Feature Selection Model: Economical Use

Ceballos Arias, Juan Camilo - jcceballoa@eafit.edu.co
Velasquez Gaviria, Diana Catalina - dvelasq8@eafit.edu.co
Castelblanco Benites, Julián - jcastelblb@eafit.edu.co
Moreno Zapata, Juan Sevastian - jsmorenoz@eafit.edu.co

Master's in Data Science and Analytics
Professor: Andres Ramirez-Hassan

October 12, 2019

1 Descriptive Analysis

A dataset with 200 records and 50 characteristics is analyzed. The price of each variable is also taken into account.

By the time the descriptive analysis is performed, it is found that if 200 surveys were going to be conducted in the second phase of the experiment, using all the variables of the dataset would meet the whole assigned budget. However, a feature and best model selection analysis is made so that the number of surveys to conduct can be maximized, minimizing the prediction error - mean squared error (mse).

To complete this task, the dataset was preprocessed standardizing all the variables, to make sure they were at the same scale before making the feature selection, and then splitting the dataset for the posterior model training and testing. For training, 70% of the data was used and for testing the 30% missing.

2 Feature Selection

Selecting the features to build the model was made by using different techniques such as filter methods (constant and quasi-constant variance, duplicated features, correlation, mutual information, univariate mse), wrapper methods (step forward, step backward, subset) and embedding methods (lasso regularization, lars, lasso Lars, elasticnet, random forest importance, gradient boosting machines).

Applying filter methods, no variables were selected because it was not found any duplicated variables nor ones with constant or quasi-constant variance.

The selection process was performed over the train dataset, in order to avoid overfitting problems. Additionally, while evaluating the methods where it was necessary to predefine the number of characteristics to select, some iterations were done over a list of possible number of them and the best option was the one which minimized the mse. Also, a cross-validation with k=5 was performed to strengthen the process.

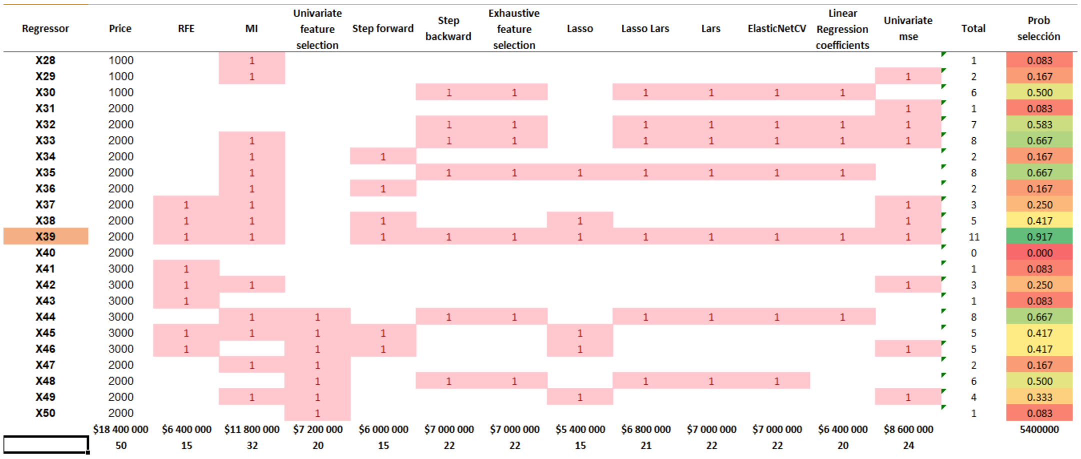
The techniques used for selecting features are mentioned as follows:

- Feature ranking with recursive feature elimination (RFE).

- Mutual information.
- Univariate feature selection.
- Step forward: iterating through different numbers of characteristics and the one with the highest AUC was chose. Furthermore, a cross-validation with k=5 was computed.
- Step backward: iterating through different numbers of characteristics and the one with the highest AUC was chose. Furthermore, a cross-validation with k=5 was computed.
- Exhaustive feature selection (subset): computed combining different numbers of characteristics which were increasing until 12 and a cross-validation with k=2, due to the algorithm's high computational cost.
- Lasso regularization: this technique was implemented with different regularization criteria: c=0.5, c=1 and c=1.5. The higher the hyper-parameter value, the lower the penalization strictness. The features selected by those three models were taken into account.
- Lasso lars: cross-validating with k=5.
- Elasticnet regularization: cross-validating with k=5.
- Lars: cross-validating with k=5.
- Linear regression coefficients.
- Univariate mse.

