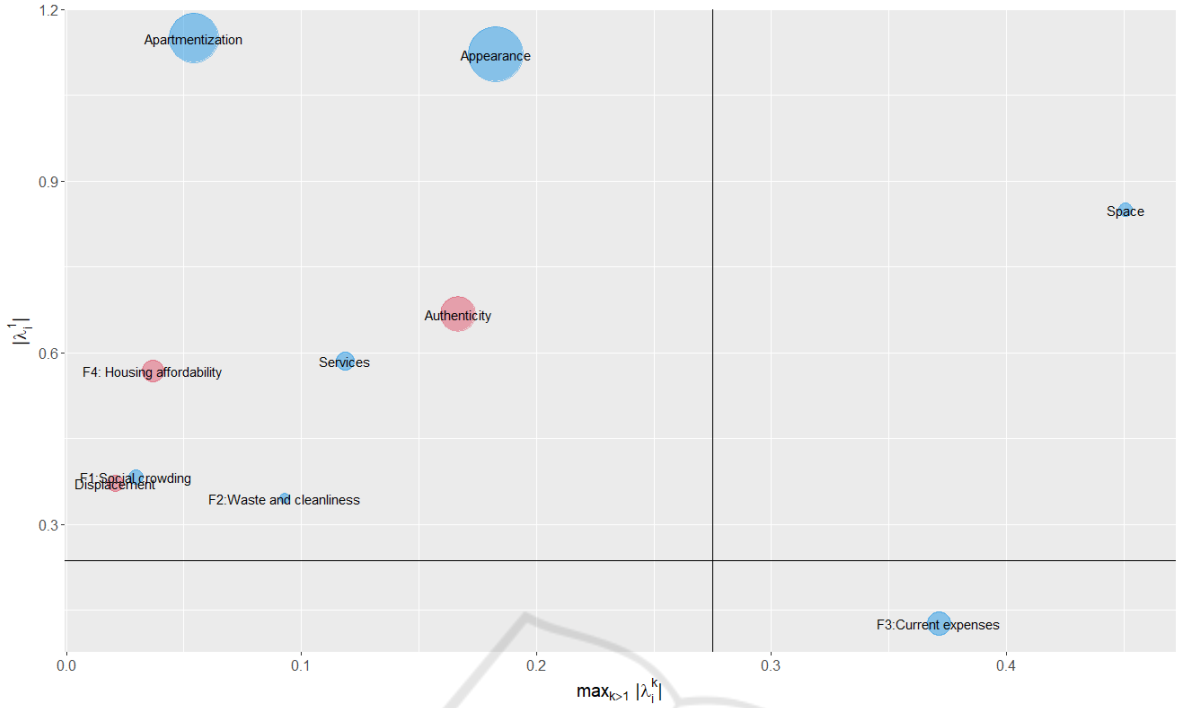

Figure 2: CCR-logit visualization example: tourism data.

The goal of the analysis is to build a model that explains the perception of overall tourism impacts by using the set of ten aforementioned variables. Since the outcome variable is a binary variable, it is reasonable to conduct logistic regression analysis, where we estimate the model

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P$$

that best explains the outcome. This also means selecting the most important predictors and explaining the relationship of each predictor with the outcome. For this reason, we perform the CCR-Logit based visualization.

Before applying the CCR-Logit algorithm we determined the value for the number of components K that provides the optimal amount of regularization. We chose the CCR model that maximizes the cross-validated area under the curve (AUC), accuracy (ACC) and sensitivity (SENSI).

We performed the cross-validation by splitting the data into 10 exclusive partitions. Each partition was used for test-training split where we estimate the model with K components ($K = 1,2,\ldots,7$). For each number of components $K, K = 1,2,\ldots,7$ we averaged the performance metrics as presented in Table 1. Based on 10-predictor models, the model with $K = 2$ components provides the maximum mean prediction metrics values.

Table 1: Performance metrics for different values of K.

K	AUC	ACC	SENSI
1	0.86	0.73	0.76
2	0.86	0.76	0.80
3	0.86	0.75	0.79
4	0.86	0.76	0.79
5	0.86	0.76	0.79
6	0.86	0.76	0.79
7	0.86	0.76	0.79

We estimated the two-component CCR model. The estimated coefficients $\lambda_i^{(k)}$, standardized coefficients $\lambda_i^{*(k)}$, and the associated p-values $\text{pv}_i^{(k)}$, for $k = 1,2$ and $i = 1,\ldots,10$, are presented in Table 2. Also, we present the standardized value of the final coefficient β_i^* in the same table.

The visualization of the results prepared accordingly to the proposed methodology in the second section is presented in Figure 2. Several conclusions can be drawn from this visualization. Three main prime predictors are Apartmentization, Appearance, and Authenticity. These predictors are located in the Prime area and have the largest overall effect. Predictors Apartmentization and Appearance have a positive overall effect on Perception, while Authenticity has a negative overall effect due to coding. Predictors F4:Housing affordability, Services, F1:Social crowding, F2:Waste and cleanliness, and Displacement are

Table 2: Estimated coefficients, p-values, and final coefficients for the two-component CCR model.

Predictor	**First component** $\lambda_i^{*(1)}$	$\text{pv}_i^{(1)}$	**Second component** $\lambda_i^{*(2)}$	$\text{pv}_i^{(2)}$	**Final coefficients** β_i^*
Apartmentization	1.149	< 0.001	0.05	0.809	0.853
Appearance	1.121	< 0.001	0.18	0.383	0.94
Authenticity	-0.667	< 0.001	-0.17	0.338	-0.607
Space	0.850	< 0.001	-0.45	0.053	0.226
Services	0.585	< 0.001	-0.12	0.509	0.313
F1: Social crowding	0.383	0.002	-0.03	0.084	0.245
F4: Housing affordability	-0.568	< 0.001	0.04	0.835	-0.368
F3: Current expenses	0.127	0.279	0.37	0.025	0.397
Displacement	-0.372	0.004	-0.02	0.903	-0.279
F2: Waste and cleanliness	0.345	0.008	-0.09	0.593	0.166

significant prime predictors having lower importance than the three aforementioned prime predictors. This lower importance is measured through the smaller overall effect presented as the size of each dot. Predictor Space is located in the Prime&suppressor area, thus it is both a significant prime and a significant suppressor variable. Since variable Space has a positive direct effect and a negative indirect effect, the total effect, measured as the normalized value of the final CCR correlation coefficient, is low. Predictor F3:Current expenses is located in the Suppressor area, thus it is a significant suppressor predictor.

We compare the results of the presented visualization and the resulting conclusions on variable importance by applying a commonly used method for comparing the relative importance of predictors in multiple regression: dominance analysis (Azen and Traxel, 2009; Budescu, 1993). Dominance analysis is a popular method to determine the relative importance of correlated variables, which ranks a given predictor by measuring how much it contributes to explaining the outcome, measured as a change in the McFadden's R^2, in all possible subset models formed by the combinations of other predictors. We present the results of the conditional and general dominance analysis. Conditional dominance is calculated as the average of the additional contributions to all subsets of models of a given model size. General dominance is calculated as the mean of average contributions across all model sizes. Results are presented in Figure 3 and Figure 4.

The outcomes of the dominance analysis reinforce the conclusions drawn from the CCR-logit visualization. Predictors Apartmentization, Appearance, and Authenticity are the three most important variables according to their average contribution based on the general dominance criterion. Space was ranked the fourth most important variable. Notice that the conditional dominance of the predictor F3:Current expenses increases as the number of variables in the model increases, which can be an indicator of suppression.

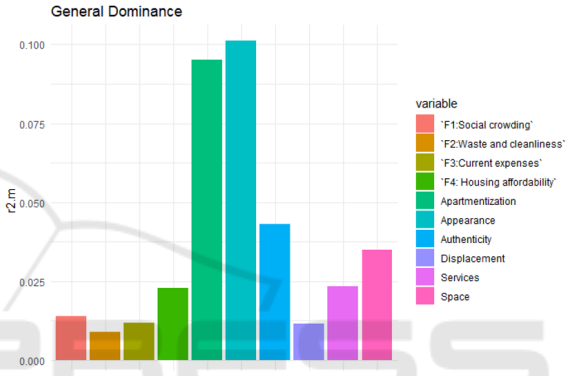

Figure 3: General dominance plot.

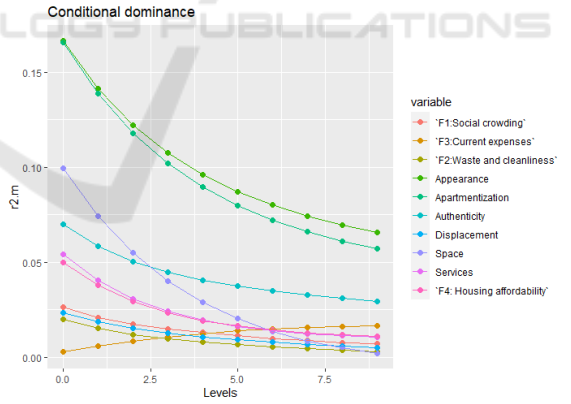

Figure 4: Conditional dominance plot.