In those methods where the regularization hyper-parameters were varied, the selection of the number of features was made depending on the case where the best mse was obtained.

Taking into account the multiple existing methods for feature selection and their pros and cons, it was thought that if there are variables which are selected by several or the majority of them, it is because those variables are important and, therefore, after evaluating all the techniques abovementioned with their different runs, a table containing the list of variables, their prices and the number of methods used was created as follows 1:



Regressor	Price	RFE	MI	Univariate feature selection	Step forward	Step backward	Exhaustive feature selection	Lasso	Lasso Lars	Lars	ElasticNetCV	Linear Regression coefficients	Univariate mse	Total	Prob selección
X28	1000		1											1	0.083
X29	1000		1										1	2	0.167
X30	1000					1	1		1	1	1	1		6	0.500
X31	2000												1	1	0.083
X32	2000					1	1		1	1	1	1	1	7	0.583
X33	2000		1			1	1		1	1	1	1	1	8	0.667
X34	2000		1		1									2	0.167
X35	2000		1			1	1	1	1	1	1	1		8	0.667
X36	2000		1		1									2	0.167
X37	2000	1	1										1	3	0.250
X38	2000	1	1		1			1					1	5	0.417
X39	2000	1	1		1	1	1	1	1	1	1	1	1	11	0.917
X40	2000													0	0.000
X41	3000	1												1	0.083
X42	3000	1	1										1	3	0.250
X43	3000	1												1	0.083
X44	3000		1	1		1	1		1	1	1	1		8	0.667
X45	3000	1	1	1	1			1						5	0.417
X46	3000	1		1	1			1					1	5	0.417
X47	2000		1	1										2	0.167
X48	2000			1		1	1		1	1	1			6	0.500
X49	2000		1	1				1					1	4	0.333
X50	2000			1										1	0.083
	$18 400 000	$6 400 000	$11 800 000	$7 200 000	$6 000 000	$7 000 000	$7 000 000	$5 400 000	$6 800 000	$7 000 000	$7 000 000	$6 400 000	$8 600 000		5400000
	50	15	32	20	15	22	22	15	21	22	22	20	24		

Figure 1: Correlation Matrix

The first column of the table contains all the variables, the second one the cost of each variable and the others represent the selection methods and contains a mark depending on whether or not the variable was selected by the method in the specific column. The last two columns of the table hold information about two measurements, the total number of times that a variable was selected by the methods and the selection probability, defined as the number of times a variable was selected divided by the total number of methods analyzed. In addition, the survey cost is calculated for every set of selected variables.

Based on the table described, alternative approaches for feature selection were used. Firstly, the variables obtaining a selection probability higher than 0.5 were taken and, secondly, the characteristics with a selection probability higher than 0.6, 0.7 and 0.9 were considered as the most important. The results are shown below:

Variables with a selection probability higher than 50% - 22 variables selected

X1 X2 X4 X5 X8 X9 X11 X12 X13 X14 X16 X1 X18 X19 X24 X30 X32 X33 X35 X39 X44 X48

Considering this selection of 22 variables, the estimated cost of a survey would be $35.000 pesos and the total for 200 surveys, $7.000.000 pesos. Much less than the assigned budget.

Variables with a selection probability higher than 60% - 16 variables selected

X1 X2 X4 X8 X11 X13 X14 X16 X17 X18 X19 X24 X33 X35 X39 X44

In this case, each survey would cost $26.000 pesos and the total cost for 200 surveys would be $5.200.000 pesos.

Variables with a selection probability higher than 70% - 8 variables selected

X1 X4 X11 X13 X16 X18 X19 X39

With this set of characteristics, the cost would be $12.000 pesos a survey, and the total cost for 200 surveys would be $2.400.000 pesos.

Variables with a selection probability higher than 90% - 3 variables selected

X11 X18 X39

Finally, with this set of variables of high selection probability, the cost per survey would be $4.000 pesos and for 200 surveys would be $800.000 pesos.