4 DISCUSSION AND CONCLUSION

Assessing predictor importance in logistic regression is an integral part of building the logistic regression

model. It is often important to distinguish between the predictors that have direct and predictors that have indirect effects on the outcome variable. We present a visualization tool that can help modelers to identify important variables in the logistic regression model while distinguishing between the prime and suppressor effects. The visualization relies on the CCR-logit approach which utilizes correlated linear combinations of the predictors instead of the original $P > 1$ predictors. This tool can be useful for determining variable importance and supporting the theoretical implications of the model by interpreting the predictor effect on the outcome. Also, it can be helpful when building regression models, for instance, in the stepwise regression procedures as an additional tool for variable selection.

From the perspective of empirical analysis presented in the application part, we can set several conclusions and recommendations for tourism sustainability monitoring practice. Apartmentization, Appearance, and Authenticity are the most important prime predictors for modeling perceptions of tourism impacts in the city of Split. Moreover, Space and F3:Current expenses are not the primary variables of interest in the context of assessing tourism sustainability in the city of Split. Still, they are important variables that should be measured and included in the analysis as control variables. By controlling for suppressors, we can obtain more accurate estimates of the unique contributions of the primary variables of interest and enhance their predictive power. Furthermore, suppressor variables indicate the presence of indirect effects in the regression model and thus help in clarifying the true nature of relationships between the variables.

The proposed methodology is presented as a tool for visualizing predictor importance in multiple logistic regression, but can be easily generalized to multiple linear regression. The generalization to multiple linear regression can be established simply by excluding the logit transformation and following the steps presented in this paper.

There are several challenges related to future improvements of the proposed visualization. First, when working with categorical predictors with more than two categories, we usually introduce dummy variables. In this case, more than one regression coefficient is related to one categorical predictor. Thus, special procedures should be developed for multiple regression models involving categorical variables with more than two categories. One of the possibilities is to take the dummy variable with the smallest p-value as a representative for each categorical variable. Secondly, valuable information missing in the proposed visualization is related to the predictive power of the predictors. It would be beneficial to present the predictive power (such as AUC, AIC) related to each predictor. For instance, we could use the prediction metrics applied on nested models to visualize the overall prediction power of a predictor, and this information could replace the normalized final CCR coefficient which was used as a measure of overall effect. This could even be a preferable choice since the final CCR coefficient can diminish the actual importance of suppressor variables. For instance, this can be the case when the suppressor has opposite regression coefficients for direct and indirect effects (such as the predictor Space in our example).

Preparing simple visualization tools for presenting predictor importance is crucial for enhancing the clarity and accessibility of complex data. By converting intricate relationships into easily interpretable visuals, these tools allow stakeholders to quickly grasp the significance of various predictors in a model. Determining predictor importance in multiple regression is sensitive to both the subjective decisions of the modeler and the inherent characteristics of the dataset. For instance, the choice of which variables to include in the model and how to handle correlations between predictors can significantly influence the results. A common approach is to exclude highly correlated predictors, focusing on the importance of the remaining variables. However, this may lead to the exclusion of important predictors. On the other hand, retaining all correlated predictors without addressing multicollinearity can result in inflated standard errors, potentially misleading the modeler into undervaluing the importance of certain predictors, even if they have a significant effect. The proposed CCR-Logistic based variable importance visualization method utilizes the full set of predictors and is capable of handling multicollinearity. Together with its simplicity, this constitutes a key advantage of the proposed visualization. This not only aids in decision-making but also ensures transparency and facilitates better communication of results to both technical and non-technical audiences. Simple visualizations foster a deeper understanding, support actionable insights, and ultimately contribute to more informed and effective data-driven decisions.

ACKNOWLEDGEMENTS

This work has been supported by the NextGenerationEU under the scientific project SURVEY+ of the Institute for Tourism, Croatia.

REFERENCES

Azen, R. and Traxel, N. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, 34(3):277–303.

Braun, M. T. and Oswald, F. L. (2011). Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behavior Research Methods*, 43(2):453–466.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3):542–551.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *American Statistician*, 59(2):127–136.

Grömping, U. (2015). Variable importance in regression models. *WIREs Comput. Stat.*, 7(2):137–152.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression (Wiley Series in probability and statistics)*. Wiley-Interscience Publication, 2 edition.

Inglis, A., Parnell, A., and Hurley, C. B. (2022). Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, 31(3):766–778.

Ludlow, L. and Klein, K. (2014). Suppressor variables: The difference between "is" versus "acting as". *Journal of Statistics Education*, 22(2):77–88.

Magidson, J. (2010). Correlated component regression: A prediction / classification methodology for possibly many features. Training.

Magidson, J. (2013). Correlated component regression: Rethinking regression in the presence of near collinearity. In *Springer Proceedings in Mathematics and Statistics*, volume 56, pages 39–56.

Matečić, I., Kesar, O., and Hodak, D. F. (2022). Understanding the complexity of assessing cultural heritage's economic impact on the economic sustainability of a tourism destination: the case of Split, Croatia. *Ekonomska misao i praksa*, 31(2):639–662.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *American Statistician*, 58(3):218–223.

Nathans, L., Oswald, F., and Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research and Evaluation*, 17(9):1–19.

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and Psychological Measurement*, 66(3):527–539.

Velicer, W. F. (1978). Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement*, 38(4):953–958.

APPENDIX

Indicator	Description	Scale
Overall attitude	Think for a moment about how tourism affects your daily life, the local economy, the environment, safety, prices, etc. Considering both the good and bad sides of tourism, do you think that life in Split is worse or better because of tourism?	Much worse (1) —– Much better (5)
Perception	Binarized overall attitude: perception=1 if Overall attitude $= 4,5$; else perception=0	Positive perception (1) vs negative perception (0)
Appearance	How does tourism development affect the appearance of the city?	It has become much uglier (1) —– It has become much more beautiful (5)
Apartmentization	What do you think about converting residential dwellings into tourist rentals, does it make life in Split worse or better?	Much worse (1) —– Much better (5)
Authenticity	How much has the character of the city changed over the past decade? Has Split lost its spirit, its authenticity?	Not at all (1) —– Completely (5)
Space	What do you think about the use of public spaces in Split? Has tourism made public spaces (promenade, city streets and squares, green areas) less or more suitable for your needs?	Much less suitable (1) —– Much more suitable (5)
Displacement	Have you, or anyone in your family/friends, moved out of the city center of Split in the last ten years?	Yes / No
Services	With intensive tourism development in Split, have the public amenities for local residents – such as kindergartens, schools, healthcare facilities, markets, and libraries – become less or more accessible?	Much less accessible (1) —– Much more accessible (5)
Noise	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Noise	A major problem (1), A minor problem (2), Not a problem at all (3)
Traffic	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Traffic congestion	A major problem (1), A minor problem (2), Not a problem at all (3)
Crowding	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Crowding on the streets/public areas	A major problem (1), A minor problem (2), Not a problem at all (3)
Transport	When you think about your daily life in Split during the tourist season, how much of a problem is the following: congestion on public transport	A major problem (1), A minor problem (2), Not a problem at all (3)
Littering	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Improperly disposed waste	A major problem (1), A minor problem (2), Not a problem at all (3)
Smell	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Unpleasant smells (from containers and waste bins)	A major problem (1), A minor problem (2), Not a problem at all (3)
Tourist behavior	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Inappropriate tourist behavior	A major problem (1), A minor problem (2), Not a problem at all (3)
Parking	When you think about your daily life in Split during the tourist season, how much of a problem is the following: Finding a parking space	A major problem (1), A minor problem (2), Not a problem at all (3)
Housing affordability	How satisfied are you with the affordability of housing in Split?	Very dissatisfied (1) —– Very satisfied (5)
Realestate prices	To what extent do you think realestate prices in Split have increased over the last five years due to tourism?	Not at all (1) —– Very much (5)
Rent	To what extent do you think rent in Split has increased over the last five years due to tourism?	Not at all (1) —– Very much (5)
Utility prices	To what extent do you think utility prices in Split have increased over the last five years due to tourism?	Not at all (1) —– Very much (5)
Grocery prices	To what extent do you think grocery prices in Split have increased over the last five years due to tourism?	Not at all (1) —– Very much (5)
Restaurant prices	To what extent do you think the prices in restaurants/cafes in Split have increased over the last five years due to tourism?	Not at all (1) —– Very much (5)