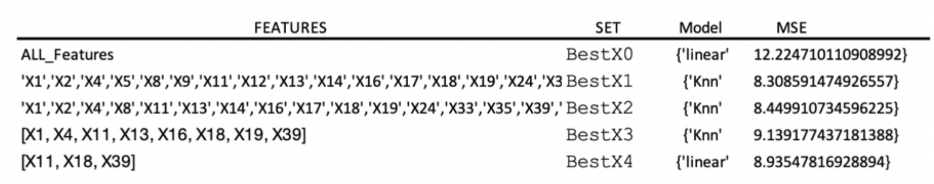
Accordingly, four additional subsets of data were built, each one containing the variables selected in every approach with the aim of looking for the best model that minimizes the mse.

3 Model Selection

Once the processes of data preparation and exploration and feature selection had finished, the modeling process was performed. Five learning models were implemented over the four abovementioned subsets: a Linear Regression, a Gradient Boosting, a Decision Tree, a Random Forest and a k-Nearest Neighbors and to evaluate the performance of every model, the test set mse of each one, was analyzed.

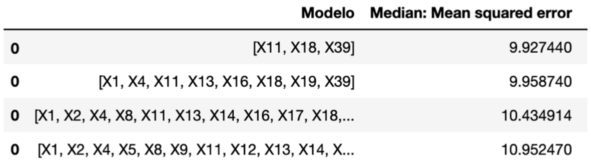
Those models were computed varying the hyper-parameters and cross-validating with k=5, and the best option of each one was picked depending on the mse.

This process was repeated 3000 times, making different partitions on the original dataset, and it was found that the best model varied significantly, but the four features set was the best one the majority of the times. Then, the four models were computed using the four different sets of variables to get a more robust estimation and the one with the best balance among the mse score, the standard deviation and the number of variables, was selected. In this case, the lowest mse values were the ones in the Linear Regression, as shown in figures 2, 3, and 4:



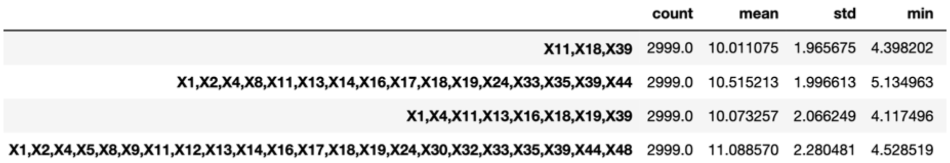
FEATURES	SET	Model	MSE
ALL_Features	BestX0	{'linear'	12.224710110908992}
'X1','X2','X4','X5','X8','X9','X11','X12','X13','X14','X16','X17','X18','X19','X24','X3	BestX1	{'Knn'	8.308591474926557}
'X1','X2','X4','X8','X11','X13','X14','X16','X17','X18','X19','X24','X33','X35','X39','	BestX2	{'Knn'	8.449910734596225}
[X1, X4, X11, X13, X16, X18, X19, X39]	BestX3	{'Knn'	9.139177437181388}
[X11, X18, X39]	BestX4	{'linear'	8.93547816928894}

Figure 2: Model Selection



	Modelo	Median: Mean squared error
0	[X11, X18, X39]	9.927440
0	[X1, X4, X11, X13, X16, X18, X19, X39]	9.958740
0	[X1, X2, X4, X8, X11, X13, X14, X16, X17, X18,...	10.434914
0	[X1, X2, X4, X5, X8, X9, X11, X12, X13, X14, X...	10.952470

Figure 3: Median MSE



	count	mean	std	min
X11,X18,X39	2999.0	10.011075	1.965675	4.398202
X1,X2,X4,X8,X11,X13,X14,X16,X17,X18,X19,X24,X33,X35,X39,X44	2999.0	10.515213	1.996613	5.134963
X1,X4,X11,X13,X16,X18,X19,X39	2999.0	10.073257	2.066249	4.117496
X1,X2,X4,X5,X8,X9,X11,X12,X13,X14,X16,X17,X18,X19,X24,X30,X32,X33,X35,X39,X44,X48	2999.0	11.088570	2.280481	4.528519

Figure 4: Standard Deviation

4 Conclusion

The best results over the test set were those gotten from the Linear Regression model with 3 variables: x11, x18 and x39. This represents a cost per survey of $4.000 pesos, so taking into account that the whole assigned budget is over $18.400.000 pesos, it would be possible to conduct a maximum of 4600 surveys in the second phase of the experiment.

It is also important to mention that if a bigger dataset is held, it is necessary to run the suggested models again and choose the one which minimizes the mse, because the sample used in this analysis is too small that could cause instability in the results.